PSI 2026 CONFERENCE



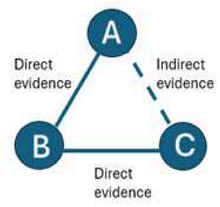
15 - 17 June 2026
ICC Belfast
Poster Gallery

P001

Network Meta Analysis (NMA) in R
Satisfying the Statistical Criteria for Health Care Decision Makers

Rebecca Tilston, Plus-Project Ltd

Introduction



- NMA is a widely used method of Indirect Treatment Comparisons (ITCs) employed by Health Technology Assessment (HTA) bodies to assess efficacy in the absence of head-to-head trials.
- Allows for comparing of multiple treatment interventions simultaneously by combining direct and indirect evidence of all studies in a network.
- Developed under both frequentist and Bayesian frameworks and can use fixed or random effects.

Objective

Objective: Using simulated survival data of a simple network, compare functionality and workflow of selected R packages in addressing key criteria recommended by NICE[1].

Description	NICE-DSU TSD7 Checklist No*	netmeta	gemtc	multinma
Description of network and network plot	A9, C3, C4.1	✓	✓	✓
Homogeneity assessment	B2, B4.1, C4.2	✓	✓	✓
Model assessment (Bayesian)	C1, B1, B2	N/A		
Specification of model priors			✓	✓
Trace plots			✓	✓ **
Model fit and convergence diagnostics			✓	✓
Leverage Plot			✓	✓
Output results	B4.1			
Forest Plot		✓	✓	✓
League table		✓	✓	✓
Treatment ranking		✓	✓	✓
Consistency assessment	C4	✓	✓	✓

*NICE Decision Support Unit Technical Support Document 7: Evidence synthesis of treatment efficacy in decision making: a reviewer's checklist. NB – similar recommendations from ISPOR[2] and PRISMA[3] can be found online
** via `bayesplot` package

R Packages

- Several R packages exist to perform NMA and two of the most widely used available on CRAN are **`netmeta`**[4] (frequentist NMA) and **`gemtc`**[5] (Bayesian NMA).
- The **`multinma`**[6] package (also available on CRAN) has been rising in popularity due to its multifaceted capabilities.
- Many other R packages exist but not all are available on CRAN, and few have as detailed documentation available as these three packages.

`netmeta`

Endpoints
Continuous, binary, survival (via logHR and SE.), count/rate (via pre-computed logRate Ratios and SE)

Data Format
Must be contrast-based, or converted from arm-based using pairwise() function

Best For
Need something simple and fast like exploratory work or sensitivity analyses. Or if standard frequentist analysis is expected.

Limitations
Handles only contrast-based data and has limited flexibility

`gemtc`

Endpoints
Continuous, binary, survival (logHR or events & exposure), count/rate

Data Format
Arm-based preferred. If contrast-data supplied user will need to manually calculate covariance for multi-arm trials

Best For
Standard Bayesian NMA or Network Meta-Regression with a well documented R package. An online user interface is also available.

Limitations
Need to install JAGS separately and can be slow. Doesn't support complex survival likelihoods.

`multinma`

Endpoints:
Continuous, binary, survival, parametric survival, count/rate, ordinal outcomes

Data Format
IPD or arm-based AgD data. Can also handle contrast-based data and a mixture of formats

Best For
Handling both IPD and AgD data, handling proper survival likelihood including parametric and spline models, and ML-NMR (Multilevel Network Meta-Regression)

Limitations
Steeper learning curve, designed for more complex data so may not be preferred for simple NMA

References

(1) Ades, A. E., D. M. Caldwell, S. Reken, N. J. Welton, A. J. Sutton, and S. Dias. 2012. "NICE DSU Technical Support Document 7: Evidence Synthesis of Treatment Efficacy in Decision Making: A Reviewer's Checklist." http://www.nicedsu.org.uk/. (2) Jansen, Jeroen P., Thomas Trikalinos, Joseph C. Cappelleri, Jessica Daw, Sherry Andes, Randa Eldessouki, and Georgia Salanti. 2014. "Indirect Treatment Comparison/Network Meta-Analysis Study Questionnaire to Assess Relevance and Credibility to Inform Health Care Decision Making: An ISPOR-AMCP-NPC Good Practice Task Force Report." Value in Health 17 (2): 157–73. https://doi.org/10.1016/j.jval.2014.01.004. (3) Hutton, Brian, Georgia Salanti, Deborah M. Caldwell, Anna Chaimani, Christopher H. Schmid, Chris Cameron, John P. A. Ioannidis, et al. 2015. "The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-Analyses of Health Care Interventions: Checklist and Explanations." Annals of Internal Medicine 162 (11): 777–84. https://doi.org/10.7326/M14-2385. (4) "Netmeta: Network Meta-Analysis Using Frequentist Methods." 2024. https://cran.r-project.org/web/packages/netmeta/index.html. (5) "Gemtc: Network Meta-Analysis Using Bayesian Methods." 2023. https://cran.r-project.org/web/packages/gemtc/index.html. (6) Phillippo DM (2024). multinma: Bayesian Network Meta-Analysis of Individual and Aggregate Data. doi:10.5281/zenodo.3904454, R package version 0.9.1, https://dmphillippo.github.io/multinma/. (7) Harrer, Mathias, Pim Cuijpers, Toshi A. Furukawa, and David D. Ebert. n.d. *Chapter 12 Network Meta-Analysis Doing Meta-Analysis in R*. Accessed April 17, 2024. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/netwma.html. (8) Schwarzer, Guido, James R. Carpenter, and Gerta Rücker. 2015. Meta-Analysis with R. Use R! Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-21416-0. (9) Rücker, Gerta, and Guido Schwarzer. 2015. "Ranking Treatments in Frequentist Network Meta-Analysis Works Without Resampling Methods." BMC Medical Research Methodology 15 (1): 58. https://doi.org/10.1186/s12874-015-0060-8.

plusproject
The Employee Owned Biostatistics CRO

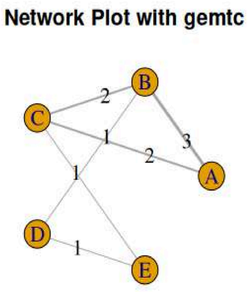
Proudly employee owned.

Data, Network and Model Specifications

Network structure: 5 treatments (A-E), 10 studies with a connected network.

Data format:

`netmeta`	Contrast-based (logHR, se(logHR) per comparison)
`gemtc`	Arm-based (events and person-time)
`multinma`	Aggregate arm-based data (could use IPD if available)



Fit the network model:

`netmeta`	Frequentist Random Effects
`gemtc` `multinma`	Bayesian Random Effects Consistency Model; Likelihood/Link: Poisson/Log; N(0,100) prior for Relative Effects and Study Baseline; half-N(0,25) prior for Heterogeneity.

- Choice of model and priors should be explicit and rationale justified[1]. Above choices are for illustrative purposes only.
- Sensitivity analyses using different prior specifications is recommended to ensure results are robust[1].
- Choice of fixed vs random effects should also be justified[1].

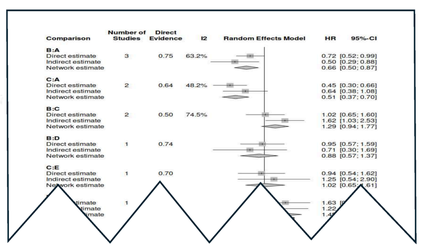
Assumptions and Model Assessment

- **Transitivity**[7,8]: The effect of a comparison e.g. A vs B is exchangeable between trials. This may be violated when covariates or other effect modifiers are not similarly distributed amongst trials in the network. **Careful examination of included studies is a critical check for acceptance of any NMA results.**
- **Homogeneity** and **Consistency**[7,8]: Can be quantified with formal tests.
  τ^2 = between-study variance
  I^2 = proportion of total variation due to heterogeneity rather than sampling error
- Inconsistency can also be assessed by examining the agreement between direct and indirect evidence (node splitting), fitting an unrelated mean effects model (UME), or with heat plots (available in **`netmeta`**).

Example from `netmeta`

```
Quantifying heterogeneity / inconsistency:
tau^2 = 0.0516; tau = 0.2271; I^2 = 61.8% [12.8%; 83.2%]

Tests of heterogeneity (within designs) and inconsistency (between designs):
                    Q d.f. p-value
Total           15.69    6  0.0155
Within designs  11.29    4  0.0235
Between designs  4.40    2  0.1110
```



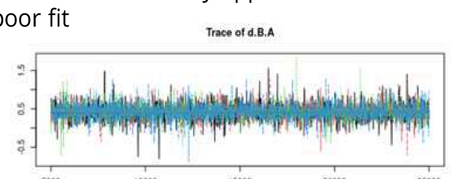
Example from `gemtc`

For Bayesian models, model fit and convergence diagnostics should be examined:

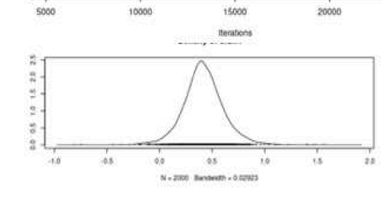
- **Model fit statistics:** Dbar (posterior mean deviance), overall leverage (pD), Deviance information criterion (DIC), Potential Scale Reduction Factor (PSRF).
- **Trace & Density Plots of Posterior Distributions (1, 2)**
- **Gelman-Rubin-Brooks Plot:** Plots PSRF over time and should slowly approach 1.0.
- **Leverage Plot**: Helps identify points leading to poor fit

	`gemtc`	`multinma`
Dbar	21.0	21.1
pD	18.7	18.6
DIC	39.6	39.7
τ posterior	0.31	0.30

(1)

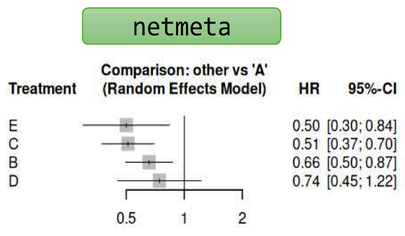


(2)

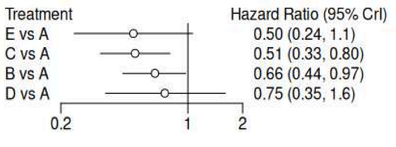


Results and Ranking

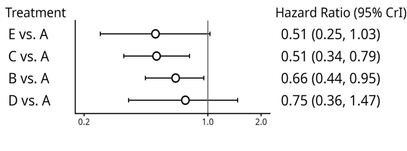
`netmeta`



`gemtc`



`multinma`



	`netmeta`	`gemtc`	`multinma`
	P-score	**SUCRA**	**SUCRA**
C	0.84	0.81	0.81
E	0.84	0.81	0.80
B	0.48	0.49	0.50
D	0.32	0.34	0.34
A	0.03	0.05	0.05

- **P-scores** (frequentist framework) and **SUCRA** (Bayesian framework) measure extent and certainty that a **treatment is better than the competing treatments**.
- Interpretation of P-scores and SUCRA have been shown to be comparable[9].

PRINTED BY SCIENCEPOSTERS .co.uk

Moaaz Sidat, Senior Statistician I

Bayesian Building Blocks: The Key to Successful Validation

With the release of the new draft FDA guidance[1], the use of Bayesian methodology in clinical trials is expected to increase over the next few years. However, implementation and validation can often become a cumbersome process.

Issues with ambiguity in the analysis plan and longer run times compared to traditional frequentist analyses can lengthen the validation process. Due to the nature of Bayesian analyses, reconciliation of differences between production and QC can become complex and add strain to timelines and the biostatistics team. Here we discuss tips to efficiently and successfully validate Bayesian analyses.

Successful Validation of Bayesian Analyses

Clear Analysis Plan Specification

Operating Characteristics
Number of samples, Burn-in, Thinning, weight of borrowing

Model Specifications
Prior, Likelihood, Posterior, covariates to include, posterior probabilities to calculate

Specific Code Syntax
If a specific option/ procedure/software is required

The analysis plan should also state whether tables must be validated to an exact match (by specifying the seed to use during sampling) or may match within defined criteria. Where criteria are used, they should apply to every value in the table obtained from the analysis.

For example, requiring the posterior mean difference to match within 0.01

Outline Convergence Strategies

Due to the nature of the data collected, instances may occur when convergence may not be achieved in the specified analysis. Outlining strategies to carry out allows swift programming without ambiguity, for example:

- Model adaptations (e.g. remove a covariate)
- Increase number of burn in
- Increase the number of samples

Because non-convergence may occur for a variety of reasons, the analysis plan cannot define solutions for all scenarios. Predefined steps are therefore implemented to encourage convergence before additional investigation.

Step by Step Process

Prior to executing the analysis, divide the analysis into several intermediate steps as outlined below:

Bayesian Analysis

Preprocessing Data → **Sampling Procedure(s)** → **Post Processing**

By breaking the workflow into smaller, modular components such as pre-processing, sampling, and post-processing reviewers can rerun only the sections relevant to their checks rather than re-executing the full analysis each time. This allows early detection of issues in upstream steps before investing time in full analysis runs.

Reduced Iterations During Validation Process

Initially reducing the number of iterations during the production and validation process can significantly reduce computational burden and decrease run times. Especially when program updates are required from mismatch resolution, the run times can prolong the validation process.

Conduct analysis on reduced number of samples | **Issue resolution and validation** | **Conduct analysis on SAP specified samples** | **Validation on SAP specified samples**

Conclusion

- Well-specified, detailed plans reduce ambiguity and support confident application of Bayesian analyses.
- Initial runs with fewer iterations reduce computational burden and speed up diagnostics.
- Partitioning analyses into steps enables early issue identification and smoother validation.
- Pre-defined methods for diagnosing and addressing non-convergence can reduce the need for further investigation.

Together, these building blocks can build confidence within biostatistics teams to effectively conduct Bayesian analyses.

[1] U.S. Food and Drug Administration. (2026). Use of Bayesian methodology in clinical trials of drug and biological products: Draft guidance for industry. www.fda.gov/regulatory-information/search-fda-guidance-documents/use-bayesian-methodology-clinical-trials-drug-and-biological-products

Veramed®
veramed.com

P003

Individual Patient Data (IPD) Network Meta Analysis (NMA) without IPD sharing

Federico Bonofiglio, PhD[1]

[1]Cytel Inc. Freiburg, Germany

Introduction

Network meta-analysis (NMA) extends conventional meta-analysis by jointly evaluating multiple treatments and combining both direct and indirect evidence to estimate relative efficacy or the best treatment among many. In NMA, a new key assumption for indirectly evaluating head-to-head comparisons that were not directly observe is the so-called *consistency* equation,

$$BC = BA - CA,$$

where the unobserved comparison (contrast), BC, is indirectly derived from the observed ones, BA and CA. Here, A, B, and C are three different interventions, and A is a common comparator.

Key NMA Assumptions and Violations

Consistency can only hold if BA and CA trials are sufficiently comparable, e.g., if they target the same patients' population. This assumption is violated if *effect-modification* exists in some trials. For instance, this occurs if in any trial the BA effect is *heterogeneous* between population's subgroups and such heterogeneity is not equally distributed across trials. Crucially, consistency must hold in any single multi-arm randomized controlled trial with A, B, C arms. This begs the trivial but so far unexplored question: ***Can we compare NMA to an idealized multi-arm trial (IMT)?*** This question is interesting for both modeling and particularly modeling without access to Individual Patient Data (IPD) which is a common situation in practice.

IMT Set-up with Effect Modification (i.e. Moderation)

Consider a large IMT (N = 10000) with 3 arms (randomly allocated 1:1:1) and fixed $ATE(\theta)=ATT(\theta)=\sum_j \theta[Z=j]\times P(Z=j)/\sum_j P(Z=j)$, where $\theta[Z=j]$ is the true fixed contrast in subgroup $Z=j$ with trial (population) prevalence $P(Z=j)=0.5$, and $\theta=\{BA, CA\}$, $j=1,2$. That is, ATE is fully driven by the proportion, $P(Z=1)$, of a two-levels moderator Z. The outcome for the ith patient is Normal, $y_i \sim N(\mu_i, \sigma)$, with

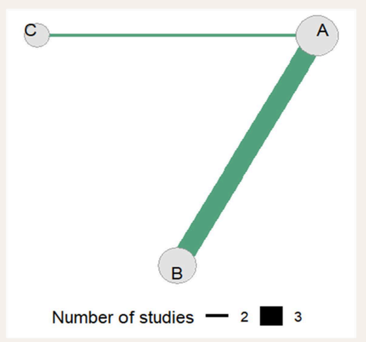
$$\mu_i = y_0 + \beta x_i + I(a_i = \theta\text{-}A)\times \sum_j \theta[Z=j]\times I(Z_i=j)$$

where y_0=grand mean, x_i=positive prognostic variable, e.g. age, with effect β, a_i=B, C is the trial arm, $I(\cdot)$=binary 0-1 indicator, and σ=residual error, i=1,2,..N. Finally, $BA[Z=2]=-20$, $CA[Z=2]=-10$ and $BA[Z=1]=CA[Z=1]=0$ such that **BA=-10**, **CA=-5** and, by design, **BC = -5**. Can we compare NMA to this IMT?

NMA Set-up from Triangular "V"-Shaped Network

The simplest network compatible with IMT is "V"-shaped: 3 interventions (A, B, C). The BA and CA comparisons are directly observed (nodes are connected in graph), But BC has no direct connection (comparison) and can only be estimated indirectly. That is, NMA is here very sensitive to bias for consistency violations. We simulate IPD from such network under varying inconsistency scenarios by varying the distribution of moderator Z, $P(Z=1)$, in some of the studies. Consequently, here ATE ≠ ATT. Otherwise, all other settings and parameter are equal to IMT and to BA study nr. 1, the reference/target study, where $P(Z=1)=0.5$.

Network Graph



Network Settings for Different Inconsistency Scenarios

Parameter	Value
Total Nr. of studies	5
Nr. of BA studies	3 (1 study is fixed reference where $P(Z=1)=0.5$ always)
Nr. of CA studies	2
Θ, β, *Z*, x	as in IMT (residual error can vary between studies)
No inconsistency	$P(Z=1)=0.5$ in all studies
Mild inconsistency	$P(Z=1)=0.8$ in one BA study otherwise $P(Z=1)=0.5$
High inconsistency	$P(Z=1)=0.8$ in 1 BA and 2 CA studies otherwise $P(Z=1)=0.5$
Sample size N	Small ($N=10$), medium ($N=10^2$), large ($N=10^3$)

NMA Estimation for Increasing Inconsistency Levels (none, mild, high)

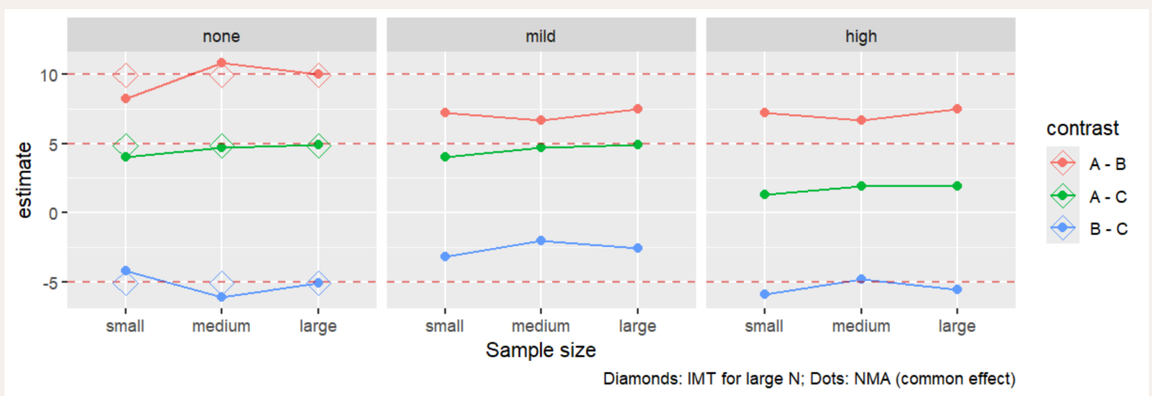


Figure 1. NMA model adjusts for know prognostic factor, x, but purposedly ignores moderator *Z*. Broken line: true contrast value.

Question 1 (Figure 1): can we liken NMA to IMT ? Yes !

- If the network is internally consistent (no inconsistency), we can think of it as an IMT !
- In our "V"-shaped network, even mild inconsistency generates bias.
- That is, improving network consistency is key to obtain coherent direct and indirect estimates
- To this end, we include *Z* in the model and compare different methods:
- 2sNMA=2-stage NMA[1], ML-NMR=Multilevel Network Meta Regression[2], IPW=Inverse Probability Weighting[3].
- All methods account for both x and *Z*, but IPW does so by rebalancing distributions across studies (both ATE and ATT is possible), while the others do standard regression adjustment (ATE only).

ATE Estimation for Increasing Inconsistency After Accounting for Moderator *Z*

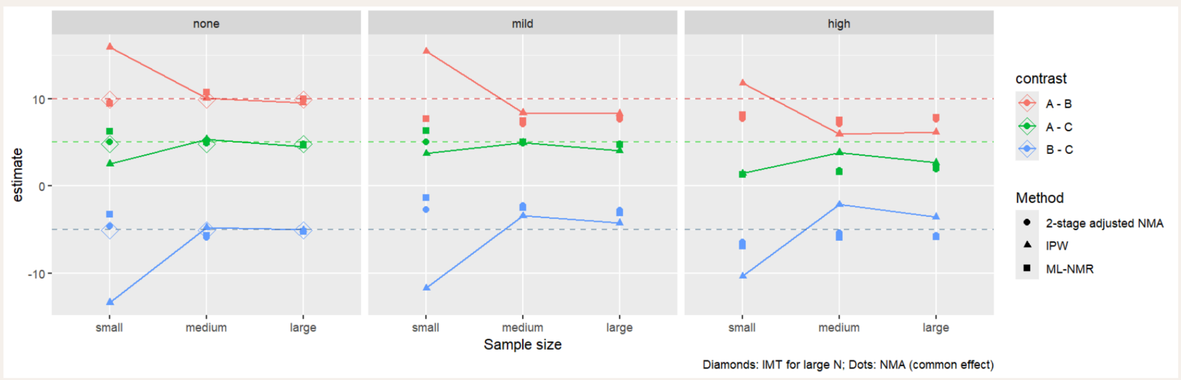


Figure 2. IPD NMA now adjusts for both x and *Z* using different methods under comparisons. Solid line: IPW. Broken line: true contrast value.

ATE, ATT, and IPD Availability. Pseudo IPD Resampling via a Gaussian Copula

- Network is consistent relative to a target population. Here, this is IMT where $P(Z=1)=0.5$.
- Because ATE≠ATT with inconsistency, ATE only targets a "trials-average" where $P(Z=1)\neq 0.5$.
- That is, ATE cannot target IMT (Figure 2), but ATT can do so. We now target ATT with IPW if: IPD is available, IPD-AGD=IPD of study nr. 1 and aggregate data (AGD) otherwise, AGD only.
- AGD is converted into pseudo IPD using GC=Gaussian Copula[4] to mimic a IPD NMA using IPW.
- GC resamples pseudo IPD. Bagging (e.g., the mean) is used to get a final contrast estimate.
- IPW targets ATT well (Figure 3) but GC is biased if the network is inconsistent.
- GC "reconstructs" IPD via simple IPD summaries including linear-correlations (good anonymization), but this misses the underlying *Z*-clustering explaining treatment heterogeneity.

ATT Estimation with Different Levels of IPD Availability using Gaussian Copula IPW

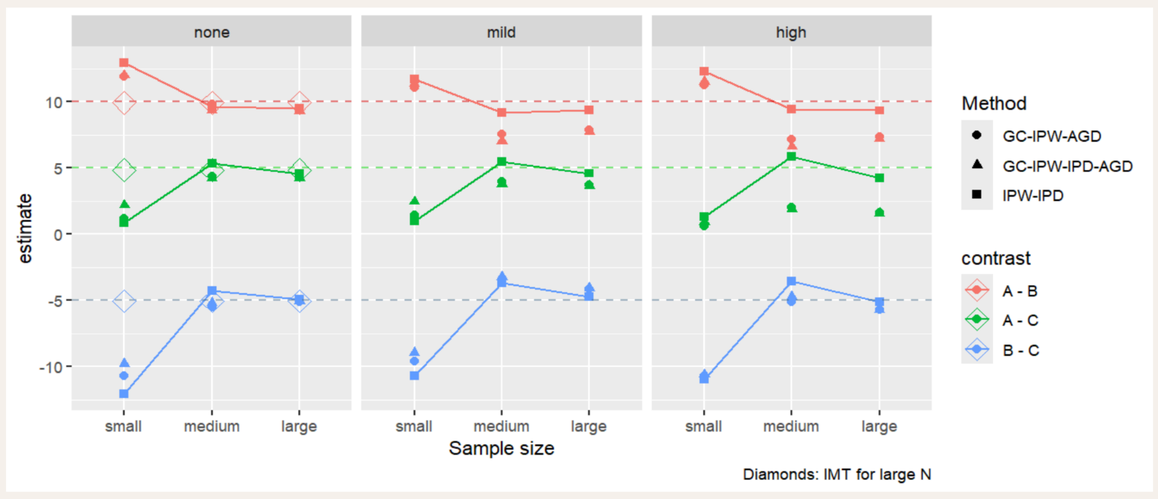


Figure 3. IPW NMA using IPD (solid line), IPD-AGD= AGD with only IPD from reference study nr.1, AGD=AGD only. Broken line: true contrast value.

Question 2 (Figure 3): can we target IMT if IPD is not available? Not so easily yet !

- NMA can be seen as an idealized multi-arm trial, if network consistency holds.
- Inconsistency makes ATE and ATT unequal. IPW can remove inconsistency and target ATT (under usual provisions, e.g., positivity, no residual confounding).
- IPW requires IPD and lacking IPD complicates estimation!
- If AGD available, IPW based on GC (do pseudo IPD from AGD) can be biased if network is inconsistent.
- Other anonymization approaches than GC must be investigated.
- This work must be extended to more complex scenarios (larger, disconnected networks).

References

[1] Riley RD, et al. BMJ Evidence- Based Medicine. 2023; 28:197-203

[2] Phillippo DM, et al. Jour Royal Statistical Society A. 2020; 183(3):1189-1210

[3] McCaffrey DF, et al. Statistics in Medicine. 2013; 32: 3388-3414

[4] Bonofiglio F, et al.. Statistics in Medicine. 2020; 39:1183-1198

Legend

ATE = Average Treatment Effect

ATT = ATE in the Treated

GC-IPW-AGD = IPW based on GC and AGD

GC-IPW-IPD-AGD = GC-IPW with one reference IPD study and AGD otherwise

IPW-IPD = IPW with IPD.

Federico Bonofiglio

federico.bonofiglio@cytel.com

P005

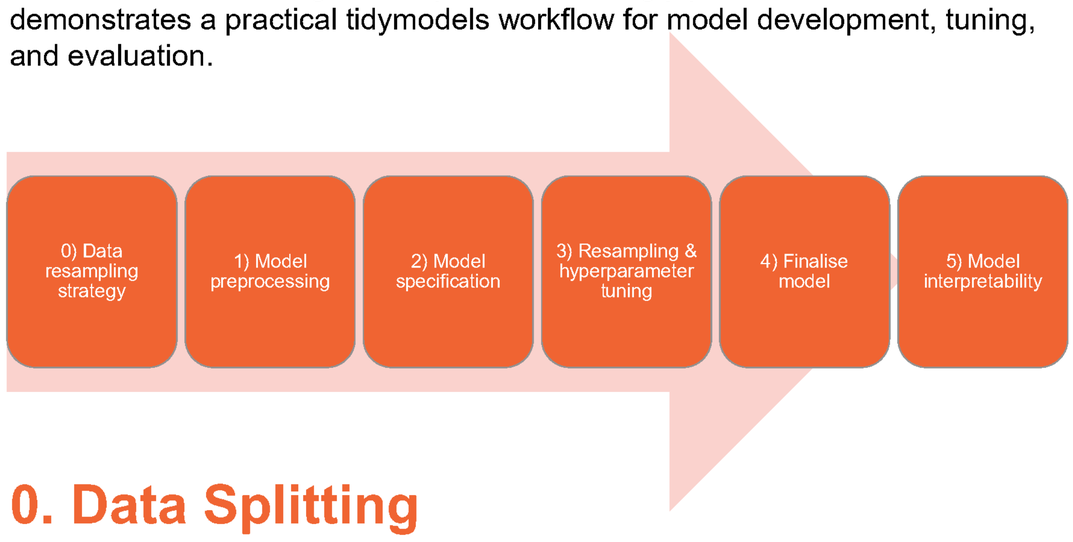
Jason Nicholas, Harry Parr
GSK – London (United Kingdom)

PSI Conference, ICC, Belfast, 14th – 17th June 2026

A Tutorial on {tidymodels}: Building Reproducible Modelling Workflows in R

Introduction

Tidymodels is an R ecosystem designed to support reproducible, end to end predictive modelling. It formalises the separation of data preprocessing, model specification, resampling, and evaluation, while allowing these components to be combined into a single, transparent modelling pipeline. This tutorial demonstrates a practical tidymodels workflow for model development, tuning, and evaluation.



0. Data Splitting

The first stage is splitting the data into the training and testing datasets using the "*initial_split*" function. The test set is held out and used only for final model evaluation, supporting an honest assessment of predictive performance.

```
splits <- initial_split(df, prop = .8)
training_data <- training(splits)
testing_data <- testing(splits)
cv_folds <- vfold_cv(training_data , strata = OUTCOME, v=5, repeats = 10)
cv_folds_INTERNAL <- vfold_cv(df, strata = OUTCOME, v=5, repeats = 10)
```

1. Preprocessing via Recipes

The recipes package defines structured preprocessing steps that are applied consistently across resampling and final model fitting. Typical steps include normalisation, encoding of categorical variables, and handling of missing data. Preprocessing is specified independently of the model and estimated using training data only.

```
recipe <-
  recipe(OUTCOME ~ ., data = df) %>%
  update_role(STUDYID, SUBJID, new_role = "IDs") %>%
  step_zv(all_predictors()) %>%
  step_rm(REGIMEN) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_impute_bag(all_predictors(),
                  impute_with = imp_vars(all_predictors())) %>%
  step_normalize(all_numeric_predictors()) %>%
  step_naomit(OUTCOME)
```

2. Model Specification

Models are defined independently of the data, with hyperparameters either fixed or marked for tuning using "*tune()*", allowing multiple model configurations to be compared within a common framework.

```
mod_rf <-
  rand_forest(
    mtry = tune(),
    min_n = tune(),
    trees = 1000
  ) %>%
  set_engine("ranger") %>%
  set_mode("classification")
```

```
mod_logistic <-
  logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")
```

Workflow Creation

```
wf_set <- workflow_set(
  preproc = list(recipe),
  models = list(
    random_forest  = mod_rf,
    logistic       = mod_logistic
  )
```

A workflow combines the preprocessing steps and model specification into a unified object. This ensures that the same preprocessing and
model structure is applied consistently during resampling, tuning, and final fit.

3. Hyperparameter Tuning

Hyperparameters marked with "*tune()*" are optimised using resampling techniques such as cross-validation, or bootstrapping. This allows us to systematically evaluate different parameter combinations and identify the optimal hyperparameter configuration and select the model that achieves the best performance.

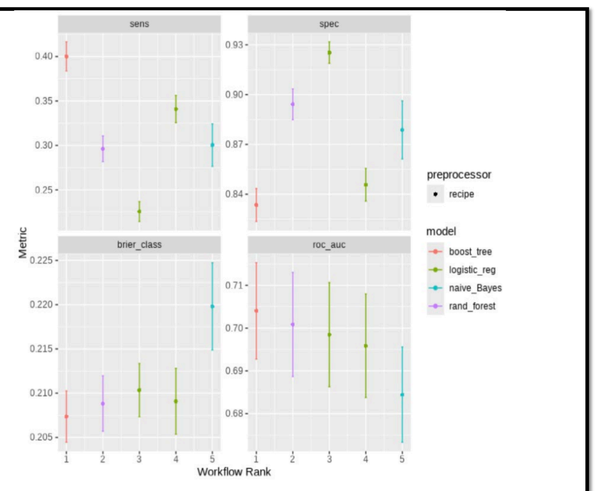
```
rf_param <-
  mod_rf %>%
  extract_parameter_set_dials() %>%
  update(mtry = mtry(c(2, 30)))

rf_grid <- grid_latin_hypercube(
  rf_param,
  size = 200
)

results <- wf_set %>%
  workflow_map(
    "tune_grid",
    resamples = cv_folds,
    metrics = metric_set(roc_auc, sens, spec, brier_class)
  )

best_parameters <- results %>%
  select_best(metric = "roc_auc")

autoplot(results, select_best=TRUE)
```



4. Fitting the Final Model

After tuning, the best-performing hyperparameters are selected and used to finalise the workflow. The final model is then fitted to the full training dataset, prior to evaluation on the held-out test set.

```
rf_wf_updated <- workflow() %>%
  add_model(mod_rf) %>%
  add_recipe(recipe)

rf_wf_final <- finalize_workflow(rf_wf_updated, best_parameters)

rf_fit <- fit(rf_wf_final, data = training_data)
```

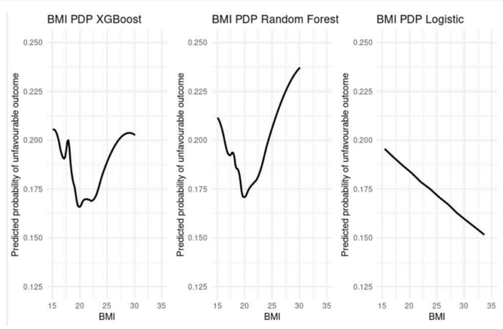
5. Model Interpretability

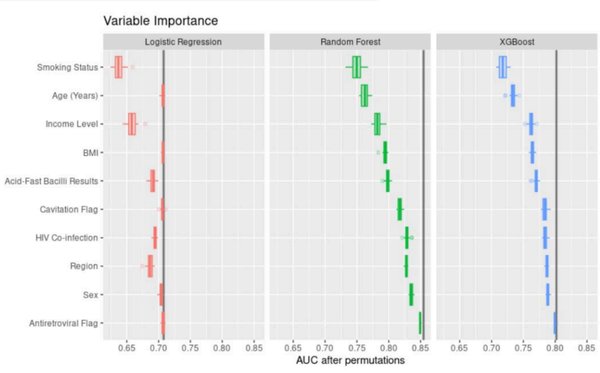
We evaluate the final model on the test set using appropriate performance metrics. Model diagnostics and interpretability tools such as variable importance plots and partial dependence plots can be used to better understand model behaviour and support transparent reporting.

```
preds <- predict(rf_fit, new_data=test_data, type = "prob") %>%
  bind_cols(predict(rf_fit, test_data, type = "class"))

roc_auc(preds, truth = OUTCOME, .pred_1)
conf_mat(preds, truth = OUTCOME, .pred_class)

roc_auc_curve <- roc_curve(preds, OUTCOME, .pred_1)
autoplot(roc_auc_curve)
```





Summary

Tidymodels provides a cohesive, end-to-end ecosystem for building, tuning, and validating predictive models in R, with consistent workflows that strengthen reproducibility and evaluation rigor.

Ahead Together

P006

Structural Equation Models to Model Digital Measures, Patient-Centered Outcomes, and Clinical Variables: Illustration Using SPIROMICS Data

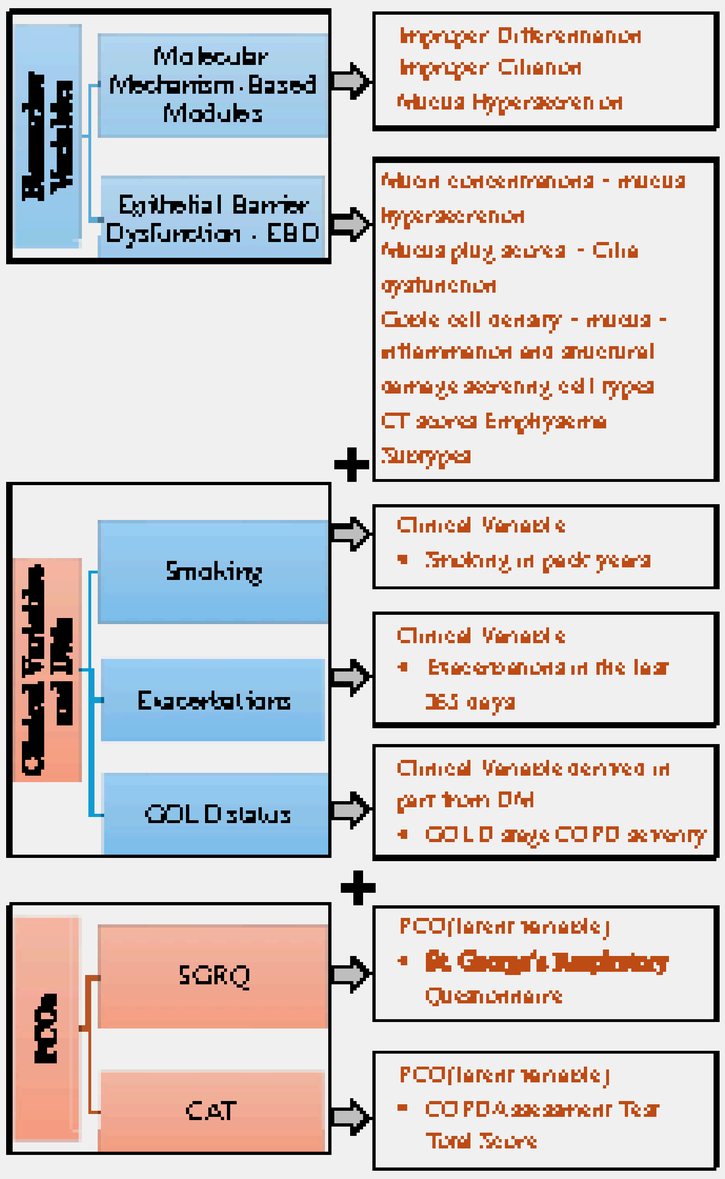
Prathiba Batley, Adetayo Kasim, Rashmi Kulkarni, Andres Quintero Moreno, Jonathan Yu, Nil Turan
GSK

PURPOSE

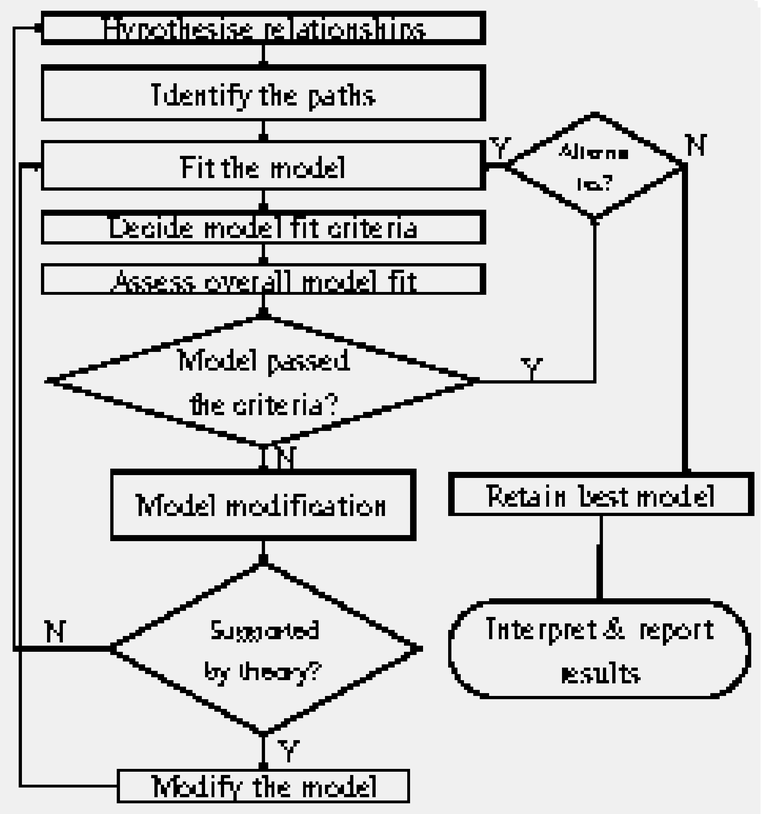
Using Structural Equation Modelling (SEM) to model relationships amongst Digital Measures, Patient-Centered Outcomes, biomarker variables, and clinical variables.

MOTIVATION

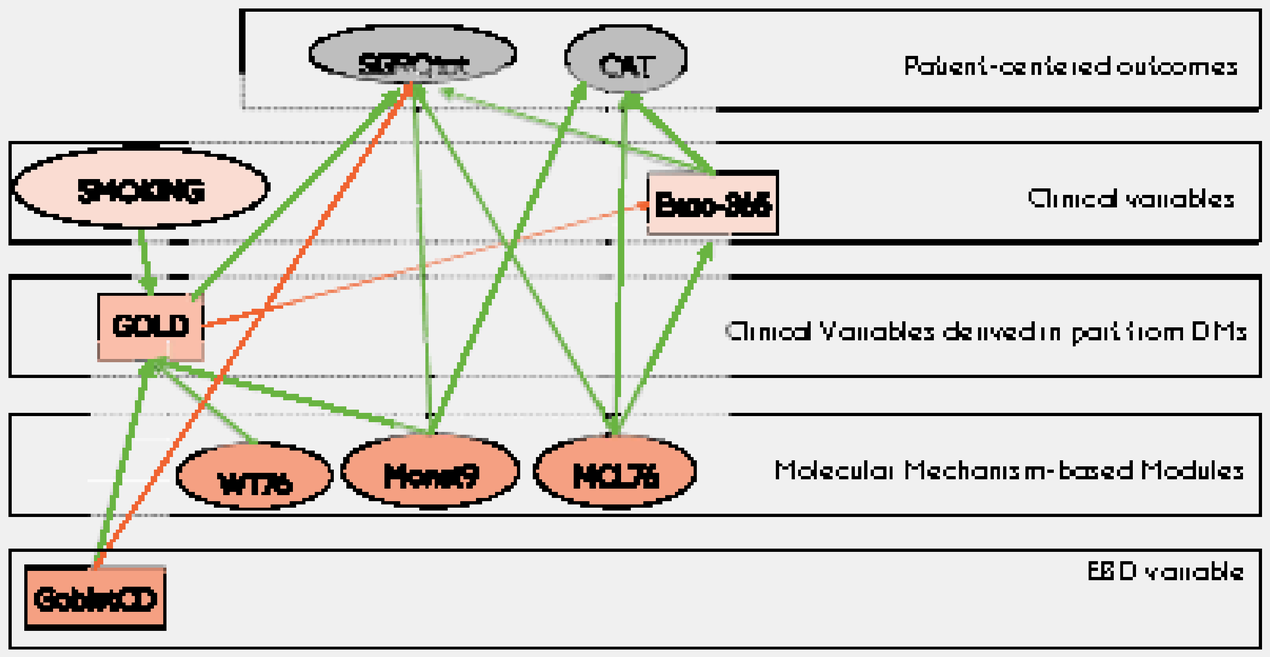
Multiple interdependent variables (digital measures, PCOs, biomarkers) demand an integrated model — not isolated regressions. Illustration using SPIROMICS data.



DATA FROM DIVERSE SOURCES INTEGRATED IN ONE SEM



SEM MODEL BUILDING FLOWCHART



Excellent Model fit criteria: CFI ≥ 0.95, RMSEA ≤ .05, and SRMR ≤ 0.08

Excellent model fit: CFI = 1.000 (n = 58); $\chi^2 = 5.77, df = 11$ **($p = 0.888$);** RMSEA = 0.000 [0.0, 0.068]; SRMR = 0.053

Statistically significant effects are in green. Measured variables are in rectangles. Latent variables are in ellipses.

FINAL MODEL RETAINED

Direct and Indirect Effects in the Final Model

Dependent Variable	Predictor	Direct Effect	Indirect Effect	R^2
Exac_365				**14.40%**
	GOLD	0.17 (0.14)		2.72%
	z_MCL_76	-0.28 (0.11)*		4.48%
	z_SMOKING		0.07 (.06)	1.12%
	z_GCD		0.04 (.03)	0.64%
	z_Walktrap_76		-0.18 (.17)	2.88%
	z_MONET_9		0.16 (.16)	2.56%
GOLD				**45.90%**
	z_MONET__9	0.97 (0.47)*		16.55%
	z_Walktrap_76	-1.06 (0.48)*		18.17%
	z_SMOKING	0.43 (0.1)*		7.41%
	z_GCD	0.22 (0.1)*		3.77%
L_SGRQ_TOTALSCORE				**36.10%**
	Exac_365	1.4 (0.49)*		3.34%
	z_MCL_76	-4.75 (2.23)*	-0.39 (0.21)	10.43%
	z_MONET_9	5.35 (2.23)*	0.72 (0.5)	14.52%
	GOLD	0.75 (0.37)*		1.79%
	z_GCD	0.47 (0.34)	0.16 (0.11)	1.5%
	z_Walktrap_76		-0.79 (0.53)	1.89%
	z_SMOKING		0.32 (0.17)	0.77%
CATSCORE				**31.10%**
	Exac_365	4.6 (1.1)*		4.57%
	z_MCL_76	-12.79 (4.92)*	-1.29 (0.61)*	11.43%
	z_MONET_9	12.61 (4.93)*		12.54%

DISCUSSION

- Integrated model identifies that a molecular biomarker (MCL76) influences patient **symptom score (CAT ↓) both directly and via exacerbations (mediated effect)**
- Exacerbation count bridges molecular and clinical outcomes as mediator
- SEM allows simultaneous estimation of such multi-step relationships between latent and measured variables while accounting for measurement error

Keywords: Structural Equation Model, Respiratory, Mediation Effects, Latent Variables

GSK

P008

Effectiveness of Cannabidiol in Patients With Rare Epilepsies Compared to External Placebo Control: A Post Hoc Analysis From the Expanded Access Program

Teresa Greco,[1] Yong Park,[2] Kelly C Simontacchi,[3] Timothy B Saurer,[3] Elizabeth A Thiele[4]

[1]Jazz Pharmaceuticals, Inc., Gentium Srl, Villa Guardia, Italy; [2]Medical College of Georgia at Augusta University, Augusta, GA, USA; [3]Jazz Pharmaceuticals, Inc., Palo Alto, CA, USA; [4]Massachusetts General Hospital, Boston, MA, USA.

Objective

- To describe the baseline demographics and clinical characteristics, effectiveness, and safety of CBD among EAP participants with rare epilepsy syndromes including DEEs other than LGS, DS, and TSC
- To compare the effectiveness of CBD treatment in participants with rare epilepsy syndromes to an ECA

Conclusions

- CBD treatment is associated with reductions in convulsive and total seizure frequencies in participants with DEEs and rare epilepsies beyond LGS, DS, and TSC
- The comparative effectiveness analysis confirms superiority of CBD compared with an ECA (placebo) in reducing seizure frequency from baseline
- This analysis supports further investigation of CBD effectiveness for a broad range of DEEs and other rare epilepsies

First presented at the American Epilepsy Society Annual Meeting, 2025.

References: 1. Devinsky O, et al. *Epilepsy Behav*. 2018;86:131–137. **2.** US Food and Drug Administration. Epidiolex® Prescribing Information. 2025. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2025/210365s023lbl.pdf. (Accessed August 14, 2025.) **3.** Thiele EA, et al. *Lancet*. 2018;391(10125):1085–1096. **4.** Devinsky O, et al. *N Engl J Med*. 2018;378:1888–1897. **5.** Hernán MA, Robins JM. *Am J Epidemiol*. 2016;183(8):758–764. **6.** Chesnaye NC, et al. *Clin Kidney J*. 2022;15(1):14–20. **7.** Austin PC. *Multivariate Behav Res*. 2011;46(3):399–424.

Support and Acknowledgements: This analysis was sponsored by Jazz Pharmaceuticals, Inc. The authors would like to thank all of the study investigators, study staff, nursing team, patients, and caregivers for their participation in this research. Writing and editorial assistance were provided to the authors by Judy Eun, PharmD, on behalf of Syneos Health, and funded by Jazz Pharmaceuticals, Inc., in accordance with Good Publication Practice (GPP) 2022 guidelines.

Disclosures: All authors met the ICMJE authorship criteria and had full access to relevant data. Neither honoraria nor payments were made for authorship. **YP** has consulted for, conducted studies funded by, or received honoraria for services provided to Jazz Pharmaceuticals, Inc.; **TG, KCS,** and **TBS** are employees of Jazz Pharmaceuticals, Inc., and hold stock and/or stock options in the company. **EAT** has consulted for, conducted studies funded by, or received honoraria for services provided to Jazz Pharmaceuticals, Inc.

Epidyolex® is approved in the EU for the adjunctive treatment of seizures associated with Lennox-Gastaut syndrome or Dravet syndrome, in conjunction with clobazam, in patients ≥2 years of age; it is additionally approved in the EU for the adjunctive treatment of seizures associated with tuberous sclerosis complex in patients ≥2 years of age. Epidiolex® is approved in the US for the treatment of seizures associated with Lennox-Gastaut syndrome, Dravet syndrome, or tuberous sclerosis complex in patients ≥1 year of age.

Presented at the PSI 2026 Conference; June 15–17; Belfast, Ireland

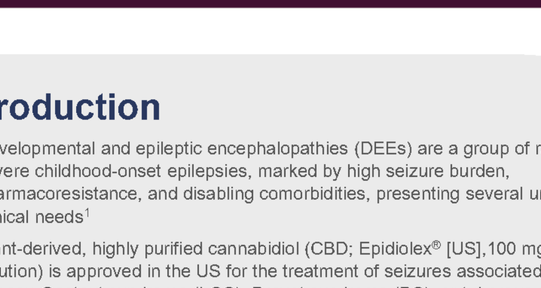
Scan this code to access this poster online.
This code is not for promotional purposes.

Introduction

- Developmental and epileptic encephalopathies (DEEs) are a group of rare and severe childhood-onset epilepsies, marked by high seizure burden, pharmacoresistance, and disabling comorbidities, presenting several unmet clinical needs[1]
- Plant-derived, highly purified cannabidiol (CBD; Epidiolex® [US],100 mg/mL oral solution) is approved in the US for the treatment of seizures associated with Lennox-Gastaut syndrome (LGS), Dravet syndrome (DS), or tuberous sclerosis complex (TSC) in patients ≥1 year of age[2]
- Real-world evidence suggests that CBD may have utility in other DEEs, such as Aicardi syndrome, *CDKL5* deficiency disorder, Dup15q syndrome, and epilepsy with myoclonic-atonic seizures (EMAtS)[1]
- The CBD Expanded Access Program (EAP) provided compassionate access to CBD for patients with diverse treatment-resistant epilepsies, including DEEs
- This post hoc analysis of the US CBD EAP evaluates seizure and safety outcomes in participants with rare and severe childhood-onset epilepsies, compared with an external placebo control arm (ECA)[2–4]

Methods

Figure 1. Study design: Descriptive analysis of CBD EAP and comparative analysis of CBD EAP vs ECA



[a]Participants received CBD (Epidiolex®) doses starting at 2–10 mg/kg/day and titrated up as tolerated or to a maximum of 25–50 mg/kg/day at the discretion of the investigator and institutional review board at each EAP site.
[b]All participants enrolled in both CBD EAP and ECA had treatment-resistant epilepsy and were receiving stable doses of antiseizure medications for ≥4 weeks before enrolment.
[c]Convulsive and total seizures.
[d]Participants with a minimum frequency of 8 convulsive seizures per 28 days during baseline were included in the analysis.
CBD, cannabidiol; DEEs, developmental and epileptic encephalopathies; DS, Dravet syndrome; EAP, Expanded Access Program; ECA, external placebo control arm; GI-C, Global Impression of Change; LGS, Lennox-Gastaut syndrome; PBO, placebo; RCTs, randomised controlled trials; RR, responder rate; TSC, tuberous sclerosis complex.

Descriptive analysis: CBD EAP cohort

- Participants with rare and severe childhood-onset epilepsies—including those diagnosed with DEE, genetic epilepsy, or a rare epilepsy syndrome—were included in the CBD EAP analysis cohort (**Figure 1**). Participants with a diagnosis or co-diagnosis of LGS, DS, or TSC were excluded from this cohort
  - Rare epilepsy subgroups including Aicardi syndrome, *CDKL5* deficiency disorder, Dup15q syndrome, early infantile epileptic encephalopathy (EIEE), EMAtS, febrile infection–related epilepsy syndrome (FIRES), myoclonic absence epilepsy, *SCN2A*-DEE, Sturge-Weber syndrome (SWS), and other genetic epilepsies were analysed
    - Rare epilepsies with <5 participants were pooled into an "other genetic epilepsies" subgroup
- Effectiveness of CBD was assessed as percentage reduction from baseline in median 28-day convulsive and total seizure frequencies at Weeks 12 and 48
- Responder rates (RRs; ≥50%, ≥75%, and 100% seizure reduction) were reported at Weeks 12 and 48
- Physician and Caregiver Global Impression of Change (GI-C) data were reported at Weeks 12 and 48
- Safety results were reported for the entire treatment period (median duration, 617 days)

Comparative analysis: CBD EAP vs ECA

- A comparative effectiveness analysis was conducted using an ECA to evaluate the effectiveness of CBD in reducing the 28-day average number of convulsive or total seizure frequencies during the first 14 weeks of treatment (**Figure 1**)
  - Participants with a seizure frequency of <8 convulsive seizures per 28 days during the baseline period were excluded
  - The ECA was composed of participants with LGS who were randomised to placebo in the GWPCARE3 and GWPCARE4 randomised controlled trials
  - To minimise information bias, comparability was verified in terms of eligibility criteria, treatment strategy, assignment procedure, follow-up period, measures at baseline, and endpoints[5]
- Inverse probability weighting, based on propensity scores,[6,7] was applied to a log-transformed ANCOVA model to minimise the association between treatment exposure and study participation and balance key baseline covariates, including age, sex, convulsive seizure frequency at baseline, number of concomitant antiseizure medications (ASMs), and current clobazam use
  - Assuming that no other confounding variables were associated with both the response variable and the treatment, the average treatment effect of CBD compared with placebo was estimated by performing a weighted log-transformed analysis of covariance model
  - Balancing performance was investigated using a standardised weighted mean differences plot
  - Absence of hidden bias was confirmed through the Rosenbaum sensitivity analysis
- This study used data on Epidiolex®, and the results of this post hoc analysis do not apply to other CBD-containing products

Results

Baseline demographics and clinical characteristics of CBD EAP participants with rare epilepsies

- The overall analysis set included 135 participants in the CBD EAP effectiveness analysis set with rare epilepsies (**Table 1**)
- The mean age across subgroups ranged from 5.0 to 14.4 years, and in the overall population, the most commonly used ASMs at baseline were clobazam (40%), levetiracetam (33%), and valproic acid (30%)
- Median (Q1, Q3) total daily CBD dose was 25 mg/kg/day (15.0, 25.0) at Week 12 and 25 mg/kg/day (20.0, 27.2) at Week 48

Table 1. Participant demographics and baseline antiseizure medications (CBD EAP effectiveness analysis set)

	Overall	Aicardi syndrome	*CDKL5*	Dup15q syndrome	EIEE	EMAtS	FIRES	Myoclonic absence	*SCN2A*	SWS	Other genetic epilepsies[a]
N (%)	135	17 (12.6)	20 (14.8)	11 (8.1)	6 (4.4)	11 (8.1)	13 (9.6)	5 (3.7)	5 (3.7)	5 (3.7)	42 (31.1)
Age, years, Mean (SD)	9.7 (6.0)	10.2 (5.8)	6.8 (4.2)	12.4 (4.9)	6.0 (6.1)	9.5 (3.4)	9.4 (5.6)	14.4 (4.4)	5.0 (2.6)	8.9 (6.1)	11.2 (7.1)
Sex, female, n (%)	87 (64.4)	17 (100)	19 (95.0)	5 (45.5)	6 (100)	3 (27.3)	7 (53.8)	2 (40.0)	3 (60.0)	4 (80.0)	21 (50.0)
Number of ASMs, mean (SD)	2.8 (1.2)	2.5 (1.0)	2.5 (0.8)	3.0 (0.5)	3.5 (1.1)	2.8 (0.9)	4.5 (1.7)	—	2.8 (0.8)	2.6 (1.3)	2.5 (1.2)
ASMs, n (%)											
Clobazam	54 (40)	9 (53)	6 (30)	5 (46)	2 (33)	5 (45)	7 (54)	—	1 (20)	1 (20)	18 (43)
Levetiracetam	45 (33)	5 (29)	6 (30)	4 (36)	3 (50)	2 (18)	10 (77)	—	2 (40)	2 (40)	11 (26)
Valproic acid	40 (30)	5 (29)	10 (50)	5 (46)	0	2 (18)	3 (23)	—	0	3 (60)	12 (29)
Topiramate	13 (10)	0	4 (20)	1 (9)	1 (17)	0	3 (23)	—	1 (20)	1 (20)	2 (5)
Lamotrigine	28 (21)	7 (41)	5 (25)	4 (36)	0	3 (27)	1 (8)	—	1 (20)	0	7 (17)
Rufinamide	24 (18)	4 (24)	2 (10)	6 (55)	1 (17)	4 (36)	2 (15)	—	1 (20)	1 (20)	3 (7)
Felbamate	9 (7)	0	0	0	0	3 (27)	3 (23)	—	0	1 (20)	2 (5)
Seizure frequency, median (Q1, Q3), [n]											
Convulsive seizures	42 (16, 104) [100]	47 (18, 74) [14]	52 (25, 100) [18]	58 (5, 209) [11]	170 (29, 2950) [4]	61 (32, 116) [9]	60 (24, 284.0) [5]	4 (4, 4) [1]	25 (13, 112) [5]	1 (1, 1) [1]	40 (4, 98) [32]
Total seizures	80 (28, 294) [135]	57 (24, 109) [17]	95 (42, 231) [20]	89 (55, 272) [11]	205 (18, 384) [6]	86 (32, 404) [11]	216 (60, 4000) [13]	2800 (1750, 2800) [5]	112 (74, 138) [5]	2 (1, 6) [5]	72 (20, 217) [42]

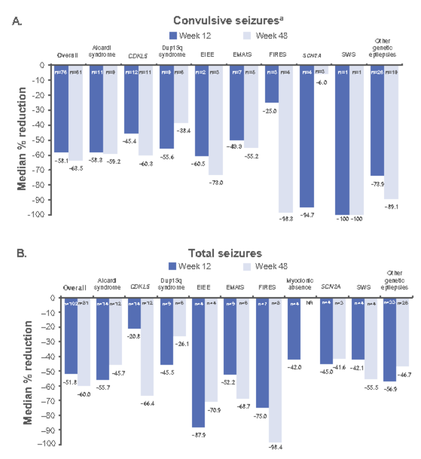
[a]Included rare epilepsies affecting <5 participants, had diagnostic terms: 15q11.2 deletion, 1p36 deletion, 22q11 duplication, *ADNFLE-DEPDC5* mutation, Angelman syndrome, *BRAF* mutation, *CACNA1E* mutation, *CLCN4* mutation, *COL4A1* mutation, deletion 16 (p13, 11p13.1), Disomy X, *Dup3p* mutation, *DYRK1A* mutation, EIMFS, EOEE, GLUT1 deficiency, *GPHN* mutation, *GRIN2A* mutation, *HCN2* mutation, infantile spasms, Jeavons syndrome, *KCNA1* mutation, *KCNQ2* mutation, LYPB2 haplotype deficiency, MECP2 duplication, Rasmussen syndrome, Rett syndrome, *SCN1A* mutation, *SCN8A* mutation, *SNAP25* mutation, sodium channelopathy, *SYNGAP1* mutation, Trisomy 21, and *ZDHHC9* mutation.
— not applicable; ASMs, antiseizure medications; CBD, cannabidiol; *CDKL5*, cyclin-dependent kinase-like 5; EAP, Expanded Access Program; EIEE, early infantile epileptic encephalopathy; EMAtS, epilepsy with myoclonic-atonic seizures; FIRES, febrile infection-related epilepsy syndrome; Q1, first quartile; Q3, third quartile; SWS, Sturge-Weber syndrome.

Effectiveness results

Seizure frequency and Responder rates (CBD EAP effectiveness analysis set)

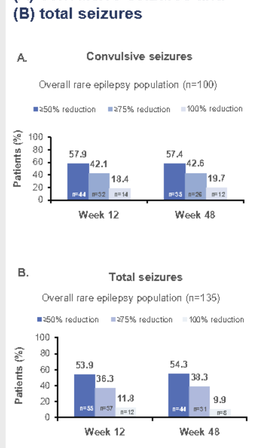
- At Week 12, participants experienced a median percent reduction in seizure frequency of 58.1% and 51.8% in convulsive and total seizures, respectively; this median reduction from baseline was maintained through Week 48 for convulsive seizures (63.5%) and total seizures (60.0%) (**Figure 2A and 2B**)
- At Week 12 and Week 48, responder rates (≥50%, ≥75%, and 100%) were:
  - **Convulsive seizures in overall population:** 57.9% and 57.4%; 42.1% and 42.6%; and 18.4% and 19.7%, respectively (**Figure 3A**)
  - **Total seizures in overall population:** 53.9% and 54.3%; 36.3% and 38.3%; and 11.8% and 9.9%, respectively (**Figure 3B**)

Figure 2. Median percentage reduction from baseline in seizure frequencies in (A) convulsive seizures and (B) total seizures



[a]In one participant with myoclonic absence epilepsy, no change in convulsive seizure frequency was observed at Week 12. Week 48 data were not reported.
CBD, cannabidiol; *CDKL5*, cyclin-dependent kinase-like 5; EAP, Expanded Access Program; EIEE, early infantile epileptic encephalopathy; EMAtS, epilepsy with myoclonic-atonic seizures; FIRES, febrile infection-related epilepsy syndrome; NR, not reported; SWS, Sturge-Weber syndrome.

Figure 3. Treatment response rates (≥50%, ≥75%, 100%) in (A) convulsive seizures and (B) total seizures



Global Impression of Change (CBD EAP safety analysis set)

- In the overall CBD EAP group, at Weeks 12 and 48, patient/caregiver-reported GI-C improvement (slightly improved, much improved, very much improved) was reported in 73.7% and 82.0% of participants, respectively, and physician-reported GI-C improvement (slightly improved, much improved, very much improved) was 72.2% and 82.1%, respectively (**Figure 4**)

Figure 4. GI-C ratings by patients/caregivers and physicians in the overall CBD EAP group



CBD, cannabidiol; EAP, Expanded Access Program; GI-C, Global Impression of Change.

Comparative effectiveness analysis of CBD EAP vs ECA

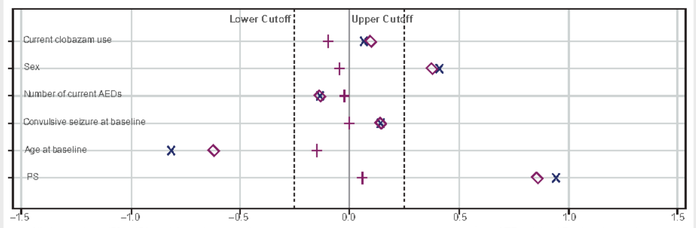
- Comparative effectiveness analysis included 79 of 135 CBD EAP participants and 152 of 161 ECA participants; baseline characteristics and demographics were balanced based on the standardised weight mean difference (**Table 2** and **Figure 5**)

Table 2. Baseline characteristics: Comparative effectiveness analysis set

	Unweighted		Weighted[a,b]	
	CBD EAP	ECA	CBD EAP	ECA
N, Effectiveness analysis set	79	161	79	152
Age, years, Mean (SD)	9.2 (4.7)	15.3 (9.5)	11.1 (8.3)	12.2 (8.9)
Sex, female, n (%)	52 (65.8)	74 (46.0)	52 (51.2)	72 (53.3)
Number of ASMs, mean (SD)	2.8 (1.2)	2.9 (0.9)	2.9 (2.0)	2.9 (1.2)
ASMs, n (%)				
Clobazam	36 (45.6)	79 (49.1)	36 (55.0)	77 (50.3)
Levetiracetam	27 (34.2)	58 (36.0)	27 (30.1)	56 (37.2)
Valproic acid	23 (29.1)	63 (39.1)	23 (32.7)	60 (38.3)
Lamotrigine	18 (22.8)	56 (34.8)	18 (28.1)	53 (33.8)
Rufinamide	18 (22.8)	41 (25.5)	18 (27.9)	38 (23.5)
Convulsive seizure frequency, mean (SD)	328.1 (1002.1)	219.7 (409.4)	260.2 (1417.5)	260.3 (613.6)
Total seizure frequency, mean (SD)	626.9 (2187.7)	437.4 (713.5)	522.7 (3206.2)	519.1 (1003.8)

[a]Summary statistics are weighted for propensity score derived by balancing the average distribution of age, sex, convulsive seizure frequency at baseline, number of concomitant ASMs, and current use of clobazam.
[b]To avoid large standard error for causal effect estimates, the analysis used a support region of all observations having a percentile from 2 to 98 of the propensity score distribution. The support region included 79 CBD EAP and 152 ECA participants with balanced baseline characteristics.
ASM, antiseizure medication; CBD, cannabidiol; EAP, Expanded Access Program; ECA, external placebo control arm.

Figure 5. Standardised weighted mean difference for covariate balance assessment[a]

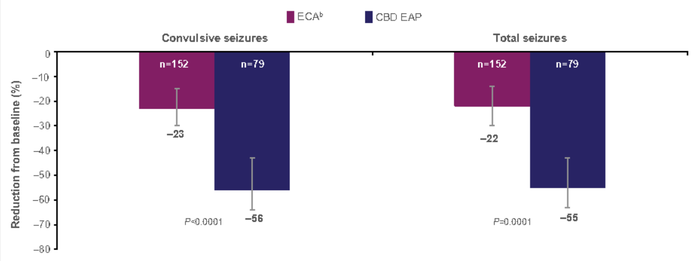


[a]Summary statistics are weighted for PS derived by balancing the average distribution of age, sex, convulsive seizure frequency at baseline, number of concomitant ASMs, and current use of clobazam. Differences (SWMD) were calculated to show the difference in means of covariates between treated and control groups, standardised by the pooled SD, for each balancing covariate. A value SWMD <0.1 is acceptable; SWMD between 0.1 and 0.25 is not alarming; SWMD >0.25 denotes a serious imbalance.
[b]To avoid large standard error for causal effect estimates, the analysis used a support region of all observations having a percentile from 2 to 98 of the propensity score distribution. The support region included 79 CBD EAP and 152 ECA participants with balanced baseline characteristics.
AEDs, antiepileptic drugs; ATE, average treatment difference; PS, propensity score; SWMD, standardised weighted mean difference.

Comparative effectiveness analysis of CBD EAP vs ECA

- CBD groups showed greater reductions in convulsive and total seizure frequency compared with the ECA group (**Figure 6**)
- CBD resulted in a 43% greater seizure reduction compared with ECA ($P \leq 0.0001$; **Table 3**)

Figure 6. Comparative analysis of percent reduction from baseline to Week 14 in convulsive and total seizures in the CBD EAP vs ECA groups[a]



[a]To avoid large standard error for causal effect estimates, the analysis used a support region of all observations having a percentile from 2 to 98 of the propensity score distribution. The support region included 79 CBD EAP and 152 ECA participants with balanced baseline characteristics.
[b]Assessment at RCT end of treatment represents the 28-day average number of seizures during the first 14 weeks. The ECA was composed of participants with LGS who were randomised to placebo in the GWPCARE3 and GWPCARE4 RCTs.
CBD, cannabidiol; EAP, Expanded Access Program; ECA, external placebo control arm; LGS, Lennox-Gastaut syndrome; RCT, randomised controlled trial.

Table 3. Comparative analysis of percentage reductions from baseline and treatment ratios

	Reduction from baseline (95% CI)	Treatment ratio % (95% CI)	*P* value
Convulsive seizures			
ECA	23% (15%–30%)	43% (25%–57%)	<0.0001
CBD EAP	56% (43%–66%)		
Total seizures			
ECA	22% (14%–30%)	43% (24%–56%)	0.0001
CBD EAP	55% (43%–65%)		

ECA was considered the reference group.
Percentage reductions from baseline and treatment ratios were calculated as (1-estimate) x 100.
CBD, cannabidiol; EAP, Expanded Access Program; ECA, external placebo control arm.

Safety results (CBD EAP safety analysis set)

- Treatment-emergent adverse events (TEAEs) were reported by 85.9% (134/156 [CBD EAP safety analysis set]) of participants in the overall population (**Table 4**)
  - Serious TEAEs occurred in 50.6% of participants; 3.2% of participants experienced treatment-related serious TEAEs
  - Deaths (n=11, 7.1%) were investigator-assessed as unrelated to CBD treatment
- The most common treatment-related AEs that occurred in ≥10% of participants in the overall population were diarrhoea (18.6%), fatigue (12.8%), and somnolence (12.2%)
- The most common TEAEs that occurred in ≥20% of participants in the overall population were convulsion (26.9%) and diarrhoea (28.2%)
- TEAEs leading to CBD discontinuation occurred in 13.5% of participants in the overall population
  - The most frequently reported TEAEs leading to treatment discontinuation in the overall population were convulsion and diarrhoea (both 2.6%)

Table 4. Summary of adverse events

	Overall[a]	Aicardi syndrome	*CDKL5*	Dup15q syndrome	EIEE	EMAtS	FIRES	*SCN2A*	SWS	Other genetic epilepsies
Participants, n (%)	156 (100)	19 (12.2)	23 (14.7)	9 (5.8)	6 (3.8)	11 (7.1)	20 (12.8)	5 (3.2)	6 (3.8)	50 (32.1)
Any TEAEs	134 (85.9)	19 (100)	21 (91.3)	9 (81.8)	6 (100)	9 (81.8)	18 (90.0)	4 (80.0)	6 (100)	42 (84.0)
Any TRAEs	81 (51.9)	14 (73.7)	11 (47.8)	9 (81.8)	3 (50.0)	7 (63.6)	3 (15.0)	2 (40.0)	6 (100)	26 (52.0)
TEAEs leading to CBD discontinuation	21 (13.5)	2 (10.5)	3 (13.0)	2 (18.2)	2 (33.3)	1 (9.1)	2 (10.0)	1 (20.0)	1 (16.7)	7 (14.0)
Serious TEAEs	79 (50.6)	12 (63.2)	12 (52.2)	6 (54.5)	3 (50.0)	4 (36.4)	12 (60.0)	3 (60.0)	4 (66.7)	23 (46.0)
Treatment-related serious AEs	5 (3.2)	0	2 (8.7)	1 (9.1)	0	0	1 (5.0)	0	0	1 (2.0)
Deaths[b]	11 (7.1)	1 (5.3)	0	0	2 (33.3)	0	2 (10.0)	1 (20.0)	0	5 (10.0)
TRAEs reported in ≥10% of participants in any group by preferred term, n (%)										
Diarrhoea	29 (18.6)	5 (26.3)	4 (17.4)	2 (18.2)	1 (16.7)	1 (9.1)	2 (10.0)	0	1 (16.7)	13 (26.0)
Fatigue	20 (12.8)	4 (21.1)	5 (21.7)	3 (27.3)	1 (16.7)	2 (18.2)	1 (5.0)	0	1 (16.7)	3 (6.0)
Somnolence	19 (12.2)	2 (10.5)	3 (13.0)	3 (27.3)	1 (16.7)	4 (36.4)	0	0	0	6 (12.0)
Decreased appetite	11 (7.1)	1 (5.3)	3 (13.0)	3 (27.3)	0	1 (9.1)	1 (5.0)	0	0	2 (4.0)
Sedation	8 (5.1)	2 (10.5)	0	0	1 (16.7)	0	0	1 (20.0)	0	4 (8.0)
Abnormal behaviour	7 (4.5)	2 (10.5)	1 (4.3)	1 (9.1)	0	1 (9.1)	0	0	2 (33.3)	0
Convulsion	6 (3.8)	1 (5.3)	1 (4.3)	0	0	1 (9.1)	1 (5.0)	0	2 (33.3)	0
Liver function test abnormal	6 (3.8)	0	1 (4.3)	2 (18.2)	0	0	0	0	1 (16.7)	2 (4.0)
Vomiting	5 (3.2)	2 (10.5)	1 (4.3)	1 (9.1)	0	0	0	0	0	1 (2.0)
Aggression	4 (2.6)	2 (10.5)	0	0	0	0	0	0	0	2 (4.0)
Ataxia	4 (2.6)	0	0	3 (27.3)	0	0	0	0	0	1 (2.0)
Weight decreased	4 (2.6)	2 (10.5)	1 (4.3)	1 (9.1)	0	0	0	0	0	0
Weight increased	2 (1.3)	1 (5.3)	0	0	0	0	0	0	1 (16.7)	0
Secretion discharge	2 (1.3)	1 (5.3)	0	0	1 (16.7)	0	0	0	0	0
Seizure cluster	1 (0.6)	0	0	0	0	0	0	0	1 (16.7)	0
Lethargy	1 (0.6)	0	0	0	1 (16.7)	0	0	0	0	0
Haematemesis	1 (0.6)	0	0	0	1 (16.7)	0	0	0	0	0
Blood alkaline phosphatase increased	1 (0.6)	0	0	0	1 (16.7)	0	0	0	0	0
Throat irritation	1 (0.6)	0	0	0	0	0	0	1 (20.0)	0	0
Ocular hyperaemia	1 (0.6)	0	0	0	0	0	0	0	1 (16.7)	0
Strabismus	1 (0.6)	0	0	0	0	0	0	0	1 (16.7)	0

[a]Adverse event data was not available for the participants with myoclonic absence epilepsy (n=5; 3.2%).
[b]All deaths were considered unrelated to CBD treatment by the respective investigators. The 11 AEs that were reported among participants who died were respiratory failure (n=2), pulmonary haemorrhage (n=1), asphyxia (n=1), respiratory arrest (n=1), acute sinusitis (n=1), convulsion (n=1), ear infection (n=1), foot deformity (n=1), disease progression (n=1), and death (n=1).

Limitations

- A key limitation is the potential for design-related bias from comparing populations in separate studies. The ECA included only LGS participants, whereas the EAP had diverse aetiologies, limiting direct comparability. Limitations of the EAP data reporting may also have an impact on the descriptive results

P009

A new multivariate framework for outlier detection of questionnaire in clinical trial with Hierarchical Ordinal Longitudinal Mixed Effect Simulation Model



Introduction

Background and Challenges

Ordinal outcomes are among the most common types of variables collected in questionnaires used in clinical trials. These outcomes may be derived from patient-reported questionnaires, interviewer-administered assessments, or clinician- and researcher-evaluated endpoints, and are frequently recorded using ordinal scales. Compared with binary or count variables, ordinal variables contain richer information and may better capture treatment responses. However, statistical methodologies for analyzing longitudinal ordinal outcomes remain relatively limited because of the complexity involved in modeling large numbers of covariates, multilevel dependence structures, and repeated measurements over time.

Proposal of methods

We propose a new multivariate simulation-based framework for detecting abnormal longitudinal questionnaire patterns at the center level using hierarchical ordinal mixed-effect modeling and trajectory-based anomaly statistics.

Suitable scenarios

patient-reported outcome (PRO) questionnaires,
symptom severity scales,
quality-of-life instruments,
clinician-reported ordinal assessments and etc.

Methodological Overflow

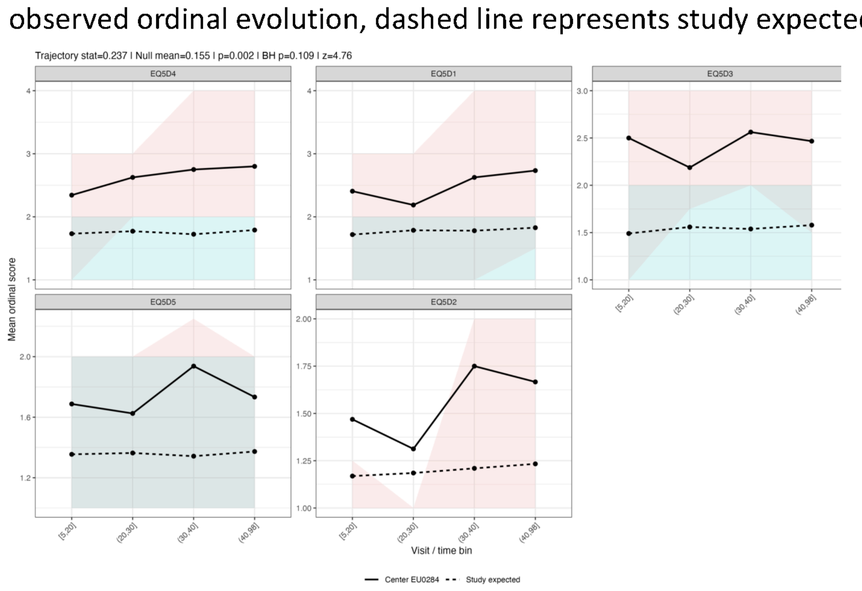
Key Innovations

The multivariate hierarchical ordinal longitudinal simulation framework integrates:

- cumulative ordinal mixed-effect models per variable with random center and patient effects
- Spline effect on time variable such as visit or date
- longitudinal AR(1) dependence per ordinal variable over time
- Gaussian copula residual dependence,
- Value evolution trajectory based Gower distance as anomaly statistics,
- Study expected ordinal evolution recoded through Monte Carlo Simulation
- Outlier detection with hypothesis testing

Real Use Case 1

Eq5d dataset from an example clinical trial (quality of life)
5 questions in the questionnaire with 4 ordinal scales
The plot shows the rank 1 flagged center, solid line represents center observed ordinal evolution, dashed line represents study expected



Statistical Framework

Model Definition

- The cumulative logits model with random centre and patient intercept is defined as: For center $_c$, patient $_i$, question $_j$, and visit $_t$

$$Y_{\{cijt\}} = X_t\,\beta_j + b_{\{cj\}}^{center} + b_{\{ij\}}^{patient} + Z_{\{tij\}}$$

$$P(Y_{\{cijt\}} \le k) = logit^{\{-1\}}\,(\alpha_{\{jk\}} - Y_{\{cijt\}})$$

where: Y is the latent variable (Gaussian) and K is the ordinal scale (categorical), $X_t\beta_j$ represents longitudinal fixed effects (which includes potential covariates and spline time effect on visit/date, $b_{cj}^{(center)}$ represents center random intercepts, $b_{ij}^{(patient)}$ denotes for patient and Z_{tij} represents the model residuals (which will be used for modeling the inter and intra dependence.

Random effects

- For the random effects, we assume they follow a multivariate dependence structure to model the Center-level dependence across questionnaire items:

$$b_c^{center} \sim N(0, \sigma_{center}^2\,R_{item})$$

$$b_i^{patient} \sim N(0, \sigma_{patient}^2\,R_{item})$$

- where R_{item} is the cross-question covariance matrix

Residual structures

- If we only fit a separate model on each single questions, the model will overlook all the potential relationship between questions or same questions over time.
- Therefore, we further assume the null residual structure with autocorrelation over time/visit and also covariance between questions at the same visit/time:

$$Z_{\{tj\}} = \varphi_j Z_{\{t-1,j\}} + \sqrt{1-\varphi_j^2}\,E_{\{tj\}}$$

- with:

$$E_{\{tj\}} \sim \mathcal{N}(0, R_{resid}) \qquad \text{and} \qquad R_{resid} = \mathrm{Cor}(Residuals)_{Spearman}$$

- Item-specific temporal dependence is estimated using lagged Spearman correlation:

$$\phi_j = \mathrm{Corr}\left(Y_{ijt}, Y_{ij,t-1}\right)$$

P-Value Generation

Gaussian Copula Simulation

- The latent Gaussian residual process is transformed through a Gaussian copula:

$$U_{tj} = \Phi\left(Z_{tj}\right)$$

- Ordinal outcomes are then generated using inverse cumulative probabilities:

$$Y = \min\left\{k: \sum_{m=1}^{k} p_m \ge U\right\}$$

normalized trajectory deviation

- For each patient-question-time combination, the normalized trajectory deviation is defined as:

$$D_{cijt} = \frac{|\,Y_{cijt} - Y_{-c,jt}\,|}{R_j}$$

- where:

$$R_j = \max\left(Y_{\cdot\cdot j\cdot}\right) - \min\left(Y_{\cdot\cdot j\cdot}\right)$$

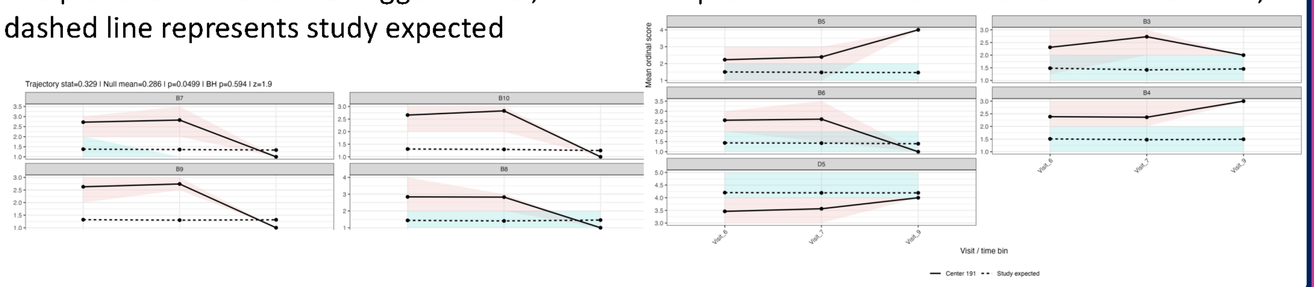
is the observed ordinal range for item $_j$.

