Bayes Factor Group Sequential Designs

Samuel Pawel Leonhard Held

Epidemiology, Biostatistics and Prevention Institute (EBPI)
Center for Reproducible Science and Research Synthesis (CRS)
University of Zurich
{samuel.pawel,leonhard.held}@uzh.ch

January 6, 2026

Abstract

The Bayes factor, the data-based updating factor from prior to posterior odds, is a principled measure of relative evidence for two competing hypotheses. It is naturally suited to sequential data analysis in settings such as clinical trials and animal experiments, where early stopping for efficacy or futility is desirable. However, designing such studies is challenging because computing design characteristics, such as the probability of obtaining conclusive evidence or the expected sample size, typically requires computationally intensive Monte Carlo simulations, as no closed-form or efficient numerical methods exist. To address this issue, we extend results from classical group sequential design theory to sequential Bayes factor designs. The key idea is to derive Bayes factor stopping regions in terms of the z-statistic and use the known distribution of the cumulative z-statistics to compute stopping probabilities through multivariate normal integration. The resulting method is fast, accurate, and simulation-free. We illustrate it with examples from clinical trials, animal experiments, and psychological studies. We also provide an open-source implementation in the `bfpwr` R package. Our method makes exploring sequential Bayes factor designs as straightforward as classical group sequential designs, enabling experiments to rapidly design informative and efficient experiments.

Keywords: Bayesian hypothesis testing, design prior, predictive power, sample size determination, sequential clinical trials

1 Introduction

A crucial decision in the design of an experiment is choosing the sample size, for example, the number of participants, animals, or measurements. A too small sample size risks yielding inconclusive results, whereas a too large sample size may be too costly, logistically challenging, or ethically problematic. Sequential designs have often been proposed as a means to reduce sample size (Kairalla et al., 2012). In a sequential design it is possible to stop an experiment early on when conclusive evidence is found in an interim analysis. For example, a clinical trial may be stopped early if there is evidence that a treatment is effective (stopping for efficacy) or if there is evidence that a treatment is ineffective or even harmful (stopping for futility).

Despite their appeal, sequential designs pose statistical and practical challenges. For example, they can bias parameter estimates (Robertson et al., 2022) or threaten the integrity of a blinded trial, as interim analyses may require preliminary unblinding (Ellenberg et al., 2019). Furthermore, the analysis of frequentist sequential designs can be unintuitive. For instance, a final p-value below the conventional 5% threshold may still be insufficient to declare efficacy if it does not cross a more stringent threshold due to repeated analyses. Many practitioners find this difficult to accept, as the same data would permit an efficacy claim if only one analysis had been conducted (Matthews, 2006).

Bayesian methods are often considered as more natural for sequential analyses (Cornfield, 1966; Jack Lee and Chu, 2012). Repeated "updating" is inherent in the Bayesian framework and Bayesian probabilities are, in principle, not affected by multiplicity issues (however, see Ryan et al., 2020; Zhou and Ji, 2023). Two popular Bayesian approaches to sequential analysis are based on posterior tail probabilities (see e.g., Berry et al., 2010; Gsponer et al., 2013; Rosner, 2020) and Bayes factors. Here we focus on sequential Bayes factor designs, which have been applied across a wide range of domains, including biomedical research (Cornfield, 1976; Spiegelhalter et al., 2004; Goodman, 2005; Johnson and Cook, 2009; Li et al., 2017; Zhu et al., 2019; Zhou et al.,

2021; Rosner et al., 2021; Moerbeek, 2021; Pourmohamad and Wang, 2022; Linde and van Ravenzwaaij, 2023), the social sciences (Schönbrodt et al., 2017; Schönbrodt and Wagenmakers, 2018; Stefan et al., 2019; Mani et al., 2021; Stefan et al., 2022, 2024), and A/B testing in the tech industry (Deng et al., 2016; Lindon and Malek, 2022). They can also be viewed as Bayesian generalizations of classical sequential designs based on likelihood ratios (Wald, 1947; Royall, 1997).

The Bayes factor

$$\text{BF}_{01} = \underbrace{\frac{\Pr(H_0 \mid \text{data})}{\Pr(H_1 \mid \text{data})} \Big/ \frac{\Pr(H_0)}{\Pr(H_1)}}_{\text{Data-based updating factor}} = \underbrace{\frac{p(\text{data} \mid H_0)}{p(\text{data} \mid H_1)}}_{\text{Relative predictive performance}}$$

is the data-based updating factor of the prior odds of two hypotheses H_0 and H_1 to the corresponding posterior odds. This update is dictated by the ratio of the data's marginal likelihood under each hypothesis. The hypothesis which better predicts the data receives more support in terms of the Bayes factor (Kass and Raftery, 1995). Bayes factors thus provide a direct and interpretable measure of relative evidence for H_0 and H_1, which, unlike posterior probabilities, does not depend on the prior probabilities of H_0 and H_1.

In a sequential Bayes factor design, one fixes two Bayes factor thresholds $k_0 > 1$ and $k_1 < 1$, the thresholds for H_0 and H_1, respectively. Common choices are symmetric thresholds, such as $k_0 = 1/k_1 = 3$ or $k_0 = 1/k_1 = 10$, corresponding to "substantial" and "strong" relative evidence under Jeffreys' conventions (Jeffreys, 1961). The experiment is then stopped for either H_0 or H_1 as soon as the Bayes factor exceeds the corresponding threshold. If H_0 and H_1 represents the absence and presence of an effect, these rules naturally accommodate stopping for futility and efficacy, respectively. Sequential Bayes factor designs thus naturally link the stopping decision to a relevant statistical evidence measure.

The top plot in Figure 1 shows a simulated trajectory of a Bayes factor (oriented in favor of H_0 over H_1). Initially, the Bayes factor fluctuates around 1, indicating absence

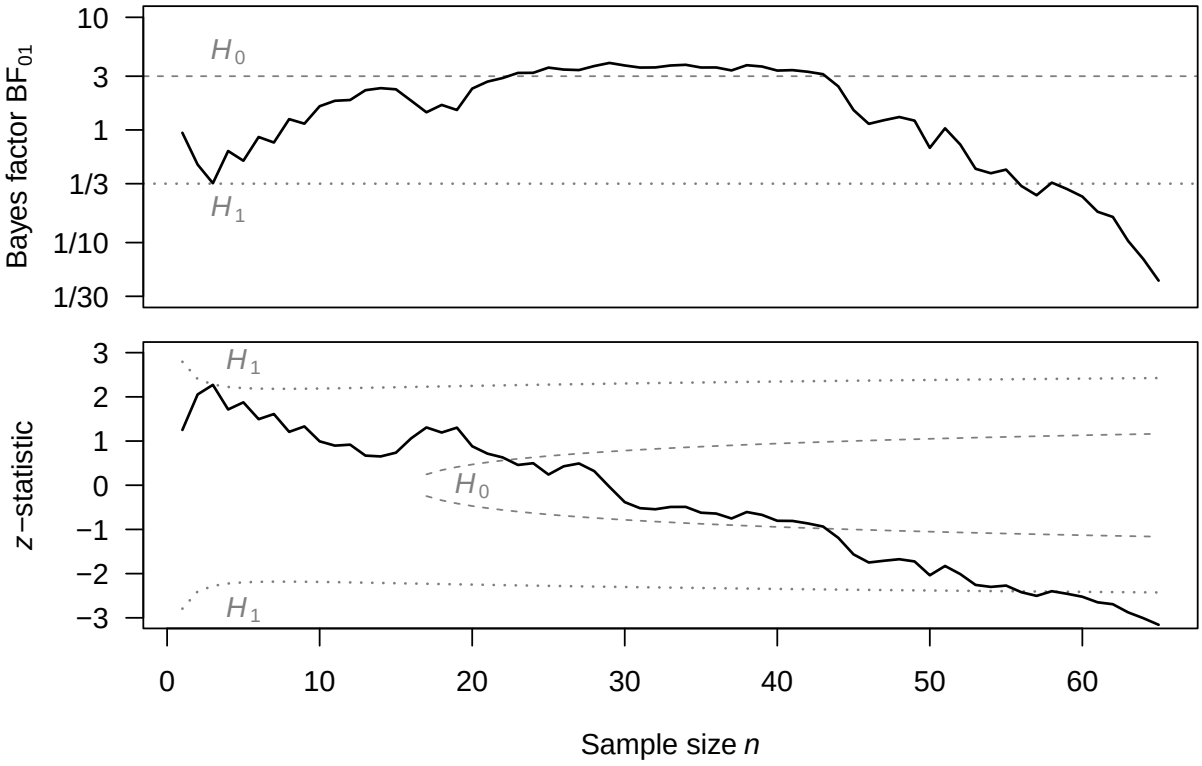


Figure 1: Illustration of a sequential Bayes factor analysis. The Bayes factor quantifying the evidence for a point null hypothesis $H_0\colon \theta = 0$ against an alternative hypothesis $H_1\colon \theta \neq 0$ is monitored as data accumulate (top). The bottom plot shows the corresponding z-statistic along with the Bayes factor stopping boundaries.

of evidence for either hypothesis. After around 20 observations, the Bayes factor surpasses the threshold of 3, indicating substantial evidence for H_0 over H_1. However, after around 50 observations, the Bayes factor quickly decreases below $1/3$ and $1/10$ – as it should since the data were simulated under the alternative hypothesis H_1. Depending on when the interim analyses were performed, the experiment would thus have been erroneously stopped for H_0 or correctly stopped for H_1.

The previous example illustrates the delicate choices that must be made when setting up a sequential Bayes factor design. These include choosing Bayes factor thresholds k_0 and k_1, deciding when to perform interim analyses, and determining the maximum sample size. Exploring these options can be challenging because computing design characteristics is typically only possible through Monte Carlo simulation methods. For example, the method implemented in the `BFDA` R package (Schönbrodt and

Stefan, 2019) involves analyzing many simulated datasets to determine how often an analysis yields correct or misleading results. While flexible, these methods are time-consuming and subject to Monte Carlo error. This limits the experimenter's freedom in prototyping and comparing different designs, which may ultimately result in the use of suboptimal designs.

In this paper, we present an alternative perspective on sequential Bayes factor designs that provides new insights and enables faster computation of design characteristics. The key idea is to consider the Bayes factor as a function of the z-statistic, the standardized difference between the observed parameter estimate and the null value. This allows us to define stopping regions in terms of the z-statistic (see the bottom plot in Figure 1). Using and extending results from classical group sequential designs, this approach enables efficient and deterministic computation of design characteristics.

In the following Section 2, we outline the general calculations of sequential Bayes factor design characteristics. We then explain a method for calculating design characteristics efficiently and without simulation when Bayes factors can be expressed as a function of the z-statistic (Section 3). Applications to clinical trials and animal experiments illustrate how the method can be used in practice (Section 4). Section 5 outlines an extension to a sequential version of the Bayes factor t-test (Gronau et al., 2020). The paper closes with concluding discussions, limitations, and opportunities for future research (Section 6). Appendix A illustrates usage of our R package `bfpwr` for planning sequential Bayes factor designs.

2 Sequential Bayes factor design characteristics

Suppose we want to plan a sequential Bayes factor design with m analyses. At each analysis $i = 1, \ldots, m$, based on the n_i observations available up to that point, a Bayes factor is computed and denoted by BF_{01}^i. Suppose that thresholds $k_0 > 1$ and $k_1 < 1$ are used and the experiment is stopped as soon as evidence for either H_0 ($\text{BF}_{01} \geq k_0$) or

H_1 ($\text{BF}_{01} \leq k_1$) is found. Four quantities are of central interest to assess the usefulness of a design:

1. **The probability of conclusive evidence for the alternative hypothesis** ($\text{BF}_{01} \leq k_1$). If data are generated under H_1 this corresponds to the probability of correct evidence for H_1 – a Bayesian analogue of the frequentist power (true positive rate). If data are generated under H_0 this corresponds to the probability of misleading evidence for H_1 – a Bayesian analogue of the false positive (type-I) error rate.

