DV04

A Quantitative Framework for Intermediate Endpoint Evaluation to De-Risk Clinical Development

Stefan P. Thoma & Claude Berge

Hoffmann–La Roche, Basel, Switzerland

Abstract

The development of new therapies is frequently hampered by clinical trials that are long, costly, and require large patient populations, particularly for chronic conditions with long-term endpoints like multiple sclerosis. This paper addresses the challenge of minimal savings in time and patient numbers when using only traditional endpoints, such as composite confirmed disability progression (cCDP), for interim and final readout. We propose the strategic use of an intermediate endpoint—an earlier marker that is reflective of clinical benefit—to provide confidence and inform the critical go/no-go decision to proceed to Phase III, or stop the Phase II early. To support this strategy, we are developing a unified, quantitative framework for the systematic evaluation of candidate intermediate endpoints in the form of a shiny application. This approach ensures that intermediate endpoints developed by different groups are fairly compared using a unified methodology. The ultimate value of this framework lies in increasing development efficiency by accelerating study conduct and reducing patient requirements, while simultaneously improving the probability of success in confirmatory clinical trials.

1 The Imperative for Smarter, Faster Clinical Trials

Optimizing the clinical development pathway is a strategic imperative for the pharmaceutical industry. The significant financial investment, long timelines, and considerable patient burden associated with bringing a new therapy to market demand innovative approaches to improve efficiency and increase the likelihood of success. This challenge is particularly acute in complex diseases like multiple sclerosis (MS), where trials designed to measure disease progression often require years of follow-up. A core opportunity for optimization lies in re-evaluating the endpoints used to make critical decisions earlier in a Phase II study.

When running Phase II studies, we want to be able to make data driven decisions as early as possible, see figure 1. Conducting interim analyses of Phase II studies can give us information on efficacy that we can use for decisions, such as:

- Stopping the study early, for lack of proof of efficacy.
- Setting up a Phase III study, as the interim results look really promising.

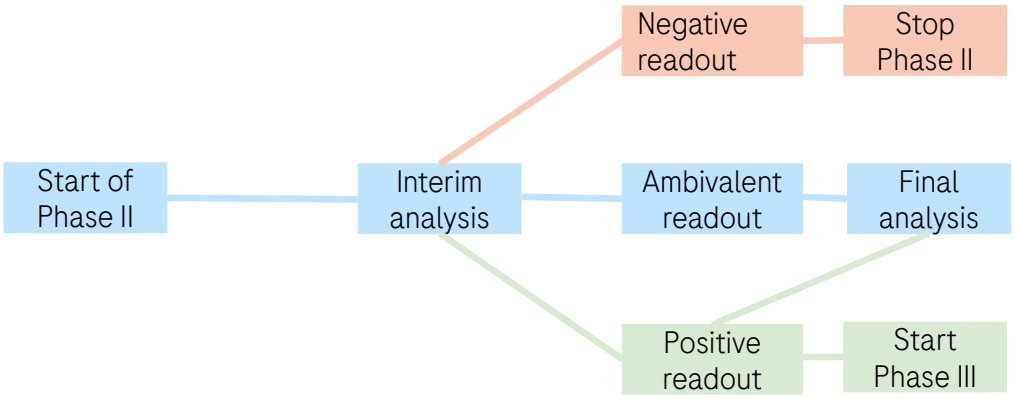

Figure 1: Phase II interim analysis decisions.

1.1 The Challenge of Traditional Endpoints in MS

In a classic clinical development plan, the same primary endpoint is often used for both the Phase II and Phase III components of a program. In multiple sclerosis, Confirmed Disability Progression (CDP) and Composite Confirmed Disability Progression (cCDP) have become standard clinical endpoints for evaluating the accumulation of disability over time. Both are accepted by regulators. For a successful Phase III study, a clinically relevant improvement must be shown on cCDP (or CDP). In this paper I will focus on cCDP, but similar issues exist for CDP as well.

cCDP measures the risk of a person's disability worsening based on any one of three key measures:

- **Confirmed Disability Progression (CDP):** An increase in a person's EDSS score that is sustained over a pre-determined time period.
- **Timed 25-Foot Walk (T25-FW):** A measure of walking speed.
- **9-Hole Peg Test (9-HPT):** A measure of hand and arm function.