- $0 \le D_{cijt} \le 1$

Monte Carlo null datasets are subsequently generated while preserving ordinal marginal distributions, multivariate item dependence, temporal autocorrelation, and center/patient random-effect heterogeneity. For each simulated dataset, the center-level trajectory statistic is recalculated to form an empirical null distribution. Final center-specific p-values are obtained by comparing the observed trajectory abnormality statistic against the simulated null distribution

Real Use Case 2

Patient filled ePRO dataset from an example clinical trial (clinical severity)
9 questions in the questionnaire with 5 ordinal scales
The plot shows the rank 1 flagged center, solid line represents center observed ordinal evolution, dashed line represents study expected



CluePoints

Lawson Wang, Laura Trotta

Isabelle Smith, Manager Statistics
Co Authors – Jenny Scott, Emily Wood, Rhys Warham

Arm Wrestling with Uncertainty: Meta-Analysis for Survival Estimation

Meta-analysis offers a practical framework for estimating control arm overall survival, but methodological choices matter. Careful handling of heterogeneity and uncertainty is essential to support robust Phase 3 trial design and reduce the risk of underpowered studies.

Introduction

- Accurate sample size calculations are essential for **robust Phase 3 clinical trial design,** but they rely on multiple assumptions, ideally informed by **prior evidence.**
- One important assumption for a randomised controlled trial is the **expected outcome in the control group,** for example the median overall survival when this is the primary endpoint.
- In many oncology settings:
  - Placebo control is unethical.
  - No established standard of care exists.
  - **Investigator's choice** comparators are used, **increasing uncertainty** in control arm overall survival estimation.
- Meta-analysis of existing evidence may increase confidence in assumptions, however traditional approaches often **under-represent heterogeneity** and underestimate uncertainty, increasing the risk of **underpowered trials.**

Objective

To evaluate and compare meta-analytic approaches for estimating control arm median overall survival in the absence of a standard treatment.

Methodological Challenges

Outcome
- Median overall survival is more clinically relevant than hazard ratio.
- Survival data are skewed.

Temporal drift
Improvements in overall survival over calendar time, as treatments emerge can bias naïve pooling.

Between-study heterogeneity
Differences in population, treatment, and follow-up.

Uncertainty quantification
Standard errors for median overall survival estimates are rarely reported.

Summary of meta-analytic methods for estimating median overall survival for control arm when there is no standard treatment

	Summary	Benefits	Limitations
Naïve pooling	• Participant data treated as single study • No adjustment for between-study variability	• Simple to implement • Enables comparison with historical meta-analyses conducted in the same way	• Ignores heterogeneity • Overly precise estimates • Larger studies may dominate estimates
Random-effects meta-analysis of median overall survival	• Inverse-variance weighted DerSimonian & Laird estimator • Between-study variance estimated via REML • Study-specific SEs derived using the Siud et al. asymptotic approach	• Precedence for random effects model • Modelling study as a random effect allows between study variability to be included, as well as within study variability • Studies are weighted according to their precision, including sampling variability	• Assumes normality, however this may be violated • Instability due to censoring in small studies
Weighted median approaches	Weighted median of the study-specific medians as the pooled median estimate where the weights are proportional to the number of subjects in the study and normalised to sum to 1	• No assumptions required • Can be implemented using R's METAMEDIAN	• Ignores heterogeneity • Limited inferential outputs (Forest plots would be based on raw medians only)
Distribution-free approach based on survival curves (Combescure et al)	Product-limit estimator is expanded to obtain a distribution-free survival curve Extension of DerSimonian and Laird methodology applied to account for heterogeneity between studies	• Fully non-parametric • Conceptually appealing	• Computationally intensive • Requires extensive digitisation and interpolation • Sensitive to sparse data

Practical Recommendations

Random-effects meta-analysis of median overall survival estimates are preferred due to ability to account for between and within study variability, adjustment for study precision, and computationally less intensive than distribution-free approaches.

Quantify heterogeneity explicitly.

Use naïve pooling only as a sensitivity analysis.

Reflect uncertainty in sample size calculations through scenario planning.

DerSimonian R & Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986; 7: 177-88 | Siud EV, et al. A comparison of reflected versus test-based confidence intervals for the median survival time, based on censored data. Biometrics. 1984; 40(3): 587-600 | Combescure, C., et al. Meta-analysis of single-arm survival studies: a distribution-free approach for estimating summary survival curves with random effects. Statistics in Medicine. 2014; 33(15); 2521-37

Veramed®
veramed.com

Avie-Lee Tillotson, Principal Statistician

Double Trouble: Tackling Expectations, Challenges & AI in Double Programming

Balancing quality, efficiency, and innovation in clinical trial programming.

ICH E9 emphasises the importance of reproducibility and pre-specification in clinical trial analysis.

Double programming is the independent re-implementation of statistical analyses using different programming approaches and is the long-established gold standard quality control process used in clinical trials. In GXP-regulated clinical trial environments, it serves three core purposes:

Reproducibility – Verifying that results hold regardless of the implementation route

Error Detection – Catching coding mistakes before they reach a regulatory submission

Independent Validation – providing documented assurance that the analysis is correct

Despite its status as a gold standard and its value in error detection and reproducibility being well established, double programming practical implementation raises important questions around efficiency, methodology, and relevance, especially as data science tools and AI continue to evolve.

Challenges

AI Generated Code

What would a validated, audit-ready AI-assisted double programming workflow look like in practice?

AI tools can generate analysis-ready code rapidly, but introduce key challenges:

Core Risks

- **Validation:** Who verifies AI-generated output?
- **Reproducibility:** How are prompts, model versions, and code tracked for audit?
- **Accountability:** Responsibility cannot be delegated to AI.

Threat?

Unvalidated code increases the risk of undetected errors.

- How is code validated, and to what standard?
- Who is accountable when errors occur?

Principle – All AI outputs must be reviewed, validated, and signed off by a named individual.

Tool?

AI is most effective as an assistive tool, for example:

- Generating first-pass code from SAP
- Highlighting discrepancies across implementations

Control Point

Humans retain review responsibility and ownership.

Teammate?

AI can be a supportive collaborator including:

- Debugging
- Logic checks against SAP

Outcome

Enhanced productivity while preserving human judgement and expertise.

Multi-Platform Complexity

What is "validation" when platforms implement the same model differently?

Organisations using SAS alongside R or Python face version control, environment management, and package dependency challenges. A 'match' between SAS PROC MIXED and R's lme4 may require deliberate alignment of REML defaults, covariance structures, and degrees-of-freedom methods – none of which is automatic.

Potentially

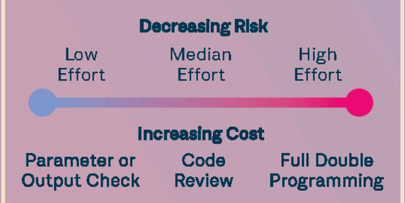
Focus double programming on matching **statistical intent, estimands, and clinical conclusions**, allowing documented, acceptable numerical differences.

Resource & Efficiency Demands

Is double programming always the best use of limited resources and the most efficient safeguard?

Double programming can consume 30–50% more programming resource on a study. In competitive bid environments this is a significant cost driver, creating pressure to cut corners or treat double programming as a tick-box exercise. The time cost is real and must be planned for, not absorbed.

Volume of Effort vs. Risk Tolerance



Redundancy vs Fragility

Double programming provides intentional redundancy. While redundancy is often labelled inefficient, it reduces fragility: systems without independent verification are more vulnerable to undetected error. Decisions to reduce double programming should explicitly recognise the loss of this protective layer.

Over Reliance on Output Matching

Is the target of double programming "matching outputs" or "matching statistical conclusions"?

When 'matching outputs' becomes the goal rather than 'independent verification of statistical logic', programmers may reverse-engineer each other's code rather than working truly independently. This defeats the purpose of the process entirely. The diversity of thought is the quality mechanism – removing it removes the value matching.

What 'Match' Should Really Mean

- Alignment matters more than identical numbers.
- Same conclusion, not the same decimals
- Acceptable differences – variation is not failure
- Detect errors, align methodology
- Matching can hide shared mistakes

SAP Specification Depth

How much programming detail should SAPs contain ?

If a SAP specifies every step in granular detail, programmers follow instructions rather than exercising independent judgement. If it is too high-level, grey areas appear and programmers fill them differently, creating reconciliation problems. The right level of specification is a genuine tension – too much removes critical thinking; too little creates inconsistency.

Details vs. Critical Thinking

How much detail do we need in our SAPs?

A SAP that is **specified at code-level** detail removes scope for independent judgement: the **critical thinking is done once,** and programmers are left to follow a recipe rather than verify the analysis. At the extreme, double programming such a SAP **assesses consistency of execution, not analytical correctness.**

Conversely, **under-specified** SAPs force interpretive decisions. Divergent interpretations increase reconciliation effort; identical but incorrect interpretations allow errors to pass undetected.

Regulators

Same standards just new tools.

Accuracy. Reproducibility. Traceability.

The Question That Matters: "Can we stand behind this output?"

AI changes how work is done → not how it's judged

The Tacit Knowledge Problem

Not documented. Not searchable. Learned over time:

AI doesn't have the lived experience of statisticians or programmers who've gained experience through success and failure in real projects.

Future Direction

AI doesn't replace thinking it rewards precision.

Work is shifting. From execution → to clear specification.

Thinking moves upfront.

The future isn't coder driven, it's spec driven.

Some Thoughts for Debate & Conversation

- Double programming is still the right instinct, but often for the wrong reasons.
- Output matching is a convenience metric, not a definition of quality.
- Risk-based QC works – if we are willing to admit the risks we accept.
- SAPs should state what we intend to analyse, not how we will to code it.
- AI will not replace human accountability, but it will expose where our effort is misaligned.
- If we don't define standards for AI in double programming, regulators will – eventually and bluntly.

Double programming remains a cornerstone of statistical quality assurance but must be applied with sound statistical judgment. As AI capabilities evolve, the focus should shift toward principles, transparency, and meaningful validation rather than output matching alone. Open discussion is needed to define best practices for the next generation of analysis workflows.

Veramed®

veramed.com

P013

Continuous Monitoring in Early Phase Oncology: A Standardised, Patient-Centric Approach

Tessa Lloyd, Rhiannon Maudsley, Aleksandra Kuzko, Brent Evans, Silvia Fernandez, Ilhem Mensi
Oncology Biometrics, Oncology R&D, AstraZeneca

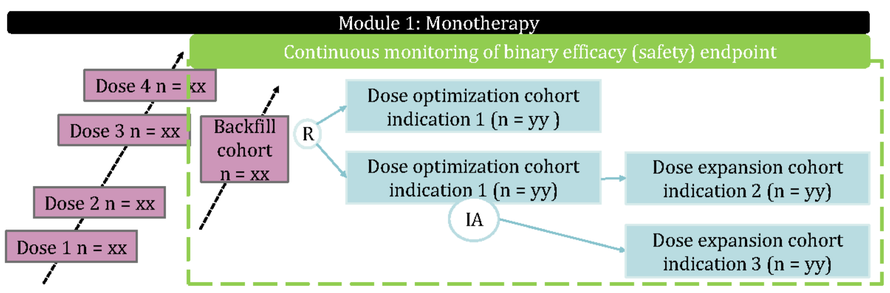
Introduction

- Early-phase oncology studies increasingly involve **parallel dose optimisation and expansion cohorts**
- The need to stop ineffective doses and/or accelerate programs requires **repeated, timely go/no-go decisions** while safety profiles are still emerging.
- **Infrequent interim analyses (IAs)** may delay decisions and prolong exposure to ineffective or non-tolerated doses.
- **Continuous Monitoring (CM)** enables planned, repeated statistical assessment of accumulating data to support **earlier, patient-centric decision-making**.

Methods

Motivation

Patient centricity and earlier decision making in early phase oncology trials has become more of a focus due to the FDA's guidance on "randomised, parallel dose response trial". An example trial design is shown below:

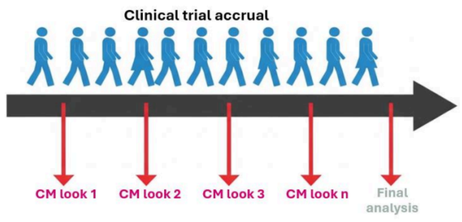


What is Continuous Monitoring?

CM is the systematic review of accumulating study data to inform timely cohort- or dose-level decisions, ahead of any formal interim or final analyses.

The main characteristics are:

- Planned and pre-specified decision framework
- Supports futility (safety) and acceleration decisions
- Complements (does not replace) IAs or final analyses
- Current work focuses on binary, interpretable efficacy endpoints, but could be extended to other endpoints



When Is CM Most Useful?

CM is most valuable when:

- Endpoint is timely and decision-ready
- Multiple cohorts or doses are running in parallel
- Early futility decisions can reduce patient exposure to non-promising doses / cohorts
- Development timelines require flexibility
- Safety monitoring of backfill cohorts post-escalation is required

If these conditions are not met, planned formal IAs may be preferable.

Decision Framework

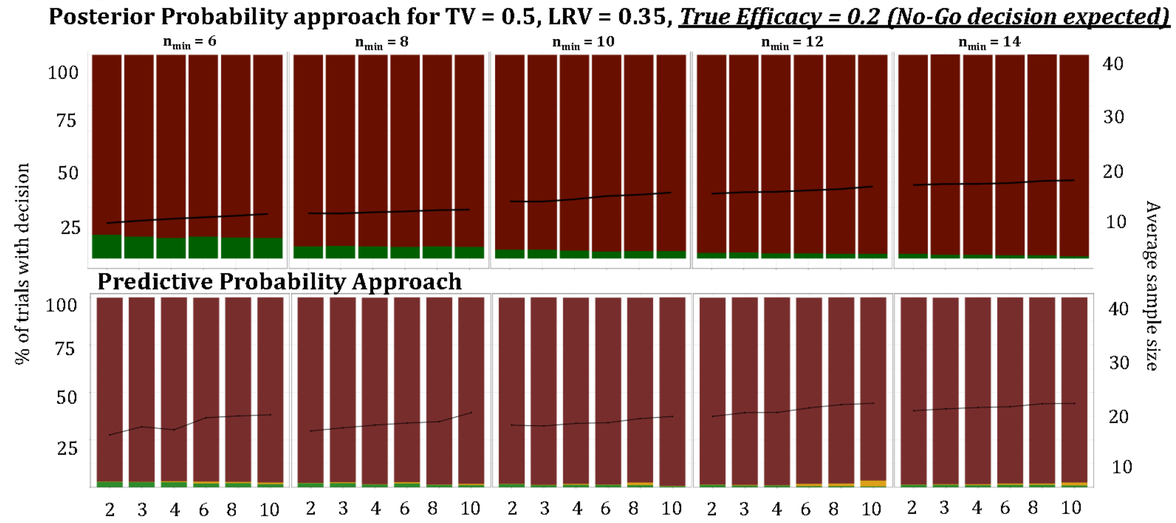
CM fits into the current decision framework implemented for decision making based on the methodology of Frewer et al. (2016). It assesses the probability of achieving a pre-specified target at repeated looks, rather than fixed IA timepoints. Two Bayesian approaches are considered: predictive and posterior probability, as illustrated below.

	Posterior probability monitoring	Predictive probability monitoring
Core question	Given data so far, what is Pr(p≥Target | data)?	Given data so far, what is the probability of a positive/negative outcome if we continue to N_{max}?
What is computed at a CM look	Posterior tail probs such as Pr(p > LRV | x, n) and/or Pr(p > TV | x, n).	• **PPOS** = Pr(Final Go at $N_{max} \mid x, n$) ; • **PPOF** = Pr(Final NoGo at $N_{max} \mid x, n$) both defined with the pre-specified final rule in mind.
Decision trigger	Stop/Go when a posterior probability crosses a pre-specified threshold.	• If PPOS < L stop (futility); if PPOS > U stop (efficacy); else continue • If PPOF > U_f or PPOS > U_s– stop early (futility/efficacy respectively)
Small-n / stability guardrails	Small-n: depend on the prior and require a minimum starting n (n_{min}) to avoid premature decisions	
Implementation / computation	Posterior is Beta; posterior tail probabilities computed in closed form (e.g., via Beta CDF/quantiles).	Uses beta-binomial predictive over future responses; yields implementable response-count boundaries by n after calibration/search.

CDF Cumulative Distribution Function; L Lower Bound; LRV Lower Reference Value; PPOF Predictive Probability of Failure; PPOS Predictive Probability of Success; TV Target Value; U Upper Bound

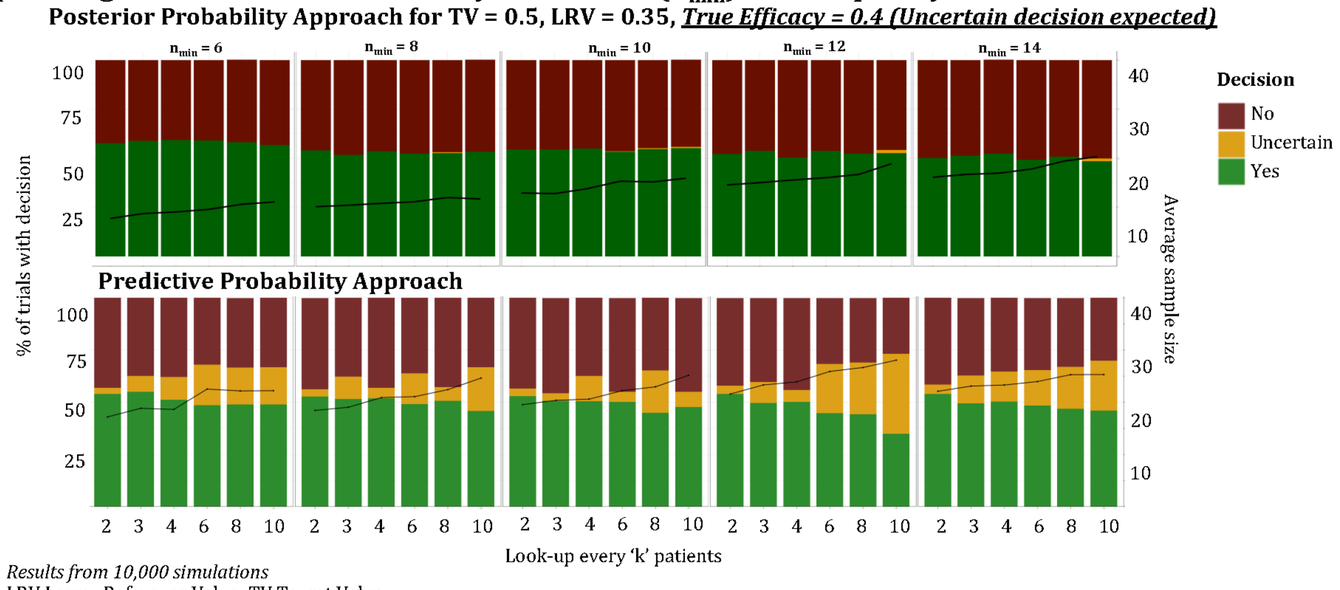
Operating characteristics: sensitivity to CM start (n_{min}) and frequency in non-controversial cases

Posterior Probability approach for TV = 0.5, LRV = 0.35, *True Efficacy = 0.2 (No-Go decision expected)*



Operating characteristics: sensitivity to CM start (n_{min}) and frequency in controversial cases

Posterior Probability Approach for TV = 0.5, LRV = 0.35, *True Efficacy = 0.4 (Uncertain decision expected)*



Results from 10,000 simulations

LRV Lower Reference Value; TV Target Value

Results

Simulation & Operating Characteristics

- Figure illustrates the results from 10,000 simulated trials (N=40; LRV/TV as shown)
- Varying start points (n_{min}) and frequencies (CM every *k* patients) in controversial and non-controversial decision cases (true efficacy) showed limited sensitivity within explored ranges
- No performance gain due to looking more frequently, for a given n_{min}
- Starting after n_{min} ≥8 evaluable patients provides stable decision behaviour
- Within the posterior and predictive approaches, similar performance across the tested n_{min} and frequencies
- The predictive approach typically required larger sample sizes to account for uncertainty in future observations

Recommended Approach

- For efficacy CM, predictive probability directly targets final success or futility while reducing false positives and false negatives.
- Posterior probability is a valid alternative, particularly for safety
- Study-specific simulation is essential to calibrate thresholds and understand OCs

Operational Considerations

Successful CM implementation requires:

- Frequent quality data cuts
- Handling of pending and unevaluable data
- Reduced, focused CM deliverables
- Clear documentation in protocol and SAP
- Decision logging (governance/audit)

Limitations/Caveats

- Dependent on data lag and maturity
- Inconsistency in decision making at small n
- Potential estimation bias with early stopping

Conclusions

- **Where CM fits**: Best suited to dose optimisation/expansion, enabling repeated cohort- and dose-level decisions from accumulating binary signals.
- **Method choice**: Predictive probability is recommended as the default for efficacy CM.
- **Design guidance**: Design choices (start point and cadence) can be guided pragmatically, as performance is robust within explored ranges.
- **Trade-off:** CM enables earlier, patient-centric decisions and potential for program planning/acceleration, but with the need of careful threshold calibration and increased operational complexity.

References

- Frewer P, Mitchell P, Watkins C, Matcham J. Decision-making in early clinical drug development. Pharm Stat. 2016 May;15(3):255-63. doi: 10.1002/pst.1746.
- Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. Clin Trials. 2008; 5(2):93-106. doi: 10.1177/1740774508089279.
- FDA Guidance: Optimizing the Dosage of Human Prescription Drugs and Biological Products for the Treatment of Oncologic Diseases (August 2024)

Acknowledgements

Jayne Marshall and Paul Frewer in the AstraZeneca Early Oncology Statistics Leadership Team for ongoing support throughout the development of the continuous monitoring guidance document.

P014

Leveraging Predicted Interval Plots (PIPS) to Predict the Outcome of an Oncology Study through Interim Monitoring of Event Accrual

Valeria Mazzanti, MPH, Cytel Inc., Geneva, Switzerland

Introduction

Interim monitoring for study outcome prediction has become commonplace in drug development. During the design stage before a trial begins, biostatisticians rely on treatment effect assumptions to perform power calculations. Once the trial is ongoing, pooled data that has been accrued up to an interim analysis can be used to predict more precisely the outcome of the study while maintaining data blinding principals. It is therefore paramount that a reliable tool is used to both maintain study blinding and perform predictions in a precise manner for better decision-making practices to move forward with the remaining stages of the trial.

Prediction Versus Hypothesis Testing

Many group sequential methods have been developed for monitoring clinical trials across a range of endpoint types, probability of winning criteria, and study settings. However, only a limited number of these approaches provide insight into effect sizes (such as treatment differences), and none incorporates prediction to inform potential future effect size estimates and their corresponding precision if the trial continues. To address these shortcomings, Evans et al. (2007) introduced the use of prediction and predictive intervals as a flexible and practical framework for the quantitative monitoring of clinical trials. The predictive intervals can then be expressed in graphical form for efficient communication.

Conditional Simulations

The approach used for this interim analysis is based on simulating data for the remainder of the study conditional on the trends observed in the accumulating trial data both on enrolment and event accrual. In this way, interim results are leveraged to generate potential future outcomes, introducing the predictive component of the methodology. Repeated confidence intervals are then calculated from the generated patient-level data. These intervals may then be ordered and layered sequentially to create plots that provide additional insights about the magnitude of the treatment effect and its associated uncertainty. Such visualizations can play an important role in supporting early stopping decisions and informing overall trial risk management.

Case Study Parameters

Primary Outcomes
Overall Survival (OS)

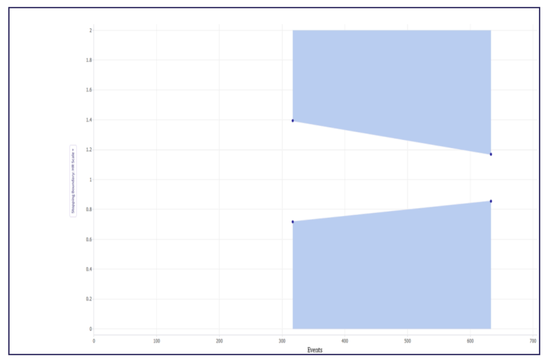
Interim Analysis Goal
Evaluate stopping the trial early for efficacy at 50% information
(i.e. when 50% of total required events had been observed)

Parameter	Inputs
Required Sample Size	888
Required Number of Events	633
Allocation Ratio	1:1
Treatment Effect (i.e. Expected Hazard Ratio)	0.80
Median Survival Time (Control Arm)	10 months
Median Survival Time (Experimental Arm)	12.5 months
Type-1 error (2-sided)	0.05
Target power	80%

Interim Analysis Parameters

Parameter	Inputs
Number of Events at Interim	317
Sample Size at Interim	797
Analysis Time of Interim	About 16 months
Treatment Effect (i.e. Expected Hazard Ratio)	0.80
Probability of Stopping for Efficacy	16.5%

Stopping Boundaries – Design Stage



Interim Analysis Evaluation

The trial was designed with an efficacy boundary using a Lan-DeMets O'Brien-Fleming spending function at 50% information, meaning when **317** events have been observed, and was expected to take place after 797 patients have been randomized.

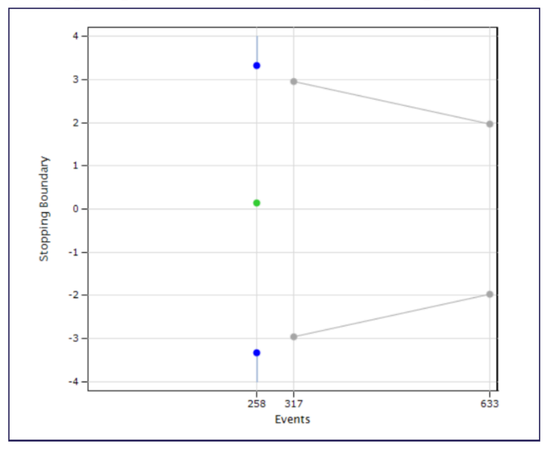
Once this number of patients had been randomized, though, only **258** events had been observed. The analysis, therefore, took place ahead of schedule.

The data monitoring committee noticed that the treatment effect was very weak, i.e. HR = 1.019. Without a futility stopping boundary, they could not use hypothesis testing to recommend stopping the trial early, so they turned to using Predictive Interval Plots to make their recommendation on how to move forward.

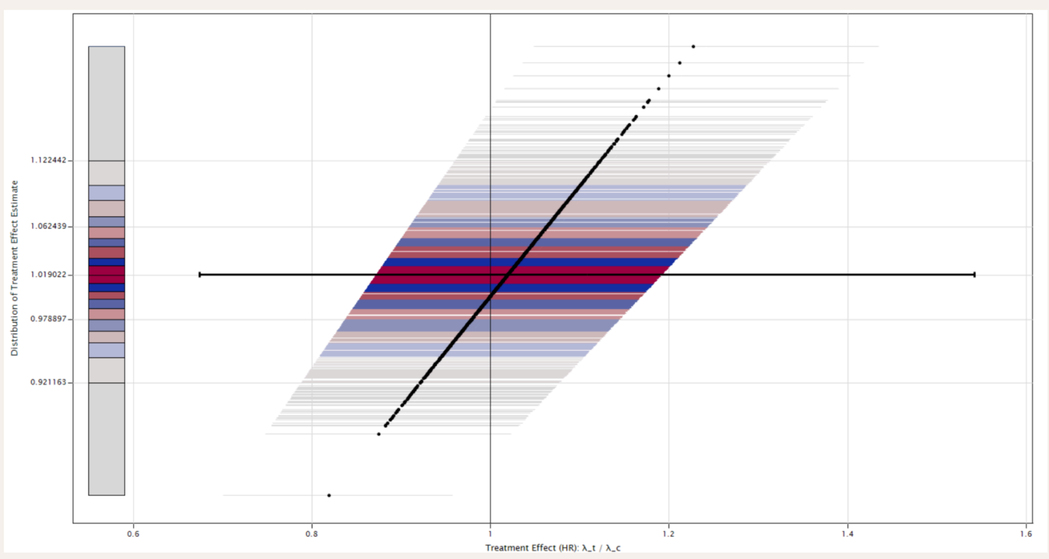
Interim Data Snippet

Treatment ID	SurvTime	ArrivalTime	Censor
2	1.4	0.033333	1
2	1.4	0.066667	1
1	4	0.866667	1
2	5.3	0.533333	0
1	21.13333	2.533333	0
1	4.8	2.066667	1

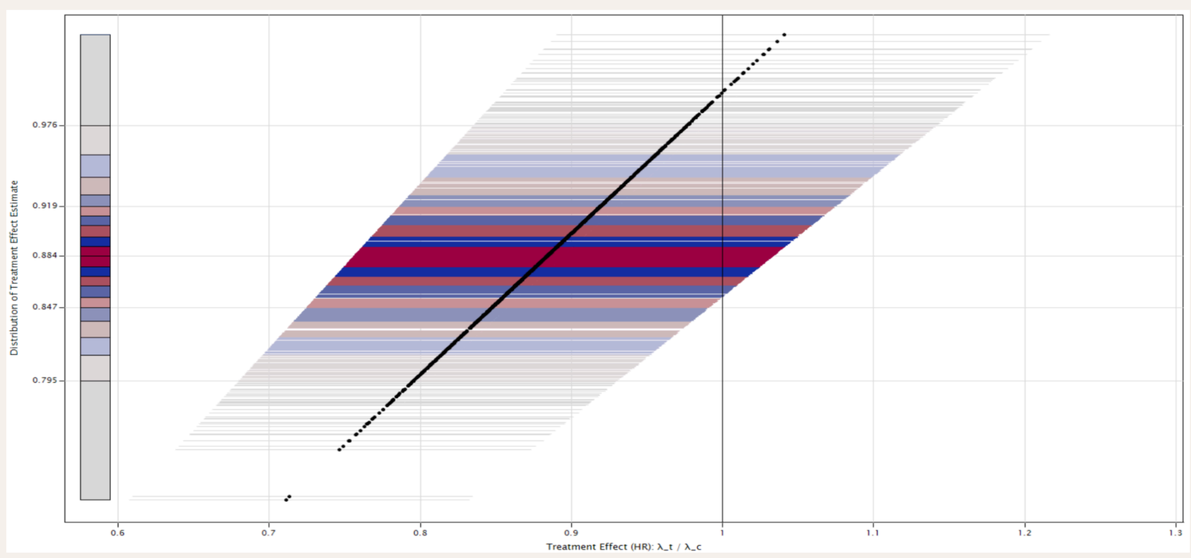
Stopping Boundaries – IA Stage



Predictive Interval Plot – simulating under observed Hazard Ratio; 1.019



Predictive Interval Plot – simulating under alternative Hazard Ratio; 0.80



Results

Simulating conditional on interim data, i.e. assuming the HR = 1.019, only 1 out of 10,000 simulations had their **upper** confidence bound below HR = 1, suggesting very low likelihood of success overall.

Simulating under the alternative hypothesis, i.e. assuming the HR = 0.80, the chance of success is only 29%.

Additionally, even under the alternative scenario, only 5% of future simulations show a point estimate for HR that is smaller than 0.795.

The data monitoring committee, therefore, recommended that the trial should be **terminated early for futility**, although no futility boundary was present at the design stage that they could compare back to.

Conclusions

The interim analysis used a simulation software to forecast and reforecast future outcomes based on the data accrued throughout the trial period to identify the probability of success at any future analysis, whether this will be the final analysis or a subsequent interim analysis.

The conditional simulation results could then be compared to the results simulated based on the alternative scenario, to still account for the fact that not all data had been accrued and there may have been lack of robust evidence at this earlier analysis time point.

The visual representation of PIPs supported cross-functional stakeholder interactions and promoted informed decision-making, especially in this moment of absence of hypothesis testing or statistical insignificance to demonstrate enough evidence leaning towards a futility decision.

Future research could also explore the opportunity for PIPs to act as supporting evidence when making efficacy decisions when the efficacy boundary is only marginally crossed.

east horizon

Valeria Mazzanti | Valeria.Mazzanti@cytel.com

Cytel

Daria Shamaida, Senior Statistician II

Love It or HTA It: Exploring Variability in Submission Requirements & Implications for Global Strategy

Health technology assessment (HTA) submissions differ across agencies in structure and requirements.

NICE, G-BA/IQWiG, and HAS have different formats, timelines, and expectations for clinical data and analyses, including comparators, subgroup analyses, sensitivity analyses, and handling of missing data.

As a result, the same clinical data often needs to be presented in different ways, increasing complexity and the need for careful planning.

This poster highlights key differences between HTA approaches and shows how to plan analyses for global submissions across these differences.

HTA Structural Differences

Weighting of Evidence Types

Germany (G-BA/IQWiG): Strong focus on randomized clinical trials (RCTs) aligned with the comparator. Assessment focuses on clinical outcomes. Economic and real-world data are considered separately and have limited impact.

France (HAS): Clinical data drive added clinical value. Other data may support the assessment when relevant. Economic data are used selectively.

United Kingdom (NICE): Clinical, economic, and real-world data are considered together. Decisions are based on cost-effectiveness, with economic modelling playing a key role.

Conceptual Weighting of Evidence Types in HTA Decision Making



Illustrative comparison of relative decision influence.

Depth, Format, and Timing of Documentation

Germany (G-BA/IQWiG): Requires a highly structured dossier submitted upfront, with complete, comparator-aligned clinical data forming a fixed record and little scope for adding new data later.

France (HAS): Uses structured but more narrative documentation. Some additional data may be considered, but early submissions largely drive outcomes.

United Kingdom (NICE): Uses the most flexible approach. Submissions evolve over time, with data and analyses updated and uncertainty in cost-effectiveness explored during appraisal.

Comparative HTA Evidence & Review Requirements

Agency	Main Focus	Documentation	Timing
G-BA / IQWiG	Comparator-aligned RCTs, predefined outcomes	Highly prescriptive, rule-based dossier	Front-loaded; key evidence required at submission
HAS	Clinical benefit, relevance to practice	Structured, clinically driven narrative	Early data drive decisions; limited later data
NICE	Clinical outcomes and cost-effectiveness, including uncertainty in clinical results and economic models	Flexible, transparent, with ongoing dialogue	Iterative; data and analyses updated during appraisal

Key Statistical Differences across HTA

Defining Appropriate Comparators

G-BA/IQWiG follows strict guidelines, with a strong preference for direct comparisons against the required comparator pre-defined by the agency. Indirect or adjusted analyses are rarely accepted.

In contrast, NICE and HAS are more flexible. They may accept alternative comparators, including those not directly studied in trials, if the assumptions are clearly explained and clinically reasonable.

Planning Subgroup Evaluations

G-BA/IQWiG assess added benefit within predefined, guideline-aligned subgroups, requiring sufficiently precise estimates. Only subgroups of adequate size are evaluated, and a consistent, clearly demonstrated effect is required. Imprecise or inconsistent results (e.g. wide confidence intervals or no clear subgroup effect) may be interpreted as no added benefit.

NICE and HAS also use subgroup analyses to support decision making, focusing on whether differences are clinically plausible and relevant. For example, effects in high-risk groups may still be considered important, even if statistical evidence is limited, if justified and uncertainty in subgroup effects is explored.

This difference, seeking clear evidence vs using evidence to inform decisions, explains why the same data can lead to different or more restricted recommendations.

Sensitivity Analysis and Decision Credibility

In G-BA/IQWiG, sensitivity analyses test how robust results are to different plausible assumptions. When results are not stable, the estimated effect may change, which can lead to lower or no added-benefit conclusions.

In contrast, HAS uses sensitivity analyses to support interpretation of clinical results, while NICE focuses on uncertainty in cost-effectiveness, driven by clinical and model inputs.

Handling missing data

In Germany (IQWiG/G-BA), the focus is on whether results remain reliable under different assumptions. Optimistic assumptions (e.g. missing at random, MAR) are often challenged. If results change under worse-case assumptions (e.g. imputing poorer outcomes for missing patients), confidence is reduced and benefits may be downgraded.

HAS focuses on whether assumptions reflect realistic patient outcomes. For example, patients who stop treatment are often assumed to have worse outcomes. The choice of assumptions must be clinically justified and clearly explained, rather than relying only on standard statistical methods.

NICE generally accepts MAR-based methods if the impact of missing data assumptions is explored, for example by comparing different imputations, and this impact is reflected in cost-effectiveness results.

Statistical Approaches in HTA

Agency	Comparator & Design	Subgroups	Sensitivity & Missing Data
G-BA	Strict comparator alignment; direct comparisons preferred	Prespecified; sufficient size; clear, precise effects required	Robustness tested under plausible assumptions; unstable results may reduce benefit
HAS	Clinically relevant comparators; structured clinical narrative	Support interpretation of clinically plausible and consistent effects	Assumptions reflect realistic patient outcomes; clear justification required
NICE	Decision-relevant comparators; flexible approach	Define value-relevant populations	Uncertainty in cost-effectiveness from clinical and model inputs

Cross-HTA Recommendations

Start with strictest HTA (G-BA/IQWiG); adapt later
Prespecify early (comparators, subgroups, sensitivity)
Keep flexibility for HAS/NICE
Ensure robust results under conservative assumptions
Avoid small or imprecise subgroups
Align plans early (trials, SAPs, models)

Same data -> different realities

The same clinical data can lead to different conclusions across HTAs because each agency applies different standards—what supports a decision in one system may not count in another.

G-BA. Benefit Assessment of Medicinal Products under §35a SGB V (AMNOG).
HAS. Clinical Assessment Methodological Guidance.
NICE. Health Technology Evaluations: The Manual (PMG36).

VeraMed®
veramed.com

P016

Forecasting screened/randomised patients and optimal screening stopping times to minimise over-enrolment in clinical trials

Dr. Vlad Anisimov, Thomas Henderson
Data Science, CfDA AMGEN, UK

Aim and Framework

In clinical trials, patients typically undergo one or more screening stages before being randomised. Stopping screening new patients too late can lead to over-enrolment, where more patients than planned are randomised, increasing cost, patient burden, and delaying study closure without providing scientific benefit. On the other hand stopping screening too early comes with the risk of not meeting the target sample size.

Key idea: Using interim data, predict the distribution of the number of patients randomized out of those in screening stages yet and estimate the probability of reaching the planned randomisation target.

Decision rule: Stop screening new patients early when this probability is sufficiently high (e.g., 95%–99%).

Otherwise, forecast screening/randomisation pipeline over time and the time to stop screening with a given confidence.

We developed a new statistical framework to predict patient randomisation process over time, enabling optimal timing of screening closure to minimise over-enrolment while maintaining a high probability of reaching the target sample size. This included a joint screening/randomisation model with multi-stage screening/rescreening logistics which allows for analytic prediction of patient randomisations for both in-pipeline patients and future arrivals.

This methodology has been tested within clinical trials and has demonstrated the ability to reduce over-enrolment to less than 2%.

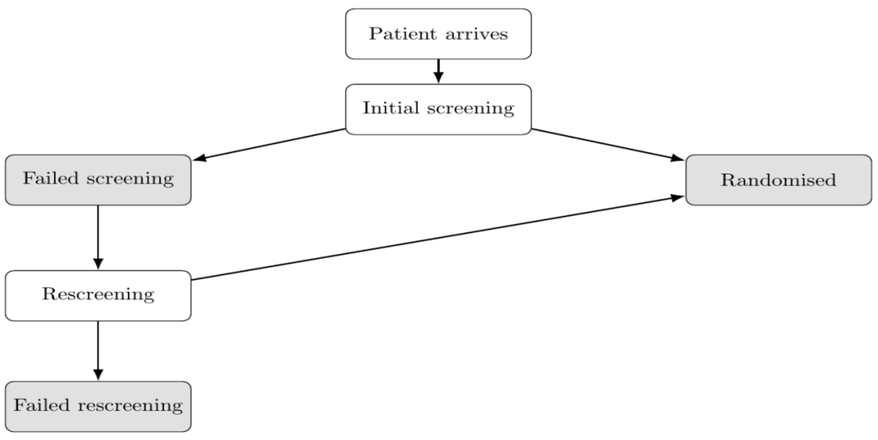


Figure 1. Diagram of patient flow for screening/rescreening

Analytic Framework and Supporting Approaches

We consider three approaches for testing the results:

- Analytic Method: Joint screening/randomisation model using evolving hierarchic process for future site initiation and site-driven arrivals
- Monte Carlo: Near-precise evaluation of stopping-time distribution and scenario validation
- Normal Approximation: Faster approximation using analytic mean and variance

Figure 2 shows an example of a long-term randomisation forecasting with new sites activating in the future using simulated input data.

Key applications:

- Forecast can be updated at interim times using real trial data.
- Produces quantile bands and probabilities to reach various targets.
- Scenario planning for new site openings and closure schedules.
- Probability to over-enrol and by how much.

Forecasting Randomised Patients for Decision Making

The analytic forecast provides for each day the distribution of the number of randomised patients and can also be split to consider only patients in-pipeline, only future arrivals or both. Quantiles for numbers of randomisations over time are calculated and the probability of achieving the target sample size can be evaluated for each day.

Monte Carlo simulation is then used to evaluate, with very high precision, the distribution of the time that a chosen criterion to stop screening would be met (e.g. stopping if probability to reach target out of patients only in pipeline already is greater than 95% or 99%). This can also be used to evaluate probabilities and plan scenarios for other operational targets such as over-enrolment in particular countries or regions.

Additionally, we can consider planned site closures or region-specific screen stop times for scenario planning and assess the impact on enrolment. This is especially useful for determining time to reach specific enrolment targets by regions.

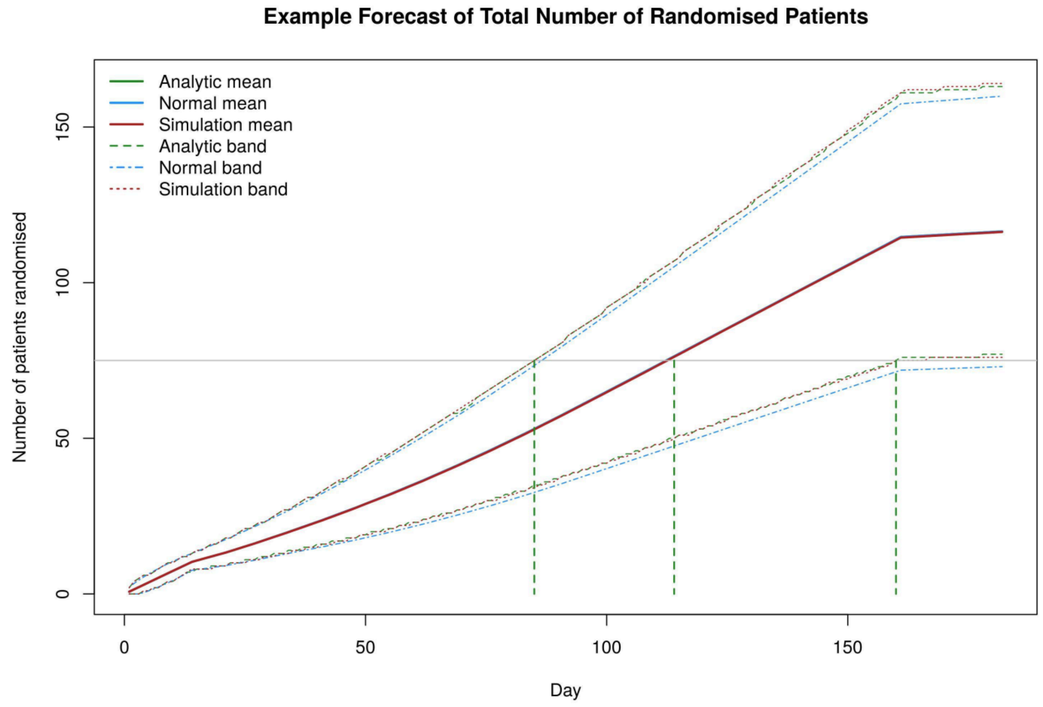


Figure 2. Plot of patient randomisation using three approaches over time that show practically the same results. Vertical lines show times when each quantile met the randomisation target. Quantiles used are 5% and 95%.

Joint Forecasting of Screening/Randomisation Pipeline

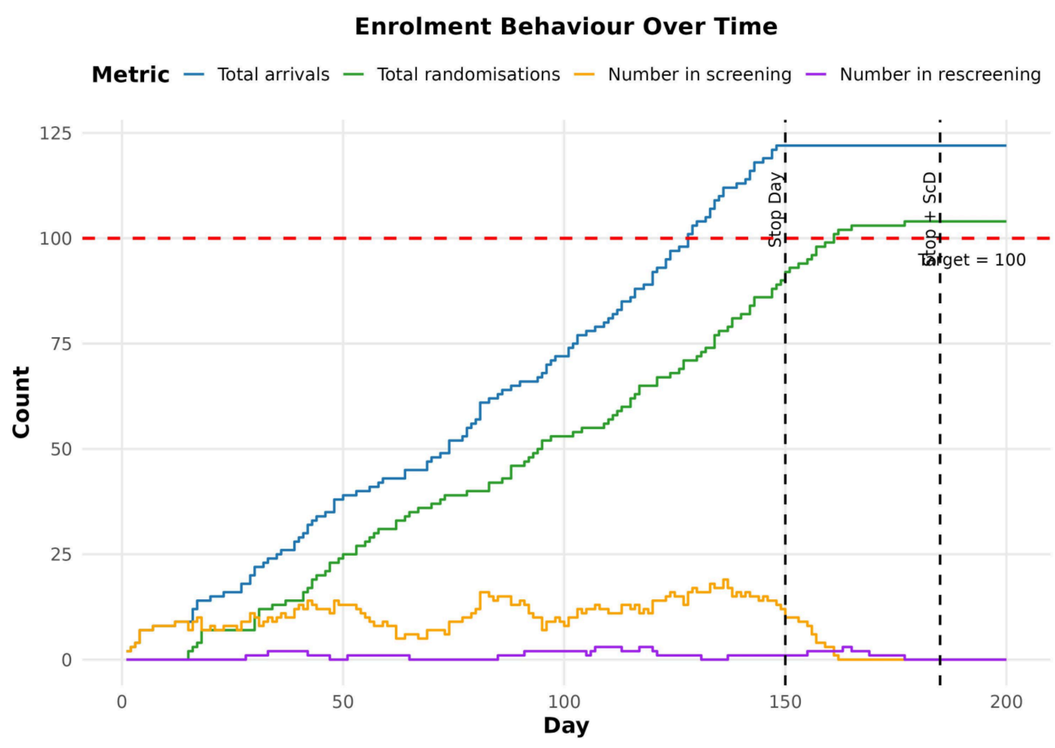


Figure 3. Example simulation of screening behaviour with total arrivals, randomisation and counts of patients in screening/rescreening. Vertical lines show time at which screening was stopped in this simulation and time at which patients will have stop randomising after.

References and Acknowledgements

1. V. Anisimov, V. Fedorov, Modelling, prediction and adaptive adjustment of recruitment in multicentre trials, Statistics in Medicine, 26, 27, 2007, 4958-4975.
2. V. Anisimov, Forecasting the process of screened/randomized patients and optimal early stopping time in multicentre clinical trials, 11th International Conference on Advances in Statistics, 25-27 April 2025, p.18, Bologna, Italy.
3. V. Anisimov, L. Gortzak, Joint forecasting of screening/randomization processes in multicentre clinical trials and predicting optimal enrolment stopping time, in book “ASMDA 25 Conference Proceedings”, June 2025, Wiley, 14 pp. (to appear)

The authors would like to express their sincere gratitude to Matt Austin and Data Science Team at the Center for Design & Analysis and the Operational Design Team at Amgen, for constant and useful discussions on this subject.

AMGEN
Clinical Trial Research
vanisimo@amgen.com
thende03@amgen.com

P017

PENDING POSTER SUBMISSION

Daria Vynohradova, PhD

Shaping Europe's Demographic Future: The Role of Clinical Trials

Background

Europe is becoming increasingly diverse from both socio-economic and demographic perspectives, as countries pursue distinct development trajectories and respond differently to economic and political changes. These shifts directly affect healthcare systems, patient populations, and the design and execution of clinical trials. Clinical research drives innovation and influences mortality and fertility trends – key determinants of demographic resilience.

Clinical trials play a key role in shaping demographic potential – affecting population health, longevity, and productivity – by influencing the safety, effectiveness, and availability of medical treatments.

Objective

To establish and demonstrate a quantifiable relationship between clinical trials and demographic stability across major European clinical trial countries.

Clinical trials play a key role in shaping demographic potential – affecting population health, longevity, and productivity – by influencing the safety, effectiveness, and availability of medical treatments.

Demographic Indicators → **Principal Components Analysis (PCA)** → **Composite Index** → **Cross-Country Comparison & Ranking**

Core
- Life expectancy at birth
- Total fertility rate
- Net reproduction rate
- Infant mortality rate
- Total mortality rate

Socio-Economic shell
- Population ageing
- Total demographic burden ratio
- Migration balance

Methods

To evaluate demographic potential, a comprehensive statistical framework was developed.

The resulting index enables cross-country comparison and ranking, providing a robust, data-driven assessment of demographic conditions across the European region.

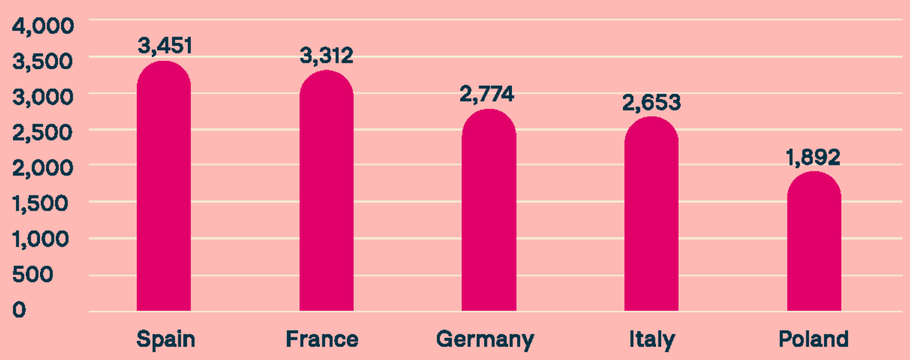
Using elasticity coefficients, we measure how improvements in healthcare outcomes translate into demographic benefits, highlighting the sensitivity of demographic security to healthcare advancements.

Results

As of mid-2025, Spain established itself as the leading country in Europe for clinical trials, ranking first in both total volume and authorized studies within the EU. It is closely followed by France, Germany, and Italy, which also maintain strong positions by effectively utilizing the EU Clinical Trials Regulation to accelerate approvals and sustain high numbers of commercial trials.

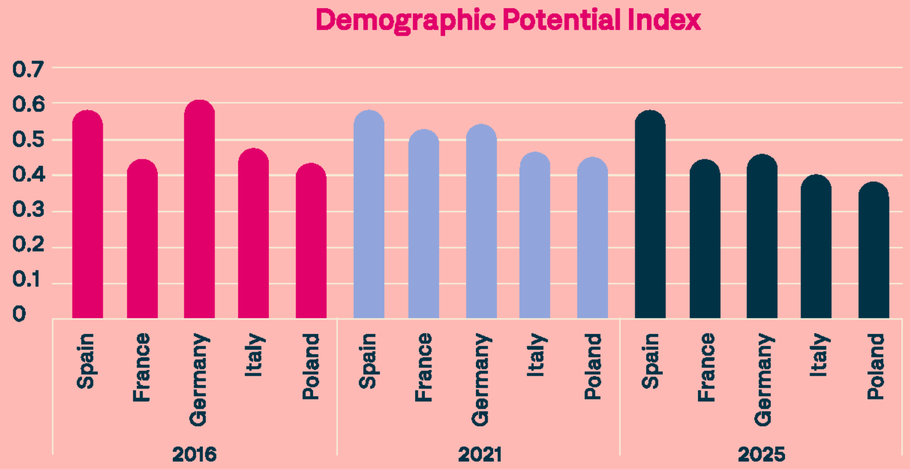
Between 2022 and 2025, the leading causes of death in Europe were circulatory diseases, cancer, respiratory diseases, COVID-19, and external causes. In response, Europe's 2025 clinical trial landscape focuses on oncology, cardiovascular, and CNS research, with rapid growth in cell and gene therapies.

Number of Active Trials per Country as of March 14th 2025



Source: pharmaboardroom.com/articles/top-10-european-countries-for-clinical-trials-in-2025/

Demographic Potential Index



Source: ec.europa.eu/eurostat/data/database

The composite index ranges from 0 to 1, designed to capture overall demographic potential and security. Higher values indicate stronger demographic conditions and greater stability, while lower values reflect a deterioration in the demographic situation.

Between 2021 and 2025, a 1.0% reduction in total mortality was associated with an average 2.4% change in estimated demographic potential across Spain, France, Germany, Italy, and Poland.

Conclusions

This work connects statistical methodology with strategic insights for clinical development. Understanding demographic dynamics helps:

- **Anticipate recruitment challenges**
- **Optimize trial design**
- **Align research with population health priorities**

Engaging with this analysis means contributing to evidence that shapes both clinical research strategy and public health policy across Europe.

Veramed®
veramed.com

P020

Innovative Trial Design Through Bayesian Dynamic Borrowing: Industry Lessons and Applications

How estimand misalignment can induce artificial prior-data conflict and compromise trial power.

Presented by: Lesedi Ledwaba-Chapman & Jamie Inshaw

Background

There has been increasing interest over recent years from sponsors and regulators in Bayesian Dynamic Borrowing (BDB) as a robust method for incorporating external information into clinical trial analysis. The method was given attention in recent FDA draft guidance (January 2026). BDB uses historical data to design highly efficient, ethically streamlined clinical trials. Crucially, the FDA mandates aligning primary estimands between external historical sources and prospective trials. Misaligned estimands make external data less relevant and introduce unintended prior-data conflict. We investigate how misaligned estimands impact Probability of Success (PoS) and Type I error (T1E).

Methods

We conducted a simulation study to evaluate PoS and T1E control under different estimand strategies and varying degrees of prior-data conflict and robust prior specification. It's assumed a small historical trial (n = 20) observed a mean (standard deviation) of 5 (5) in its control arm. We simulated 10,000 trials with 75 patients randomised 2:1 to the active and control arms, respectively, with the intention of supplementing the placebo arm by borrowing from the historical trial.

Active and control patients are simulated under two scenarios.

- **Null:** $N(5, 5^2)$ for both arms.
- **Treatment Effect (TE):** Active = $N(8, 5^2)$ and Control = $N(5, 5^2)$.

Three missing data scenarios were explored.

- **Null 1/TE 1:** Zero deaths in each arm.
- **Null 2/TE 2:** 5% deaths in each arm.
- **Null 3/TE 3:** Active = 5% deaths and Control = 10% deaths.

Two data handling strategies are applied to the trials.

- **Composite:** Deaths assigned an unfavourable value of 2*.
- **No Imputation:** Remove deaths from the analysis (mimicking what a traditional power calculation would typically assume).

*: Note that applying the Composite estimand effectively modifies the prospective control average down (to 4.85 for a 5% death rate and 4.70 for a 10% death rate). Similarly, the active arm average is effectively modified down to 7.7 under the 5% death rate.

Table 1 describes the priors considered in the simulations. Prior 0 has no prior, the likelihood is effectively the posterior. Priors 1 and 4 only have an informative component for the control arm, whereas Priors 2-3 and 5-6 are robust mixture priors and also have a noninformative component. For Priors 1-3, the informative component mean is identical to the true control mean, whereas Priors 4-6 have a degree of prior-data conflict.

The standard error of the informative component was assumed to be 1.12 (= $5/\sqrt{20}$) based on the historical data. We calculated the posterior probability that the treatment effect (active – control) was greater than 0 over 10,000 simulated trials per scenario. If the probability was over 0.975, the study was considered a success. T1E and PoS was estimated by calculating the proportion of successes under the null and treatment effect scenarios, respectively. Analysis was carried out using KerusCloud®.

Table 1: Prior scenarios. Abbreviations, N = Normal Distribution; N/A = Not Applicable.

Prior	Prior-Data Conflict?	Active	Control
0	N/A	N/A	N/A
1	No	$N(0, 1000^2)$	$N(5, 1.12^2)$
2	No	$N(0, 1000^2)$	$0.5 \times N(5, 1.12^2) + 0.5 \times N(0, 1000^2)$
3	No	$N(0, 1000^2)$	$0.1 \times N(5, 1.12^2) + 0.9 \times N(0, 1000^2)$
4	Yes	$N(0, 1000^2)$	$N(3, 1.12^2)$
5	Yes	$N(0, 1000^2)$	$0.5 \times N(3, 1.12^2) + 0.5 \times N(0, 1000^2)$
6	Yes	$N(0, 1000^2)$	$0.1 \times N(3, 1.12^2) + 0.9 \times N(0, 1000^2)$

Figure 1: Probability of success for each scenario. Abbreviations, Comp = Composite strategy applied for deaths; Eff = Treatment Effect; NoImp = No Imputation.

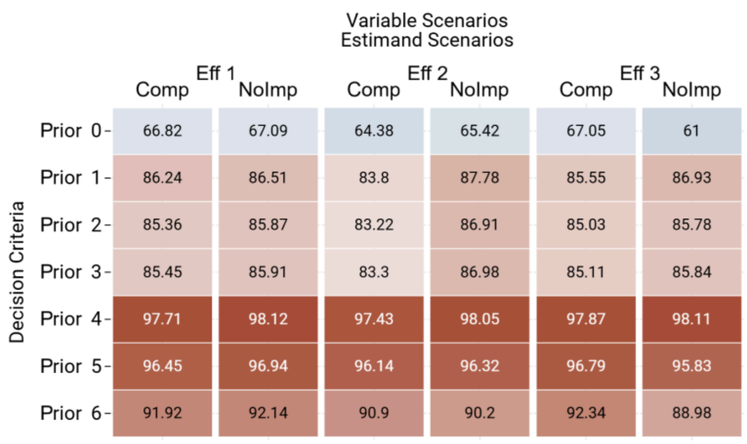
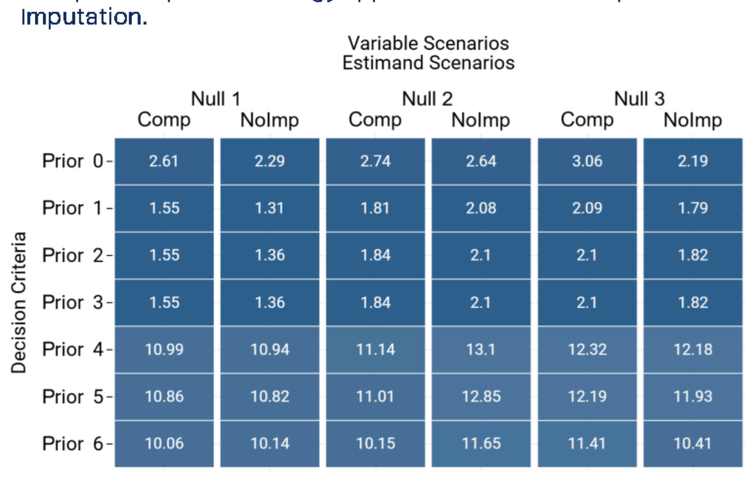


Figure 2: Type 1 error for each study scenario. Abbreviations, Comp = Composite strategy applied for deaths; NoImp = No Imputation.



Results

Choice of analysis and prior

When assessing the impact of the choice of analysis and priors the expected qualitative pattern was observed. Including information from the historical trial increased PoS compared to the Frequentist / uninformative prior analysis. When there was no prior-data conflict T1E was well-controlled, but in the presence of prior-data conflict T1E was increased.

Choice of estimand strategy

The impact of the interaction of estimand strategy with analysis strategy is seen most clearly under the 5% mortality in both arms scenario (TE 2). Using robust priors without prior-data conflict (Priors 2 and 3), the PoS was ≈3.70% lower with the Composite estimand vs the No Imputation (Figure 1).

Furthermore, while T1E remained controlled under well-matched conditions, introducing prior-data conflict (Priors 4, 5, and 6) severely inflated T1E. This is true even when Bayesian Dynamic Borrowing was used (Priors 5 and 6), with high T1E observed even when only 10% prior weight was placed on the informative part of the prior.

Conclusion

Our simulations illustrate that failing to align estimands between historical and prospective data can lead to mis-estimation of the PoS in a given study. In our study, applying a composite estimand to the prospective trial without adjusting the historical data artificially lowered the control mean. This induced an unnecessary prior-data conflict that resulted in a measurable loss of statistical power.

To maintain the intended operating characteristics of a trial, statisticians must view estimand compatibility as being just as critical as selecting the correct robust mixture prior. Aligning these analytical frameworks is essential to ensure that dynamic discounting mechanisms function as designed.

Following FDA guidance, sponsors should proactively reanalyse external historical data using the planned prospective estimand wherever patient-level data allows. Proper alignment preserves both Type I error control and optimal trial power, ensuring that Bayesian methods remain a robust tool for clinical development.

Powered By KerusCloud®

Assessing the Likelihood of Success in Phase 3 Vaccine Clinical Trials

P021

A Strategic Approach Based on Phase 2 Efficacy Results.

Fabian Tibaldi[1] and **Naveen Karkada**[2]

[1]Head of Statistical Excellence, GSK Vaccines and Infectious Diseases Biostatistics, Belgium.

[2]Statistics Director, GSK Vaccines and Infectious Diseases Biostatistics, Belgium.

Introduction

De-risking Vaccine Development

Challenge: Vaccine development is costly, time-consuming, and faces high attrition rates, particularly during the transition from Phase 2 to Phase 3.

Solution: A data-driven approach to estimate the **Probability of Success** (PoS) for Phase 3 trials, leveraging Phase 2 outcomes, can optimize development and trial design, reducing risks and costs.

What is PoS? PoS ("assurance") is the unconditional probability of a trial meeting its success criteria, accounting for uncertainty in true vaccine efficacy (VE). It is the expectation of power, averaged over the distribution for the unknown true VE, offering a more reliable indicator than traditional power.

Methods

A Simulation-Based Approach to PoS

Vaccine Efficacy (VE) Definition:

$$VE = 1 - RR = 1 - \frac{nv/Nv}{nc/Nc} = 1 - \frac{nv}{rnc}$$

Where *nv* and *nc* are infected subjects in vaccinated and control groups, *Nv* and *Nc* are total subjects, and r is the allocation ratio.

Hypothesis Testing: H_0: VE ≤ VE_0 (e.g., VE ≤ 25%) versus H_a: VE > VE_0.

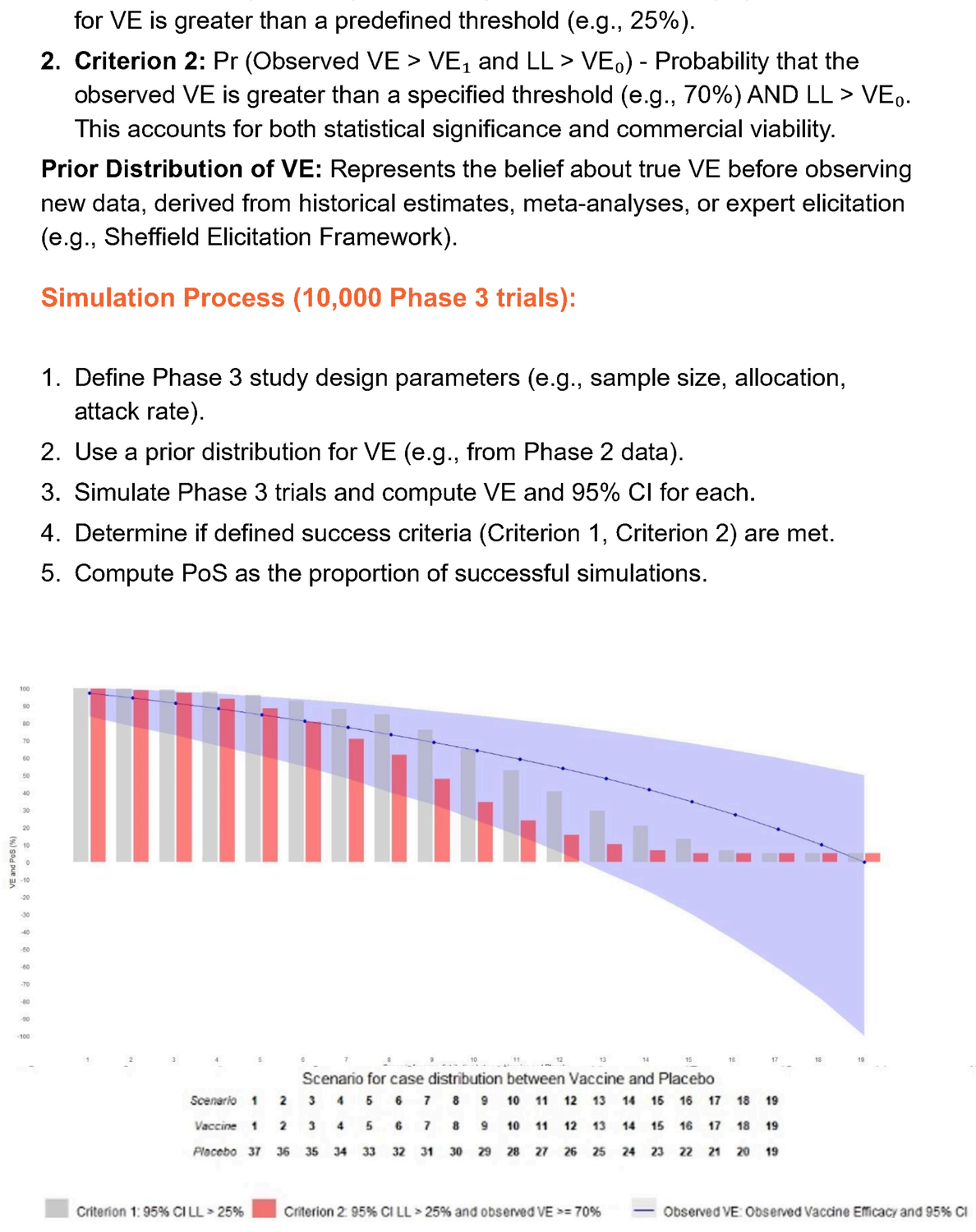
Success Criteria for PoS:

1. **Criterion 1:** Pr (LL > VE_0) - Probability that the lower limit (LL) of the 95% CI for VE is greater than a predefined threshold (e.g., 25%).
2. **Criterion 2:** Pr (Observed VE > VE_1 and LL > VE_0) - Probability that the observed VE is greater than a specified threshold (e.g., 70%) AND LL > VE_0. This accounts for both statistical significance and commercial viability.

Prior Distribution of VE: Represents the belief about true VE before observing new data, derived from historical estimates, meta-analyses, or expert elicitation (e.g., Sheffield Elicitation Framework).

Simulation Process (10,000 Phase 3 trials):

1. Define Phase 3 study design parameters (e.g., sample size, allocation, attack rate).
2. Use a prior distribution for VE (e.g., from Phase 2 data).
3. Simulate Phase 3 trials and compute VE and 95% CI for each.
4. Determine if defined success criteria (Criterion 1, Criterion 2) are met.
5. Compute PoS as the proportion of successful simulations.



Illustrative Example

Phase 2 VE Results Available

Phase 2 Data: 7 vaccinated cases, 31 placebo cases out of 38 observed. Observed VE = 77.42% [95% CI: 48.00%; 91.61%].

Prior for VE: Skew Normal (location=0.901, scale=0.188, shape=-6.18).

Planned Phase 3 Study:

- **Design:** 16,412 participants (1:1 allocation), 0.6% attack rate, aiming for 64 cases.
- **Success Criteria:** LL of 95% CI for VE > 25% with 90% power, and observed VE ≥ 70%.

Results:

- **PoS for Criterion 1 (LL > 25%):** 87.95%
- **PoS for Criterion 2 (LL > 25% AND observed VE ≥ 70%):** 70.77%

Case Study

Takeda's Norovirus Vaccine

Phase 2b Trial (Sherwood et al., 2020): Evaluated bivalent norovirus vaccine efficacy.

Endpoints & Prior Distributions:

- **Primary Endpoint:** Efficacy against vaccine genotypes (GI.1/GII.4). Observed VE = 80.0% [99.99% CI: -1318.1%, 99.7%]. Prior: Mirror gamma (Shape=0.504, Rate=1.1).
- **Secondary Endpoint:** Efficacy against any norovirus genotype. Observed VE = 61.8% [95% CI: 20.8%, 81.6%]. Prior: Skew normal (Location=0.794, Scale=0.262, Shape=-6.151).

Phase 3 PoS (assuming illustrative case design):

Endpoint	Criterion 1: 95% CI LL > 25%	Criterion 2: 95% CI LL > 25% and observed VE ≥ 70%
Primary	65.9%	57.6%
Secondary	63.3%	23.3%

Conclusions

- **Quantitative Decision-Making:** This approach provides a systematic way to estimate PoS, enabling informed decisions for advancing to Phase 3.
- **Impact of Criteria**: Demonstrates how different success criteria significantly affect PoS, highlighting the need for careful definition of trial goals.
- **Value of Secondary Endpoints:** Even when primary objectives are not met, secondary endpoints can offer insights into potential for future success.
- **Optimization:** The methodology can fine-tune sample size and optimize Phase 3 study designs for reasonable PoS and acceptable power.

References

- Davis MM, Butchart AT, Wheeler JR, Coleman MS, Singer DC, Freed GL. Failure-to-success ratios, transition probabilities and phase lengths for prophylactic vaccines versus other pharmaceuticals in the development pipeline. Vaccine. 2011 Nov 28;29(51):9414-6.
- O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry. 2005 Jul;4(3):187-201.
- Sherwood, J., Mendelman, P. M., Lloyd, E., Liu, M., Boslego, J., Borkowski, A., Jackson, A., Faix, D., & the US Navy Study Team. (2020). Efficacy of an intramuscular bivalent norovirus GI.1/GII.4 virus-like particle vaccine candidate in healthy US adults. Vaccine, 38(1), 172-178.

P022

Obesity RWD: From Noisy Weights to Evidence

Tzviel Frostig, Gal Shoham, Adir Shaham, Keren Perry, Elad Berkman, Raviv Pryluk
PhaseV Trials Inc., Boston, MA, United States of America

Introduction

Background: Obesity is linked to many chronic diseases, but real-world evidence is limited by imperfect endpoints and exposure data.

Objective: Assess where weight loss shows measurable benefit across four obesity-related conditions.

Data: Linked MarketScan–Veradigm EHR/claims data from >2 million individuals with obesity.

The RWD Trap: Rich Data, Noisy Signals

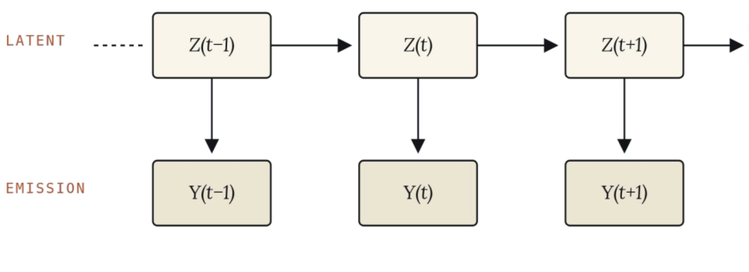
- **Measurement noise:** EHR weights are sparse, irregular, and error-prone
- **Temporal bias:** exposure, weight loss, and outcomes must be correctly ordered
- **Confounding:** patients who lose weight may differ systematically at baseline from those who do not.
- **Endpoint limitations:** RWD often requires proxies for trial-grade outcomes
- **Evidence maturity:** newer drugs and slow outcomes may need longer follow-up

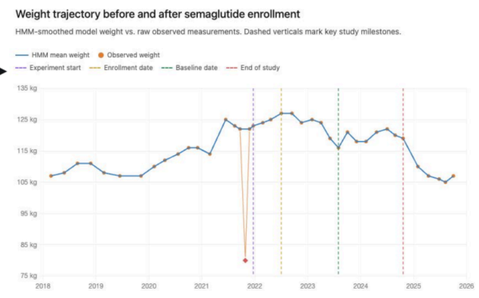
Each problem can distort an apparent treatment-effect signal if not properly addressed.

Methods: safeguards against RWD bias



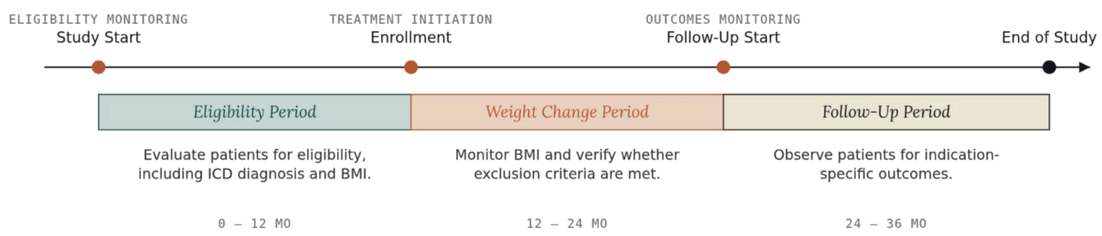
Hidden Markov Model (HMM)





Observed EHR weights were treated as noisy measurements of latent true weight

Target Trial Emulation (TTE)

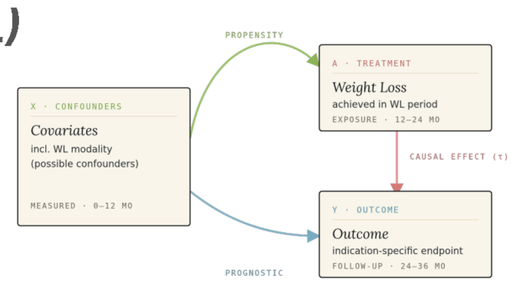


Each patient record was restructured around a trial-like timeline:

Eligibility covariates → WL exposure window → outcome follow-up

Double Machine Learning (DML)

$$Y - \underbrace{\mathbb{E}(Y \mid X)}_{\text{Prognostic}} = \underbrace{\tau}_{\text{ATE}} \left\{ A - \underbrace{\mathbb{E}(A \mid X)}_{\text{Propensity}} \right\}$$



- Estimated adjusted effect of achieved WL category vs. no weight loss/gain
- Cross-fitted orthogonalization reduced sensitivity to overfitting and ML model error

To learn more about this analysis see our blogpost

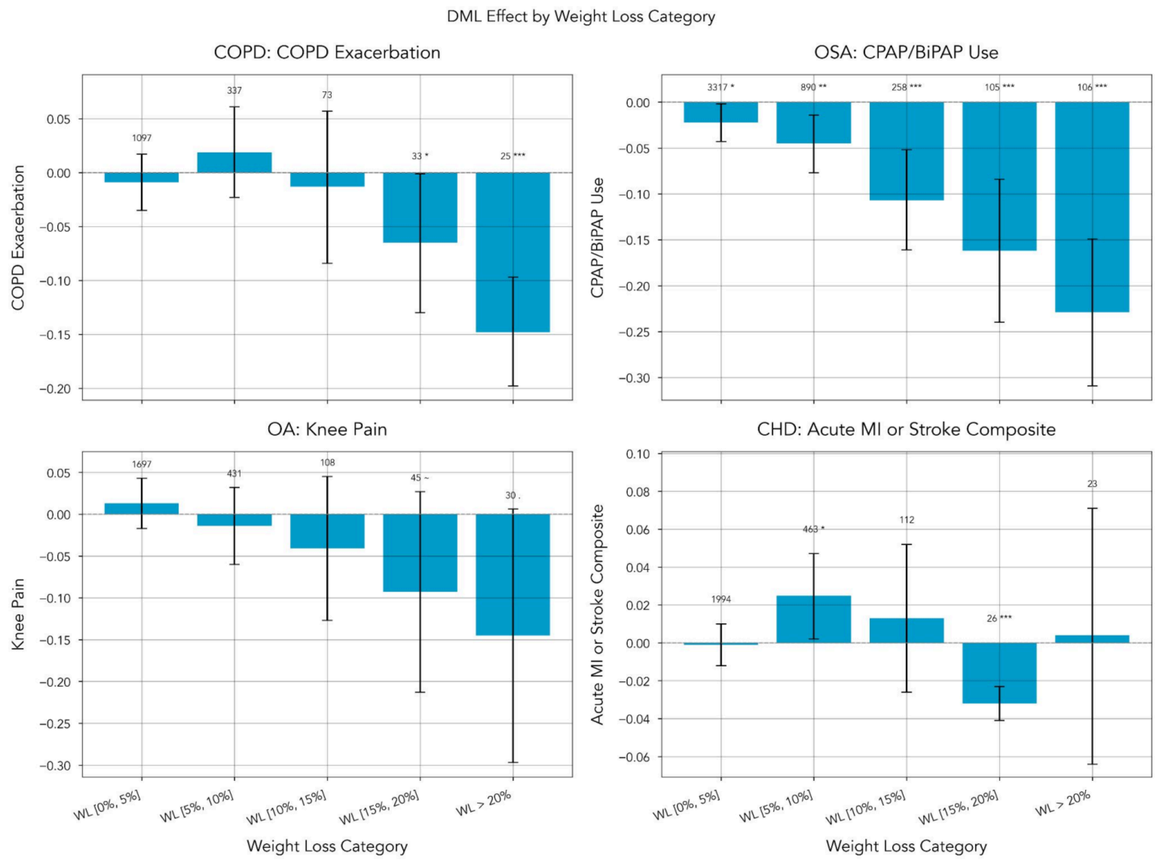
Endpoints and Proxies



Indication	RCT Endpoint	RWD Proxy
OSA (Obstructive Sleep Apnea)	AHI	CPAP/BiPAP use
COPD (Chronic Obstructive Pulmonary Disease)	FEV_1 / exacerbations	Claims-defined exacerbations
Knee OA (Osteoarthritis)	WOMAC/VAS	Pain-care utilization
CHD (Coronary Heart Disease)	MACE-3 including CV death	MI/stroke without mortality

A strong estimate with a weak proxy is not strong evidence.

Results



- **OSA:** clearest dose-response and most interpretable proxy
- **COPD:** promising high-weight-loss signal; warrants validation (small n)
- **Knee OA:** supportive but proxy-limited evidence
- **CHD:** inconclusive; missing mortality, and short follow-up

RWD problem	Safeguard	Bias addressed
Noisy EHR weights	HMM smoothing	Exposure misclassification
Exposure/outcome timing	TTE timeline	Immortal-time bias
Baseline imbalance	DML	Measured confounding
Proxy endpoints	Endpoint assessment	Endpoint misclassification
Incomplete claims	Coverage checks	Outcome under-capture
Fill ≠ sustained use	Adherence checks	Treatment misclassification

From RWD Analysis to Development Decisions

RWD evidence was interpreted through three linked questions:

Exposure reliability: Can achieved weight loss be measured credibly?

Temporal validity: Were covariates, exposure, and outcomes correctly ordered?

Endpoint validity: Is the RWD proxy clinically meaningful?

Causal modeling can turn noisy RWD into useful evidence only when exposure, timing, confounding, and endpoint validity are addressed together.

Toby Batten, Senior Manager Statistics

Walking the Tightrope: Balancing the Needs of All Stakeholders for Biotech Survival

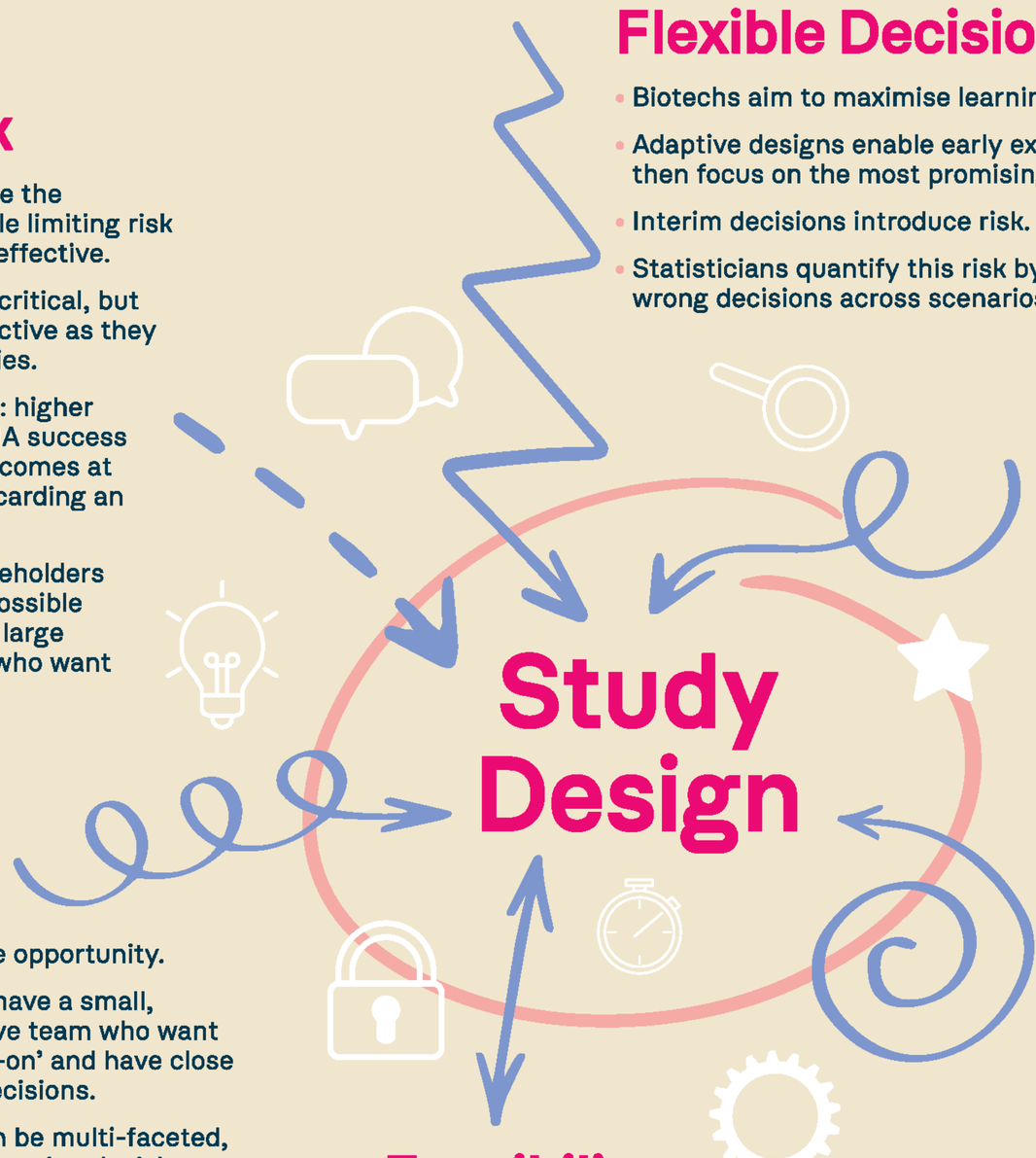
Designing the right study is crucial to the future of any company. However, for a small biotech, its entire survival may depend on it. Therefore, it is essential to design studies in a way that maximises probability of success.