2. **The probability of conclusive evidence for the null hypothesis** ($\text{BF}_{01} \geq k_0$). If data are generated under H_1 this corresponds to the probability of misleading evidence for H_0 – a Bayesian analogue of the false negative (type-II error) rate. If data are generated under H_0 this corresponds to the probability of correct evidence for H_0 – a Bayesian analogue of the true negative rate.

3. **The expected sample size at termination**. A lower expected sample size is desired to reduce the actual sample size of the experiment.

4. **The standard deviation (or variance) of the sample size at termination**. A lower standard deviation is desired to better anticipate the actual sample size of the experiment.

The probability of finding evidence for H_1 can be decomposed

$$
\begin{aligned}
&\Pr(\text{Evidence for } H_1) \\
&= \Pr(\text{BF}_{01}^1 \leq k_1) && \text{Evidence for } H_1 \text{ in analysis 1} \\
&\quad + \Pr(k_1 < \text{BF}_{01}^1 < k_0, \text{BF}_{01}^2 \leq k_1) && \text{Evidence for } H_1 \text{ in analysis 2} \\
&\quad \vdots \\
&\quad + \Pr(k_1 < \text{BF}_{01}^1 < k_0, \ldots, k_1 < \text{BF}_{01}^{m-1} < k_0, \text{BF}_{01}^m \leq k_1) && \text{Evidence for } H_1 \text{ in analysis } m.
\end{aligned}
$$

Similarly, the probability of finding evidence for H_0 can be expressed as

$$
\begin{aligned}
&\Pr(\text{Evidence for } H_0) \\
&= \Pr(\text{BF}_{01}^1 \geq k_0) && \text{Evidence for } H_0 \text{ in analysis 1} \\
&\quad + \Pr(k_1 < \text{BF}_{01}^1 < k_0, \text{BF}_{01}^2 \geq k_0) && \text{Evidence for } H_0 \text{ in analysis 2} \\
&\quad \vdots \\
&\quad + \Pr(k_1 < \text{BF}_{01}^1 < k_0, \ldots, k_1 < \text{BF}_{01}^{m-1} < k_0, \text{BF}_{01}^m \geq k_0) && \text{Evidence for } H_0 \text{ in analysis } m.
\end{aligned}
$$

The expected sample size is

$$
\begin{aligned}
&\text{E}(n) \\
&= n_1 \times \{\Pr(\text{BF}_{01}^1 \leq k_1) + \Pr(\text{BF}_{01}^1 \geq k_0)\} && \text{Stop in analysis 1} \\
&\quad + n_2 \times \{\Pr(k_1 < \text{BF}_{01}^1 < k_0, \text{BF}_{01}^2 \leq k_1) + \Pr(k_1 < \text{BF}_{01}^1 < k_0, \text{BF}_{01}^2 \geq k_0)\} && \text{Stop in analysis 2} \\
&\quad \vdots \\
&\quad + n_m \times \Pr(k_1 < \text{BF}_{01}^1 < k_0, \ldots, k_1 < \text{BF}_{01}^{m-1} < k_0) && \text{Stop in analysis } m
\end{aligned}
$$

and likewise for $\text{E}(n^2)$, based on which also the variance of the sample size can be computed by

$$\text{Var}(n) = \text{E}(n^2) - \text{E}(n)^2.$$

Instead of the variance, the coefficient of variation $\text{COV}(n) = \sqrt{\text{Var}(n)}/\text{E}(n)$ may also be of interest to compare the variability of the sample size across designs with differing expected sample sizes. Each of these quantities thus consists of a (weighted) sum of analysis-wise stopping probabilities. These stopping probabilities are generally hard to compute due to the accumulating data having a complex distribution with dependence across analyses. For this reason, Monte Carlo simulation is typically used for approximating design characteristics. However, in the following we will show that by assuming a particular form of Bayes factor and data, fast and simulation-free computation is possible.

3 Bayes factors based on *z*-statistics

When testing hypotheses related to an unknown parameter θ, many types of Bayes factors BF_{01} can be expressed as a function of the z-statistic $z = (\hat{\theta} - \theta_0)/\sigma$, where $\hat{\theta}$ is an estimate of θ, σ is the estimate's standard error (typically of the form $\sigma = \lambda/\sqrt{n}$ with λ^2 a unit variance and n the effective sample size), and θ_0 is the null value related to the null hypothesis H_0 (typically $\theta_0 = 0$). Table 1 shows several Bayes factor types for which this is possible. These are related to a general class of Bayes factors based on test statistics (Johnson, 2005; Held and Ott, 2018). Assume now that we are interested in the critical z-values(s) for which $\text{BF}_{01} = k$ for some $k > 0$. For many Bayes factors in Table 1, critical values can be derived in closed-form (see the right column).

Although closed-form solutions are convenient, they are not necessary for computing design characteristics as described in the following sections. For example, the Bayes factor that contrasts a point null hypothesis ($H_0 \colon \theta = 0$) with a directional alternative hypothesis ($H_1 \colon \theta > 0$) (bottom row in Table 1) does not have an analytically available critical z-value. However, numerical root-finding can be used to determine it, as the Bayes factor can be expressed as a function of the z-statistic.

Figure 2 shows criticial z-values associated with different Bayes factors indicating strong evidence for the alternative over the null hypothesis ($\text{BF}_{01} = 1/10$). The critical z-values from two commonly used group sequential design (GSD) methods, Pocock and O'Brien-Fleming, are shown as comparison (see e.g., Chapter 2 in Jennison and Turnbull, 1999). Both ensure control of the type-I error rate at the conventional $\alpha = 0.025$ (one-sided) in a sequential design with 5 equally-spaced analyses. Section 3.2 will describe how the type-I error rate of sequential Bayes factor designs can be evaluated. We can see that the value and shape of the critical value boundaries differs across the different Bayes factors. The directional null vs. directional alternative Bayes factor ($H_0 \colon \theta \leq 0$ vs. $H_1 \colon \theta > 0$; black) shows near constant critical values across all analyses. In contrast, the point null vs. point alternative Bayes factor ($H_0 \colon \theta = 0$ vs. $H_1 \colon \theta = 0.1$; yellow) shows decreasing critical values, while the point null vs. directional alternative

Table 1: Different types of Bayes factors for data in the form of an estimate $\hat{\theta}$ of the parameter θ with standard error σ, which is assumed to be normally distributed $\hat{\theta} \mid \theta \sim \mathrm{N}(\theta, \sigma^2)$, and the critical values for the corresponding z-statistic $z = \hat{\theta}/\sigma$ so that the Bayes factor equals a threshold $\mathrm{BF}_{01} = k$.

Bayes factor	Critical z-value(s)
Directional null ($H_0\colon \theta \leq 0$) vs. directional alternative ($H_1\colon \theta > 0$) with marginal normal prior $\theta \sim \mathrm{N}(\mu, \tau^2)$ $$\mathrm{BF}_{01} = \frac{1 - \Phi(\mu_*/\tau_*)}{\Phi(\mu_*/\tau_*)} \Big/ \frac{1 - \Phi(\mu/\tau)}{\Phi(\mu/\tau)}$$ with $\tau_*^2 = 1/(1/\sigma^2 + 1/\tau^2)$ and $\mu_* = (z_i/\sigma + \mu/\tau^2)\tau_*^2$	$$z_{\mathrm{crit}}(k) = \left(\Phi^{-1}\left[\left\{ k \frac{1 - \Phi(\mu/\tau)}{\Phi(\mu/\tau)} + 1 \right\}^{-1} \right] \sqrt{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} - \frac{\mu}{\tau^2} \right) \sigma$$
Point null ($H_0\colon \theta = 0$) vs. point alternative ($H_1\colon \theta = \mu$) $$\mathrm{BF}_{01} = \exp\left(\frac{\mu^2}{2\sigma^2} - \frac{z\mu}{\sigma} \right)$$	$$z_{\mathrm{crit}}(k) = \frac{\mu^2/\sigma^2 - \log k^2}{2\mu/\sigma}$$
Point null ($H_0\colon \theta = 0$) vs. two-sided alternative ($H_1\colon \theta \neq 0$) with normal prior under alternative $\theta \mid H_1 \sim \mathrm{N}(\mu, \tau^2)$ $$\mathrm{BF}_{01} = \sqrt{1 + \frac{\tau^2}{\sigma^2}} \exp\left[-\frac{1}{2} \left\{ z^2 - \frac{(z - \mu/\sigma)^2}{1 + \tau^2/\sigma^2} \right\} \right]$$	$$z_{\mathrm{crit}-}(k) = M - \sqrt{X} \;\text{ and }\; z_{\mathrm{crit}+}(k) = M + \sqrt{X}$$ with $M = -\mu\sigma/\tau^2$ and $X = \{\mu^2/\tau^2 + \log(1 + \tau^2/\sigma^2) - \log k^2\}(1 + \sigma^2/\tau^2)$
Point null ($H_0\colon \theta = 0$) vs. directional alternative ($H_1\colon \theta > 0$) with truncated normal prior under alternative $\theta \mid H_1 \sim \mathrm{N}(\mu, \tau^2)_{(0,+\infty)}$ where the subscript denotes truncation of the distribution to the interval $(0, +\infty)$ $$\mathrm{BF}_{01} = \sqrt{1 + \frac{\tau^2}{\sigma^2}} \exp\left[-\frac{1}{2} \left\{ z^2 - \frac{(z - \mu/\sigma)^2}{1 + \tau^2/\sigma^2} \right\} \right] \frac{\Phi(\mu/\tau)}{\Phi(\mu_*/\tau_*)}$$	z_{crit} not analytically available but can be determined with numerical root-finding

Bayes factor ($H_0 : \theta = 0$ vs. $H_1 : \theta > 0$; light-blue) shows increasing critical values. The constant Pocock GSD critical values (green) thus align with the shape of the directional null vs. directional alternative Bayes factor (black) whereas the O'Brien-Fleming GSD critical values (dark-blue) align with the shape of the point null vs. point alternative Bayes factor (yellow).

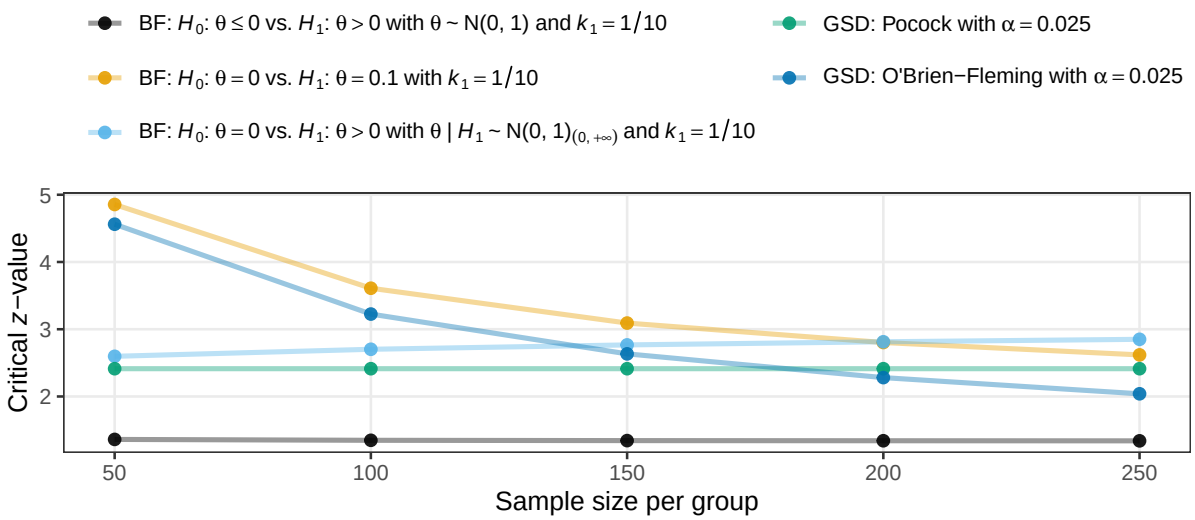


Figure 2: Critical z-values where $\text{BF}_{01} = 1/10$ for different types of Bayes factors (BF) from Table 1 and group sequential designs (GSD). The tested parameter is the standardized mean difference between two normally distributed populations with know variance and standard error of the form $\sigma = \sqrt{2/n}$ where n is the sample size per group.