While cCDP is a clinically meaningful and well-established endpoint, its use as a time-to-event measure presents significant operational challenges. Dichotomizing continuous outcomes, such as the components of cCDP, reduces the amount of information in the data [1]. Because the events (i.e., confirmed worsening of disability) accumulate slowly, gaining enough information to make good decisions based on interim results requires either a large sample of patients, or a later interim readout. This inherent characteristic makes trials costly, lengthy, and resource-intensive, creating a major bottleneck in the development pipeline.

1.2 A proposed solution: Intermediate Endpoints

To address these challenges, we propose the strategic use of intermediate endpoints (IEs) to de-risk clinical programs. An IE is defined as a marker or an endpoint that can be assessed or observed at the futility analysis. It is 'intermediate' in the sense that it is measured sometime between the initial exposure to a treatment and the ultimate clinical progression of the disease, which remains the clinical endpoint of study.

The role of the IE in this context is to be predictive of the final clinical benefit and provide evidence of a potential treatment effect, thereby building confidence to proceed to a definitive Phase III trial. Crucially, the IE does not need to satisfy the strict traditional criteria of a

true surrogate endpoint. Instead, it can serve as a "necessary but not sufficient" condition for observing a treatment effect on the final Phase III endpoint. The primary objective is to leverage an IE that is reached earlier to inform a more robust go/no-go decision after Phase II, thereby saving time and resources while increasing the overall probability of program success. To do this effectively requires a systematic and quantitative method for evaluating potential IEs.

2 A Framework for Intermediate Endpoint Evaluation

The effectiveness of an IE hinges on its relationship with both the drug's mechanism of action and the final clinical outcome. Based on our analysis, a strong candidate for an IE should possess the following key characteristics:

- **Early Drug Effect:** The treatment must demonstrate a measurable effect on the IE relatively early in the trial timeline, well before the final clinical endpoint would typically be assessed.
- **Predictive Value** The ability to detect a treatment effect on the IE must provide meaningful and reliable information about the probable effect of the drug on the final clinical endpoint at a later time point.
- **Correlation:** The IE, when measured at an earlier (interim) time point , must be statistically correlated with the clinical endpoint measured at the final readout.

2.1 Statistical Foundation

Let's call the time point of the interim analysis, t_1; the time point of the final readout, t_2.

The clinical endpoint is the time to disease worsening. The treatment effect is operationalized as the hazard ratio (HR) between the control arm and the treatment arm. The HR can be computed both at the interim analysis (HR_{t1} the reference to make an early decision) and at the final readout, HR_{t2} for determining the success of the study.

We assume that the IE is a continuous measure of the marker of interest, measured at timepoint t_1. We further assume that the variability of the IE readout is similar in the treatment and control arms, and we compute the treatment effect to be standardized, i.e. mean change between the treatment arms, divided by the standard deviation:

$$\Delta IE(t_1) = \frac{E(IE(t_1)|\text{treatment}) - E(IE_i(t_1)|\text{control})}{\sqrt{\text{var}(IE(t_1))}}$$

We can then jointly model the treatment effects $log(HR_{t2})$, $log(HR_{t1})$, and ΔIE_{t1} using equation 1.

$$\begin{pmatrix} \log(\hat{HR}_{t2}) \\ \log(\hat{HR}_{t1}) \\ \Delta\hat{IE}_{t1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \log(HR) \\ \log(HR) \\ \Delta IE_{t1} \end{pmatrix}, \begin{pmatrix} \sigma^2(\log(\hat{HR}_{t2})) & \dots & \dots \\ \text{cov}(\log(\hat{HR}_{t2})), \log(\hat{HR}_{t1})) & \sigma^2(\log(\hat{HR}_{t1})) & \dots \\ \rho_{12} \cdot \sigma(\log(\hat{HR}_{t2})) \cdot \sigma(\Delta\hat{IE}_{t1}) & \rho_{11} \cdot \sigma(\log(\hat{HR}_{t1})) \cdot \sigma(\Delta\hat{IE}_{t1}) & \sigma^2(\Delta\hat{IE}) \end{pmatrix} \right) \tag{1}$$

In the context of our framework, the relevant parameters in 1 are all specified by the user. This is relatively straightforward for the treatment effects. The variances of the treatment effects are also clear – both for the HR, and for ΔIE – and so is the covariance between HR_{t1} and

HR_{t2}. The correlation coefficients ρ_{11} and ρ_{12} are more challenging. These are the correlation coefficients that describe the relationship between the IE_{t1} and the HR_{t2} (HR_{t1}, respectively). We are mostly interested in ρ_{12}, the relationship between IE_{t1} and HR_{t2}, and consider ρ_{11} fixed. ρ_{12} is specified by the user and should ideally be estimated from data.

$$\rho_{12} = Corr(\Delta IE(t_1), HR(t_2))$$

ρ_{12} can be approximated by the patient level within-study correlation between biomarker and clinical endpoint:

$$\rho_{12} \approx Corr(IE_i(t_1), log(t_i))$$

Where t_i is the event time for patient i.