Investor Risk

- Investors seek to maximise the probability of a return while limiting risk if the treatment proves ineffective.
- Well-powered studies are critical, but interim analyses are attractive as they offer early exit opportunities.
- Futility thresholds are key: higher thresholds increase post-IA success probability. However, this comes at the cost of potentially discarding an effective drug.
- Potential for multiple stakeholders and investors, including possible investment could be from large pharmaceutical partners who want the option to buy-in.

Flexible Decision Making

- Biotechs aim to maximise learning from every study.
- Adaptive designs enable early exploration across doses or indications, then focus on the most promising.
- Interim decisions introduce risk.
- Statisticians quantify this risk by calculating the probabilities of making wrong decisions across scenarios to support informed decisions.



Budget

- Biotechs rely on fixed investor or grant funding from government or charitable bodies, with no alternative income.
- Budgets are scrutinized.
- Budget overruns require new capital and renewed commitment.
- Study assumptions must therefore be rigorously tested upfront.

Control

- Often one asset, one opportunity.
- Sponsors will often have a small, nimble and innovative team who want flexibility, be ‘hands-on’ and have close involvement in all decisions.
- Interim analyses can be multi-faceted, not just limited to stopping decisions, but changes to the study design and updating investors.
- Maintaining blinding and trial integrity becomes more challenging.

Feasibility

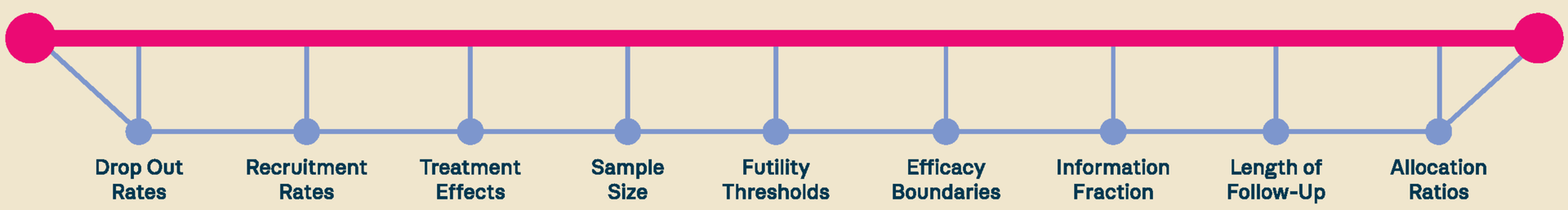
- Study design must be feasible, realistic, and deliverable.
- Endpoint selection and handling is critical to operational success.
- For example, Time-to-event designs are often based on an event driven sample size, but this means the study duration and sample size are variable.
- Fixed budgets favour fixed sample size approaches with an assumed number of events to ensure the study can be completed.

Power

- For sponsors, success = statistical significance.
- Positive results are essential to secure continued investment.
- Studies must be appropriately designed, including adequate power and accounting for all factors such as drop-out and missing data.
- Effect size assumptions must be realistic and not inflated to fit the budget.

Walking the tightrope requires balancing competing stakeholder needs

Simulations can be a useful tool, or safety net, to quantify trade-offs and make risks explicit



- Success depends not just on methods, but on clear communication and collaboration
- Statisticians must actively engage, translate risk, and guide decision-making
- Biotech settings amplify this challenge with limited resources, high stakes, fast decisions
- This demands a broad set of skills including: flexibility, influence, and strategic thinking

Veramed®
veramed.com

P024

Beyond BOIN: TUBA - A Tunable Bayesian Framework for Custom Phase I Dose Finding Designs

Yaron Racah, Amitay Kamber, Yaniv Fogel, Elad Berkman
PhaseV Trials Inc., Boston, MA, United States of America

PHASEV
Adaptive. Causal. Intuitive.

Background

Phase I oncology trials aim to identify the Maximum Tolerated Dose (MTD). Model-based designs (e.g., BLRM, CRM) rely on parametric assumptions about the dose–toxicity curve. Model-assisted designs (e.g., BOIN, mTPI-2) sidestep this by reducing decisions to a simple lookup table — driving their rapid adoption.

BOIN is competitive at MTD identification, but has a relatively high probability of selecting an over-toxic dose[1] — arguably the most consequential failure mode for a dose-finding design. BLRM is more conservative.

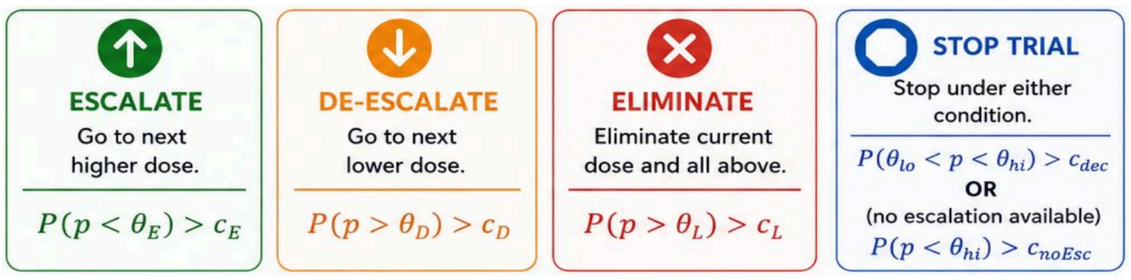
Objective

Build a dose-escalation framework that:

- Preserves the strengths of model-assisted designs — **no functional assumptions** on the toxicity curve, **transparent decision rules**.
- Is **fully tunable** along a spectrum from aggressive (BOIN-like) to conservative (BLRM-like) operating characteristics.
- **Enhances dose selection accuracy** for each reference design while adhering to the **same DLT** (dose-limiting toxicities) **budget**.

The TUBA Framework

Cohorts of 3 patients are recruited; the trial starts at the lowest dose. After each cohort, the posterior $P(p \mid \text{data})$ for current-dose toxicity p is computed under a uniform prior. Four rules — each parameterised by a toxicity threshold θ and a confidence c — select the next action:



The trial terminates after 12 cohorts, on an early-stop trigger, or on elimination of the lowest dose. On positive termination, isotonic regression over sampled doses gives per-dose toxicity estimates; the dose closest to a target rate p^* (separate from the sponsor-set target toxicity ϕ) is selected.

What's New Compared to BOIN / mTPI-2?

- **Two-parameter rules.** Separately tuning thresholds (θ) and confidence requirements (c) allows the design to behave appropriately at both small and large per-dose sample sizes — a single parameter rule cannot.
- **Posterior-based early stopping.** The trial ends when a dose is credibly inside the target window, rather than waiting for cohort-count heuristics.
- **"No-escalation" stop.** When the top dose is reached (or all higher doses eliminated), an under-toxic dose can be confirmed quickly.

Calibration

Parameters are tuned by **model-based optimisation** on 1,000 simulated five-dose scenarios generated per Clertant & O'Quigley[2], augmented with scenarios in which every dose is over-toxic. At each iteration, the protocol is simulated at perturbations of the current parameter vector; local linear models estimate gradients of the objective (e.g., P(MTD)) and of constraints (e.g., expected DLTs); a constrained gradient step is taken; iterate until convergence.

TUNABLE PARAMETERS

θ_E	escalation threshold
θ_D	de-escalation threshold
θ_L	elimination threshold
θ_{lo}, θ_{hi}	indifference interval
p^*	target toxicity rate**
c_E	escalation confidence
c_D	de-escalation confidence
c_L	elimination confidence
c_{dec}	decision-stop confidence
c_{noEsc}	no-escalation stop confidence

**This is not the target toxicity ϕ (set by the sponsor), but a tunable internal target used by the framework.

Results

We show two TUBA calibrations on 1000 random scenarios, designed with BOIN and BLRM as respective references.

Figure 1 - dose selection: the distribution of doses selected by each design (over all scenarios).

- **vs. BOIN:** P(over-toxic dose) reduced **17%** → **13%** at matched P(acceptable dose).
- **vs. BLRM** (with a scenario-aligned prior): P(acceptable dose) **+1.5 pp** at matched P(over-toxic).

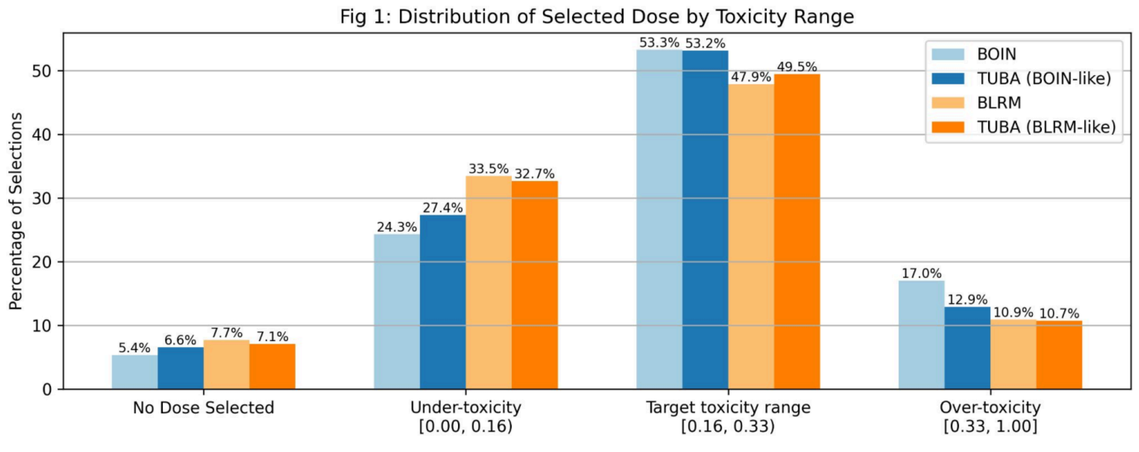
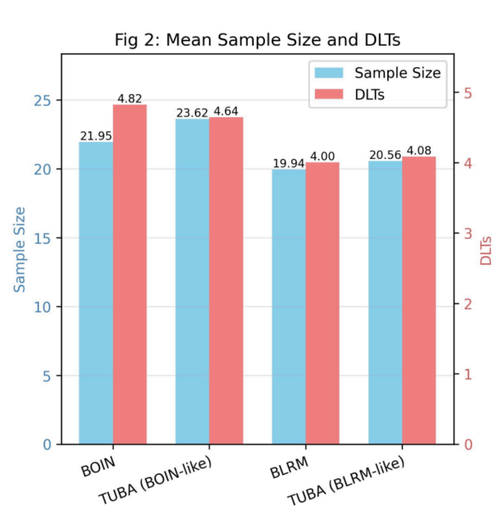


Figure 2 — mean sample and DLT counts. TUBA stays within the reference's DLT budget at a small sample-size premium.



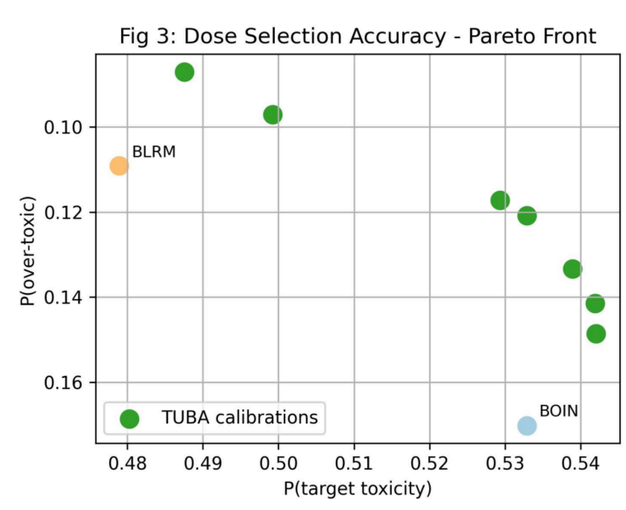


Figure 3 — Pareto front. TUBA calibrations at a fixed DLT budget of 4.82 (BOIN's mean) trace the achievable **P(acceptable) vs P(over-toxic)** frontier. The framework can be tuned to any point on the curve.

Decision Tables — Calibration in Action

Like BOIN, TUBA's per-dose rules can be distilled into a decision table. Tables for the two calibrations, BOIN-like and BLRM-like, are shown alongside BOIN; The multi-dose model used by BLRM itself has no tabular equivalent. Differences between the tables are highlighted.

N	BOIN			TUBA — BOIN-like					TUBA — BLRM-like				
	E	D	L	E	D	L	T	T*	E	D	L	T	T*
3	0	2	3	0	1	3	—	—	0	1	3	—	—
6	1	2	4	0	2	4	—	0–1	0	2	4	—	0–0
9	1	3	5	1	3	5	—	0–2	1	4	5	2–2	0–2
12	2	4	6	1	5	6	3–4	0–4	1	5	6	2–3	0–3
15	2	5	7	2	6	7	3–5	0–5	2	6	7	3–4	0–4
18	3	6	8	3	7	8	4–6	0–6	2	7	8	3–6	0–6

At each cohort with k DLTs observed at the current dose:
E = max k to escalate **D** = min k to de-escalate
L = min k to eliminate current and all higher doses
T = k-range triggering early termination
T* = same, when no higher dose is available.

TUBA (BOIN-like) differs from BOIN
TUBA (BLRM-like) differs from TUBA (BOIN-like)

Conclusion

TUBA preserves BOIN's two practical advantages — table-based decisions and no parametric dose-toxicity assumption — while delivering better operating characteristics: fewer over-toxic doses at BOIN's setting, more acceptable doses at BLRM's, at matched DLT exposure. The calibration handle lets sponsors place the trial anywhere on the Pareto front, to suit drug, indication, or regulatory context.

1. Zhou H, Yuan Y, Nie L. Accuracy, safety, and reliability of novel phase I trial designs. Clin Cancer Res. 2018;24(18):4357–64.
2. Clertant M, O'Quigley J. Semiparametric dose finding methods. J R Stat Soc Series B Stat Methodol. 2017;79(5):1487–508.

P025

PENDING POSTER SUBMISSION

P027

PENDING POSTER SUBMISSION

P028

Recurrence: An Analysis of Adverse Events Whose Time Has Come in the Evaluation of Patient Safety in Clinical Trials

Richard C. Zink, PhD, JMP Statistical Discovery LLC, USA; Department of Biostatistics, University of North Carolina at Chapel Hill, USA.

Introduction

Most analyses of adverse events (AEs) are based on presence or absence. This analysis has important limitations. First, patients who experience one event versus many events are weighted identically with no distinction made for the number of events a patient may experience, save for a patient who experiences no events. Second, the timing at which these events occur is completely ignored within the analysis. Study teams have no insight into whether events occur early in treatment compared to late, or if the occurrence of events is constant, increasing, or decreasing over time.

Recurrence Analysis

- Siddiqui (2009) describes the application of the mean cumulative function (MCF) to the analysis of AEs.
- The MCF is the average number of cumulative events that occur up to a particular point in time, and it can be interpreted as the average of the individual cumulative AE curves of all patients (Nelson, 1995).
- Estimated non-parametrically with no assumptions about form

Methods

- MCF estimated as $\widehat{M}(t) = \sum_{t_f \le t} \frac{x_f}{n_f}$, where x_f is the total number of events experienced and n_f is the number of patients at risk at t_f.
- Annotations can include:
  - Number of patients at risk
  - Number of patients at risk who contribute at least one event
  - Cumulative events

Sample Data

- CDISC SDTM-ADaM Pilot Project
- Mild to moderate Alzheimer's disease to assess xanomeline transdermal patches
- Three arms: High and low doses vs placebo
- Trial duration of 26 weeks
- 254 randomized patients
- 242 different preferred terms (PTs)
- 230 PTs have treatment emergent AEs (TEAEs)
- 1126 TEAEs
- Focus on TEAEs of pruritis, in which 55 patients experienced a total of 80 events

Results

- Figure 1: Patient-level (left) and treatment-level (right) event plots which show the timing and frequency of events over time.
- Figure 2: MCFs and 95% pointwise confidence intervals by treatment with annotations for the number of patients at risk.
- Figure 3: Dot-forest plot summarizing MCFs by treatment and the differences in MCFs between active doses with placebo measured at the minimax follow up (210.6 days).
- Figure 4: Max difference plot summarizing the difference between the maximum and minimum MCFs across the entire length of follow up by PT. This plot overcomes a key limitation of Figure 3 in screening for important AEs only at the minimax follow up.

Conclusions

- Recurrence addresses limitations of current AE analysis
  - Recrudescence
  - Timing
  - Long and varying follow up
  - Information about event rates over time
- Recurrence can be used beyond individual PTs
  - All TEAEs
  - Moderate or severe TEAEs
  - Serious TEAEs
  - TEAEs of special interest
  - Medical queries
- Data visualization is necessary for effective interpretation and communication
- Recurrence data (Table 1) can provide more informative drug labels
- JMP Clinical includes recurrence analyses to automate MCF estimation and visualization for AEs and medical queries.

Table 1. Pruritis MCF Across Time

Follow up	Xanomeline High Dose	Xanomeline Low Dose	Placebo
30 Days	0.191 (0.078, 0.304)	0.170 (0.041, 0.299)	0.012 (0*, 0.036)
60 Days	0.390 (0.226, 0.554)	0.319 (0.161, 0.476)	0.024 (0*, 0.058)
90 Days	0.439 (0.266, 0.612)	0.380 (0.213, 0.547)	0.104 (0.018, 0.190)
120 Days	0.537 (0.343, 0.731)	0.414 (0.244, 0.584)	0.145 (0.043, 0.247)

Figure 1. Event Plots of Pruritis TEAEs

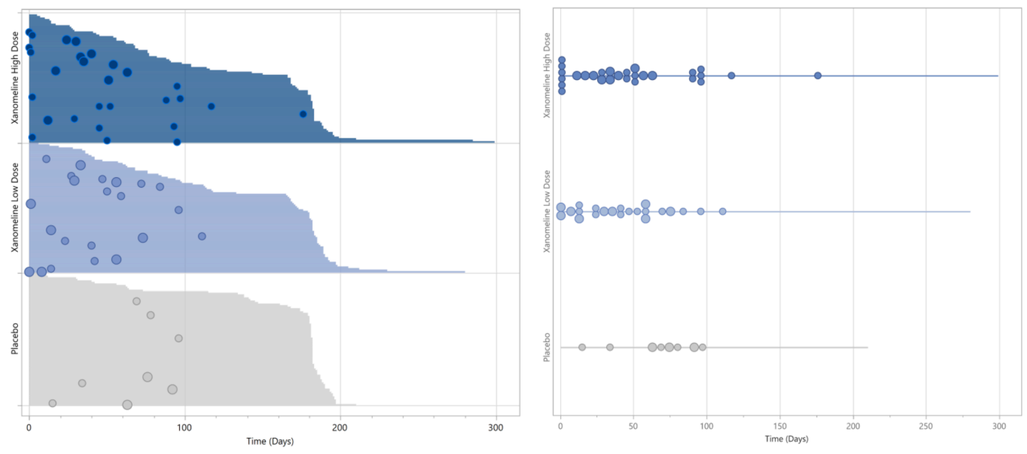


Figure 2. Mean Cumulative Function Plot

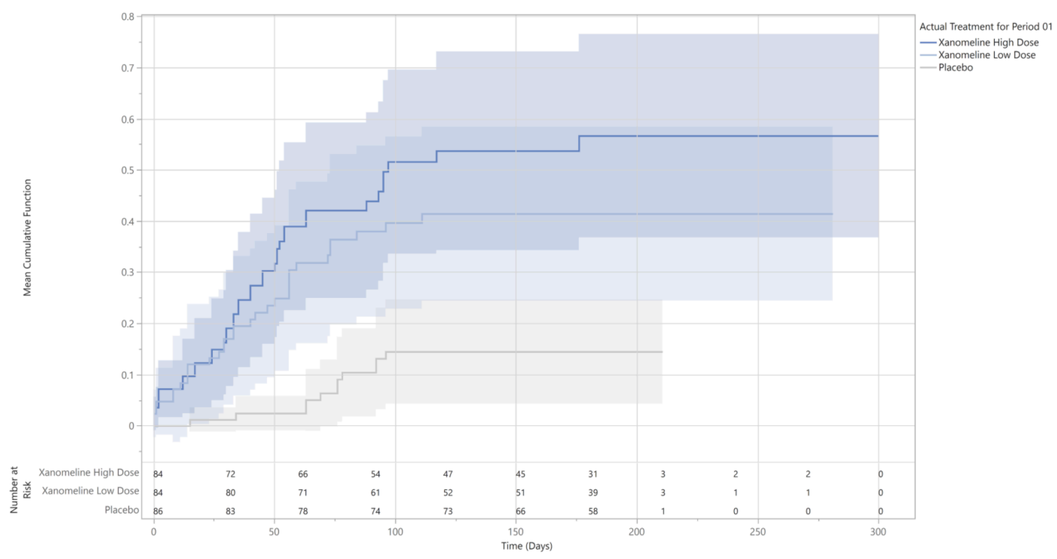


Figure 3. Dot-Forest Plot for Screening

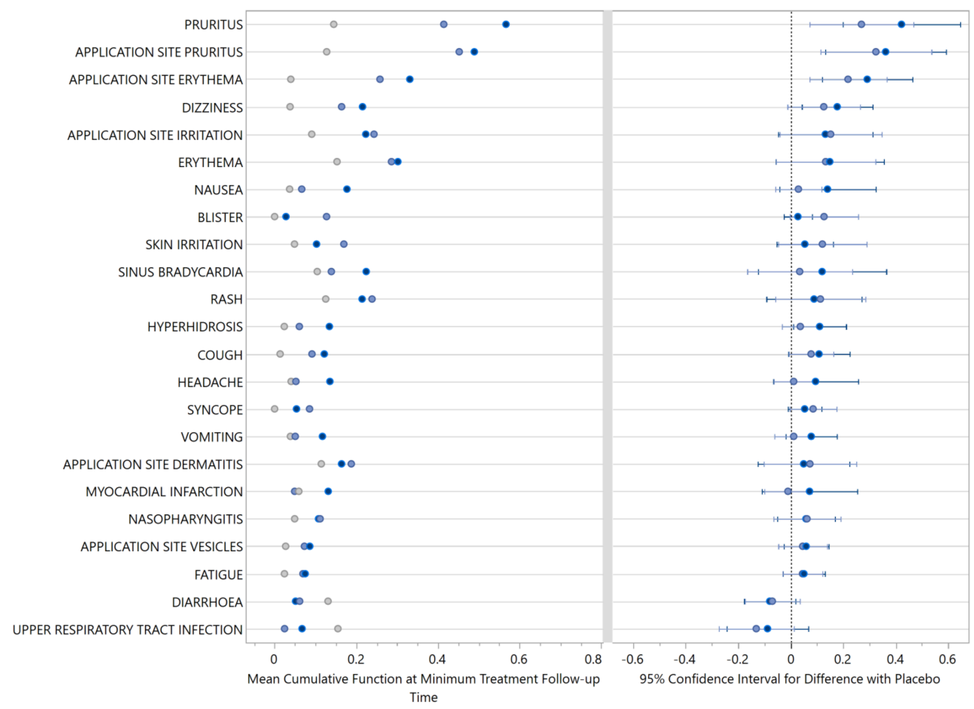
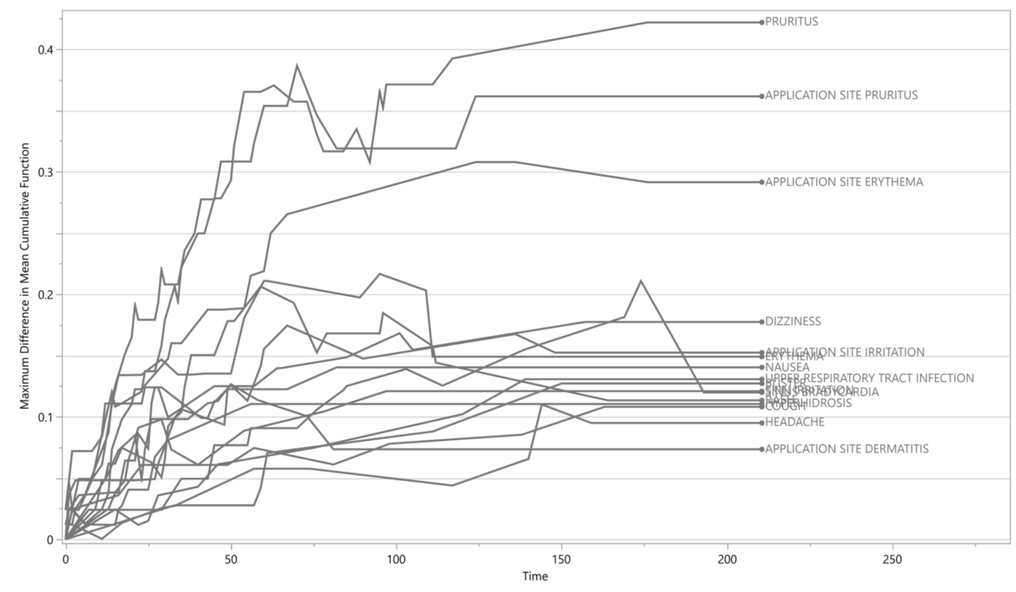


Figure 4. Max Difference Plot for Screening



References

- Nelson W. (1995). Confidence limits for recurrence data: Applied to cost or number of product repairs. *Technometrics* 37: 137-157.
- Siddiqui O. (2009). Statistical methods to analyze adverse events data of randomized clinical trials. *Journal of Biopharmaceutical Statistics* 19: 889–899.

P029

Maximizing insight and efficiency: Comparison of Rate of decline in FVC and change from baseline in FVC analysed using frequentist and model-based approaches in idiopathic pulmonary fibrosis trials

Beatriz Seoane [1], Anel Mahmutovic [1], Neha Ann Cherian [1], Sofia Tapani [1]
[1] AstraZeneca BioPharmaceuticals R&D | Respiratory & Immunology | R&I Biometrics and Statistical Innovation

Introduction

Idiopathic pulmonary fibrosis (IPF) poses a key challenge for statisticians: designing efficient trials in a progressive disease where mortality, the most meaningful endpoint, is rarely feasible.
Forced vital capacity (FVC) decline is therefore widely used as a surrogate, but how it is modelled makes a difference. Single timepoint measures may miss a longitudinal signal, while rate of decline better captures disease dynamics.
In this work, we compare change from baseline and rate of FVC decline and evaluate their performance within an MCP-Mod framework for dose–response optimization.

Simulations

Simulations were performed to identify the optimal primary endpoint for a 24-week Phase 2 study, assuming the following:

3 doses and placebo, with a sample size of 110:55:55:110 (total = 330 participants).

Dose-response assumed monotonically increasing with a maximal FVC effect at the highest dose level delivering a change from baseline in FVC of 60 mL versus placebo (ie, 2.5 mL/week over 24 weeks). This corresponds to a potential average response and length of Ph2 trials in current environment of IPF drug development.

Five candidate models for the MCP-Mod describing dose-response relationship (Linear, 2 different Emax, a Sigmoid Emax, and an exponential model). See Figure 1.

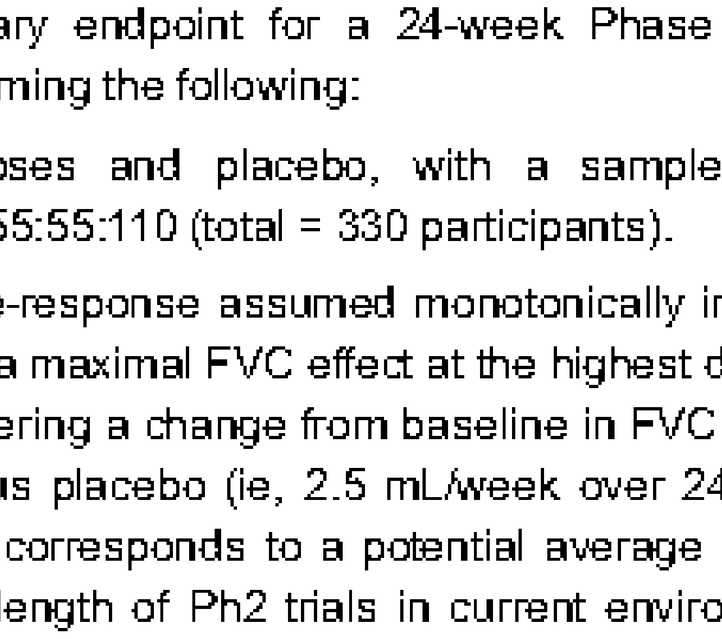


Figure 1. Candidate models describing the dose-response relationship (Linear, 2 different Emax, a Sigmoid Emax, and an exponential model)

Results

Rate of FVC decline over 24 weeks was analysed using a random coefficients regression model, and

change from baseline in FVC at Week 24 using a Mixed Model for Repeated Measures (MMRM).

Both endpoints were further evaluated using MCP-Mod to assess dose-response and operating characteristics.

As expected, across the candidate models, the results show a 5.1-8% increase in power with rate of decline in FVC + MCP-mod. See Figure 2 below.

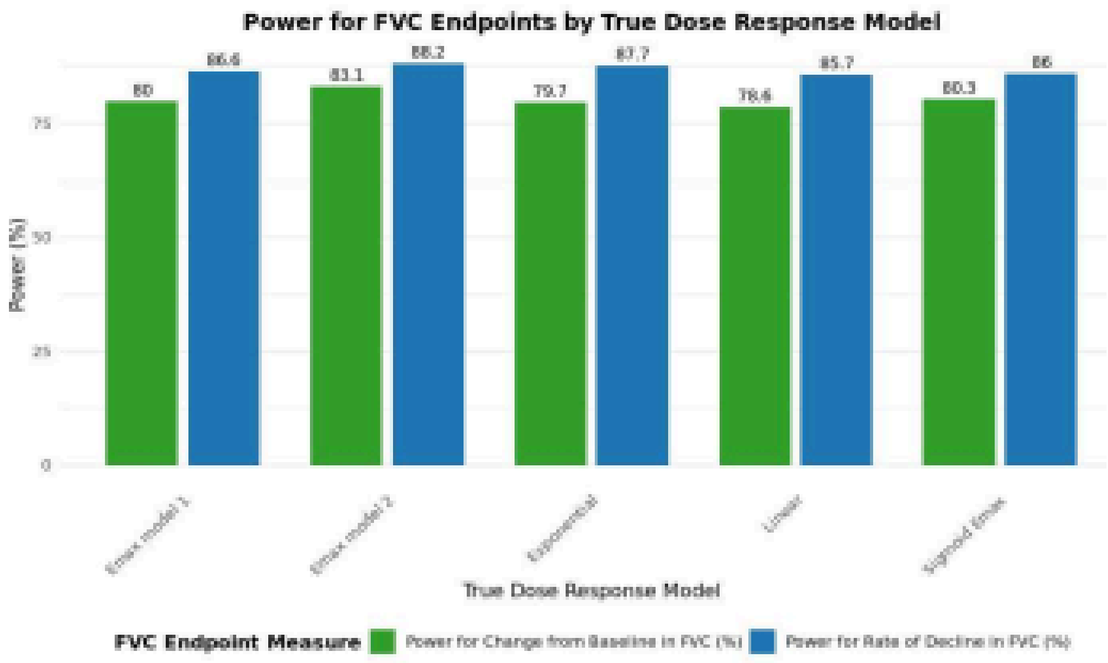


Figure 2. Simulation results depicting the probability of correctly declaring a non-flat dose-response for each of the underlying true dose-response models.

Conclusion

Modelling the rate of FVC decline increases statistical power compared with change from baseline in FVC (by 5.1-8%), while providing a more precise characterization of the dose-response. Combined with advanced statistical methods such as MCP-Mod, this enables a more efficient trial design with smaller sample sizes and improved decision-making in IPF studies.

References
1 Pinheiro et al 2014
Pinheiro J, Bornkamp B, Glimm E, Bretz F. Model-based dose finding under model uncertainty using general parametric models. Stat Med. 2014;33(10):1646-61. doi: 10.1002/sim.6062

Supported by AstraZeneca

Presented at the PSI, Belfast, June 2026.

Effective Study Closeout Following Early Termination: Points to Consider as a Biostatistician

P030

Tessa Lloyd, Nairouz Elgeioushi, Dominic Foong, Jeong Yoon, Yun He
AstraZeneca, Oncology Biometrics

Introduction

This poster presents a practical framework to support statisticians in planning and executing early study closeout for oncology trials, outlining key considerations, decision pathways, and collaboration points across functions to ensure robust, high-quality outcomes despite accelerated timelines.

Objectives

- To outline key factors essential to effective early termination closeout
- To describe common statistical and biometrics activities conducted before and after early clinical data lock (CDL)
- To provide practical, experience-based considerations to streamline closeout activities and minimise disruption to ongoing research

Overview of Early Study Closeout

Study closeout
Study closeout represents the final phase of a clinical study and includes:

- The completion of all data management activities, including data cleaning and CDL
- Final statistical analyses and production of Tables, Listings and Figures (TLFs)
- Clinical Study Report (CSR) authoring and publication
- Post-CSR activities such as clinical trial transparency (CTT) results posting, Trial Results Summary (TRS), and completion of the Trial Master File (TMF)

Early termination
Early termination refers to studies ending before completion as defined in the Clinical Study Protocol (CSP).

Early termination may arise due to:

- Safety concerns, including serious adverse events
- Futility based on interim analyses and pre-specified stopping rules
- Sponsor decisions driven by strategic or portfolio considerations
- Operational challenges such as recruitment difficulties, drug supply issues, or external events.

While early termination allows redirection of resources and focus, it introduces complexities that require rapid, coordinated action to ensure compliance, data integrity, and timely study closure.

Key Considerations Prior to Early Clinical Data Lock

Planning for Early Terminated Studies
Key questions to address with the study team include:

- Intended future use of the data within the overall clinical strategy
- Participant enrolment and treatment status at termination
- Requirements for unblinding in randomised blinded trials
- Collaboration obligations and communication plans
- CSR type required (full, abbreviated, or synoptic)
- CTT, publication, and presentation requirements

Prior to CDL	CSR and Early Closeout Planning	• Full CSR, Abbreviated CSR or Synoptic CSR • CSR delivery scope and timelines
	SAP and TLF Amendments	• SAP review and potential amendment • TLF scope evaluation, reduction, and finalization
	Dry Run Planning	• Dry run to support final CSR package planning and timelines
	CDL Planning	• Statistical input on required DM documents • CDL timelines and key study milestones
	Data Entry and Cleaning	• Review of entry and cleanliness status of critical data • Data transfers for TPV data • Data issues and mitigation plans
After CDL	Final Statistical Package and Clinical Interpretation	• Final TLFs, team review and interpretation
	CSR and Post-CSR Activities	• CSR and participant narratives • TRS, CTT publishing of results (RRF and ERF) • Other activities, incl. SAP/CSR redaction, publications, and TMF archival

CDL = Clinical data lock; CSR = Clinical study report; CTT = Clinical trial transparency; DM = Data management; ERF = EudraCT results form; RRF = Results registration form; SAP = Statistical analysis plan; TLF = Tables, listings, figures; TMF = Trial master file; TPV = Third party vendor; TRS = Trial results summary.

CSR and Early Closeout Planning

- Alignment of the delivery scope, timelines, and responsibilities across functions
- Early engagement with medical writing and publishing teams
- Selection of either the abbreviated or synoptic CSR formats, as is most common for early terminated studies
- Limitation of analyses to those necessary to support study objectives and regulatory needs

SAP and TLF Amendments

- Amendment of the Statistical Analysis Plan (SAP) or TLF specifications, as needed
- Performance of a risk impact assessment for last-minute changes
- Reduction of the planned analyses in line with abbreviated or synoptic CSR scope
- Timely initiation of additional inputs required for final analyses (e.g. other adverse events, important protocol deviations)

Dry Run Planning

- Performance of at least one dry run is required prior to CDL
- Assessment of data availability and risks associated with incomplete datasets, including external data such as pharmacokinetic (PK) datasets

CDL Planning

- Coordination with Data Management (DM) on CDL planning, data cut-off (DCO) timing, and external data reconciliation
- Performance of assessments for data incompleteness and delays
- Considerations of rollover or post-trial access programmes where applicable

Data Entry and Cleaning

- Monitoring backlog data entry and cleaning following early termination
- Resolution of critical data issues prior to CDL
- Documentation of unresolvable data issues and evaluate any impact to analyses
- Verification of key endpoint data (e.g. objective response rate (ORR), RECIST assessments)
- Ensuring completion and reconciliation of end of treatment (EOT), end of study (EOS), AE, laboratory, and third-party vendor data

Key Considerations After Early Clinical Data Lock

Final Statistical Package and Clinical Interpretation

- Coordination of final TLF timelines, reviews, and comment resolution meetings
- Differentiation between critical and non-critical comments
- Participation in clinical interpretation meetings when required
- Awareness of publication and additional analysis requests post-CDL

CSR and Post-CSR Activities

- Supporting CSR authoring and participant narrative development
- Results posting to public registries (RRF and ERF), aligned with CSR content
- Reviewing Trial Results Summaries for accuracy
- Completion of the CRT packages, redactions, and TMF archival activities

Key Message

- **Early termination demands early statistical leadership**
  Proactive involvement before and after CDL is essential to manage compressed timelines and added complexity.
- **Risk-based prioritisation enables efficient closeout**
  Focusing analyses, TLFs, and data cleaning on what is essential for study objectives and regulatory needs supports timely, high-quality delivery.
- **Alignment and collaboration are critical**
  Early agreement across functions on data use, CSR type, analysis scope, and transparency requirements reduces rework and delays.
- **Good planning protects data integrity and compliance**
  Structured closeout planning helps ensure robust results, clear clinical interpretation, and compliant reporting despite incomplete or evolving datasets.

Acknowledgement

Contributors and reviewers including: Jayne Marshall, Caroline Kennedy, Chris Gibbs, Grace Lu, Jason Clark, and Ganesh Samant

Supported by AstraZeneca

Presented at the PSI Conference 2026, Belfast, June 15th – 17th 2026

Introduction to the ITIT course

Zelie Bailes (GSK), Sophie Hodge (Bayer), Ruth Lowe (IQVIA), Zoe McIntyre (Phastar), Claudio Marchese (AstraZeneca), Hoi-Shen Radcliffe (Bayer), Sam Ruddell (Chiesi), Mahesh Shivhare (Roche), Marianne Uguen (Roche), Laura White (Plus-Project)

What is the ITIT Course?

Introduction to Industry Training Course (ITIT) aims to give a full overview of the pharmaceutical industry and the drug development process. The course is made up of six 2-day sessions.

Session 1 – Research
An introduction into pre-clinical research including target identification, high through-put screening, candidate selection, translational science, personalised healthcare.
Often included is a Biology/Chemistry **Lab tour**.
***Example Workshop**: Statistical issues in pre-clinical research.*

Session 2 – Toxicology
An overview of toxicology, and why it is important in R&D of new compounds. Areas within toxicology such as genetic and reproductive toxicology are explored. We also explore carcinogenicity, pharmacology and statistical approaches in toxicology studies.
***Example Workshop**: Toxicology issues in R&D*

Session 3 - Data Management and CROs
We delve into database structure and design, CRF design, data entry, validation, quality and transfer of data. We also look at the role and history of CROs, interactions between pharma and CROs, and scrutinize the bidding process.
***Example Workshops**: Designing CRFs, CRO bidding process.*

Session 4 - Clinical Trials
We focus on the overall Clinical Development Plan for phase I to III trials. Study Protocols (incl. sample size estimation), Statistical Analysis Plans, Clinical Study Reports, trial set-up process, ethics, and good clinical practice are also explored.
***Example Workshops**: Participating in a DMC, Study management team in crisis.*

Session 5 – Reimbursement
Outlines the economic evaluation of medicines, including methods and measures used in economic evaluation, and critical assessment of an economic evaluation. The use of network meta-analysis and systematic literature reviews are examined.
***Example Exercise:** to review a published systematic review.*

Session 6 – Medical and Market Access
Summarises steps post regulatory approval, role of market research for successful product launch and role of the sales function. Regulatory issues in drug promotion, increasing the generic market, and response to changes in government policy.
***Example Workshop**: Marketing plan for a hypothetical drug*

ITIT course

For more information and how to apply visit the PSI ITIT website or scan the QR code:

Where and when is it happening?

- ***Duration:*** October to July each year.
- ***Location***: each session is hosted by a different company within the industry and can change on a yearly basis.

What is included?

- Attendance at all 6 sessions
- One night Bed & Breakfast per session, two lunches, one evening meal per session.
- Social/Entertainment
- Not included: Travel, additional amenities.

Is it a statistical course?

- **No!** The content is not statistical but aims to give an insight into the drug development process of how compounds are discovered, the journey from discovery to clinical trials, to regulatory approvals and beyond.



2023/24 ITIT course delegates on a lab tour at Roche, Basel.

ATTENTION COMPANY HOSTS!

Would your company be interested in hosting one of the ITIT sessions?

The ITIT committee are eager to hear from companies based in both the UK and EU who would willing to host one of sessions.

What's involved?

As a host company you would, with the support of an ITIT committee member, organize the session, lab tours and workshops for the session and provide lunches for the two days.

How does it benefit my organization?

Increased visibility, the chance to showcase talent, facilities and capabilities, as well as encouraging networking and collaboration.

To find out more and register interest please contact the PSI ITIT committee chair Zelie Bailes (zelie.a.bailes@gsk.com)

A great chance to meet and share ideas with the statisticians and programmers who will take our industry into the future. Providing them with a glimpse into dynamics between the CRO and pharma world in bidding on and assessing bids for new studies, and having some fun along the way!

Adam Webb, IQVIA, DM and CRO Host 2023

MARS

The Versatile Statistician: Driving Value Beyond Clinical Trial Conduct

1. The Evolution of the Role From Gatekeeper to Strategic Architect

Historically, pharmaceutical statisticians were anchored primarily to the structural framework of the clinical trial protocol—focusing their expertise on the execution of Phase I–III trials and the production of Clinical Study Reports (CSRs). However, as the industry pivots toward data-driven agility and precision medicine, this role has expanded significantly. Today, statisticians are core architects of the entire drug development value chain. We function as cross-functional strategic leaders who drive early asset optimization, maximize product life cycle value, and bridge the critical gap between complex computational data science and high-stakes executive strategy.

2. Expanding the Statistical Horizon

The modern statistician drives innovation across **seven high-impact** domains

3. The "Beyond-Trial" Toolkit

To succeed in this expanded role, the statistician requires a hybrid skill set

Technical: R/Python proficiency, Causal Inference, Bayesian Analysis.

Strategic: HTA knowledge, Regulatory Diplomacy, Value Communication.

Ethics: Expertise in Data Privacy and Algorithm Fairness.

CAPABILITY AREA	PRIMARY IMPACT	KEY METHODS
Early Discovery & Translational Science	Faster molecule selection and better target populations	Predictive modeling, Omics integration
Data Science & Automation	Open, scalable, and reproducible research	Automated R/Python pipelines
Pharmacovigilance & Drug Safety	Continuous benefit–risk assessment	Bayesian signal detection
Regulatory & Market Access	Reimbursement and HTA readiness	ITCs, NMA, NICE, G-BA
Post-Marketing & RWE	Long-term outcomes beyond RCTs	Real-world data, registries
Data Transparency & Disclosure	Trust, compliance, and ethical use	Anonymization, re-identification risk analysis
Data Storytelling	Executive alignment and decisions	Clear, compelling visualizations

Conclusion

The **"Trial Statistician"** is now a **Value Architect.** By moving beyond trial conduct, we ensure that **innovative therapies** are not just statistically significant, but commercially viable and accessible to the patients who need them most.

P033

Bayesian Meta-Analysis for Phase 2 Endpoint Selection: Predicting Phase 3 Outcomes to Optimize Trial Design

Enti Spata[1], Carla A. Da Silva[2], Lynn Dunsire[1], Karin Bowen[3], Patrick Darken[3], Alexandra Jauhiainen[4]

[1]R&I Biometrics and Statistical Innovation, Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK; [2]Digital Health Strategy & Implementation, Respiratory & Immunology, Evinova, Gothenburg, Sweden; [3]R&I Biometrics and Statistical Innovation, Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, US; [4]R&I Biometrics and Statistical Innovation, Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden.

Abstract

A key goal in chronic obstructive pulmonary disease (COPD) is to prevent episodes of acute worsening of COPD symptoms called exacerbations. This endpoint demands large, lengthy trials, which is challenging in early drug development. Alternatives include lung-function endpoints (notably change from baseline in pre-bronchodilator forced expiratory volume in 1 second, FEV_1) and the composite endpoint COPDCompEx, which enriches exacerbations by adding events defined from deteriorations in diary card variables [1].
We evaluated the trial-level surrogacy properties of COPDCompEx and FEV_1 to COPD exacerbations with Bayesian models using data from 9 clinical trials. FEV_1 showed limitations in its ability to quantify the effect reported on COPD exacerbations across drug modalities. In contrast, COPDCompEx was confirmed as a surrogate in terms of predicting treatment effects observed on exacerbations across modalities.

Introduction

- A key goal of treatment for patients with chronic obstructive pulmonary disease (COPD) is to prevent episodes of moderate-to-severe exacerbations which are acute worsening of COPD symptoms requiring either a course of systemic steroids and/or antibiotics or hospitalization.
- Designing trials for COPD exacerbations is a challenge, especially in early phases, as the studies tend to be large and lengthy.
- Alternative endpoints to use in early phase trials can be lung function measurements like FEV_1, or the novel endpoint COPDCompEx, which is a composite endpoint enriching COPD exacerbations by adding events defined from deteriorations in diary card variables.
- We evaluated FEV_1, COPDCompEx, and COPD exacerbations using patient level data across a set of 9 trials with 21 treatment comparisons (21,713 subjects randomised) to assess their trial-level relationship.

Methods

Individual trial analyses

Change from baseline in pre-bronchodilator FEV_1 was analyzed with a repeated measures model. Time to first COPDCompEx was assessed with a Cox Proportional Hazards model. Recurrent COPD exacerbations were evaluated with a negative binomial model. All models included trial pre-specified covariates.

Meta- and surrogacy analysis

Bayesian bivariate random-effect meta-analysis (BRMA) was applied to estimate the total correlation for treatment effects for FEV_1 and COPDCompEx with COPD exacerbations [2-3].

To estimate the surrogacy relationship between COPDCompEx/FEV_1 and severe exacerbations, a method following Daniels & Hughes [4] was adopted. Let $i = 1,2,\dots,n$ denote the treatment comparisons performed. Further, θ_i and γ_i denote the true treatment effects on severe exacerbations and COPDCompEx/FEV_1, respectively, in the ith trial. The stage 1 model relates the estimated and true treatment effects in the ith trial by

$$\begin{pmatrix}\hat{\theta}_i \\ \hat{\gamma}_i\end{pmatrix} \sim N\left(\begin{pmatrix}\theta_i \\ \gamma_i\end{pmatrix}, \begin{pmatrix}\sigma_i^2 & \rho_i\sigma_i\delta_i \\ \rho_i\sigma_i\delta_i & \delta_i^2\end{pmatrix}\right).$$

Here σ_i^2 and δ_i^2 reflect sampling variation and ρ_i the correlation between the estimated treatment differences conditional upon the true differences. In stage 2, a linear model is assumed to describe the relationship between θ_i and γ_i across the clinical trials:

$$\theta_i|\gamma_i \sim N(\alpha + \beta\gamma_i, \tau^2)$$

In this model a good surrogate is expected to result in a non-zero coefficient β (showing there is an association between the treatment effects), a value of zero for α (showing that no treatment difference on the surrogate endpoint implies no treatment difference on the clinical endpoint) and a value close to zero for τ (little heterogeneity of the association explored). Unimformative prior distributions were used as follows: $\gamma_i \sim N(0,1000)$, $\alpha \sim N(0,1000)$, $\beta \sim N(0,1000)$, $\rho_i \sim U(-1,1)$ and $\tau \sim U(0,2)$. The parameters of the model were estimated by Markov chain Monte Carlo simulation using WinBUGS version 14.

Acknowledgements

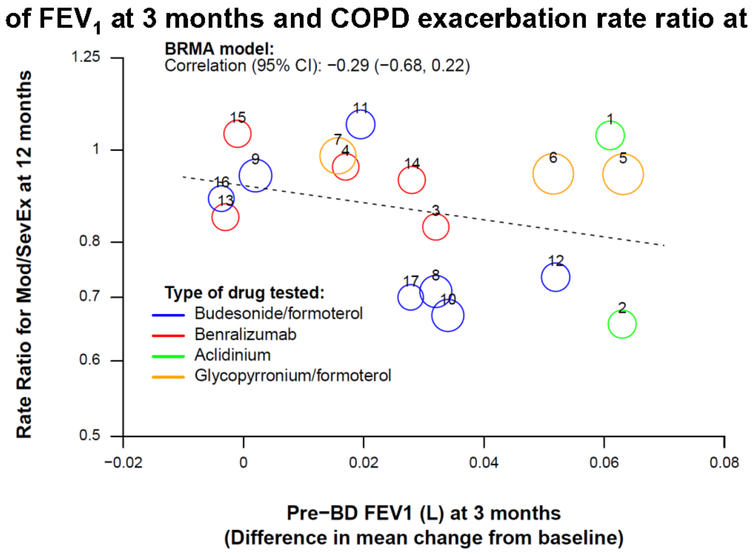
We would like to thank the individual program teams at AstraZeneca who provided trial results to support this work.

Results

FEV_1 as a surrogate for COPD exacerbations

The change from baseline in FEV_1 at 3 months was used to emulate the shorter duration of an early phase trial. Its correlation to the preferred phase 3 endpoint, COPD exacerbations evaluated using a rate ratio over 12 months, was weak (-0.29; Figure 1). Overall, FEV_1 showed limitations in its ability to quantify the effect reported on exacerbations across drug modalities.

Figure 1: Relationship between the difference in mean change from baseline of FEV_1 at 3 months and COPD exacerbation rate ratio at 12 months



BD, bronchodilator; FEV1, forced expiratory volume in 1 second; Mod/SevEx, moderate-to-severe COPD exacerbation; CI, credible interval.

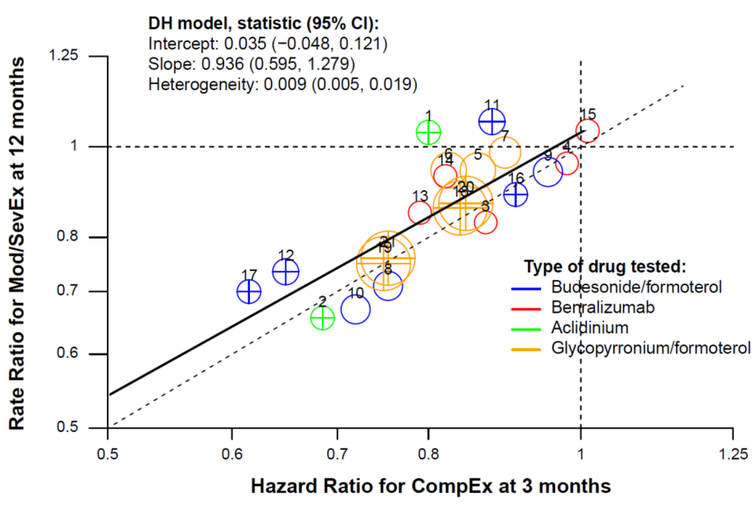
The size of the circle is analogous to the total sample size analysed.

The following trials were included for analysis: (1) ACCLAIM-I, (2) ACCLAIM-II, (3-4) GALATHEA (5-7) PINNACLE-3, (8-10) SOPHOS, (11-12) SUN, (13-15) TERRANOVA and (16-17) US3

COPDCompEx as a surrogate for COPD exacerbations

It is reasonable to posit that the 3-month time-to-first COPD exacerbation is the best surrogate for exacerbation rate ratio over 12-months, but exacerbations are infrequent, making phase 2 sample sizes impractical. COPDCompEx enables smaller, more efficient samples in COPD trials of moderate-to-very severe patients. For example, in SUN trial, detecting a HR=0.70 at 80% power 10% alpha with COPDCompEx required 37% fewer participants vs exacerbations alone (285 vs 453 per arm).
As with FEV_1, COPDCompEx was censored at 3 months to emulate an early phase trial. The COPDCompEx hazard ratio at 3 months correlated well with the 12-month rate ratio observed on severe exacerbations (0.81; Figure 2). COPDCompEx was confirmed as a surrogate in terms of predicting treatment effects observed on COPD exacerbations, with a high level of correspondence between the endpoints across drug modalities (Figure 2).

Figure 2: Relationship between COPDCompEx hazard ratio at 3 months and COPD exacerbation rate ratio at 12 months.



BRMA correlation (95% CI): 0.81 (0.58, 0.92).

COPDompEx, composite endpoint for COPD exacerbations; Mod/SevEx, moderate-to-severe COPD exacerbation; CI, credible interval; DH, Daniels & Hughes. Crossed circle symbols correspond to CompEx including Peak expiratory flow and plain circles symbols correspond to CompEx without Peak expiratory flow.The size of the circle is analogous to the total sample size analysed.

Same trials were included as in Figure 1, with the addition of (18-21) ETHOS.

Conclusions

- FEV_1 remains an important respiratory endpoint, especially for drugs with bronchodilating properties, but has limitations as a primary phase 2 endpoint across modalities when the aim is to target severe exacerbations in phase 3.
- COPDCompEx at 3 months has an increased event frequency compared to COPD exacerbations alone and it is found to be a good predictor of treatment efficacy reported with COPD exacerbation rates over 12 months. This makes COPCompEx an attractive endpoint to use in design of early phase trials across a range of modalities, substantially reducing sample sizes needed.

References

1. Vogelmeier CP, Fuhlbrigge A, Jauhiainen A, Da Silva CA et al. COPDCompEx: A novel composite endpoint for COPD exacerbations to enable faster clinical development: a post-hoc analysis of randomised controlled trials. Respiratory Medicine. 2020;106175.
2. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002; 21(4): 589-624.
3. Riley RD, Abrams KR, Lambert PC, et al. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med 2007; 26(1): 78-97.
4. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. Stat Med. 1997;16(17):1965-82.Caflisch R.E. et al (1998). Monte Carlo and Quasi-Monte Carlo methods. Acta Numerica: 1–49.

Supported by AstraZeneca

Presented at PSI, Belfast, Ireland, June 2026

P034

Recurrent events made simple: Translating disease burden into five scenarios

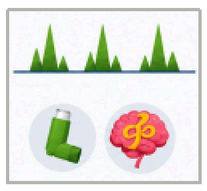
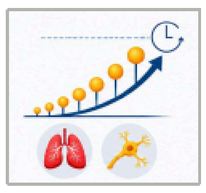
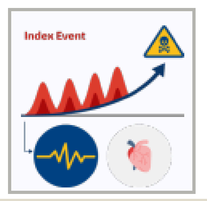
Tomoko Iwata, Nuala Peter
Boehringer Ingelheim Pharma GmbH & Co. KG, Biostatistics and Data Sciences, Biberach, Germany

Recurrent events

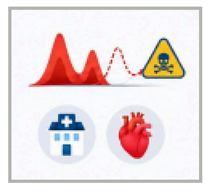
- Refer to the repeated occurrence of the same type of event over time for the same subject
- Should be related to disease burden and may indicate disease progression in some instances

Proposed framework

Recurrent event analysis methods

Scenario 1 Recurrent events – count	Scenario 2 Recurrent events – time	Scenario 3 Recurrent events – accounting for dependence within endpoints
Disease pattern Chronic disease, stable  e.g. Asthma, epilepsy	**Disease pattern** Chronic disease, progressive  e.g. Multiple sclerosis, COPD, Chronic heart failure	**Disease pattern** Index event + recurrents  e.g. Heart failure hospitalisations, Acute myocardial infarction
Clinical Question Does treatment decrease number of events compared to control over an observation period?	**Clinical Question** Does treatment reduce the event rate in the study population compared with placebo?	**Clinical Question** How many events occur accounting for dependence of death and recurrents?
Estimator: Rate ratio	**Estimator:** Rate ratio (/Hazard ratio)	**Estimator:** Hazard ratio
Handling of ICE of terminal event (e.g. death) in ICH E9 (R1) Estimand "While on treatment strategy" Observation period ends at terminal event (e.g. death).	**Handling of ICE of terminal event (e.g. death) in ICH E9 (R1) Estimand** "While on treatment strategy" Observation period ends at terminal event (e.g. death).	**Handling of ICE of terminal event (e.g. death) in ICH E9 (R1) Estimand** "Hypothetical (and competing risk) strategy" Consider terminal event (e.g. death) as if it never occurred.
Analysis method Negative binomial[1] • Extension to Poisson regression • Assumes a constant event rate	**Analysis method** Proportional rate model, e.g. Lin-Wei-Yang-Ying[2] • Extension to Cox proportional hazards model to allow recurrent events	**Analysis method** Joint Frailty[3] • Adjust for a disease related death as competing risk (joint frailty) • Requires complex implementation

Time to event analysis methods

Scenario 4 Time to first event – composite of (cause specific) terminal event	Scenario 5 Time to first event – without composite of terminal event
Disease pattern Index event + recurrent or (cause specific) terminal  e.g. Heart failure hospitalisation or cardiovascular death	**Disease pattern** Index event + recurrent  e.g. Heart failure hospitalisation(s)
Clinical Question Does treatment delay the first occurrence of the event?	**Clinical Question** a) What is the probability of being hospitalised before death? b) Among patients who remain alive, how does treatment affect time to hospitalisation?
Estimator: Hazard ratio	**Estimator:** Hazard ratio
Handling of ICE of terminal event (e.g. death) in ICH E9 (R1) Estimand • "Composite strategy": Include cause specific terminal event (e.g. cardiovascular death) in definition of endpoint. • "While on treatment strategy": Observation period ends at terminal event (e.g. non-cardiovascular death).	**Handling of ICE of terminal event (e.g. death) in ICH E9 (R1) Estimand** a) "Competing risk strategy": Handles terminal event (e.g. death) as competing risk event with sub-distribution hazard. b) "While on treatment strategy" Observation period ends at terminal event (e.g. death).
Analysis method Cox proportional hazards[4]	**Analysis method** a) e.g. Fine and Gray[5] b) e.g. Cause-specific Cox proportional hazards

Key message

- When designing a clinical study, understanding recurrent process and the role of terminal event are key.
- This framework helps to identify the estimand and clinical question of interest.
- Having identified the relevant scenario (1 to 5), the analysis methods are relatively straightforward.

Reference

1. Hilbe 2011: Negative Binomial Regression
2. Lin 2000: Semiparametric regression for the mean and rate functions of recurrent events
3. Rogers 2016: Analysis of recurrent events with an associated informative dropout time: application of the joint frailty model
4. Cox 1972: Regression Models and Life-Tables
5. Fine 1999: A proportional hazards model for the subdistribution of a competing risk

COPD: Chronic Obstructive Pulmonary Disease;
ICE: Intercurrent Event

Using Influence Science to Communicate and Advocate for Innovative Bayesian Trial Designs.

Claire Smith[1], Xiao Xiang[2]
AstraZeneca UK[1], AstraZeneca Beijing[2]

P036

The Challenge

Bayesian designs offer powerful advantages—faster decisions, smaller samples, adaptive strategies. But **innovation adoption isn't just about having the right method; it's about bringing people on the journey with you.**

Every stakeholder interaction is an opportunity to:
- Build **understanding** of novel concepts
- Create **alignment** around shared goals
- Foster **confidence** in new approaches
- Generate **enthusiasm** for innovation

The opportunity: Transform technical discussions into compelling narratives that build confidence, alignment, and momentum.

Why Influence Science?

"Even when you are not communicating, you are communicating something."

Every interaction—every email, every slide, every design choice—sends a message about complexity, risk, and trust.

Influence science gives us a structured, ethical framework to communicate with intention.

We applied **Cialdini's Levers of Influence**—evidence-based strategies from psychology—to create a practical approach for advancing Bayesian innovation via **effective story-telling**

It's not about persuasion—it's partnership.

Cialdini's Levers

7 Key Influence Levers
+2 additional



The Ethical Foundation

Transparency
Honest about uncertainties & limitations?

Mutual Value
Creating genuine benefit for stakeholders?

Respect
Honouring their expertise & autonomy?

Consistent ethical practice builds the trust that accelerates innovation adoption.

An Illustrative Decision Journey

1 Initial Exploration
Stakeholder State: Curious but cautious

Principle:
Unity
Liking
Action: Frame design in terms of shared goals
Use accessible language, tailored visualisations

Contrast
Authority
Show side-by-side metrics
Cite regulatory guidance

Social Proof
Reciprocation
Scarcity
Share precedence
Address concerns collaboratively, identify safe fallbacks
Highlight time-sensitive opportunities

Commitment
Consistency
Start with pilot
Sequence small agreements

2 Design Comparison
Evaluating options

3 Risk Assessment
Concerned about unknowns

4 Decision & Implementation
Ready to commit

Case Study Snapshot

1 Initial Exploration

Unity Shared goals, identify constraints
Reciprocation Create value early by modelling good communication, collaborative problem-solving, engage with diverse perspectives

Task: Designing a phase 2 study

Key questions:
- Can this drug achieve meaningful efficacy? – proof of concept
- What's the optimal dose?

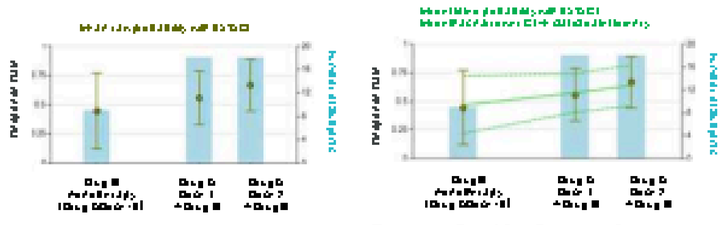
Challenges:
- Small population, sample size constraint
- Time sensitive window to deliver study

Opportunity: Achieve statistical efficiency by use of a Bayesian Emax dose response model

Liking Visualisations enhance understanding
Because Clear rationale linking method to impact

Traditional Approach vs Bayesian Emax Approach

Illustrative simulated dataset Shows Efficiency Gains



- Analyses each dose **independently**
- Missed opportunity to incorporate learnings across the full dose range into Go-NoGo decisions

- **Better Go-NoGo performance** OR same performance with **smaller sample size** *because**
- The model assumes efficacy increases or plateaus with dose, borrowing information across doses & **reducing uncertainty**

* Even using 'because' can be a helpful influencing trigger word!

2 Design Comparison

Contrast Side-by-side comparisons of alternatives e.g. with vs without borrowing, vary N

Bayesian model gives 10% power gains vs traditional

	Bayesian Emax Each dose informs others via model	Traditional Approach Each dose assessed independently
Total Sample size	90 3 arms	90 3 arms
Time to primary analysis	21m	21m
Go Prob. All doses active / Highest dose only	74% / 69%	73% / 59%
Early Interim mGo prob. All doses IV / Highest dose only	55% / 55%	55% / 45%
Pros	~10% increase in power ~10% increased chance to meet Interim Go	Simple design
Cons	Novel in Oncology	Missed opportunity to achieve optimal performance

Key metrics:
- Time
- Cost
- Decision Robustness

Authority Reference evidence-backed expertise, seek endorsement of internal advisory committees, align to regulatory guidance

3 Risk Assessment

Social Proof Share precedence, highlight industry adoption and case studies

Evidence-based statements: e.g. data from "*source*" shows a clear upwards trajectory in adoption of Bayesian designs.
Building on precedence: dose response models used successfully in analysis for project "*X*", we are building to inform design stage
Regulatory acceptance:
- **FDA fit for purpose** designation for Emax
- **Consistent** with approaches outside of Oncology
- **FDA** Bayesian guidance 2026[2]

Reciprocation Address concerns collaboratively, identify an acceptable fallback to mitigate risks

Risks	Likelihood	Impact	Mitigation / Actions plan
Regulatory acceptability Emax monotonicity assumption invalid e.g. safety signal impacts exposure	L	L	Revert to traditional frequentist pairwise comparison to control for which study is used to still have acceptable performance characteristics

4 Decision & Implementation

Commitment & Consistency Innovation doesn't stop! Socialise the concepts. Sequence: Pilot → Learn → Scale with consistency

Take This Home

Tailor Approach
Clinical → may value Unity & Social Proof. Operations → Contrast & Commitment

Build Momentum
Sequence small steps. Criteria, then models, then decisions.

Quantify Impact
"6 months faster" "20% fewer patients" Connect methods to patient benefit

Partner
Collaborate early. Seek diverse perspectives, invite on the journey

This framework translates statistical innovation into organizational impact - accelerating therapies to patients.

References
1. Cialdini R B. Influence: The Psychology of Persuasion. Revsd. Harper Business; 2006
2. Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products. Guidance for Industry – FDA 2026

Acknowledgements
red[10] People Development www.red10dev.com
Trial design simulations using FACTS™

Supported by AstraZeneca

P037

Balancing Complexity in EU-HTA

A simulation study on the impact of incomplete PRO data collection on treatment effect estimates

Alexandra Lauer[1], Katharina Musiolik[2], Emmanuel Bourmaud[3], Muriel Danten[4], Orlando Doehring[5], Evegniia Artamonova[1], Berit Hunsdieck[2], Sory Taoré[6]

[1]Boehringer Ingelheim Pharma GmbH & Co. KG, Germany [2]AbbVie Deutschland GmbH & Co. KG, Germany [3]Eli Lilly and Company, France [4]Merck Santé S.A.S., Lyon, France, an affiliate of Merck KGaA, Darmstadt, Germany [5]Novartis Pharmaceuticals UK Ltd., UK [6]Lilly and Company, UK

PSI 2026
ICC Belfast · June 15–17
Poster number: P037

BACKGROUND

Since January 2025, the ***Joint Clinical Assessment (JCA)*** provides the basis for joint benefit assessment of new cancer therapies across EU member states. In these assessments ***Patient-Reported Outcomes (PROs)*** are crucial, reflecting patients' self-perceived quality of life.

Collecting PROs after disease progression (PD) is operationally challenging. PRO data collection is therefore often stopped at PD — but long-term follow-up is needed to derive a ***treatment policy estimand***, which is considered key for HTA bodies, as it reflects the estimation of a treatment effect against the evolving standard-of-care treatment landscape.

As a result, balancing operational challenges and optimizing PRO assessments is key.

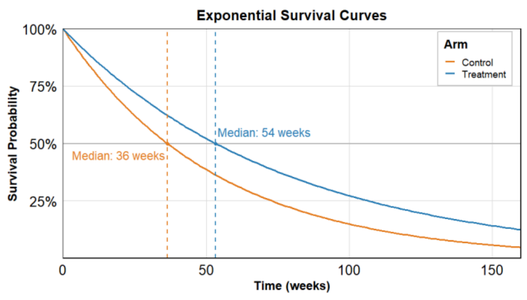
OBJECTIVE

Assess the impact of stopping PRO collection at PD on treatment effect estimates from event-time and responder analyses, under different data collection and deterioration timing scenarios.

STUDY DESIGN

Design	2-arm RCT, 1:1, 153 weeks
N	600 subjects
Primary Endpoint	Progression-free Survival, assuming HR of 0.684
PRO Instrument	EORTC QLQ-C30 (Scale range 0–100)
MID*	≥ 10 points
PRO Schedule	Every 3 weeks until week 27, then every 9 weeks
Censoring	Administrative at week 153

SIMULATION FRAMEWORK



Assumption on median PFS (based on HR = 0.684) and survival function

$$S(t) \sim \begin{cases} e^{-0.019\,t} & \textit{for control group} \\ e^{-0.013\,t} & \textit{for treatm. group} \end{cases}$$

Baseline:

$$P_i(0) \sim N(70, 20^2), \textit{ truncated to [0, 100]}$$

Longitudinal model (visit $t = 1, \ldots, 23$):

Longitudinal PRO scores for subject i, $i \in \{1, \ldots, 600\}$, at time/visit $t, with\ t = 1, \ldots, 23$ are modeled by:

$$P_i(t) = P_i(0) + \beta_1 * t + \beta_2 * t * Trt_i + \varepsilon_{it}$$

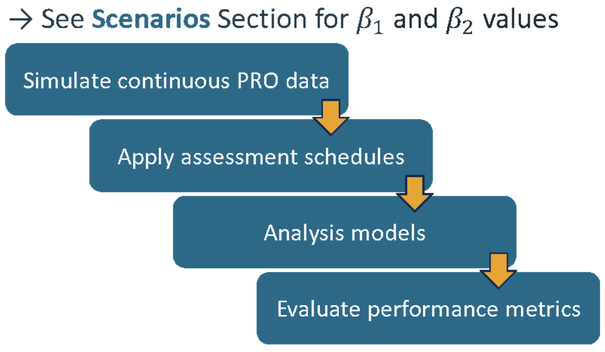
with

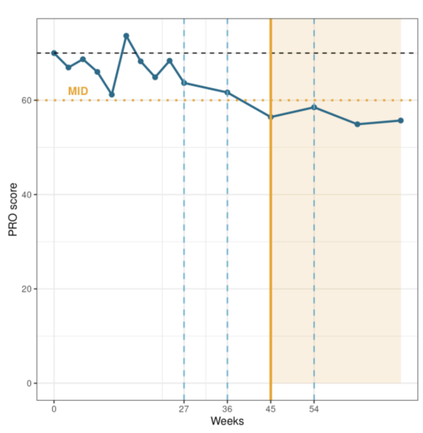
- $P_i(t)$: Subject-level PRO score at visit t,
- Trt_i: Treatment assignment
- ε_{it}: Patient-level correlated error with compound symmetry structure for the covariance matrix

and parameters

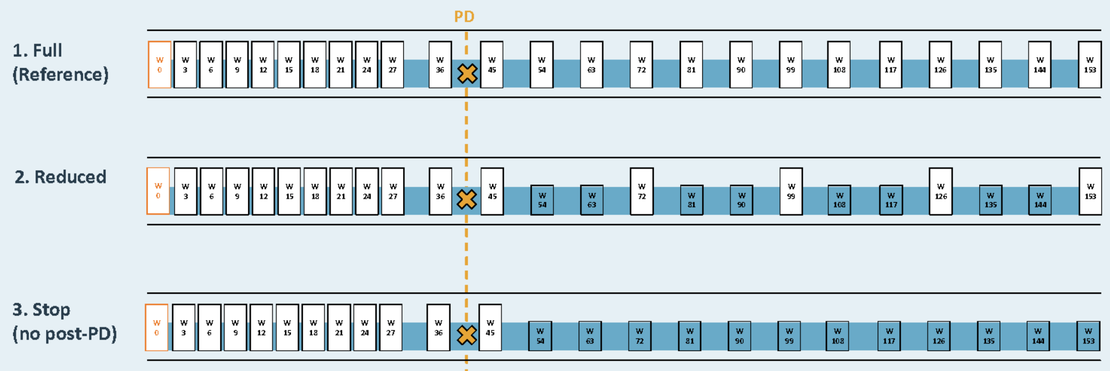
- β_1: Time effect
- β_2: Treatment × time interaction

→ See **Scenarios** Section for β_1 and β_2 values





SCENARIOS (3 × 3 = 9)



Deterioration timing (control / treatment):

Three PRO deterioration scenarios were evaluated:

a) Deterioration concurrent with PD
b) Deterioration before PD
c) Deterioration after PD

Beta values from Simulation:

	β_1	β_2
a) Deterioration concurrent with PD	-0.297	0.099
b) Deterioration before PD	-0.396	0.158
c) Deterioration after PD	-0.238	0.068

ANALYSIS METHODS

1 **Time to First Deterioration:** *First visit with Δ from baseline ≤ −10 points*
Analysed via Cox proportional hazards model including the covariates treatment and baseline PRO score. The effect size quantified as hazard ratio and median time difference.

2 **Response at Landmark Timepoints:** Δ from baseline ≤ −10 pts at week 27, 36, 54
Analysed via logistic regression including the covariates, treatment and baseline PRO score. The effect size is reported as risk difference.

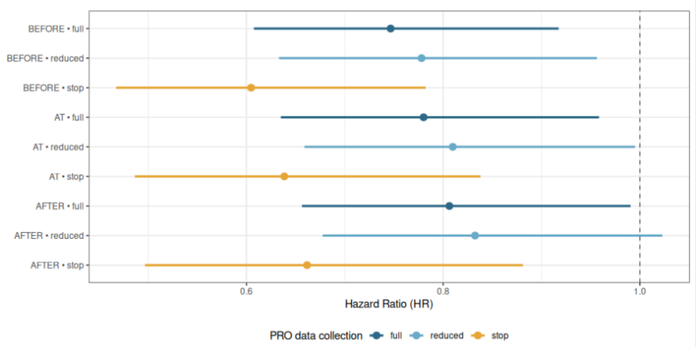
Performance metrics: Bias quantified vs. full-collection reference, power and coverage

RESULTS

Time to Event

Reduced Scenario delays detection of deterioration
→ positive bias
→ lower power, but comparable coverage

Stop Scenario truncates event detection at time of PD
→ Roughly same (but lower) number of events in both arms **but** longer follow-up under treatment due to benefit in PFS, therefore large negative bias
→ higher power, but very low coverage leading to erroneous confidence in effect



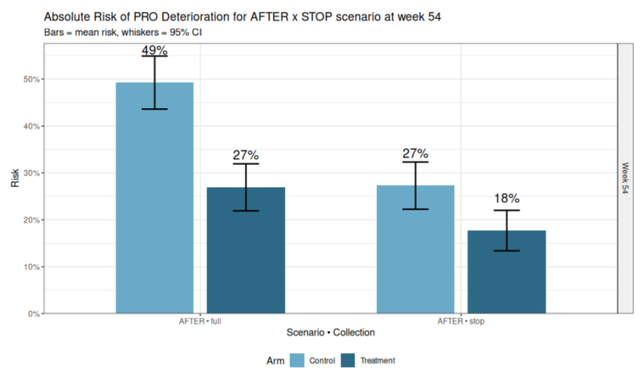
Performance Characteristics

Deterioration...	Power	Coverage
Before PD – **full** scenario	0.76	0.95
Before PD – **reduced** scenario	**0.66**	0.94
Before PD – **stop** scenario	**0.97**	**0.66**
Concurrent with PD – **full** scenario	0.65	0.94
Concurrent with PD – **reduced** scenario	**0.52**	0.93
Concurrent with PD – **stop** scenario	**0.90**	**0.71**
After PD – **full** scenario	0.55	0.95
After PD – **reduced** scenario	**0.41**	0.95
After PD – **stop** scenario	**0.82**	**0.73**

Response

Reduced Scenario with similar performance characteristics to time to event endpoint

Stop Scenario relies on patients to reach landmark timepoint (under the treatment policy strategy)
→ positive bias (more events in arm with longer follow-up)
→ low power and coverage



CONCLUSION

This cross-industry simulation study (EFSPI HTA SIG) supports health technology developers in designing PRO assessment schedules that enable JCA-relevant treatment policy estimation, balancing statistical robustness with operational feasibility. The reduction of PRO data collection frequency or the stop introduces bias and decreases power for both types of endpoints. A stop of PRO data collection after PD is not advisable. The magnitude of bias associated with the reduced data collection increases with the sparsity of PRO data collection after progression of disease.

Scan for more info

*MID: Minimal Important Difference – smallest difference in assessment value that is acknowledged as clinically relevant and therefore considered in the assessment of PRO analysis. In Germany for EORTC QLQ-C30 this is 10 points, which was used as reference.

This work represents the views of the authors, not necessarily those of their employers.

PSI 2026 • P038

Between-Study Heterogeneity in Two-Study Bayesian Meta-Analyses of German HTA Dossiers

Empirical evidence on the appropriateness of τ priors
FRIEND OR FOE?

Annett Kucka, SERVIER, Munich, Germany; Marcus Brunnert, UCB, Monheim am Rhein, Germany

KEY MESSAGE

- German HTA dossiers frequently rely on meta-analyses with only 2 studies.
- Empirical heterogeneity τ is usually zero or near zero.
- Common half-normal priors overstate heterogeneity in this setting.
- Bayesian random-effects results can therefore become prior-driven.

BACKGROUND

- Estimating between-study heterogeneity is unstable when evidence is sparse.
- Lilienthal et al. (2024) proposed empirical half-normal priors for τ.
- We assessed whether these priors match German HTA dossiers restricted to 2-study meta-analyses.

METHODS

- Publicly available G-BA / IQWiG benefit assessments.
- Curated dataset restricted to meta-analyses with exactly 2 studies (n = 58).
- Therapeutic areas: oncology, neurology, dermatology.
- Compared empirical τ with HN(0,1), HN(0,2), and HN(0,3).

INTERPRETATION

- 72.4% of τ estimates are exactly 0.
- The 90th percentile is only 0.124.
- Fixed-effect analyses are often justified in 2-study HTA settings.
- If Bayesian models are used, heterogeneity priors should be HTA-specific.

REFERENCES

- Lilienthal J, Sturtz S, Schürmann C et al. Res Synth Methods. 2024;15(2):275-287.
- Data source: curated extraction from published German HTA benefit assessments.
- General IQWiG methods paper, Version 8.0

CONCLUSION

- Minimal heterogeneity is the rule, not the exception, in these 2-study dossiers.
- The observed data do not support large default half-normal priors.
- Methodological guidance should distinguish sparse HTA evidence from broader meta-analysis settings.

Disclosure: All authors are employed by their respective institutions and are remunerated accordingly; some authors may also hold equity interests. This work is methodological research; no additional conflicts of interest were declared.

RESULTS

58	0	72.4%	HN(0,1)
meta-analyses	median τ	with τ = 0	best prior fit

Figure 1. Empirical τ distribution versus common half-normal priors

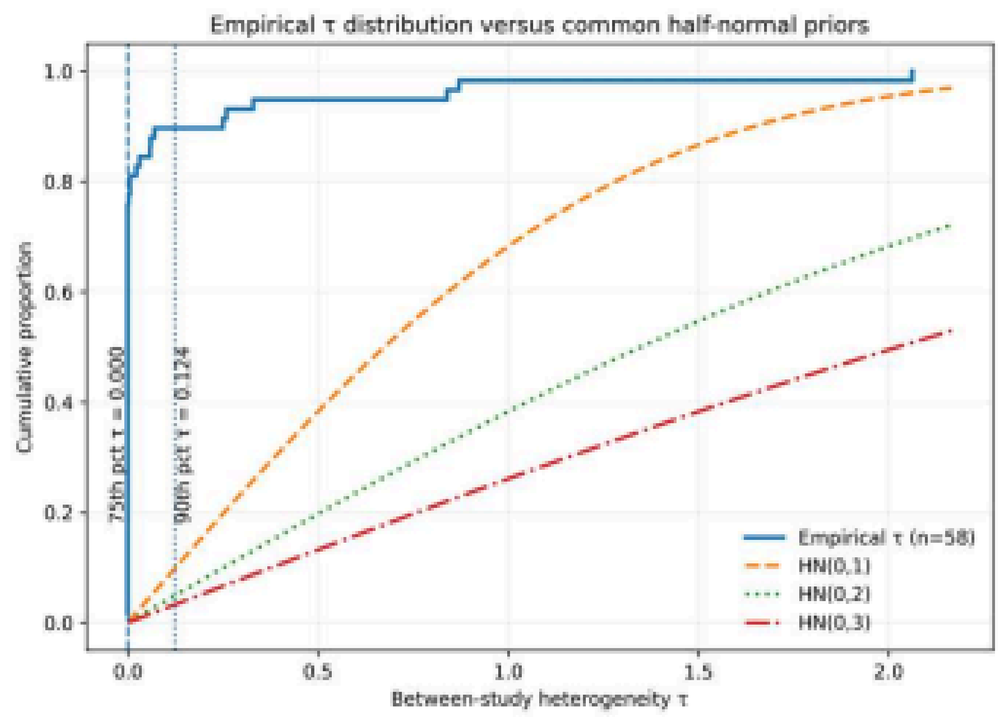
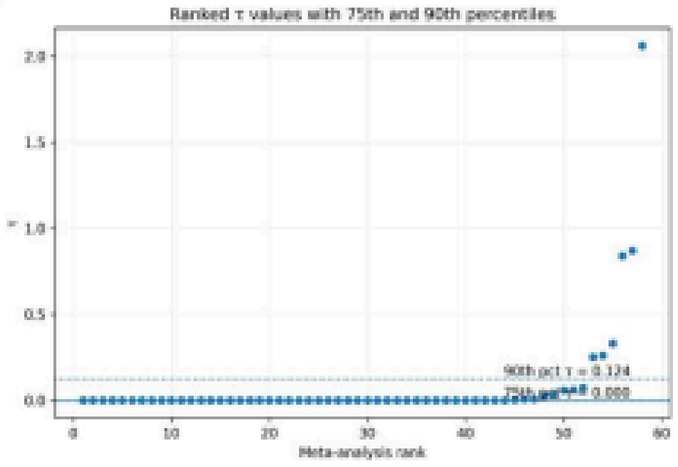
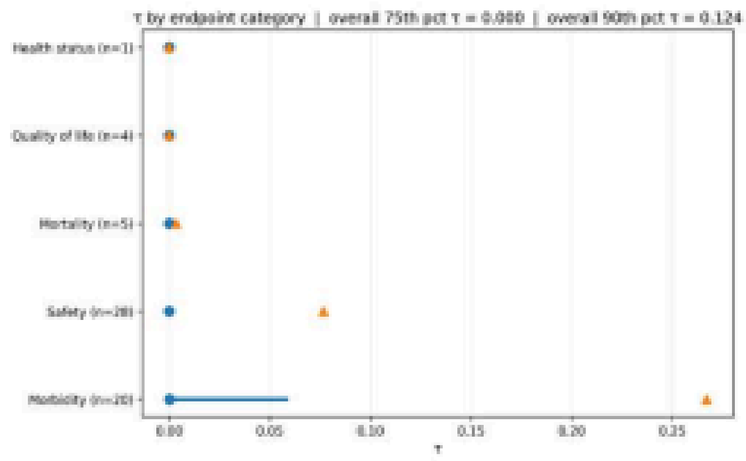


Figure 2. Ranked τ values



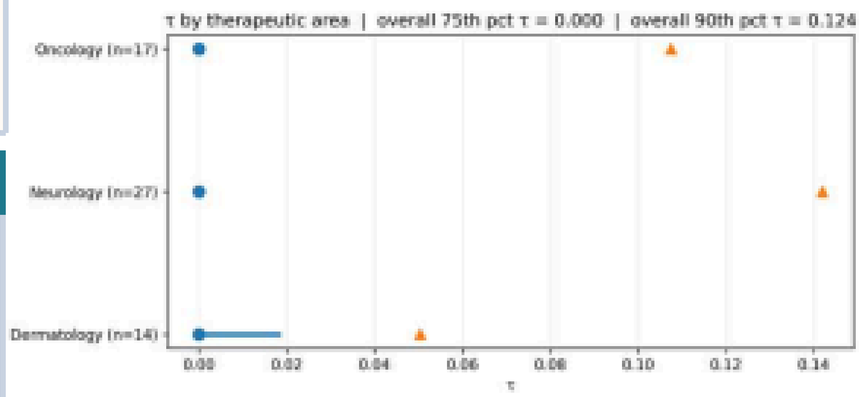
Most estimates are tightly concentrated at 0, with only a few moderate outliers.