3.1 Stopping regions based on *z*-statistics

Denote by z_i the z-value at analysis i and by $z_{i,\text{crit}}(k)$ the critical value(s) for which the Bayes factor equals the threshold $\text{BF}_{01} = k$. For Bayes factors with only one critical z-value (e.g., directional null vs. directional alternative) the H_0 stopping condition $\text{BF}^i_{01} \geq k_0$ is equivalent to $z_i \leq z_{i,\text{crit}}(k_0)$ whereas the H_1 stopping condition $\text{BF}^i_{01} \leq k_1$ corresponds to $z_i \geq z_{i,\text{crit}}(k_1)$. The continuation condition $k_1 < \text{BF}^i_{01} < k_0$ is equivalent to $z_{i,\text{crit}}(k_0) < z_i < z_{i,\text{crit}}(k_1)$. This means that the z-statistic stopping region at analysis i for the vector of cumulative z-statistics $(Z_1, \ldots, Z_i)^\top$ consists of an i-dimensional rectangle (an "i hyper-rectangle"). Integrating this region over the joint

distribution of the z-statistics gives the probability to stop at analysis i. For example, for $i = 1$ this is simply the interval $[z_{1,\text{crit}}(k_1), +\infty)$ while for $i = 2$ we have the rectangle $(z_{1,\text{crit}}(k_0), z_{1,\text{crit}}(k_1)) \times [z_{2,\text{crit}}(k_1), +\infty)$, see the blue and green regions in the left plot in Figure 3.

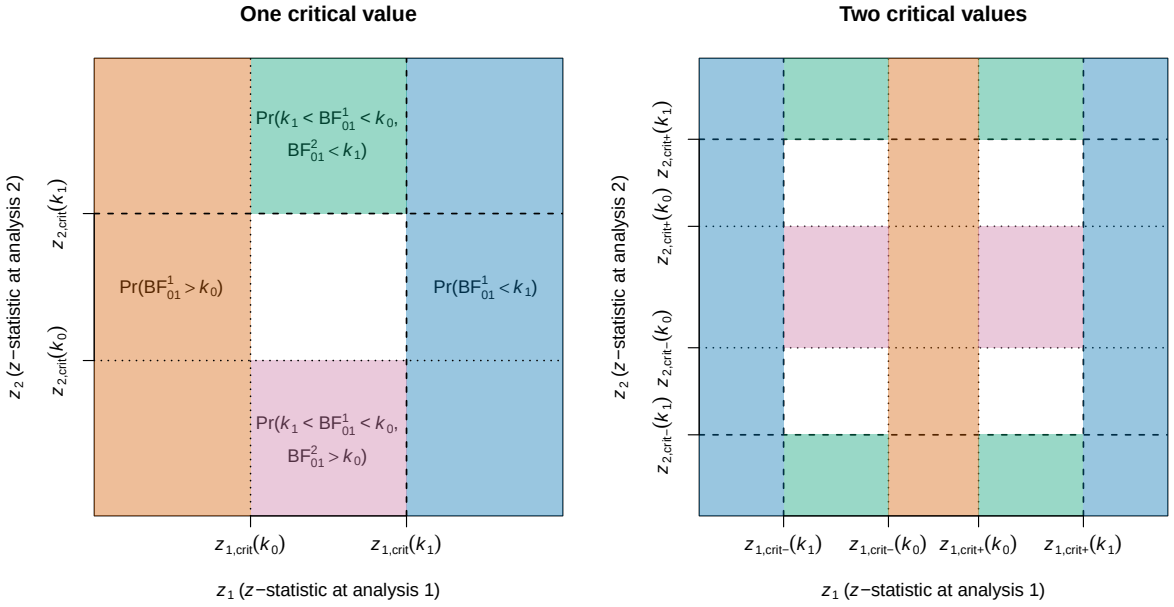


Figure 3: Illustration of critical z-values such that $\text{BF}_{01} \geq k_0$ (dotted lines) and $\text{BF}_{01} \leq k_1$ (dashed lines) for Bayes factors with one critical value (left) and two critical values (right) such that $\text{BF}_{01} = k$. Integrating the colored regions produces the indicated analysis-wise stopping probabilitites. For example, integrating the green region gives the probability to not stop in the first analysis but stop for H_1 in the second analysis.

For Bayes factors with two critical values (e.g., point null vs. two-sided alternative), the conditions are more complicated. The H_0 stopping condition $\text{BF}^i_{01} \geq k_0$ is equivalent to the z-statistic being in the interval around zero $z_i \in [z_{i,\text{crit}-}(k_0), z_{i,\text{crit}+}(k_0)]$ whereas the H_1 stopping condition $\text{BF}^i_{01} \leq k_1$ corresponds to the z-statistic being outside the interval $z_i \notin (z_{i,\text{crit}-}(k_1), z_{i,\text{crit}+}(k_1))$. The continuation condition $k_1 < \text{BF}^i_{01} < k_0$ is equivalent to the z-statistic being in between these two regions $z_i \in (z_{i,\text{crit}-}(k_1), z_{i,\text{crit}-}(k_0)) \cup (z_{i,\text{crit}+}(k_0), z_{i,\text{crit}+}(k_1))$. This means that at analysis i, the stopping region consist of the union of multiple i hyper-rectangles, their number doubling with each analysis. For example, in analysis 1 we have two intervals of z-statistics for which the Bayes factor is $\text{BF}_{01} \leq k_1$, see the blue regions in the right plot of Figure 3.

In analysis 2, the number of rectangles doubles to four (green rectangles in the right plot of Figure 3). This means that with an increasing number of interim analyses, the number of hyper-rectangles grows exponentially as every i hyper-rectangle at analysis i splits into two $i+1$ hyper-rectangles at analysis $i+1$.

While the exponentially growing number of hyper-rectangles poses certain computational challenges, we did not observe any numerical issues in our applications when exhaustively computing the probability of all hyper-rectangles in designs with even up to ten analyses. In many fields, designs involving such a large number of analyses are rarely encountered. For example, the most common number of interim analyses in clinical trials is only one or two, if any (Stevely et al., 2015). Moreover, for Bayes factors with only one critical z-value (e.g., one-sided tests), the number of hyper-rectangles increases only linearly with the number of analyses. Exhaustively computing their probabilities therefore poses no computational problems, even for a very large number of analyses.

3.2 The predictive distribution of the *z*-statistics

In order to compute the probability for the z-statistics to fall within a given region, we need to know their distribution. Suppose that m analyses produce the sequence of z-statistics $\{Z_1, \ldots, Z_m\}$. Under typical conditions, these have a *canonical distribution* (Jennison and Turnbull, 1999, Chapter 3), that is, is a distribution of the form

1. $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^\top$ has a m-variate normal distribution
2. $\mathrm{E}(\boldsymbol{Z}) = \theta \boldsymbol{I}$ where θ is the true parameter and $\boldsymbol{I} = (\sqrt{I_1}, \ldots, \sqrt{I_m})^\top$ is the vector of square-rooted information levels. The information level is typically the inverse of the squared standard error σ^2. For example, for a normal mean, the information level is $I_i = 1/\sigma^2 = n_i/\lambda^2$, where n_i is the sample size at analysis i and λ^2 the known variance of one observation.
3. $\mathrm{Cov}(Z_i, Z_{i+j}) = \boldsymbol{\Sigma}_{i,i+j} = \sqrt{I_i / I_{i+j}}$ for $j \geq 0$. For example, for a normal mean the

covariance is $\text{Cov}(Z_i, Z_{i+j}) = \sqrt{n_i / n_{i+j}}$.

In short, the vector of z-statistics has distribution

$$\boldsymbol{Z} \mid \theta \sim \text{N}_m \left(\theta \boldsymbol{I}, \boldsymbol{\Sigma} \right). \tag{1}$$

Assuming a fixed value for θ, one can compute the probability of a z-statistic vector falling into a given stopping region using standard numerical approaches for calculating multivariate normal integrals. Due to the canonical distribution, this calculation can even be further simplified via a recursive one-dimensional integration algorithm, which is also used extensively in the calculation of stopping probabilities in classical group sequential designs (Armitage et al., 1969; Jennison and Turnbull, 1999). However, from a Bayesian perspective, it seems more natural to account for uncertainty of θ by specifying a *design prior* distribution (O'Hagan and Stevens, 2001), sometimes also known as sampling prior (Wang and Gelfand, 2002; Psioda and Ibrahim, 2018), instead of a fixed value. Taking the design prior to be a normal prior $\theta \sim \text{N}(\mu_d, \tau_d^2)$, the marginal (or prior-predictive) distribution of $\boldsymbol{Z}$ is then

$$\boldsymbol{Z} \mid \mu_d, \tau_d^2 \sim \text{N}_m \left(\mu_d \boldsymbol{I}, \boldsymbol{\Sigma} + \tau_d^2 \boldsymbol{I}\boldsymbol{I}^\top \right), \tag{2}$$

see Appendix B for a derivation. The distribution (2) reduces again to the canonical distribution if a point prior at μ_d is assigned (i.e., if $\tau_d \downarrow 0$). However, for non-degenerate priors ($\tau_d > 0$), the vector of accumulating test statistics $\boldsymbol{Z}$ exhibits higher correlations than under the canonical distribution, which increases with increasing τ.

In contrast to the analysis prior, which is typically "weakly-informative" or "objective" in some sense, the design prior should represent genuine knowledge and uncertainty at the design analysis in order to accurately estimate stopping probabilities and sample sizes. However, setting the design prior to a point mass ($\tau_d \downarrow 0$) allows us to study the frequentist operating characteristics of the sequential design. For example, specifying a point prior at $\mu_d = 0$ and computing the probability to find evidence for

H_1 gives the frequentist type-I error rate. Despite the fact that this probability may be unrealistic from a Bayesian perspective, showing that a sequential Bayes factor design is appropriately calibrated (i.e., has type-I error rate below a conventional level, e.g., 5%) may be required from certain stakeholders, such as a regulatory authorities in drug development (U.S. Food and Drug Administration, 2010; Campbell, 2020).

Assuming there are m analyses, the probability of evidence for $H_i \in \{H_0, H_1\}$ is

$$\Pr(\text{Evidence for } H_i) = \sum_{j=1}^{m} \Pr\{\boldsymbol{Z}_{1:j} \in \boldsymbol{S}_j^i \mid \mu_d \boldsymbol{I}_{1:j}, (\boldsymbol{\Sigma} + \tau_d^2 \boldsymbol{I}\boldsymbol{I}^\top)_{1:j,1:j}\}$$

where $1 : j$ indicates indexing of the first j elements and $\boldsymbol{S}_j^i$ is the set of z-statistic stopping regions for H_i at analysis j. For example, assuming that $m = 3$ and that there is only one critical value so that $\text{BF}_{01} = k$ (e.g., a directional null vs. directional alternative Bayes factor in Table 1), we have the stopping regions for H_1

$$\boldsymbol{S}_1^1 = [z_{1,\text{crit}}(k_1), +\infty] \quad \boldsymbol{S}_2^1 = \begin{bmatrix} z_{1,\text{crit}}(k_0), & z_{1,\text{crit}}(k_1) \\ z_{2,\text{crit}}(k_1), & +\infty \end{bmatrix} \quad \boldsymbol{S}_3^1 = \begin{bmatrix} z_{1,\text{crit}}(k_0), & z_{1,\text{crit}}(k_1) \\ z_{2,\text{crit}}(k_0), & z_{2,\text{crit}}(k_0) \\ z_{3,\text{crit}}(k_1), & +\infty \end{bmatrix}$$

corresponding to hyper-rectangles in one-, two-, and three-dimensional space. While under the canonical distribution, these probabilities could be expressed as recursive one-dimensional integrals, this is not possible anymore with design priors where $\tau_d > 0$ due to the distribution not being canonical anymore. However, we have found that computation of even very large-dimensional integrals (e.g., designs with 20 analyses, which is unrealistic in practice) is still very efficient with modern implementations of the multivariate normal distribution, such as the `mvtnorm` R package (Genz and Bretz, 2009). Such extreme designs will be demonstrated in Sections 4.2 and 5.