However, t_i is censored, more often than not. Therefore, deriving the correlation is not straightforward and poses one of the biggest challenges of this approach. We propose three different ways of coming up with a plausible value of ρ_{12}.

1. *No data available:* If no data from a randomized controlled trial is available to estimate the correlation ρ_{12}, a range of plausible correlation should be considered. This range should be determined collaboratively with clinicians and statisticians involvement, and should ideally cover small correlations for context.

2. *Data of one randomized controlled trial is available*: In this case, we propose bootstrapping the correlation coefficient, i.e. resample subjects from your population, and estimate both IE_{t1} and HR_{t2}. Repeat this process many times. Then compute ρ_{12} and a corresponding 95% confidence interval on the resampled estimate, then use the confidence interval as the range of plausible values for ρ_{12}.

3. *Multiple randomized controlled trials available:* Estimate the endpoints IE_{t1} and HR_{t2} separately for each study. Directly compute ρ_{12} and a corresponding 95% confidence interval on the estimates, then use the confidence interval as the range of plausible values for ρ_{12}.

2.1.1 Evaluating the IE

To get the operating characteristics required to evaluate the IE, users must specify the Null and Alternative hypothesis, and the decision criteria. *Null hypothesis H0:* The treatment has no effect on the clinical endpoint, $HR_{t2} = 1$, nor on the IE, $E(IE_i(t_1) - IE_i(t_R)|\text{treatment}) = E(IE_i(t_1) - IE_i(t_R)|\text{control})$.

Alternative hypothesis H1: There is a treatment effect on the clinical endpoint, $HR_{t2} \neq 1$, and on the IE, $E(IE_i(t_1) - IE_i(t_R)|\text{treatment}) \neq E(IE_i(t_1) - IE_i(t_R)|\text{control})$. Specifically, the treatment effects under H1 are defined by the application user.

For early decision making, users must specify the futility threshold for both ΔIE_{t1} and HR_{t1}. If the treatment effect at t_1 is weaker than the specified threshold, the drug is considered futile, and the trial should be stopped.

With the joint distribution from equation 1, one can assess various probabilities. For example the conditional power, i.e. the probability to reject H0 at the final analysis, given observed interim results of the IE, see equation2, or the probability of a successful trial, given the decision criteria specified.

$$P\left(\text{reject H1 at final} \middle| \Delta IE_{t1} = \Delta IE_{t1}^{observed}\right) \tag{2}$$

1. *Probability of a successful trial:* We can compute the probability to have a successful readout at t_2 with and without introducing a futility analysis at t_1. We can then compare

the probability of a successful readout using a specific IE (and specific decision criteria) at interim with the probability of a successful readout using the clinical endpoint HR (and specific decision criteria) at interim. This helps us evaluate the usefulness of an IE with regards to the probability of a successful readout, but also helps to power studies correctly, taking the futility analysis on respective endpoints into consideration.

2. *Conditional Power:* A measure that captures the change in certainty of success at the final readout, conditional on the observed interim result, either on HR or on IE. It helps study teams understand the impact of observing a larger or smaller than assumed effect size of the IE on the probability of a successful readout, and how this is affected by ρ_{12}.

There are additional characteristics of an IE we can compute, i.e. loss of power, but those will not be discussed in this report. For other characteristics of interest, see [2].

3 The Design of the Shiny Application

The application does not rely on data, only on model inputs specified by the user. As this application does not rely on data, using the *teal* package would not have made sense. I mention this, as *teal* is currently the standard way to create shiny applications in the Roche data science department. While the application is already being used in clinical development, it is also still a work in progress.

3.1 User Input

Inputs are structured as: *Study design*, including number of subjects, recruitment, time of final readout, and time of the interim analysis. *Hypotheses*, where you specify the expected treatment effects on an event based clinical endpoint at the final readout, the expected treatment effect on the continuous IE at the interim analysis, and their correlation. *Decision Criteria*, including the type-1 error at the interim analysis and the final analysis, and futility thresholds on the clinical endpoint and on ΔIE.