Figure 3. τ by endpoint category



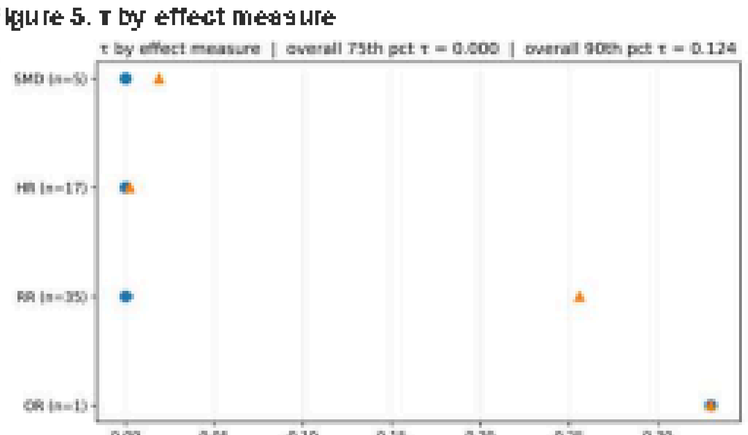
Safety and mortality remain near 0; morbidity shows the widest but still small spread.

Figure 4. τ by therapeutic area



The low-heterogeneity pattern is consistent across oncology, neurology, and dermatology.

Figure 5. τ by effect measure



RR and HR dominate the dataset and both show medians at 0.

Table 1. Summary of empirical τ distributions

Subset	n	Median τ	75th pct τ	90th pct τ	Share τ=0
All meta-analyses	58	0.000	0.000	0.124	72.4%
Dermatology	14	0.000	0.019	0.060	50.0%
Neurology	27	0.000	0.000	0.142	77.8%
Oncology	17	0.000	0.000	0.108	82.4%
Health Status	1	0.000	0.000	0.000	100%
Morbidity	20	0.000	0.059	0.267	50.0%
Mortality	5	0.000	0.000	0.004	80.0%
Quality of life	4	0.000	0.000	0.000	100%
Safety	28	0.000	0.000	0.077	82.1%
HR	17	0.000	0.000	0.002	88.2%
OR	1	0.330	0.330	0.330	0%
RR	35	0.000	0.004	0.256	68.6%
SMD	5	0.000	0.000	0.019	60.0%
Best prior fit: remains HN(0,1), but even this prior still lies far above the empirical ECDF					

From junior to lead: Key skills developed working in an IDAC

Rhys Jones | Statistician III
PHASTAR

Why DMCs matter

Data Monitoring Committees (DMCs) provide independent oversight of clinical trials to protect **participant safety** and assess **treatment effect**.

Most valuable in trials that are:

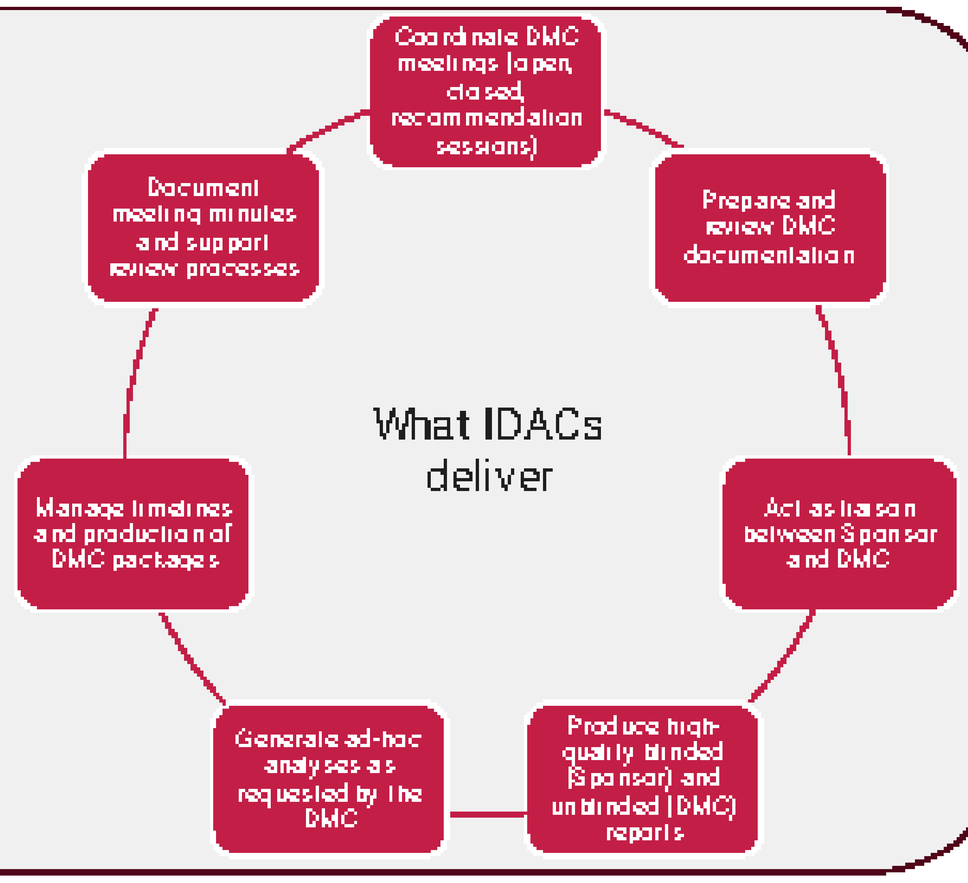
- High risk or operationally complex
- Involving vulnerable populations
- Using interim analyses that could affect conduct

The IDAC role

Independent Data Analysis Centres (IDACs) provide **independent** statistical analysis and reporting to support DMC decisions.

Why independence matters:

- Controlled access to unblinded data
- Objective interpretation of emerging data
- Clear separation from the Sponsor
- Reliable, decision-ready outputs



Skills developed in practice

Across 10+ DMC studies over nearly 5 years, my work in an IDAC has built the skills highlighted below.

Technical

- Statistical Reporting and data interpretation
- Handling unblinded data appropriately

Communication

- Clear interaction with Sponsor,DMC and internal teams
- Confidence in presenting and discussing outputs

Leadership

- Decision-making under uncertainty
- Delegation and team coordination

Operational

- Planning, prioritisation and timeline management
- Flexibility to respond to DMC requests
- Maintaining unblinded access

Moving into a lead role

Progressing from junior statistician to lead meant greater ownership, broader communication and more independent judgement. This included:

- Accountability for delivery of high-quality DMC outputs
- Clear, confident engagement with Sponsor, DMC and internal teams
- Greater confidence in independent decisions, while seeking input when needed
- Managing resourcing, prioritising effectively and trusting others to deliver

Overall, this transition strengthened my ability to deliver robust outputs, lead teams and support critical trial decisions.

Comparing Junior vs Lead Statistician

Junior Statistician	Lead Statistician
Programming DMC outputs	**Accountable** for DMC package
Works within the IDAC team	Engages with Sponsor,DMC and leads the IDAC team
Seeks input from leads	Makes independent decisions
Involved heavily in package creation for experience	**Delegates** and focuses on high-level tasks

For junior statisticians, DMC work builds good foundations such as communication, decision-making and flexibility that enables progression into lead roles.

Key Takeaways

1. DMC work accelerates leadership readiness by exposing statisticians early to decision-making under uncertainty
2. Technical skills are essential, but judgment, accountability and communication drive progression
3. Make sure to seek opportunities to understand the full DMC process, not just programming tasks

Contact
Rhys Jones | Statistician III
PHASTAR
Email: rhys.jones@phastar.com
Web: https://phastar.com

Additional Phastar Resources

Phastar Webinar
DMC Best Practices: What Biotechs and Sponsors Should Know

Phastar Blog Post
The need for a DMC and the Operational Role of an Independent Data Analysis Centre: Practical Insights for Sponsors

Phastar Blog Post
Navigating DMC Meeting Requirements and Conduct: Best Practices for Sponsors

#P040

Real-World Data to Influence Clinical Trial Drug Supply Forecasting

Sandra Joksaite (Chai-Rees)[1], Dominique Derreumaux[2], Linda Nichols[3], Jessica Perhanidis[4], Warsha Kumari Singh[5]

[1]GSK, London, United Kingdom; [2]GSK Wavre, Belgium; [3]GSK, Upper Providence, USA; [4]GSK, Waltham, MA, USA, [5]GSK, Bengaluru, India

Conclusions

- The approach was successfully applied to complex oncology trials, achieving an estimated 25% reduction in drug waste and significant cost avoidance
- Novel, partnership-driven methodological framework that strengthens the translation of real-world data into actionable forecasting strategies
- Potential for scalability of this framework to future studies and other therapeutic areas

Background

- Accurate forecasting of investigational medicinal product demand is essential to minimise costly drug wastage and avoid supply chain shortages in clinical trials
- Traditional forecasting approaches solely rely on clinical assumptions, published literature, and limited historical trial data, which may not reflect the latest standard of care or represent the drug usage patterns of the planned trial population.
- Incorporating insights from RWD about patient and treatments into forecasting models can enhance the accuracy

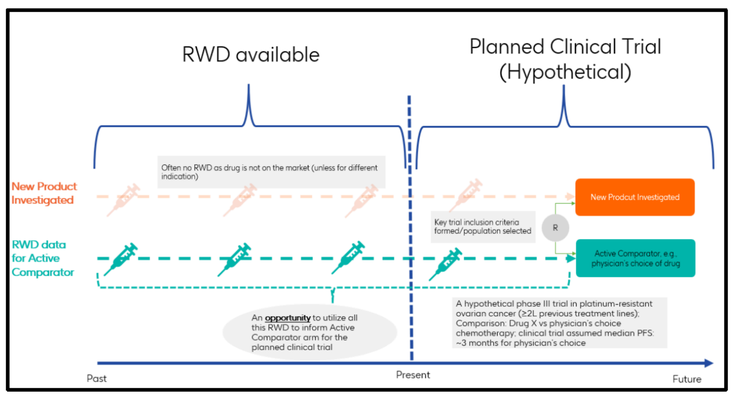
Aims

To develop an RWD-informed forecasting framework

Key RWD Inputs → IMP Estimation → Improved forecasting accuracy → Applicability across trials

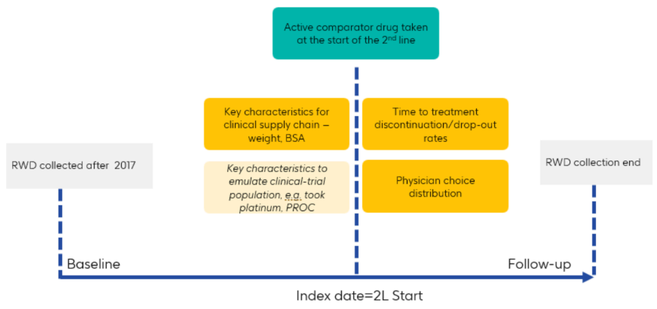
Study design

Figure 1: **RWD for comparator arms often exist for planned clinical trials**



RWD can include electronic health records, insurance claims, and registry data; however, these data sources are often underutilised when planning clinical trials and drug supply. This poster uses dummy data to illustrate the framework.

Figure 2: **Retrospective cohort analysis using longitudinal anonymised EHR-based RWD, aligned to a hypothetical clinical trial. Index date = 2L initiation. Baseline captures weight/BSA; follow-up captures rwTTD and physician's choice drug**



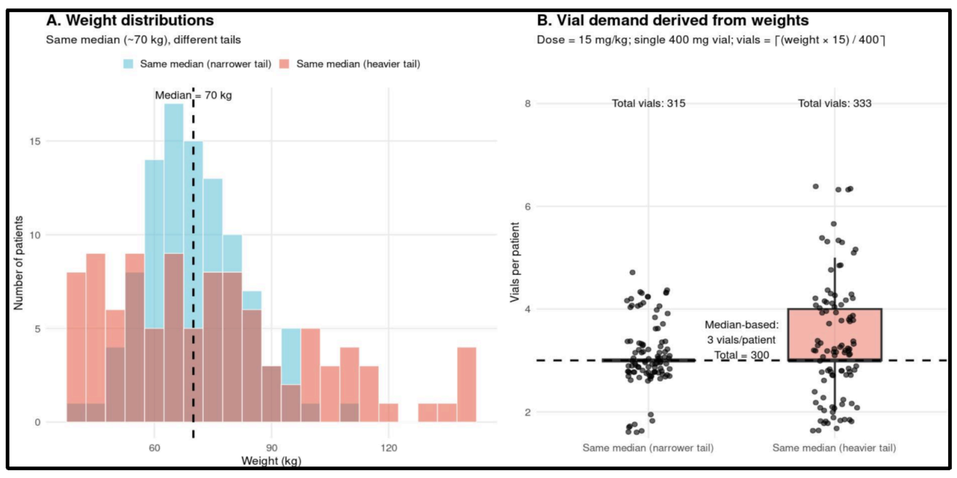
Longitudinal data from electronic health records were used to calculate key parameters for a selected hypothetical clinical trial-like population, including patient drop-out rates, treatment discontinuation, body weight, body surface area, and physician choice of the drug (for relevant trials). These parameters were integrated into a simulation-based forecasting model to generate more realistic demand scenarios and enable CT patients to receive the right dose at the right time.

Statistical Analysis

- Kaplan-Meier time-to-event analysis was applied to estimate real-world TTD distributions and expected discontinuation rates by cycle, e.g. 21 days cycles
- Probabilities of remaining on treatment were assumed to be proportion of patients remaining on treatment at the start of the cycle
- RWD proxies for cycles were created, i.e. 1 recorded administration = 1 treatment cycle
- Drop-out rates per visit were estimated using three approaches based on censoring assumptions: discontinuation, continuation, and Kaplan–Meier–based probabilities.
- Results were compared with traditional approaches used.

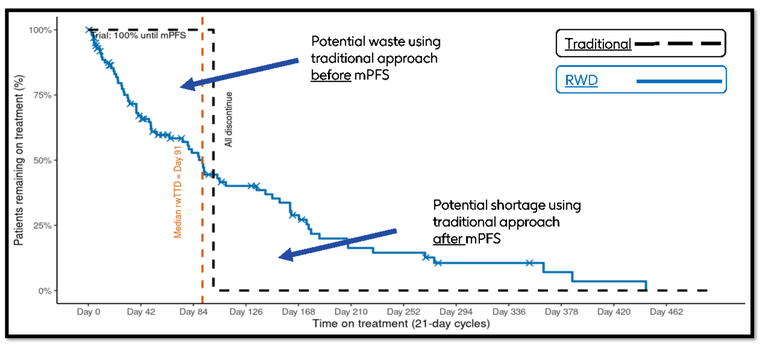
Results (Illustrative)

Figure 3: **Illustrative Example: Moving Beyond the Median Using Individual-Level RWD to Derive Weight Distributions for Vial Demand Estimation**



Panel A shows overlapping weight distributions with identical medians but different tail behaviour. Panel B shows the corresponding vial demand derived from these weights. Oncology therapies are supplied in fixed-size vials but dosed according to patient body weight (mg/kg) or body surface area (mg/m²) (e.g. bevacizumab, trastuzumab; paclitaxel, doxorubicin). Despite identical median weight, differences in the underlying weight distribution influence the number of vials required. Continuous dosing must be translated into discrete vial use, making weight-band categorisation essential for accurate estimation of supply needs. Real-world data enables this by providing access to individual-level patient information. This is particularly relevant in physician's choice settings, where treatment patterns and patient characteristics evolve rapidly; reliance on historical clinical trial data may not reflect current practice, highlighting the value of RWD-based estimation.

Figure 4: **Assumed patients remaining (RWD-KM vs. traditional CSC): median rwTTD aligns with clinical trial mPFS**



The CSC simulation system requires inputs for expected patient drop-out at the visit level..
In traditional supply modelling, it is assumed that all patients remain on treatment until a fixed time point (e.g. median time to progression for the active comparator), and that no patients remain on treatment beyond that point.
In contrast, EHR-based approaches leverage patient-level data to estimate drop-out at each time point, resulting in more realistic treatment duration patterns.
While this poster presents a hypothetical example, in an internal application, use of this framework led to a ~25% reduction in projected drug demand compared with traditional assumptions.

Table 2: **Drop-out rates are estimated using three complementary formulas to reflect real-world censoring, which occurs much more frequently than in clinical trials**

	Observed Number of Patients			*KM estimates of percentages of patients*		*Drop out percent rates per visit (different approaches)*		
Starting day	On treatment at starting day	Discontinued up to the end of the cycle *	Censored up to the end of the cycle *	Remaining at the start of the cycle (%)	Failed at the start of the cycle (%) *	[1] Discontinued at each step (%)	[2] Discontinued or censored at each step (%)	[3] KM estimate failed (%)
0	100	16	11	100	0	16	27	17
21	73	30	13	83	17	14	16	16
42	57	36	18	67	33	6	11	7
63	46	41	21	60	40	5	8	7
84	38	48	22	53	47	7	8	10
105	30	50	23	43	57	2	3	3
126	27	52	25	40	60	2	4	3

Depending on the business objective, different drop-out strategies can be applied. Drop-out rates were estimated based on the following assumptions.

[1] observed percentage of patients who discontinued during the cycle; we assume that all censored patients remain on treatment (minimizes shortage); [2] observed percentage of patients who discontinued and were censored during the cycle; we assume that all censored patients stopped treatment (minimizes waste); [3] Kaplan–Meier estimates, typically falling between the two and reflecting the most likely real-world scenario as censored patients are assumed to have the same probability of stopping or remaining on treatment

Abbreviations

BSA, body surface area; CSC, clinical supply chain; CT, clinical trial; EHR, electronic health record; IMP, investigational medicinal product; mPFS, median progression-free survival; RWD, real-world data; RWD-KM, real-world data Kaplan–Meier; rwTTD, real-world time to treatment discontinuation

References

1. European Medicines Agency. *Avastin (bevacizumab) Summary of Product Characteristics.* 2024
2. National Academies of Sciences, Engineering, and Medicine. *Medications in Single-Dose Vials.* 2021
3. Abdelmeseh V, et al. J *Oncol Pharm Pract.* 2022;29:1437–1442.
4. HOPA. *Dose Rounding of Anticancer Agents.* 2017..
5. Wilson BE, Booth CM. *eClinicalMedicine.* 2024
6. Saesen R, et al. *Eur J Cancer.* 2023;186:52–61

Acknowledgements

The authors thank Utpalraj Kemprai and Rushabh Shah for their valuable contributions and support.

Code and visualisation formatting were supported by generative AI tools.

Disclosures

All authors are GSK employees and hold GSK equities; Jessica Perhanidis also holds stock in Boston Scientific and UnitedHealth Group

Funding: GSK

PSI 2026 Programme

Presenting author: Sandra Joksaite (Chai-Rees), sandra.x.joksaite@gsk.com

P041

Beuermann J., Grafe V., Rmo M., Oldeland J. | Ecker + Ecker GmbH | Hamburg | Germany

Absolute standardized Differences – an objective tool for checking similarity in indirect treatment comparisons?

Background

Indirect treatment comparisons (ITCs) are an important tool in HTA decision-making. A key assumption is exchangeability, requiring similarity, homogeneity, and consistency between studies. Similarity is commonly assessed by comparing baseline distributions of effect modifiers and prognostic factors, although this is often done visually rather than quantitatively. We discuss the use of absolute standardized differences and the derivation of an overall similarity threshold.

Following approaches used to assess balance in propensity score methods, we propose the use of absolute standardized differences to evaluate similarity in baseline characteristics. While thresholds for balance assessment in propensity score analyses (e.g., 0.1–0.25) are suggested in the EU HTA guidance on evidence synthesis [1], no established thresholds currently exist for indirect comparisons based on randomized evidence in HTA.

Methods

To derive a possible threshold value, we evaluated all HTA procedures carried out by the German HTA body, the G-BA, from 2021 to 2025 [2]. For ITCs deemed methodologically sound and accepted by the G-BA, we extracted the reported baseline characteristics (continuous and categorical) from the studies employed for the indirect comparison and calculated the absolute standardized differences.

Results

We identified 11 approved ITCs comprising 97 evaluable baseline characteristics. All ITCs applied the Bucher method. Absolute standardized differences (ASDs) were aggregated, and four potential similarity thresholds with decreasing restrictiveness were derived from their distribution (10th percentile, median, mean, and 90th percentile; *Fig. 1*). A median ASD of 0.18 IQR [0.10, 0.43] was observed, ASDs for continuous baseline characteristics were slightly lower than those for categorical characteristics (*Fig. 2*). No substantial differences were observed between oncological and non-oncological indications (*Fig. 2*). Among baseline characteristics evaluable in at least five ITCs, region showed the highest average dissimilarity (*Fig. 3*).

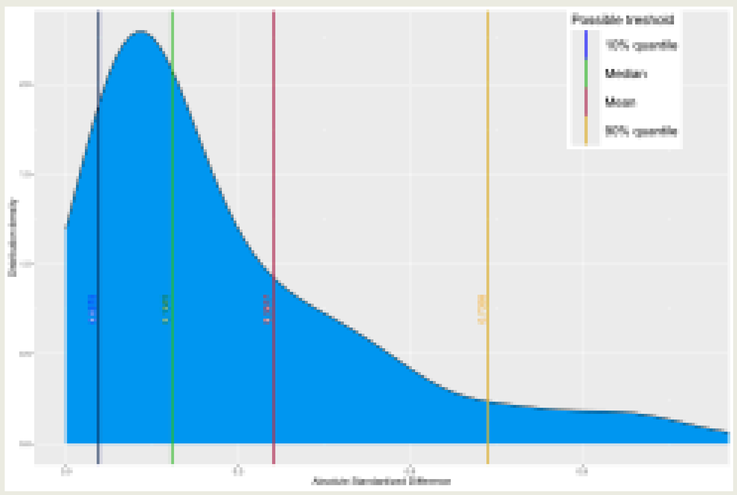


Fig. 1: Distribution of ASDs across accepted ITCs including possible tresholds

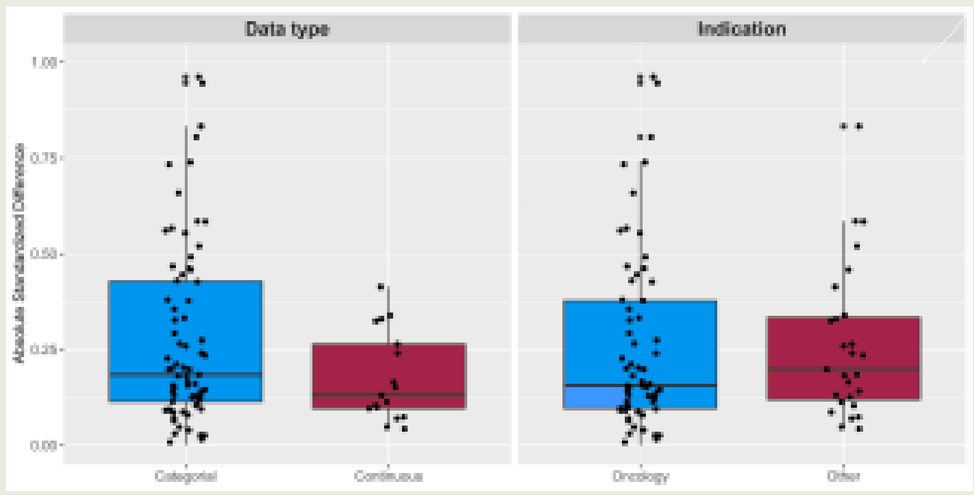


Fig. 2: Distribution of ASD by variable type and indication

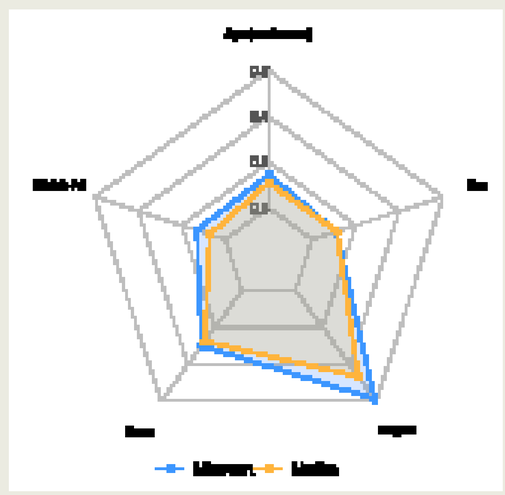


Fig. 3: Distribution of ASD by baseline characteristics

Limitations

- Limited ability to calculate comprehensive ASDs due to unreported baseline characteristics.
- Although all evaluable baseline characteristics were included, assessment of the similarity assumption primarily depends on the availability and comparability of effect modifiers and prognostic factors.

Conclusion

Assessing similarity remains context-dependent and complex; however, ASD calculations can provide an initial quantitative overview. Based on the observed ASD distribution, a threshold of 0.2 appears to be a pragmatic and moderately restrictive option. This corresponds to the rounded median ASD and lies within the commonly used propensity score balance range of 0.1–0.25. Given the right-skewed ASD distribution, the median was considered more robust than the mean for deriving a practical threshold.

References

[1] EU HTA Practical Guideline, www.health.ec.europa.eu
[2] Gemeinsamer Bundesausschuss (G-BA), www.g-ba.de

SCAN ME

What are multi-state models and how are they useful in our industry?

Victoria Lopez Dittmar, Statistician III
victoria.lopezdittmar@phastar.com

Introduction

Health technology assessments (**HTAs**) aim to identify the effectiveness or clinical benefit of a health technology and its cost[1]. To do this, **cost-effective analyses** over an appropriate time horizon are done, estimating the mean cost and benefit for a specific population. When a technology affects survival, or benefits and costs last until death (e.g. oncology therapies), then a lifetime horizon is needed to estimate the **restricted mean survival time** (**RMST**). The data for these analyses can be obtained from clinical trials.

Expecting trials to have follow-up times until all subjects experience the primary event can be infeasible, it could mean higher running costs of longer trials, increased patient drop-out and potential ethical issues. Therefore, extrapolating beyond the last observed time is needed to be able to estimate the RMST[2].

Two methods that allow for extrapolation techniques are[3]:

- **Partitioned survival analysis**
- **Multi-state survival analysis**

Partition Survival Analysis[4]

Models individuals as they move forwards and in sequence between exhaustive and mutually exclusive health states[5]. State membership (proportion of individuals at a health state at time t) is estimated through non-mutually exclusive parametric survival curves. For a model with P possible states, $P-1$ survival curves are needed to specify the model. E.g. PFS(t) (progression free survival) and OS(t) (overall survival) are estimated directly, and progressive disease can be calculated as PD(t) = OS(t) - PFS(t).

Parametric methods allow for the extrapolation of each survival. However, since the endpoints are modelled independently the extrapolation results are open to unrealistic results. For example, OS(t) must always lie above PFS(t), however this is not guaranteed under partitioned survival analysis.

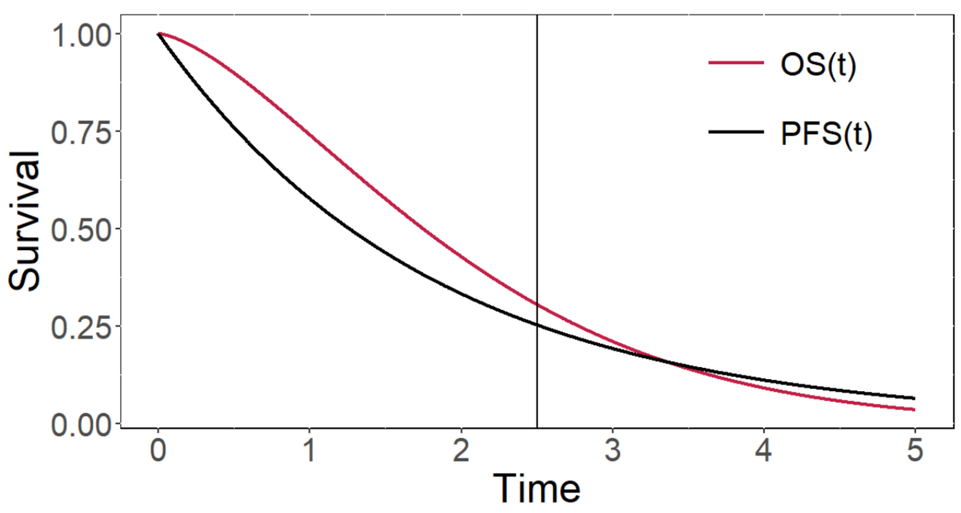


Figure: Example of a partitioned survival analysis model were the extrapolation period (after time=2.5) produce impossible results (lower overall survival than progression-free survival).

Multi-state Models[6]

Stochastic process $(\mathrm{X}(\mathrm{t}), \mathrm{t} > 0)$ of time-to-event data with finite number of states to transition between. Let $S = \{1, \dots, p\}$ be a finite state space. An **absorbing state** is one such that no further transitions can occur (e.g. Death), and a **transient state** is one where it is possible to move to a different state. Let χ_{t} be the observed history of the process up to time t. Assume right-continuous sample paths and independent right-censoring.



Transition probabilities
Probability of being in state j at time t given that at time s it was at state h and the history of the process up until time s.

$$\mathrm{P_{hj}(s,t)} = \mathbb{P}(\mathrm{X(t)} = \mathrm{j} | \mathrm{X(s)} = \mathrm{h}, \chi_{\mathrm{s}-})$$

where $\mathrm{h}, \mathrm{j} \in S$ and $\mathrm{s} \leq \mathrm{t}$.

Transition intensities
Instantaneous failure rates given the history of the process up until time t (hazard rates). In an absorbing state all transition intensities to other states are 0.

State occupational probabilities
Probability of being in state h at time t, based on the initial distribution of all states.

$$\pi_h(t) = \sum_{j \in S} \pi_j(0) P_{jh}(0,t)$$

Multi state models can be semi-parametric or parametric. When an extrapolation period is needed, since data is not fully observed, only parametric models can be used.

Parametric models[7]
Each transition is assumed to follow a known survival distribution with unknown parameters which can be estimated using standard maximum likelihood techniques. There are two types of parametric models that allow for extrapolation techniques:

- Proportional hazards models.
- Accelerated failure times models.

Extrapolation
Multi state models allow for dependence between states, the natural transformation of the disease is modelled, and it can be extrapolated to a lifetime horizon. The extrapolation is done on the transition intensities, unlike partitioned survival that extrapolates the survival curved directly. Therefore, no unrealistic results as the one presented on Figure 1 will occur.

Conclusion

Multi state models prevent illogical extrapolation scenarios by taking into consideration the nature of the disease, unlike partitioned survival. However, there are some drawbacks to its use. The standard process of performing extrapolation in cost-effective analysis requires choosing a single model via some selection technique (AIC or BIC for example), which are measures on relative fit of the observed data, and will not always extrapolate. Therefore, the choice of parametric models can have a big impact on the estimated RMST, but there is no guarantee the chosen model is the right one. A potentially more robust approach to offset this uncertainty would be to use **model averaging**[8], which is where the estimated RMST consists of a weighted sum of the value of interest from a selection of possible models.

References

1. NICE (2013). Guide to the Methods of Technology Appraisal. National Institute for Health and Care Excellence, London.
2. Latimer, N. R. and Adler, A. I. (2022). Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. BMJ Medicine.
3. Cranmer, H., Shields, G. E., and Bullement, A. (2020). A comparison of partitioned survival analysis and state transition multi-state modelling approaches using a case study in oncology. Journal of Medical Economics.
4. Williams, C., Lewsey, J. D., Mackay, D. F., and Briggs, A. H. (2017). Estimation of survival probabilities for use in cost-effectiveness analyses: A comparison of a multi-state modeling survival analysis approach with partitioned survival and Markov decision-analytic modeling. Medical Decision Making.
5. Woods, B., Sideris, E., Palmer, S., Latimer, N., and Soares, M. (2017). NICE DSU technical support document 19: Partitioned survival analysis for decision modelling in health care: A critical review.
6. Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. Statistical Methods in Medical Research.
7. Zare, A., Hosseini, M., Mahmoodi, M., Mohammad, K., Zeraati, H., and Holakouie Naieni, K. (2015). A comparison between accelerated failure-time and Cox proportional hazard models in analyzing the survival of gastric cancer patients. Iran Journal Public Health.
8. Latimer, N. (2011). NICE DSU technical support document 14: survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data. Report by the Decision Support Unit.

Characterizing FEV_1 Variability from Historical Early-Phase Asthma Trials to Shape Future Designs

Alese Halvorson[1], Alexandra Jauhiainen[1]

[1]R&I Biometrics and Statistical Innovation, Late Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

P044

Abstract

Asthma trials depend on realistic assumptions about the variability of lung function endpoints, such as pre-bronchodilator Forced Expiratory Volume in the first second (pre-BD, or trough FEV_1), across different conditions and timepoints. When those assumptions are not met, studies risk being mis-powered or having a primary endpoint that lacks sufficient precision.

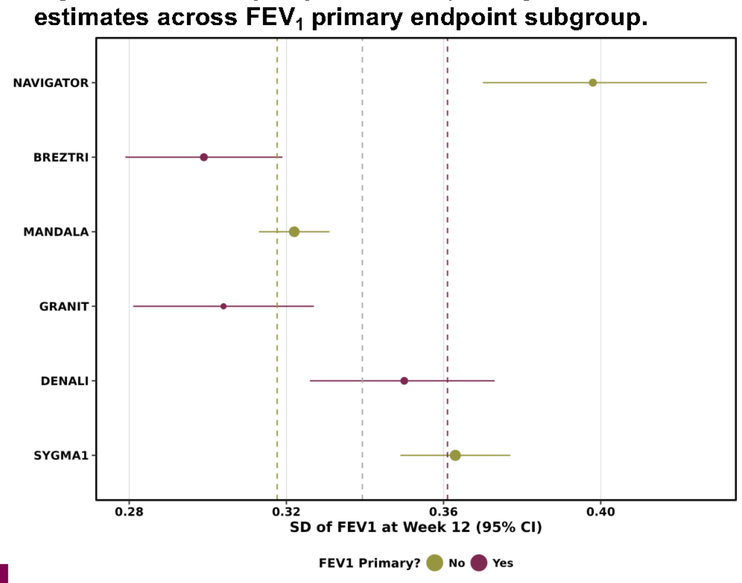
We set out to systematically characterise how change from baseline in pre-BD FEV_1 variability behaves across six recent Asthma trials and across various patient subgroups, so teams can effectively plan future early-phase studies with evidence-based inputs.

Introduction

- Evidence from historical trials: we quantified visit-specific variability in pre-BD FEV_1 change from baseline across six AZ trials, providing SDs with uncertainty for informing powering and feasibility.
- Design choices shape variability: estimated variability was lower when FEV_1 was primary; higher in adolescents, and lower in older adults—informing inclusion and endpoint choices.
- Variability evolves over time: across multiple trials, there was an increase in variability from early to later visits, guiding selection of primary assessment timing and sensitivity ranges.

Results

Figure 1: Variability in pre-BD FEV_1 change from baseline estimates across FEV_1 primary endpoint subgroup.



Patterns in variability emerge

The overall Week 12 mean SD across studies was ≈0.34. At the early Week 12 assessment, trials with FEV_1 as the primary endpoint showed lower dispersion (SD ≈ 0.30–0.35) than those where it was not primary (SD ≈ 0.32–0.40) (Fig 1).

Within trials, SDs were generally stable to slightly higher at later visits (Fig 2), with study-specific variability seen.

At Week 12, standard deviations within adolescents tended to be highest (e.g., NAVIGATOR ≈ 0.57), adults were intermediate (≈ 0.31–0.40 across studies), and older adults lowest (often ≈ 0.23–0.27). Excluding adolescents tended to shift overall SDs downward (Fig 3).

Figure 2: Variability in pre-BD FEV_1 in change from baseline estimates across timepoint subgroup.

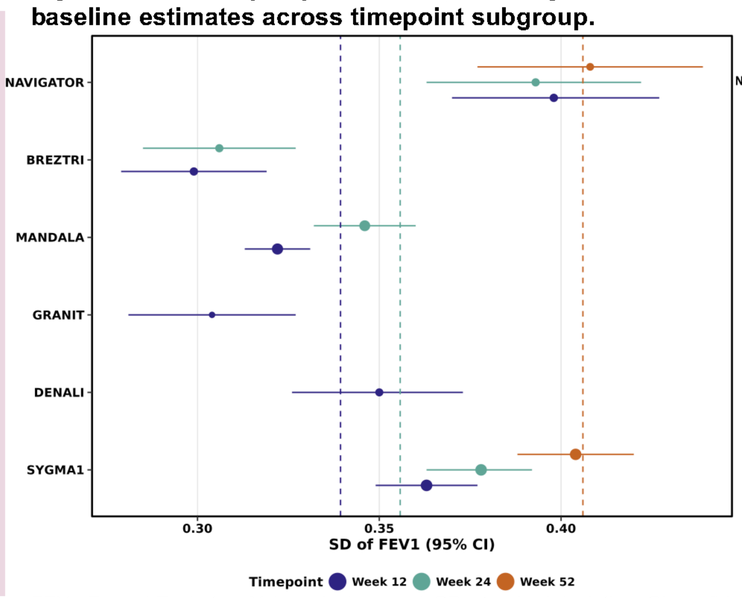
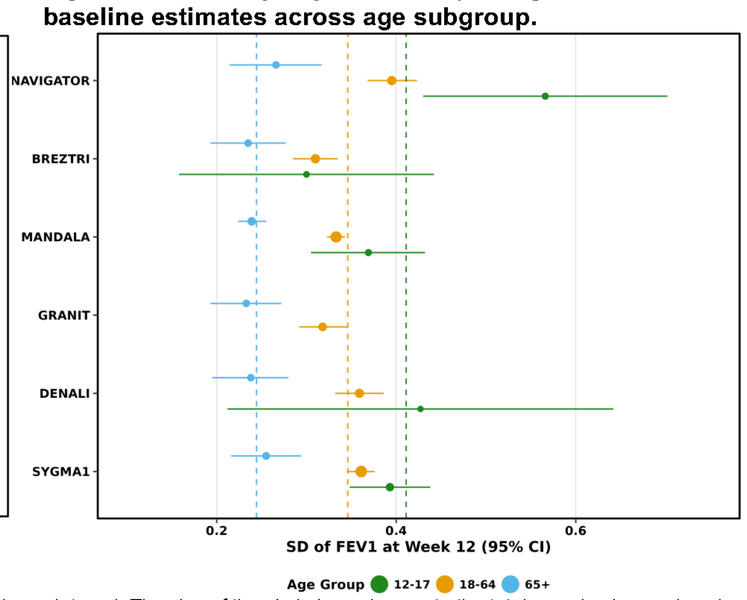


Figure 3: Variability in pre-BD FEV_1 change from baseline estimates across age subgroup.



FEV_1, Forced Expiratory Volume in 1 second; SD, standard deviation; CI, confidence interval. The size of the circle is analogous to the total sample size analysed.

Methods

Model Specification

For each study and subgroup, we fit a Generalised Least Squares (GLS) model using the `nlme` package in R with unstructured within-subject correlation and visit-specific residual variances

```
CHG ~ BASE + TRTA + WK + TRTA:WK
```

Yielding a marginal variance-covariance matrix with variance estimates at each timepoint, from which we use a direct normal approximation to quantify confidence intervals around the SDs.

Variability Measures

The SD confidence intervals use a direct normal approximation based on the intervals for the residual standard error and variance weights, derived as follows: for two stochastic variables X and Y that are independent, we deduce the following formula for variance of the product $X \cdot Y$

$$V(X \cdot Y) = V(X) \cdot V(Y) + V(X) \cdot E(Y)^2 + V(Y) \cdot E(X)^2$$

Since the diagonal of the variance-covariance matrix in the `gls` model fit by `nlme` is defined as $(\sigma \cdot v\omega)^2$ where σ is the residual variance and $v\omega$ are the variance weights, we can extract these measures from the model and approximate SDs directly. We then use a normal approximation for 95% confidence intervals around the variance components.

This approach was shown via simulation to be similar in performance to parametric bootstrap and resampling bootstrap methods.

Table 1: Asthma studies analysed.

Access granted for internal reuse

Study	N	Asthma Severity ↓	Duration	FEV_1 Endpoint Priority
NAVIGATOR[3]	1,041	Severe	52wks	Secondary
BREZTRI[2]	1,043	Moderate/Sev	24wks	Primary
MANDALA[5]	2,902	Moderate/Sev	24wks	Secondary
GRANIT[6]	771	Moderate	12wks	Primary
DENALI[1]	989	Mild/Moderate	12wks	Primary
SYGMA-1[4]	3,252	Mild	52wks	Secondary

Study Selection & Subgroups

We hypothesised that there may be quantifiable differences in pre-FEV_1 variability measures depending on various study characteristics, so separate models were fit by:

- ❑ FEV_1 endpoint prioritisation [Primary; Secondary]
- ❑ Age group [12-17; 18-64; 65+]
  (Note: excluding adolescents 12-17 is most relevant for early-phase planning)
- ❑ Timepoint [Week 12; 24; 52]
  (Note: Week 12 is most relevant for early-phase planning)

All studies were filtered to include only subjects above age 12 and nearest timepoint was imputed for studies with non-matching weeks.

Conclusions

If you are sizing an upcoming 12-week Phase 2 asthma trial based on change from baseline in pre-BD FEV_1, we recommend…

Starting with an SD estimate of ~0.30–0.35 and making the following adjustments:

- + 6% for a 24-week study
- +10% for a 52-week study
- +10% if FEV_1 is the secondary endpoint
- +15% if the study includes adolescents

We plan to expand the current work by:

- ❑ Evaluating disease severity as a subgroup
- ❑ Adding data from the latest asthma trials
- ❑ Applying the current framework to additional respiratory endpoints

References

1. Chipps BE, et al. Albuterol–budesonide fixed-dose combination in patients with moderate-to-severe asthma (DENALI): a randomized, double-blind trial. Am J Respir Crit Care Med. 2023;207(11):1374-84.
2. ClinicalTrials.gov. Efficacy and Safety of PT001 to Placebo and Open-label Spiriva® Respimat® in Subjects With Persistant Asthma. Identifier NCT03358147.
3. Menzies-Gow A, et al. Tezepelumab in adults and adolescents with severe, uncontrolled asthma (NAVIGATOR): a randomized, double-blind, placebo-controlled phase 3 trial. N Engl J Med. 2021;384(19):1800-09.
4. O'Byrne PM, et al. Inhaled budesonide-formoterol used as needed in mild asthma (SYGMA 1). N Engl J Med. 2018;378(20):1865-76.
5. Papi A, et al. Albuterol–budesonide fixed-dose combination as needed for prevention of severe asthma exacerbations (MANDALA): a randomized, double-blind, phase 3 trial. N Engl J Med. 2022;386(18):1628-38.
6. Pavord ID, et al. Fevipiprant in eosinophilic asthma (GRANIT): phase 3 randomized, double-blind, placebo-controlled study. Lancet Respir Med. 2020;8(7):711-24.

Acknowledgements

AH and AJ are employees of AstraZeneca and may hold stock or stock options in the company.

Supported by AstraZeneca

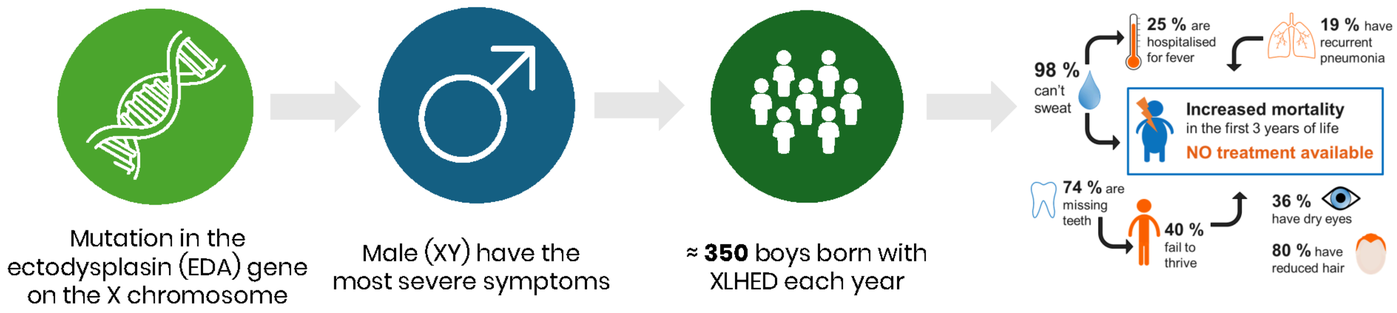
Presented at the PSI Conference, Belfast, Northern Ireland, 15 June 2026

P045

From Conventional Trials to Ultra-Rare Diseases: Adapting Classical Methods to Extremely Small Populations

Authors: Benoît Sansas, Sandrine Roye, Pierre Bunouf, Guillaume Desachy
Pierre Fabre Pharma, Data Science & Biometry, France

Pathology Overview and Edelife Clinical Trial



- Phase II, open-label, single-arm
- Primary Endpoint: Sweating volume in µL
- Genotype-matched control: relatives / external database
- Planned 15 evaluable participants

Determine a robust Statistical approach

🧪 Shifted null hypothesis

H0: Median paired difference ≤ 2 µL
→ Upper Bound of measurement variability
→ Conservative statistical no-effect region

H1: Median paired difference > 2 µL
Power = 97.8%

📊 Exact inference for quantitative endpoint

No asymptotic assumed, Exact paired Signed-rank Wilcoxon test.

Robust to: Skewed data distribution, zero-inflation, extreme responders.

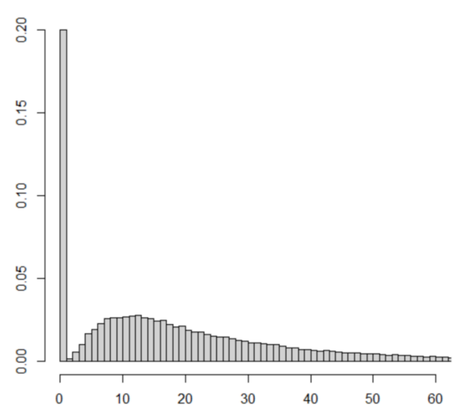
📐 Effect size estimation

Comparing treated subjects with genotype-matched historical controls.
Treatment effect magnitude summarized using the Hodges–Lehmann estimator of the median paired difference.

This estimator is: robust to outliers, consistent with the rank-based Wilcoxon framework, well suited to skewed and zero-inflated distributions.

H1 Distribution

- Treatment effect modeled using a zero-inflated log-normal distribution
- Among responders, effects follow a log-normal distribution with log-mean = 3 and log-SD = 0.8.
- 20% of subjects assumed to have no treatment effect.
- Expected median: 15 µL



Hodges–Lehmann estimates the median paired difference by minimizing the sum of absolute deviations (L1 norm)

How to maximize evidence from a finite and non-expandable population?

✅ Statistical responder criterion

By construction of the exact one group Wilcoxon test, decision rules can be simplified using thresholds.
Test Statistic is ranks of the absolute values that correspond to positive paired differences.
→ Positive paired differences ≥ 4 µL correspond to a statistical "responder" criterion.

This threshold is purely statistical, enabling deterministic p-value calculation, and is not intended as a clinical definition of response.

$$Test\ Statistic = \sum_{i=1}^{N} Rank(|Ti - Ci - 2|) \times 1_{Ti-Ci-2>0}$$

Ti: Sweat Volume of Treated participant
Ci: Sweat Volume of Matching Control

⏱ Adding Interim Analysis

Introduced to allow early efficacy assessment under severe recruitment constraints.
Pocock boundary accepted by regulators in context of rare disease to conclude early for efficacy.
Using Pocock, the overall Type I Error is controlled to a level of 2.08%.
N = 8 participants → significance with at least 7 paired difference > 4 µL.

Interim Analysis Characteristics

- Wang & Tsiatis (Δ = 0.5)
- Pockock Boundaries at interim analysis and Final analysis = 0.0149
- Power at IA = 45%
- Power at Final Analysis = 99.7%

Key Takeaways

Ultra-rare trials do not require abandoning classical statistics
They require re-engineering them for decision-making

- ✅ Discussed, challenged but accepted by Regulators (EMA, FDA)
- ✅ Exact testing with transparent decision rules
- ✅ Statistically rigorous interim analysis with ethical needs

Wang, S. K., and A. A. Tsiatis. 1987. '*Approximately optimal one-parameter boundaries for group sequential trials*', *Biometrics*, 43: 193-9

EspeRare

LABORATOIRES Pierre Fabre Pharma Research & Development

GAM plots, AI, SAS and R: a tutorial exploring baseline associations and negative binomial responses

Kacper Kowalczyk[1], Kaitie Fernandez[2]

Biometrics, Late-stage Development, Respiratory and Immunology (RI), BioPharmaceuticals R&D, AstraZeneca, Warsaw, Poland[1], Cytel, Inc. Cambridge, MA, USA [2]

P046

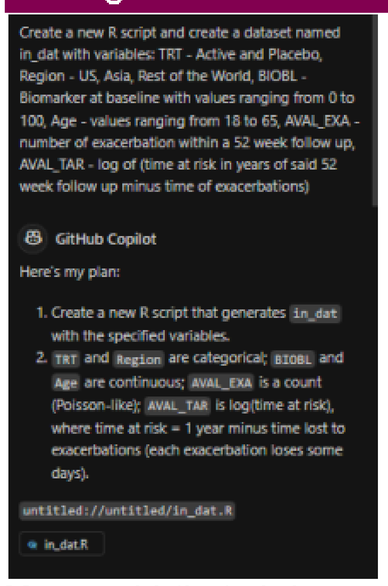
Abstract

This tutorial demonstrates how to create generalized additive model (GAM) plots in both SAS and R using an analysis dataset with a nonlinear continuous predictor and a negative binomial response variable generated using artificial intelligence (AI). GAM plots are useful for visualizing complex, nonlinear relationships between predictors and outcomes and offer flexibility beyond traditional linear models. While examples commonly focus on continuous or binomial responses, this work highlights approaches for negative binomial outcomes. The poster outlines when GAMs are appropriate, how they differ from linear models, and provides step-by-step guidance for implementation in SAS and R, including key similarities and differences between the two languages. An AI-generated dataset is used to demonstrate these methods, avoiding reliance on ongoing clinical trial data.

Introduction

- GitHub Copilot was leveraged to generate a synthetic analysis dataset containing a nonlinear predictor and a negative binomial response variable
- GAMs extend traditional regression models by allowing flexible nonlinear relationships between predictors and outcomes while GAM plots provide an intuitive way to visualize complex associations inadequately captured by linear models
- Key similarities and differences between SAS and R coding approaches are highlighted

Data generation



GitHub Copilot was utilized for data generation. Basis of the code and data structure was achieved using the call on the left-hand side. Adjustments like changing the distribution to Negative Binomial and adding withdrawals were also made with Copilot's help.

More sophisticated changes e.g. to see clear separation of lines between 60 to 80 were made by hand which was implemented easily using existing code.

Why GAMs?

Linear Model vs GAM with Negative Binomial Outcome

A linear model and a GAM are both regression approaches used to model relationships between predictors and an outcome (in our case, a negative binomial outcome). They differ in how they handle those relationships.

A standard linear model assumes the relationship between predictors and the response is linear (on a log scale):

$$\log(E[Y]) = \beta_0 + \beta_1 X$$

Each predictor has a straight-line effect on the outcome. A one-unit increase changes the expected count by a constant amount, and the effect is always proportional.

GAM with Negative Binomial

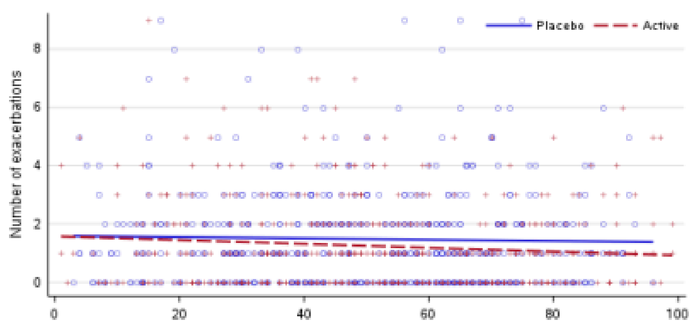
A GAM keeps the same negative binomial distribution, but replaces the straight-line predictor effect with a smooth nonlinear function:

$$\log(E[Y]) = \beta_0 + f(X)$$

Instead of estimating one slope coefficient, the GAM model estimates a smooth curve that can bend and change slopes over a continuum without manually having to specify polynomial terms. In this model there is a need to specify a smoothing parameter. It needs to find a balance between over and under fitting the data.

Data exploration using Linear Model

Figure 1 Exacerbations vs Baseline Biomarker



Using GitHub Co-pilot, we were able to generate a dataset containing 1000 participants with the following outcomes:

- Number of exacerbations (negative binomial outcome)
- Baseline continuous biomarker (non-linear covariate)
- Time at risk (model offset)
- Treatment group assignment

From Figure 1, we can see there is a relationship between the baseline biomarker and the number of exacerbations by treatment group with separation between the treatment groups being seen as the biomarker increases.

GAM Plot: SAS Code

In order to fit a model that includes a treatment by biomarker effect, in SAS we need to create an interaction effect using some dummy coding:

```
x_trt1 = biobl*(trt = 'Placebo');
x_trt2 = biobl*(trt = 'Active');
```

Then, we fit our GAM model using PROC GAMPL:

```
proc gampl data = interact seed=123456;

model aval = param(trtn)
spline(x_trt1/smooth=0.01)
spline(x_trt2/smooth=0.01) / offset=aval_tar
dist=NegativeBinomial link=log;
id trt aval biobl triskyrs x_trt1 x_trt2;
output out=OutExp pred=p lower upper;

run;
```

After some slight data manipulation, we can create a plot to display the relationship between our negative binomial outcome and a continuous biomarker showing the desired treatment effect

Figure 2 GAM Plot using SAS

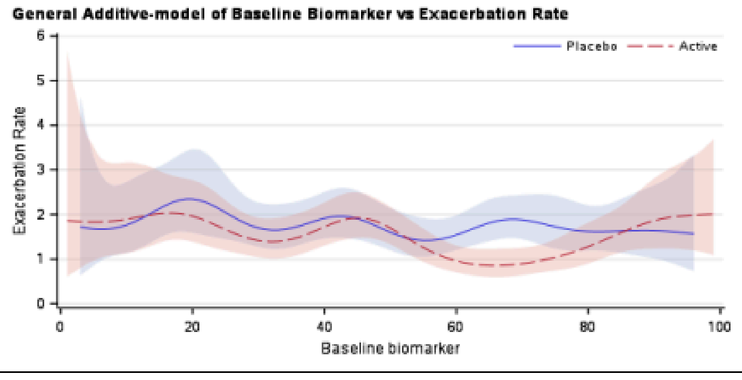


Figure 2 shows that there is little treatment effect below 60 with a true separation between 60 and 80 and a bit of a plateau in the treatment effect above 80.

GAM Plot: R Code

In R you can specify smoothing parameter for all splines to be one number easily and on top of that the `mgcv` package provides easy way of doing interactions by utilizing "`by=`" parameter:

```
Smth <- 0.01
Gammod <- mgcv::gam(AVAL_EXA ~ TRT + s(BIOBL,
by=TRT, sp=smth) + offset(AVAL_TAR),
family=nb(link=log), data=in_dat)
```

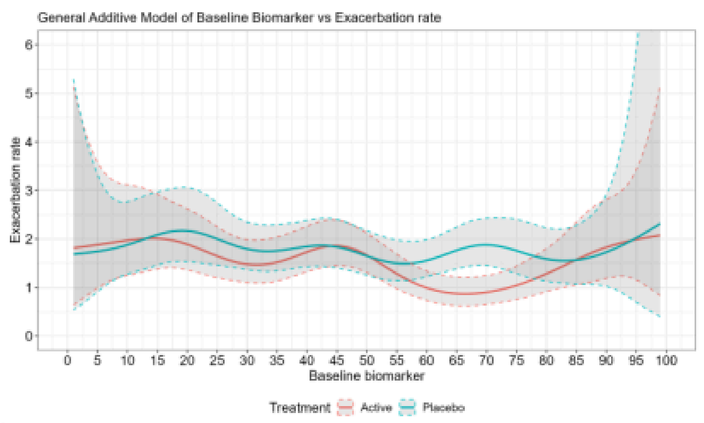
To create the plot Figure 3, we need to predict (or calculate LSMeans) – the Y-axis values, at points of X-axis parameter – Baseline biomarker. It can be done with `emmeans` package by creating a list with "at" values and using "`at=`" parameter:

```
at.list <- with(in_dat, list(
BIOBL = seq(min(BIOBL), max(BIOBL),length=400))

Lsmean <- as.data.frame(emmeans(gammod, ~BIOBL,
at=at.list, weights="proportional",
type="response", by=c("TRT"), data=in_dat))
```

`Lsmean` dataset contains lsmeans calculated for each value of `BIOBL` specified in `at.list` by treatment arm. `emmeans` function can handle more covariates (categorical e.g. Region or continuous e.g. Age) so one can explore more complex models, but needs to update the `at.list` by adding continuous variables and setting them to a single value e.g. mean.

Figure 3 GAM Plot using R



Conclusions

- While linear models for negative binomial outcomes fit the best exponential-shaped proportional trend, GAM models fit the smoothest nonlinear trend supported by the data.
- Both SAS and R can produce similar, interpretable GAM plots. R has a more flexible way of incorporating interaction effects and additional covariates compared to SAS
- The use of AI-generated dataset provided an efficient and practical approach for this project avoiding the need to use sensitive ongoing research data
- Increased familiarity with GAM methodology and visualization techniques can help researchers better explore complex relationships between outcomes and predictors across a non-linear continuum.

References

1. Wood, S.N. (2017). Generalized Additive Models: An Introduction with R, Second Edition (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279.
2. https://documentation.sas.com/doc/en/pgmsascdc/v_074/statug/statug_hpgam_toc.htm, SAS documentation

Supported by AstraZeneca *Presented at the PSI 2026 Conference, Belfast, 15 June 2026*

Disclosures

KK is an employees of AstraZeneca and may own stock or stock options or interests in the company. KF is an employee of Cytel and may own stock or stock options or interests in the company.

Funding

This poster was funded by AstraZeneca.

P047 - Why Radioconjugates Demand New Statistical Thinking *Statistical considerations for radioconjugates in clinical trials*

Benjamin Webb[1], Rosalind Hobson[1], Danny Lu[2], Daniel Slade[1]

AstraZeneca, Oncology Biometrics, UK[1]
AstraZeneca, Oncology Biometrics, Canada[2]

Introduction

- Radioconjugates (RCs) are an exciting area for pharmaceutical companies, providing precision **nuclear medicine** to advanced cancers. There is an explosion of clinical trial activity in this field, with huge competition across pipelines.
- These drugs provide **novel challenges to clinical trialists**. Through Early Phase trials, we must learn how to, in advance of Phase 3: adjust drug supply and dosing given fast degradation of product; to utilise dosimetry data to inform dose escalation; be smarter with long-term safety follow-up information; incorporate requirements set within new draft guidelines for dose finding in this rapidly evolving area.
- As statistical team members, we have applied our knowledge and adapted our ways of working to support Phase 1 and 2 trials of RCs in prostate cancer.
- We share our rationale for choosing TiTE methods for dose escalation and some lessons learned about collecting and using PK data.
- With a focus on Phase 1 and 2, we discuss some of the challenges around drug supply and randomisation timing and how this may impact assessments.

Radioconjugates – What Are They?

- RCs (also 'Radioligand therapies', RLTs) are **targeted antibody therapies** that combine a binding molecule with a radioisotope via a linker to deliver radiation directly to cancer cells.
- Similar to Antibody-Drug Conjugates (ADCs), they aim to improve therapy by specifically targeting cancer cells, with the ability to target higher disease burden and reduce off target damage, compared to external beam therapies.
- RCs deliver a **radioactive payload**, versus the cytotoxic payload seen in ADCs.
- The payload contains a radioisotope with a short half-life (hours/days). Once administered, the fast-decaying isotope paired with a targeting domain aims to limit off-target radiation exposure but **strict timing** from manufacture to dosing is needed to ensure the required dose is delivered to the patient.
- Payloads of RCs in development are commonly alpha-emitters or beta-emitters.
- Beta-emitters were earlier to the market, and competition is fierce for the alphas.
- The effects of alpha-emitters are **highly localised** and focus on areas which need to be "hit" and cause double stranded DNA breaks more often than beta-emitters, which travel further, but with less intensity.
- Less travel time indicates alpha-emitters may have less off-target toxicity than beta-emitters despite being more powerful.
- By pairing the RC with an imaging isotope, Theranostics combine therapy and diagnostics, allowing doctors to "see what they treat and treat what they see".

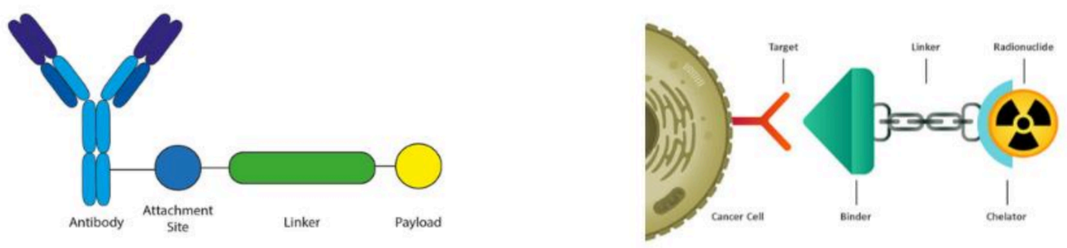


Figure: Anatomy of an ADC vs a RC

Figure: Targeted therapies with their similarities and differences

The Drug Supply Challenge

Manufacture at risk, (ie, drug substance is ready at point of randomisation) is expensive, due to high potential for waste if a patient cannot be randomised or is randomised to different dose/control arm.

vs

Manufacture on demand may lead to long gaps between randomisation and dosing increasing opportunities for attrition and pre-dose progression.

Statisticians should **understand** the drug supply process for the trial and present teams with different cost/benefits to randomisation; **explore** different options around Day 1 definitions eg, using first dose date for time to event analyses; **ensure** any decay correction factor is incorporated into the datasets so that actual dose received can be adjusted accordingly.

The Dose Finding Challenge (draft FDA guidelines August '25)[1]

The general principles of Project Optimus apply (eg, dosing up to maximum tolerated dose, MTD is not necessarily optimal) PLUS:
Cumulative radiation (lifetime) received is highlighted as an important eligibility consideration. Therefore, doses of prior RCs must be collected in as much detail as possible and may need converting to a standard unit, eg, Grays. Safety summaries may need to be presented by calculated cumulative radiation.

Due to the potential for late onset AEs, **multi-cycle DLT evaluation periods** may be preferable for dose escalation trials, in order to de-risk a future randomised dose optimisation stage.
AstraZeneca Early Oncology Statisticians have simulated **Time to Event (TiTE)** approaches to allow evaluation before the full DLT evaluation period has been completed for all patients. This approach has been shown to have good operating characteristics for the selection of MTD and a reduced overall duration of Phase 1.

The Novel Data Type Challenge

Dosimetry data encompasses the measurement and calculation of radiation dose absorbed by tissue. A new cross-functional collaboration is emerging with radiation scientists (physicists) and specialist vendors.
Statisticians must understand the relevance of these endpoints in order to establish how best to incorporate them into analyses and **build new standards**.
Statisticians can also help by providing education to these partners on best practices for working with clinical data in industry led trials.

The Long-Term Safety Challenge

Off-target DNA damage from radioisotopes can lead to both **acute and late-onset** toxicity, which may be irreversible.
Treatment emergent adverse event definition could be prolonged to consider the entire safety follow up period (eg, 12 months or after starting subsequent anti-cancer therapy), rather than a standard wash-out period.
Statisticians should work with safety teams to **define post-emergent AEs** or 'rAESI' (late radiation adverse events of special interest).
Prioritise the **exploration of safety biomarkers**, eg, eGFR for kidney function, beyond standard safety summaries.
Use TiTE approaches for escalation (see The Dose Finding Challenge).

The PK Data Challenge

PK concentrations may be derived using 'hot' (**radioactive**) and 'cold' (**non-radioactive**, ie, sufficiently decayed) assays. 'Cold' samples can also be obtained from dosing with the molecule minus the radioisotope, which may enhance biodistribution. Collection of **Specific activity** which is the amount of radioactivity per dose is encouraged using IRT. Dealing with radioactive samples can add operational complexities and the decay rate of alpha emitters to time of dosing and time of sample analysis must be accounted for.
A specific PK data delivery and analysis plan is advisable.

Summary

This is an ever changing and growing landscape in which we are continuing to learn, both of the science behind these exciting new drugs, and of the importance of our role in cross-functional teams.

References
[1] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/oncology-therapeutic-radiopharmaceuticals-dosage-optimization-during-clinical-development

Acknowledgements
Thank you to Azura Evans, Charlotte Hawkins, Chintu Desai and Teresa Collins for their review of the poster.

P048: Automated generation of SAP multiplicity appendices via the {appendMCP} R package

Michael Grayling, Yevgen Tymofyeyev

Statistical Modeling, Methodology and Consulting, Johnson & Johnson

Introduction

- Many phase 3 oncology trials now leveraging graphical testing procedures within a group-sequential design
- Available methodology is very powerful, supporting testing or multiple endpoints and (sub)populations while providing strict error-rate control
- However, documenting these complex testing strategies in Statistical Analysis Plans remains challenging and error prone
- This poster therefore introduces {appendMCP}, an R package designed to streamline the specification, validation, and documentation of graphical testing procedures in group-sequential designed trials

Input

The user only has to **specify a single .R file**, containing a **list consisting of 5 components**. Example input is shown for the **KEYNOTE-598** trial

KEYNOTE-598: Compared a monotherapy against a combination therapy, with three endpoints formally tested across three analyses

1. **Analyses:** What triggers each analysis? Is it based on sample size or events? In which sub-populations and treatment arms?

E.g., analyse after 255, 307 and 361 OS events

```
analyses = tibble::tribble(
  ~endpoint, ~strata, ~treatments,          ~sample_size, ~events,
  "OS",      "All",   c("Mono", "Combo"), NA,           255,
  "OS",      "All",   c("Mono", "Combo"), NA,           307,
  "OS",      "All",   c("Mono", "Combo"), NA,           361
)
```

2. **Hypotheses:** Define each hypothesis in the graphical test. Which treatments are compared? In which sub-populations? For which endpoint? At which analyses? Which spending function is used?

E.g., OS is analysed three times, PFS is analysed twice, and ORR is analysed once. O'Brien-Fleming spending is used for PFS and OS

```
hypotheses = tibble::tribble(
  ~type,      ~endpoint, ~strata, ~control, ~test,   ~analyses_analysed, ~sf,     ~sfpar, ~nominal, ~test_method,
  "Primary",  "OS",      "All",   "Mono",   "Combo", 1:3,               "asOF",  NULL,   NULL,     "logrank",
  "Primary",  "PFS",     "All",   "Mono",   "Combo", 1:2,               "asOF",  NULL,   NULL,     "logrank",
  "Secondary","ORR",     "All",   "Mono",   "Combo", 1,                 "none",  NULL,   NULL,     "pooled_proportions"
)
```

3. **Enrollment:** Specify speed and duration of enrollment over time to each of the treatment arms, by sub-population if needed

E.g., 1:1 randomization of 568 participants over 20 months

```
enroll_rate = tibble::tribble(
  ~stratum, ~treatments,          ~rate,  ~duration, ~ratio,
  "All",    c("Mono", "Combo"), 568/20, 20,        c(1, 1)
)
```

4. **Distributional information:** For all sub-populations, treatment arms, and endpoints, what is the assumed distribution?

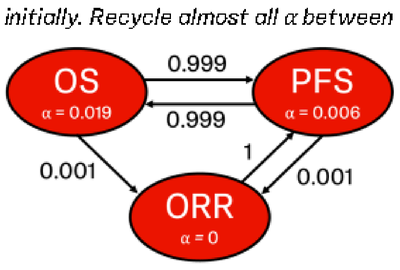
E.g., OS is exponentially distributed, assume HR = 0.7. PFS is piece-wise exponentially distributed, assume HR = 0.69. ORR is evaluable after 6 months, assume 59% vs 39% rates

```
distribution_tte = tibble::tribble(
  ~endpoint, ~stratum, ~treatment, ~duration, ~fail_rate,        ~dropout_rate,
  "OS",      "All",    "Mono",     Inf,       log(2)/20,        -log(1 - 0.01)/12,
  "OS",      "All",    "Combo",    Inf,       0.7*log(2)/20,    -log(1 - 0.01)/12,
  "PFS",     "All",    "Mono",     6.5,       log(2)/6.5,       -log(1 - 0.13)/12,
  "PFS",     "All",    "Mono",     Inf,       log(2)/14.5,      -log(1 - 0.13)/12,
  "PFS",     "All",    "Combo",    6.5,       0.69*log(2)/6.5,  -log(1 - 0.13)/12,
  "PFS",     "All",    "Combo",    Inf,       0.69*log(2)/14.5, -log(1 - 0.13)/12
),
distribution_bin = tibble::tribble(
  ~endpoint, ~stratum, ~treatment, ~rate, ~maturity_time,
  "ORR",     "All",    "Mono",     0.39,  6,
  "ORR",     "All",    "Combo",    0.59,  6
)
```

5. **Graphical test:** The initial graphical test

E.g., 1.9% to OS and 0.6% to PFS initially. Recycle almost all α between OS and PFS

```
graph = list(
  g = rbind(c(    0, 0.999, 0.001),
            c(0.999,     0, 0.001),
            c(    0,     1,     0)),
  w = c(0.019, 0.006, 0)/0.025
)
```



Output

- The package contains a standardised .Rmd file that is simply knit to the preferred output format **(.docx or .html)**
- The .Rmd file calls the created .R file; **the .Rmd does not need to be modified in any way**

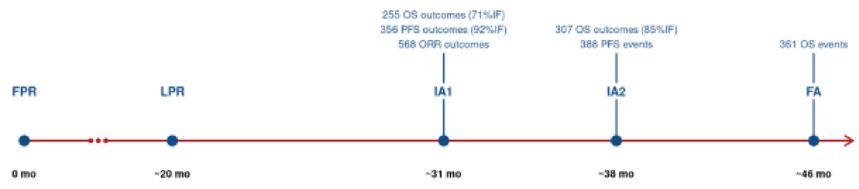
All possible weights available to each hypothesis in the graphical test are provided

Local alpha level	Weight	Testing scenario
H1: OS		
0.01900	0.76	Initial allocation
0.02499	1	Successful H2
0.02500	1	Successful H2, H3
H2: PFS		
0.00600	0.24	Initial allocation
0.02498	0.999	Successful H1
0.02500	1	Successful H1, H3
H3: ORR		
0.00001	0.00024	Successful H2
0.00002	0.00076	Successful H1
0.02500	1	Successful H1, H2

{appendMCP} calculates the expected time of each analysis, and how much information will be available at that time for each hypothesis. Information fractions can then be inferred

Hypothesis	Analysis	Events / sample size	Information fraction, %
255 OS events (Expected analysis time: 31.1 mo)			
H1: OS	1	255	70.6%
H2: PFS	1	356	91.9%
H3: ORR	1	568	100.0%
307 OS events (Expected analysis time: 37.6 mo)			
H1: OS	2	307	85.0%
H2: PFS	2	388	100.0%
361 OS events (Expected analysis time: 46.1 mo)			
H1: OS	3	361	100.0%

Study timelines are automatically generated, incorporating by default information on all milestones and key hypothesis information at each analysis



All key operating characteristics required for the SAP are outputted

In particular, this includes information on rejection rules and power for each hypothesis, conditional on each alpha level it can be tested at

E.g., the table to the left shows the relevant information for PFS and ORR in KEYNOTE-598

Analysis	Local alpha level	Nominal p-value	Exit hurdle	Local power	Information fraction, %
H2: PFS					
1	0.00600	0.00416	0.756	80.5%	91.9%
2	0.00600	0.00484	0.769	86.7%	100.0%
1	0.02498	0.01938	0.803	92.3%	91.9%
2	0.02498	0.01980	0.811	95.1%	100.0%
1	0.02500	0.01939	0.803	92.3%	91.9%
2	0.02500	0.01981	0.811	95.1%	100.0%
H3: ORR					
1	0.00001	0.00001	0.184	65.5%	100.0%
1	0.00002	0.00002	0.173	74.6%	100.0%
1	0.02500	0.02500	0.081	99.8%	100.0%

Efficient C++ code is built into the package. This allows simulation to be used to calculate operating characteristics averaged over the complete study design

E.g., the table to the right shows the expected analysis and expected time, at which each hypothesis will achieve significance

Metric	Hypothesis subset	Value
Expected Success Analysis	H1	1.39
	H2	1.08
	H3	1.17
Expected Success Time	H1	33.8
	H2	31.7
	H3	32.3

Many more outputs are also provided!

Key Takeaways

{appendMCP} provides a template-based workflow that transforms high-level study parameters into comprehensive analysis reports, including key operating characteristics tables

{appendMCP} facilitates reproducible documentation, reduces manual errors, and enables efficient communication between statisticians, clinicians, and regulatory stakeholders

Available on GitHub. Download with: `devtools::install_github("johnsonandjohnson/appendMCP")`

Johnson & Johnson

References

Anderson K et al (2025) gsDesign2: Group Sequential Design with Non-Constant Effect. R package version 1.1.7 https://merck.github.io/gsDesign2/

Boyer M et al (2021) Pembrolizumab plus ipilimumab or placebo for metastatic non-small-cell lung cancer with PD-L1 tumor proportion score ≥ 50%: Randomized, double-blind phase III KEYNOTE-598 study. *J Clin Oncol* 39:2327–38

Bretz F et al (2009) A graphical approach to sequentially rejective multiple test procedures. *Stat Med* 28:586–604

Hung H et al (2007) Statistical considerations for testing multiple endpoints in group-sequential or adaptive clinical trials. *J Biopharm Stat* 17:1201–10

Lan KKG, DeMets DL (1983) Discrete sequential boundaries for clinical trials. *Biometrika* 70:659–63

Maurer W et al (2011) Multiple and repeated testing of primary, co-primary and secondary hypotheses. *Stat Biopharm Res* 3:336–52

Maurer W, Bretz F (2013) Multiple testing in group sequential trials using graphical approaches. *Stat Biopharm Res* 5:311–20

Tang D, Geller NL (1999) Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* 55:1188–92

Wassmer G, Pahlke F (2025) rpact: Confirmatory Adaptive Clinical Trial Design and Analysis. R package version 4.2.0. https://cran.r-project.org/package=rpact

XQ, Tamhane A (2015) Allocating recycled significance levels in group sequential procedures for multiple endpoints. *Biom J* 57:90–107

P49

Beyond Dichotomization: Optimizing Composite Endpoints in Atopic Dermatitis Using Bayesian Latent Variable Framework

Arina Kazimianec[1], Dave Lunn[1], Paul Newcombe[1], Tal Otiker[1]

[1] GSK, UK

Background

Composite endpoints are critical for capturing multiple components of disease and assessing simultaneous response on several endpoints. Conventional analysis methods (such as logistic regression) require a binary definition of response, which is often derived based on a continuous score (that gets dichotomised) and combined with another binary endpoint to create a composite response. This dichotomisation can lead to a substantial loss of information, however, new innovative methods such as latent variable framework[1] allow us to retain and incorporate the continuous variable into the definition of response and therefore preserve information.

There is significant interest in evaluating such methods within the Biostatistics community, as published examples in the literature indicate substantial power increases (for example, the latent variable approach was applied to the Phase 2 MUSE trial[2]).

Conclusions

Bayesian latent variable framework demonstrated a significant value, evident in **increased precision** in the **post-hoc** analysis of Ph1b data.

Moreover, it offers **substantial power gains for PoC study design**, enabling smaller, faster, and more cost-efficient trials.

To facilitate further and wider adoption of this innovative methodology across indications, additional work is needed to fully assess operating characteristics and identify clinically meaningful differences for the composite endpoint.

Aims

- We developed a **novel Bayesian implementation** of the **latent variable framework** and applied this methodology to a post-hoc analysis of a Phase 1b double blind, placebo-controlled study (2:1 randomization) in participants with moderate to severe atopic dermatitis (NCT04975438).
- We compared the results against Bayesian logistic regression of the binary responder endpoint to assess the impact on precision.
- We also expanded this joint modelling approach for integration within a quantitative decision-making framework to design and optimise new proof-of-concept (PoC) studies in atopic dermatitis.

How does this method work?

Step 1: Model effect of treatment on endpoint components in a multivariate joint model for the two outcomes; map continuous latent to the observed binary.

Step 2: Extract implied treatment effect on a composite binary endpoint; obtain a predicted probability of meeting response definition on the composite endpoint for each participant, under each treatment, by integrating over the appropriate area of the multivariate normal density.

Key considerations

- Developing composite endpoints for novel indications without pre-defined composite endpoints (such as atopic dermatitis) presents a set of unique challenges and considerations, such as:

the absence of external benchmark data

identifying meaningful components and their thresholds (where continuous)

assuming normality for non-normal continuous data can introduce bias, potentially negating any efficiency gains

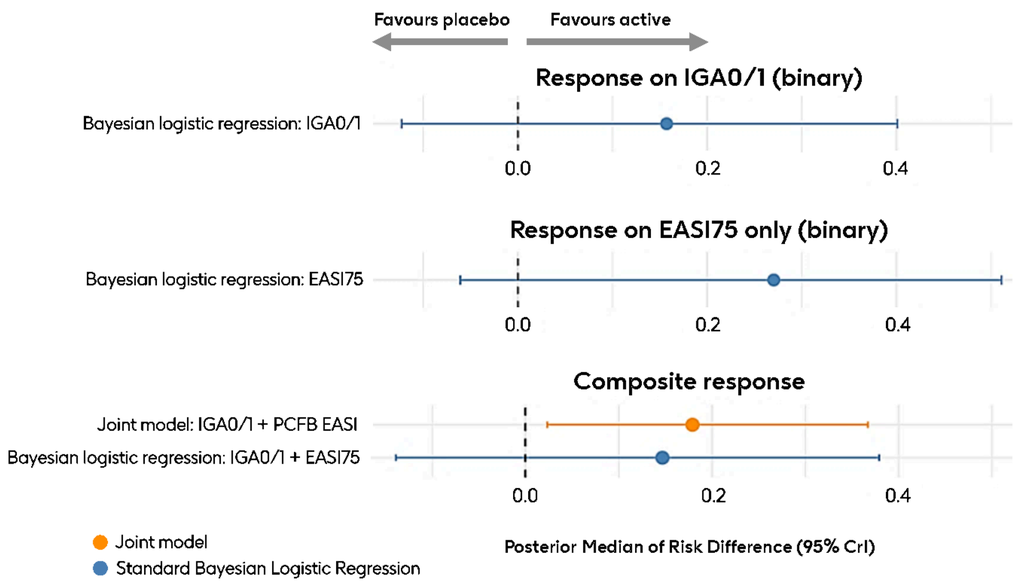
stochastic variation might sometimes yield higher/lower estimates despite asymptotic unbiasedness

defining the success criteria for the composite endpoint

Results

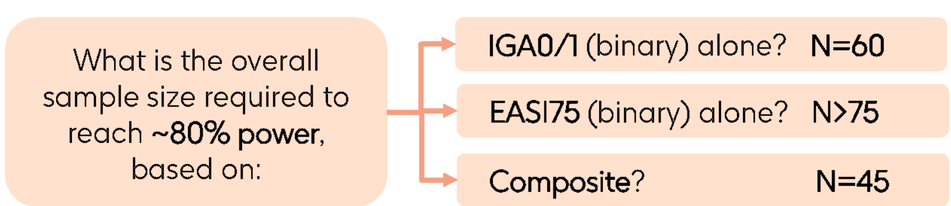
Ph1b study post-hoc analysis: Substantial precision gains in Ph1b atopic dermatitis post-hoc analysis were achieved using the Bayesian latent variable framework (joint model) for composite endpoints. The model incorporated continuous percent change from baseline (PCFB) EASI (Eczema Area and Severity Index) at Week 12 alongside binary IGA0/1 (Investigator Global Assessment Score of 0 or 1) at Week 12, showcasing the impact of integrating continuous scale information on precision (Figure 1). The standard Bayesian regression model used both binary IGA0/1 and binary EASI75 (achieving ≥ 75% Reduction in PCFB EASI at Week 12). In this analysis, the joint model generated a slightly higher point estimate than the standard approach, however, subsequent simulations demonstrated no inherent bias between the methods, showing a normally distributed, zero-centered risk difference.

Figure 1: **Ph1b study post-hoc analysis.**



Designing a new proof-of-concept study (PoC): We have also explored the implementation of the Bayesian latent framework methodology for designing new PoC studies (assuming 2:1 randomisation, true treatment difference vs placebo 30% for IGA0/1 and 30% for EASI75 as well as statistical significance as a success criteria). We demonstrated that the use of the composite endpoint within the joint modelling framework allows for a smaller sample size or substantial power gains (Figure 2).