4 Applications

We will now illustrate planning of sequential Bayes factor designs using case studies from clinical trials and animal experiments.

4.1 The Low-PV trial

The Low-PV trial (Barbui et al., 2021) assessed if the drug ropeginterferon alfa-2b could help low-risk polycythaemia vera (Low-PV) patients keeping haematocrit levels (the volume percentage of red blood cells in blood) within a safe range. The study was designed to have three analyses after 50, 100, and 150 patients had been followed-up, respectively. The design assumed response rates of $\pi_0 = 50\%$ and $\pi_1 = 75\%$ under H_1 in the control and treatment groups, corresponding to an odds ratio of $\text{OR} = 3$. The study was stopped after the second interim analysis because the group sequential stopping bounds for efficacy were crossed. Table 2 summarizes the by-analysis results.

Table 2: Results from the Low-PV trial (Barbui et al., 2021). Shown are by-analysis sample sizes (n) and estimated response probabilities ($\hat{\pi}$) in control (subscript 0) and treatment groups (subscript 1), respectively, along with estimated odds ratio ($\widehat{\text{OR}}$) with corresponding 95% confidence interval (CI), z-value, and decision based on group sequential stopping bounds. The right-most column gives the Bayes factor contrasting $H_0\colon \text{OR} = 1$ to $H_1\colon \text{OR} = 3$ based on an approximately normal likelihood of the estimated log OR.

Analysis	n_0	n_1	$\hat{\pi}_0$	$\hat{\pi}_1$	$\widehat{\text{OR}}$ (95% CI)	z	Decision	BF_{01}
1	26	24	57.7%	87.5%	5.1 (1.2 to 21.6)	2.23	Continue	1/9.2
2	50	50	60.0%	84.0%	3.5 (1.4 to 9.0)	2.60	Stop (efficacy)	1/27.9

Assuming a normal likelihood for the estimated log odds ratio $\hat{\theta} = \log \widehat{\text{OR}}$ and using the hypotheses specified by the trial investigators, we can compute the Bayes factor from Table 1 contrasting the point hypotheses $H_0\colon \theta = 0$ to $H_1\colon \theta = \log(3) \approx 1.1$. This leads to a Bayes factor $\text{BF}_{01}^1 = 1/9.2$ in the first analysis and $\text{BF}_{01}^2 = 1/27.9$ in the second. Had we used a threshold of $k_1 = 1/10$, the study would have thus been stopped for H_1 after the second analysis, resulting in the same decision as was made by the trial investigators based on a classical group sequential design.

Suppose the trial has not been carried out and we want to plan it using a sequential Bayes factor design. Assuming allocation of n patients to each group, the approximate standard deviation of the estimated log odds ratio $\hat{\theta}$ can be derived via the delta method to be $\sigma = \sqrt{1/\{n\pi_0(1-\pi_0)\} + 1/\{n\pi_1(1-\pi_1)\}}$. Once data are observed, it is estimated through the usual standard error obtained from plugging in the estimated rates $\hat{\pi}_0$ and $\hat{\pi}_1$ (Bland, 2000). Assuming a normal distribution for $\hat{\theta}$ around the true log odds ratio θ with standard deviation σ, implies a canonical distribution for $z = \hat{\theta}/\sigma$ with information level $I = 1/\sigma^2$. We can therefore apply the results from Section 3 to compute sequential Bayes factor design characteristics.

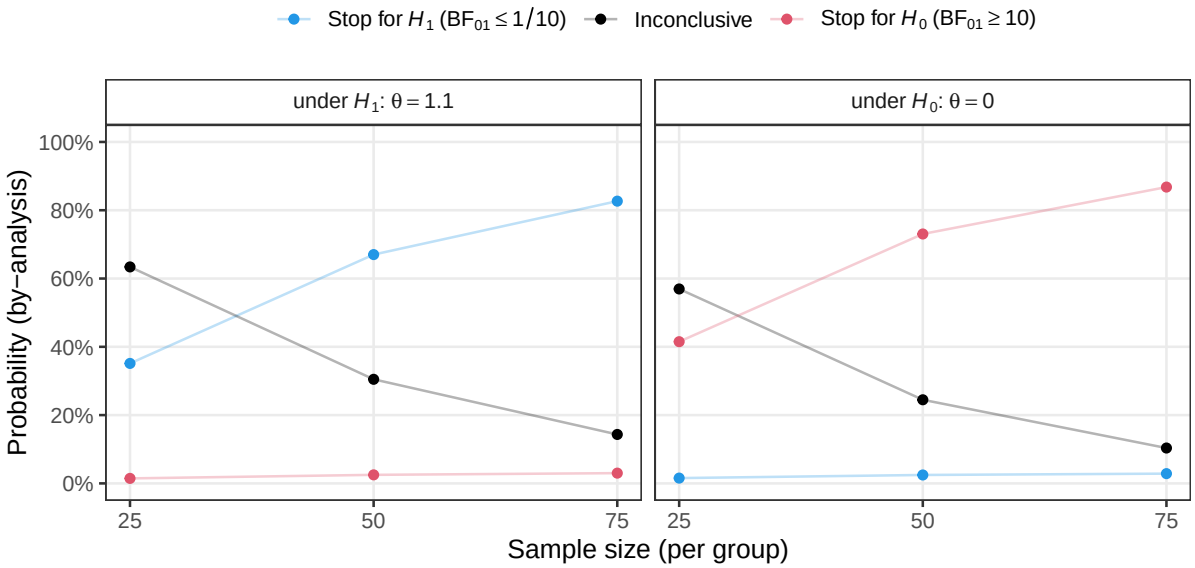


Figure 4: Sequential Bayes factor probabilities for the Low-PV trial (Barbui et al., 2021).

Figure 4 shows the probability of obtaining conclusive evidence for H_1 (blue), H_0 (red), or to remain inconclusive (black) by each of the three analyses, assuming that data are generated under H_1 ($\theta = 1.1$; left plot) or under H_0 ($\theta = 0$; right plot). Three equally-spaced analyses at 25, 50, and 75 patients per group are considered, as in the original trial. We can see that the probabilities of conclusive evidence for the corresponding true hypotheses are very similar under both hypotheses. Both surpass 80% after the third analysis, but remain below 90%. To achieve a probability of 90%, the maximum sample size and/or the number of interim analyses must be further in-

creased.

Keeping the number of analyses fixed at three, we can use numerical root-finding (e.g., `uniroot` in R) to determine the maximum sample size size to achieve a 90% probability of conclusive evidence. This leads to $n_3 = 87$ under H_0 and $n_3 = 102$ under H_1, translating into 29 and 34 additional patients per group at each analysis, respectively. Taking the maximum of the two would ensure that the trial produces conclusive evidence with high probability.

Figure 5 shows design characteristics varying the maximum samples size and also the numbers of analyses. This includes also only one analysis which corresponds to a fixed design. Note that our approach can compute these design characteristics in a few seconds, whereas simulating all these designs would introduce Monte Carlo error and take substantially longer as for every combination a sequential trial has to be simulated many times. From the top-row plots, we can see that the probability of obtaining "correct" evidence (i.e., $\text{BF}_{01} \leq 1/10$, if H_1 is true and $\text{BF}_{01} \geq 10$, if H_0 is true) increases with increasing maximum sample size (x axis). Increasing the number of analyses (color) also increases the probability of obtaining correct evidence for low maximum sample sizes. However, for maximum sample sizes higher than around 110, increasing the number of analyses actually decreases the probability of obtaining correct evidence. This is because a design with more analyses has also higher chances of producing misleading evidence (i.e., $\text{BF}_{01} \geq 10$ if H_1 is true and $\text{BF}_{01} \leq 1/10$ if H_0 is true) and stopping for the wrong hypothesis (second-row plots).

The probability of misleading evidence first increases and then decreases again with increasing maximum sample size. Since H_0 and H_1 are point hypotheses, the probability of misleading evidence is bounded by $k = 1/10$, which is known as the "universal bound" (Royall, 1997). However, across all choices of the number of analyses, the probability is much lower than the bound, remaining below 5% across all maximum sample sizes and number of analysises, though bringing it below 2.5% (the conventional type-I error rate for one-sided tests) requires substantial increases in the maximum sample

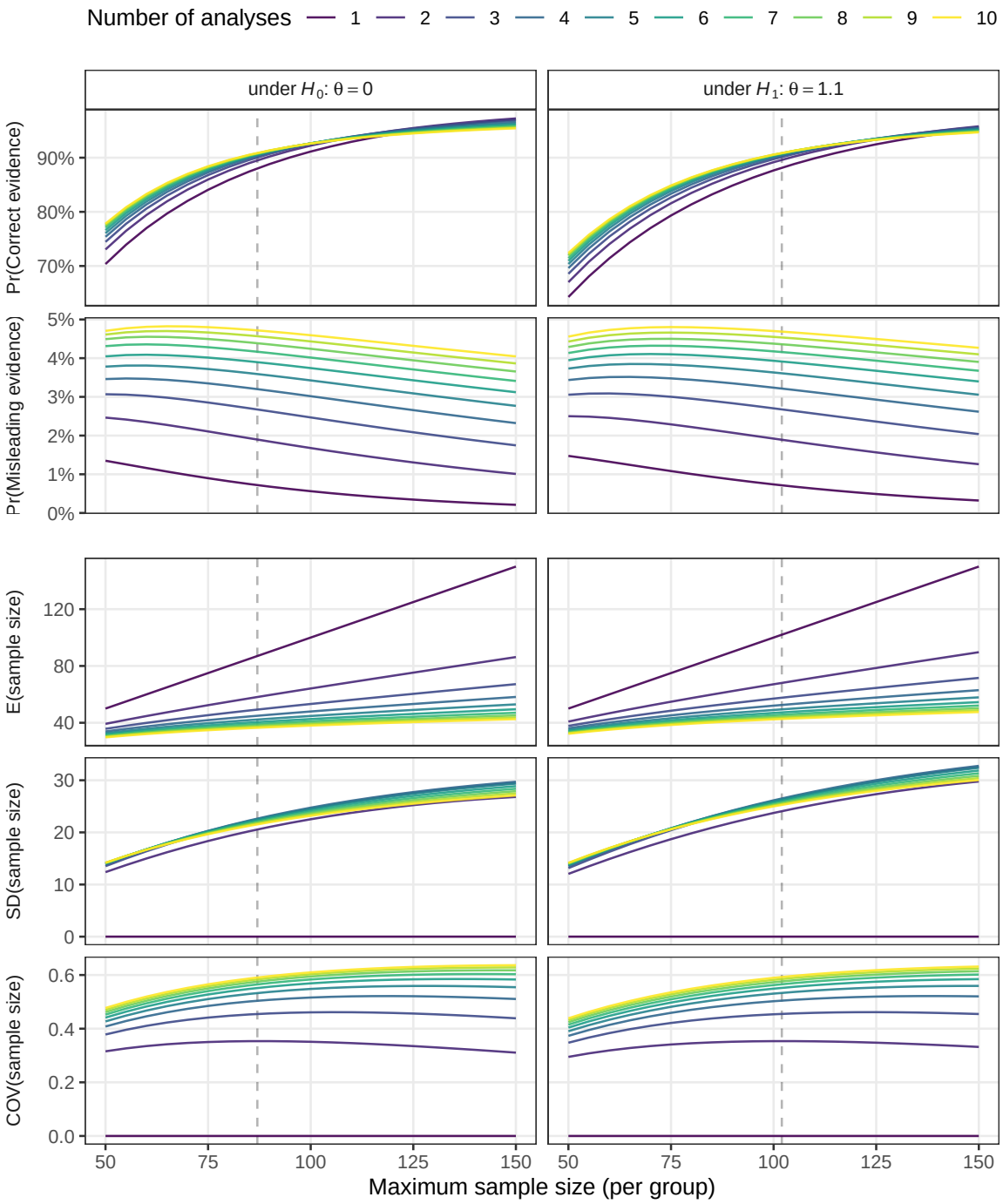


Figure 5: Sequential Bayes factor design characteristics for the Low-PV trial (Barbui et al., 2021) for differing maximum sample sizes and number of analyses. The dashed lines show the maximum sample sizes corresponding to 90% probability of correct evidence for 3 analyses.

size for designs with many analyses (not shown).