This structure allows for a straightforward specification of trial parameters. To reduce operating and programming complexity, the inputs are limited to the most important aspects of the trial design. While the application does not allow users to specify every aspect of a potential trial, the specification options are flexible enough to generate insights on the potential usability of an IE.

The outputs do not rely on heavy simulations. Still, producing the results can take a couple of seconds. To let the user know that computations are happening, loading animations have been added. This is a very basic way of making the application more usable, and (thanks to *shinycssloaders*) takes no effort at all. Additionally, some intermediate results that require less computation are displayed instantly. This includes the amount of information available at interim analysis given the study design parameters – a good sanity check for users when specifying their inputs.

The computations are entirely deterministic. Rerunning the same specifications yield the same results. To increase the ease of sharing and reproducing results, we have added a specification download/upload button. Once study parameters are configured, the downloaded specifications can be shared among collaborators, and they can easily be stored for later use.

3.2 Outputs and Results

The results are structured in different tabs, visually representing different operating characteristics. Generally, we contrast the results of using the IE at the interim analysis with using the primary clinical endpoint at the interim analysis, where the clinical endpoint at interim analysis serves as a baseline for reference.

3.2.1 Futility Analysis

In figure 2 the black line represents the study power without any futility analysis. The orange line represents the probability of a successful final readout for a study that is also conducting a futility analysis using the clinical endpoint and a user specified futility threshold for the clinical endpoint. The blue band represents the probability of a successful final readout for a study that is conducting a futility analysis on the IE treatment effect. The width of the band represents plausible values for the correlation between the interim treatment effect on the IE and the final readout on the clinical endpoint. The figure clearly visualizes the treatment effect size requirement on the IE to be useful as an intermediate endpoint in this study setup.

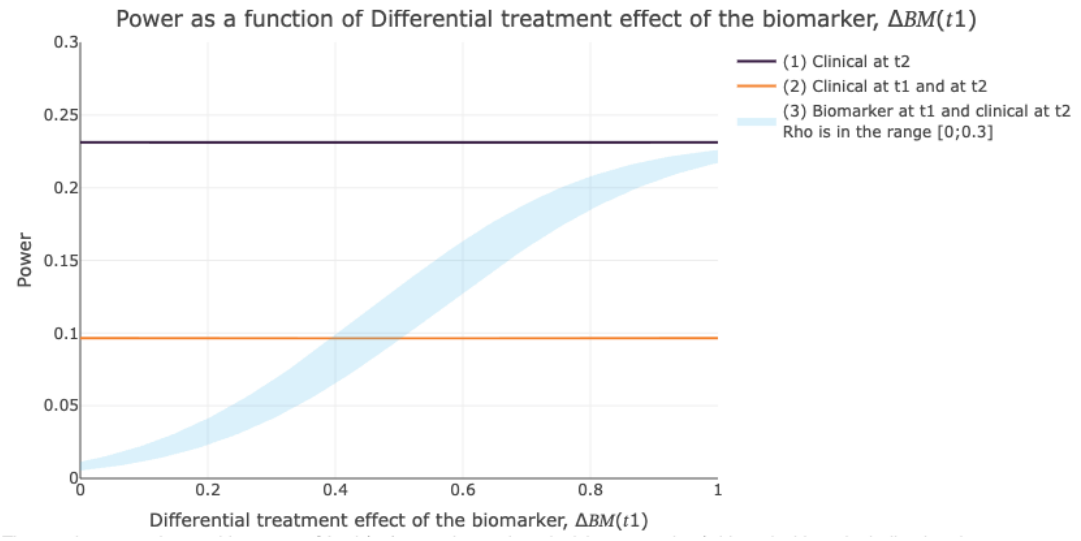

Figure 2: Power as a function of the IE treatment effect at the interim analysis.

For figure 3, the black and orange line represent the same as for figure 2 above. The blue line now represents a fixed, user specified, treatment effect on the IE at the interim analysis. This line highlights how (for a fixed treatment effect on IE) the benefit of using the IE is dependent on the correlation between the interim treatment effect of the IE and the final readout on the clinical endpoint. The blue shaded area represents the plausible (user specified) correlation values. In the example shown here, with an expected treatment effect of 0.2, using the IE at futility, is even worse than using the clinical endpoint itself, both in the blue shaded area of plausible correlations, but even for very high correlations. Hence, such an IE would not be sufficient.