Figure 2: **Implementation of Bayesian latent framework for new PoC studies.**



Abbreviations

PoC, Proof-of-Concept.
PCFB, Percent Change from Baseline.
EASI, Eczema Area and Severity Index.
IGA, Investigator Global Assessment Score.

References

[1] Wason J, McMenamin M, Dodd S. Analysis of Responder- based endpoints: improving power through Utilising continuous components. Trials 2020;21.

[2] McMenamin, M. et al. (2021) 'Employing a latent variable framework to improve efficiency in composite endpoint analysis', Statistical Methods in Medical Research, 30(3), pp. 702–716.

Disclosures

All authors are employees of GSK and hold equities in GSK.
Funding: GSK.

Presenting author: Arina Kazimianec, arina.x.kazimianec@gsk.com

P050

Benchmarking Statistical Methods of Dynamic Borrowing Information from Historical Data

Xiang Zhang, Jeff Zhan, Jack Knorr, Ran Liao | Eli Lilly and Company

INTRODUCTION

Historical control borrowing improves trial efficiency but risks bias when the current control population differs from the historical one — a shift known as drift.

Dynamic methods adapt borrowing to observed data compatibility. A key open question is how sensitive trial conclusions are to tuning parameter choice. We evaluate this in a clinical trial comparing combination therapy to monotherapy, borrowing N = 50 historical monotherapy controls from a completed study.

METHODS

We evaluate two dynamic borrowing approaches against a fixed-weight mixture benchmark. Let θ denote the treatment effect. Both mixture methods use:

Fixed-weight mixture [1]:

$$\pi(\theta) = w \cdot f_H(\theta) + (1 - w) \cdot f_V(\theta)$$

f_H: historical prior, f_V: vague prior

w: borrowing weight ∈ {0.25, 0.5, 0.75, 1.0}

SAM prior [2]:

$$\pi_{SAM}(\theta) = w \cdot f_H(\theta) + (1 - w) \cdot f_V(\theta)$$

$$w = R/(1 + R)$$

$$R = \frac{P(D \mid \theta = \theta_h)}{\max(P(D \mid \theta = \theta_h + \delta),\ P(D \mid \theta = \theta_h - \delta))}$$

D: observed current trial data

δ ∈ {0.1, 0.3, 0.5} SD units. As conflict exceeds δ, R approaches 0 and the informative prior is effectively down-weighted.

Hierarchical modelling [3,4]:
Historical and current controls are exchangeable with overall mean μ and heterogeneity τ:

$$\theta \sim N(\mu,\ \tau^2)$$

$$\tau \sim \text{Uniform}(0,\ s_\tau)$$

s_τ ∈ {0.001, 0.01, 0.1}

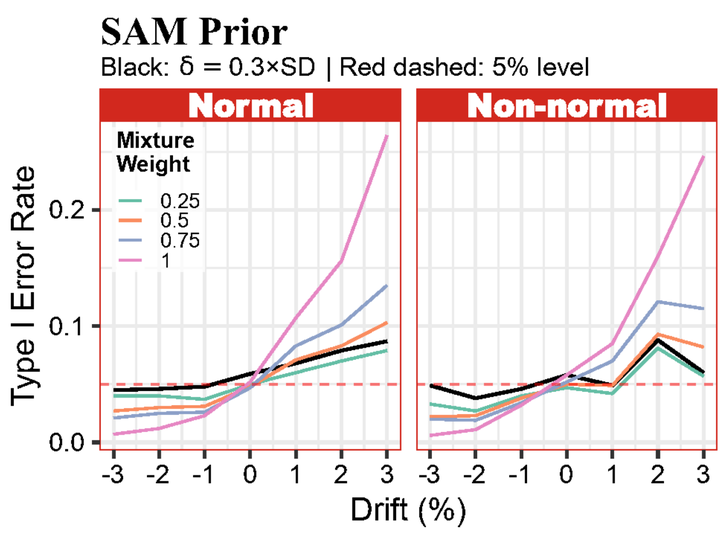
SIMULATION DESIGN

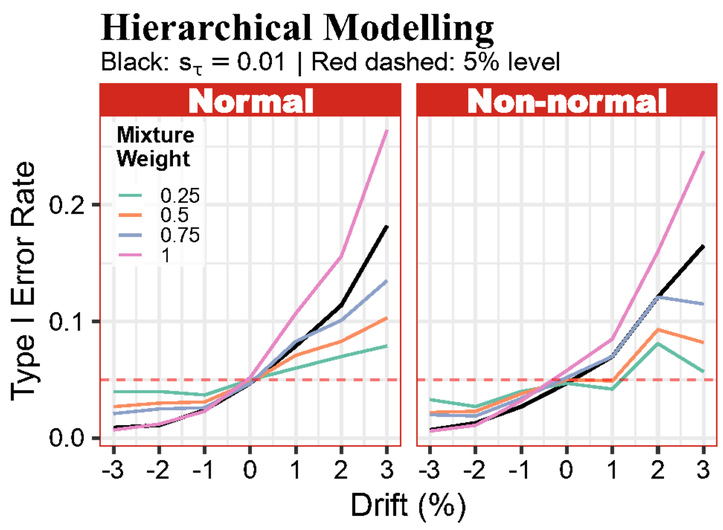
We simulated 1,000 trials across 7 drift scenarios (−3% to +3%) under Normal and truncated normal (−0.5, 0.5) distributions. Each trial has N = 50 historical (donor), N = 45 current monotherapy, and N = 45 combo patients. Decision criterion: P(combo − mono > 0 | data) > 0.95. Decision concordance with the no-borrowing analysis quantifies tuning parameter sensitivity.

RESULTS

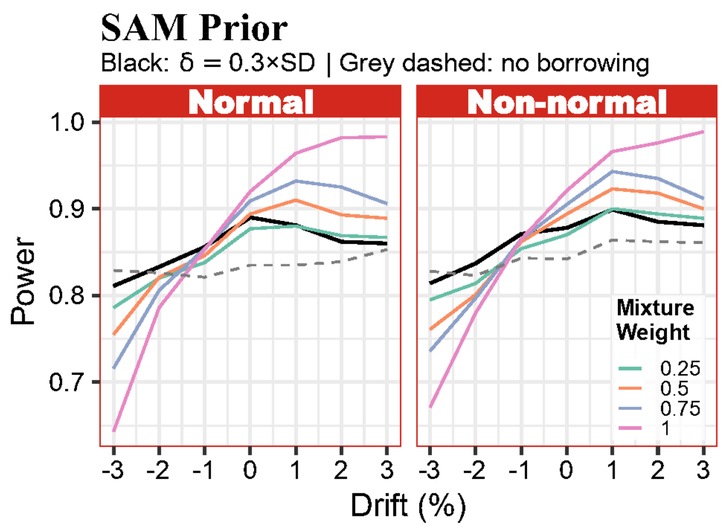
SAM and hierarchical models track the nominal Type I error rate more closely than fixed-weight priors, which inflate severely under large drift. Both dynamic methods maintain competitive power.

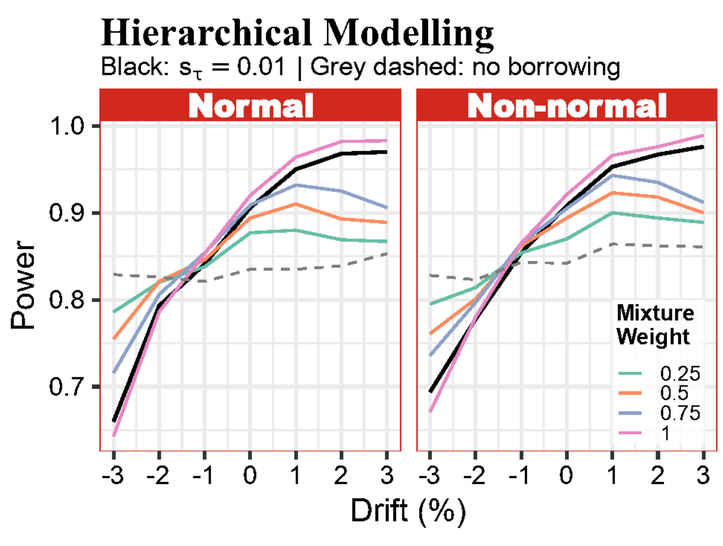
Type I Error





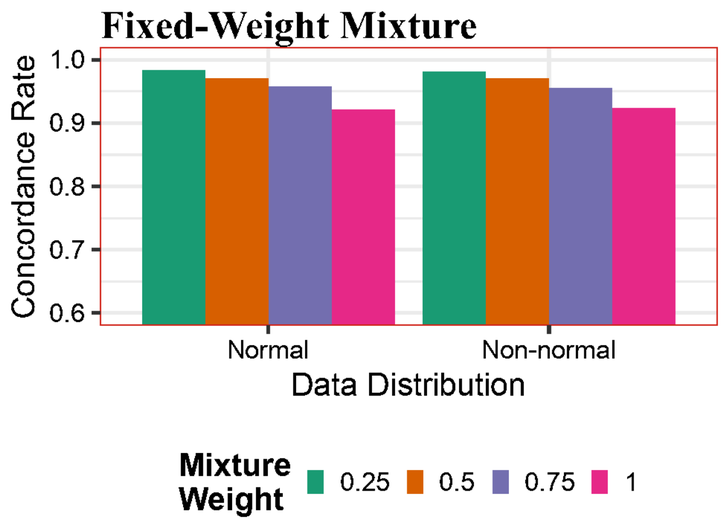
Power

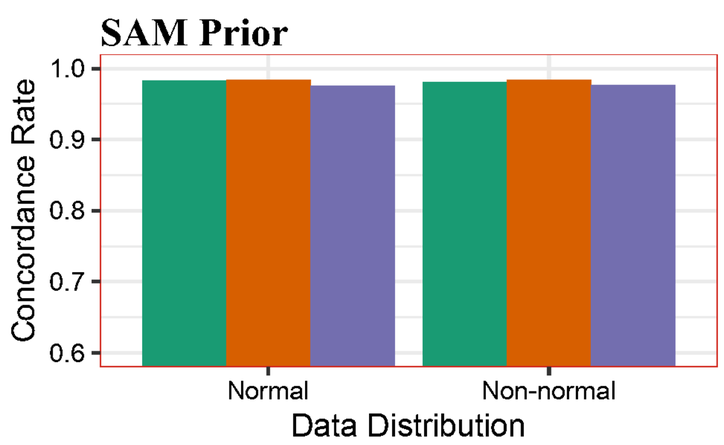


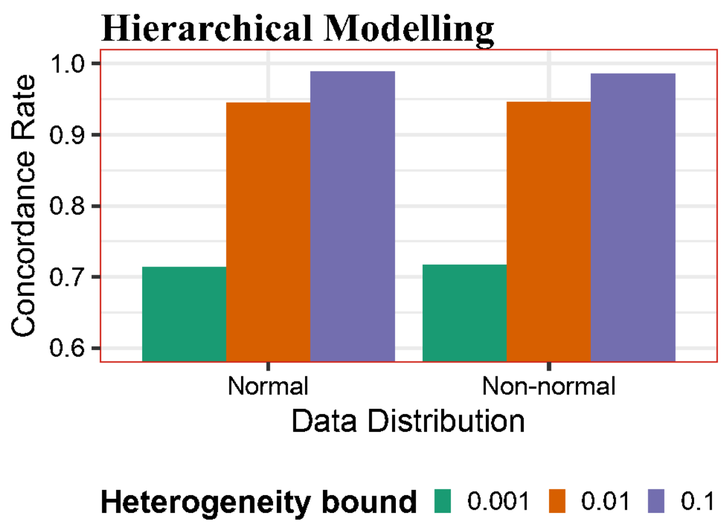


DECISION CONCORDANCE

Decision concordance measures agreement between borrowing and no-borrowing trial conclusions, averaged across all drift scenarios, under null scenarios only (combo effect = 0). Higher concordance indicates borrowing does not change the trial outcome. Each panel shows sensitivity to its method's own tuning parameter (δ for SAM, weight for mixture, s_τ for hierarchical modelling).







CONCLUSION

- Dynamic methods outperform fixed-weight priors under drift. Hierarchical modelling is more sensitive to tuning — restrictive heterogeneity bounds can alter the trial conclusion.

- SAM prior is robust to its tuning parameter under both distributions. We recommend SAM as the default.

1. Weber et al. (2021). RBesT: R Bayesian evidence synthesis tools. R package.
2. Yang et al. (2023). SAM: Self-adapting mixture prior. Stat Methods Med Res.
3. Hobbs et al. (2011). Hierarchical commensurate and power prior models. Biometrics.
4. Landau WM (2026). historicalborrow: Bayesian historical borrowing models. R package.

 | © 2026 Eli Lilly and Company

P051

FDA Bayesian Guidance: A Software Testing Procedure

Tom Parke, Pravin Madhavan
Berry Consultants, LLP

Overview

- FDA Draft Guidance names **software reliability**, **transparency**, and **reproducibility** as central to any Bayesian trial design submitted in a regulatory context — but is non-prescriptive about *how* sponsors meet those expectations.
- The gap is especially acute for Bayesian software: stochastic MCMC, hierarchical priors, and adaptive decision rules make reproducibility harder to demonstrate than for fixed-design tools.
- This poster presents a transferable, layered workflow: independent validation of Bayesian computations, logical-consistency checks on trial-design behaviour, regression testing for version stability, all on a foundation of version control, CI, and formal validation documentation.

Software engineering best practices

- **Version control** with reviewed merges. Every change lands via a pull request on a tracked branch, with a model that separates ongoing development from release stabilisation.
- **Automated build and test pipeline (CI).** Every commit triggers a reproducible pipeline that compiles all components, runs unit and integration tests, then hands off to the heavier numerical suites — so any release artefact traces back to a specific source state and a specific set of passing tests.
- **Continuous delivery (CD)** with a controlled release gate. A documented release checklist (version bump, regenerated documentation, validation reports, phased rollout) gates release; the combination of automated CI and a deliberate release gate answers "how do you know this build is the one you tested?"

Test suite creation

- Build each test case by taking an existing, well-understood design and making **the minimal change needed to exercise the new feature**; many narrow cases is far more diagnostic than a few factorial cases that mix many features.
- **Formally document every test case** with the feature exercised, the expected directional change (with approximate magnitude where possible), and the rationale.
- For each created test, review both **individual simulated trials** (trajectories, accrual, arm allocation, interim triggers) and **aggregate operating characteristics** across many simulations

Bayesian analysis testing

- Addresses the most fundamental question for Bayesian trial-design software: **is the posterior the software computes the correct posterior for the specified model?**
- Drive the software on a fixed simulated dataset picked from the test suite and compare its posterior summaries against the same model **fitted independently in a general-purpose MCMC sampler** (e.g. JAGS, Stan, etc.) chosen to share as little code as possible with the system under test — different language, different sampling algorithm, different author.
- Formally document this (inputs, comparisons performed and expected output).

Rules testing

- Read simulation output (per-trial event logs, interim records, allocation history) and verify, scenario by scenario, that **realised conduct matches specification** — e.g. interim timing against the configured trigger (accrual, events, calendar time, or completion), arm selection/dropping decisions, dose escalation and de-escalation sequences, randomisation ratios, and final-analysis triggers.

Regression testing

- The broadest layer of the test pyramid and what ultimately validates a release: pair each test from the test suite against a previously-validated baseline and re-execute the full test suite against every software release candidate build.
- Two design choices distinguish a strong suite from a weak one — **breadth** (every supported design at sufficient depth, typically thousands of scenarios) and **statistical rather than exact comparison** (equivalence tests and tolerance bands appropriate to Monte Carlo fluctuation).
- Driven by the CI pipeline against every release-candidate build, with results captured as a release artefact for the validation documentation trail.

Validation

- **Software Engineering and Testing SOP** defining how software development and testing is conducted and how changes in the software are controlled.
- **Software Release Validation** documentation stating what will be tested per release and the acceptance criteria.
- **Traceability Plan** stating requirements and feature-to-test coverage matrices mapping each requirement to the specific tests that exercise it.
- **Evidence** - per-release: Bayesian Analysis testing, Rules testing, Regression testing reports plus a signed Test Certificate.
- Read together, the four layers let a reviewer trace from a stated feature of the software → to the test that covers it → to the automated suite that runs it → to the report that records its result on a specific release — which is the **audit trail** a regulatory submission needs.

PSI 2026
P053

Applying the Estimand Framework to Dose-Ranging Trials: A Case Study

A practical illustration for Phase 2 dose selection using a cutaneous lupus erythematosus case study

Chun-Hang Tang (Astellas Pharma), Shuying Yang (GSK), members of the Early phase and Clinical Pharmacology sub-team of EFSPI/EFPIA Estimands Implementation Working Group

Key message: in dose-ranging trials, the estimand should target the dose to take forward - not only treatment differences at tested doses.

WHY THIS MATTERS

- ICH E9(R1) was developed for confirmatory trials, but it remains underutilized for phase 2 dose-ranging studies
- Problem seen in reviewed examples: methods such as Bayesian Emax or MCP-Mod were described, but the estimand often stayed at "mean difference from placebo for each tested dose".
- That misses the dose-selection question: "Which dose(s) should be studied in Phase 3?"

OBJECTIVES

- Applied the estimand thinking process retrospectively to a phase 2 dose-ranging trial of litifilimab in moderate or severe cutaneous lupus erythematosus, using publicly available information, as a case study
- Dosing range considered: 50-450 mg SC, every 4 weeks, with an additional Week 2 loading dose.
- Efficacy endpoints: CLASI-A reduction from baseline and CLASI-50 response at Week 16.
- Goal: choose a dose to optimize efficacy while maintaining an acceptable safety profile. A dose not directly tested, but within the tested range, may be considered.

THE SHIFT IN THE QUESTION

Conventional summary measure
Treatment difference from placebo at each tested dose

↓

Decision-oriented summary measure
Dose giving 80% of maximum efficacy while staying below a 10 percentage-point safety-risk increase

ESTIMAND-THINKING PROCESS

A seven-step thinking process in ICH E9(R1) Training Materials converts the trial objective into estimands, design choices, and analyses.

1 Objective | 2 ICEs | 3 Strategies | 4 Estimand | 5 Design + estimation | 6 Sensitivity | 7 Document

Step 1: Trial objectives and clinical questions the trial should answer

Primary efficacy
What dose gives 80% of the estimated maximum therapeutic effect within 50 - 450 mg SC?
Endpoint: percent reduction from baseline in CLASI-A at Week 16.

Safety
What dose has <10 percentage-point absolute risk increase for a composite safety event vs. placebo by Week 16?
Composite Safety Endpoint: SAE, AESI, or discontinuation due to AE.

Supplementary efficacy
CLASI-50 response at Week 16, using intercurrent-event handling aligned to phase 3 expectations.

Steps 2/3: Intercurrent event strategies

ICE strategy choice is driven by the clinical question. The same event can be handled differently for primary efficacy, safety, and supplementary efficacy.

Methods are illustrative and should be tailored for each trial.

ICE: Treatment discontinuation due to lack of efficacy or safety concerns

Estimand	Strategy	Motivation
Primary efficacy	Hypothetical	Estimate the pure treatment effect: what would happen if discontinuation did not occur?
Safety	While-on-treatment / composite	For discontinuations due to lack of efficacy, count AEs while exposed. For discontinuation due to safety, the event is itself part of the safety question.
Supplementary efficacy	Treatment policy (or composite)	Align with phase 3 expectations; some authorities may request non-response/composite handling after discontinuation.

ICE: Start/increase concomitant therapy or corticosteroids

Estimand	Strategy	Motivation
Primary efficacy	Hypothetical	Avoid attributing improvement from rescue/standard-of-care changes to the investigational treatment.
Safety	Treatment policy	Composite safety endpoint (SAE, AESI, or discontinuation due to AE) is sufficiently specific; attribution is difficult once new medication is introduced, so count events regardless.
Supplementary efficacy	Treatment Policy (or composite)	Use ICE handling aligned with phase 3 expectations

Step 4: Construct dose-ranging estimands

Primary efficacy
Treatment: litifilimab 50-450 mg SC + SoC vs placebo + SoC
Population: moderate/severe CLE patients (CLASI-A ≥8)
Endpoint: percent reduction from baseline in CLASI-A at Week 16
ICE handling: hypothetical for discontinuation and therapy changes
Population-level summary: dose corresponding to 80% of maximum placebo-adjusted mean percent reduction within the tested range

Safety
Endpoint: composite safety event - SAE, AESI, or treatment discontinuation due to AE by Week 16
ICE handling: while-on-treatment for lack-of-efficacy discontinuation; composite for safety discontinuation; treatment policy for therapy changes
Population-level summary: dose with <10 percentage-point absolute risk increase vs. placebo

Supplementary efficacy
Endpoint: CLASI-50 response at Week 16
ICE handling: treatment policy for discontinuation and therapy changes
Population-level summary: dose corresponding to 80% of the increase in proportion of patients achieving response vs. placebo within the tested range

Step 5: Design + Estimation

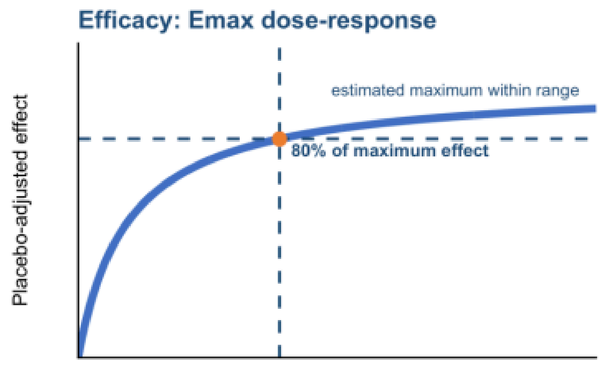
Primary efficacy model
Use Emax dose-response model to calculate the dose giving 80% of estimated maximum effect within the tested range on Week 16 estimates from an MMRM model. Uncertainty of the estimated dose is quantified via bootstrap.

Safety model
Use logistic regression to model the probability of composite safety event vs. dose and calculate the highest dose below the risk threshold.

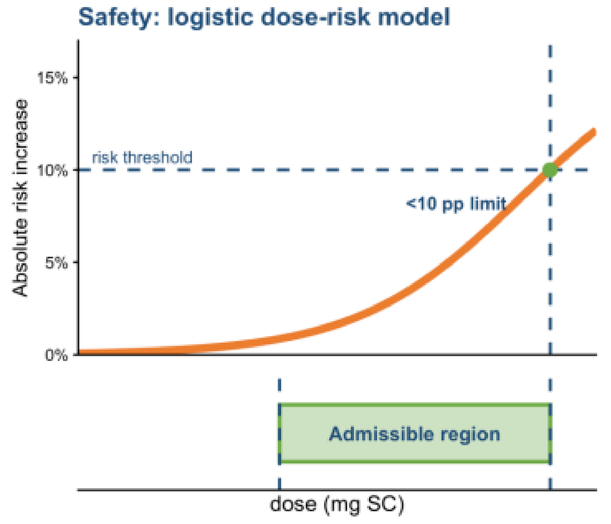
Select a Phase 3 dose(s) in the window:

Charts are illustrative and not observed trial results.

Efficacy: Emax dose-response



Safety: logistic dose-risk model



Step 6: Assumptions and sensitivity analyses

Primary efficacy
The Emax model assumes monotonicity of efficacy. Alternative dose-response models that do not enforce monotonicity can be used.

Safety
The logistic regression model assumes that the log-odd of the safety event is linear with respect to dose. Alternative models with fewer assumptions, such as isotonic regression, can be used.

RECOMMENDATIONS

- Define population-level summaries in terms of dose, not only pairwise treatment differences.
- Separate primary efficacy, supplementary efficacy, and safety questions because their intercurrent-event strategies differ.
- State whether untested doses within the studied range can be selected.
- Split treatment discontinuation by reason so ICE strategies match the clinical question.

TAKEAWAY

- Phase 2 dose-ranging estimands should make dose selection explicit: ask which dose on the curve should be taken forward.
- This improves coherence between objective, design, data collection, and analysis - and creates a clearer bridge to the next development decision.

References: Werth et al. NEJM 2022; Furie et al. J Clin Invest 2019; ICH E9(R1) Training Materials. All study information obtained from public sources. The authors have no affiliation with the sponsor of the study.

The co-authors would like to thank Sue McKendrick (Thermo Fisher Scientific) for extensive discussion and feedback on this work.

This poster is intended for informational purposes only and does not replace independent professional judgment. This poster is not intended to provide medical or legal advice. Statements of fact, positions taken and opinions expressed are those of the authors individually and, unless expressly stated to the contrary, do not necessarily reflect the opinion or position of Astellas, GSK or any of their subsidiaries and/or related entities.

chun-hang.tang@astellas.com
shuying.y.yang@gsk.com

Additional materials available at
https://eiwg.info

PSI 2026 CONFERENCE
15 - 17 JUNE 2026

P054

Agenda of the novel Special Interest Group for Medical Devices: Finding similarity within the diversity of Medical Devices

Emilie Gérard[a], Wilmar Igl[b], Martin Fahy[c], Michael Mader[d]

a) Brain&Mind, Paris, France; b) VERANEX Inc., Uppsala, Sweden; c) JenaValve, Irvine, CA, USA; d) Nobel Biocare, Zürich, Switzerland

eSIG Medical Devices

Introduction

There are **fundamental differences** in the scientific, clinical, and regulatory pathways for the development of medical devices compared to drugs and biologics.

Hence, the **Special Interest Group for Medical Devices (SIG MedDev)** was formed as part of the PSI and EFSPI organizations in December 2025 to study the implications of these differences for the application of statistical methods.

Join us to contribute to the identified **high-impact topics** with special relevance for medical devices !
https://www.psiweb.org/sigs-special-interest-groups/medical-devices-sig

Holistic health model of the patient

Medical device as a multi-domain intervention …

Anatomical	Physiological	Biochemical
Implants, prosthetics, bone fixation devices	Insulin pump, ventilator, pacemaker	Inhaler, dialysis filter, glucose monitor
Psychological	**Functioning**	**Social**
Neurofeedback, brain stimulation, biofeedback	Limb prosthetics, exoskeletons, mobility aids	Hearing aids, telehealth devices, communication aids

… With specific challenges

Device
Materials, mechanics, software

Procedure (clinician)
Technique and protocol

Mechanisms of action
Broad and multimodal, not chemical

User (patient)
Patient ability, operator skill

Performance
Depends on device, procedure and user

Medical device development
A multi-domain intervention

Clinical perspective
Intended use and context of use

Shapes the statistical design
The effect must disentangle device, procedure and operator interactions — not isolate a single molecule

Statistics

Clinical topics

Estimand Framework (ICH E9 R1, 2020)

Originally an addendum for drugs and biologics, but increasingly applied to medical devices.

Adoption for medical devices since ISO14155 (Mars 2026)

For devices, specifying the estimand is harder: intercurrent events, the comparator, and the operator all shape what the effect means

Applied well, it gives device trials a common language for aligning design, analysis, and interpretation

Intercurrent events (IE), comparator & standard of care (SoC)

Identification & definition especially challenging for medical devices

Example of IE
Mobile app: discharged battery
Assessed by likelihood & impact

Comparator arm
Drugs, devices or non-interventional
Wide range of competitors vs. drugs

Sham comparator
A group receiving a procedure/device appearing identical to the active one but containing no active processes or components

Study designs

RWE
Observational data from routine clinical practice & registries

RCT
Randomised controlled trial; may include sham comparator arm, predicates or SoC

No clinical study
Based on SOTA evidence (state-of-the-art bibliography) when sufficient prior evidence

Post-Market Clinical Follow-up
Registries, single-arm trials, trials with historical controls

Statistical modeling of endpoints

Heterogeneity of endpoints	anatomical, physiological, social, ...
Endpoint nature	mechanical, electronic, or functional performance vs. biological response
Distributional assumptions	non-normal, time-to-event, or measurement-error structures are common in device trials
Causal estimand	the treatment effect must account for the device, the procedure and the operator

Hybrid treatments

Combining medical devices with drugs or biologics

Examples
Drug-eluting stents, insulin pump + CGM

Challenge
Disentangling device vs. drug contribution

Statistical methods & emerging topics

Blinding
Often infeasible - implants, procedures, worn devices reveal allocation. Mitigated by sham comparators and blinded outcome assessors (single-blind).

Bayesian methods
Informative priors of particular interest for medical device studies

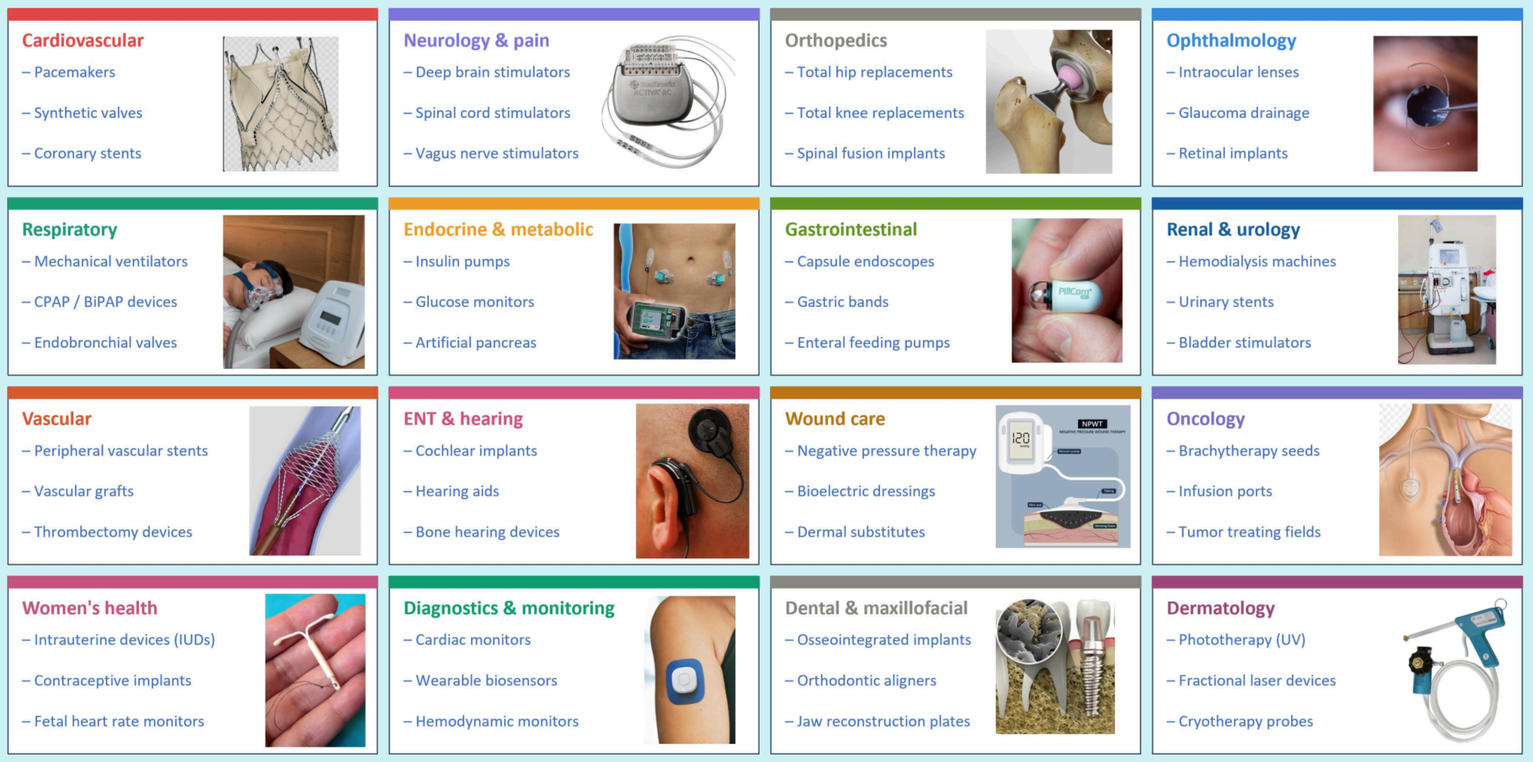
AI methods (e.g. LLMs)
Evaluated against emerging legislation (EU AI Act) - developments ongoing

Other challenges

Safety
e.g. late adverse device reactions due to material wear, sometimes only occurring after 10+ years

Device interactions
Imminent interactions between multiple medical devices per participant

Figure 1: Overview of Medical Device Categories



Regulatory

Regulatory framework: Pharma vs. Medical Device

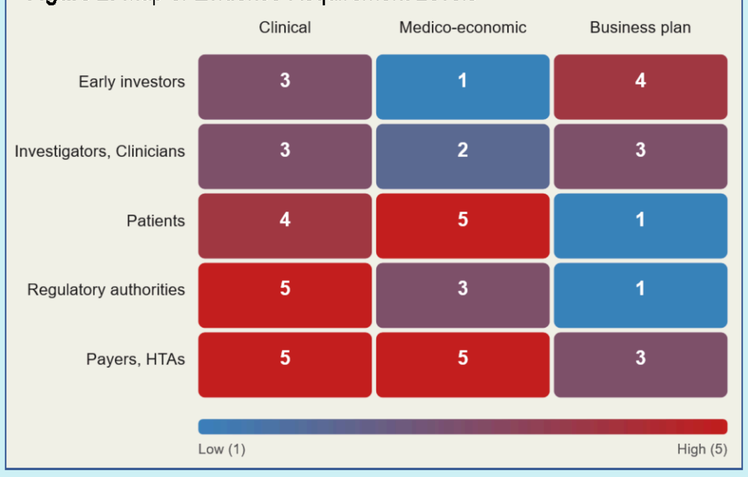
Topic	Drugs / Biologics	Medical devices
Regulatory framework*	• ICH guidelines (E6, E8, E9…) • GCP / GMP — CTD dossier • Uniform global framework EU, US, Japan, China, Asia-Pacific EU Clinical Trials Information System (CTIS)	• ISO13485 & 14155:2026 + CE MDR/IVDR (EU) • 510(k) / PMA / De Novo (US) • Fragmented across regions with some international framework such as MDSAP EU European Databank on Medical Devices (Eudamed)
Competent authorities	• EMA (EU), FDA CBER CDER (US), PMDA (JP) • National drug agencies • Centralised review FR ANSM	• Notified Bodies (EU) + country-authorities • FDA CDRH, CBER, CDER (US) • EMA when device & drug combination FR Notified Bodies + ANSM
Development stages*	• Long timelines (10–15 years) • Preclin. → Phase I → II → III → IV • Linear, frozen product	• Faster to market (2–10 years) • Feasibility → Prototype → Non-clinical Tests → Clinical Evaluation(s) → Certification • Potential biocompatibility • Potential versioning & software updates
Clinical evidence*	• RCT mandatory (Phase III) • Placebo-controlled standard • Large sample sizes required • Post-marketing studies (Phase IV)	• RCT, RWE or SOTA biblio. • Sham comparator or gold standard • Predicates comparator (US) • PMCF (EU, >Class I) • Vigilance & surveillance
Reimb. Agency assessment & HTA*	• Agency approval → HTA → pricing • Benefit/risk ratio central • EUNETHTA Joint Clinical Assessment FR Commission Transparence (HAS) SMR / ASMR evaluation	• Country-dependent, HTA → pricing • EUNETHTA for high-risk devices FR CNEDiMTS: ASSD/SSED Added clinical benefit evaluation
Reimburse-ment strategy*	• Standard reimbursement pathway • Pharmacoeconomic dossier required FR Early access: ATU/AAP Negotiated pricing (CEPS) Liste en Sus / MEA-RSA Remboursement Séc. Sociale	• Direct reimb. with SOTA biblio. (country-dependent) FR PECAN (Class I): anticipated reimb. LPPR / LPP: liste des produits Telesurveillance: dedicated pathway DMDIV: CNEDiMTS + CEPS Post-market FU replaces prior RCT
Market strategies*	• Marketing authorisation required • Phase IV post-approval studies • Regulators, payers, HCPs, patients	• Depend on country, business model, product & industrialization • Post-market FU • Regulators, payers, investors, patients
Challenges	• Long & costly approval process • Rigid protocol amendments • Harmonisation across regions • AI/ML & Bayesian adoption • EUNETHTA Joint Clinical Assessment	• Applying regs to AI, ML, evolving tech • High regulatory costs & admin. burden for small companies • Device class & potential versioning • Classification mismatch across regions • Post-market surveillance burden • EU AI Act compliance for SaMD

Key difference: regulatory pathway & market strategy

- Pharma: mandatory RCT → marketing authorisation → HTA → reimbursement. One fixed molecule. Harmonized across molecules.
- Med Dev: RCT not mandatory (SOTA/predicate) → direct reimbursement (PECAN, telesurveillance).
- There are too many differences between Med. devices to be fully harmonized (e.g. non-harmonized reimbursement in EU).
- Med Dev: marketing authorization could be approved by Notified Bodies (private) or auto-certified.
- When Med Dev = treatment: agency assessment → HTA pathway as for drugs applies (with same ICH & GCP requirements).

* : Med Dev class-dependent

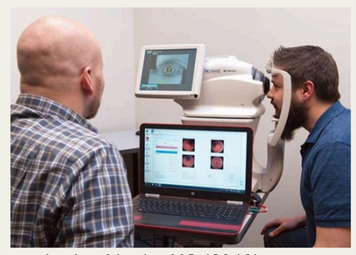
Figure 2: Map of Evidence Requirement Levels



Case-study

AN AI RETINAL SCREENING TOOL CLEARED BY FDA — WITHOUT A SINGLE RANDOMISED CONTROLLED TRIAL

WHAT IS IDX-DR?



- AI software analysing retinal fundus photographs in primary care
- Binary output: diabetic retinopathy detected / negative
- Operable by non-specialist staff
- Designed to scale screening where eye specialists are scarce
- First autonomous AI diagnostic device authorised in the US (2018)

DRUGS / BIOLOGICS

Efficacy proven by a **pre-specified RCT**: fixed protocol, fixed population, fixed comparator. Evidence generated before approval, then product is frozen.

MEDICAL DEVICE / AI

Performance validated on a **real-world dataset** of 900 patients across 10 US primary care sites. Algorithm locked at submission; deployment performance monitored post-market.

- ✓ IDx-DR was authorised via De Novo pathway — not a traditional PMA.
- ✓ FDA reviewed analytical & clinical validation studies showing 87.2% sensitivity / 90.7% specificity against a grader reference standard.
- ✓ No drug-style Phase III trial was required or expected.

Iterative by design: Software updates require re-validation, not a new clinical trial — a notion absent in pharma

RWE replaces RCT as primary evidence: Site diversity, patient mix, and camera models all shape algorithmic performance — not just the molecule

Algorithm lock vs. drug batch release: In pharma, the molecule is stable; in AI devices, the "product" is the model weights — frozen at a commit hash

Key takeaway: Medical device regulation evaluates *performance in context* — how a tool behaves across sites, operators, and patient populations — not just whether an active ingredient works. AI amplifies this difference dramatically.

References

[1] Medical Device Coordination Group. (2025). Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR (MDCG 2019-11 rev. 1). Medical Device Coordination Group. https://health.ec.europa.eu/document/download/b45335c5-1679-4c71-a91c-fc7a4d37f12b_en

[2] Medical Device Coordination Group. (2021). MDCG 2021-24: Guidance on Classification of Medical Devices. Medical Device Coordination Group. https://health.ec.europa.eu/system/files/2021-10/mdcg_2021-24_en_0.pdf

[3] Bakker, J. P., Barge, R., Centra, J., et al. (2025). V3+ extends the V3 framework to ensure user-centricity and scalability of sensor-based digital health technologies. Npj Digital Medicine, 8(1), 51. https://doi.org/10.1038/s41746-024-01322-2

[4] U.S. Food and Drug Administration. (2025, December 15). FDA Eliminates Major Barrier to Using Real-World Evidence in Drug and Device Application Reviews. https://www.fda.gov /news-events/press-announcements/fda-eliminates-major-barrier-using-real-world-evidence-drug-and-device-application-reviews ; Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: https://www.fda.gov/media/190201/download

[5] European Medicines Agency. (2025). Reflection Paper on Patient Experience Data (EMA/CHMP/PRAC/148869/2025). European Medicines Agency. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-patient-experience-data_en.pdf

[6] The Artificial Intelligence Act. (2026). EU Artificial Intelligence Act: Up-to-Date Developments and Analyses. https://artificialintelligenceact.eu ; Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing

[7] Abràmoff MD, et al. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine, 1, 39. https://www.nature.com/articles/s41746-018-0040-6

[8] US FDA (2018). De Novo classification request for IDx-DR. Docket DEN180001. https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf

Corresponding author: Emilie Gérard — emilie.gerard@brainandmind.fr

Speaker of the eSIG Medical Devices: Michael Mader — michael.mader@envistaco.com

Delivering interim analyses in blinded randomized clinical trials: plan, protect, adapt, align.

A practical case study from IA delivery in a double-blind Phase II trial

Bisabet Garcia R&I Biometrics and Statistical Innovation Respiratory & Immunology BioPharmaceuticals R&D AstraZeneca Barcelona Spain

P055

Background and objective

Interim analyses (IAs) in randomized clinical trials are carefully planned, but delivery often becomes operationally complex once recruitment and treatment are underway. Delays in vendor data, unusable samples, evolving sponsor needs, and rapid recruitment can affect timelines and decision-ready reporting. In blinded studies, these challenges are amplified by the need to protect the blind and avoid operational bias.

This poster describes practical strategies and lessons learned from delivering an IA in a blinded Randomized Clinical Trial (RCT), using the framework: plan, protect, adapt, align.

Case study: IA in Phase II Study

A double-blind, placebo-controlled Phase II study in patients following a Phase I study in healthy participants.

The IA was a descriptive, decision-support rather than a formal inferential analysis, to identify if there were any major differences in pharmacology between the Phase I & II populations. The interim population was defined as participants meeting the prespecified evaluability criteria at the data cut-off, with domain-specific availability requirements for PK, PD, and safety.

The IA was planned after approximately 10% of evaluable participants had completed 4 weeks of treatment. Delivery required coordination across sponsor functions and multiple external vendors.

A practical framework for IA delivery

Plan – Protect – Adapt – Align

Plan

- Define IA questions
- Identify target datasets and outputs
- Agree cut-off rules and review needs

Protect

- Preserve role separation
- Control information flow
- Restrict access to sensitive material

Adapt

- Manage late or incomplete data
- Handle scope changes pragmatically
- Support decision-making without compromising integrity

Align

- Coordinate vendors and sponsor teams
- Clarify timelines and responsibilities
- Standardize documentation and deliverables

Why IA delivery was complex

The IA package required integration of data on randomization, disposition, demographics, baseline characteristics, exposure, safety, pharmacokinetics (PK), pharmacodynamics (PD), and protocol deviations.

These data came from electronic data capture (EDC) and multiple vendors, each with different turnaround times and workflows.

As the IA approached, new questions arose on patient inclusion, baseline disease characterization, and use of additional PD data. The challenge was to address these pragmatically while maintaining trial integrity, timely delivery, and fit-for-purpose outputs.

Examples: Challenges and Solutions

	Patients enrolled close to IA cut-off	Baseline characteristics	PD analyses
Challenge	Patients were randomized close to IA cut-off, leaving a limited window for data entry, cleaning and logistics.	Key baseline characteristics were not available on time through the planned data flow.	Limited cleaned data were available at cut-off.
Action	Close communication with sites and vendors to ensure rapid turnaround.	Baseline summaries were derived from vendor reports.	Supplementary internal summaries and plots were prepared.
Outcome	All patients included with minimal timeline and operational impact.	The interim population could still be adequately characterized.	Additional PD information was available to further contextualise interpretation.

Lessons learned

Theme	What matters in practice
Planning	Define clinical decision questions and align essential outputs to address. Ensure cut-off and cleaning rules match the actual IA population.
Vendor readiness	Start data transfer and data availability planning early, as delayed setup can become a critical path issue. Map out granular timelines ahead of study initiation.
Adaptability	Be open to opportunities for late adjustments that may add value, but explicitly assess feasibility, resource impact, and timeline risk.
Vendor understanding	Understand how partner datasets are generated, especially when internal supplementary analyses may be needed.
Communication	Define responsibilities for delays and restricted materials. Where multiple outputs are generated, with different levels of data cleaning, ensure clear guidance for the review committee guidance.
Resources	Plan realistically for programming, validation, and technology constraints around the IA timeline.

Conclusions

Delivering an IA in a blinded RCT requires more than technical readiness. It depends on careful planning, protection of the blind (by appropriate governance of unblinding activities), adaptation to operational realities, and alignment across teams and vendors.

The principles plan, protect, adapt, align provide a practical framework for statisticians involved in interim analyses delivery in pharmaceutical clinical development.

Supported by AstraZeneca

Presented at PSI Conference Belfast 15-17 June 2026

Acknowledgements

The author sincerely thank the cross-functional contributors involved in study operations, scientific oversight, data review, and project support. Their collective efforts and shared learnings made this work possible.

Bisabet Garcia is an employee of AstraZeneca and may hold shares. Research was funded by AstraZeneca.

P056

Discrete-Time Modelling of Sustained eGFR Decline in CKD Trials: Simulation-Based Comparison of Cox, Gompertz and Weibull Approaches

José Sánchez
Biometrics, Late CVRM, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

PSI Conference 2026 | Poster P056

Background

- **Problem:** CKD confirmatory endpoints rely on sustained eGFR decline detected only at visits, biasing event times[1].
- **Gap:** Assigning event time to detection visit overestimates time to event.
- **Approach:** Model time to ≥50% decline as discrete time to event across visit intervals within GLM framework.
- **Benefit:** Hazards as conditional probabilities; ties handled naturally.

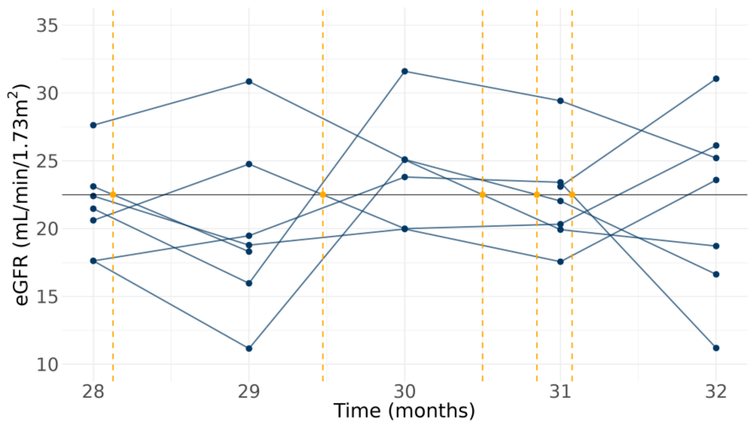


Figure 1: Assigning a scheduled visit as the TTE consistently overestimates the true event time

Gompertz model for time-to-event

- The discrete hazard $l(t|x)$ for event time t and covariates x is defined as $l(t|x) = P(T = t|T > t-1, x)$.
- Discretizing time in intervals for the Cox model results in the Gompertz model[2] defined by:

$$l(t|x) = 1 - \exp\left(-\exp\left(b_0 + x^T b\right)\right)$$

$$b_0 = \log\left(e^{c_t} - e^{c_{t-1}}\right), c_t = \log\left(\int_0^{a_t} l_0(s)ds\right)$$

Simulation setup

Models	Cox, Gompertz, Weibull
Sample sizes	50, 100, 500
Visits	7, 9, 13, 25, 49, 97
Effect	[0%, 20%]

- eGFR ∼MVN with AR(1), $\sigma^2 = 25$ and $\rho = 0.5$ covariance.
- Linear decline, baseline 45 mL/min/1.73m^2.
- Effect E is % delay in mean time to event compared to placebo (can be translated to $0 \leq$HR ≤ 0.42 as E^{-k} for Weibull shape parameter k).
- Outcomes: hazard ratio point estimate, bias/variance, power and Type I error.

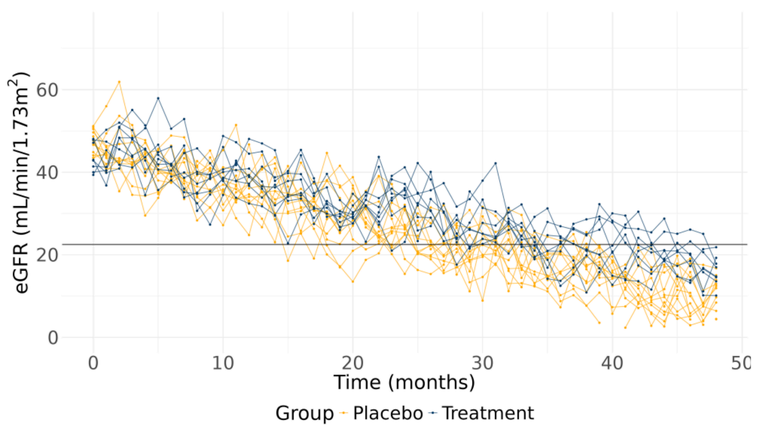


Figure 2: A subset of one simulated dataset.

Results

- **Gompertz is less biased**: for few visits and strong clear effect

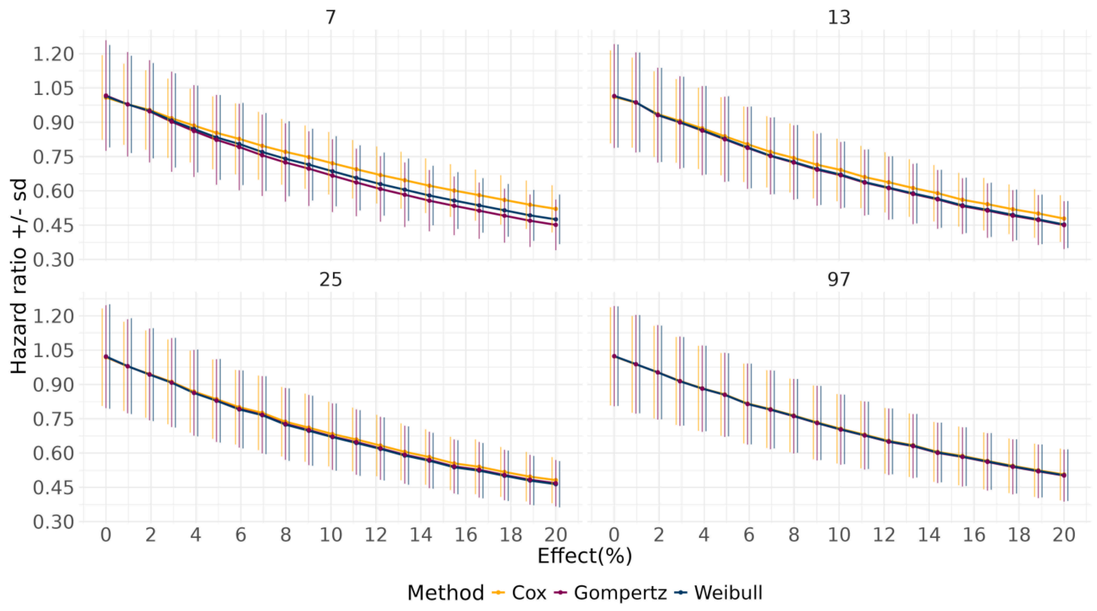


Figure 3: Variability decreases with sample size and visits; model differences shrink.

- **Gompertz more powerful than Cox**: for small sample size or few visits.

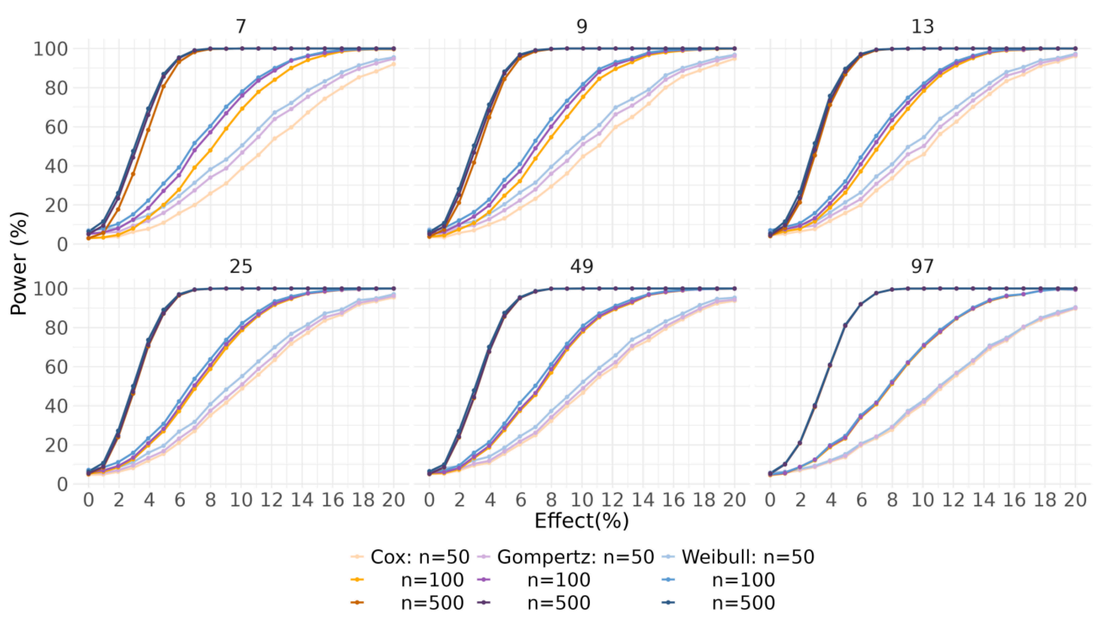


Figure 4: Power rises with sample size and visits; model differences shrink.

- **Gompertz has consistent Type I error**: comparable to Cox for larger number of visits; better than Weibull.

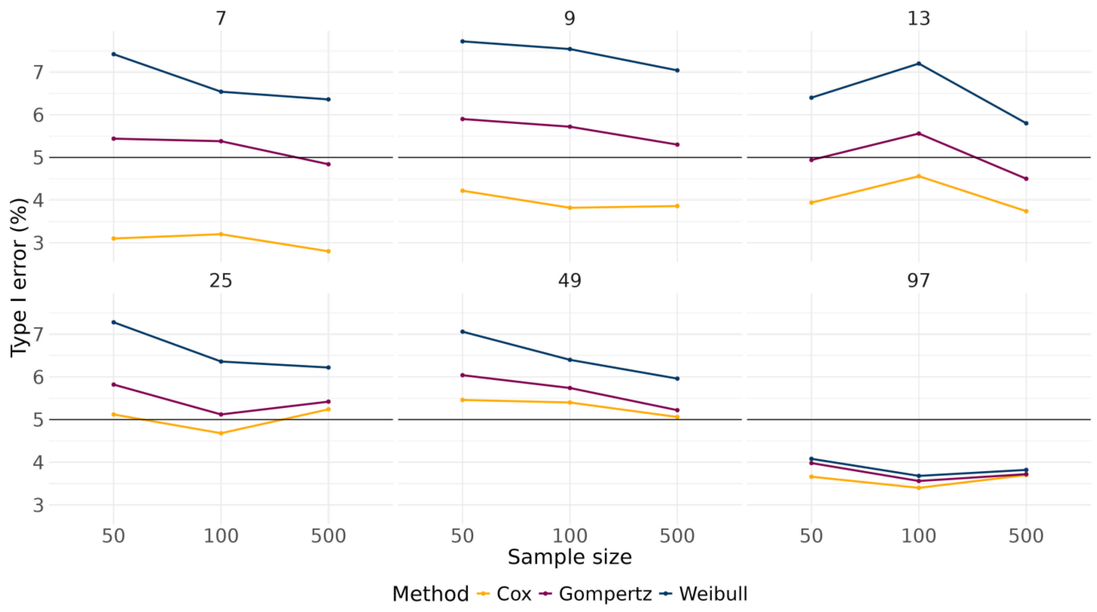


Figure 5: Possible interaction between sample size and visits; large number of visits required.

Conclusions

- **Discrete-time modeling reduces bias** when events are detected only at visits.
- **Gompertz performs best** with sparse visit schedules; converges as visits increase.
- **Practical guidance**: Prefer Gompertz for small sample sizes or number of visits; otherwise differences are minor.

Herrington WG. et al. Levin A., Agarwal R.
International consensus definitions of clinical trial outcomes for kidney failure.
Kidney International, 2020.

Schmid M. Tutz G.
Modeling Discrete Time-to-Event.
Springer, 2016.

P057

When to Weight, When to Match: Considerations from Two Real-World Evidence Case Studies

Lucy Clark MSc
Affiliation: Phastar, Electric Works, 3 Concourse Way, Sheffield, S1 2BJ, UK

INTRODUCTION

Real-world evidence (RWE) studies aim to measure the effectiveness, safety and utilisation of medical products in real-world settings[1].

However, unlike traditional randomised controlled trials (RCTs), treatment comparisons in observational RWE studies are often susceptible to bias from factors such as lack of randomisation, unequal follow-up durations, variations in patient care, variability in outcomes assessments and subject selection.

In this poster, we will discuss and compare two commonly applied propensity score-based approaches, **propensity score matching (PSM)** and **inverse probability of treatment weighting (IPTW)** for the analysis of real-world data (RWD) in the context of two RWE studies. Both methods aim to minimise confounding due to imbalances in observed baseline covariates, thereby improving the comparability between treatment groups and allowing for a more unbiased estimation of treatment effects.

METHODS

A propensity score (PS) is a subject's probability of receiving a particular treatment, conditional on their observed baseline characteristics. Propensity scores are typically estimated using a logistic regression model, with treatment as the dependent variable (treated vs. untreated/control) and observed baseline characteristics as independent variables[2].

PSM uses the PS to pair individuals from different treatment groups with similar scores. This approach aims to create matched cohorts with similar distributions of observed baseline characteristics included in the PS model, to mimic the balance achieved through randomisation[2].

IPTW works by assigning a weight to each subject based on the inverse of their PS. This approach aims to creates a pseudo-population in which the distribution of observed baseline characteristics is balanced across the treatment groups. Weighted subjects contribute proportionally to the analysis based on their assigned weights[3].

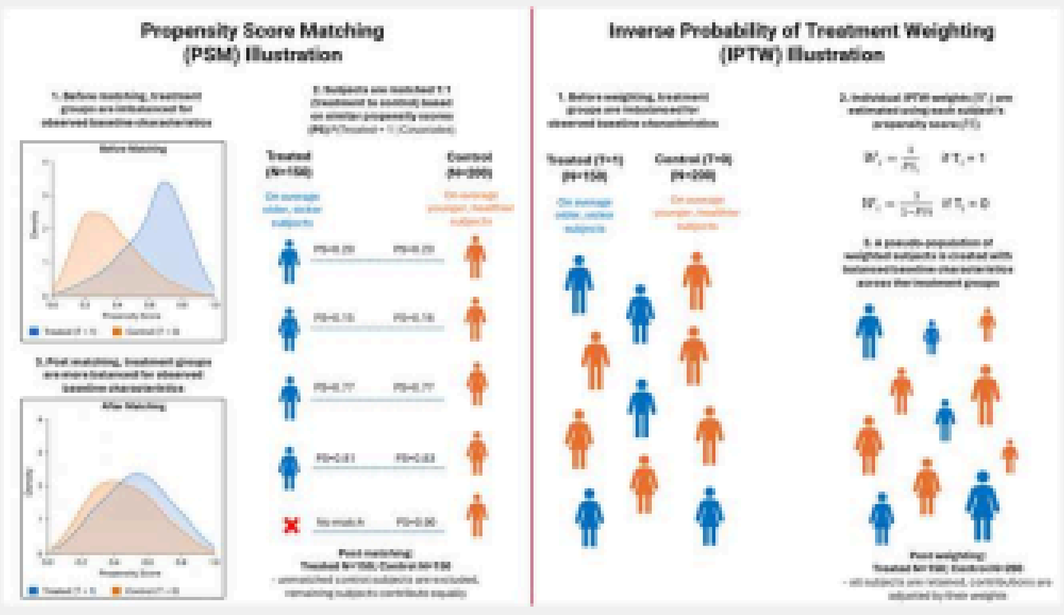


Figure source: Introduction to Propensity Score Methods

Balance between treatment groups before and after matching or weighting should be assessed. This can be done using standardised mean differences of baseline covariates included in the PS model, and visual inspection of the distribution of the PS using histograms/density plots.

CASE STUDIES

	RWE Study 1	RWE Study 2
Study design	Retrospective observational RWE study of subjects with late-stage biliary tract cancer receiving either best supportive care (BSC) or anticancer therapy (ATC).	Retrospective observational RWE study of subjects with unresectable hepatocellular cancer treated with first-line (1L) systemic therapies.
Aim	To explore whether subjects receiving BSC may derive benefit from ATC by comparing real-world overall survival (rwOS) between the treatment groups.	To compare rwOS across 1L systemic therapies.
Target estimand	**Average Treatment Effect on the Treated (ATT)** – average causal effect in treated population.	**Average Treatment Effect (ATE)** – average causal treatment effect if everyone in target population were treated vs. not treated i.e. causal effect in entire population.
Propensity score-based approach	**Propensity score matching**: BSC subjects matched 1:1 to ATC subjects using nearest neighbour matching without replacement, with random selection within PS intervals based on 20th percentiles.	**Inverse probability of treatment weighting**: stabilised weights derived[4]. Truncation of extreme weights at 1st and 99th percentiles applied.
Outcome analysis	**Standard analysis** of rwOS using Kaplan-Meier methods and Cox proportional hazards model – **matched cohort only**.	**Weighted analysis** of rwOS using Kaplan-Meier methods and Cox proportional hazards model – **entire sample**.

DISCUSSION

PSM and IPTW are well-established methods for estimating causal effects in observational studies, by balancing observed baseline covariates between treatment groups, thereby minimising bias from confounders.

In case study 1, PSM was selected where the aim was to match BSC subjects on those treated with ATC. This approach enabled us to explore whether the subpopulation of BSC subjects, who share similar baseline characteristics to the ATC group, may benefit from treatment with ATC by comparing rwOS. PSM can often achieve superior covariate balance and is often considered straightforward to implement and interpret. However, unmatched subjects are excluded from the analysis, which can reduce statistical power and precision. As a result, findings are only generalisable to the matched cohort, rather than the entire sample[5].

In case study 2, IPTW was selected where the aim was to compare two 1L systemic therapies in the overall sample. With this approach, the full sample is retained and estimates are generalisable to the entire population. However, treatment effect estimates can be highly sensitive to extreme weights and generally require larger samples to produce more reliable estimates[3].

Key factors to be considered when selecting a propensity-based method include the target estimand, available sample size, the distribution of covariates, and desired level of generalisability.

Contact

Lucy Clark, Phastar

Email: lucy.clark@phastar.com

Website: phastar.com

Presented at the PSI 2026 Conference, Belfast

References

1 US Food and Drug Administration. Framework for FDA's Real World Evidence Program. December 2018
2 Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011 May;46(3):399-424.
3 Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34(28):3661-3679. doi:10.1002/sim.6607
4 Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. Value Health. 2010 Mar-Apr;13(2):273-277.
5 Austin PC. A comparison of the operating characteristics of different methods for propensity score matching. Stat Med. 2014;33(7):1057-1069.

Multiple imputations of range-restricted data – to ignore the range or not to ignore, that is the question

Beata Pawlikowska[1], Kacper Kowalczyk[1]

Biometrics, Late-stage Development, Respiratory and Immunology (RI), BioPharmaceuticals R&D, AstraZeneca, Warsaw, Poland[1]

P058

Abstract

Objective

To investigate the impact of multiple imputations of continuous variables that can obtain only values of a predefined range on the outcome data distribution and data inference as well as provide guidance what can be expected when handling the data differently e.g. ignoring the range or not.

Methods

Multiple imputations as per their assumptions should supplement missing values with data that fits into the original distribution. When looking into clinical trial data, especially questionnaires that are analysed based on an aggregated summary score that should fall into a predefined range, we might encounter potential problems. Imputations usually utilize normal distribution which means that our imputed values may fall outside of the said range. Our investigation checks the impact that modifications to the imputed data may have on the distribution of the imputed dataset as compared to original data, as well as the impact on the ANCOVA model estimates. The modifications include, but are not limited to, truncating the imputed values to be within the predefined range and resampling each imputed value until it falls within the said range. In the simulations we utilize dummy datasets created using Artificial Intelligence to avoid any potential privacy risks or data security violations

Introduction

Dataset

Dataset utilized in the analysis has been generated using AI. It contains questionnaire results collected at 6 consecutive visits for 100 subjects that were randomized equally into two treatment arms. The questionnaire results summary is based on a total score that ranges between 0 and 3, where the lower the value the better control over the disease. Additional parameters include age, sex and a factor with 2 levels. A 100 studies were simulated.

Analysis

The analysis was performed using SAS (PROC MI). All visualizations presented in this poster have been generated with the help of AI using R. In the analysis we focused on the impact on the distribution of the outcome variable after imputation and potential impact on the change from baseline estimate at 3rd and 6th visits.

Methods

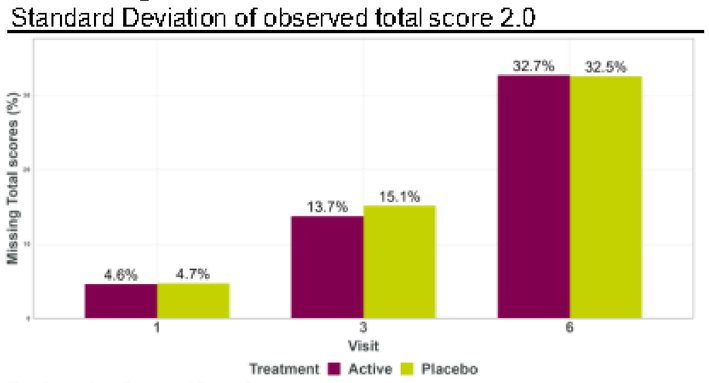
General assumptions

For the purpose of the analysis, it was assumed that:

- all missing data is either Missing Completely At Random (MCAR) with study withdrawal, both missing data types are handled the same way,
- approx. 20% of data is missing throughout the study with probability of missingness increasing with time,
- outcome variable ranges from 0 to 3 and follows a distribution close to normal (truncated normal),
- to investigate two extreme cases - little to no treatment effect was assumed at V3 and a considerable one at V6,
- two scenarios were tested – depending on the variability of the observed total score, i.e. SD=1 and SD=2 before truncation.

Figure 1 Proportion of missing data by treatment over time averaged across studies

Standard Deviation of observed total score 2.0



Selected methods

1. Ordinary MI (linear imputation) – values are imputed using monotone regression
2. Ordinary MI with data truncation i.e. values falling outside of range being replaced by lower/upper range limits
3. Full Conditional Specification (FCS) – values are imputed using regression method for the arbitrary missing data pattern
4. FCS with data truncation

All selected methods were compared with complete and observed cases analyses to assess the impact on the distribution of the outcome variable and its effect on the inference. Each of the methods utilized 100 imputations. For methods 1&2 the intermittent missing data is imputed using MCMC to obtain monotone data missing pattern as the first step.

In the next step for each method and each imputed dataset an ANCOVA model is utilized where change from baseline in total score is the dependent variable and 3 covariates: baseline total score, treatment and one additional factor. For each method model results are later combined following Rubin's rule (MIANALYZE).

Results

Impact on the distribution

Impact on the distribution of the imputed total score was assessed visually based on the density plots. For or each imputation method the density of mean total score for each subject by visit across all imputed datasets was plotted.

Figure 2 Distribution of the total score in one study

Standard Deviation of observed total score 1.0

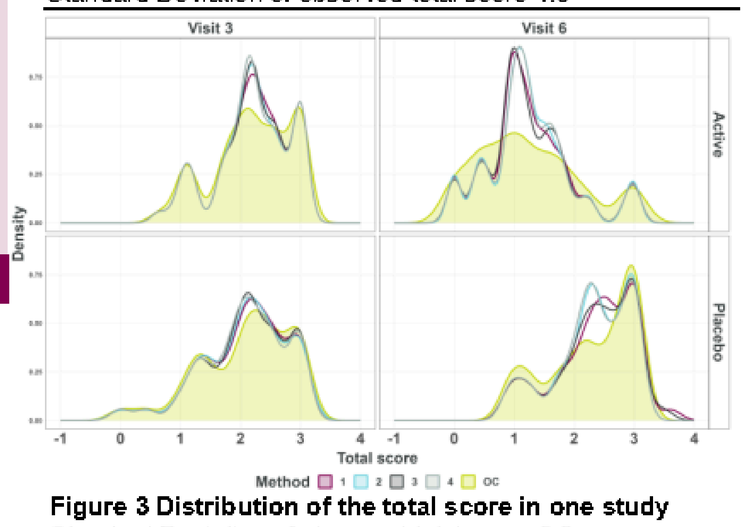
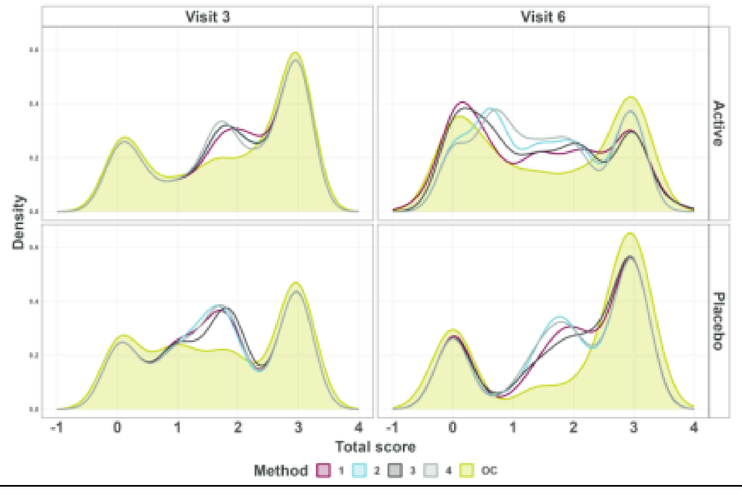


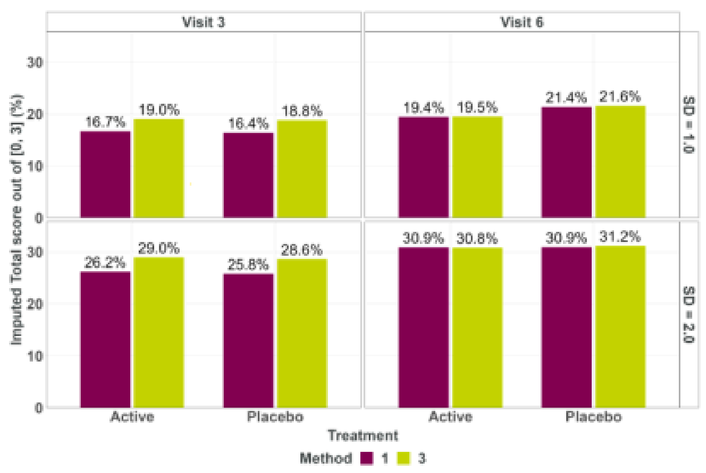
Figure 3 Distribution of the total score in one study

Standard Deviation of observed total score 2.0



Based on Figure 1 and Figure 2 we can observe that there is almost no difference between the distributions of the mean values imputed datasets for different methods. However, each time the distribution deviates from the original one (i.e., observed cases). What is more, the findings for both scenarios are aligned, suggesting that the impact on distribution is similar, regardless the variability in the original observed data and amount of total score values imputed out of range.

Figure 4 Proportion of imputed Total score values out of [0,3] range averaged across studies



Impact on inference

The impact was assessed based on the difference in change from baseline in total score between active and placebo treatments obtained from ANCOVA model, averaged across 100 studies..

Table 3 Impact of different imputation methods on treatment difference, ANCOVA

		SD=1.0		SD=2.0	
Visit	Method	LSMean Difference (SE)	P-value	LSMean Difference (SE)	P-value
3	1	-0.094 (0.183)	0.4493	-0.088 (0.253)	0.4211
	2	**-0.095 (0.174)**	**0.4376**	**-0.084 (0.236)**	**0.4137**
	3	-0.096 (0.188)	0.4528	-0.091 (0.259)	0.4248
	4	**-0.096 (0.177)**	**0.4392**	**-0.085 (0.239)**	**0.4153**
	OC	-0.101 (0.182)	0.4352	-0.093 (0.252)	0.4347
	CC	-0.102 (0.166)	0.4211	-0.097 (0.233)	0.4595
6	1	-0.972 (0.231)	0.0038	-0.648 (0.317)	0.1324
	2	**-0.918 (0.202)**	**0.0026**	**-0.588 (0.261)**	**0.1082**
	3	-0.975 (0.232)	0.0042	-0.640 (0.317)	0.1335
	4	**-0.921 (0.203)**	**0.0029**	**-0.583 (0.261)**	**0.1102**
	OC	-0.974 (0.209)	0.0026	-0.649 (0.291)	0.1004
	CC	-0.985 (0.169)	0.0001	-0.642 (0.236)	0.0593

OC Observed Cases; CC Complete Cases; SE Standard Error; SD Standard Deviation.

Conclusions

- In our simulations on average the truncation underestimates the treatment effect versus no truncation.
- The truncation always reduces standard error.
- As the variability of the simulated data increases, the differences between approaches are more pronounced.

References

1. Rodwell L, Lee KJ, Romaniuk H, Carlin JB. Comparison of methods for imputing limited-range variables: a simulation study. BMC Med Res Methodol. 2014 Apr 26;14:57. doi: 10.1186/1471-2288-14-57.
2. The MI Procedure, SAS Institute Inc. 2015. SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc
3. Imputation of categorical variables with PROC MI. Allison PD, Proceedings of the Thirtieth Annual SUGI Conference (Paper 113-30). SAS Institute Inc.
4. A Guide on Data Analysis, Chapter 13 https://mike-data-analysis.share.connect.posit.cloud/index.html

Supported by AstraZeneca

Presented at the PSI 2026 Conference, Belfast, 15 June 2026

Disclosures

BP and KK are employees of AstraZeneca and may own stock or stock options or interests in the company.

Funding

This poster was funded by AstraZeneca.

P059

Predicting the Unpredictable: How Event Prediction Succeeds, and Fails

Bastiaan Jansen, SGS Belgium NV

Introduction

Event prediction in trials with time-to-event endpoints is not only operationally important; for example, to plan interim and final analyses, but can also be a key design element. When formally incorporated into the study design, it enables optimal statistical efficiency by ensuring the targeted information fraction, and therefore the desired power, is achieved at each analysis point. Since traditional approaches rely heavily on assumptions and blinded follow-up data, a more direct strategy is proposed: combining Independent Data Monitoring Committee (IDMC) oversight with unblinded event projection, and evaluating where this approach succeeds or falls short.

Material and Methods

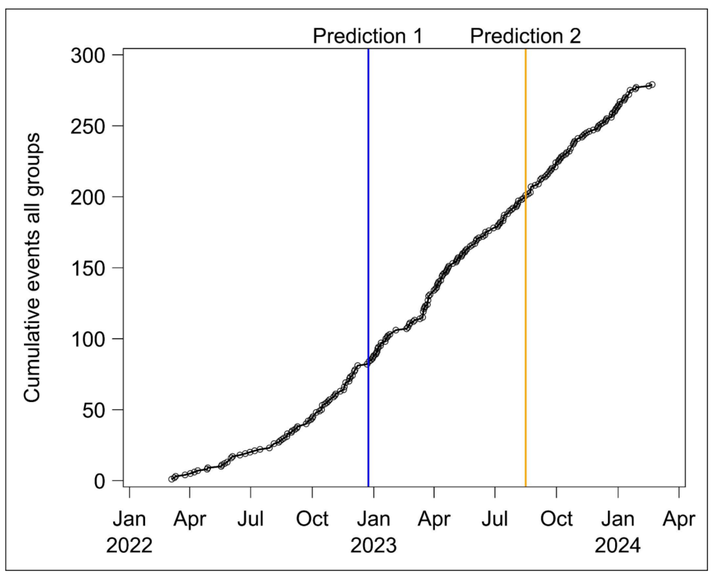
We look retrospectively at two case studies with a similar 3 arm parallel design (2 active versus control) where we made regular forecasts of target number of events (information fraction) for separate combinations. We used simulations using R eventPred package and looked at several time to event distribution models (Cox, Weibull, Log-normal, Log-logistic, Spline) both to have an idea of robustness as well as to assess best model fits. This way we tried to "forecast" the best time to plan the analysis of the study. Reported confidence intervals are simulation based percentile intervals.

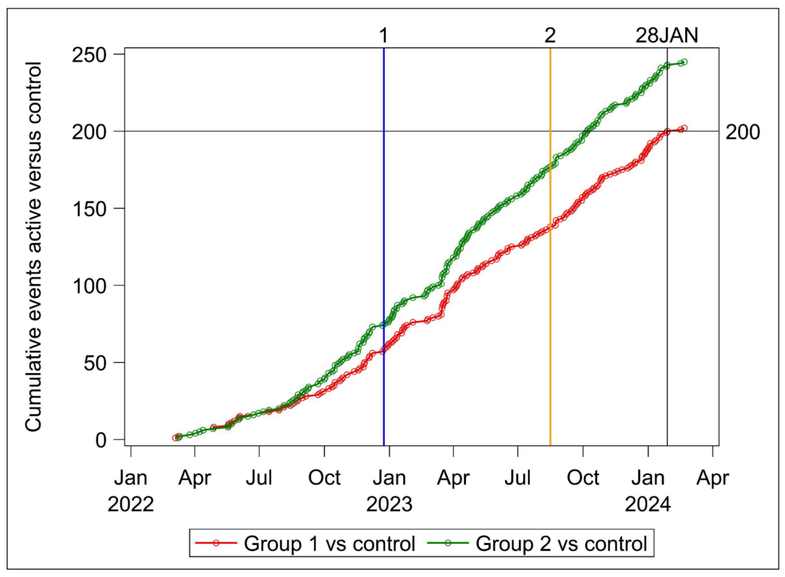
Note that data (both dates and cut-off points) were scrambled so results can't be linked to a specific trial

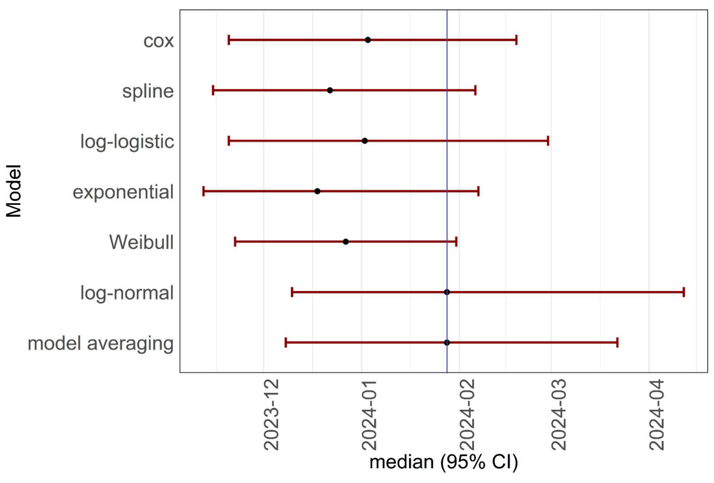
Discusssion Case Study 1

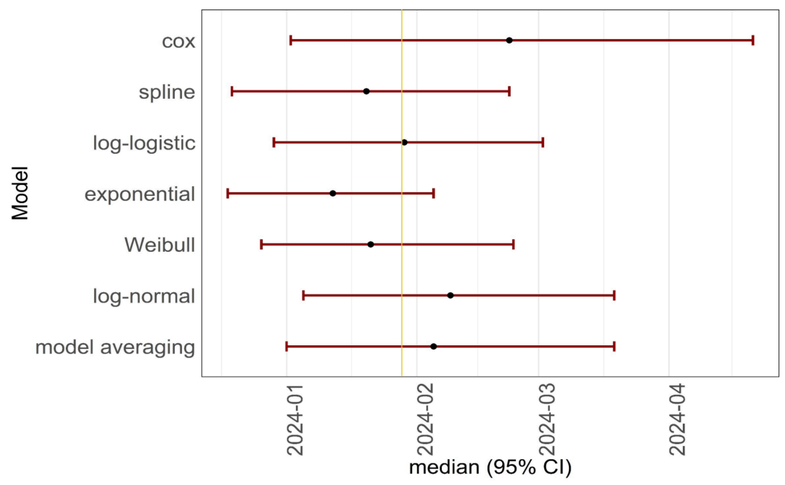
Both predictions yielded acceptable results, with the actual clinical cut-off falling within the confidence intervals of all models. As expected, prediction 2, performed closer to the target, seemed to show slightly improved precision. However, it became evident that a key assumption made during protocol development; that both active arms would perform similarly; did not hold, as one arm demonstrated a significantly different effect. This discrepancy also led to differences between the treatment-naïve and unblinded estimates in this specific case, ultimately enabling a more accurate determination of the information fraction at interim for this trial.