Increasing the number of analyses reduces the expected sample size (third-row plots) due to the potential early stopping, while having a non-monotone effect on

the standard deviation of the sample size (fourth-row plots) – increasing for smaller maximum sample sizes and decreasing for larger maximum sample sizes. Dividing the standard deviation by the expected sample size gives the coefficient of variation (bottom-row plots), which enables sample size variability to be compared while accounting for differing expected sample sizes. For a given maximum sample size, the coefficient of variation increases with increasing number of analyses, as more sample sizes at which stopping is possible become available. At the same time, for a given number of analyses, increasing the maximum sample size initially increases the coefficient of variation until it reaches a maximum, after which it decreases again. However, for a low number of analyses (e.g., two or three) this change is relatively small so that the coefficient of variation remains almost constant.

In sum, compared to a fixed design, a sequential design with even only two or three analyses can drastically improve efficiency. It reduces the expected sample size and increase the probability of correct evidence, while only slightly increasing the probability of misleading evidence and the variability of the sample size.

4.2 Rat experiment on weight loss

In animal research, just as in clinical trials, every additional observation carries significant ethical weight. Reducing sample size is hence of paramount interest, which is one pillar of the 3R ("Replace, Reduce, Refine") principles in animal research (Russell and Burch, 1959). We will now reanalyze data from a preclinical experiment with rats, which was retrospectively analyzed by Kang et al. (2025). In the experiment, rats were randomly assigned to different dose levels of a candidate drug or to a control group to estimate the drug's effect on weight loss. Information that could be used to identify the drug or study has been removed by Kang et al. (2025) due to confidentiality issues. The top plot of Figure 6 shows the measured weight loss values across groups. As can be seen, the low dose group (yellow) differs little from the control group (black), whereas the medium (blue) and high dose (green) groups show much higher weight loss.

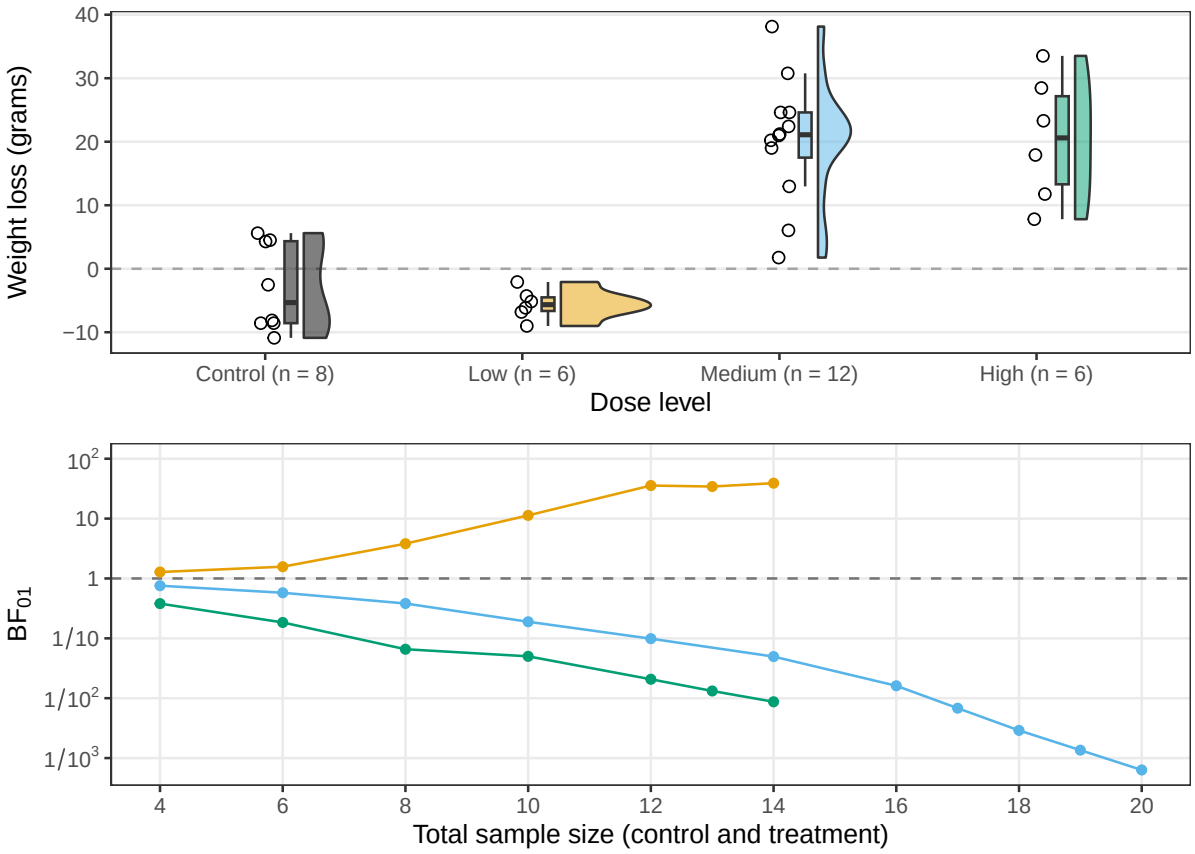


Figure 6: Weight loss in rats assigned to different levels of a candidate drug (Kang et al., 2025) (top plot). The bottom plot shows a sequential Bayes factor analysis contrasting the null hypothesis of no mean difference ($H_0\colon \theta_i = 0$) to a mean difference of 5 grams ($H_0\colon \theta_i = 5$) for each treatment group $i \in \{\text{low, medium, high}\}$ indicated by the color. At each step an observation from control and treatment group is added until the maximum group size is reached.

In the analysis of Kang et al. (2025), the effects of interest were the mean differences in weight loss between the treatment groups and the control group, denoted by θ_i with $i \in \{\text{low, medium, high}\}$. The investigators defined a mean difference of at least 5 grams as effective, which we will now use to specify the alternative hypothesis. The Bayes factor contrasting the point hypotheses $H_0\colon \theta_i = 0$ to $H_1\colon \theta_i = 5$ is shown in the bottom plot of Figure 6 for each treatment group (color). An observation from each group is added at each step until the maximum group size is reached. In each step, the mean difference $\hat{\theta}_i = \widehat{\mathrm{E}}(Y_i) - \widehat{\mathrm{E}}(Y_{\text{control}})$ and its standard error $\sigma_i = \sqrt{\widehat{\mathrm{Var}}(Y_i)/n_i + \widehat{\mathrm{Var}}(Y_{\text{control}})/n_{\text{control}}}$ are estimated from the available data, which are then used to compute the Bayes factor from Table 1. Note that the order

in which the data were collected is unknown, which is why a random permutation is shown here. Appendix C demonstrates that also for other permutations of the data, similar results are obtained. As the sample size increases, the Bayes factor in the low dose group (yellow) increases and reaches $\text{BF}_{01} > 10$ after 10 rats, suggesting strong evidence for the absence of an effect over its presence. In contrast, the Bayes factors in the medium (blue) and high (green) dose groups decrease with increasing sample size, surpassing the threshold for strong evidence of $\text{BF}_{01} < 1/10$ after 12 and 8 rats, respectively. Thus, if the experiment had been conducted using this sequential Bayes factor design, $8 - \max\{10, 12, 8\}/2 = 2$ rats from the control group and $6 + 12 + 6 - (10 + 12 + 8)/2 = 9$ rats from the treatment groups could have been saved, totalling to 11 saved rats.

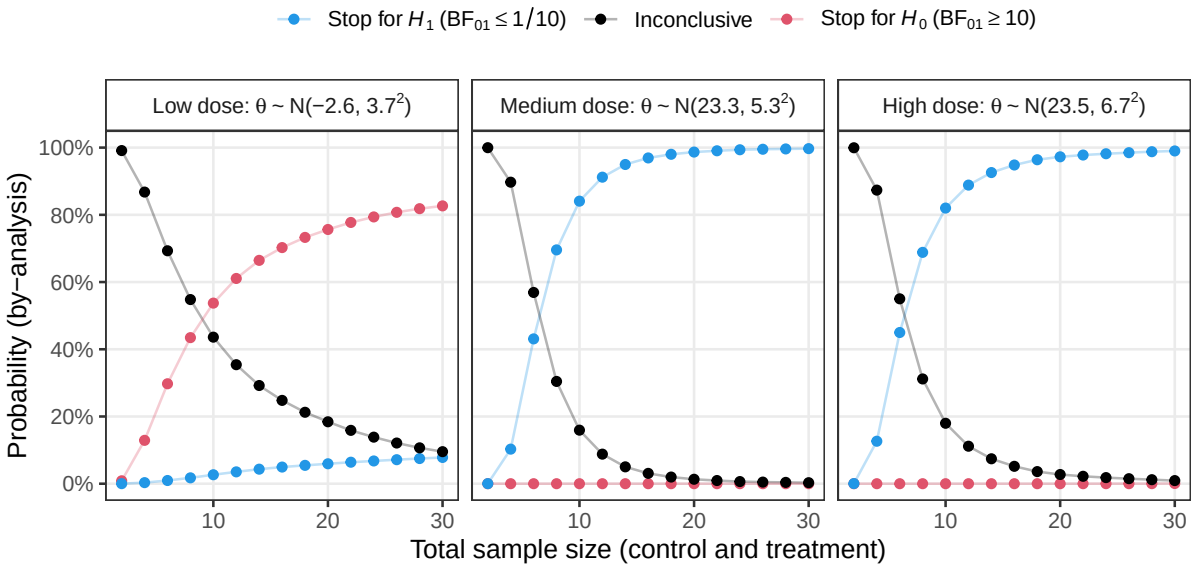


Figure 7: Sequential Bayes factor probabilities for replications of weight loss rat experiments (Kang et al., 2025). Probabilities are computed assuming a design prior corresponding to posterior distributions for the mean difference θ based on data from the original experiments (indicated in the plot panels). Analyses are assumed to be conducted after each pair of rats from control and treatment groups.

Suppose that before moving on to trials with human participants, we want to replicate these findings in independent experiments – a practice which is generally recommended in preclinical studies to rule out false positives (Piper et al., 2019; Drude et al., 2021). In the Bayesian design of a replication study, it is natural to use a design prior

based on data from the original study to plan the replication (Micheloud and Held, 2022; Anderson and Kelley, 2022; Pawel et al., 2023). Figure 7 shows stopping probabilities for replication study designs in which an analysis is performed after each pair of rats from the control and treatment groups. These designs assume a normal design prior distribution for the mean difference θ, centered around the estimated mean difference from the original experiment, with a standard deviation equal to the standard error of the estimate (see plot panels). Such a design prior can be motivated as the posterior distribution of the mean difference based on the original data and a flat prior. Note that if we would also use this design prior as the analysis prior (instead of the point hypothesis $H_1 : \theta = 5$), a "replication Bayes factor" (Verhagen and Wagenmakers, 2014; Pawel and Held, 2022) would be obtained. However, here we will consider testing the same alternative hypotheses as in the original study.