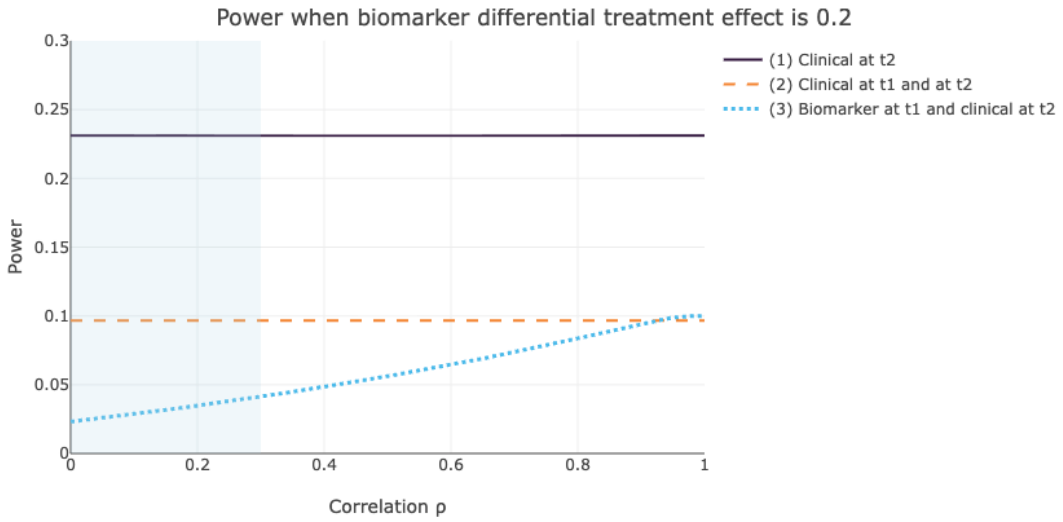

Figure 3: Power as a function of the correlation between the IE treatment effect and the clinical endpoint at the final readout.

3.2.2 Conditional Power

The green lines in figure 4 show the conditional power (positive final readout) given an observed effect size on the IE for the upper and lower limits of plausible correlations. This assumes that the true treatment effect on the IE at the interim analysis is as the user specified. The flat light green line shows that if there is no correlation between the treatment effect on the IE at the interim analysis, the conditional power is independent on the observed effect on the IE at the interim analysis. This makes sense: You cannot gain knowledge on the clinical endpoint by observing an endpoint (at interim) that is not correlated with the clinical endpoint. If the endpoints *are* correlated, as seen on the bright green line, the conditional power increases when observing a larger treatment effect on the IE at the interim analysis. The type-I error rises as well with an increased observed effect on IE, keep in mind though that the type-I error assumes that there is no treatment effect, observing large treatment effects on IE is therefore very unlikely.