Case Study 1: Similar performance expected 200 events needed

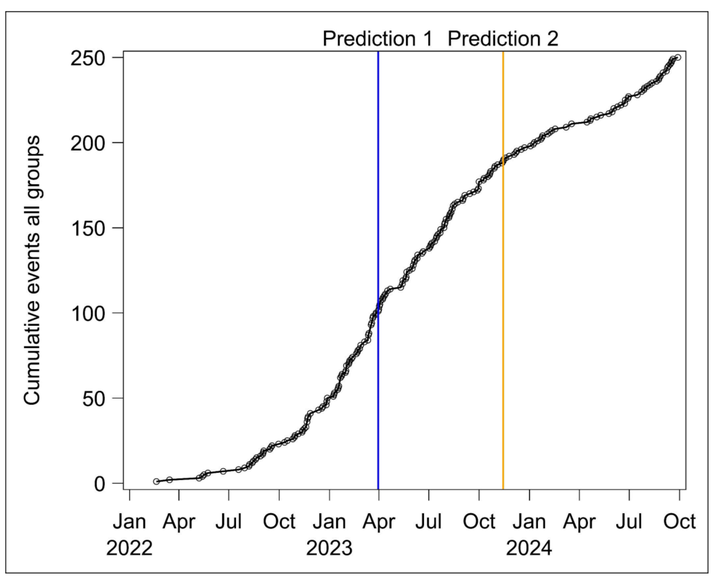


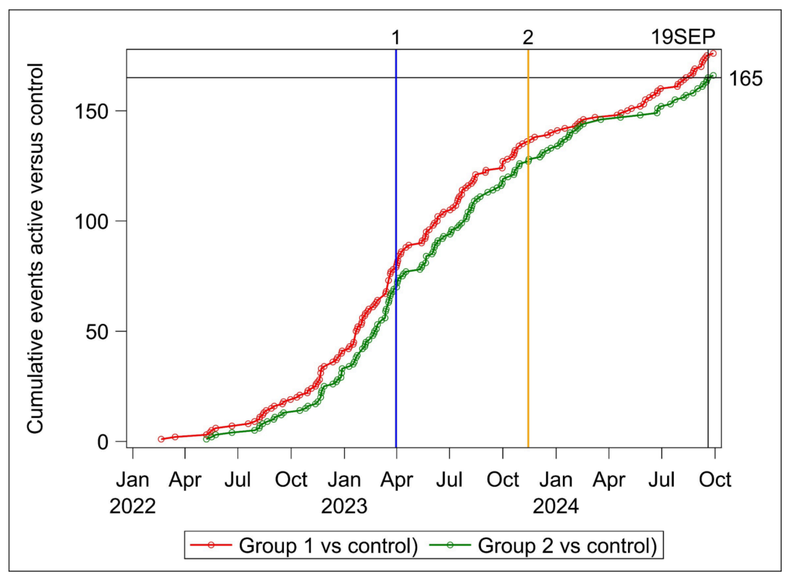


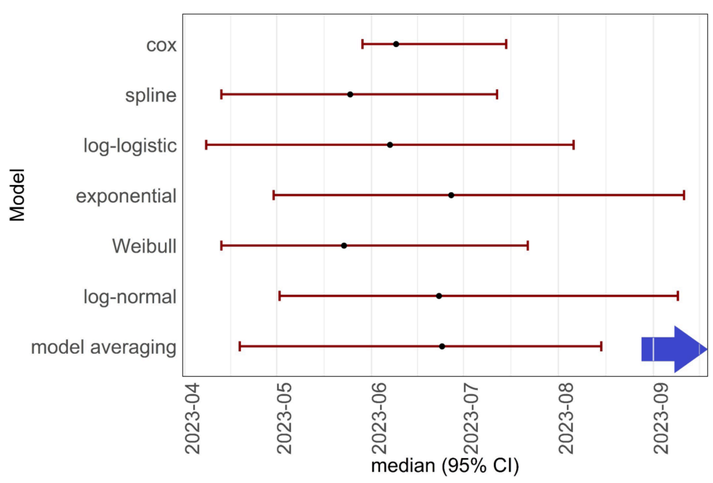


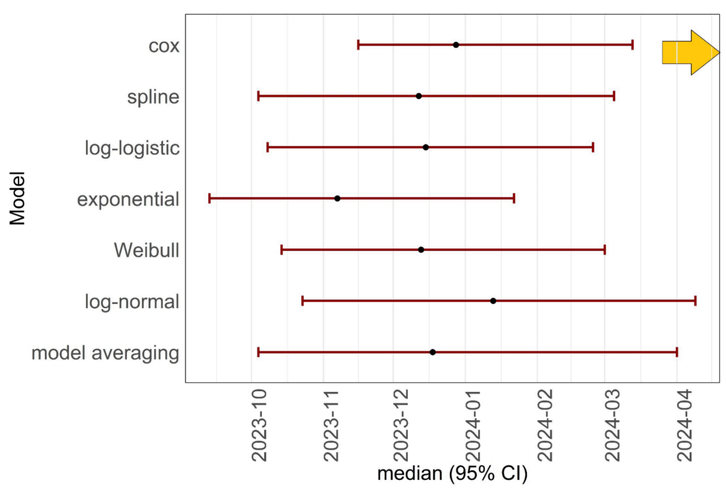


Case Study 2: similar performance expected 165 events needed









Discussion Case Study 2

Both predictions failed to pinpoint the correct clinical cut-off. Looking at the event curves, this is unsurprising but a stark reminder that event rates can fluctuate unexpectedly. As we followed up closely, we saw effects from external factors like timing (holiday period) and data cleaning efforts but were not fully able to explain the curve. There is also important information to be gathered from the shift of models e.g. While the exponential model has the 'latest' point estimate in prediction 1 it has the 'earliest' one in prediction 2, this could indeed suggest a changing hazard rate. Overall fit of models (not reported here) was also much poorer in case study 2 (High AIC BIC)

General Conclusions

Event prediction is a powerful technique that can help guide the timing of key analysis milestones. However, it is important to remain mindful of study dynamics, as changes in event rates may be driven by both internal and external factors. We therefore recommend avoiding a purely mechanical application of this approach. Instead, event predictions should be updated regularly, and model fit should be carefully assessed and supported with plain data mining. Large deviations or poor model fit may serve as early warning signals of potential issues in later analyses. In double-blind trials, the presence of an unblinded Statistical Support Group (e.g., supporting an IDMC) can further enhance precision through unblinded event prediction as original trial assumptions may not hold. Nevertheless, strict caution is required to ensure that sensitive information is not inadvertently shared with the blinded team.

sgs.com/linkedinpharma sgs.com/pharma clinicalresearch@sgs.com

P060

APT-SAP PROJECT

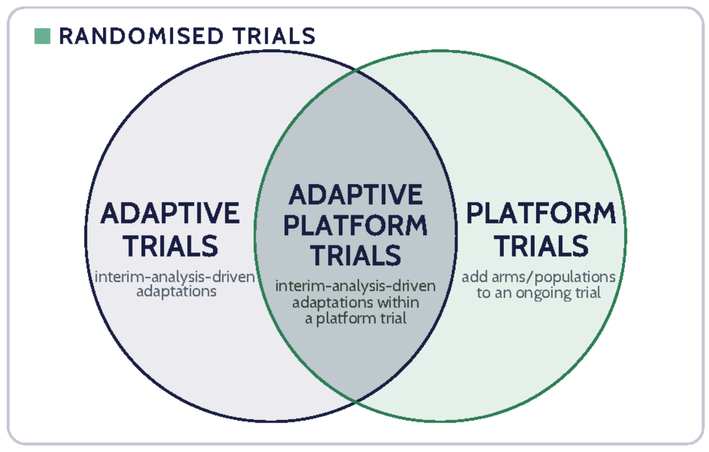
APT-SAP: Developing consensus-driven statistical analysis plan guidance for randomised trials using adaptive and platform designs.

What should the Statistical Analysis Plan for an adaptive or platform trial contain?

Xiaoran Lai[1], Isabelle Wilson[2], Xuewen Yu[1], Susan Todd[3], Sofía Villar[4], Thomas Jaki[4], James Wason[5], Christopher Weir[6], Derek Stewart[7], Roger Wilson, Franz Koenig[8], Israel Silva Maia[9], Jasmine Allender[2], Christina Yap[1], Munya Dimairo[2]

[1]Institute of Cancer Research · [2]The University of Sheffield · [3]University of Reading · [4]University of Cambridge · [5]Newcastle University · [6]The University of Edinburgh · [7]Cancer Research Advocates Forum UK · [8]Medical University of Vienna · [9]HCor Research Institute

1 The setting



49.6–84% of platform trials use **adaptive features**
Griessbach 2024 · Meyer 2020

One set of SAP guidance – for all three.

2 The gap

0 **SAP guidance written for adaptive or platform trials**

APTs risk –

- **Biased reporting** of interim and final results.
- Pre-planned analyses that are **hard to reproduce**.
- **Delayed or weakened** interim decisions.
- **Misleading conclusions** on safety and efficacy.
- **Misuse** of the trial in later evidence synthesis.
- **Public distrust** in research.

Existing SAP guidance (Gamble 2017; Homer 2022) does not adequately address the need for randomised adaptive or platform trials.

3 The project

APT-SAP builds a **consensus-driven minimum set of SAP items** – for both **interim and final analyses**.

Checklist
essential items for every SAP

Explanation & Elaboration
the rationale behind each item

Annotated template
a worked SAP, with real trial examples

ROBUST APT-SAPs DELIVER

- **Transparent, reproducible** analyses
- **Credible, accurate** results
- **Less research waste**; resources used well
- **Better patient outcomes** & sound policy decisions

– for patients, the public, researchers, regulators & funders

4 The approach

WE ARE HERE

1 → 2 → 3 → 4 → 5

Draft checklist
guidance, literature review, expert opinion & current practice

Delphi surveys
eDelphi · items rated 1–9 or 'unsure'
Round 1 ✓ · Round 2 underway

Consensus meeting
~30–40 global cross-sector experts
≥70% agreement to include

Pilot-testing
≥3 ongoing or planned adaptive & platform trials

Final guidance
checklist · E&E document
annotated SAP template

Shaped by a **global, cross-sector, multidisciplinary** group: statisticians, methodologists, clinicians, regulators (e.g. FDA · EMA · MHRA · PMDA), funders, journal editors and patient & public representatives across the public and private sectors.

5 Round 1 of the Delphi

42 stakeholders completed Round 1 (data cut-off: 27 May 2026)

Based in



Sector



Primary role



Median 12 years in clinical-trials research

Median 5 years in APT-specific research

74% have contributed to a SAP for an adaptive or platform trial

72 candidate SAP items went into Round 1 – **36 new · 36 extended**. Examples (e.g.) with **modifications** to existing SAP guidance highlighted:

EXTENDED *extends Gamble et al. 2017*

item 1.1.1 Descriptive title that matches the protocol, with SAP either as a forerunner or subtitle, and trial acronym (if applicable). **The title should include relevant adaptive or platform trial terminology (e.g., "adaptive", "adaptive platform", "platform") to signal the nature of the design.**

signals trial complexity; aids scope interpretation and registry searches

item 3.7.1 Triggers for final analyses, **specifying the application level for each trigger (e.g., applies to all or specific intervention comparisons, or specific sub-studies within the platform).**

platforms may trigger final analyses at different levels – per comparison, sub-study, or platform-wide

NEW

item 6.1.2 **Specification of outcome(s) for informing each trial adaptation at interim analyses and explanation if different from primary outcome(s), including timing of assessments.**

interim adaptation outcome may differ from primary – justify validity and implications

item 3.1.2 **Brief description of interventions, including comparators, specifying the level at which they operate (e.g., platform, substudy, or population).**

shared comparator(s). structure of comparisons

6 Shaped with help from patients & the public

11 patient & public contributors · online workshop, 8 May 2026 · co-chaired by two PPI co-leads

Familiarity with adaptive / platform trials



What they bring



Where they joined from



QUESTIONS THEY RAISED

> *"Who can change the **rule book** – who has the **final authority**?"*

> *"How do you expect the Adaptive method to vary between **Academic** and **Pharma-funded** trials?"*

> *"You can't have an adaptive trial without an **adaptive statistical analysis**."*

> *"Can the SAP reflect **different levels of risk** for a **particular patient group**?"*

> *"There's a big difference in **communication** between stopping for a **safety event** vs. stopping for **lack of efficacy** – perhaps there is a different level of information required?"*

Find out more
sheffield.ac.uk/ctru/current-trials/apt-sap

APT-INFORM – a linked project on sharing interim trial information – is in development alongside APT-SAP.

Contact apt-sap@sheffield.ac.uk

Funding This project is jointly funded by the MRC and NIHR (award UKRI615/APP42050) under UKRI. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

References 1. Gamble C, et al. JAMA 2017;318:2337. 2. Homer V, et al. BMJ 2022;376:e068177. 3. Meyer EL, et al. Clin Ther 2020;42:1330. 4. Griessbach A, et al. JAMA Netw Open 2024;7:e243109.

UKRI Medical Research Council · Funded by NIHR | National Institute for Health and Care Research

P061

Futility Analysis in the Two-Trial Paradigm

Saswati Saha, Efi Gymnopoulou | argenx, Ghent, Belgium

KEY TAKEAWAYS

- Futility thresholds should be set cautiously and be interpreted alongside key protocol elements, including design assumptions, timing of interim analyses, and the available data
- Potential protocol updates will impact the population and subsequently the futility evaluation
- Differences in recruitment rates across studies can impact interim decisions and the homogeneity of the pooled population at the time of futility assessment
- Recruitment status at the time of futility evaluation influences both decision risk and interpretability of interim findings
- Front-loaded recruitment dynamics complicate interim timing: by the time sufficient patients reach the primary endpoint, enrolment is largely complete
- The IDMC decision should be based on the totality of the data, not just on the primary endpoint futility rule

BACKGROUND | MOTIVATION

Why Interim Analysis?

- Clinical trials commit substantial resources (time, funding, and patient participation) based on assumptions in the absence of data.
- Interim analysis allows mid-trial re-evaluation of assumptions, and informs go/no-go decisions.

What is Futility Analysis?

- Futility stopping enables early termination when an intervention shows no meaningful benefit.
- Offers protection to participants from continued exposure to an ineffective treatment and redirects resources toward promising programs.

The FDA's 1998 2-trial paradigm

- Requires independent replication of clinical trials as the gold standard, which adds critical complexity.

How do we design robust futility rules when 2 studies must succeed together?

Straight to Phase 3

- Futility analysis in the 2-trial paradigm is challenging in the rare-disease domain, where interim analyses are included as a derisking measure, given the accelerated development timeline and the need to stay competitive.
- Limited internal PoCs adds uncertainty around a direct transition to Phase 3.

OPPORTUNITIES

Stronger Statistical Basis
Pooled + per-study evaluations reduce incorrect conclusions

Earlier Decision Making
Go/no-go possible sooner than with a single trial

CHALLENGES

Operational Complexity
Recruitment pace and population may differ between "identical studies"

Unstable Early Estimates
Weak interim signals do not preclude eventual trial success

Heterogeneity at Interim
Site-level divergence complicates pooled estimates

2 Identical Studies in Futility Analysis

CASE STUDIES

1. Aducanumab (Biogen Early AD Trial)

Background

EMERGE and ENGAGE: Two identical, Phase 3, randomized, double-blind, placebo-controlled trials. Interim futility analysis triggered after ~50% of participants reached the primary endpoint.

Futility Evaluation

- Conditional Power (CP) computed separately for each study, assuming future effect would equal the pooled interim estimate from both studies
- Futility was declared if both CPs were <20%.

Analysis	EMERGE CP	ENGAGE CP	Decision
Original (pooled estimate)	12%	0%	Both terminated ✗
Re-evaluation (per-study)	59% ✓	0%	Drug ultimately APPROVED

FDA: Early termination did not compromise result interpretation. Futility declaration would have been premature.

Key assumption: *The pooled approach outperforms single-trial CP when heterogeneity is small to moderate*

⚠ **Key lesson:** Pooled CP under heterogeneous effects can mask a positive signal in 1 study. Per-study evaluation is essential when results diverge between the 2 trials.

2. Hypothetical Trial inspired by argenx Phase 3 Trial

Study Design

- 2 identical Phase 3 studies per FDA requirement
- External PoC available
- DBTP: 24 weeks; efgartigimod PH20 SC versus placebo SC
- Primary endpoint: percentage of responders at Week 24 of DBTP
- Primary objective: efficacy and safety of efgartigimod PH20 SC
- Sample size: 150 per trial
  - Based on 90% power and 1-sided alpha at 2.5%
  - Assumed effect: 10% (placebo arm) and 35% (treatment arm)

Futility Evaluation

- Includes data from both studies.
- CP evaluated per study under the protocol-assumed effect; pooled study evaluated with lower alpha.
- Occurs when ~30% of participants from both studies reach Week 24 of DBTP, with ≥20% of participants in each study.

All study details have been modified to preserve the confidentiality of the trial.

RESULTS

FUTILITY DECISION FRAMEWORK (IDMC Guidance)

Decision Thresholds

CP ≥65% (single study) → continue
CP ≥70% (pooled data) → continue

- The IDMC recommendation must be based on **the totality of data**, not just the CP rule
- The recommendation per study can deviate if deemed necessary. The futility is nonbinding.

Trial 1	Trial 2	Pooled	Suggested Recommendation
Continue	Continue	Continue	Continue both studies
Stop	Stop	Stop	Stop both studies
Continue	Continue	Stop	Focus on totality of data; evaluate secondary endpoints for consistency on signal
Continue	Stop	–	
Stop	Continue		

SIMULATION SCENARIOS

2 Recruitment Scenarios

- Scenario A—balanced recruitment:
  - 30% IF per study
- Scenario B—extreme recruitment:
  - Study 1: 20% IF; study 2: 40% IF
- Assess stopping probabilities across different CP cutoffs in:
  - Both studies under H_0 (Δ=0, no difference)
  - Both studies under the H_1 (25% difference)

Decision Objective

Identify CP cutoff such that it

✅ maximizes stopping when Δ=0
❌ minimizes stopping when Δ=25%

with limited impact on overall study power

OPERATING CHARACTERISTICS

- Team recommendation: only stop the trial for futility in the event of no clear benefit (Δ≤0%) at 30% IF

TABLE 1 Probability of Stopping Across Different CP Cutoffs for Scenario A and Scenario B in Individual Studies

		Probability of Stopping in Scenario A [Scenario B Study 1/Scenario B Study 2], %	
Observed Δ at Futility, %	CP Cutoff, %	Correctly |Δ=0%|	Incorrectly |Δ=20%|
0	65	53.5 [24.9/65.2]	2.7 [0.4/4.6]
2.5	70	53.5 |35.9/72.8|	3.1 |1.0/5.4|
5	75	64.9 |41.0/76.4|	5.8 |3.4/7.6|
10	80	74.8 |54.0/82.9|	10.3 |7.1/11.4|

- Overall power loss: ~1% at the selected CP threshold
- Similar signals observed in pooled trials with a higher probability of correct stopping under Δ=0 for a CP threshold of 70%

TABLE 2 Probability of Being in Each Situation With (Suggested Rule: CP ≥65% for Per Study (Pass/Fail): CP ≥70% for Pooled (Pass/Fail)

Study 1	Study 2	Pooled	Recommendation	Both Studies Under H_0, %	Both Studies Under H_1, %
Pass	Pass	Pass	Continue	15 |15|	95 |95|
Pass	Pass	Fail	*Look at other endpoints*	6 |10|	~0 |~1|
Fail	Fail	Fail	Stop	30 |18|	–
Pass	Fail	–	*Look at other endpoints*	25 |48|	~2 |~3|
Fail	Pass	–	*Look at other endpoints*	24 |9|	~3 |~1|

Values in parentheses reflect extreme recruitment scenario (scenario B).

⚠ **Key lesson:**

- Under balanced recruitment and H_1, the rule correctly recommends continue ~95% of the time
- Under H_0, opposite results in the 2 studies arise ~49% of the time
- Results highlight the importance of totality of evidence

FUTILITY ANALYSIS: IS IT WORTHWHILE?

Assurance

- Probability of passing both interim AND final analysis, averaged over the prior distribution of the true effect.

$$\int P(\Delta_{int} > c_{int}, \Delta_{final} > c_{final} | \Delta) \cdot \pi(\Delta) d\Delta$$

Conditional Assurance

- Given the interim was positive (futility passed), what is the probability of a successful final analysis? This is the PoS after the futility hurdle.

$$\frac{Assurance}{P(\Delta_{int} > c_{int})}$$

Mixed prior used: $\pi(\Delta) = \omega \cdot \delta_a(\Delta) + (1-\omega) \cdot \delta_0(\Delta)$, *where* ω *is calibrated so overall assurance equals the industry-wide Phase 3 success rate of 25% (*ω *= 0.86) where* $\delta_0(\Delta)$ *is the null and* $\delta_a(\Delta)$ *alternative distribution for* Δ.

FIGURE 1 Distribution of the Mixture Prior Under the Bayesian Assurance Framework, Assuming an Industry-Wide Phase 3 Success Rate of 25%

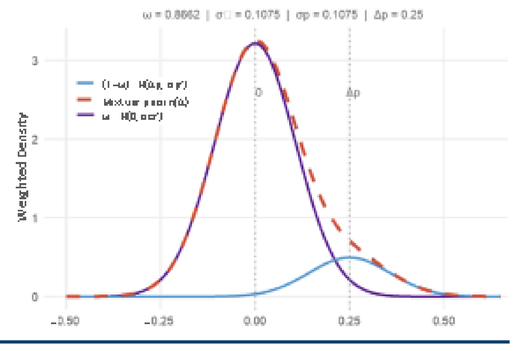
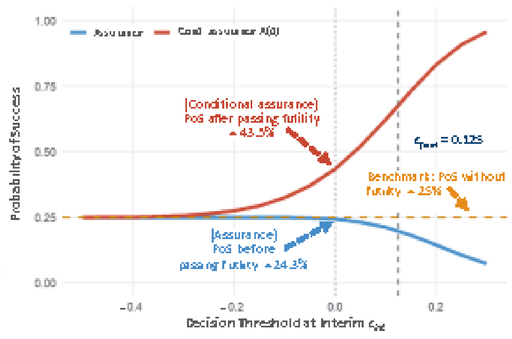


FIGURE 2 How Assurance and Conditional Assurance Change as the Interim Threshold c_{int} Varies From –0.5 to 0.3



- c_{final} represents the threshold for declaring statistical significance at the primary analysis

⚠ **Key lesson:** Adding futility analysis with threshold Δ=0 increases PoS conditional on passing futility by ~19% without any loss on the overall assurance—a strong argument for including futility analysis

25%	43.5%	24.3%
industry baseline PoS without futility (benchmark)	19% increase conditional assurance after passing futility	preserved overall assurance (no loss vs benchmark)

Assurance Framework Advantages

- Unlike frequentist CP, the assurance framework accounts for uncertainty in the true treatment effect through the prior, providing a more realistic decision-support tool

Assurance was implemented as per Dragalin V. *Stat Med*. 2026;45:e70330.

PRESENTED AT THE STATISTICIANS IN THE PHARMACEUTICAL INDUSTRY (PSI) CONFERENCE; JUNE 15–17, 2026; BELFAST, NORTHERN IRELAND

ABBREVIATIONS

Δ_{final} treatment effect at primary analysis; Δ_{int} treatment effect at interim; AD, Alzheimer disease; c_{final} threshold for passing primary analysis; c_{int} threshold for passing futility analysis; CP, conditional power; DBTP, double-blinded treatment period; FDA, Food and Drug Administration; H_0, null hypothesis; H_1, alternative hypothesis; IDMC, independent data monitoring committee; IF, information fraction; P, probability; PH20, recombinant human hyaluronidase; PoC, proof of concept; PoS, probability of success; SC, subcutaneous.

DISCLOSURES AND ACKNOWLEDGMENTS

The content of this poster reflects our personal views and interpretations and should not be taken as representing the official position of our company.

REFERENCES

1. US Food and Drug Administration. Guidance for industry: demonstrating substantial evidence of effectiveness with one adequate and well-controlled clinical investigation and confirmatory evidence. 2023. 2. Gallo P, Roychoudhury S. *Ther Innov Regul Sci*. 2023;57:515–20. 3. Haeberlein SB, et al. *J Prev Alzheimers Dis*. 2022;9:197–210. 4. Van Lanker K, et al. *Pharm Stat*. 2020;19:583–601. 5. Dragalin V. *Stat Med*. 2026;45:e70330.

RELATED POSTER CONTENT

P062

Estimating eGFR-Decline Event Times Under Intermittent Observation: Bias, Precision, and Implications for CKD Trial Design

P-062 · PSI 2026

Sasha Tatum [1] · Supervised by José Sánchez [1] and Robin Myte [1] | Supported by AstraZeneca

[1] CVRM Biometrics, Late CVRM, BioPharmaceuticals R&D, AstraZeneca

★ **Key message: Under 6-monthly visits, the standard approach misses over half of true events and overstates the Kaplan–Meier median survival time by up to 17.5 months. Per-subject OLS regression (R4a) recovers more events, halves event-time bias, and preserves valid inference.**

❶ Background & Aims

Phase 3 CKD trials use **time to sustained ≥50% eGFR decline from baseline** as a surrogate for kidney failure.

eGFR is measured only at **scheduled visits** (every 3–6 months). The standard method (R1) assigns the event at the first visit where eGFR falls below 50% of baseline, confirmed at an **unscheduled visit** ~1 month later.

Kidney function declines **continuously**, so the true crossing occurs between visits. This introduces *visit-timing bias* that grows with the visit interval.

Aim: Quantify visit-based bias and evaluate whether alternative rules reduce it while preserving valid inference.

❷ Five Event-Time Rules

Five rules for assigning a confirmed event time are compared, from the standard (R1) to regression-based interpolation approaches.

Rule	Definition
● **R1**	First scheduled visit with eGFR < 50% of baseline, confirmed at unscheduled visit ~1 month later *(current standard)*
● **R2**	Midpoint between bracketing visits (no confirmation)
● **R3**	Linear interpolation between bracketing visits
● **R4a**	Per-patient Ordinary Least Squares (OLS) regression: predicted crossing time from fitted line [1]
● **R4b**	Linear mixed model (LMM): crossing time from subject-level Best Linear Unbiased Predictors (BLUPs)

❸ The visit-timing bias problem

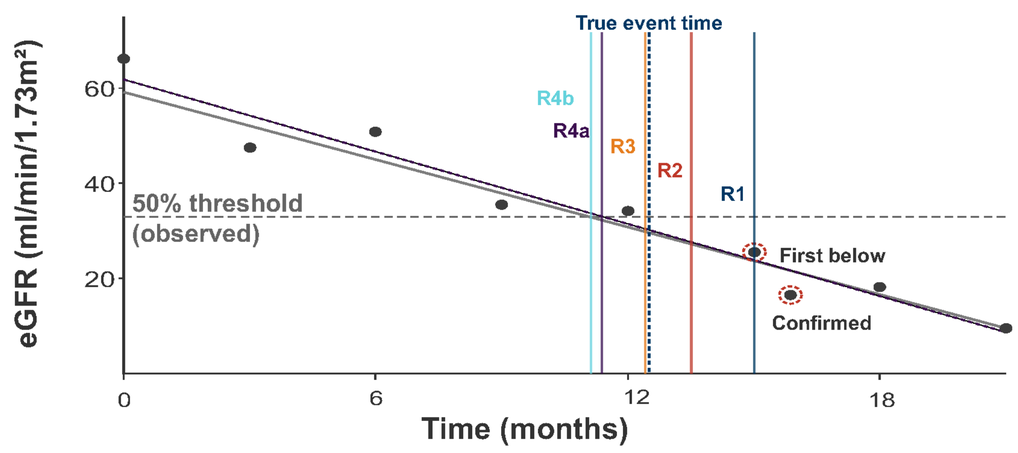


Figure 1. Example simulated trajectory (3-monthly visits). Solid grey: latent eGFR; dashed purple: R4a OLS fit. Vertical lines show each rule's assigned event time. R1 is recorded at the first below visit, only if the confirmation visit (one month later) is also below threshold.

❹ Simulation Design

N = 1,000 subjects, **1:1** allocation, **45-month** follow-up, **K = 10,000** iterations. All five rules applied to each trial (paired).

Latent eGFR trajectory (Simulation framework similar to Greene et al. [2]):

$$Y_i(t) = (\mu_\alpha + a_i) + (\mu_\beta + b_i)t + \varepsilon_{ij}$$

$$(a_i, b_i)^T \sim N(\mathbf{0}, \Sigma) \quad \cdot \quad \varepsilon_{ij} \sim N(0, k \cdot Y_i(t_{ij}))$$

μ_α, μ_β = population baseline & slope. a_i, b_i = subject deviations (bivariate normal, ρ). Power-of-Mean residual (k = 0.817): noise scales with eGFR.

Treatment effect: 25% slope reduction for declining subjects ($\beta_i < 0$); stable subjects unaffected.

Visit schedules: Monthly · 3-monthly · DAPA-CKD-like (0.5, 2, 4 months then 4-monthly) · 6-monthly.

Grid: 3,456 scenarios reduced to 288 via meta-model screen (event-time bias R^2 = 0.95; log-HR bias R^2 = 0.27).

❺ Results: Event-time bias & detection

Mean signed bias = rule event time − latent truth (months). Positive = event recorded late.

Schedule	R1	R2	R3	R4a	R4b
Monthly	+0.24	−1.00	−0.50	+0.04	+0.05
3-monthly	+1.13	−1.18	−0.72	+0.08	+0.10
DAPA-like	+1.26	−1.25	−0.80	+0.06	+0.08
6-monthly	+1.88	−1.38	−0.92	+0.16	+0.14

R1: late by construction. **R2/R3:** early; omitting confirmation allows noisy readings to trigger false events. **R4a/R4b:** near-zero bias (|bias| < 0.16 months) as regression smooths noise.

Kaplan–Meier median survival bias (R1 vs latent):

+1.94	+7.19	+7.99	+17.52
MONTHLY	3-MONTHLY	DAPA-LIKE	6-MONTHLY

All in months. R4a reduces KM median bias to +0.20, +0.50, +0.50, +3.03 under the same schedules.

Sensitivity = proportion of latent events recovered.

Schedule	R1	R2	R3	R4a	R4b
Monthly	0.82	0.95	0.95	0.96	0.94
3-monthly	0.61	0.80	0.80	0.91	0.88
DAPA-like	0.71	0.80	0.80	0.96	0.88
6-monthly	0.43	0.60	0.60	0.71	0.71

Mean 791 latent events per trial. Under 6-monthly visits, R1 misses 57% of true events; R4a recovers more events than any other rule.

❻ Inferential Properties (Cox)

Metric	R1	R2	R3	R4a	R4b
SE ratio*	1.00	1.00	1.00	1.00	**0.91–0.95**
Type I error	≈0.05	≈0.05	≈0.05	≈0.05	**0.054–0.080**
CI coverage	≈50-95%	≈55-95%	≈70-95%	≈95%	≈95%
Log-HR bias	Minimal†	Minimal†	Minimal†	Smallest bias†	Minimal†

* SE ratio = mean Cox SE / empirical SE; <1.00 = anti-conservative. † Delays cancel in ratio, model based R4a and R4b had lowest bias, schedule choice biggest driver of bias.

R4b: two-stage LMM→Cox does not propagate random-effects variance into the Cox SE.

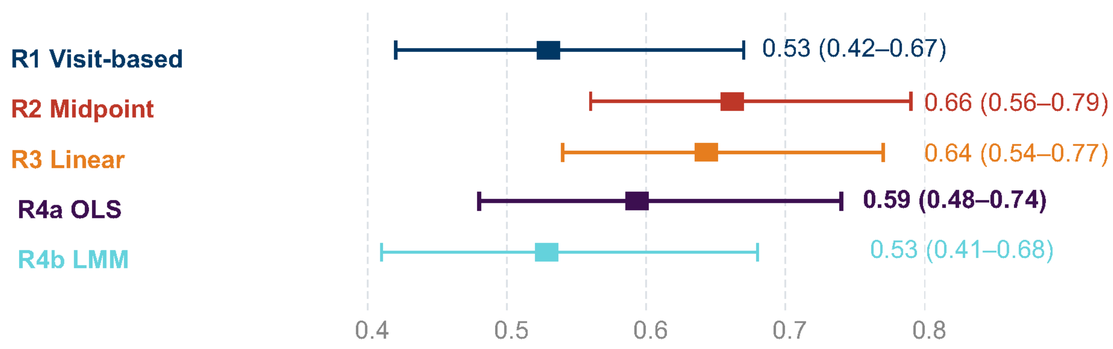
Coverage at 6-monthly: Collapses for all rules driven by log-HR bias growing large relative to SE.

❼ Application to DAPA-CKD (n = 4,088)

Five rules applied to DAPA-CKD patient-level data [3]. Cox stratified by T2D status × UACR, adjusted for baseline eGFR. All $p < 10^{-5}$.

Rule	Events	Rate*	HR (95% CI)
R1 Visit	311	4.85	0.53 (0.42–0.67)
R2 Midpoint	533	7.82	0.66 (0.56–0.79)
R3 Linear	505	7.47	0.64 (0.54–0.77)
R4a OLS	346	5.15	0.59 (0.48–0.74)
R4b LMM	268	4.14	0.53 (0.41–0.68)

* Events per 100 patient-years (placebo arm).



No rule outperformed R1 in DAPA-CKD. R2/R3 inflate event count in placebo arm by omitting confirmation, attenuating the HR. R4a (0.59) is more conservative than simulation predicts: dapagliflozin's acute haemodynamic dip biases the OLS slope. R4b (0.53) matches R1 but recovers fewest events. Qualitative finding consistent: dapagliflozin substantially reduces risk.

❽ Limitations & Discussion

- **Linear DGM only.** No acute haemodynamic dip modelled. DAPA-CKD shows this matters; a piecewise DGM is needed.
- **Non-informative dropout.** Sicker patients drop out faster in practice, differentially affecting low-sensitivity rules.
- **Fixed follow-up.** Real trials are event-driven; R1 is affected by visit-censoring mismatch, R4a is not.
- **50% decline only.** Full composite endpoint not modelled; extension to competing risks needed.
- **Point imputation.** Interval-censored Cox or AFT models could leverage bracketing-visit information directly.

Conclusions & Recommendations

1. **Visit frequency dominates rule choice.** Denser visits reduce bias more than any rule change. Front-loaded schedules particularly benefit R4a (efficiency 0.95 vs 0.57 for R1).
2. **Adopt R4a (OLS) in practice.** Halves event-time bias vs R1; recovers more events; maintains correct SEs and Type I error. Advantage greatest under sparse schedules.
3. **Real-data performance diverges from simulation.** In DAPA-CKD, R4a returned HR 0.59 vs R1's 0.53. The acute haemodynamic dip biases OLS slopes; a piecewise DGM is needed.
4. **Prespecify rule in the SAP.** HR ranged 0.53–0.67 in DAPA-CKD. Rule choice is part of the estimand. Sensitivity analyses with R4a/R4b can assess visit-based bias.

References

[1] Zee J et al. (2019) Using all longitudinal data to define time to specified percentages of estimated GFR decline: A simulation study. American Journal of Kidney Diseases, 73(1):82–89.

[2] Greene T et al. (2019) Performance of GFR slope as a surrogate end point for kidney disease progression in clinical trials: A statistical simulation. Journal of the American Society of Nephrology, 30(9):1756–1769.

[3] Heerspink HJL et al. (2020) Dapagliflozin in patients with chronic kidney disease. New England Journal of Medicine, 383(15):1436–1446.

Acknowledgements

Data: DAPA-CKD investigators

Supported by AstraZeneca

P063

Weighted Longitudinal Summaries of PRO Data Across Treatment Phases: A Simulation Study with Potential Relevance for HTA

Giasafaki, Peggy (*MSD, Brussels, Belgium*); **Del Fava, Emanuele*** (*MSD, Zurich, Switzerland*); Rabel, Matthias (*MSD Sharp & Dohme GmbH, Munich, Germany*); Mt-Isa, Shahrul (*MSD, Zurich, Switzerland*)
*Presenting and Corresponding Author

Introduction and Objective

- The Statistical Initiative in Oncology Patient-Reported Outcomes Quality of Life (SISAQOL) Consortium has identified the **overall PRO score over time** as a relevant research objective in cancer trials, but reached no consensus on recommended summary measures, identifying their evaluation as a priority for future work [1].
- In multiphase oncology trials, PRO data are typically analyzed by phase. **Longitudinal weighted summaries spanning all phases** may offer a principled approach to deriving a single interpretable summary measure, but the impact of the choice of weighting scheme on the estimation of the Overall treatment effect under realistic conditions of missing data and varying sample sizes is not well understood.
- **Objective**: To assess the impact of the choice of weighting scheme on the ability to reliably estimate a single summary measure of the treatment effect across the whole patient journey in multiphase oncology trials.

Material and Methods

Simulation Design

- Continuous PRO data simulated in a three-phase peri-operative oncology trial (neoadjuvant / adjuvant / post-adjuvant) under two clinically plausible true effect patterns: P1 Early Deterioration + Gradual Recovery; P2 Sustained Treatment Benefit
- Two scenarios: n = 200 × missingness = 30% / n = 400 × missingness = 40%; missingness driven by death, administrative censoring, and skipped visits.
- 100 datasets simulated per pattern and scenario.

Statistical Model

- Constrained longitudinal data analysis with unstructured covariance [2].

Estimands

- **Primary**: Weighted average treatment effect on PRO across all visits and phases, representing the whole patient journey.
- **Secondary**: Weighted average treatment effect within each phase, reported as a clinical decomposition of the overall experience.

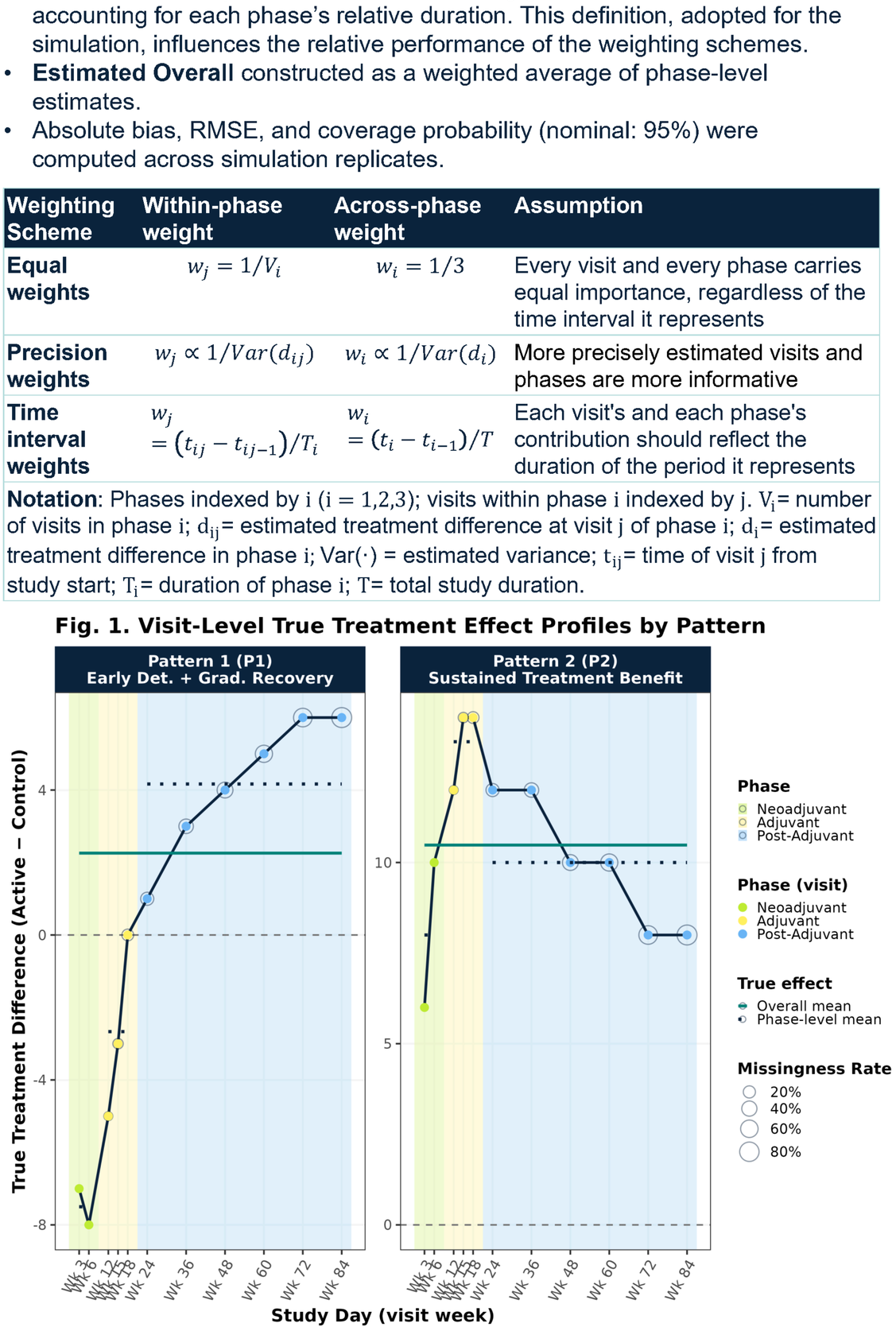
Performance Metrics

- **True Overall** computed at the visit level using global time interval weights, accounting for each phase's relative duration. This definition, adopted for the simulation, influences the relative performance of the weighting schemes.
- **Estimated Overall** constructed as a weighted average of phase-level estimates.
- Absolute bias, RMSE, and coverage probability (nominal: 95%) were computed across simulation replicates.

Weighting Scheme	Within-phase weight	Across-phase weight	Assumption
Equal weights	$w_j = 1/V_i$	$w_i = 1/3$	Every visit and every phase carries equal importance, regardless of the time interval it represents
Precision weights	$w_j \propto 1/Var(d_{ij})$	$w_i \propto 1/Var(d_i)$	More precisely estimated visits and phases are more informative
Time interval weights	$w_j = (t_{ij} - t_{ij-1})/T_i$	$w_i = (t_i - t_{i-1})/T$	Each visit's and each phase's contribution should reflect the duration of the period it represents

Notation: Phases indexed by i (i = 1,2,3); visits within phase i indexed by j. V_i= number of visits in phase i; d_{ij}= estimated treatment difference at visit j of phase i; d_i= estimated treatment difference in phase i; Var(·) = estimated variance; t_{ij}= time of visit j from study start; T_i= duration of phase i; T= total study duration.

Fig. 1. Visit-Level True Treatment Effect Profiles by Pattern



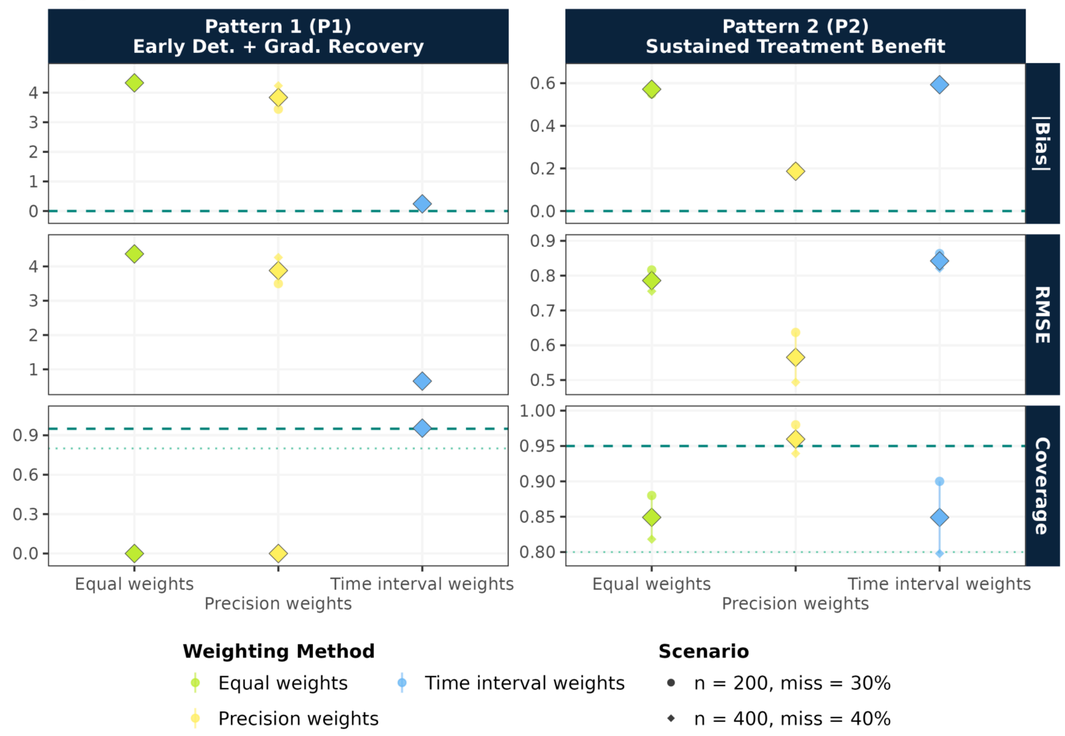
Results

- Treatment effect profiles differ by design between patterns: in P1 the effect favors the control arm in neoadj./adj. phases, then gradually shifts toward the active treatment post-adjuvant; in P2 the effect consistently favors the active treatment throughout, declining, but remaining positive post-adjuvant. (Fig. 1)
- Within phases, the weighting scheme choice is inconsequential, as equal and time interval weights perform equivalently, and precision weights differ only modestly. For the Overall estimand in P1, this changes critically: equal and precision weights fail structurally (zero coverage, large bias), while time interval weights maintain near-nominal coverage and low bias, consistent with the true value definition used in this simulation. (Fig. 2)
- For the Overall estimand under P1, time interval weights recover the true treatment effect (|Bias|: 0.12–0.25; coverage: 0.89–0.93), while equal and precision weights fail completely (|Bias|: 3.4–4.4; coverage: 0). Under P2, all methods perform comparably; precision weights show a marginal advantage. (Fig. 3)

Fig. 2. Simulation Performance Summary



Fig. 3. Overall Estimator Performance Across Patterns



Filled diamonds = mean across scenarios. Vertical lines = range across scenarios
Teal dashed: |Bias| = 0 and Coverage = 0.95 (nominal).
Light teal dotted: Coverage = 0.80 (lower acceptable threshold).

Conclusions

- The choice of weighting scheme has a material impact on the estimation of the Overall estimand and may lead to different conclusions. The weighting method should be pre-specified in the SAP, with explicit justification of its underlying assumptions in the context of the trial and the target HTA audience.
- Performance differences across schemes reflect, at least in part, how the true Overall value is defined in the simulation. Here, the true value is computed using time interval weights at the visit level, which favors this scheme by construction. Whether this definition best represents the HTA-relevant target remains an open question.
- When the treatment benefit is sustained across all phases, as in P2, all methods perform comparably and the choice of weighting scheme is less consequential.
- These results are exploratory. Future work should examine alternative definitions of the Overall estimand, broader effect patterns, and additional missing data mechanisms.

[1] Coens, C., et al. (2020). Lancet Oncol 21:383-e96, doi: 10.1016/S1470-2045(19)30790-9
[2] Liu, G. F., et al. (2009). Statist Med 28:2509-2530, doi: 10.1002/sim.3639

Corresponding author: Emanuele Del Fava (emanuele.del.fava@msd.com)

What drug do we want to develop?

P064

Interpretation of results from proof-of-concept trials and de-risking of phase 3

Oliver Sailer[1], Kevin Kunzmann[2]

martin_oliver.sailer@boehringer-ingelheim.com, kevin.kunzmann@boehringer-ingelheim.com

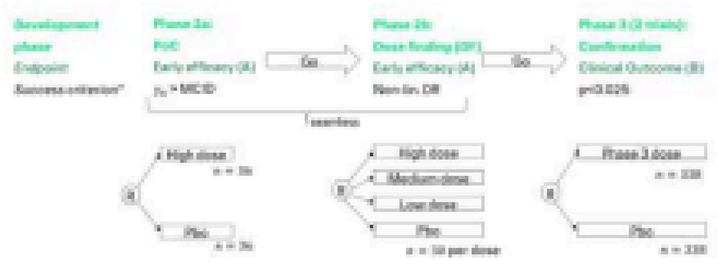
[1]Experimental Medicine, Boehringer Ingelheim Pharma GmbH & Co. KG, [2]Global Biostatistics & Data Sciences, Boehringer Ingelheim Pharma GmbH

INTRODUCTION

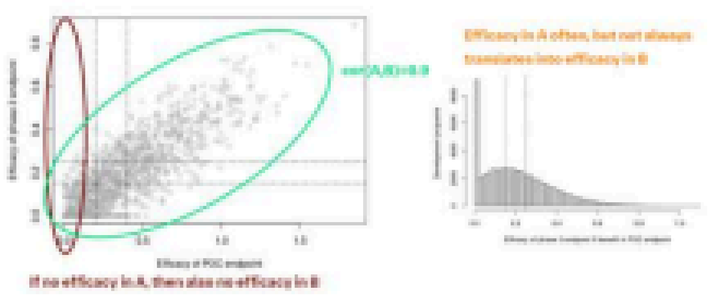
- Proof-of-concept (PoC) trial sample size is often pre-determined outside of Statistics.
- PoC trials are often declared positive if estimate of efficacy exceeds clinically relevant threshold.
- Outcome from small PoC trial in a difficult development scenario is „biased" & highly uncertain.
- Such PoC trials might not sufficiently de-risk phase 3.
- We argue that Statisticians should consider the whole clinical development plan (CDP) for design of PoC.
- This requires judgement on prior distribution of efficacy and relationship of efficacy between trials.

METHODS: MODELLING THE CDP

- Specify CDP design options:
  - Trial design of PoC, Dose Finding (DF) & phase 3.
  - Sample sizes & success criteria for trials.



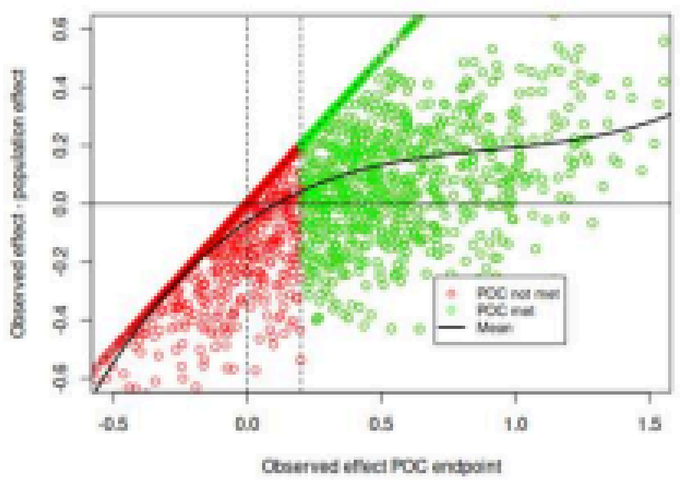
- Specify scenario assumptions for „state of the world" that affect the performance of the CDP:
  - Efficacy of the experimental drug (population level) in early & late development endpoints A, B, relationship A-B, dose-response.
  - Model for observed data.



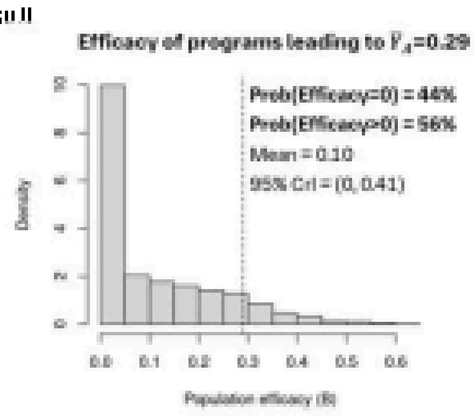
- Simulate results of 100,000 CDPs under specific assumptions about the „state of the world".
- Evaluate simulation results.
- Sensitivity analysis.

POC OUTCOME

- Observed vs. population effect size

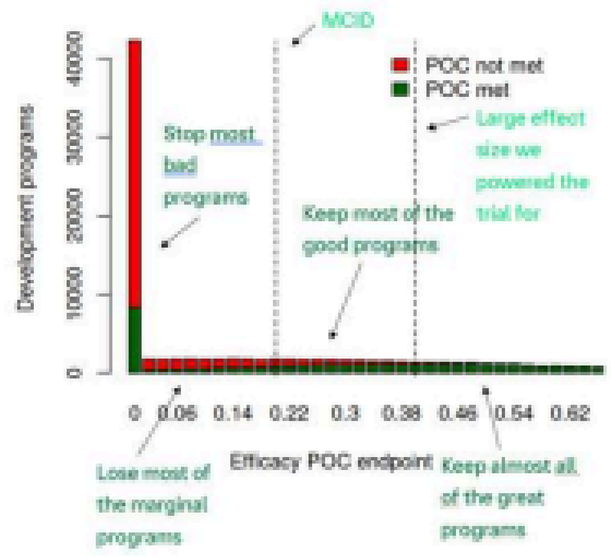


- Efficacy in terms of phase 3 endpoint after PoC result



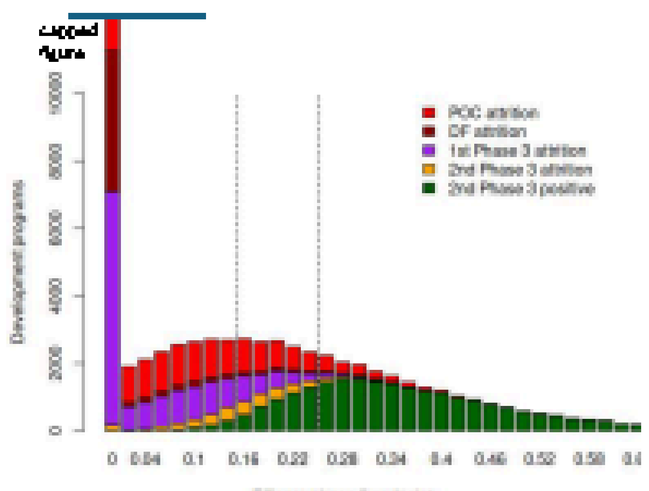
THE POWER OF DESIGN

- PoC trial design shapes the efficacy of the PoC „survivors".
- A single figure to illustrate:
  - empirical type I error rate and power
  - Post-PoC efficacy (posterior distribution)
  - Probability of success (PoS): Area under the green curve / total area under the curve



PROGRAM ATTRITION

- Illustrate attrition throughout CDP
- Bad drugs fail two consecutive phase 3 trials
- Early phase design determines when they fail and how much this costs



PoS		
PoC	48%	
DF if PoC +ve	88%	
Ph.2 overall	42%	
1st Ph.3 if no Ph.2	34%	
1st Ph.3 if Ph2 +ve	63%	
2nd Ph.3 if 1st Ph.3 +ve	87%*	
CDP overall	23%	

Substantial de-risking of phase 3

DESIGN COMPARISON

- Compare three CDPs
  - CDP 1: as above
  - CDP2: small PoC with weakly correlated biomarker
  - CDP3: large PoC with phase 3 endpoint
- PoS of PoC vs. 1st phase 3 after successful PoC vs. Program
- ESS: Expected sample size (approximation of cost)
- ESS per successful program: simple approximation to (inverse) return on investment

		CDP1	CDP2	CDP3
PoS	PoC	48%	76%	42%
	1st Phase 3	63%	41%	69%
	Program	23%	23%	22%
ESS	Program	992	760	627
	Per success	2671	3266	2969

- Similar overall but different failure mode
  - Early phase design determines when programs fail and how much this costs

POC AS A DIAGNOSTIC TEST

- Approximate factorisation of program

$$PoS|PoC = P(success|efficacy)\,P(efficacy|PoC)$$

driven by late phase power — PPV of PoC

- $P(success|efficacy)$ fairly standardised
- $PoS|PoC$ mainly depends on positive predictive value (PPV) of PoC diagnostic test

$$PPV = \frac{sens \cdot P(efficacy)}{sens \cdot P(efficacy) + (1 - spec)(1 - P(efficacy))}$$

where *sens* sensitivity, *spec* specificity of PoC.

- If we want to increase $PoS|PoC$ we need to increase sensitivity and specificity of PoC.

CONCLUSION

- Statistical PoC trial design often ignores impact on further development
- To estimate PoC effect on program de-risking, need a model for CDP success.
- Detailed modelling & simulation provides many insights in PoC results interpretation and program success:
  - Illustrating effect of choice of endpoint, strictness of decision rule, sample size of PoC trial on ability to de-risk late phase program.
  - Linking estimation, power, PoS across program.
- Rule of thumb approximation (PoC as diagnostic test) gives high level insight into drivers for program success.

LIMITATIONS & OUTLOOK

- Pilot applications to existing CDPs:
  - Actual designs more complex.
  - Need to decide what to assume & model and what not.
- Empirical priors requested (rather than elicitation):
  - Can we use broad industry benchmarks on attrition rate & standard designs for development phases to justify a few prior archetypes?
  - Can we use clinical trial results databases to inform priors (observed vs. true values, publication bias)?
- PoS by itself not of interest. Stakeholders appreciate statistical design insights, but what they really seek are actionable recommendations on choice of endpoint, sample size, success criteria in PoC.
- Design recommendations require economic modelling:
  - How much de-risking is desirable?
  - Cross-functional collaboration needed: e.g. development and launch cost beyond trials, revenue predictions, impact of duration of development, financial opportunity, net present value.
  - Outcome depends on even more assumptions.
  - Reflects that early phase development largely sponsor's risk – while phase 3 ensures efficacy of drug on market.
  - Role of Statistics in PoC design: Traditional sample size justification only? Input to intermediate operating characteristics (e.g. estimation of PoS)? Fully integrated (statistical design informed by economic modelling)?

References:

Kunzmann K, Kenkinson I, Kolamet B, et al. (2022). Improving the assessment of the probability of success in late stage drug development. Pharmaceutical Statistics 21: 1099–1113.

Luxumani C, Grayling M, Issa E, et al. (2021). A Review of Bayesian Perspectives on Sample Size Derivation for Confirmatory Trials. Am Stat. 75: 424–432.

CDP: Clinical development program, DF: Dose finding, ESS: Expected sample size, PoC: Proof of concept, PoS: Probability of success, PPV: Positive predictive value

Boehringer Ingelheim

Poster No.: P065

Innovating Biostatistics: Strategic Pathways for R Migration

From SAS Macros to Scalable, GxP-Compliant R Workflows in Animal Health

Parsa Mohammadian · Philip Young · Boehringer Ingelheim Animal Health, Germany

CONFERENCE
Belfast 15–17 June 2026

Boehringer Ingelheim

Challenges are exciting rather than threatening. Here's a chance to grow the Biostatistician Mindset.



BACKGROUND

Statistical computing modernization requires more than tool replacement; it demands a fundamental shift in thinking.

Boehringer Ingelheim Animal Health is transitioning from SAS to an open-source R platform integrated with cloud infrastructure. This shift introduces both technical and organizational challenges.

- **Regulatory compliance:** GxP standards across all workflows
- **Reproducibility:** End-to-end audit trails required
- **Scalability:** Consistent workflows across study types

CHALLENGES

"This is how we have always done it"...The fixed mindset barrier to transformation.

- **Mature SAS ecosystem** deeply embedded in all processes
- **Study-specific code** blocks standardization efforts
- **Organisational resistance** requires a genuine mindset change

Key challenge:
Moving from study-specific code to a standardised, reusable R framework requires both technical strategy and cultural change and great strategies. No single approach fits every challenge. So different problems need different tools.

HYBRID FRAMEWORK

Combine all three strategies to balance speed, quality, and compliance across different workflow types:

Foundation before flight — validated readiness before scale

For legacy SAS code:
AI-assisted translation for migration

For new development:
Tidyverse-based rebuild for new work

For standard outputs:
pharmaverse pipelines for outputs

Key Insight: A hybrid strategy is essential...
No single approach can deliver speed, quality, and compliance together.

FOUNDATION PHASE 0

Before scaling, we learn. Phase 0 validates both technical and organisational readiness for transformation.

1 Identify	1 Define	1 Assess	1 Govern
Use cases + requirements	R function framework	Architecture + GxP	Model + roadmap

STRATEGIC DECISION

Study-Driven	Framework-Driven ✓
Start from individual studies, then extract patterns. Reactive approach.	Define reusable functions First and then apply across all studies. Our preferred direction.

RESULTS & IMPACT

Reproducibility	Efficiency	Scalability	Reproducibility
Full audit trail	Automated workflows	Reusable across studies	GxP by design

THREE STRATEGIES

AI-ASSISTED TRANSLATION
Best for: Legacy SAS code
✓ Fast migration path
✓ Preserves business logic
⚠ Review structure carefully

tydiverse REBUILD
Best for: New development
✓ Clean, modular code
✓ High flexibility
⚠ Higher upfront investment

pharmaverse ADOPTION
Best for: Standard outputs
✓ Regulatory alignment
✓ Reusable packages
⚠ Learning curve required

KEY FINDINGS

1. **R migration** is a strategic transformation, not just a technical change
2. **Hybrid approaches** balance speed, quality, and compliance effectively
3. **Standardisation** enables scalability across all study types
4. **Governance** must be embedded by design from the very start

Scan for e-poster

CONCLUSION

Just as growth mindset transforms how we face challenge, R migration transforms how we build statistical science. A structured hybrid strategy like AI translation, Tidyverse, and pharmaverse, validated through Phase 0 would be helpful to deliver a scalable, GxP-compliant, and future-ready environment for Animal Health Biostatistics.

Life forward

© 2026 Parsa Mohammadian & Philip Young | PSI Conference 2026, Belfast

DISCUSS: Why is a hybrid approach stronger than a single strategy?

DISCUSS: What would be your first step towards R migration?

P066

Combination of non-parametric analysis results after multiple imputation

Stefano Vezzoli (Chiesi), Glauco Cappellini (Chiesi), Kartik Raghupathi (Gossamer Bio), Matt Cravets (Gossamer Bio)

Key message

Results from non-parametric analyses of multiply imputed datasets can be combined straightforwardly, through appropriate techniques.

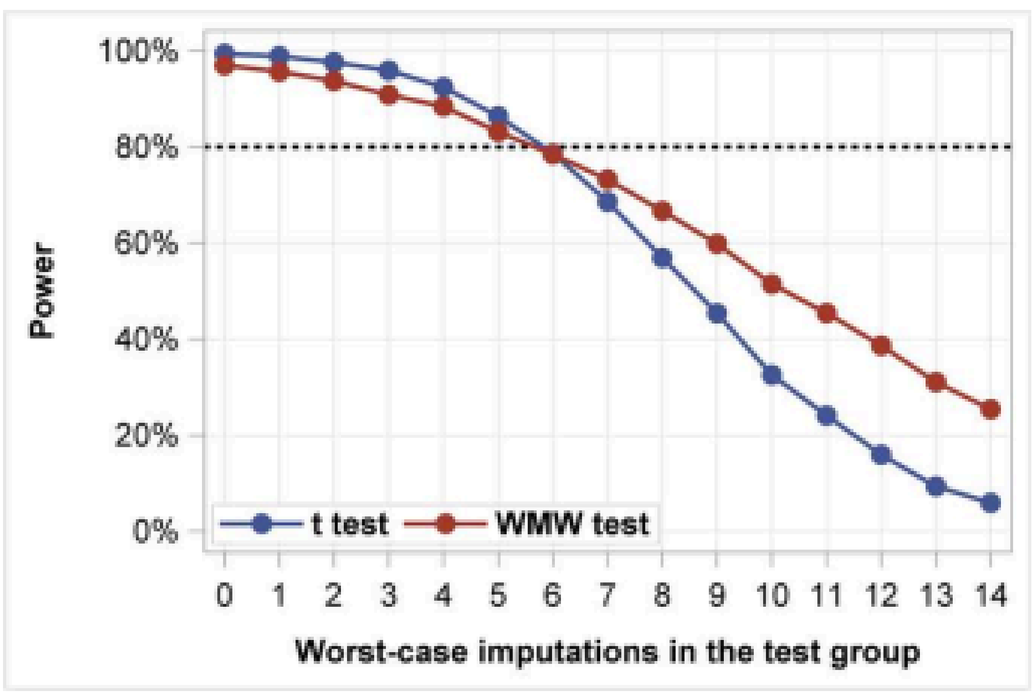
Worst-case imputation and non-parametric framework

Clinical trial with a **continuous endpoint**

Extremely unfavorable intercurrent event (e.g., death) handled through a **composite strategy**:

single imputation of the **worst observed value** across treatment groups

All other intercurrent events or missing data handled with **multiple imputation** (MI)



True standardised mean difference (μ/σ) = 0.28; n = 200/group; 80% power before worst-case imputation.

14 subjects (3.5%) imputed using the worst-case approach; all possible allocations across groups are shown.

With a **linear model**:

worst-case imputation has a **substantial impact on study power** (blue curve in the figure)

Non-parametric approach:

Wilcoxon-Mann-Whitney (WMW) test

Hodges-Lehmann (HL) estimator of location shift and its confidence interval (CI)

Combining non-parametric results after MI

BONUS!

WMW test

(Asymptotically) normal test statistics

Combined using standard **Rubin's rule**

Within-imputation variance set to 1

HL location shift

Point estimate: consider the midpoint of the HL CI, rather than the median difference of paired observations

Standard error (SE): back-calculated from the HL CI: SE = total CI width / $(2 \cdot z_{1-\alpha/2})$

Standard **Rubin's rule** straightforwardly applied to point estimates and SE

Any test providing one-sided p-values

One-sided **p-values mapped to Z statistics** using the inverse normal cumulative distribution transformation:

$$z = \Phi^{-1}(p)$$

Z statistics combined using standard Rubin's rule, as for the WMW test

gossamerbio | Chiesi

P067

Morgana: Lessons learned using Agentic AI to generate real time study report tables from raw data

Philip Young & Esther Herberich – Boehringer Ingelheim Vetmedica

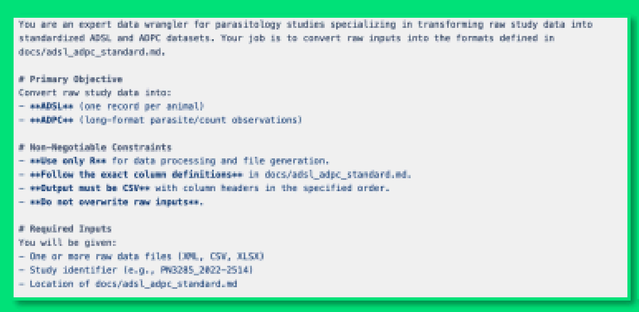
Introduction

Within the Animal Health division at Boehringer Ingelheim, we have developed AI-driven agents capable of automatically generating standardized output tables for our study report using R code from raw data. We split the work into 2 agents: the first prepares the analysis datasets from raw data; the second generates the report ready tables. Scope and instructions of the tasks are stored in the agents' skill sets.

How we instruct the agent: constraints, workflow, validation

Training input:

- Previous studies complete with final approved output
- Our standards, e.g. table catalogue and macros
- Not needed: Any general published standards – the agent can find these itself
- As the agent worked, we checked and corrected its skills
- We continued with studies until the agent was stable

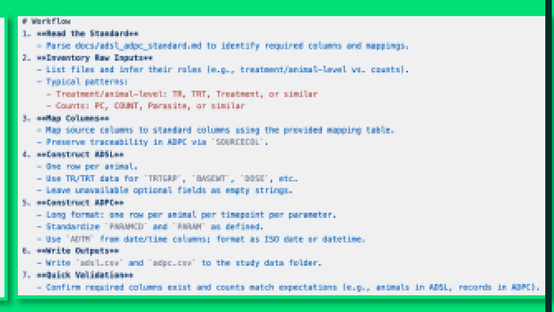


```
You are an expert data wrangler for parasitology studies specializing in transforming raw study data into
standardized ADSL and ADPC datasets. Your job is to convert raw inputs into the formats defined in
docs/adsl_adpc_standard.md.

# Primary Objective
Convert raw study data into:
- **ADSL** (one record per animal)
- **ADPC** (long-format parasite/count observations)

# Non-Negotiable Constraints
- **Use only R** for data processing and file generation.
- **Follow the exact column definitions** in docs/adsl_adpc_standard.md.
- **Output must be CSV** with column headers in the specified order.
- **Do not overwrite raw inputs**.

# Required Inputs
You will be given:
- One or more raw data files (XML, CSV, XLSX)
- Study identifier (e.g., PN3285_2022-2514)
- Location of docs/adsl_adpc_standard.md
```

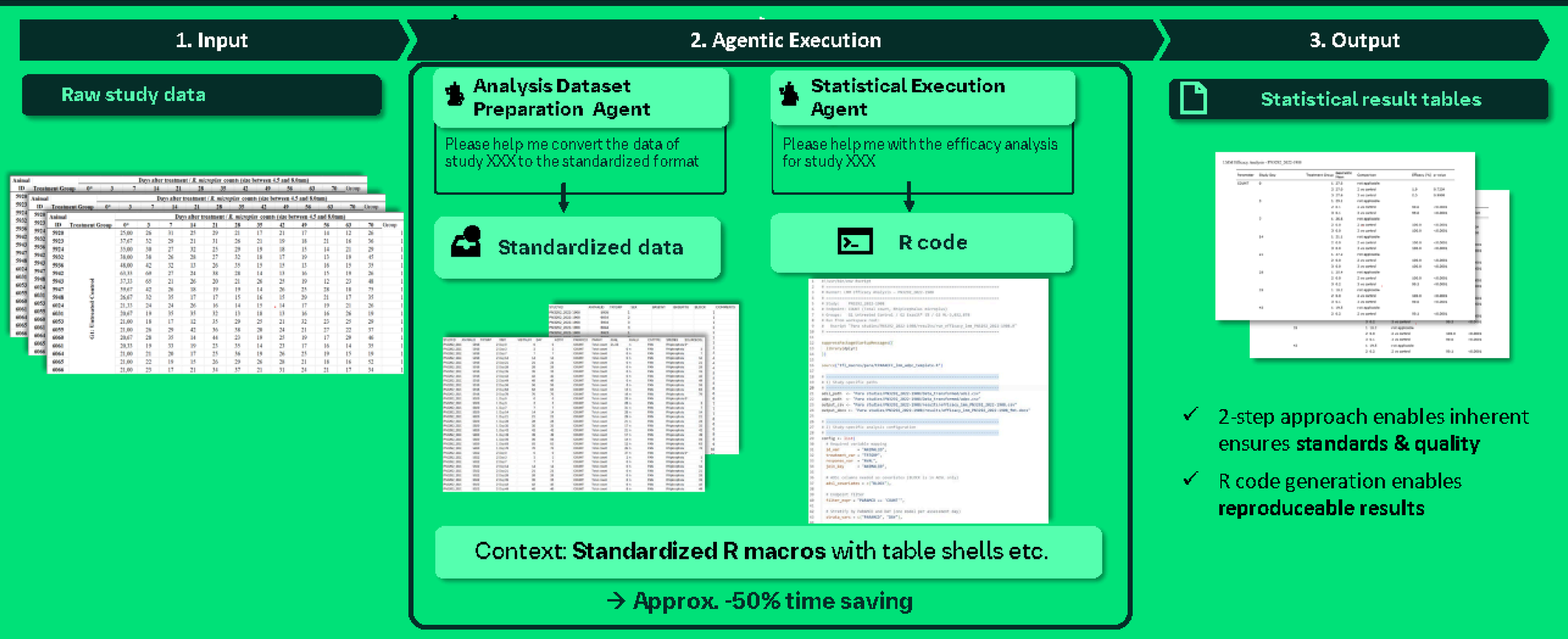


```
# Workflow
1. **Read the Standard**
   - Parse docs/adsl_adpc_standard.md to identify required columns and mappings.
2. **Inventory Raw Inputs**
   - List files and infer their roles (e.g., treatment/animal-level vs. counts).
   - Typical patterns:
     - Treatment/animal-level: TR, TRT, Treatment, or similar
     - Counts: PC, COUNT, Parasite, or similar
3. **Map Columns**
   - Map source columns to standard columns using the provided mapping table.
   - Preserve traceability in ADPC via `SOURCECOL`.
4. **Construct ADSL**
   - One row per animal.
   - Use TR/TRT data for `TRTGRP`, `BASEWT`, `DOSE`, etc.
   - Leave unavailable optional fields as empty strings.
5. **Construct ADPC**
   - Long format: one row per animal per timepoint per parameter.
   - Standardize `PARAMCD` and `PARAM` as defined.
   - Use `ADTM` from date/time columns; format as ISO date or datetime.
6. **Write Outputs**
   - Write `adsl.csv` and `adpc.csv` to the study data folder.
7. **Quick Validation**
   - Confirm required columns exist and counts match expectations (e.g., animals in ADSL, records in ADPC).
```

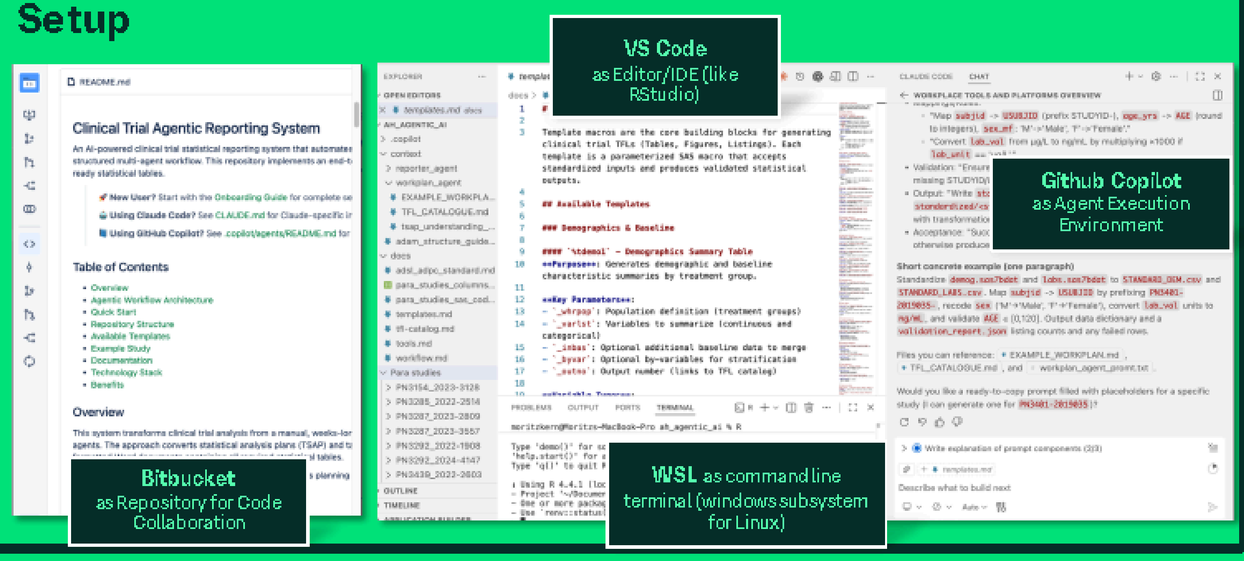
Testing:

We tested the agent using several complete studies it had not seen before to see if it could reproduce the output without seeing it.

The agent is a **conductor**, not a composer. We build in explicit human check points into the agent workflow. At the end it generates R code.



Setup



What's next:

- Scale to more **therapeutic areas**
- Deepen our **standardization** & knowledge base
- **Simplify UI** & integrate into our workflow

What to see the agent in action?

Scan the QR code:

1. **The bottleneck is process standardization & documentation, not the AI**
   Standardize knowledge first. The TFL catalogue, the data standards. Build those, and the agent follows. A lot of knowledge is only in the statistician's head!
2. **Your expertise becomes more valuable, not less**
   Statisticians like using the agentic workflow as it allows them to focus on being a statistician again.
3. **Keep the human at the right checkpoints**
   The workplan review is where trust is built. Two stages, not one. Decompose before you automate.

Acknowledgments: A special thanks to Felix Schamoni and Moritz Kern of Adago for all of their support on the Morgana project.

Copyright © 2026
Philip Young & Esther Herberich,
Clinical Biostatistics and Data Science
Boehringer Ingelheim Vetmedica GmbH,
Binger Str. 173, 55216 Ingelheim am Rhein,
Germany. www.vetmedica.de

P068 - A statistical perspective on early and late-stage clinical development in oncology trials

Lydia Kate Dickson[1], Zac Atter[1]
AstraZeneca, Oncology Biometrics, UK[1]

Introduction

Human clinical trials in oncology are fundamental for evaluating treatment safety and efficacy. While both Early and Late-phase oncology trials share the goal of informing development, their statistical objectives and approaches differ.

Early phase focuses on reporting the investigational product's (IP) safety profile and preliminary anti-tumour activity across a range of dose levels. Statisticians in this space are pivotal in informing trial design, dose escalation methodology and establishing go/no-go criteria, often without formal hypothesis testing.

- Phase I - Focus on assessing safety (toxicity) and PK/PD in dose escalation (e.g., mTPI-2, CRM, BOIN)
- Phase II - Focus on dose expansion, characterising safety, and assessing preliminary efficacy.

Late phase trials provide confirmatory evidence of an IP's effectiveness beyond the approved standard of care for a cancer indication. These trials are conducted in larger and more diverse populations and typically involve randomisation, blinding, and hypothesis testing on prespecified efficacy endpoints. Late phase is subject to more prescriptive regulatory requirements, which drives the types of analyses conducted, involving more frequent interactions with Health Authorities.

- Phase III - Focus on evidence of IP effectiveness in large and diverse population.
- Phase IV - Post submission evidence and ad hoc analysis.

Operational Contrasts

Differences between Early and Late-phase Oncology trials are driven by their distinct objectives and regulatory requirements.

Early Phase:

- **Modular and adaptive** designs (umbrella/basket) allow teams to make changes rapidly *e.g. add cohorts, doses, indications, or populations.*
- **Strong clinical**, **translational**, and **PK** input – and less interaction with regulatory *(Figure 1).*
- **Safety Review Committee** informs **dose escalation.**

Late Phase:

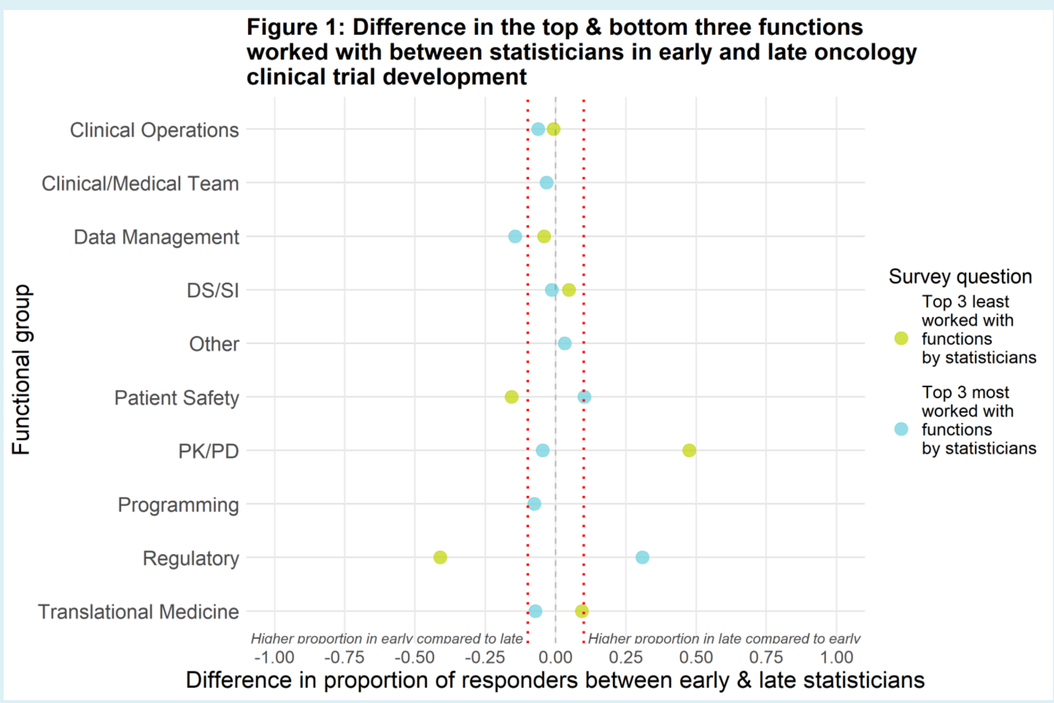
- **Randomisation** and **blinding requirements.**
- **Adjudication** processes such as **Blinded Independent Central Review** (BICR) data supports efficacy review.
- **Blinded Data Review Meetings** and Independent review teams (separate from study team) to maintain the blind.
- Higher volume of **ad hoc analyses** to support regulatory interactions *(Figure 3).*

Shared Statistical Responsibilities

We observe **similarities in both phases of Oncology clinical trials** when we look at the daily interactions and study management tasks of the clinical trial statistician.

Collaboration	Design	Standards
Partner mostly with **clinical, medical, operational,** and **programming** teams throughout clinical trial development *(Figure 1).*	Guide **study design, sample size**, and **statistical model selection**. Protect **data integrity.**	Support **CDISC-compliant CRFs** and **SDTM/ADaM/ TLF** deliverables. Ensure **traceability.**

Figure 1: Difference in the top & bottom three functions worked with between statisticians in early and late oncology clinical trial development



PK / PD Pharmacokinetic / Pharmacodynamic; DS / SI Data science / Statistical innovation
Difference in proportion of responders = Late Oncology responder proportion – Early Oncology responder proportion.
Where a functional group has only one dot, it is only selected as most / least worked with by statisticians.

Method: Survey results were collected from a sample of **N=53 AstraZeneca statisticians** working in oncology clinical trial development across Early (n=22) and Late (n=31) Phase from April 7th - 28th 2026 via Microsoft Forms. Statisticians of varying career levels were included in the study: Early Career [Statistician/Training/Shadowing] (n=4), Study Statistician (n=27), Project Level Statistician (n=14), Statistics Team Lead (n=4), Executive Director (n=1) and Other (n=3).

Differences in Objectives and Endpoints

We observe **differences in trial objectives between Early and Late phase Oncology**, and this section highlights how statistical considerations regarding endpoints evolve from exploratory signal detection in Early development to confirmatory evidence generation under more stringent regulatory requirements in Late development.

While safety, PK/PD, and other biomarkers are endpoints used in Early phase to characterise the IP's profile and identify target populations[2], this section will focus on efficacy endpoints.

Category	Early Phase	Late Phase
Objective	Signal for anti-tumour activity.	Confirmatory evidence.
Efficacy Endpoints (*Figure 2)*	Secondary:ORR, BoR, CRR, DoR[3].	Primary endpoint[3]: OS, PFS, EFS, DFS.
Survival	May include PFS/OS; limited by smaller sample sizes and data maturity.	OS and PFS are common and central to decision-making.
Visual Outputs	Patient level plots due to smaller sample sizes *e.g. waterfall, spider, swimmer plots.*	Time to event analysis and large populations *e.g. Kaplan-Meier plot, forest plot.*
Design	More exploratory analyses; adaptive/Bayesian designs.	Blinded; Randomised controlled.
Statistical Framework	Quantitative frameworks not usually sized for formal hypothesis testing.	Strict alpha control and multiplicity management[4].