For the medium and high dose groups, the probabilities of finding strong evidence for the alternative over the null hypothesis ($\text{BF}_{01} \leq 1/10$; blue curves) quickly increase with increasing sample size. They reach 80% after a sample size of 10 in both cases. In contrast, for the low dose group, the probability of finding strong evidence for the null over the alternative hypothesis ($\text{BF}_{01} \geq 10$; red curve) increases more slowly, requiring nearly 30 rats to reach 80%. While the probability of misleading evidence remains nearly 0% for the medium and high dose groups (red curves), it increases considerably for the low dose group (up to almost 10%, blue curve) since the design prior based on the original data does not fully rule out positive effects in the neighborhood of the alternative hypothesis ($H_1 : \theta = 5$).

In sum, the developed sequential Bayes factor design method enables rapid calculation of key design characteristics while accounting for parameter uncertainty. This can potentially lead to more efficient designs. For instance, rather than allocating an equal number of rats to all groups, one could allocate fewer rats to the medium and high dose groups and more rats to the low dose group, thereby ensuring a high probability of informative inferences across all groups.

5 The sequential Bayesian *t*-test

The Bayes factor version of the t-test is a popular approach for sequential hypothesis testing in the social sciences. Assuming normally distributed data with unknown variance, Gronau et al. (2020) proposed an "informed" t-test Bayes factor that can take prior information into account through informative prior distributions. The Bayes factor is

$$\text{BF}_{01} = \frac{\text{T}_\nu(t \mid 0, 1)_{(-\infty, +\infty)}}{\int_{-\infty}^{+\infty} \text{NCT}_\nu(t \mid \theta\sqrt{n})\, \text{T}_\kappa(\theta \mid \mu, \tau)_{[a,b]}\, \text{d}\theta} \tag{3}$$

where t is the observed t-statistic, n is the effective sample size (the actual number of observations/pairs for one-sample/paired t-tests, or half the harmonic mean of the group sizes for the two-sample t-tests), and $\text{T}_\nu(\cdot \mid \mu, \tau)_{[a,b]}$ is the density of the location-scale t distribution with location μ, scale τ, and corresponding degrees of freedom ν, truncated to the interval $[a, b]$. $\text{NCT}_\nu(\cdot \mid \lambda)$ is the density of the non-central t distribution with non-centrality parameter λ. When the hyperparameters are set to $\kappa = 1$, $\mu = 0$, $a = -\infty$, $b = +\infty$, the prior becomes a Cauchy distribution, and the Bayes factor reduces to the widely used "Jeffreys-Zellner-Siow" (JZS) Bayes factor (Jeffreys, 1961; Zellner and Siow, 1980). The JZS Bayes factor is typically used with a scale of $\tau = 1/\sqrt{2}$, as e.g., implemented in the `BayesFactor` R package (Rouder et al., 2009).

Design calculations for the sequential Bayes factor t-test can also be embedded in the proposed z-statistic framework. The Bayes factor (3) is not available in closed-form but requires one-dimensional numerical integration. Therefore, the critical t-value such that $\text{BF}_{01} = k$ must be determined numerically, as demonstrated by Pawel and Held (2025) and Wong and Tendeiro (2025) for fixed designs. Assuming that the effective sample size is large enough (e.g., $n \geq 30$), the t distribution can be approximated by a normal distribution. Specifically, for large enough n, we have that $t \mid \theta \sim \text{N}(\theta\sqrt{n}, 1)$ where θ is the standardized mean (difference). Therefore, the vector of accumulating t-statistics approximately follows the canonical distribution from Section 3.2 with information level $I_i = n_i$ at analysis i, enabling the computation of

stopping probabilities and other design characteristics via numerical multivariate normal integration.

Schönbrodt and Wagenmakers (2018) describe an application of the t-test Bayes factor to sequential experiments in psychology. They consider an extreme design involving 61 analyses, in which an analysis is performed after every additional pair of participants in treatment and control groups until the maximum sample size $n_{61} = 100$ is reached or the experiment stopped before. They set asymmetric Bayes factor thresholds $k_0 = 6$ and $k_1 = 1/30$ and specify a normal design prior for the standardized mean difference $\theta \sim \mathrm{N}(0.5, 0.1^2)$ to account for parameter uncertainty. Finally, for the analysis they assume a default JZS Bayes factor with scale $\tau = 1/\sqrt{2}$ truncated to positive standardized mean differences ($a = 0$, $b = +\infty$).

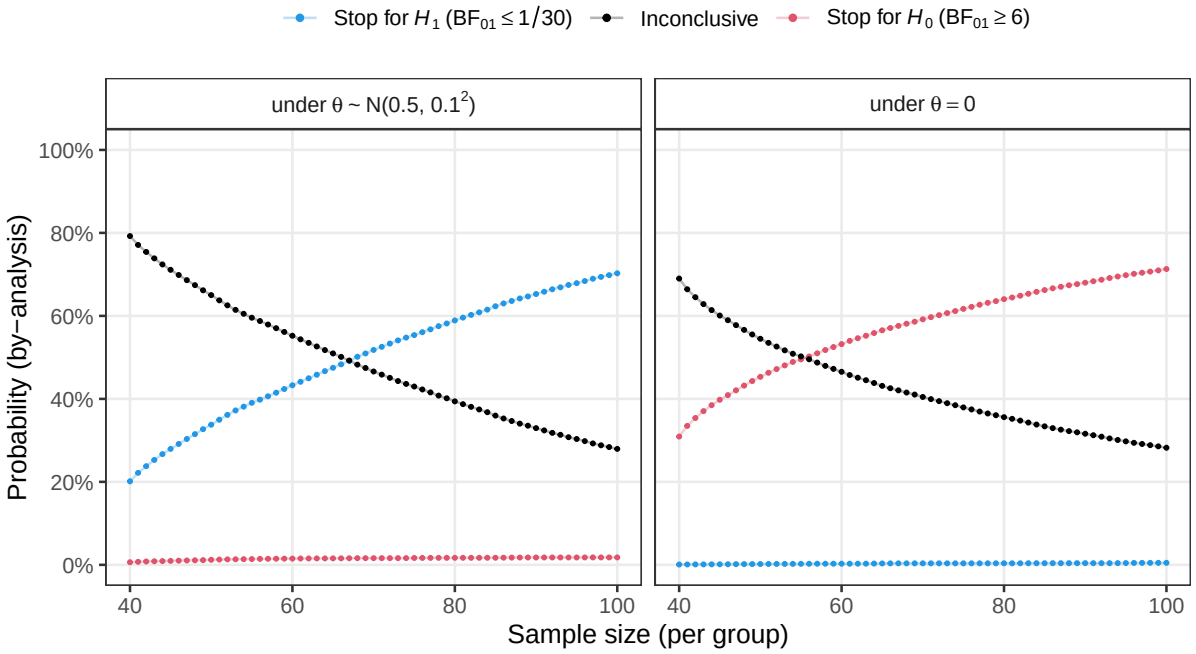


Figure 8: Sequential Bayes factor design probabilities based on t-test Bayes factor for design from Schönbrodt and Wagenmakers (2018).

Figure 8 shows the stopping probabilities for the design from Schönbrodt and Wagenmakers (2018). Despite the many interim analyses, the calculations took only a few seconds, whereas simulation-based calculations using the `BFDA` package took around 20 minutes on the same computer. Under the specified normal design prior (left plot),

we obtain the stopping probabilities for H_1 and H_0 as 70.3% and 1.8%, respectively, as well as an average sample size of 69.4. These are very close to the 70.6%, 1.6%, and 69 reported by Schönbrodt and Wagenmakers (2018), which were estimated with simulation. Similarly, assuming no effect ($\theta = 0$, right plot), we obtain 0.5% to stop for H_1, 71.3% to stop for H_0, and an average sample size of 65.7, which are again very close to the 0.6%, 70.9%, and 66 reported by Schönbrodt and Wagenmakers (2018). In sum, the z-statistic perspective enables quick and accurate calculation of key design characteristics of sequential designs based on the Bayes factor t-test, one of the most commonly used sequential Bayes factor tests.

6 Discussion

Bayes factors are natural tools for sequential data analysis: Data can be repeatedly analyzed without concerns about multiplicity, stopping decisions are naturally linked to interpretable updates of prior to posterior odds of competing hypotheses, and, unlike posterior tail probabilities, there is no need to specify prior probabilities for competing hypotheses. Despite these advantages, the broader adoption of sequential Bayes factor designs has been limited, one potential reason being the difficulty of computing their design characteristics. Existing approaches typically rely on extensive simulation, which can be computationally costly, sensitive to Monte Carlo error, and inconvenient when exploring many design options.

In this paper, we introduced a general approach that overcomes these limitations by expressing Bayes factors as functions of z-statistics and and extending results from classical group sequential design theory. This perspective showed that Bayes factor stopping rules correspond to sets of hyper-rectangles regions in the space of accumulating z-statistics. Under the canonical z-statistic distribution, the probability of these regions can be computed efficiently using multivariate normal integration, eliminating the need for simulation. The resulting computations are fast, accurate, and scalable to

designs with many interim looks. The approach also naturally incorporates design priors, enabling experiments to account for parameter uncertainty at the design stage. Traditional fixed parameter calculations are a special case, thus enabling flexible exploration of both Bayesian and frequentist operating characteristics within a unified framework.

There are, however, some limitations. The method requires that the Bayes factor can be expressed as a function of the z-statistic and that the accumulating z-statistics follow, at least approximately, a canonical multivariate normal distribution. Situations where these assumptions are not met (e.g., binary data with small sample size and/or extreme probabilities) may still require simulation or other numerical methods for computing design characteristics. Developing modifications for these situations could be one avenue for future research.

To conclude, many sequential Bayes factor designs can be planned as rapidly and reliably as classical group sequential designs. Our accompanying R package `bfpwr` implements these methods and offers experimenters a practical tool for designing efficient and informative studies.

Acknowledgments

We thank Wong Tsz Keung, Riko Kelter, and František Bartoš for valuable comments on drafts of the manuscript. We thank Torsten Hothorn for help with calculating multivariate normal probabilities with `mvtnorm::lpmvnorm`. We thank Tony Pourmohamad for pointing us to the data from the rat experiment. The acknowledgment of these individuals does not imply their endorsement of the paper.

Conflict of interest

We declare no conflict of interest.

Software and data

Data from the Low-PV trial were extracted from Table S5 in the supplement of Barbui et al. (2021). Data from the weight loss rat experiment were extracted from Figure 4 in Kang et al. (2025). Code and data to reproduce our analyses are openly available at `https://github.com/SamCH93/bfgsd`. A snapshot of the repository at the time of writing is available at `https://doi.org/10.5281/zenodo.XXXXXX`. We used the statistical programming language R version 4.5.2 (2025-10-31) for analyses (R Core Team, 2025) along with the `ggplot2` (Wickham, 2016), `dplyr` (Wickham et al., 2023), `mvtnorm` (Genz and Bretz, 2009), `ggpubr` (Kassambara, 2023), `ggrain` (Allen et al., 2021), `xtable` (Dahl et al., 2019), `rpact` (Wassmer and Brannath, 2016), and `knitr` (Xie, 2015) packages.

References

M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, J. van Langen, and R. A. Kievit. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Research*, 4(63), 2021. doi:10.12688/wellcomeopenres.15191.2.

S. F. Anderson and K. Kelley. Sample size planning for replication studies: The devil is in the design. *Psychological Methods*, 2022. doi:10.1037/met0000520.

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235, 1969. doi:10.2307/2343787.

T. Barbui, A. M. Vannucchi, V. De Stefano, A. Masciulli, A. Carobbio, A. Ferrari, A. Ghirardi, E. Rossi, F. Ciceri, M. Bonifacio, A. Iurlo, F. Palandri, G. Benevolo, F. Pane, A. Ricco, G. Carli, M. Caramella, D. Rapezzi, C. Musolino, S. Siragusa, E. Rumi, A. Patriarca, N. Cascavilla, B. Mora, E. Cacciola, C. Mannarelli, G. G. Loscocco, P. Guglielmelli, S. Betti, F. Lunghi, L. Scaffidi, C. Bucelli, N. Vianelli, M. Bellini, M. C. Finazzi, G. Tognoni, and A. Rambaldi. Ropeginterferon alfa-2b versus phlebotomy in low-risk patients with polycythaemia vera (Low-PV study): a multicentre, randomised phase 2 trial. *The Lancet Haematology*, 8(3):e175–e184, 2021. doi:10.1016/s2352-3026(20)30373-2.