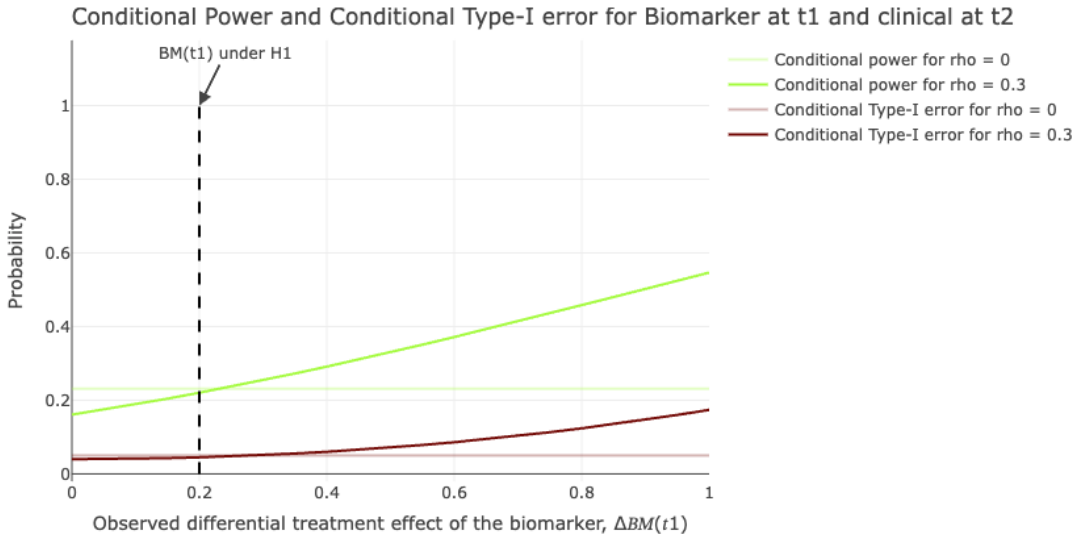

Figure 4: Conditional power given an observed IE at the interim readout

Figure 5 shows the same for observed effects on the clinical endpoint at the interim analysis. Since observing the clinical endpoint at the final readout uses all data on the clinical endpoint evaluated at the interim readout, the two endpoints are directly linked. Therefore, a larger observed treatment effect on the clinical endpoint at the interim analysis will always increase the conditional power at the final readout. Conditional power even reaches a probability of 1 on the right hand of the plot, but this is misleading: The observed treatment effect on the clinical endpoint is computed as a hazard ratio, the x-axis should really be on a logarithmic scale. What we see on the right hand side of the plot is an infinitely large observed effect size.

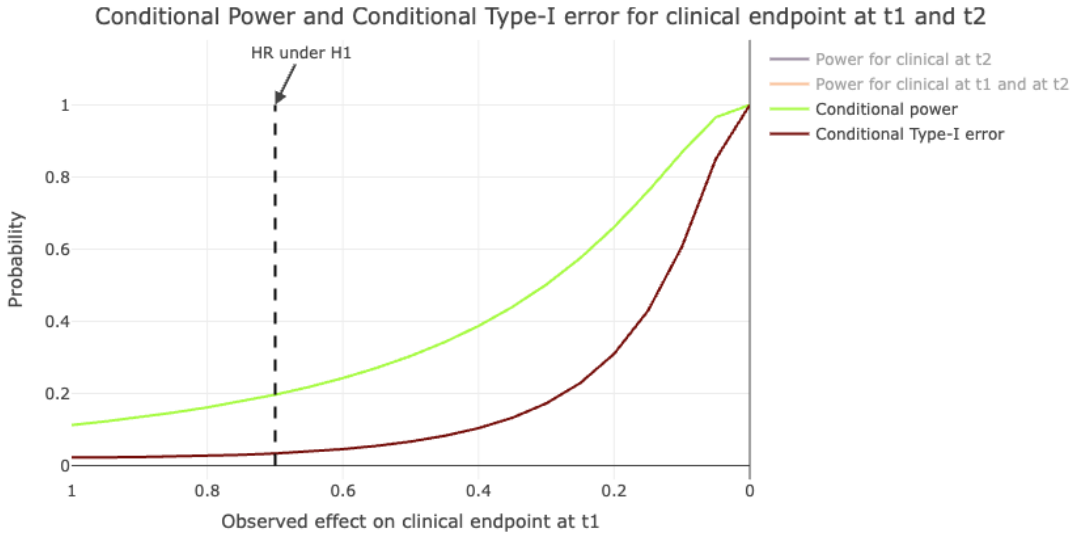

Figure 5: Conditional power given an observed clinical endpoint at the interim readout

4 Conclusion

The challenge of long and expensive clinical trials remains a primary obstacle in drug development. This paper has outlined a quantitative framework designed to address this inefficiency through the strategic use of intermediate endpoints. By providing a standardized approach, this framework moves beyond ad-hoc assessments and empowers teams to systematically evaluate candidate biomarkers and endpoints. This methodology enables earlier and more informed go/no-go decisions, which can accelerate development timelines, reduce resource expenditure, and ultimately increase the probability of success in large, confirmatory Phase III trials. The framework is designed to be therapeutic area (TA) independent and can be implemented in a user-friendly application, facilitating its broad adoption and helping to forge a more efficient path for bringing innovative medicines to patients.

Acknowledgments

We would like to express our gratitude to the statisticians Daria Rukina, Gian-Andrea Thanei and Anton Belousov for developing the methodology used in this application.

An AI Large Language Model was used to assist in the drafting, editing, and structuring of this paper based on source materials and a detailed outline provided by the authors. Specifically:

Google. (2025). LM Notebook (Gemini 1.5 Pro, October 9, 2025 version) [Large language model]. https://gemini.google.com

Contact Information

stefan.thoma@roche.com

References

[1] Douglas G Altman and Patrick Royston. The cost of dichotomising continuous variables. *BMJ : British Medical Journal*, 332(7549):1080, May 2006.

[2] Paul Gallo, Lu Mao, and Vivian H. Shih. Alternative Views On Setting Clinical Trial Futility Criteria. *Journal of Biopharmaceutical Statistics*, 24(5):976–993, September 2014.