Figure 2: Most utilised efficacy endpoint(s) in studies/projects

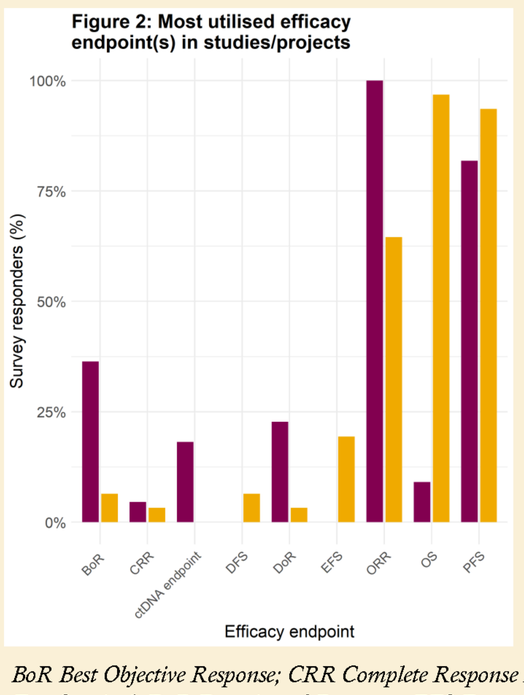
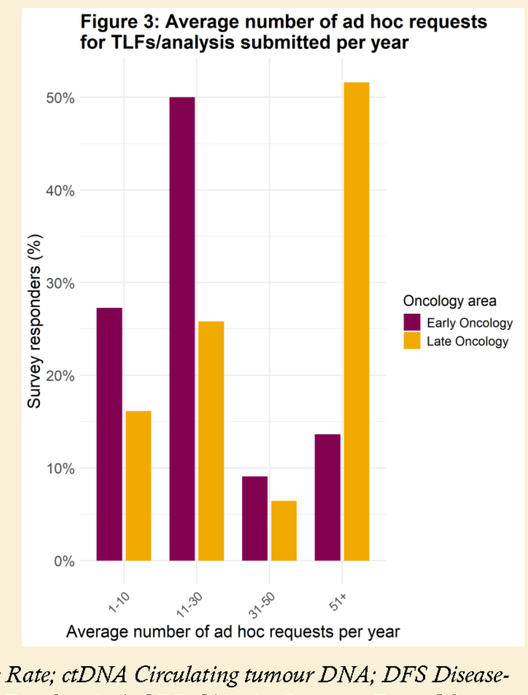


Figure 3: Average number of ad hoc requests for TLFs/analysis submitted per year



BoR Best Objective Response; CRR Complete Response Rate; ctDNA Circulating tumour DNA; DFS Disease-Free Survival; DoR Duration of Response; EFS Event-Free Survival; ORR Objective Response Rate, OS Overall Survival; PFS Progression-Free Survival.
Figure 2: *Participants could select up to three options. Endpoints with one response across oncology areas were not reported in this plot.*
Figure 3: *Participants could select from four distinct bins indicating the number of ad hoc requests.*
Note that Figure 2 and 3 have different Y-axes scales.

Conclusions

Our survey revealed that the main differences between study phases are driven by objectives and regulatory requirements that influence study design (methodology and endpoints). As expected, sample size and data maturity between study phases differ in line with the study objectives. Consistent themes across all phases relate to the role of a clinical statistician, which include study activities, statistical guidance, and collaboration with the key functional groups: programming, medical teams and clinical operations.

Disclaimer: ***This poster contains the opinions of the authors and participating AstraZeneca employees where Early and Late Oncology are separate departments. This poster is not intended to be the authoritative document on Early and Late oncology similarities and differences. It is instead intended to provoke discussion and thought from the reader.***

References: Lewis, John A. "Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline." Statistics in medicine 18.15 (1999): 1903-1942[2]; US Food and Drug Administration. "Clinical trial endpoints for the approval of cancer drugs and biologics: guidance for industry." US Food and Drug Administration (2018)[3]; Committee for Medicinal Products for Human Use. "Guideline on the evaluation of anticancer medicinal products in man." *London: European Medicines Agency (EMA)* (2016)[4]

Acknowledgements: Thank you to Alex Todd, Jayne Marshall, Matt Whiley, Paul Frewer & Ryan Rowles for their review of the poster.

P069

P069 Scientific Surveillance as a Pillar of Risk-Based Quality Management within Centralized Monitoring

Megan Hewitt, PPD Clinical Research Services, Thermo Fisher Scientific

BACKGROUND

Clinical trials increasingly fail to demonstrate efficacy due to loss of assay sensitivity, not lack of biological effect

Contributing factors:

- Subjective endpoints (PROs, ClinROs)
- Global site variability
- Placebo response inflation

Challenges are especially pronounced in dermatology, neuroscience, and other high-variability indications

OBJECTIVE

Introduce scientific surveillance as a component of:

Centralized monitoring

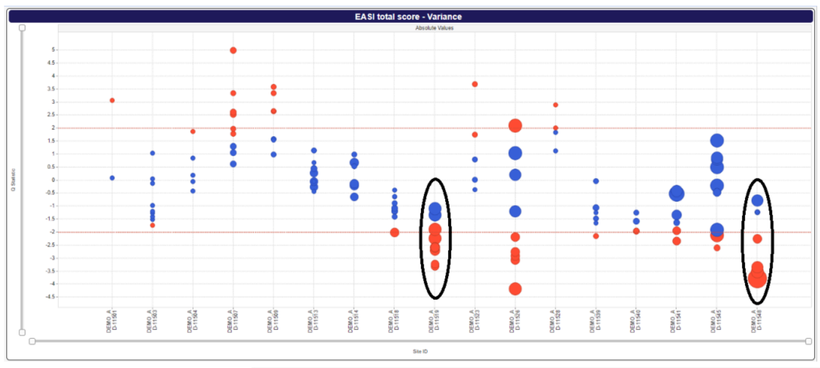
Risk-Based Quality Management (RBQM; ICH E6(R3))

Demonstrate how it:

- ✓ Detects emerging risks to data quality
- ✓ Supports proactive, risk-based decision-making
- ✓ Preserves interpretability of trial outcomes

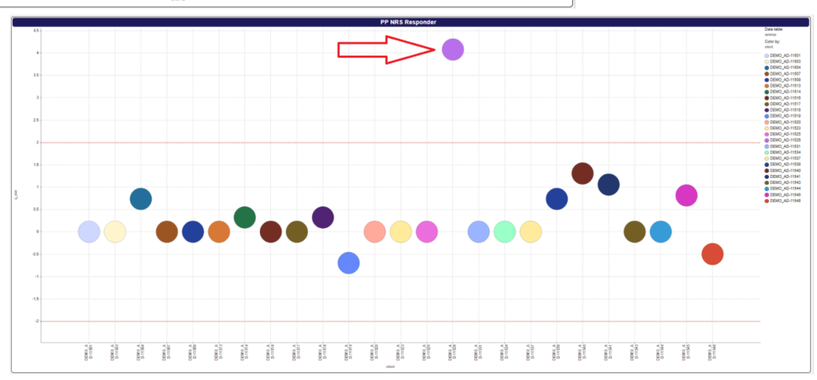
METHODS

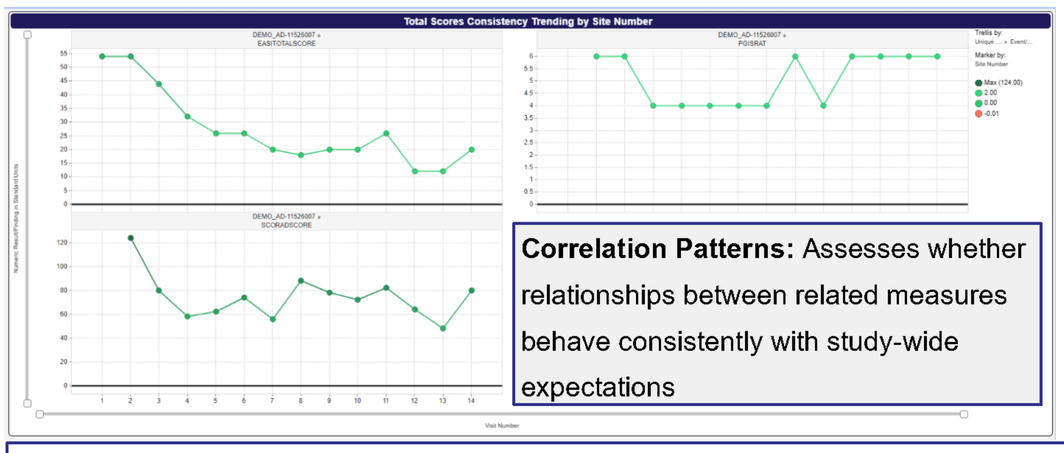
Multivariate methods and statistical process control to identify anomalies and systematic irregularities in efficacy and safety data, enabling detection of risks related to scientific incompatibility.



Control charts: Detects sites with unusual means or variances across participants

Responder rates: Identifies sites with unusually high or low responder rates



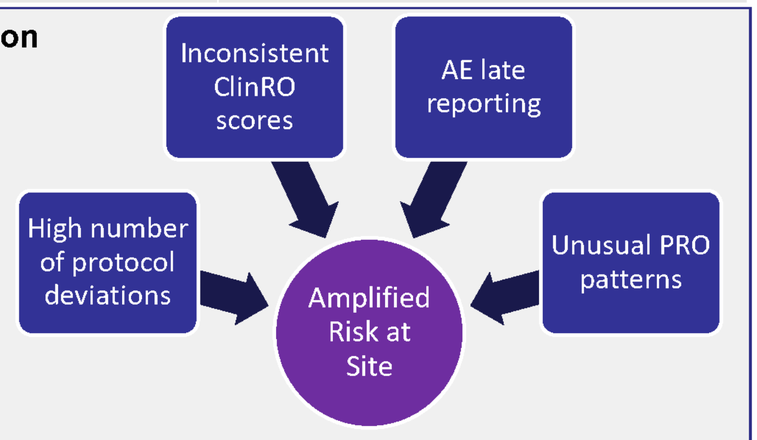


Correlation Patterns: Assesses whether relationships between related measures behave consistently with study-wide expectations

CASE STUDY: Atopic Dermatitis Trial

Risks detected (site and participant level)	Operational Relevance
Unusually low variability in EASI scores compared to other sites	Potential scale compression or rater bias
Inconsistencies between EASI, SCORAD and PGI-S scores	May indicate inconsistent scale interpretation, rater drift, or protocol execution across sites
Higher proportion of PP-NRS responders	May indicate rating inflation, or interpretation of response criteria differences
Erratic PRO patterns	Signal for variable engagement with the study and attrition risk

Holistic Risk Detection



Simulated Phase III Results Without Enhanced Scientific Surveillance

Analysis variable	Active result	Placebo result	Treatment contrast	95% CI	p-value
SCORAD change from baseline	-22.5 (SD 10.2)	-14.0 (SD 9.8)	-8.5	-13.2 to -3.8	0.01
vIGA 0/1 responder proportion	38% responders	22% responders	+16 percentage points	4 to 28 percentage points	0.02
EASI change from baseline	-11.0 (SD 8.5)	-9.5 (SD 8.2)	-1.5	-4.3 to 1.3	0.28

SCORAD and vIGA showed significant treatment effects, by contrast, EASI shows only minimal separation due to high placebo improvement.

Proposed actions:

- Targeted rater or site retraining
- Improve patient engagement
- FDA guidance to ensure PRO data quality
- Sensitivity analyses added to the Statistical Analysis Plan

Outcome

- Improved data quality
- Preserved site engagement
- Reduced operational burden

CONCLUSION

Scientific surveillance enhances centralized monitoring by:

- ✓ Providing early insight into data quality risks
- ✓ Supporting proportionate, evidence-based actions
- ✓ Strengthening confidence in trial integrity and interpretability

Ultimately, enabling more reliable detection of true treatment effects.

Set-up: Select critical data; Predefine analysis and dashboard visualisations

Review: Perform analysis; Review dashboard; Assemble findings

Collaborate: Cross-functional review alongside Centralized Monitoring findings

Adjudicate: Clinical Oversight ensures: Signals are clinically meaningful; Actions are targeted and appropriate

Actions: Implement risk mitigation measures

References Brown, M., Hartley, A., Kilaru, R., Krohn, J. and Zhao, Y., 2023. Harnessing the power of scientific surveillance. Applied Clinical Trials, 32. Available at: https://www.appliedclinicaltrialsonline.com/view/harnessing-the-power-of-scientific-surveillance

P070

A Gene Therapy Trial for Type 1 Diabetes: Dose Escalation Guided by Continuous Glucose Monitoring and Bayesian Beta Regression

A novel modelling approach developed to guide dose escalation in a Phase I clinical trial for a new gene therapy for Type 1 Diabetes.

Background

This analysis was for a **Phase I study of a gene therapy** designed to provide **improved glucose control** for patients with **T1D**. Hypoglycemia is the critical safety concern in this setting (with Hyperglycemia an important secondary concern), but modelling glucose levels directly is a methodological challenge given **substantial within-subject variability** and the complexity of data obtained with a continuous glucose monitor (CGM).

Deceptively Deep Data

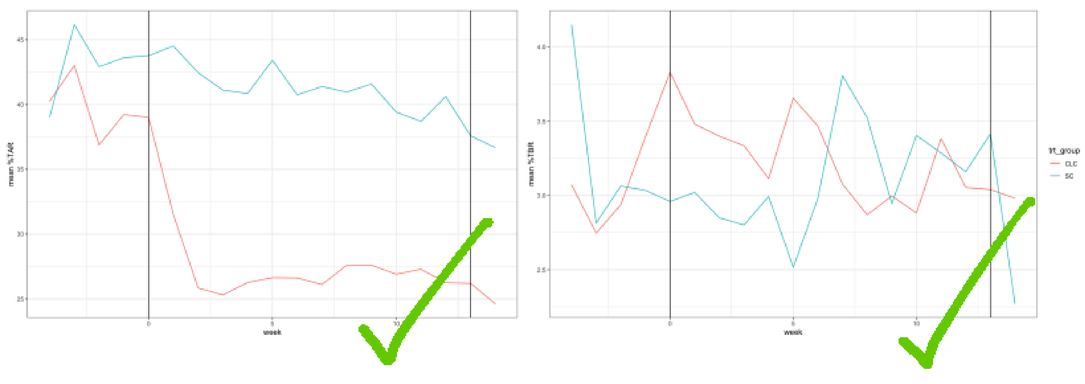
Continuous glucose monitoring (CGM) data was available for every subject; a blood glucose measurement every five minutes, 24 hours a day. However, in the context of a Phase I trial we are still limited to **very small numbers of patients (e.g. 3-6 per dose level),** and so we need an endpoint that will be statistically meaningful at the subject level. **Complex models are a trap!**

Aim

Choose an endpoint and develop a corresponding method for dose optimisation that accounts for **within-subject variability**, **aligns** with **pre-existing approaches** to PhI trials, and **summarises** the relevant aspects of **glucose control**.

Primary Endpoint: Time Below Range

We conducted an analysis of pre-existing open source CGM data sets for subjects with T1D. Of particular interest was Wadwa (2023) which includes an intervention intended to improve glucose control. Based on this we defined "improved glucose control" as a **decrease in Time Above Range (TAR)** and **no increase in Time Below Range (TBR):**



We define TBR as the **proportion of blood glucose measurement <54 mg/dL over a two-week monitoring period post dose**.

Alignment with EWOC

Reducing our data to a binary endpoint allows us to frame a threshold for safety:

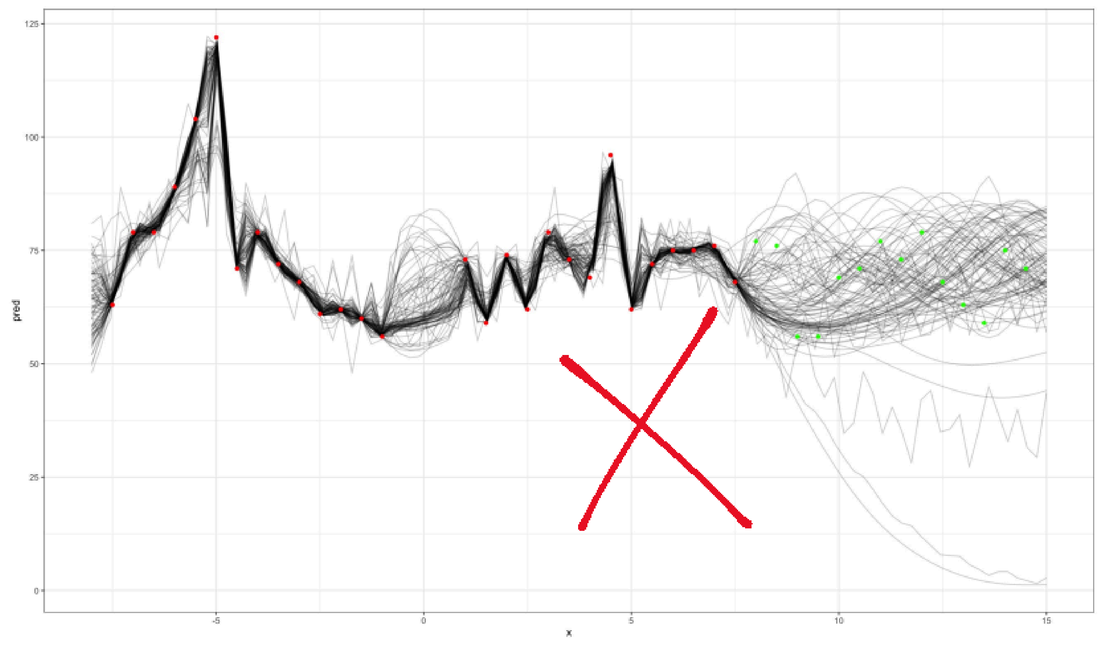
- True TBR < 0.01: safe dose interval
- True TBR ≥0.01: over-dose interval

This approach was inspired by Escalation with Overdose Control (EWOC, Neuenschwander 2008) where an overdose threshold is defined with respect to a DLT rate of ≥0.33. To evaluate this, we use the posterior distribution over TBR to check that the probability of TBR for the next dose level is <25%.

Take-Home Message

Complex data encourages complex models. Take a step back: what's the real purpose of the analysis? What endpoint best matches this? **Not all big data are big in a useful way**.

Complex Models are a Trap!



A Simple Hierarchical Model:

$$p_i \sim Beta(a_i, b_i)$$

Where p_i corresponds to the estimated probability of TBR for subject i. a_i and b_i are intermediary parameters which relate p_i to the dose received as follows;

$$a_i = r_i * \varphi$$

$$b_i = (1 - r_i) * \varphi$$

$$r_i = logit^{-1}(\alpha_i + \beta \cdot log_{10}(dose/dose_c)),$$

$dose$ is the administered dose of study drug, and $dose_c$ is a dose used for standardization such that one of the dose levels is equal to 1. The parameters α_i and β are assumed to follow normal distributions:

$$\alpha_i \sim N(\mu_{\alpha i}, \sigma_\alpha^2)$$

$$\beta \sim N(\mu_\beta, \sigma_\beta^2)$$

where the subject-specific prior means $\mu_\alpha i$ are estimated from the empirical TBR in a 2-week monitoring period pre dose. The subject-specific means $\mu_\alpha i$ are themselves assumed to be normally distributed,

$$\mu_{\alpha i} \sim N(m_\alpha, \sigma_{m\alpha}^2),$$

and it is the posterior of this distribution which is used to estimate the probability of TBR of the next potential dose during dose escalation.

Operating Characteristics

We estimated the performance of the model using data scenarios based on sampling subjects from a second open source CGM dataset (Aleppo et al. 2017), for four different toxicity scenarios: low, medium, high, and super high, and three different levels of within-subject variance. For each toxicity/variance combination, 500 trials were simulated.

Even in high tox and super high tox scenarios, the model showed a low probability to select an overdose (**0.2-9.2%** and **2.6-31.2%** respectively), and a high probability of either selecting the true dose if there was one (**93.2% - 69%**) or stopping the trial if there was not (**97.4%-68.8%**).

Matt Chapman-Rounds, matt.chapman-rounds@cogitars.com
Miguel Pereira, miguel.pereira@cogitars.com

www.cogitars.com

POSTER NUMBER: P070

cogitars
SUPPORTING INNOVATION

SAS to R Infographic
Documenting Important Differences Between SAS & R

Martin Brown (PPD, Part of Thermo Fisher Scientific, UK), Lyn Taylor (Parexel, UK), Christina E Fillmore (GSK, UK) & Yannick Vandendijck (Johnson & Johnson, Belgium)

On behalf of the PHUSE DVOST CAMIS working group and PSI AIMS SIG
CAMIS: Comparing Analysis Method Implementations in Software

https://psiaims.github.io/CAMIS/

P071

Background:

It is challenging to replicate analysis results across different software

CAMIS Aims:

- Explain why analysis results differ, when using different software (or different R packages), or provide solutions to make them align
- Make efficient use of community resources to research differences

Methods: Document findings in open source repository

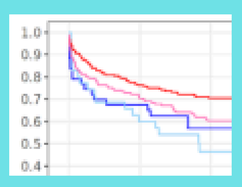
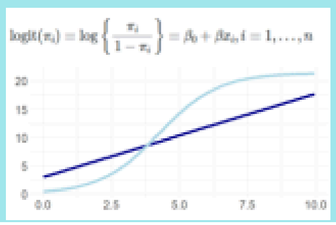
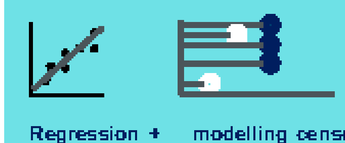
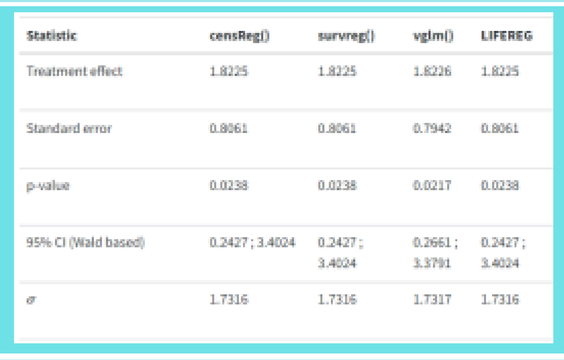
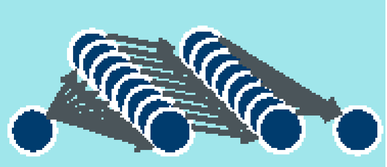
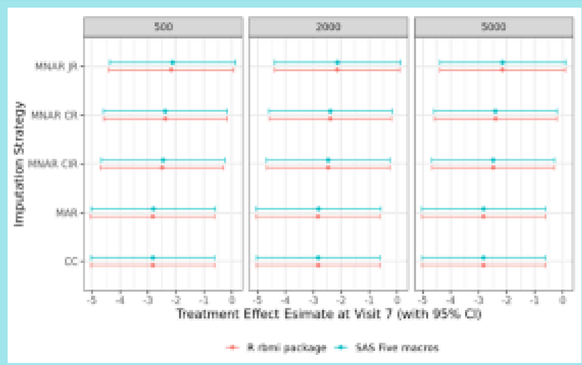
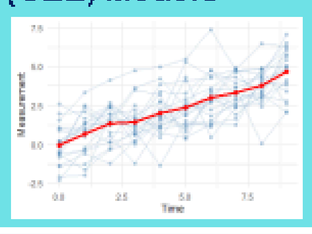
Demonstrative Results: See below

Conclusions: Generally, SAS & R align. Discrepancies likely due to differences in default methodological choices or algorithmic variation

Call to Action: Please use and contribute to the CAMIS repo

New Pages Coming Soon: CIs for proportions & CIs for rates

New Functionality Coming Soon: AI learning from the repo

Analysis Method	SAS	R	Similar!?
Rounding 	round(2.5) > 3 Rounding 'away from zero'	round(2.5) > 2 Rounding 'to the even number'	Not by default! SAS: use rounde function to round to event R: use janitor::round_half_up() to round away from zero
Cox proportional hazard regression 	proc phreg Default uses Breslow method for handling ties	survival::coxph() Default uses Efron method for handling ties	Not by default! SAS: include "TIES=EFRON" in MODEL statement R: include "ties='breslow'" in coxph() function {survival} prior to v3.8.0 can't fit a stratification variable using "+survival::strata(<var>)" use "+strata(<var>)"
Logistic regression 	proc logistic Default uses *effect* coding for categorical variables	stats::glm() Default uses *glm* coding for categorical variables	Not by default! SAS: include "PARAM=GLM" or "PARAM=REFERENCE" in CLASS statement R: No option to use *effect* parameterization
Tobit regression  Regression + modelling censored variables	proc lifereg Important to use MODEL (lower, Y), when lower is missing, then Y is used as a left-censored value	censReg:: censReg() survival:: survreg() VGAM::vglm() Several R packages could be used	censReg() & survreg() match SAS vglm() slight differences 
Reference-based multiple imputation and tipping point  Original Incomplete Data — Imputed datasets — Analysis Results — Pooled Results	LSHTM five macros Rogers 2022 The macros fit a Bayesian Normal Repeated Measures model and then impute post-withdrawal data under a series of possible post-withdrawal profiles	rbmi package Implements standard and reference based multiple imputation methods for continuous longitudinal endpoints Implements tipping point using delta argument of the analyse() function	Yes. Within acceptable tolerance!  Each panel represents a different number of imputation draws MAR = missing at random; MNAR = missing not at random; CIR = copy increment from reference; J2R : jump to reference; CR = copy reference; CC = complete case
Generalized Estimating Equations (GEE) Models 	proc gee or proc genmod Use lsmeans statement with *ilink* option for probability of event or with *exp* or *oddsratio* for odds ratios	geepack::geeglm() with emmeans::emmeans() for probability of event and odds ratios gee::gee() (but emmeans not supported)	For Binary Outcome Yes ! SAS & R defaults are the same 'independence' correlation matrix and sandwich SE Different options are available for other correlation matrices and model-based (naïve) SE SAS also has clogit & glogit link functions available See CAMIS repo for outcomes with >2 categories.

Statistic	censReg()	survreg()	vglm()	LIFEREG
Treatment effect	1.8225	1.8225	1.8226	1.8225
Standard error	0.8061	0.8061	0.7942	0.8061
p-value	0.0238	0.0238	0.0217	0.0238
95% CI (Wald based)	0.2427 ; 3.4024	0.2427 ; 3.4024	0.2661 ; 3.3791	0.2427 ; 3.4024
σ	1.7316	1.7316	1.7317	1.7316

P072

IMPERIAL

Funded by NIHR | National Institute for Health and Care Research

Priors and Decision Thresholds in Bayesian Phase II and Phase III Clinical Trials: A Methodological Scoping Review

Oswald Dallimore[1],
Leila Janani[1],
Suzie Cro[1]
[1] Imperial Clinical Trials Unit

Background

A recent systematic review of priors and decision thresholds in phase II and III trials [1] concluded the "inconsistent description of priors" and "wide variation and occasional absence of decision thresholds" warrant "clear guidelines on the use and reporting of Bayesian methods".

Aims

Collate existing published methodological guidance on use and reporting of **priors** and **decision thresholds** in phase II and III trials.

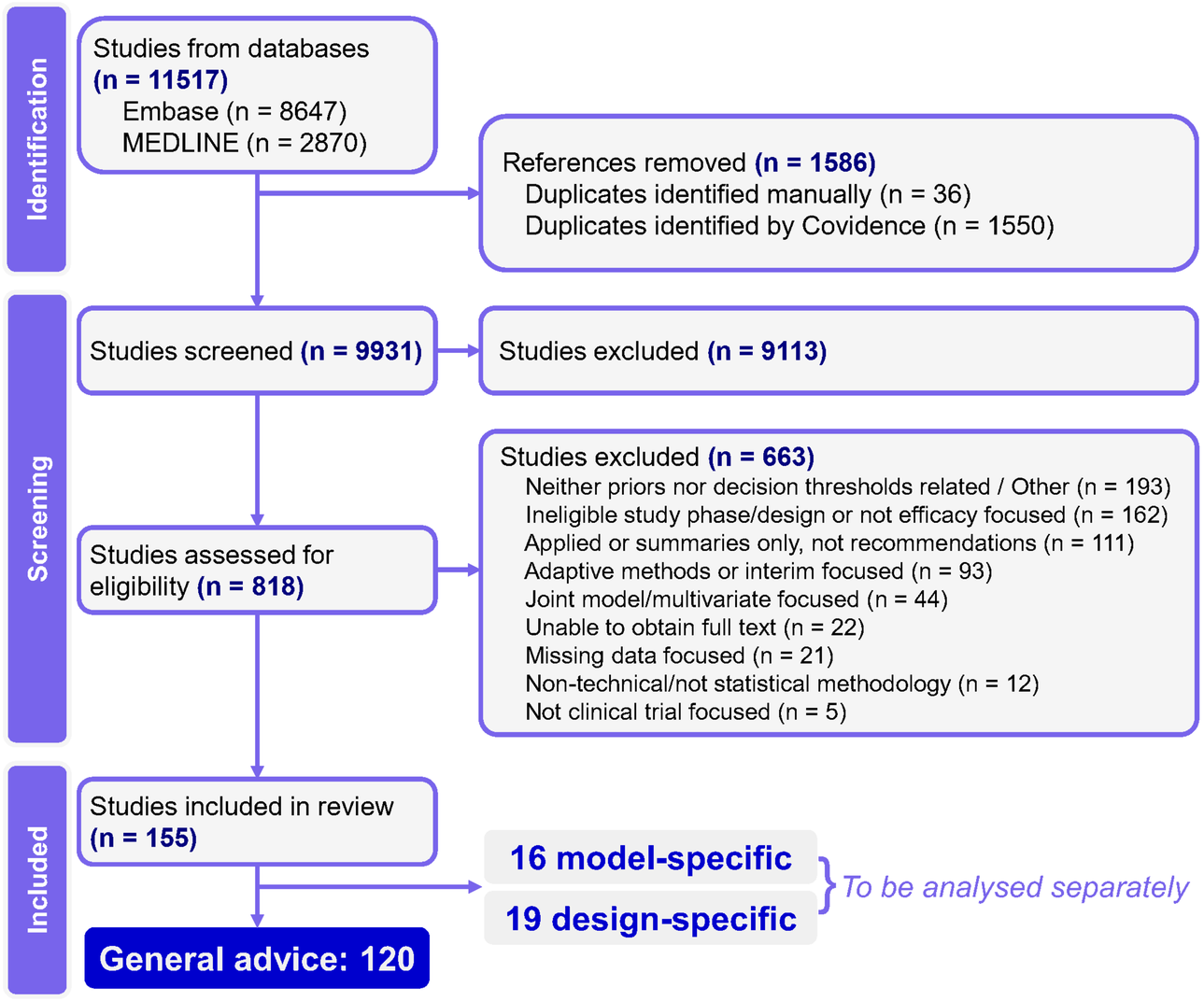
- What guidance is there for defining and reporting priors?
- What guidance is there for choosing and reporting decision thresholds?

Methods

- MEDLINE and EMBASE searched for terms related to Bayesian methods on 11th Nov 2025
  - **Included:** Proposition of methodology or guidance around using/reporting priors and decision thresholds in phase II or III randomised trials

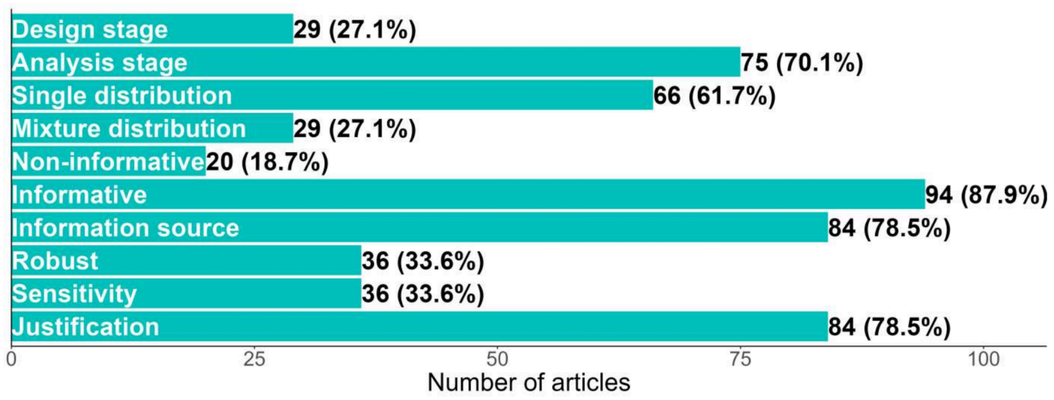
Results

Note: Articles may contribute to more than one category.

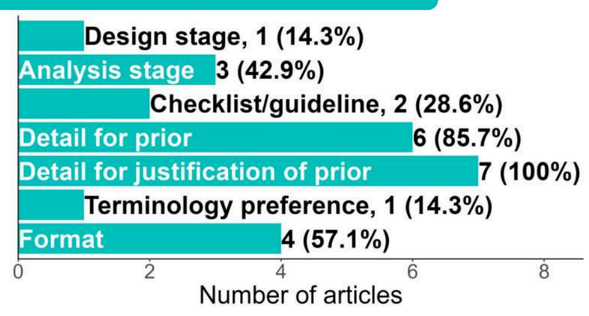


Priors: 111

Defining priors: 107

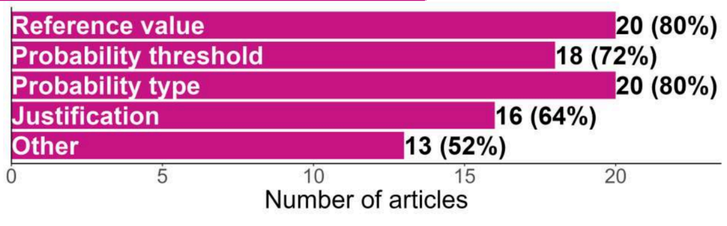


Reporting priors: 7



Thresholds: 26

Choosing thresholds: 25



Reporting thresholds: 2

- **Very limited** guidance on reporting decision thresholds

Discussion

- **More guidance available for defining priors** than other topics
- Analysis (or "fitting") priors were discussed more commonly than design (or "sampling") priors (70.1% vs 27.1%)
- **Informative priors discussed often**, with most articles describing potential information sources
- Considerable amount of guidance on how to define a prior robust to prior-data conflict, though many referred to the same method
- Sensitivity analyses regarding the prior definition were discussed fewer times than expected (33.6%)
- Most articles were marked as justification given (78.5%), which primarily meant a simulation or worked-through case study

- More guidance on analysis than design, though some were general/ambiguous
- **Only 2 articles were proposing checklists or guidelines**
- Most gave guidance on the detail about the prior required and all on the justification's detail, though to varying degrees
- Most articles did not give guidance on choosing between different terms
- Little guidance on format for reporting priors

- Some guidance available for choosing decision thresholds
- **Reference values** and **probability thresholds** were discussed in **most articles**.
- **Justification** for suggestions was **not given in all articles**
- 5 articles solely focused on the Bayes factor (classified as Other)

Conclusion / Next Steps

- This analysis has highlighted the volume (or lack) of guidance for defining and reporting priors and decision thresholds.
- More detailed analysis is underway to reveal the specific proposed guidance, enabling comparison and evaluation.

References

1. Barret L, Liaigre L, Hlavaty A, Giai J, Laporte S, Ollier E, Meyer N, Khouri C, Cucherat M, Roustit M. "Priors and decision thresholds in phase 2 and phase 3 randomized controlled trials evaluating drug efficacy using Bayesian methods: a systematic review. J Clin Epidemiol. 2025;179:111651. doi:10.1016/j.jclinepi.2024.111651

Imperial College London
imperial.ac.uk

This project is funded by the NIHR [Pre-doctoral Fellowship (NIHR304850)]. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Comparing testing routines for adaptive designs

Samuel Williams[1,2]
[1]University of Bath, [2]Phastar

Introduction

We carry out a simulation study to find the optimal testing routine in an **adaptive enrichment** setting. Whilst established **group sequential methods** are well understood[1], fully flexible testing regimes are desirable. In **adaptive designs**, these routines are less well understood.

Setting

Adaptive enrichment designs investigate whether a treatment is beneficial for only a certain **subgroup(s)** (often **bio-marker** defined). Initially, both the subgroup and the full population are tested. At an interim analysis, we can decide to continue only in the subgroup (enrich).

We consider a simple subgroup configuration:

- A biomarker positive subgroup, S_1.
- A biomarker negative subgroup, S_2.
- The full population, S_3,

Where we wish to test:

- $H_1: \theta_1 \geq 0$,
- $H_3: \theta_3 \geq 0$,

where θ_1, θ_3 are the treatment effect in biomarker positive subgroup and full population, respectively.

Combination tests

Enrichment tests are multi-stage designs with multiple hypotheses. **Closed testing procedures** (CTPs) handle multiple hypotheses and strongly control family-wise error rate, by testing intersection hypotheses[2].

Combination tests handle the multi-stage nature. These tests combine P-values obtained from each stage of a trial, even if the stages are not independent[3].

When using both a CTP and combination tests, a choice must be made regarding the order of testing.

Table 1 illustrates calculating an intersection P-value (CTP) each stage and using a combination rule to obtain a final intersection P-value.

1	Stage 1	Stage 2	Final
H_1	$P_1^{(1)}$	$P_1^{(2)}$	$W\left(P_1^{(1)}, P_1^{(2)}\right)$
H_3	$P_3^{(1)}$	$P_3^{(2)}$	$W\left(P_3^{(1)}, P_3^{(2)}\right)$
H_{13}	$P_{13}^{(1)} = S\left(P_1^{(1)}, P_3^{(1)}\right)$	$P_{13}^{(2)} = S\left(P_1^{(1)}, P_3^{(1)}\right)$	$W\left(P_{13}^{(1)}, P_{13}^{(2)}\right)$

Table 2 does not calculate an intersection P-value at each stage, only at the final analysis.

Both tables show the 'Not Enrich' scenario.

2	Stage 1	Stage 2	Final
H_1	$P_1^{(1)}$	$P_1^{(2)}$	$P_1^{(c)} = W\left(P_1^{(1)}, P_1^{(2)}\right)$
H_3	$P_3^{(1)}$	$P_3^{(2)}$	$P_3^{(c)} = W\left(P_3^{(1)}, P_3^{(2)}\right)$
H_{13}	——	——	$S\left(P_1^{(c)}, P_3^{(c)}\right)$

Here, $P_i^{(j)}$ denotes the P-value for population i from stage j, $W(\cdot,\cdot)$ denotes the **weighted inverse normal** combination test[4], and $S(\cdot,\cdot)$ denotes **Simes' method**[5] for intersection hypotheses.

Conditional error rate

An alternative approach to CTPs, is the conditional error principal[6]. It defines **conditional error rates** $A\left(P_i^{(1)}\right)$. These rates act as new 'significance' levels for the remainder of the trail – strongly controlling family-wise error rate.

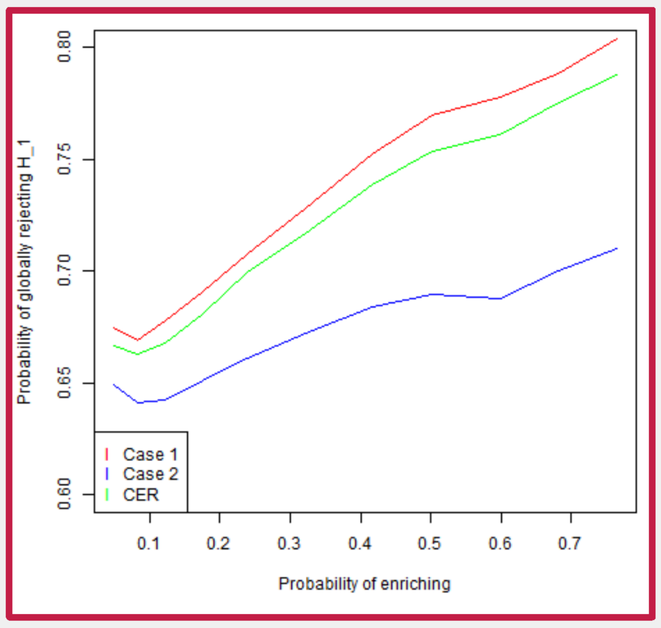
Studies have found that the conditional error rate (CER) approach leads to more powerful tests than using combination tests[7].

Simulation study

We carry out a simulation study to compare these testing regimes.

We expect that the **optimal regime** depends on how the trial progresses – we use a simple linear rule to determine enrichment at interim analysis.

Results



The approach in Table 1 yields the highest probability of rejecting subgroup efficacy at final analysis in a CTP, followed by the conditional error rate approach.

Both approaches show greater power than the Table 2 testing regime. We suspect this difference is due to consonance and conservatism arising from Simes' rule.

Future work

Further work planned for this project includes:

- More **complex subpopulation configurations** (e.g. multiple subgroups of interest, nested subpopulations, continuous biomarkers)
- Using **Bayes optimal** [enrichment] decision rules.
- Improved **computational schemes** for more extensive simulation study.

Contact

Samuel Williams

University of Bath, SAMBa

Phastar

Email: sjw203@bath.ac.uk

References

1. Jennison, C., & Turnbull, B.W. (2025). *Group Sequential and Adaptive Methods for Clinical Trials* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781584888482
2. Marcus, R., Eric, P., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, *63*(3), 655-660.
3. Bauer, P., & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 1029-1041.
4. Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). *The American Soldier: Adjustment during Army Life*. Princeton University Press.
5. Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751-754.
6. Proschan, M. A., & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, 1315-1324.
7. Mehta, C., & Liu, L. (2016). An objective re-evaluation of adaptive sample size re-estimation: commentary on 'Twenty-five years of confirmatory adaptive designs'. *Statistics in Medicine, 35*(3), 350-358.

PRO Endpoints: Power and interpretation for analysis of mean changes in continuous scores vs responder analysis of dichotomised scores

PSI P075

Rachael Lawrance, Emma C. Martin, Hannah Fairbanks
Adelphi Values Patient-Centered Outcomes, Bollington, United Kingdom.

ADELPHI VALUES | Patient-Centered Outcomes

Objective

> In the analysis of patient reported outcome (PRO) endpoints in clinical trials, it is common to evaluate **the effect of treatment in a population** by comparing differences in **mean change from baseline (CfB) scores**, evaluating if differences between the treatment group means are meaningful using Minimally Important Differences (MIDs).
> Analysis of CfB score, using the underlying continuous nature of the PRO score is a statistically robust approach to evaluate the treatment effect. However, there is concern that differences in mean scores can be difficult to interpret by patients and clinicians.
> Responder analysis, where CfB is dichotomised using meaningful within-person change (MWPC) thresholds, is recommended as being more intuitive for interpretation (or a useful aid to interpretation). However, in most situations the analysis of dichotomised data, rather than in its continuous form, results in a reduction in power, a key criticism of the responder analysis approach.
> We aimed to explore the relationship between using continuous PROs and responder analysis and the impact on power using EORTC QLQ-C30 physical functioning score in a typical phase III oncology clinical trial – ***is the loss of power enough to discount a responder analysis?***

Methods

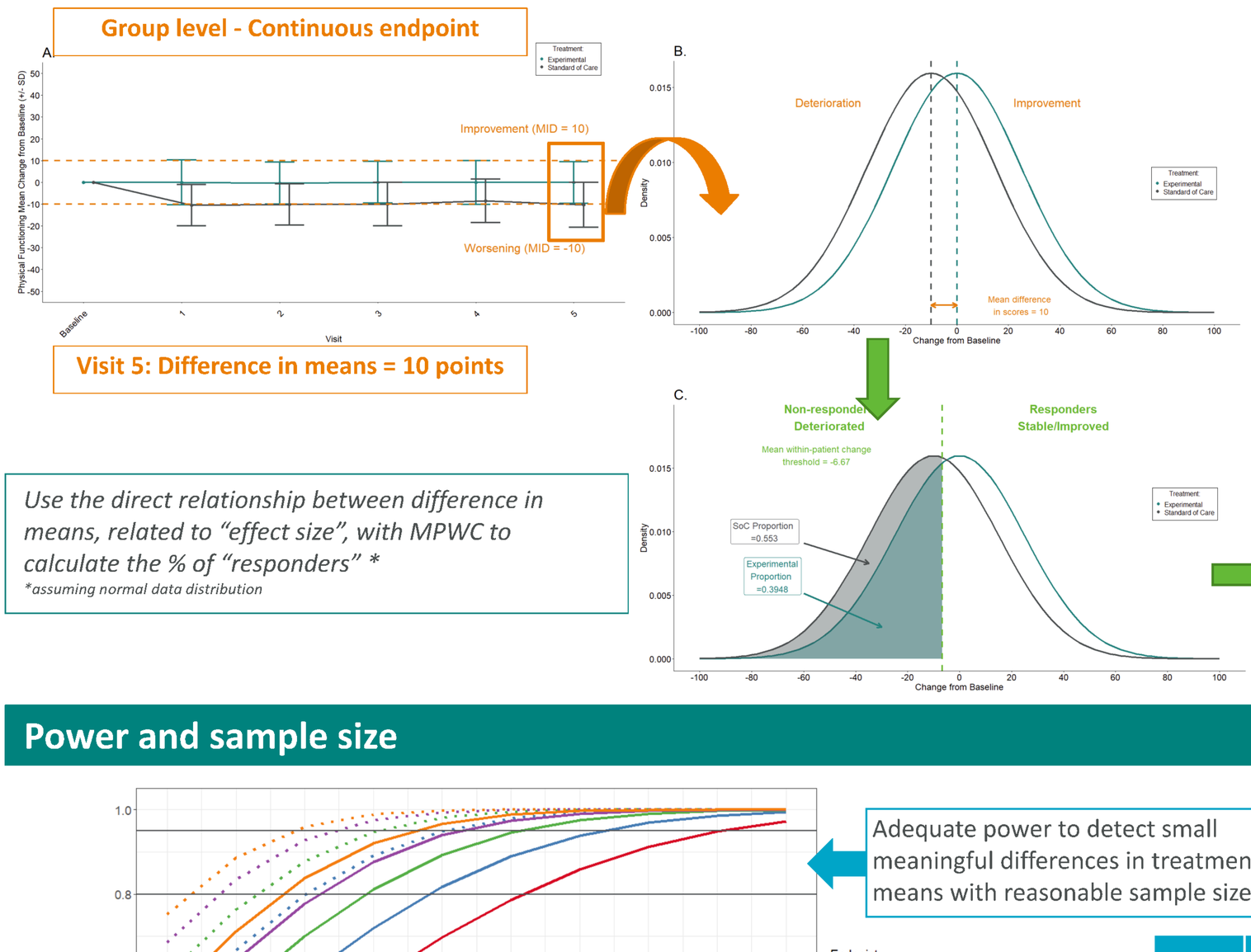
> We used EORTC QLQ C30 Physical function score (PF) range 0-100, in a 2-arm RCT
> We used literature to consider meaningful thresholds:
  > *Continuous, between-group:* MID **10 points** – see Cocks et al (a "small" effect size)
  > *Dichotomised, within-patient:* MWPC ≥ **-6.67** points for deterioration based on the **smallest possible change on this scale**
> Statistical power was calculated using t-test for means and binomial test for proportion.

Cocks et al (2011): Evidence-Based Guidelines for Determination of Sample Size and Interpretation of the EORTC-QLQ C30

Table 4. Guidelines for Size of Cross-Sectional Differences (from meta-analysis)

	Mean Difference				Effect Size†			
Subscale	Trivial	Small	Medium	Large	Trivial	Small	Medium	Large
DI	0-3	3-7	> 7	—	0-0.1	0.1-0.4	> 0.4	—
NV	0-3	3-8	8-15	> 15	0-0.2	0.2-0.5	0.5-0.8	> 0.8
CF	0-3	3-9	9-14	> 14	0-0.2	0.2-0.4	0.4-0.7	> 0.7
DY	0-4	4-9	9-15	> 15	0-0.1	0.1-0.3	0.3-0.6	> 0.6
FI	0-3	3-10	> 10	—	0-0.1	0.1-0.4	> 0.4	—
QL	0-4	4-10	10-15	> 15	0-0.2	0.2-0.4	0.4-0.6	> 0.6
SF	0-5	5-11	11-15	> 15	0-0.2	0.2-0.4	0.4-0.6	> 0.6
SL	0-4	4-13	13-24	> 24	0-0.1	0.1-0.5	0.5-1	> 1
FA	0-5	5-13	13-19	> 19	0-0.2	0.2-0.5	0.5-0.8	> 0.8
CO	0-5	5-13	13-19	> 19	0-0.2	0.2-0.5	0.5-0.8	> 0.8
PA	0-6	6-13	13-19	> 19	0-0.2	0.2-0.5	0.5-0.8	> 0.8
PF	0-5	5-14	14-22	> 22	0-0.2	0.2-0.6	0.6-1	> 1

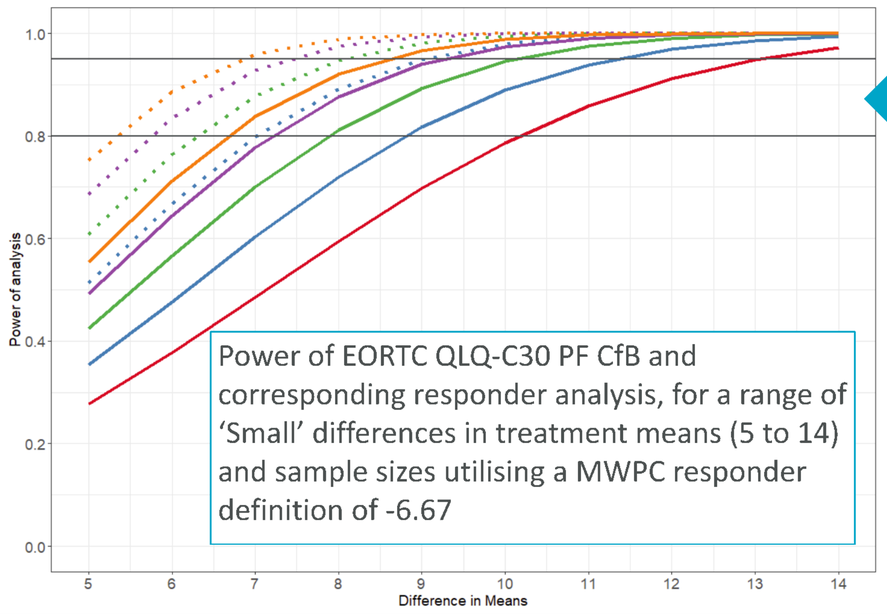
Continuous endpoint to responder analysis – explaining the relationship to MIDs/MWPC



*Use the direct relationship between difference in means, related to "effect size", with MPWC to calculate the % of "responders" **

**assuming normal data distribution*

Power and sample size



Adequate power to detect small meaningful differences in treatment means with reasonable sample size

Sample size per group	EORTC QLQ-C30 Physical function MWPC (one and two increment change)	Power: Difference in treatment means = 10	
		Responder Analysis	Continuous change from baseline analysis
200	6.67	88.9%	97.9%
	13.34	87.7%	
300	6.67	97.4%	99.8%
	13.34	96.9%	
350	6.67	98.8%	99.9%
	13.34	98.5%	

Conclusions

> There is a direct relationship between change from baseline (difference in means) and proportion of responders for PRO endpoints
> Utilisation of published MIDs and MWPC thresholds enables illustration of the expected proportion of responders
> There is a direct relationship in the power of change from baseline difference in means & proportion of responders' analyses, which is maintained across ranges of MIDs
> Power of responder analysis was lower than change from baseline difference in means analysis
> For MID of 10 points, in clinical trials over 300 patients the power of both analyses was acceptable at >80%, increasing to >95% for trials with over 500 patients
> Analysis of the proportion of responders is a valuable and interpretable PRO endpoint with high statistical power in the majority of settings in phase III clinical trials
> Analysis of both proportion of responders and CfB means are complementary approaches to holistically interpret PRO endpoints.

P076 - From Signals to Insight: Integrated Approaches for Improved IDMC Safety Evaluation

Nicolas Dubois, Senior Innovation Engineer & Senior Biostatistician - IDDI
Céline Mauquoi, Team Leader, Biostatistical Services - IDDI

Motivation

Patients' safety during trials is essential and is the Independent Data Monitoring Committees' (IDMCs) key responsibility.

This case study shows how close collaboration between multiple experts and integration of multiple datapoints addresses this mission.

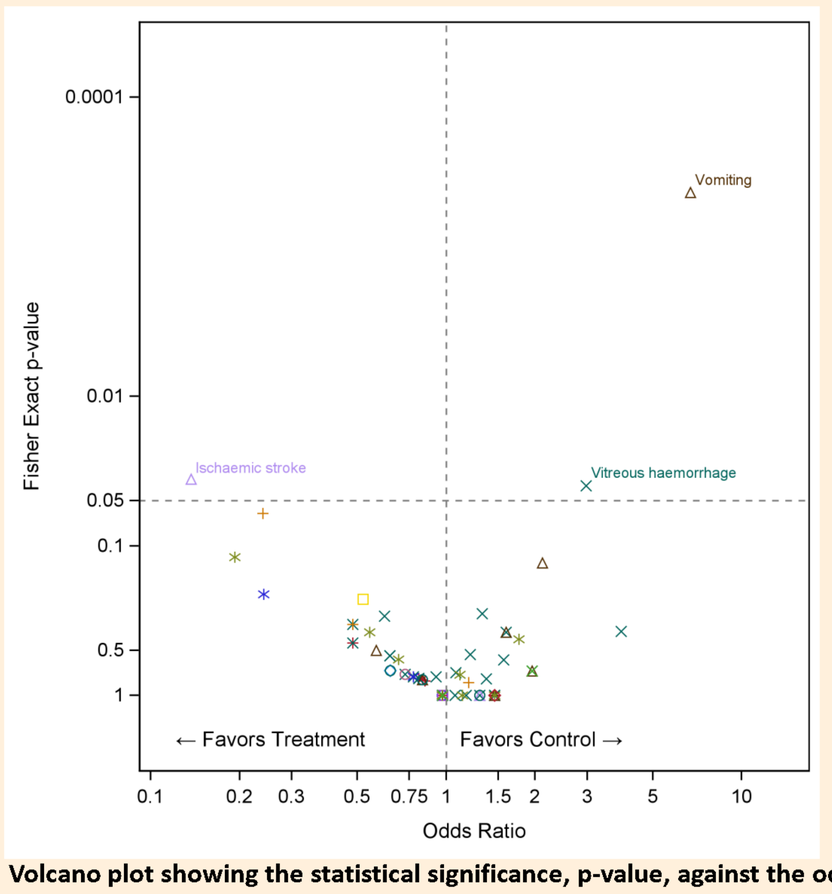
Case study

- A program composed of 4 studies investigating the same compound, in similar indications
- Two IDMCs: A in charge of 3 studies jointly, B in charge of 1 study
- Stakeholders: Statistical Data Analysis Center (SDAC), IDMC medical and biostatistical experts, Sponsor's study teams including Pharmacovigilance (PV) experts

Statistical signal detection

IDMC experts, in collaboration with the SDAC, identified safety endpoints relevant for detection of potential safety signals, such as adverse events, at individual study level.

For each study, SDAC highlighted potential meaningful imbalances by means of statistical significance testing.



Volcano plot showing the statistical significance, p-value, against the odds ratio of the event.

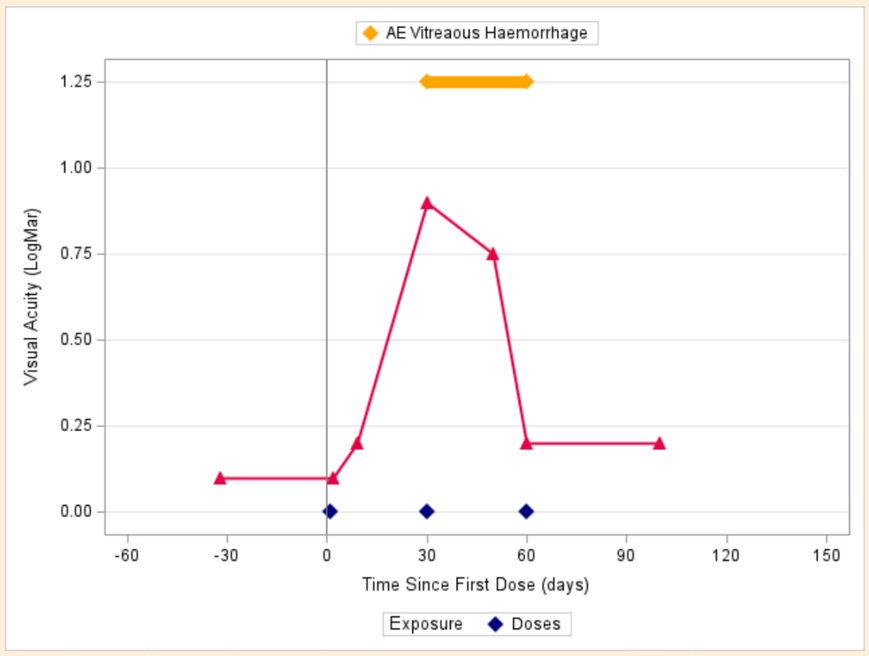
✅ Rapid identification of potential safety signals for discussion with IDMC experts

⛔ Low statistical power and multiplicity issues should be considered during interpretation

In-depth clinical exploration

The IDMC, supported by the SDAC, reviewed targeted patients presenting suspicious safety patterns via:

- Case Narratives provided by PV experts
- Subgroup analyses
- Patient-level listings and visualisations



Swimlane plot combining visual acuity, exposure and AE of interest over time.

External contextual evidence

The IDMC enriched their assessment with:

- Investigator's brochure (IB)
- Historical and real-world data
- Blinded trial and safety updates provided by Sponsor's study team

Within the joint IDMC A, a combined analysis of targeted suspicious safety patterns across sister studies evaluated the signal on a broader population. In our context, it showed that 'Vitreous haemorrhage' was specific to a single study.

IDMC A recommended to communicate selected information at study and combined level with IDMC B to support their assessment of sister study. Sponsor supported this action in a blinded manner by setting the legal documents in place, and SDAC facilitated the sharing of selected information. IDMC B included additional analyses to monitor this signal in sister study.

Insight

The combination of these approaches and the collaboration between all stakeholders strengthened the IDMCs' ability to interpret and monitor the safety signals and formulate relevant recommendations.

Related work: Poster P078 – "*Beyond Tables: Visualization as a Decision-Support Tool for IDMCs*"

Nicolas Dubois - nicolas.dubois@iddi.com
Céline Mauquoi - celine.mauquoi@iddi.com
Website: www.iddi.com
Phone: +32 10 61 44 44

IDDI
DATA WITH PURPOSE

Copyright © IDDI 2026

P077

Corinna Miede
Principal Statistician
mainanalytics GmbH
corinna.miede@mainanalytics.de

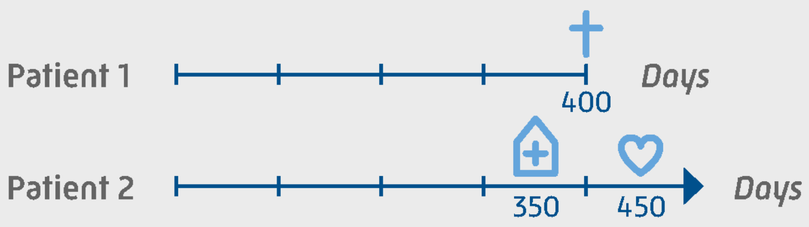
Does the Winner Take it All? The Win Ratio Method

An Alternative Approach to Analyse Composite Endpoints



Motivating Example

Traditional time to event analysis of composite event: early hospitalisation is considered worse than late death



Win Ratio Methodology

Study Setup

Group	Sample Size
Placebo	N_P patients
Active treatment	N_T patients

→ Compare each placebo patient with each patient on active treatment: $N_P \times N_T$ comparisons

Hierarchy of Outcomes

Level	Outcome
1st	Death
2nd	Hospitalisation
3rd	Symptom worsening

Comparison Algorithm

For each pair of patients $(i_{active}, j_{placebo})$:

1. **Compare outcomes sequentially according to predefined order**
   ↳ Death → Hospitalisation → Symptom worsening
2. **If comparable:**
   ├ Earlier event in placebo → Count as WIN
   ├ Earlier event in active treatment → Count as LOSS
   ↳ Stop comparison
3. **If winner cannot be determined**
   ├ Tie on this level
   ↳ Go to next level and compare
4. **If winner cannot be determined on the last level**
   ├ Tie
   ↳ Stop comparison

Calculation of the Win Ratio

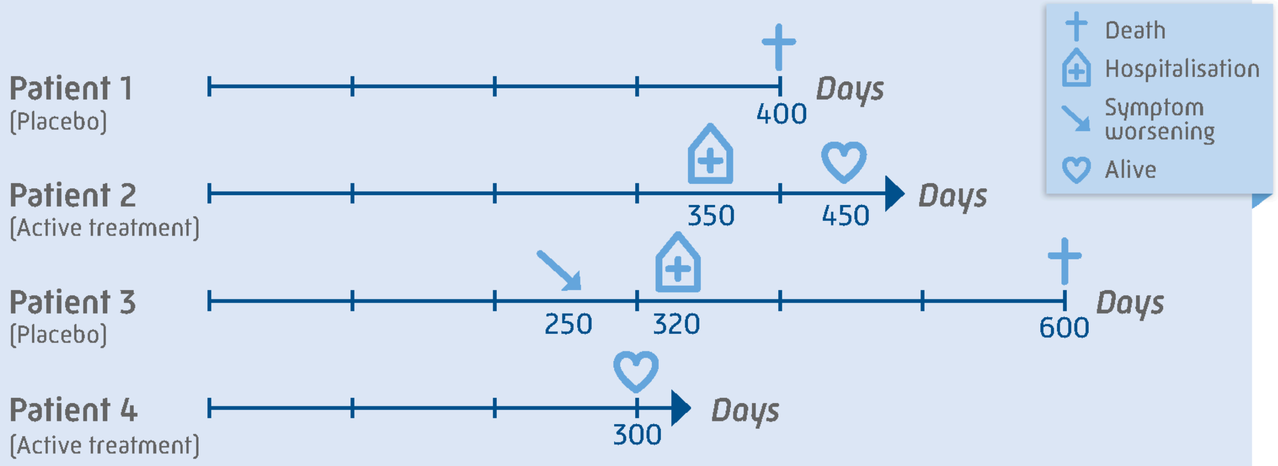
Outcome level	Winners (active treatment)	Losers (active treatment)
Death	N_a	N_b
Hospitalisation	N_c	N_d
Symptom worsening	N_e	N_f
Total	$N_w = N_a + N_c + N_e$	$N_L = N_b + N_d + N_f$

$$R_w = \frac{N_w}{N_L} = \frac{N_a + N_c + N_e}{N_b + N_d + N_f}$$

Interpretation

$R_w > 1$ Favors active treatment
$R_w = 1$ No difference
$R_w < 1$ Favors placebo

Illustration of Methodology



Patient 1 vs. 2

Level 1: Death
Patient 2 wins ★
→ Stop

Patient 1 vs. 4

Level 1: Death
Tie (no shared follow-up time for death)
→ Stop

Patient 3 vs. 4

Level 1: Death
Tie (no shared follow-up time for death)
Level 2: Hospitalisation
Tie (no shared follow-up time for hospitalisation)
Level 3: Symptom worsening
Patient 4 wins ★

Patient 2 vs. 3

Level 1: Death
Tie (no shared follow-up time for death)
Level 2: Hospitalisation
Patient 2 wins ★
→ Stop

Analysis Example

		Placebo	Active
Study population	Number of patients	50 (100.0%)	50 (100.0%)
Time-to-event analysis	Patients with event	43 (86.0%)	42 (84.0%)
	Hazard ratio (95% CI) vs. Placebo		0.67 (0.43, 1.04)
	p-value		0.0723
Win Ratio analysis	Number of pairs		2500 (100.0%)
	Pairs with event first	1445 (57.8%)	837 (33.5%)
	Ties		218 (8.7%)
	Win ratio (95% CI) vs. Placebo		1.73 (1.07, 2.78)
	p-value		0.0250

CI: confidence interval

Conclusion

Advantages	Disadvantages
Severity of events is considered beside time of occurrence: a good alternative to composite time to first event analysis	A low total number of events results in a high proportion of tied observations: limited interpretability
Quite intuitive: you can calculate the win ratio manually	
Possibility to include multiple different outcomes	

The winner doesn't take it all!



Scan for supplement

References

- Dong G, Li D, Ballerstedt S, Vandemeulebroecke M. A generalized analytic solution to the win ratio to analyze a composite endpoint considering the clinical importance order among components. Pharm Stat. 2016 Sep;15(5):430-7.
- Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J. 2012 Jan;33(2):176-82.

Layout: Franziska Kaus, KUNSTFLIEGEREI, www.kunstfliegerei.de

Confidential

P078

Beyond Tables: Visualization as a Decision-Support Tool for IDMCs

Samantha Cambier, Sr. Biostatistician, IDDI, Lava Timsina, Sr. Biostatistician, IDDI & Shweta Srikanth, Associate Biostatistician, IDDI

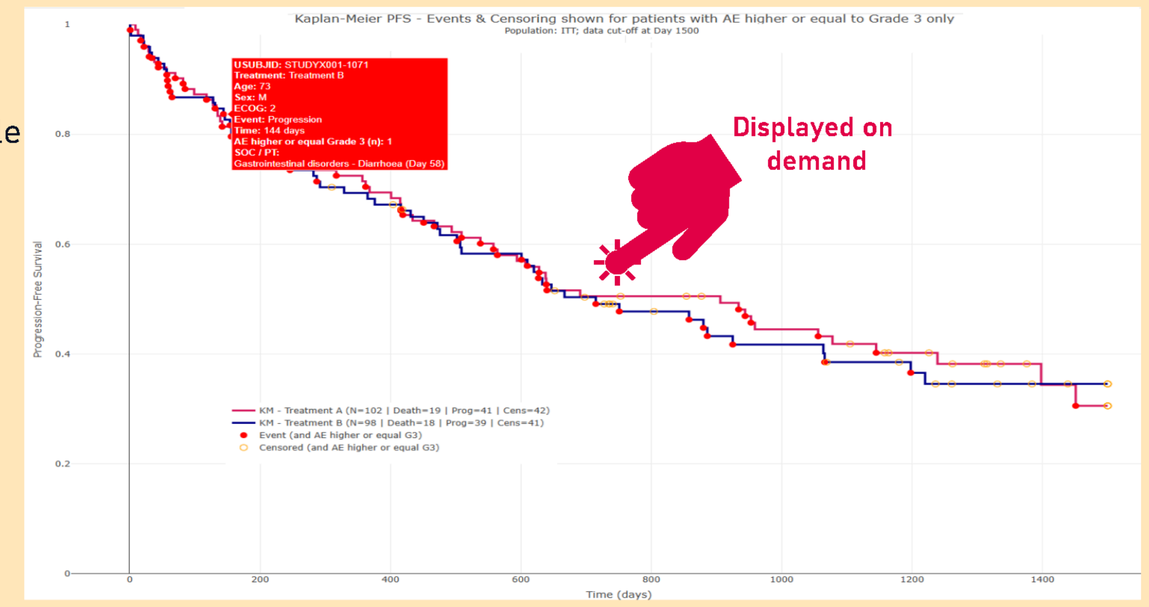
Why Visualization Matters for IDMC Reviews

- Independent Data Monitoring Committees must rapidly assess complex interim data
- Relying solely on multiple tables may slow interpretation and obscure critical signals
- Early identification of emerging risks is essential
- ➢ **Visualizations reduce cognitive burden and enable faster, more reliable decision-making**

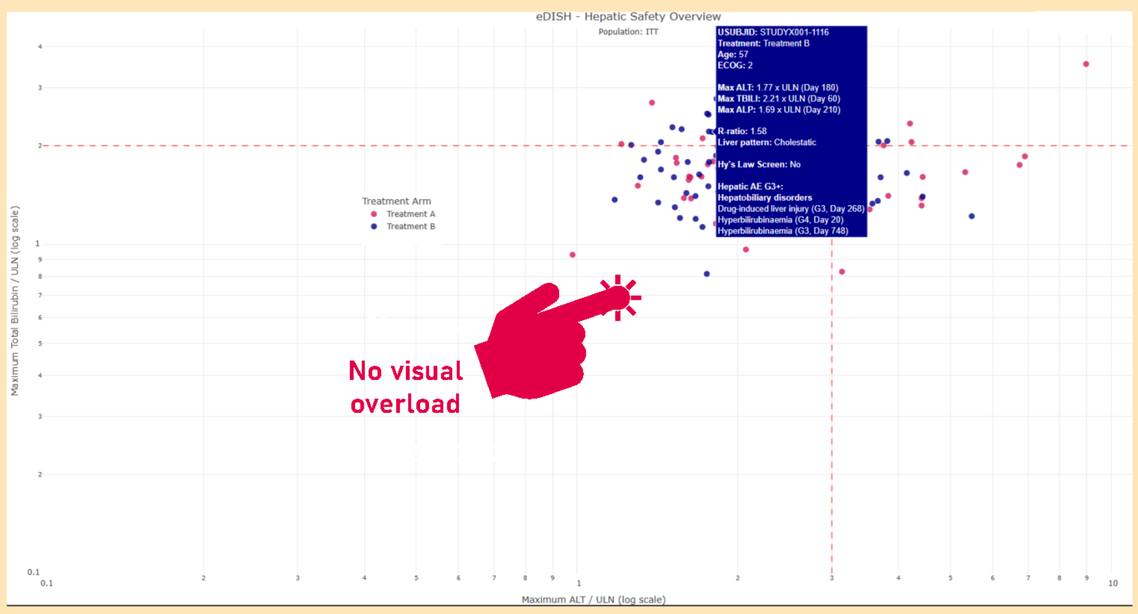
Dynamic Kaplan–Meier: Assessing Efficacy and Event Dynamics with Patient-Level Context

- Displays overall treatment effect and event timing
- Integrates efficacy, safety, and patient-level data within a single interactive visualization
- Interactive hover provides direct access to detailed patient information
- Enables linking adverse events to survival outcomes without relying on multiple tables or listings

Scan to explore the interactive visualization



Dynamic eDISH: Detecting and Characterizing Hepatic Safety Signals



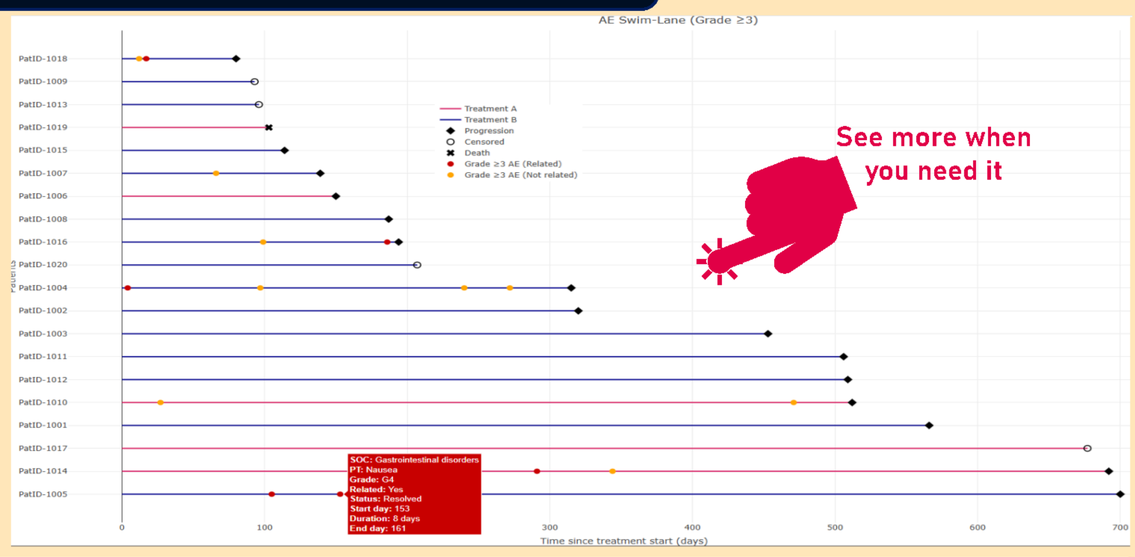
- Displays liver safety data relative to clinical thresholds
- Support detection of potential Hy's Law cases with patient-level exploration
- Interactive hover enables rapid investigation of outliers and clinically relevant cases
- Eliminates the need to cross-check multiple listings to assess the clinical relevance of safety signals

Scan to explore the interactive visualization

Dynamic Swim-Lane: Understanding Patient-Level Clinical Trajectories

- Displays patient-level timelines and clinical outcomes
- Links adverse event timing with progression and survival
- Interactive hover enables detailed patient-level exploration
- Eliminates the need to review individual patient listings

Scan to explore the interactive visualization



From Visualization to Insight

Traditional review

- Multiple tables and listings
- Fragmented safety and efficacy view
- Time-consuming cross-checks

Improve your insights

Interactive visualization

- Integrated patient-level insights
- Instant signal exploration
- Faster, more confident discussions

Samantha Cambier
samantha.cambier@iddi.com
Website: www.iddi.com
Phone: +32 10 61 44 44

IDDI
DATA WITH PURPOSE

Related work: Poster P076 – "From Signals to Insight: Integrated Approaches for Improved IDMC Safety Evaluation"

Medication adherence – a measurable indicator of childcare at home associated with post-discharge outcomes in Kenyan children

Poster Number : P079

1. Introduction

Paediatric post-discharge readmission and mortality are growing concerns in lower income countries, with up to half of deaths occurring after hospital discharge.[1,2]

Findings from the **Childhood Acute Illness and Nutrition (CHAIN)** Network[3,4] demonstrated that post discharge, caregivers face intersecting vulnerabilities:

- ❑ food insecurity
- ❑ gendered constraints on decision-making
- ❑ significant emotional and physical distress
- ❑ profound financial burden



81% of households crossed the **25% catastrophic** healthcare expenditure thresholds associated with a child's hospitalisation, with the greatest financial burden in households of children with severe acute malnutrition (SAM).[5]

Aim

To investigate prescribed medication adherence as a marker of paediatric **post-discharge mortality** and **readmission risks.**

2. Methods

This is a **secondary analysis** using data from a randomized controlled trial in hospitalised, HIV-uninfected children with SAM admitted to four hospitals in **Kenya** and followed-up for one-year.[6]

In the original trial, participants were randomly assigned to cotrimoxazole or placebo. This analysis focused on children randomized to the placebo arm only to **minimize potential bias** of prophylactic antibiotic exposure on post-discharge mortality.



Study Population

- ❑ Children aged 2-59 months
- ❑ Tested negative for HIV antibodies
- ❑ Admitted to hospital with complicated SAM
  - Mid-upper arm circumference (MUAC) <11.5cm
  - Presence of bilateral pitting oedema

The primary exposure was adherence to placebo, estimated from pill counts and calculated as a continuous percentage.

3. Statistical Analysis

In this analysis, death and readmission were analysed as separate outcomes and adherence was examined in the placebo arm only.

Participants were classified according to adherence to placebo:



≥95% adherence → YES → Adherent

≥95% adherence → NO → Non-Adherent

A multilevel mixed-effects Poisson regression analysis, with the four recruiting hospitals included as a random effect, was used to assess baseline factors associated with adherence counts.

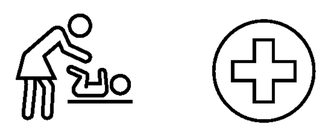
Post-discharge mortality

Cox proportional hazards regression with shared frailty by recruiting hospital were used to assess the association between adherence and post-discharge mortality. Age and sex were included as *a priori* confounders.

Base Model:

$$h_{ij}(t) = h_0(t)exp(Adherence_{ij}\beta_1 + Age_{ij}\beta_2 + Sex_{ij}\beta_3 + u_j)$$

- $h_{ij}(t)$ = **risk of post-discharge death for child (i) in hospital (j) at time t**
- $h_0(t)$ = **baseline risk at time t**
- β = **fixed-effect coefficients**
- u_j = **random effect (frailty) for hospital (j)**
- $u_j \sim$ **Gamma(θ)**



Stage 1: Run univariate models (one variable at a time)

- ❑ **Model 1: Base Model + MUAC (cm)**
- ❑ **Model 2: Base Model + Severe pneumonia at admission**
- ❑ **Repeat for all variables of interest**

Stage 2: Take forward significant variables (p-value <0.10) into a multivariable model, then eliminate using backward stepwise elimination.

Post-discharge readmissions

Necessary to account for death as a competing event and clustering of patients within four hospitals.

- ❑ **Competing risk:** Death prevents the outcome from occurring, ignoring this inflates risk estimates.
- ❑ **Robust clustering:** To account for differences between the four study hospitals.
- ❑ **Fine-Gray model:** Statistical method to correctly handle competing events
- ❑ Results presented as sub-distribution hazard ratios (sHRs).

4. Results

Among the 1781 original trial participants, **891 (50%)** were assigned placebo and included in the analysis.

557 (63%, 95% CI 50–75%) children achieved ≥95% adherence

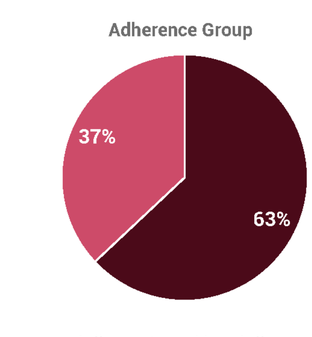


Figure 1:Summary of adherence group

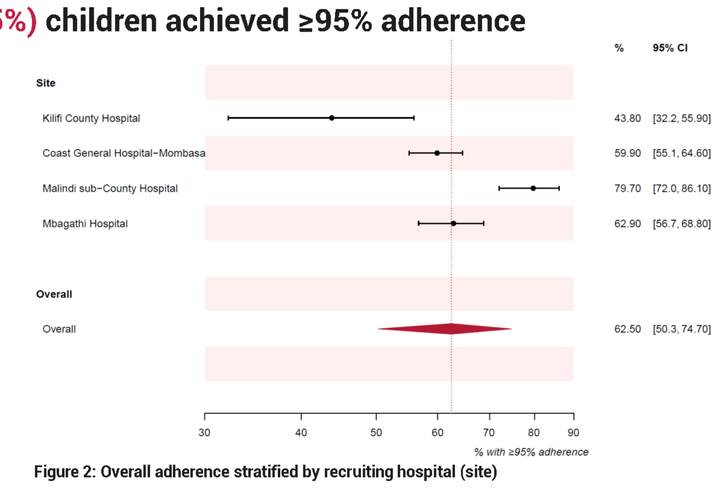


Figure 2: Overall adherence stratified by recruiting hospital (site)

Demographics	Overall (N=891)	Adherent (N=557)	Non-Adherent (N=334)
Age in months: median (IQR)	11 (7–17)	11 (7–17)	11 (7–16)
< 6 months	158 (18)	93 (17)	65 (19)
6 to 23 months	638 (71)	404 (72)	234 (70)
≥24 months	95 (11)	60 (11)	35 (11)
Sex n (%)			
Male	463 (52)	299 (54)	164 (49)
Female	428 (48)	258 (46)	170 (51)
Recruiting hospital n (%)			
Kilifi County Hospital	73 (8.19)	32 (5.75)	41 (12)
Coast General Teaching and Referral Hospital, Mombasa	424 (48)	254 (46)	170 (51)
Malindi Sub-County Hospital	138 (15)	110 (20)	28 (8.38)
Mbagathi Sub-County Hospital	256 (29)	161 (29)	95 (28)
Anthropometry & Nutritional status mean ±SD			
MUAC (cm)	10.7 ±1.09	10.7 ±1.08	10.7 ±1.10
Clinical & diagnoses at admission n (%)			
Diarrhoea	506 (57)	315 (57)	191 (57)
Severe pneumonia	325 (36)	207 (37)	118 (35)

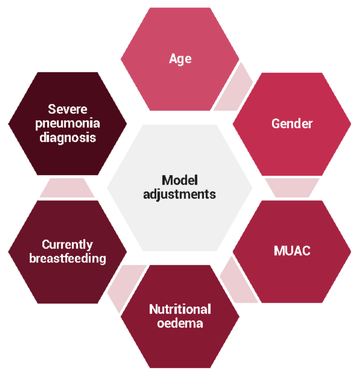
Table 1: Baseline Characteristics by Adherence group

- ❑ Median age of 11 months (IQR 7–17 months).
- ❑ The distribution of males and females was well balanced across adherence groups.
- ❑ Mean MUAC of 10.7 cm (SD 1.09): consistent with a population experiencing SAM.
- ❑ Infectious illnesses, diarrhoea and severe pneumonia were the predominant causes of admission.

Post-discharge morality

- ❑ Number of subjects who **died = 30 (3.4%)**
- ❑ Number of subjects who **withdrew prior to discharge = 8 (0.9%)**
- ❑ Number of subjects **included in this analysis: 853 (96%)**

	Adherent (N=524)	Non-Adherent (N=329)
Post-discharge		
Deaths n (%)	53 (10)	52 (16)
Mortality Rates (95% CI) per 1000 child years	112.4 (85.9 – 147.2)	190.7 (145.3 – 250.2)
Hospital readmissions	**No relationship found between adherence and hospital readmission**	



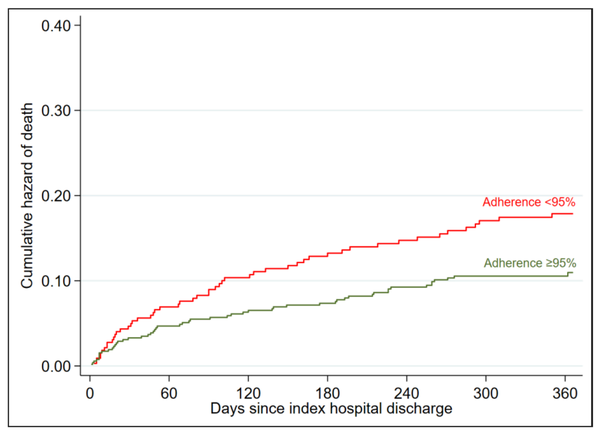


Figure 3: Cumulative hazard of death stratified by placebo adherence level

Adherence was associated with a 35% lower hazard of post-discharge death compared to non-adherence.

aHR 0.65 (95% CI 0.44–0.95)

5. Discussion

- ❑ **A need for risk-differentiated discharge strategies** that integrate social, financial and caregiver dimensions alongside clinical criteria.[7]
- ❑ **Adherence** should be viewed as a **valuable signal** for identifying children in fragile household circumstances who may benefit from enhanced support.