S. M. Berry, B. P. Carlin, J. J. Lee, and P. Muller. *Bayesian Adaptive Methods for Clinical Trials*. Chapman & Hall/CRC, 2010.

J. M. Bland. Statistics notes: The odds ratio. *BMJ*, 320(7247):1468–1468, 2000. doi:10.1136/bmj.320.7247.1468.

G. Campbell. FDA regulatory acceptance of Bayesian statistics. In *Bayesian Methods in Pharmaceutical Research*, pages 41–51. Chapman and Hall/CRC, 2020. ISBN 9781315180212. doi:10.1201/9781315180212-2.

J. Cornfield. Sequential Trials, Sequential Analysis and the Likelihood Principle. *The American Statistician*, 20:18–23, 1966. doi:10.1080/00031305.1966.10479786.

J. Cornfield. Recent methodological contributions to clinical trials. *American Journal of Epidemiology*, 104(4):408–421, 1976. doi:10.1093/oxfordjournals.aje.a112313.

D. B. Dahl, D. Scott, C. Roosen, A. Magnusson, and J. Swinton. *xtable: Export Tables to LaTeX or HTML*, 2019. URL `https://CRAN.R-project.org/package=xtable`. R package version 1.8-4.

A. Deng, J. Lu, and S. Chen. Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 243–252, 2016. doi:10.1109/dsaa.2016.33.

N. I. Drude, L. Martinez Gamboa, M. Danziger, U. Dirnagl, and U. Toelch. Improving preclinical studies through replications. *eLife*, 10, 2021. doi:10.7554/elife.62101.

S. S. Ellenberg, T. R. Fleming, and D. L. DeMets. *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. John Wiley & Sons, 2nd edition, 2019. doi:10.1002/9781119512684.

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg, 2009. ISBN 978-3-642-01688-2.

S. N. Goodman. Introduction to Bayesian methods I: measuring the strength of evidence. *Clinical Trials*, 2(4):282–290, 2005. doi:10.1191/1740774505cn098oa.

Q. F. Gronau, A. Ly, and E.-J. Wagenmakers. Informed Bayesian *t*-tests. *The American Statistician*, 74(2):137–143, 2020. doi:10.1080/00031305.2018.1562983.

T. Gsponer, F. Gerber, B. Bornkamp, D. Ohlssen, M. Vandemeulebroecke, and H. Schmidli. A practical guide to Bayesian group sequential designs. *Pharmaceutical Statistics*, 13(1):71–80, 2013. doi:10.1002/pst.1593.

L. Held and M. Ott. On *p*-values and Bayes factors. *Annual Review of Statistics and Its Application*, 5(1), 2018. doi:10.1146/annurev-statistics-031017-100307.

J. Jack Lee and C. T. Chu. Bayesian clinical trials in action. *Statistics in Medicine*, 31(25):2955–2972, 2012. doi:10.1002/sim.5404.

H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 1961. 3rd edition.

C. Jennison and B. W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall, 1999.

V. E. Johnson. Bayes factors based on test statistics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(5):689–701, 2005. doi:10.1111/j.1467-9868.2005.00521.x.

V. E. Johnson and J. D. Cook. Bayesian design of single-arm phase II clinical trials with continuous monitoring. *Clinical Trials*, 6(3):217–226, 2009. doi:10.1177/1740774509105221.

J. A. Kairalla, C. S. Coffey, M. A. Thomann, and K. E. Muller. Adaptive trial designs: a review of barriers and opportunities. *Trials*, 13(1), 2012. doi:10.1186/1745-6215-13-145.

J. Kang, T. Koulis, and T. Pourmohamad. Sample size reduction in preclinical experiments: A Bayesian sequential decision-making framework. *Journal of Biopharmaceutical Statistics*, pages 1–16, 2025. doi:10.1080/10543406.2025.2556680.

R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430): 773–795, June 1995. doi:10.1080/01621459.1995.10476572.

A. Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2023. URL `https://CRAN.R-project.org/package=ggpubr`. R package version 0.6.0.

W. Li, M.-H. Chen, X. Wang, and D. K. Dey. Bayesian design of non-inferiority clinical trials via the Bayes factor. *Statistics in Biosciences*, 10(2):439–459, 2017. doi:10.1007/s12561-017-9200-5.

M. Linde and D. van Ravenzwaaij. baymedr: an R package and web application for the calculation of Bayes factors for superiority, equivalence, and non-inferiority designs. *BMC Medical Research Methodology*, 23(1), 2023. doi:10.1186/s12874-023-02097-y.

M. Lindon and A. Malek. Anytime-valid inference for multinomial count data. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2817–2831, 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/12f3bd5d2b7d93eadc1bf508a0872dc2-Paper-Conference.pdf`.

N. Mani, M. S. Schreiner, J. Brase, K. Köhler, K. Strassen, D. Postin, and T. Schultze. Sequential Bayes factor designs in developmental research: Studies on early word learning. *Developmental Science*, 24(4), 2021. doi:10.1111/desc.13097.

J. N. Matthews. *Introduction to Randomized Controlled Clinical Trials*. Chapman & Hall/CRC, second edition, 2006.

C. Micheloud and L. Held. Power calculations for replication studies. *Statistical Science*, 37(3): 369–379, 2022. doi:10.1214/21-sts828.

M. Moerbeek. Bayesian updating: increasing sample size during the course of a study. *BMC Medical Research Methodology*, 21(1), 2021. doi:10.1186/s12874-021-01334-6.

A. O'Hagan and J. Stevens. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, 21(3):219–230, 2001. doi:10.1177/02729890122062514.

S. Pawel and L. Held. The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):879–911, 2022. doi:10.1111/rssb.12491.

S. Pawel and L. Held. Closed-form power and sample size calculations for Bayes factors. *The American Statistician*, 79(3):330–344, 2025. doi:10.1080/00031305.2025.2467919.

S. Pawel, G. Consonni, and L. Held. Bayesian approaches to designing replication studies. *Psychological Methods*, 2023. doi:10.1037/met0000604.

S. K. Piper, U. Grittner, A. Rex, N. Riedel, F. Fischer, R. Nadon, B. Siegerink, and U. Dirnagl. Exact replication: Foundation of science or game of chance? *PLOS Biology*, 17(4):e3000188, 2019. doi:10.1371/journal.pbio.3000188.

T. Pourmohamad and C. Wang. Sequential Bayes factors for sample size reduction in preclinical experiments with binary outcomes. *Statistics in Biopharmaceutical Research*, 15(4):706–715, 2022. doi:10.1080/19466315.2022.2123386.

M. A. Psioda and J. G. Ibrahim. Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*, 20(3):400–415, 2018. doi:10.1093/biostatistics/kxy009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. URL `https://www.R-project.org/`.

D. S. Robertson, B. Choodari-Oskooei, M. Dimairo, L. Flight, P. Pallmann, and T. Jaki. Point estimation for adaptive trial designs I: A methodological review. *Statistics in Medicine*, 42(2): 122–145, 2022. doi:10.1002/sim.9605.

G. L. Rosner. Bayesian adaptive designs in drug development. In *Bayesian Methods in Pharmaceutical Research*, pages 161–184. Chapman and Hall/CRC, 2020. doi:10.1201/9781315180212-8.

G. L. Rosner, P. W. Laud, and W. O. Johnson. *Bayesian Thinking in Biostatistics*. Chapman and Hall/CRC, 2021. ISBN 9781439800102. doi:10.1201/9781439800102.

J. N. Rouder, P. L. Speckman, D. Sun, R. D. Morey, and G. Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, 2009. doi:10.3758/pbr.16.2.225.

R. Royall. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London New York, 1997. ISBN 9780412044113.

W. M. S. Russell and R. L. Burch. *The Principles of Humane Experimental Technique*. Methuen, London, U.K., 1959.

E. G. Ryan, K. Brock, S. Gates, and D. Slade. Do we need to adjust for interim analyses in a Bayesian adaptive trial design? *BMC Medical Research Methodology*, 20(1), 2020. doi:10.1186/s12874-020-01042-7.

F. D. Schönbrodt and E.-J. Wagenmakers. Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1):128–142, mar 2018. doi:10.3758/s13423-017-1230-y.

F. D. Schönbrodt, E.-J. Wagenmakers, M. Zehetleitner, and M. Perugini. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2): 322–339, 2017. doi:10.1037/met0000061.

F. D. Schönbrodt and A. M. Stefan. *BFDA: An R package for Bayes factor design analysis (version 0.5.0)*, 2019. URL `https://github.com/nicebread/BFDA`.

D. J. Spiegelhalter, R. Abrams, and J. P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley, 2004.

A. M. Stefan, Q. F. Gronau, F. D. Schönbrodt, and E.-J. Wagenmakers. A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods*, 51(3):1042–1058, 2019. doi:10.3758/s13428-018-01189-8.

A. M. Stefan, F. D. Schönbrodt, N. J. Evans, and E.-J. Wagenmakers. Efficiency in sequential testing: Comparing the sequential probability ratio test and the sequential Bayes factor test. *Behavior Research Methods*, 54(6):3100–3117, 2022. doi:10.3758/s13428-021-01754-8.

A. M. Stefan, Q. F. Gronau, and E.-J. Wagenmakers. Interim design analysis using Bayes factor forecasts. *Psychological Methods*, 2024. doi:10.1037/met0000641.

A. Stevely, M. Dimairo, S. Todd, S. A. Julious, J. Nicholl, D. Hind, and C. L. Cooper. An investigation of the shortcomings of the CONSORT 2010 statement for the reporting of group sequential randomised controlled trials: A methodological systematic review. *PLOS ONE*, 10(11):e0141104, 2015. doi:10.1371/journal.pone.0141104.

U.S. Food and Drug Administration. Guidance for the use of Bayesian statistics in medical device clinical trials, 2010. URL `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials`.

J. Verhagen and E.-J. Wagenmakers. Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology*, 143:1457–1475, 2014. doi:10.1037/a0036731.

A. Wald. *Sequential Analysis*. Wiley, New York, 1947.

F. Wang and A. E. Gelfand. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17 (2):193–208, 2002. doi:10.1214/ss/1030550861.

G. Wassmer and W. Brannath. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer, New York, 2016.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023. URL `https://CRAN.R-project.org/package=dplyr`. R package version 1.1.4.

T. K. Wong and J. N. Tendeiro. On a generalizable approach for sample size determination in Bayesian *t* tests. *Behavior Research Methods*, 57(5), 2025. doi:10.3758/s13428-025-02654-x.

Y. Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL `https://yihui.org/knitr/`. ISBN 978-1498716963.

A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa*, 31(1):585–603, 1980. doi:10.1007/bf02888369.

T. Zhou and Y. Ji. On Bayesian sequential clinical trial designs. *The New England Journal of Statistics in Data Science*, pages 136–151, 2023. doi:10.51387/23-nejsds24.

Y. Zhou, R. Lin, and J. J. Lee. The use of local and nonlocal priors in Bayesian test-based monitoring for single-arm phase II clinical trials. *Pharmaceutical Statistics*, 20(6):1183–1199, 2021. doi:10.1002/pst.2139.

L. Zhu, Q. Yu, and D. E. Mercante. A Bayesian sequential design for clinical trials with time-to-event outcomes. *Statistics in Biopharmaceutical Research*, 11(4):387–397, 2019. doi:10.1080/19466315.2019.1629996.

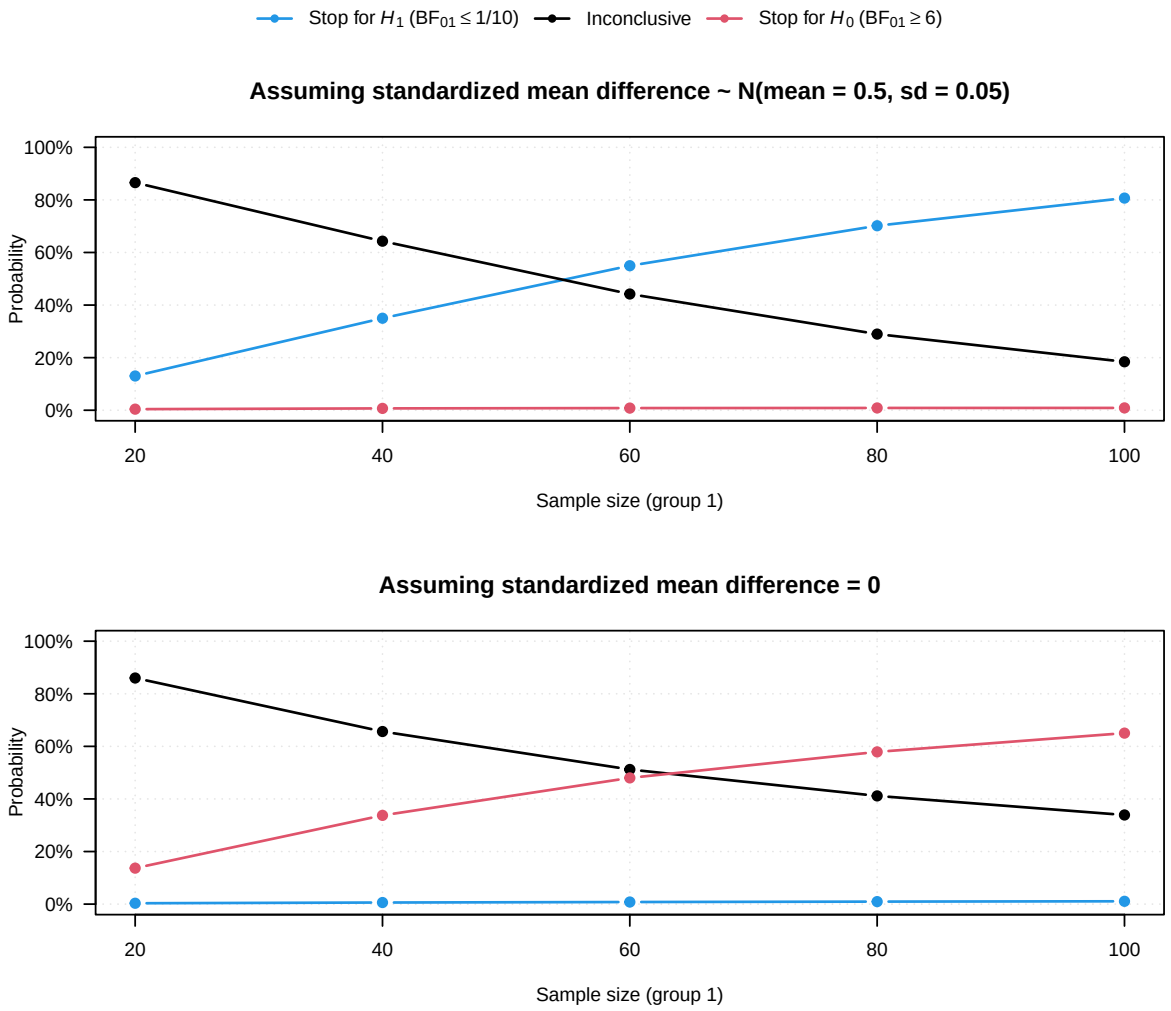
Appendix A The R package bfpwr

The following code excerpt shows how the `bfpwr` R package can be used to compute design characteristics of a sequential JZS (*t*-test) Bayes factor design.

```r
## group sequential design features not in CRAN version yet
remotes::install_github(repo = "SamCH93/bfpwr", subdir = "package", ref = "gsd")
library(bfpwr) # load package
## set up sequential t-test Bayes factor design
design <- ptbf01seq(
    k1 = 1/10, # Bayes factor threshold for H1
    k0 = 6, # Bayes factor threshold for H0
    type = "two.sample", # two-sample t-test
    n = seq(20, 100, 20), # per-group sample sizes at analyses
    ## specify one-sided Jeffreys-Zellner-Siow analysis prior
    plocation = 0, pscale = 1/sqrt(2), pdf = 1, alternative = "greater",
    ## specify normal design prior around SMD = 0.5 with small stand. deviation
    dpm = 0.5, dpsd = 0.05
)
design # print design summary

##
## Sequential Bayes Factor Design
## --------------------------------
## H0:                SMD (stand. mean diff.)  = 0
## H1:                SMD (stand. mean diff.)  > 0
## Analysis prior:    SMD|H1 ~ t(location = 0, scale = 0.7071, df = 1)_+
## Design prior:      SMD ~ N(mean = 0.5, sd = 0.05)
## BF thresholds:     H1 if BF01 <= 1/10, H0 if BF01 >= 6
## Number of looks:   5
## Sample sizes 1:    20, 40, 60, 80, 100
## Sample sizes 2:    20, 40, 60, 80, 100
##
##
## Stagewise cumulative probabilities:
##  Stage Pr(H1 stop) Pr(H0 stop) Pr(inconclusive)
##      1      0.1302      0.0041           0.8656
##      2      0.3500      0.0070           0.6430
##      3      0.5497      0.0082           0.4421
##      4      0.7017      0.0087           0.2897
##      5      0.8068      0.0088           0.1843
##
## Expected sample size 1: 64.8083
## Expected sample size 2: 64.8083
## Standard deviation of sample size 1: 28.3783
## Standard deviation of sample size 2: 28.3783
##
## NOTE:  BF01 < 1 indicates evidence for H1 over H0

plot(design) # plot design under design prior (top) and under H0 (bottom)
```



Appendix B Marginal distribution of the *z*-statistics

The z-statistic vector can be represented as $\boldsymbol{Z} \mid \theta = \theta \boldsymbol{I} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathrm{N}_m(\boldsymbol{0}, \boldsymbol{\Sigma})$ and independent of θ. Since $\boldsymbol{I}$ is fixed and also $\theta \sim \mathrm{N}(\mu_d, \tau_d^2)$, the marginal distribution of $\boldsymbol{Z}$ is also normal. By the law of total expectation, its expectation is

$$\begin{aligned} \mathrm{E}(\boldsymbol{Z}) &= \mathrm{E}\{\mathrm{E}(\boldsymbol{Z} \mid \theta)\} \\ &= \mathrm{E}(\theta \boldsymbol{I}) \\ &= \mathrm{E}(\theta) \boldsymbol{I} \\ &= \mu_d \boldsymbol{I}. \end{aligned}$$

Similarly, applying the law of total covariance, its covariance is

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{Z}) &= \mathrm{E}\{\mathrm{Cov}(\boldsymbol{Z} \mid \theta)\} + \mathrm{Cov}\{\mathrm{E}(\boldsymbol{Z} \mid \theta)\} \\ &= \mathrm{E}(\boldsymbol{\Sigma}) + \mathrm{Cov}(\theta \boldsymbol{I}) \\ &= \boldsymbol{\Sigma} + \boldsymbol{I}\,\mathrm{Cov}(\theta)\boldsymbol{I}^\top \\ &= \boldsymbol{\Sigma} + \tau_d^2 \boldsymbol{I}\boldsymbol{I}^\top.\end{aligned}$$

Appendix C Sensitivity analysis for rat experiment

Figure 9 and Table 3 show results from sensitivity analyses regarding the random permutation of the data set from Kang et al. (2025), for which the original collection order is unknown. The 11 rats saved reported in Section 4.2 are representative, or even a conservative estimate compared to the the permutation distribution in Figure 9. Similary,

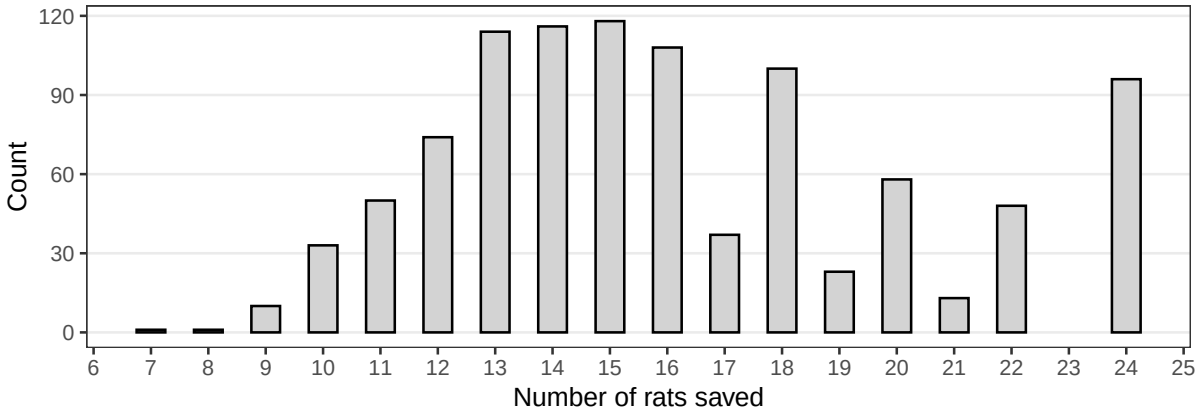


Figure 9: Number of rats saved across 1'000 random permutations of the data order when using a sequential Bayes factor analysis.

the decisions to stop for H_0 in the low dose and to stop for H_1 in the other groups reported in Section 4.2 corresponds to the majority of decisions across the permutations (Table 3).

Table 3: Proportion of sequential Bayes factor stopping decisions for H_0 or H_1 across 1'000 random permutations of the data order.

Treatment group	Stop for H_0	Stop for H_1	Indecisive
Low	89.5%	10.5%	0%
Medium	0%	100%	0%
High	0%	100%	0%

Computational details

```
cat(paste(Sys.time(), Sys.timezone(), "\n"))

## 2026-01-06 09:22:22.311216 Europe/Zurich

sessionInfo()

## R version 4.5.2 (2025-10-31)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 24.04.3 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0  LAPACK version 3.12.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Zurich
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] ggrain_0.1.0  tidyr_1.3.1   dplyr_1.1.4   scales_1.4.0  rpact_4.3.0
##  [6] ggpubr_0.6.2  ggplot2_4.0.1 mvtnorm_1.3-3 xtable_1.8-4  bfpwr_0.2
## [11] knitr_1.50
##
## loaded via a namespace (and not attached):
##  [1] rappdirs_0.3.3        generics_0.1.4        rstatix_0.7.3
##  [4] magrittr_2.0.4        evaluate_1.0.5        grid_4.5.2
##  [7] RColorBrewer_1.1-3    backports_1.5.0       Formula_1.2-5
## [10] purrr_1.2.0           viridisLite_0.4.2     abind_1.4-8
## [13] cli_3.6.5             ggpp_0.5.9            rlang_1.1.6
## [16] cowplot_1.2.0         remotes_2.5.0         withr_3.0.2
## [19] rootSolve_1.8.2.4     tools_4.5.2           polynom_1.4-1
## [22] ggsignif_0.6.4        qrng_0.0-10           curl_7.0.0
## [25] broom_1.0.10          vctrs_0.6.5           R6_2.6.1
## [28] lifecycle_1.0.4       car_3.1-3             MASS_7.3-65
```

```
## [31] pkgconfig_2.0.3       RcppParallel_5.1.11-1 pillar_1.11.1
## [34] gtable_0.3.6          glue_1.8.0            Rcpp_1.1.0
## [37] xfun_0.54             tibble_3.3.0          tidyselect_1.2.1
## [40] highr_0.11            farver_2.1.2          labeling_0.4.3
## [43] carData_3.0-5         lamW_2.2.5            compiler_4.5.2
## [46] S7_0.2.1
```