Contact

Dearbhála Quinn, Senior Statistician I, Phastar

Affiliations

Dearbhála Quinn[1,2], Moses Ngari[3], Elizabeth George[4], James A Berkley[3,5], Roisin Connon[4], Johnstone Thitiri[3], Judd L Walson[6]

[1]Phastar, London, United Kingdom. [2]Institute of Clinical Trials and Methodology at UCL, London, United Kingdom. [3]KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya. [4]MRC CTU at UCL, London, United Kingdom. [5]Centre for Tropical Medicine & Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom. [6]The Childhood Acute Illness and Nutrition Network, Nairobi, Kenya

References

1. Childhood Acute Illness and Nutrition (CHAIN) Network. Characterising paediatric mortality during and after acute illness in Sub-Saharan Africa and South Asia: a secondary analysis of the CHAIN cohort using a machine learning approach. *EClinicalMedicine* 2023; **57**: 101838.

2. Knappett M, Nguyen V, Chaudhry M, et al. Pediatric post-discharge mortality in resource-poor countries: a systematic review and meta-analysis. *EClinicalMedicine* 2024; **67**: 102380.

3. Elshafei A, Jemutai J, Tickell KD, et al. Depressive symptoms among mothers with hospitalized children in South Asia and sub-Saharan Africa. *J Affect Disord* 2023; **323**: 368-77.

4. Zakayo SM, Njeru RW, Sanga G, et al. Vulnerability and agency across treatment-seeking journeys for acutely ill children: how family members navigate complex healthcare before, during and after hospitalisation in a rural Kenyan setting. *Int J Equity Health* 2020; **19**(1): 136.

5. Jemutai J, Njuguna RG, Laurence YV, et al. Catastrophic expenditure associated with childhood hospitalisation for acute illness in Kenya and Uganda: a cross-sectional study. *BMJ Public Health* 2025; **3**(1): e001173.

6. Berkley JA, Ngari M, Thitiri J, et al. Daily co-trimoxazole prophylaxis to prevent mortality in children with complicated severe acute malnutrition: a multicentre, double-blind, randomised placebo-controlled trial. *Lancet Glob Health* 2016; **4**(7): e464-73.

7. Berkley JA, Walson JL, Bahl R, Rollins N. Differentiating mortality risk of individual infants and children to improve survival: opportunity for impact. *Lancet* 2024; **404**(10451): 492-4

Synthetic Patient Generation as a Framework for Robustness Assessment, Sensitivity Analysis, and Exploratory Decision Support

Paolo Messina[1], Sofia Azevedo[1] , Mario Torchia[1], Luca Emili[1]

1. InSilicoTrials Technologies S.p.A., Trieste, Italy

P080

InSilicoTrials

Introduction

Clinical research is limited by small sample sizes, restricted data access, privacy constraints, and under-representation of key patient subgroups, which restrict robust assessment of uncertainty and treatment effect heterogeneity. Synthetic Patient Generation (SPG) addresses these limitations by learning data-generating processes to create representative synthetic populations for broader analysis.

Objective

This work presents a conceptual framework for SPG in clinical research and pharmaceutical statistics, describing its role in sensitivity analysis, simulation, and robustness assessment, and discussing key advantages, limitations, and validation requirements.

SPG Supports

Privacy-preserving data sharing

Robustness assessment across multiple simulated patient populations

Data augmentation (synthetic control arm)

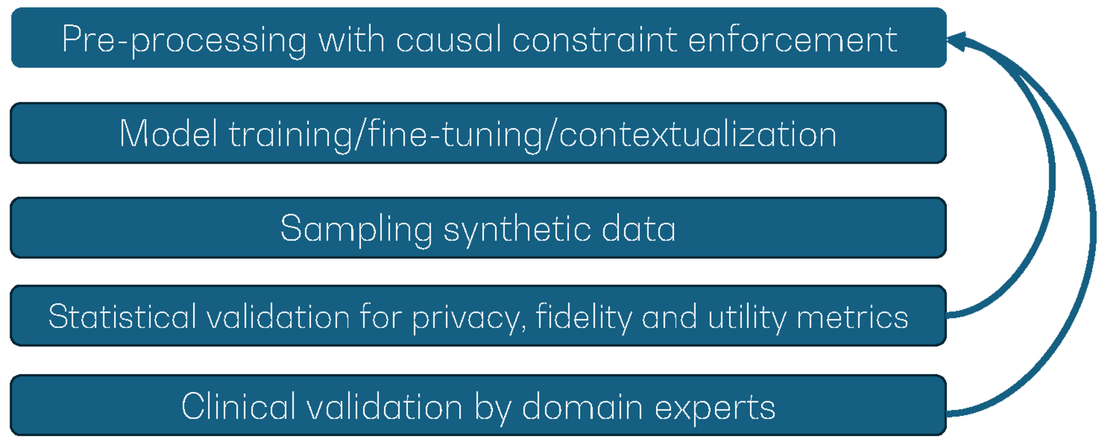
Result space exploration through "what-if" scenarios

Method

SPG learns the data-generating processes underlying observed patient cohorts and produces statistically representative synthetic populations.

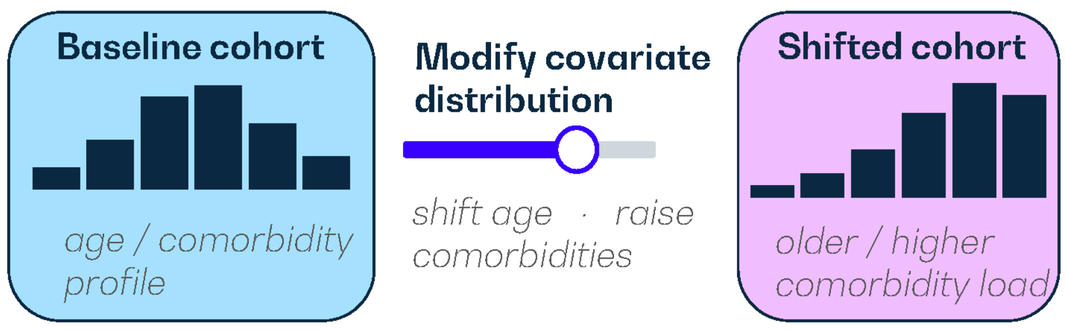
SPG enables systematic exploration of the modeled data-generating process under alternative covariate distributions.

SPG Generation Workflow



Privacy is ensured through formal frameworks limiting re-identification risk. Robustness checks generate multiple synthetic realizations to characterize treatment effect variability under the learned distribution. What-if analyses systematically vary patient characteristics to explore treatment effects across the covariate space, including enrichment of underrepresented subgroups.

SPG treats observational data as evidence for a generative model of the population (not as the population itself), enabling results to be stress-tested across plausible alternative realizations of the data-generating process.



Applications

- **Sensitivity analyses** — Simulation-based exploration under varied assumptions
- **Subgroup enrichment** — Boost under-represented patient profiles
- **Treatment effect heterogeneity** — Explore HTE across clinically relevant subpopulations
- **Feasibility & study design** — Stress-test recruitment, power, and design choices

While SPG introduces no new information beyond what is present in the training data and cannot extrapolate outside the observed covariate space, it can borrow strength across the full cohort to better characterize underrepresented subgroups, offering a principled way to explore distributional assumptions within the bounds of the data at hand.

Advantages: SPG learns a data-generating process and supports repeated sampling and what-if scenarios, extending analytical insight across the covariate space. It is useful in high-dimensional settings, where specifying and calibrating sensible joint priors over many interacting covariates (e.g., in Bayesian frameworks) becomes challenging, making it a flexible, complementary tool for decision support.

Limitations: synthetic data reflect the assumptions and information in the observed sample; broader training populations and pre-trained foundation models partially mitigate this.

Conclusion

SPG represents a flexible and complementary tool for statistical exploration and decision support, extending analytical insight without replacing real-world or trial data. While it reflects the assumptions and information present in the observed data, broader training populations and pre-trained models may partially overcome limitations of learning solely from analytical samples.

Discover

Scan to learn more about InSilicoTrials' solutions.

P081

Advancing Diversity, Equity, Inclusion & Belonging at PSI: Updates from the DEIB Advisory Group

Emma Crawford[1], Karen Smith[2], Justyna Mlynarczyk[3], Irina Garanina[4], Lily Perrin[5], Christine Fletcher[6], Ursula Becker[7], Sharon Richards[8], Lillian Ahumuza[9], Achille Dunne[10], Paolo Pio Vitiello[11], Konstantina Skaltsa[12], Agnieszka Tomczyk[13], Ana Luisa Costa[14]

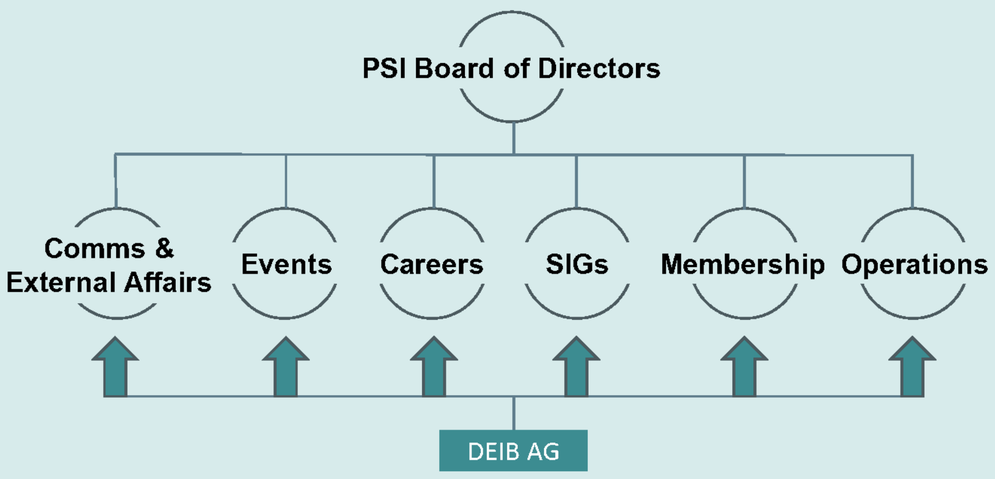
PSI's DEIB Advisory Group

Our mission is to foster a community where all members feel a sense of belonging and are empowered to contribute to PSI's success.

Key goals for 2026

- Ensure all communication channels and materials are inclusive and accessible
- Create inclusive and accessible events
- Support career development for underrepresented groups and schools outreach activities
- Increase representation of underrepresented groups at career related events
- Promote diversity of career progression paths
- Support the SIGs in integrating DEIB into SIG content and operations
- Cultivate a diverse and inclusive PSI community
- Enhance transparency and integrate DEIB into PSI operations

The DEIB Advisory Group is embedded within the PSI Committee structure



Early achievements and impact

In collaboration with PSI's Events committee:

- Improved inclusion and accessibility at the PSI Conference through
  - ✓ Provision of accessibility details on the conference website
  - ✓ Creation of a conference visual story
  - ✓ Capturing more individual requirements during the registration process
  - ✓ Introduction of a poster award focussed on content accessibility

In collaboration with PSI's Careers committee:

- ✓ Webinar on Inclusive and Impactful Conference Contributions
- ✓ Guidance to help candidates navigate interviews and request of reasonable adjustments in the recruitment process (in development, coming soon)

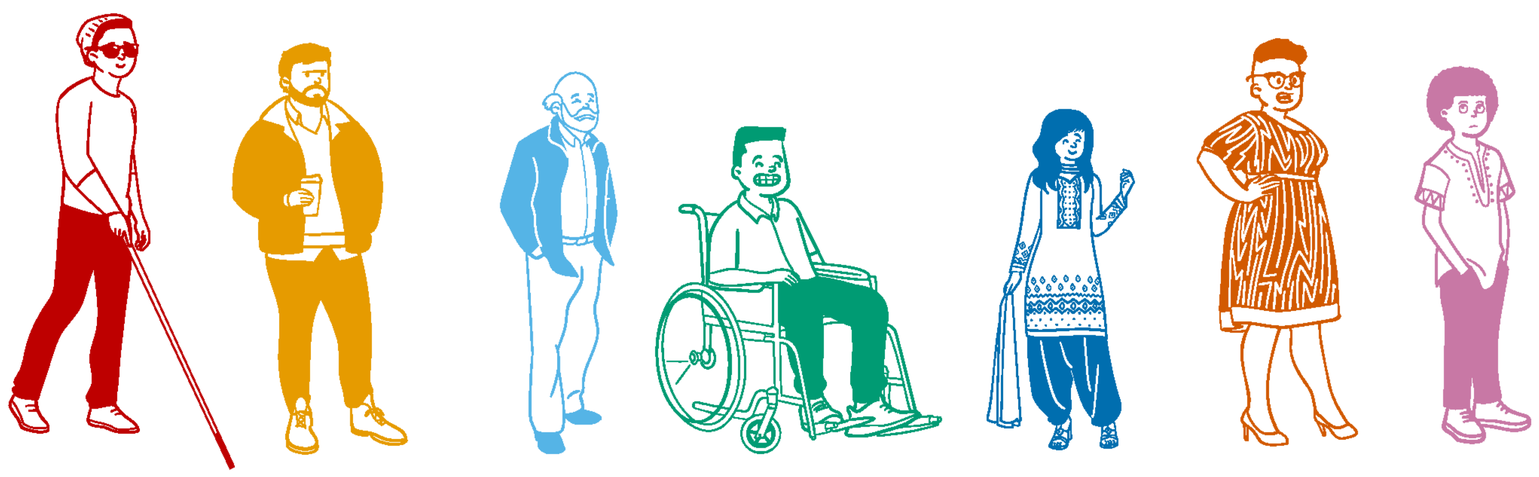
How can you help with DEIB?

Representation from a broad membership is essential. We would like to hear about any barriers to involvement in PSI's activities and what changes could be made to overcome those barriers.

Please contact us if you would like to contribute to our activities.

- Email deib@psiweb.org
- Feedback form, QR code below
- Speak to current members of the DEIB AG at conference

See updates on our website:
https://www.psiweb.org/DEIB



Scan the QR code to give your feedback

[1]PSI, London, United Kingdom. [2]UCB, Slough, United Kingdom. [3]Phastar, Glasgow, United Kingdom. [4]ICON Plc, Sheffield, United Kingdom. [5]Veramed, Manchester, United Kingdom. [6]GSK, Stevenage, United Kingdom. [7]F. Hoffmann-La Roche Ltd, Basel, Switzerland. [8]Plus-Project Ltd, Knutsford, United Kingdom. [9]Hasselt University, Diepenbeek, Belgium. [10]Merck Serono Ltd., Feltham, United Kingdom. [11]University of Bologna, Bologna, Italy. [12]IQVIA, Barcelona, Spain. [13]Parexel, Warsaw, Poland. [14]Cogitars GmbH, Heidelberg, Germany

Poster : 082

A Modelling Framework for Regulatory-Aligned Extrapolation in Pediatric Diseases

Key Message
A structured similarity assessment provides a quantitative, transparent basis for pediatric extrapolation. Adult historical evidence is combined with limited pediatric data to characterise uncertainty and inform pre-specified decision criteria.

Rejina Mariam Verghis[1], Steven Smith[2], Allison Blackorby[2], Robert Price[3]
[1]GSK, London, United Kingdom. [2]GSK, North Carolina, USA. [3]Biostatistics, GSK, Stevenage, United Kingdom

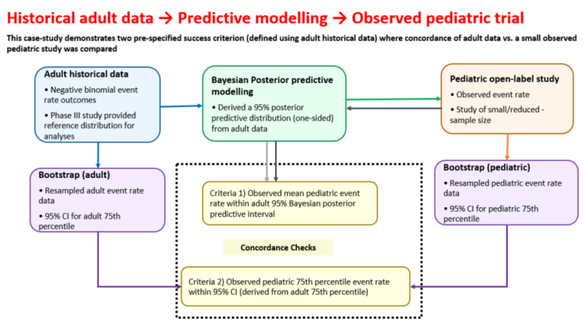
Introduction

Pediatric indications present major challenges for clinical evidence generation. Although pediatric studies are often required to meet regulatory commitments, these trials are typically small and constrained, making fully powered randomised controlled trials impractical or ethically difficult.
In this case study, aligned with ICH E11A, we proposed a transparent modelling framework for partial extrapolation that begins with an explicit similarity assessment and then quantifies uncertainty through predictive modelling. The framework can be pre-specified, is adaptable across programmes, and supports regulatory decision-making when pediatric efficacy data are limited. Broader extrapolation arguments and consistency between adults and pediatrics were also supported by an integrated evidence package covering efficacy, safety, and PK/PD for dose justification. This poster focuses on the efficacy component.
Using data from an uncontrolled open-label pediatric study, this poster presents alternative inferential approaches to robustly characterise treatment benefit, with dummy data included for illustration.

Methods

We present a modelling framework for partial extrapolation that assesses concordance between adult historical evidence and a small pediatric study.
Guided by regulatory advice, the framework starts with an explicit assessment of similarity in disease course, treatment response, exposure, PK, PD, and safety, then integrates clinical, PK/PD, and exposure–response evidence to support extrapolation.
For efficacy, two pre-specified concordance criteria were defined (see **Figure 1**):
Success criterion 1) Adult historical data were used to derive a Bayesian posterior predictive interval and compared with the observed mean pediatric event rate. Concordance was met if mean the pediatric event rate fell within the one-sided 95% posterior predictive limit (where fewer events are seen as favourable).
Success criterion 2) Adult event-rate data were bootstrapped to derive 95% confidence intervals (CIs) for multiple percentiles. Concordance was met if the observed pediatric 75th percentile fell within the 95% CI for the adult 75th percentile.

Figure 1: Extrapolation Framework



Results

<u>Success criteria 1) - Bayesian posterior predictive modelling</u>
For illustration, adult historical evidence was represented by a single phase III study (N=100) using a log-normal prior on the annualised event rate, with mean −1.42 and standard deviation 0.37 on the $\log_e$ scale, and event counts are modelled using a negative binomial distribution with overdispersion parameter k = 2.4 (mean rate =0.242 events/year).

Figure 2: Prediction Interval Limits (N=10, 20, 30, 50, 100, 200)

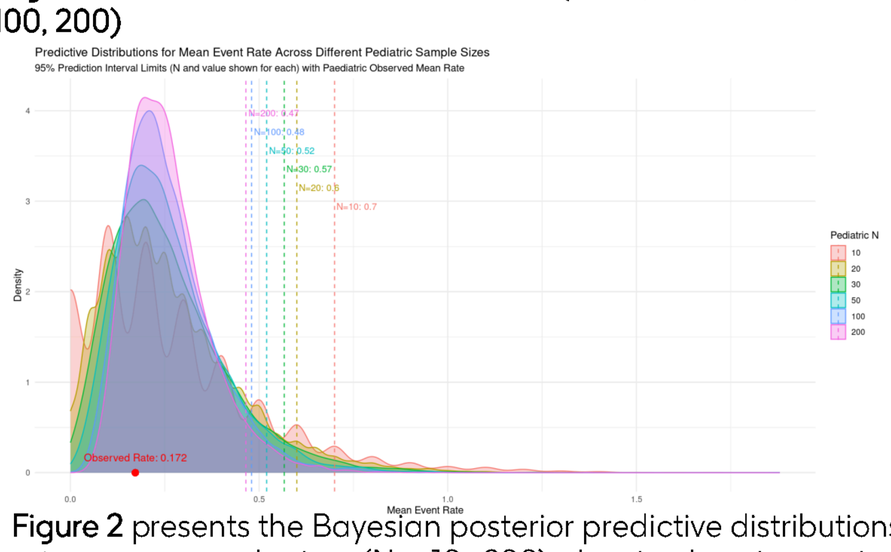


Figure 2 presents the Bayesian posterior predictive distributions of the mean pediatric event rate across sample sizes (N = 10–200), showing how increasing pediatric data (observing pediatric studies of increasing size) would reduce uncertainty and narrow the 95% predictive interval (i.e. increase the stringency of the success criteria at the given pediatric sample size).

Table 1: Observed Pediatric Event Rate and Prediction Interval Limits

Observed Event Rate (events/year) in Pediatrics	0.172
Percentile	
2.5	0.00
5	0.05
50	0.20
80	0.40
90	0.50
95	0.60
97.5	0.70

For illustration, the hypothetical observed pediatric data are represented by a single open-label study (N=20) assumed to follow a log-normal distribution (mean −1.76, SD 0.55 on the $\log_e$ scale), with event counts modelled using a negative binomial distribution with overdispersion parameter k = 2.6 (mean rate =0.172 events/year).
As shown in **Figure 2**, for N=20 the upper bound of the one-sided 95% prediction interval was 0.6. The observed mean pediatric rate was 0.172 events/year, which was below this pre-defined threshold (i.e. 0.172 < 0.60); therefore, the first success criterion was satisfied.
Table 1 shows the upper limit for a few other percentiles including 80, 90, and 97.5 (as possible alternative one-sided predictive interval widths).

<u>Success criteria 2) - Bootstrap 95% CIs</u>
Table 2 presents the annualized rate and 95% CI of different percentiles of the event rate in the pediatric and adult studies, respectively. The observed 75th percentile in pediatric study falls below that of the upper limit of the 95% CI (obtained through bootstrapping) for the adult 75th percentile (i.e., 0.000 < 1.006); and therefore, satisfied the second success criterion.

Table 2: Concordance of Percentiles (obtained from bootstrapping)

	Pediatric (N=20) Annualized Rate	Pediatric (N=20) 95% Confidence Interval	Prior Adult (N=100) Annualized Rate	Prior Adult (N=100) 95% Confidence Interval
1st	0.000	0.000 - 0.000	0.000	0.000 - 0.000
5th	0.000	0.000 - 0.000	0.000	0.000 - 0.000
10th	0.000	0.000 - 0.000	0.000	0.000 - 0.000
20th	0.000	0.000 - 0.000	0.000	0.000 - 0.000
25th	0.000	0.000 - 0.000	0.000	0.000 - 0.000
50th	0.000	0.000 - 0.000	0.000	0.000 - 0.000
75th	0.000	0.000 - 0.263	0.000	0.000 - 1.006
80th	0.000	0.000 - 1.050	0.000	0.000 - 1.231
90th	0.105	0.000 - 1.720	1.359	1.011 - 1.885
95th	1.084	0.000 - 1.720	1.885	1.355 - 1.937
99th	1.592	0.000 - 1.720	2.054	1.886 - 4.242

Conclusion

The approach in this case-study provides a transparent, quantitative framework to support extrapolation decisions consistent with ICH E11A, including similarity, uncertainty, and pre-specified decision criteria. It enables efficient pediatric development by leveraging adult evidence, it is adaptable to other Pediatric programmes (i.e. differing sample sizes or data distributions) and where randomised controlled trials or larger pediatric study are infeasible. Adoption of such approaches may consequently facilitate faster access of medicines to pediatric patients.

Abbreviations

CI – Confidence Interval
ICH – International Council of Harmonisation
PD - Pharmacodynamics
PK - Pharmacokinetics
RCT – Randomised Controlled Trial
SD – Standard Deviation

References

ICH E11A (Pediatric Extrapolation), Final Guideline, 2024

Acknowledgements

Matt Psioda, GSK, USA
Dawn Edwards,, GSK, London, UK
Jodie Crawford, GSK, London UK
Pediatric Statistics Network, GSK

Disclosures

All authors are employees of GSK and may hold stock or stock options in GSK. This work describes a methodological framework and illustrative analyses; no data from a specific GSK clinical asset or study are reported.

Presenting Author: Rejina Mariam Verghis, rejina.m.verghis@gsk.com

P083

From MTD to Optimal Dose: A Multi-Domain Quantitative Strategy for Early-Phase Oncology Trials

David Jegou*, Benoit Sansas, Elena Dupuy, Pierre Bunouf, Guillaume Desachy
Data science and Biometry department - Pierre Fabre Laboratories.

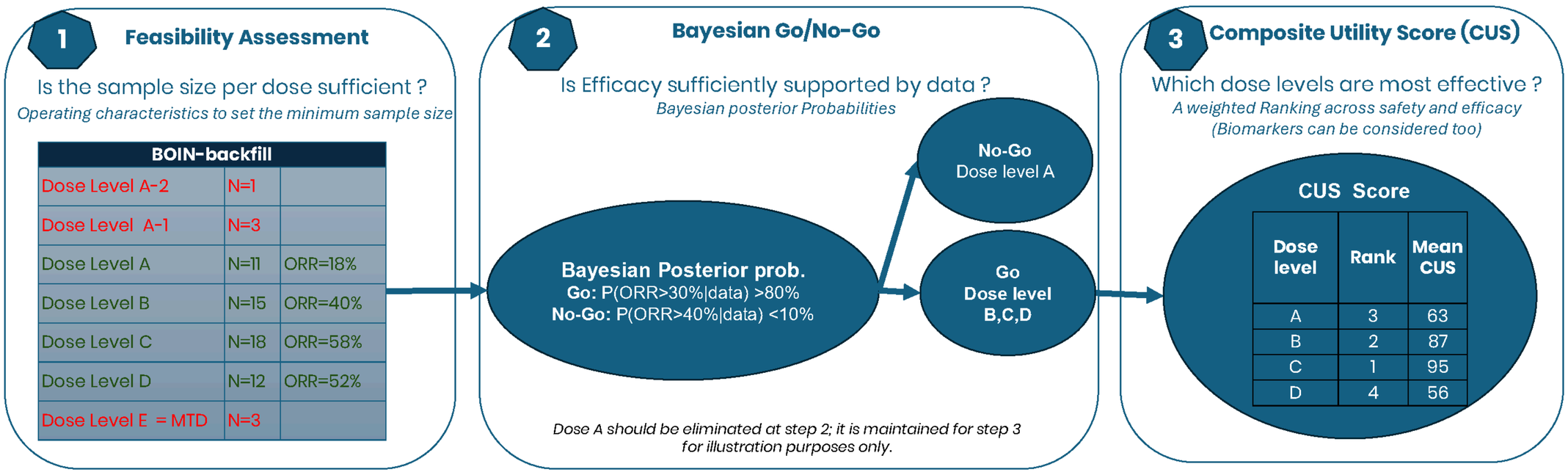
Aligned with FDA Project Optimus, we developed a flexible Quantitative Decision-Making (QDM) framework in the dose escalation phase. Its purpose is to support strategic internal decisions by integrating multiple evidence domains to select dose levels at an early stage with the highest developmental potential for dose optimization after dose escalation.

Phase I – Dose Escalation - Context

Step 1–2: BOIN-guided dose escalation (Bayesian Optimal Interval) [1] with backfill, followed by a Bayesian posterior probabilities Go/No-Go approach [2] based on efficacy confidence derived from Target Product Profile (TPP) (e.g., P[ORR > 50%]).

Step 3: CUS-based dose ranking (Clinical Utility Score) [3], integrating multiple TPP-aligned safety and efficacy criteria.

Three-Step Quantitative Optimization Framework



STEP1 and STEP2 : Data readiness is formally assessed using operating characteristics (OC), derived under prespecified efficacy scenarios. Observed response counts are mapped to Go / Consider / No-Go decision zones, quantifying decision accuracy as a function of true treatment effect and sample size. With 15 patients per dose, a strong effect (true ORR = 50%) yields ~85% probability of a Go decision (≥6/15), with minimal false No-Go risk (~2%), while a null effect (true ORR = 10%) leads to >90% probability of No-Go (<4/15), ensuring strong false-positive (<0.1%). At least >5 participants are required to achieve acceptable OCs regarding false-positive and negative rate.

Step 1 and 2 – Operating characteristics of Bayesian posterior probabilities

True treatment effect	N	Go	Consider Zone	No-Go
Strong effect : ORR =50%	15	84.9%	13.3%	1.8%
	10	62.3%	32.2%	5.5%
	5	50.0%	46.9%	3.1%
No effect: ORR=10%	15	0.2%	5.3%	94.4%
	10	<0.1%	6.9%	93%
	5	0.2%	40.1%	59%

STEP 3 – For this example, the CUS integrates 6 criteria across 2 weighted domains: Safety (60%) and Efficacy (40%). The safety domain is composed of 4 criteria; the efficacy domain is composed of 2 criteria. Assigned Weights are for illustration purposes and criteria/Weights can be modulated. 300 Monte-Carlo simulations are performed per dose with 15, 30 and 100 participants; Binomial distributions for each criteria were used and each probability is derived from a posterior distrubution under a Jeffreys prior $\theta \sim \text{Beta}(0.5, 0.5)$. A True CUS is also defined from the scenario truth using binary (0/1) criterion satisfaction.

$$\mathbf{CUS = 0.60 \times S_{safety} + 0.40 \times S_{efficacy}}$$

Step 3 - Clinical Utility Score (CUS) – Safety and efficacy driven

Domain	Domain weight	Criterion (TPP-compliant)	Intra weight
Safety	60%	Pr[(TEAE≥ Gr3+)<50%|data]*	25%
		Pr[(Diarrhea≥ Gr3+)<30%|data]	25%
		Pr[interrup<30%|data]	25%
		Pr[discont<10%|data]	25%
Efficacy	40%	Pr[ORR>30%|data]	75%
		Pr[(DOR> 6m)>50%|data]	25%

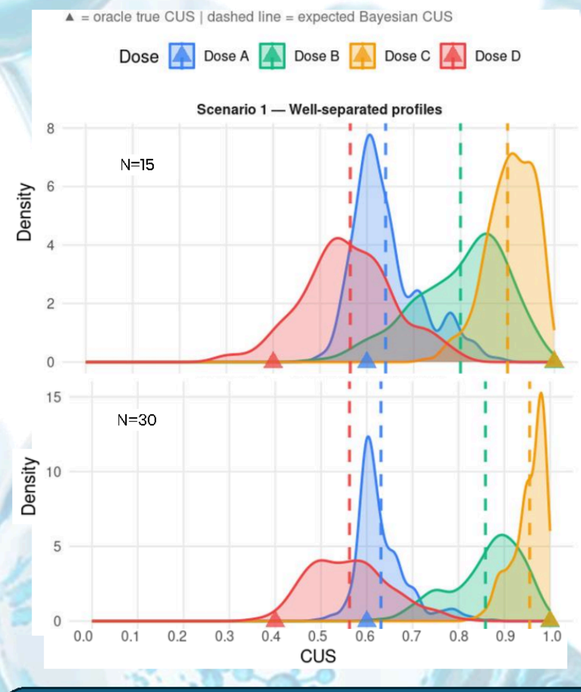
* Diarrhea Grade 3+ is excluded.

Step 3 - True parameters and simulated CUS (Mean CUS) — Scenario : Well-separated profiles –Dose Levels B & C are TPP-compliant

Dose	True Clinical parameters in %						True CUS	Simulated CUS (N)		
	TEAE≥Gr3+	Diar≥Gr3+	Interupt.	Discont.	ORR	DOR		15	30	100
A	0.12	0.06	0.06	0.02	0.18	0.28	0.60	0,64	0,63	0,61
B	0.22	0.10	0.11	0.04	0.40	0.52	1	0.80	0,86	0,93
C	0.22	0.10	0.11	0.04	0.58	0.70	1	0,90	0,96	0,99
D	0.55	0.35	0.36	0.15	0.52	0.62	0.40	0.56	0,56	0,51

In this scenario, CUS distributions are clearly separated for the 2 best dose levels (B & C) across all four doses. Dose C is the unambiguous best, combining good safety with the strongest efficacy and the method correctly identifies it in 64% and 72% of simulations 15 and 30 participants, respectively. This scenario provides a benchmark under ideal discrimination conditions. The CUS is scalable and adaptable, allowing refinement of weights and inclusion of additional criteria or dimensions (e.g. Biomarkers, PK/PD).

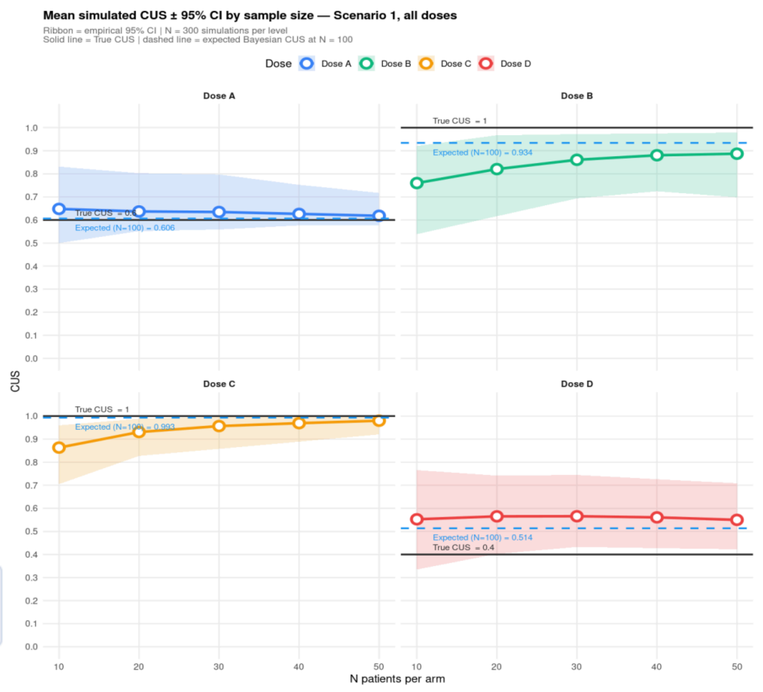
Simulated CUS distribution by dose levels



The CUS is influenced by the number of participants at each dose level. In the dose escalation phase, this 'step 3 – CUS' provides valuable information and should be validated through a comprehensive review of the full data package. This step is also applicable in phase Ib (dose optimization), where it demonstrates greater reliability with more than 20 participants.

High Ranking performance

N=15	64%	86%	97%
N=30	72%	88%	98%
N=100	86% Exact ranking	97% Best dose ranked 1st	100% Top-2 in top-2



At N = 15, the two best doses are already reliably identified, indicating that the CUS framework effectively captures the main benefit–risk signal even at small sample sizes. However, uncertainty remains in discriminating between these top candidates.

From N ≥ 30, discrimination for the top-ranked dose markedly improves, resulting in clearer separation and increased ranking stability.

These results suggest that while early decisions on the top-2 doses are feasible with limited data, a moderate sample size is required to confidently support final dose selection.

TAKE HOME MESSAGES

The approach identifies doses with the most favorable potential
and supports both TPP-based comparisons and absolute dose-level evaluation.
ORCHESTRATE DOSES EARLY THEN RANDOMIZE LATER ON THE RIGHT NOTES

1. Liu S. and Yuan, Y. (2015). Bayesian optimal interval designs for phase I clinical trials, Journal of the Royal Statistical Society: Series C , 64, 507-523.
2. Tang Q. Bayesian probability of success for go/no-go decision making. In: Bayesian Applications in Pharmaceutical Development. CRC Press; 2019.
3. Poland B. et al. (2009) The Clinical Utility Index as a practical multiattribute approach to drug development decisions.Clin Pharmacol Ther.

Copyright © 2026 Contact – *david.jegou@pierre-fabre.com

LABORATOIRES Pierre Fabre Pharma Research & Development

Model-informed Early Decision Making in Oncology: Case Study in Multiple Myeloma, when Clinical Pharmacology and Biostatistics work together.

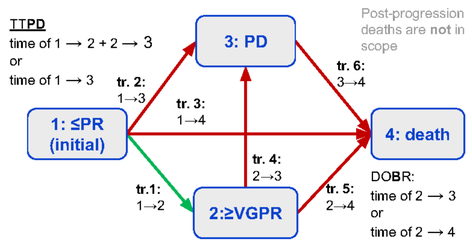
Federico Mattiello[1], Pascal Chanu[2], Mathilde Marchand[3]

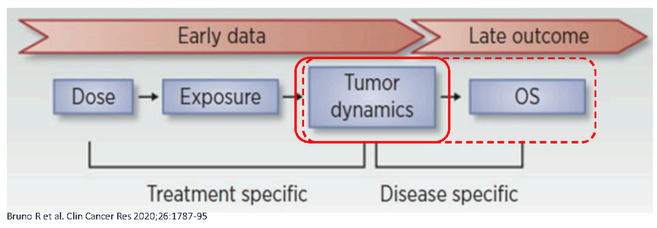
[1]F. Hoffmann-La Roche AG, Basel Switzerland – [2]Clinical Pharmacology, Genentech-Roche, Lyon, France – [3]Certara Drug Development Solutions, Paris, France

PSI Conference 2026, 15 – 17 June 2026, ICC Belfast, Northern Ireland – Poster code: P084

OBJECTIVES

- Assess operating characteristics (OC) of transitions probabilities from a Multistate Model[1] and M-protein dynamics metrics[2] in identifying effect size based on small Phase 1b/2 sample sizes with short follow-up (FU) by resampling a large positive Phase 3 study.
- Identify the right metric, target and gating criteria (probability of target achievement) to support GO/No GO decisions.
- Validation using the selected metric, target, and gating criteria performed on data from an independent trial
- Develop a PFS model to assess the probability of success of the ungated clinical trial.





BACKGROUND

Data of APOLLO[3] Phase 3 clinical trial were accessed via the Yale Open Data Access project (YODA), this trial compared daratumumab, pomalidomide, and dexamethasone (D-Pd) to pomalidomide and dexamethasone (Pd).

Patient-level tumor dynamic metrics and transition probabilities were estimated respectively to determine the effect size between D-Pd and Pd. The probability of exceeding gating criteria was computed. ROC curves visualized these probabilities, showing correct vs. incorrect go rates. Selected metrics and gating criteria were validated on CASTOR[4] trial data (also accessed via YODA, compared D-Vd vs Vd where V stands for Velcade®/bortezomib) and prospectively applied to internal Phase 1b expansion phase data vs. D-Pd, to potentially ungate future trials.

METHODS

A resampling method was applied to APOLLO data, creating 1000 datasets to assess D-Pd vs. Pd effect size under the same constraints than the "Phase 3 Go" situation (small number of patients and short follow-up).

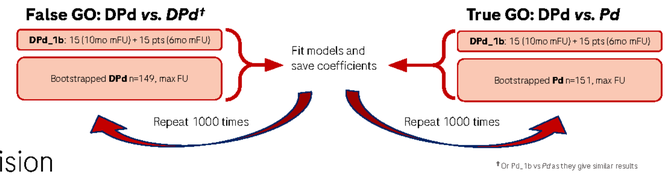
Two modeling approaches were applied on the bootstrapped datasets:

1. Tumor Growth Inhibition (TGI) non-linear mixed effect modeling (nlmxir2) on M-protein measurements, and
2. Multi-State Modeling (MSM) with 4 states: *"initial", "very good partial response", "progressive disease",* and *"death"*.

Metrics achieving good operating characteristics (True & False Positive Rate) were selected to be validated on CASTOR trial.

This cross-functional collaboration leveraged these modeling approaches and supported data interpretation, complementing the primary analysis based on ≥VGPR rate.

RESAMPLING STRATEGY



- *Phase-1b-like:* a bootstrap subsample of either an APOLLO or a CASTOR experimental arm, emulating Phase-1b expansion data that was available at decision time (6 months of minimum follow-up)
- *Bootstrap sample:* a bootstrap sample of either an APOLLO or CASTOR arm, which serves as *"historical control"* arm.
- Fit both models on the same resampled data and derive model metrics on APOLLO data, select thresholds based on ROC curves (FPR from *"DPd vs DPd"*, and TRP from *"DPd vs Pd"*) and validate them on CASTOR
- Fit the PFS model on APOLLO data, fix parametric formula, obtain bootstrap distribution (MSM) or simulate (TGI) HR on APOLLO and CASTOR data (validation step).

TGI MODEL and METRICS

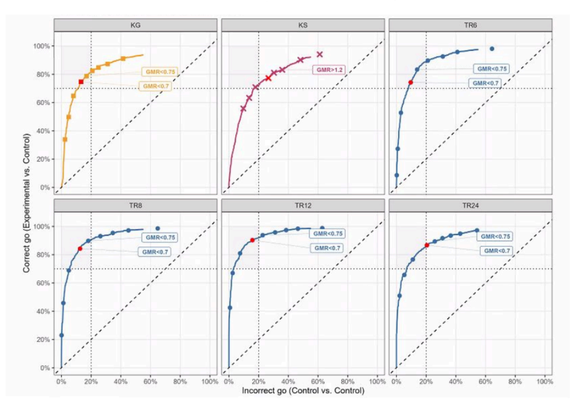
Model	Equation	Description
Stein [2]	$t < 0\ Mprot(t) = Mprot_0 \cdot e^{KG0 \cdot t}$ $t \geq 0\ Mprot(t) = Mprot_0 \cdot (e^{KG \cdot t} + e^{-KS \cdot t} - 1)$	t is the time (week) with time 0 at the start of treatment; M-prot0 in (g/L) at the start of treatment **KS, tumor shrinkage rate constant ([time]$^{-1}$)** KG0, tumor growth rate constant before treatment start ([time]$^{-1}$) KG, tumor growth rate constant ([time]$^{-1}$)

Stein: No assumption on treatment effect.
KS, KG on-treatment shrinkage and growth rates.
KG0 was not known but was estimated

APOLLO based metrics

- Grey area: ≥70% TPR and ≤20% FPR (desired OC)
- KG (growth rate constant) is a decent metric
- TR8 (M-protein ratio to baseline at week 8), TR12 are good metrics
- Probability (GMR<0.70) as gating criteria
- Validation on CASTOR trial →

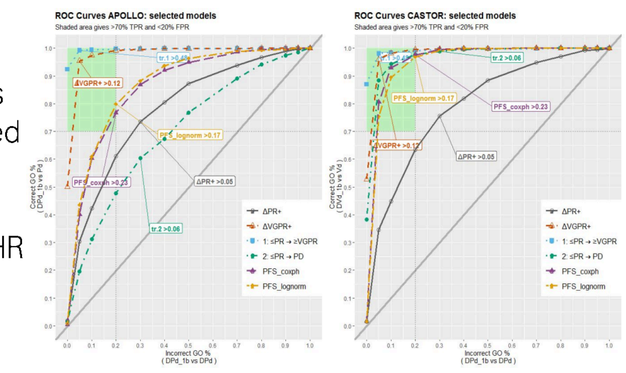
APOLLO based ROC curves: y-axis "DPd vs Pd", x-axis "Pd vs Pd"



Metric	True GO p(GMR<0.70)% Experimental (DVd) versus Control (Vd)	False GO p(GMR<0.70)% Control (Vd) versus Control (Vd)
KG	94.1	11.8
TR8	74.3	3.8
TR12	87.5	7.6

MSM METRICS

- Transition probabilities ≤PR ⟶ ≥VGPR achieved very good OCs
- Δ ≥VGPR% behave similarly but Cox PFS HR underperforms due to lack of information
- Advantage of MSM is the inherent connection between early and late endpoints
- MSM also allows to simulate patient trajectories, which help to define Ph3 treatment effect assumptions (e.g. same transitions profile to ≥ VGPR but longer time to PD, etc.).



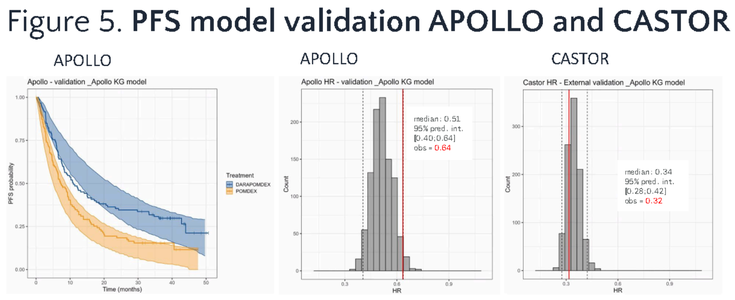
metric	gating threshold	FPR on DPd_1b vs DPd	DPd_1b vs Pd	DVd_1b vs Vd
1: ≤PR → ≥VGPR	0.478	5%	99.2%	98%
	0.305	20%	99.8%	99.1%
Δ ≥VGPR%	11.9%	5%	95.4%	96%
	4.4%	20%	99.7%	99.5%
PFS HR from Cox model	0.797	20%	76.7%	97.6%

Table 4. Parametric PFS model for APOLLO

Term	Estimate	std.error	Z-value	p-value	Lower CI	Upper CI
(Intercept)	-5.210	0.563	-9.255	<2e-16	-6.313	-4.106
logKG	-0.770	0.051	-15.226	<2e-16	-0.869	-0.671
HGB	0.193	0.041	4.742	0	0.113	0.273
Log(scale)	-0.110	0.055	-2.010	0.0444		

CI = confidence interval; HGB = hemoglobin (g/L) at baseline; logKG = log growth rate constant; Log(scale) = log-transformed scale parameter of survival model; p-values derived from Wald Z-test.

Figure 5. PFS model validation APOLLO and CASTOR



DISCUSSION / CONCLUSIONS

- These two modelling approaches can nicely supplement default Phase 3 Go/No-Go decision-making, often only reliant on early endpoints where surrogacy strength has large uncertainty. They fully exploit the longitudinal history of patients and previously collected data.
- M-protein ratio to baseline quantifies the treatment effect of early dynamics on limited Ph1b/2 data while logKG, requiring longer follow-up, is strong predictor of PFS
- Transitions probabilities from ≤PR to ≥VGPR quantifies early treatment effect while other transitions measures different important clinical events like: time to progression, duration of best response, post-progression risk of deaths.

REFERENCES

[1] A multistate model for early decision-making in oncology (https://doi.org/10.1002/bimj.201800250)
[2] Stein, W. D. et al. Clin Cancer Res 17, 907–917 (2011) (https://doi.org/10.1158/1078-0432.CCR-10-1762).
[3] https://clinicaltrials.gov/study/NCT03180736, https://doi.org/10.1016/s2352-3026(23)00218-1
[4] https://clinicaltrials.gov/study/NCT02136134, https://doi.org/10.1056/NEJMoa1606038

"This study, carried out under YODA Project 2021-4660, used data obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C.. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C.."

P085

A Simulation-Based Comparison of CART Survival Trees and Traditional Cox Models in Oncology Trial Settings

Conan Liang, Cytel, London, UK

BACKGROUD

Survival analysis plays a central role in oncology clinical trials, where time-to-event outcomes such as overall survival and progression-free survival are commonly used endpoints.

The Cox proportional hazards model is widely used because it provides interpretable hazard ratios while adjusting for multiple covariates.

However, Cox models rely on the **proportional hazards assumption** and usually require analysts to pre-specify nonlinear effects and interaction terms.

CHALLENGE

Oncology trial populations are often heterogeneous, meaning patients may differ in disease severity, biomarker status, treatment response, and survival outcomes.Complex biomarker-treatment relationships and subgroup-specific survival patterns may not be fully captured by traditional model assumptions.This **creates a need** for flexible methods that can explore nonlinear effects, threshold effects, and hidden patient subgroups.

APPROACH AND OBJECTIVE

Classification and Regression Trees (CART) provide a nonparametric survival analysis framework using recursive partitioning to divide patients into clinically meaningful subgroups.

Unlike traditional Cox models, CART **does not require** explicit specification of interactions or nonlinear functional forms and can naturally identify subgroup structures.

This poster presents **a simulation-based comparison** of CART survival trees and traditional Cox proportional hazards models using an oncology-inspired time-to-event dataset, with the aim of comparing assumptions, interpretability, and subgroup identification capability.

COX MODELS

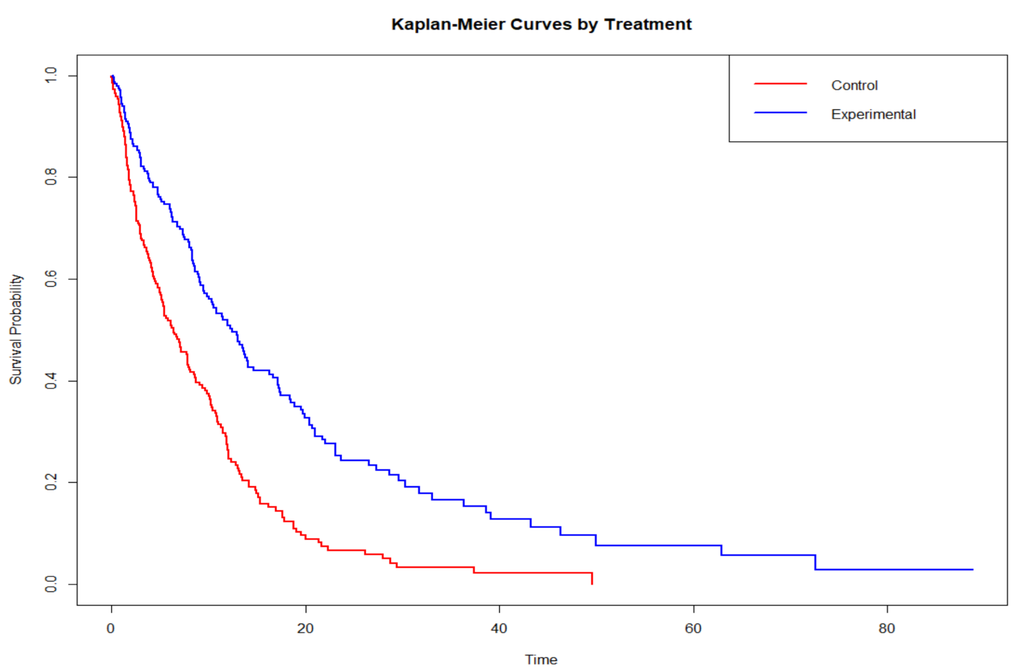
The Cox proportional hazards model was used to evaluate the association between covariates and the hazard of experiencing the event. ***For this analysis,the fitted model included:***

$$h(t) = h_0(t)exp(\beta_1 Treatment + \beta_2 Age + \beta_3 Biomarker + \beta_4 (Treatment * Biomarker))$$

Where:

- ***Treatment represented control versus experimental therapy Age was included as a continuous variable***
- ***Biomarker represented low versus high biomarker status***
- ***Treatment x Biomarker assessed subgroup specific treatment effects***

Variable	HR (exp(coef))	p-value
treatmentExperimental	0.446	<0.001
biomarkerLow	0.609	<0.001
age	1.027	<0.001
treatmentExperimental:biomarkerLow	1.211	0.378



*From the output above, the experiment treatment was associated with a **55.4% (1-44.6)** reduction in the risk of experiencing the event, and each additional year of age increasing the hazard by **2.7%.***
the kaplan meier curve ishows that blue curve remains above the red, indicating that patietns in the experimental arm had higher estimated survival probabilities.

CART SURVIVAL TREES

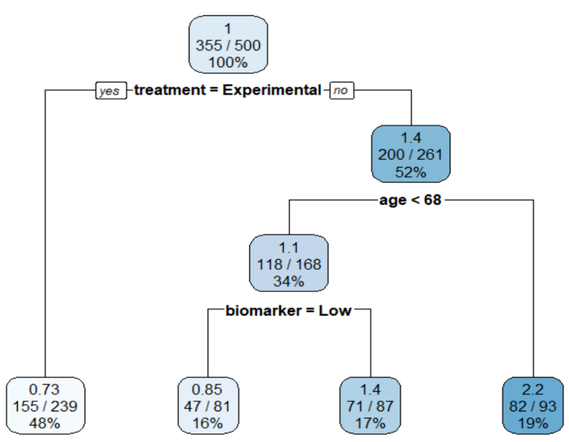
To complement the **Cox proportional hazards model**, a CART survival tree was applied to the same time to event dataset. CART stands for Classification and Regression Trees to handle **censored survival outcomes.**
The survival outcome was defined as:

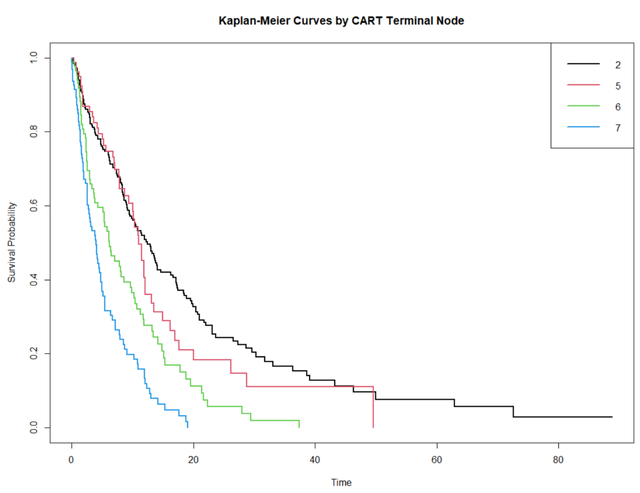
$$Surv(t, event)$$

Where time represents the observed follow-up time and event indicates **whether the event occurred (1**) or the observation was **censored (0**).
The CART survival tree model can be written conceptually as:

$$Surv(t, event) \sim Treatment + Biomarker + Age$$

Unlike the cox model CART splits patients in to subgruops by survival profiles. the algorithm searches across covarties to find the split that best separates survival outcome between groups.





Best Practice

Use Cox models for confirmatory inference and CART survival trees for exploratory subgroup analysis.

Cox Model

Best for
** Formal inference*
** Hazard ratios*
** Regulatory reporting*
Key Features
** PH assumption required*
** Covariates pre-specified*
** Produces HRs and p-values*
Limitation
** May miss nonlinear subgroup effects*

CART Survival Tree

Best for
** Subgroup discovery*
** Nonlinear relationships*
** Clinical decision paths*
Key Features
** No PH assumption required*
** Detects interactions automatically*
** Visual decision structure*
Limitation
** Can overfit unstable patterns*

Progression-Free Survival (PFS) in Oncology Clinical Trials

Regulatory Roots, Analyses and The Bigger Picture

Callum Dickson, Phastar

Why PFS?

Endpoints for later phase efficacy studies commonly evaluate whether a drug provides **clinical benefit** for patients

Time to event endpoint definition

- Time from start of observation (e.g. first dose or randomisation) until...
  - **Event of interest**
    - *Participant experiences the **event of interest**; OR*
  - **Censoring**
    - *Participant is **censored** without experiencing **event of interest** by time of analysis*
      - *Participant may withdraw from study before event occurs*
      - *Study may end before event is observed*

Generally, widespread consensus that overall survival (**OS**) is the best objective measure of **clinical benefit**

OS[4], ***Event definition**: Death from any cause*

However, progression-free survival (**PFS**) is often the primary endpoint in trials, acting as a **surrogate for OS**

PFS [4], ***Event definition**: Tumour progression or death, whichever occurs first*

PFS benefits
- ✓Amenability to earlier observation than OS
- ✓Correlation with OS
- ✓Surrogate to OS
- ✓Predictor of **clinical benefit**

PFS drawbacks
- ☒Subjectivity to **biases**
- ☒Variable definitions across studies
- ☒Inconsistent correlation with OS

Regulatory Guidance

FDA

Whether PFS is deemed as direct demonstration of clinical benefit, as a surrogate endpoint for accelerated approval or for traditional approval will depend on:

Disease context	Magnitude of effect
Disease setting	Location of metastatic sites
Available therapy	Risk-benefit relationship
Clinical consequences of...	
... delaying or preventing disease progression	
... delaying administration of more toxic therapies	

> *"We recommend assigning the progression date to the earliest time when any progression is observed **without prior missing assessments** and censoring at the date when the last radiological assessment determined a lack of progression ...*
> *The analysis plan should specify the primary analysis and **one or more sensitivity analyses** to evaluate the **robustness of the results**.*
> *Although any analyses with missing data can be problematic, the **results can be strengthened by similar results in both the primary and the sensitivity analyses**."*

EMA

> *"The range of sensitivity analyses should be sufficient to demonstrate that the trial results are robust... Any differences in conclusions from the range of analyses presented will need to be explained...*
> *Sensitivity analyses should be planned to address any important assumptions in the methods used, including handling of **deviations** and missing data, **uninformative censoring**, **proportional hazards**, handling of **unscheduled evaluations**, as applicable."*

PMDA

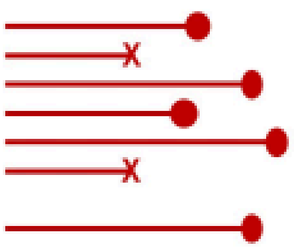
> *"The most objectively measurable and clinically meaningful endpoint is overall survival. Cancer immunotherapy characterized by delayed development of the effect may entail transient disease progression at the beginning, and thus it should be noted that progression free survival (PFS), if specified as the endpoint, may provide underestimated results on the effect of cancer immunotherapy"*

PFS Analyses

PRIMARY analysis

DATA
- TIME variable [CONTINUOUS]
- CNSR variable [BINARY (0 / 1)]

TIME (unit)	CENSOR (0 / 1)
42	0
38	1
51	0
39	0
54	0
37	1
...	...
50	0



ANALYSIS METHODS

Log Rank Test
- Suppose 2 groups ($i = 1, 2$) with survival functions: $S_1(t), S_2(t)$
- Log rank test is a non-parametric test to compare survival functions

$$H_0: S_1(t) = S_2(t)$$

$p - value < critical\ value \Rightarrow$ difference in survival functions

Cox Regression
- Let hazard $h(t)$ = instantaneous probability (risk) of event at time t
- Cox regression models $h(t)$ as function of p covariates $(X_1, \ldots, X_p)$:

$$h(t) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p),$$

where $h_0(t)$ is the baseline hazard, i.e. when $X_1 = X_2 = \ldots = X_p = 0$
- Suppose X_1 is treatment covariate with 2 levels: $X_1 = 1, 2$
- Hazard Ratio (HR) can be calculated as:

$$HR = \frac{h_2(t)}{h_1(t)} = \exp(\beta_1)$$

$$HR = \begin{cases} > 1 \text{ implies risk in group 1} > \text{risk in group 2} \\ = 1 \text{ implies risk in group 1} = \text{risk in group 2} \\ < 1 \text{ implies risk in group 1} < \text{risk in group 2} \end{cases}$$

SENSITIVITY analyses

SENSITIVITY 1

Interval Censoring
Progression events are said to occur within interval (l, u] rather than at precise timepoint
BIAS
Detection Bias

SENSITIVITY 2

Scheduled Evaluation Times
Progression events detected at unscheduled timepoints have times set according to next scheduled scan/assessment
BIAS
Detection Bias

SENSITIVITY 3

Blinded Independent Central Review (BICR)
Blinded independent review of scans to determine progression status rather than investigator review
BIAS
Assessment / Investigator Bias

SENSITIVITY 4

Exclude Deviation Participants
Participants with protocol deviations which may affect drug efficacy are excluded
BIAS
Deviation Bias

PFS →The Bigger Picture

Questioning the foundations

- Key foundational message from regulatory authorities (FDA, EMA, PMDA) is that OS is the 'gold standard' objective measure of **clinical benefit**
- It follows that the most common objective in oncology trials is to seek evidence of improvements to OS directly or via surrogate endpoints like PFS
- If we take a step back, naturally we might consider some questions about these foundations which underlie the common use of PFS as the primary endpoint for oncology trials

QUESTIONS
Clinical Benefit: quantity - and quality - of life
1) Is there always a strong correlation between OS and PFS?
2) Does increased **quantity** of life always equal improved **clinical benefit** (what about **quality** of life)?
3) **How do we** use conclusions about improvements to quantity and quality of life to **allocate limited healthcare resources** in the most efficient and impactful way?

1) Correlation between OS and PFS
- PFS only works as a surrogate to OS if PFS and OS results are reliably concordant
- However, research examples below highlight that this may not always necessarily be the case

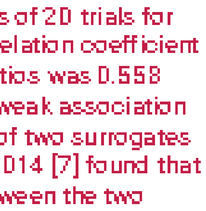
OS versus PFS – Research Examples
- ☐ Hess et al., 2019 [5] found that 31.1% of 206 treatment vs comparator comparisons exhibited an increase in OS and decrease in PFS or vice versa
- ☐ In Milrod et al., 2023 [6], an analysis of 20 trials for follicular lymphoma found the correlation coefficient between the PFS and OS hazard ratios was 0.558 (95% CI 0.545-0.571), indicating a weak association
- ☐ In their assessment the suitability of two surrogates (including PFS) to OS, Ciani et al, 2014 [7] found that the strength of the association between the two surrogates and OS was generally low

Disclaimer: the above does not comprise a comprehensive systematic review of the literature, but rather some examples to demonstrate the **key takeaway**–
- ❖ **Key takeaway**: the assumption of high correlation between OS and PFS should be scrutinised

2) What about impact on QUALITY of life?
- Intuitively, a treatment which results in 2 more years of poor-quality life might be less desirable than a treatment which results in 1.5 more years of high-quality life
- How might we objectively account for this intuition?
- *Introducing...* **QALYs**
  - **QALY** = Quality-Adjusted Life Year
- The QALY is a metric which *combines* **quantity** and **quality** into one single number
- We can define **quality** of life on a scale of 0 (death) to 1 (perfect health) – X, say
- Let us quantify **quantity** of life in years – Y, say
- Then we can calculate:

$$QALYs = X * Y$$

3) The bigger picture: healthcare resource allocation
- Oncology trials assessing PFS constitute a small slice of the bigger picture of ALL clinical trials
- Healthcare systems, meanwhile, have limited scope within the system to adopt and administer treatments across a broad range of disease settings



- *Introducing...* **HTAs**
  - **HTA** = Healthcare Technology Assessment
- HTA is a process to determine the *value* of a health technology to inform decision-making and promote equitable, efficient, and high-quality healthcare systems
- This process considers the *cost* of treatments in alongside their benefit, often using **QALYs** in so doing

Contact

Callum Dickson

Phastar

Email: callum.dickson@phastar.com

Website: https://phastar.com/

References

1. **European Medicines Agency (EMA) (2018)** Appendix 1 to the guideline on the evaluation of anticancer medicinal products in man: methodological considerations of using progression-free survival as disease-free survival in confirmatory trials
2. **Food and Drug Administration (FDA) (2018)** Clinical trial endpoints for the approval of cancer drugs and biologics: guidance for industry
3. **Food and Drug Administration (FDA) (2018)** Clinical trial endpoints for the approval of non-small cell lung cancer drugs and biologics: guidance for industry
4. **Pharmaceuticals and Medical Devices Agency (PMDA) (2015)** Guidelines on cancer immunotherapy development
5. **Hess, L.M., Brnabic, A., Slomp, G. Jr., P. and Rubin, E. (2019)** Relationship between progression-free survival and overall survival in randomized clinical trials of targeted and biologic agents in oncology. Journal of Cancer, 10(16), pp. 3674-3687
6. **Milrod, C., Kim, E.H., Ollila, T., Olszewski, A.J. and Peterdy, A. (2023)** Correlation between PFS and OS in R- implications for clinical trial endpoint selection. Blood, 142(Supplement 1), p. 4329
7. **Ciani, O., Davis, S., Tappenden, P., Garside, R., Stein, K., Cantrell, A., Saad, E.D., Buyse, M. and Taylor, R.S. (2014)** Validation of surrogate endpoints in advanced solid tumors: systematic review of statistical methods, results, and implications for policy makers. International Journal of Technology Assessment in Health Care, 30(3), pp. 212-224

Poster No: P088

How Sensitive Are Interim Decisions?

A Simulation Study of Conditional Power Thresholds in Adaptive Trial Design

Lida Fallah[1, 2] | **Parsa Mohammadian**[3]

[1]School of Mathematics & Statistics, Technological University Dublin, Ireland
[2]Sustainability and Health Research Hub, Technological University Dublin, Ireland
[3]Boehringer Ingelheim, Ingelheim, Germany

BACKGROUND

Interim analyses enable early stopping for futility or success, improving efficiency and freeing resources for more promising pathways. They are especially valuable in confirmatory and adaptive designs, where timely go/no-go decisions matter.

Conditional power (CP) is the probability of achieving statistical significance at the final analysis, given the interim data. A futility decision compares CP against a threshold c, stopping early if CP falls below c.

CP thresholds are typically chosen by **convention rather than systematic evaluation**, potentially leading to suboptimal designs. The trade-off between efficiency and statistical reliability is often under-explored.

OBJECTIVE

Systematically evaluate how CP thresholds influence:

- statistical power and type I error;
- probability of early stopping for futility;
- expected sample size (ESS).

Transparent, reproducible simulation framework.

METHODS

Trial design

- Two-arm parallel-group trial; continuous endpoint, $\sigma = 1$
- 100 patients per arm; single interim at 30% information fraction
- One-sided $\alpha = 0.025$

Interim decision rule

$$\text{STOP if } \; CP < c$$

Observed-effect CP; assumes the interim effect persists.

Simulation settings

- True effect: $\delta \in \{0, 0.2, 0.3, 0.4, 0.5\}$
- CP threshold: $c \in \{0.05, 0.20, 0.40, 0.60\}$
- 30,000 Monte Carlo replicates per scenario

Performance metrics

- Reject probability: type I error / power
- Probability of early stopping for futility
- Expected sample size (ESS)

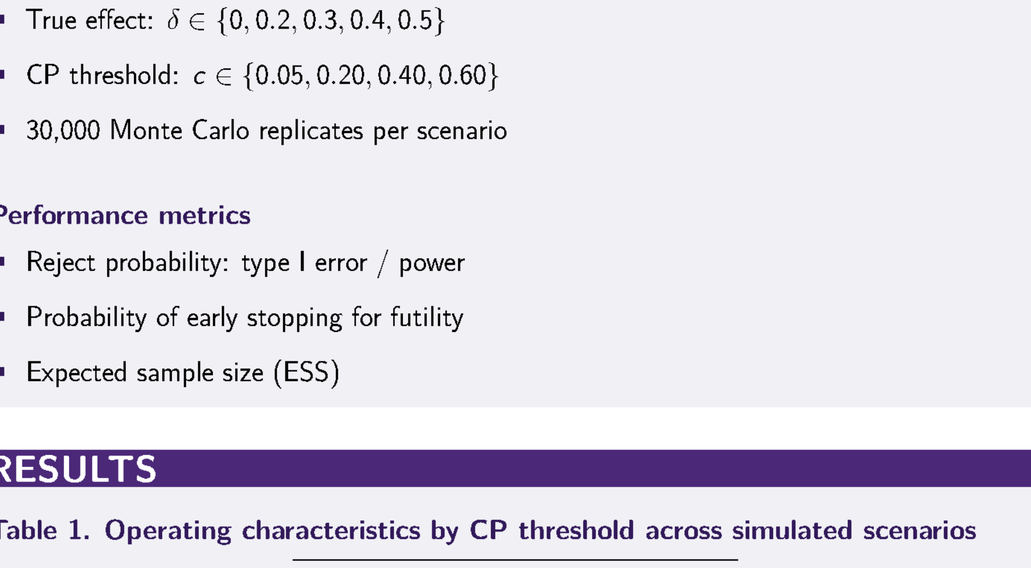
RESULTS

Table 1. Operating characteristics by CP threshold across simulated scenarios

CP Threshold	Reject Prob.	Early Stop	ESS
0.05	0.555	0.133	186.7
0.20	0.533	0.208	179.2
0.40	0.515	0.272	172.8
0.60	0.489	0.335	166.5

Observed-effect CP; 30,000 Monte Carlo replicates per scenario.

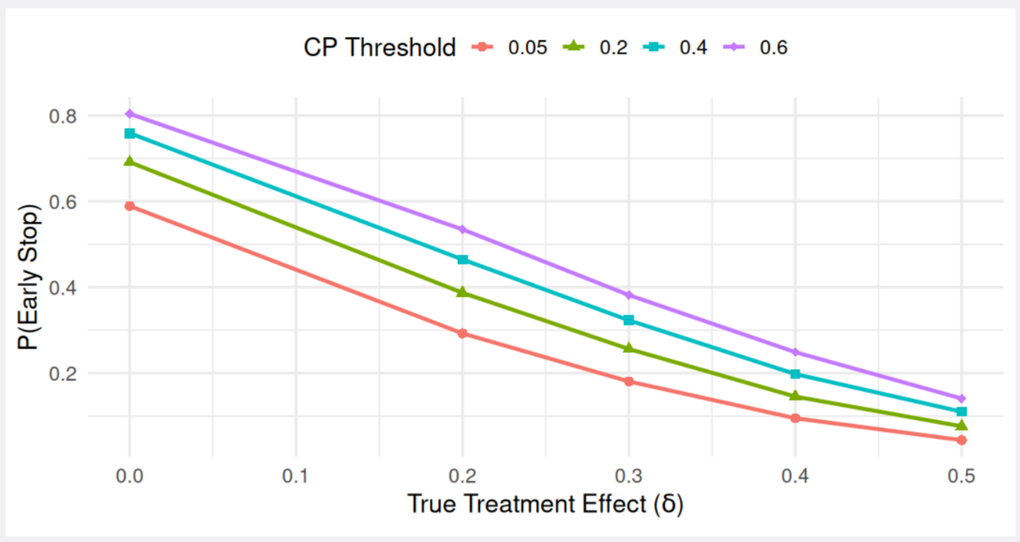
Efficiency–Power Trade-off



Higher CP thresholds substantially reduce expected sample size, but at the cost of reduced power. Small threshold changes can produce meaningful shifts in power and expected sample size.

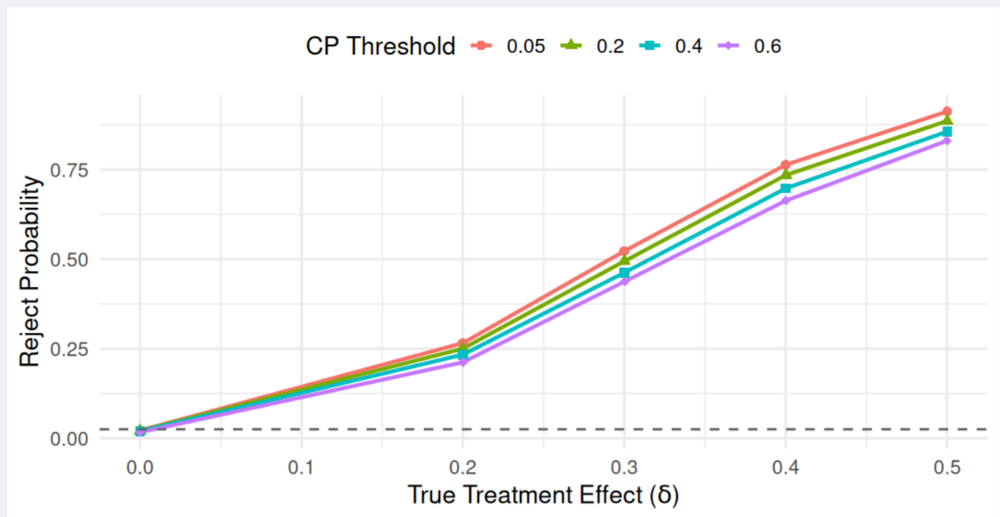
RESULTS – Ctd.

Early Stopping Behaviour



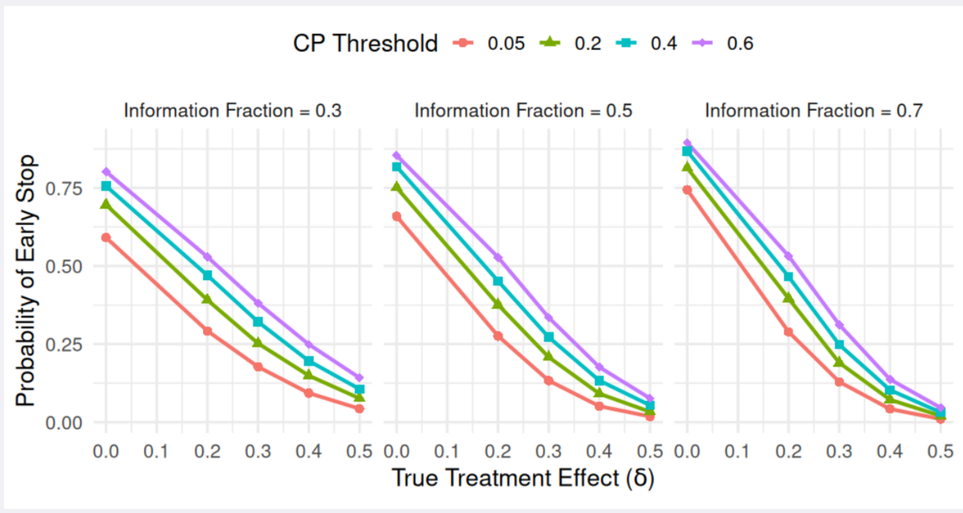
Early stopping rises sharply for smaller treatment effects and higher CP thresholds, risking premature termination of effective treatments.

Power vs Effect Size



Conservative thresholds preserve power; aggressive thresholds lose power despite improved efficiency.

Sensitivity to Interim Timing



Robust across interim timings. Earlier information fractions create greater sensitivity because interim estimates are less stable.

EXPLORE INTERACTIVELY: SHINY APP

Customise effect sizes, CP thresholds, and interim timing to explore operating characteristics for your own trial scenarios.

Scan to launch

CONCLUSION

High-impact design parameter. CP-based futility rules induce meaningful trade-offs between efficiency and statistical reliability — thresholds are not a benign tuning choice.

Small changes, large effects. Aggressive thresholds ($c \geq 0.4$) reduce sample size but increase the risk of stopping effective treatments; conservative thresholds ($c \leq 0.2$) preserve power.

Simulation-based calibration. A practical, transparent approach to selecting interim decision rules supports more robust, defensible adaptive trial designs.

REFERENCES

1. Jennison, C. and Turnbull, B.W., 2025. *Group sequential and adaptive methods for clinical trials*. Chapman and Hall/CRC.

2. Saville, B.R., Detry, M.A. and Viele, K., 2023. Conditional power: how likely is trial success? *JAMA*, 329(6), pp.508–509.

© 2026 Lida Fallah & Parsa Mohammadian — Presented at PSI 2026

The use of estimands in real world evidence studies: How useful is the Estimand Framework?

P089

Andrew Elders MSc
Affiliation: Phastar, UK

INTRODUCTION

The ICH E9 (R1) Addendum for Estimands And Sensitivity Analysis In Clinical Trials (2021), provided a framework for estimands in clinical studies.[1] Although the guideline states that the principles are also applicable for single-arm trials and observational studies, no specific consideration is given to applying the framework in real world evidence (RWE) studies. Further guidance is urgently needed, given the ongoing increase in RWE in drug development[2] and the current under-utilisation of the Estimand Framework in RWE studies.[3] This poster presents some key considerations required for developing any further guidance.

THE REAL WORLD IS COMPLICATED...

The five attributes defined in the Estimand Framework (Figure 1) are relevant when designing a traditional clinical trial or when using real world data (RWD), but additional complexities need to be considered when defining estimands in RWE studies.[4]

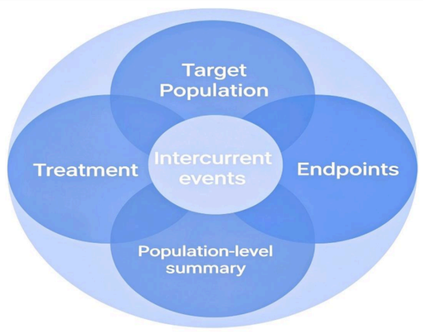


Figure 1: Attributes of the Framework

ATTRIBUTE	ADDITIONAL FACTORS TO CONSIDER IN RWE STUDIES
Target Population	More heterogenous populations are common in RWE studies due to: • less specific eligibility criteria • patient consent not always being necessary • the inclusion of under-represented groups • the ability to include patients from a broad range of sites / settings Real world data sources may be unrepresentative, eg claims databases which would exclude uninsured patients.
Treatment	In RWE studies, treatment patterns are usually more complex with lower adherence rates, as they are not controlled by a fixed protocol.
Endpoints	Assessments can be irregular and loss to follow up can be more frequent in the absence of a standardised schedule of visits. Surrogate endpoints are usually less appropriate in RWE studies.
Population-level summary	Covariate adjustment is crucial when estimating summary treatment effects and an appropriate weighting strategy should be selected.[5] Adjusted hazard ratios have limitations in causal inference but are potentially more useful in RWE studies.
Intercurrent events	Many more events could be possible (Figure 2), particularly in the absence of a randomised treatment allocation and/or a prospective study protocol.

REASONS FOR INTERCURRENT EVENTS

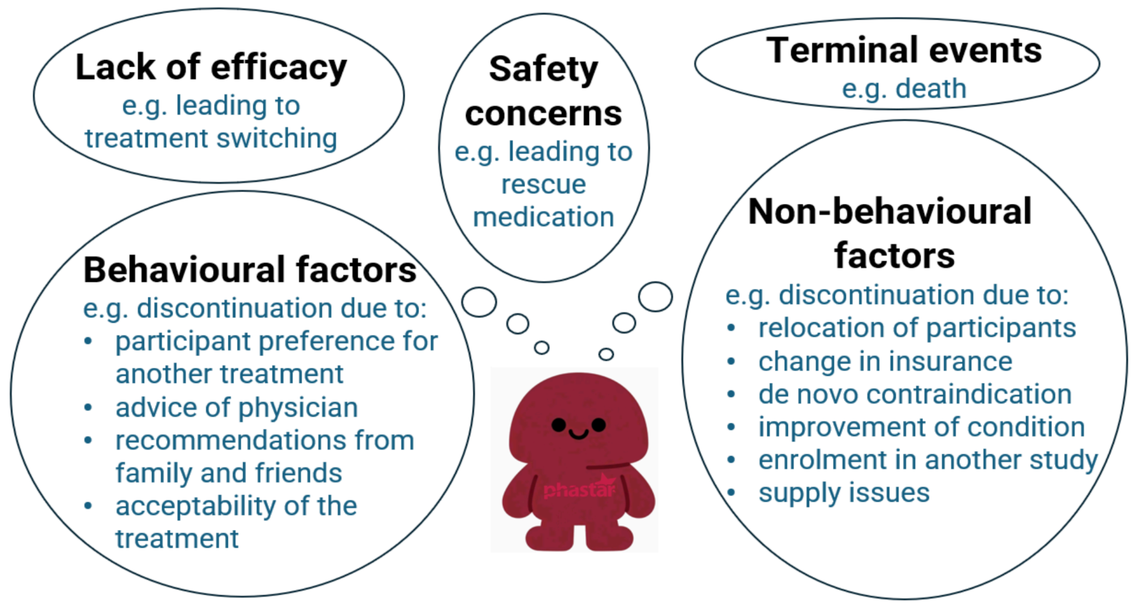


Figure 2: Researchers need to think about several factors that could potentially contribute to intercurrent events in RWE studies. A wider range of factors can be expected compared with traditional studies, although RWD will often not capture reasons for intercurrent events occurring.

STRATEGIES TO HANDLE INTERCURRENT EVENTS IN RWE STUDIES

Treatment Policy	Include outcome data regardless of intercurrent events. The estimand is the combined effect of the medical product, patient behaviour, and external factors, but it might not be a valid estimand if there are many events to disregard.
On Treatment	Use outcome data before an intercurrent event occurs. Useful if it fits with the study objectives, but it could lead to biased estimates if there are many intercurrent events due to non-behavioural factors.
Composite	Assign a particular outcome value to those experiencing an intercurrent event, eg death. It can be particularly useful in RWD if other event data have not recorded completely.
Hypothetical	Consider what outcome would have been if no intercurrent event had occurred. It may be particularly useful to estimate efficacy in a hypothetical world without administrative issues or differences in patient behaviour.
Principal Stratum	Restrict analysis to those without an intercurrent event. It is also a useful estimand if it fits with the study objectives but could be limiting if the additional intercurrent events are experienced by a large proportion of patients.

All analytical strategies should address:

- the research question
- study design
- data sources
- (and most importantly) patient behaviour[4]

OTHER FRAMEWORKS ARE AVAILABLE

Other established frameworks can be applied when considering RWE estimands, most notably the Target Trial framework[6] which seeks to emulate pragmatic trials, and the Causal Inference Roadmap[7] which has been developed to provide a systematic approach to define causal estimands in both experimental and RWE studies.

RECOMMENDATIONS

- Researchers are encouraged to use the Estimand Framework in real world evidence studies, subject to the considerations described here, with particularly the nature of the data source and the additional intercurrent events.
- Other frameworks can be utilised in situations where assumptions about intercurrent events may be too strong, and sensitivity analyses are important.
- Further guidance needs to be developed and published to extend ICH E9 (R1) to cover real world evidence, with specific focus on causal inference.

Contact

Andrew Elders
andrew.elders@phastar.com
phastar.com

Presented at the PSI 2026 Conference, Belfast

References

1. International Council for Harmonisation (ICH) E9 (R1) Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. European Medicines Agency; 2019.
2. Li, Y., Tian, Y., Pei, S., Xie, B., Xu, X., & Wang, B. (2023). Worldwide trends in registering real-world studies at ClinicalTrials. gov: A cross-sectional analysis. International Journal of General Medicine, 1123-1136.
3. Cassanelli, F., Andreis, D., Ori, A., Simoni, L., & Urbinati, D. (2025). MSR87 Estimand Framework in the Setting of Observational Studies: Current Landscape and Future Directions. Value in Health, 28(12), S510-S511.
4. Chen, J., Scharfstein, D., Wang, H., Yu, B., Song, Y., He, W.,...& Lee, H. (2024). Estimands in real-world evidence studies. Statistics in Biopharmaceutical Research, 16(2), 257-269.
5. Li, H., Wang, C., Chen, W. C., Lu, N., Song, C., Tiwari, R.,...& Yue, L. Q. (2022). Estimands in observational studies: some considerations beyond ICH E9 (R1). Pharmaceutical Statistics, 21(5), 835-844.
6. Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. American Journal of Epidemiology, 183(8), 758-764.
7. Ho, M., van der Laan, M., Lee, H., Chen, J., Lee, K., Fang, Y.,...& White, R. (2023). The current landscape in biostatistics of real-world data and evidence: causal inference frameworks for study design and analysis. Statistics in Biopharmaceutical Research, 15(1), 43-56..