UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Informed Bayesian survival analysis

Bartoš, F.; Aust, F.; Haaf, J.M.

DOI
10.1186/s12874-022-01676-9

Publication date
2022

Document Version
Final published version

Published in
BMC Medical Research Methodology

License
CC BY

Link to publication

Citation for published version (APA):
Bartoš, F., Aust, F., & Haaf, J. M. (2022). Informed Bayesian survival analysis. *BMC Medical Research Methodology*, *22*, Article 238. https://doi.org/10.1186/s12874-022-01676-9

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Oct 2026

Bartoš *et al.*
BMC Medical Research Methodology (2022) 22:238
https://doi.org/10.1186/s12874-022-01676-9

BMC Medical Research Methodology

RESEARCH

Open Access

Informed Bayesian survival analysis

František Bartoš[1,2*], Frederik Aust[1] and Julia M. Haaf[1]

Abstract

Background: We provide an overview of Bayesian estimation, hypothesis testing, and model-averaging and illustrate how they benefit parametric survival analysis. We contrast the Bayesian framework to the currently dominant frequentist approach and highlight advantages, such as seamless incorporation of historical data, continuous monitoring of evidence, and incorporating uncertainty about the true data generating process.

Methods: We illustrate the application of the outlined Bayesian approaches on an example data set, retrospective re-analyzing a colon cancer trial. We assess the performance of Bayesian parametric survival analysis and maximum likelihood survival models with AIC/BIC model selection in fixed-n and sequential designs with a simulation study.

Results: In the retrospective re-analysis of the example data set, the Bayesian framework provided evidence for the absence of a positive treatment effect of adding Cetuximab to FOLFOX6 regimen on disease-free survival in patients with resected stage III colon cancer. Furthermore, the Bayesian sequential analysis would have terminated the trial 10.3 months earlier than the standard frequentist analysis. In a simulation study with sequential designs, the Bayesian framework on average reached a decision in almost half the time required by the frequentist counterparts, while maintaining the same power, and an appropriate false-positive rate. Under model misspecification, the Bayesian framework resulted in higher false-negative rate compared to the frequentist counterparts, which resulted in a higher proportion of undecided trials. In fixed-n designs, the Bayesian framework showed slightly higher power, slightly elevated error rates, and lower bias and RMSE when estimating treatment effects in small samples. We found no noticeable differences for survival predictions. We have made the analytic approach readily available to other researchers in the RoBSA R package.

Conclusions: The outlined Bayesian framework provides several benefits when applied to parametric survival analyses. It uses data more efficiently, is capable of considerably shortening the length of clinical trials, and provides a richer set of inferences.

Keywords: Bayesian, Survival analysis, Model-averaging, Bayes factor, Historical data

Background

There has been a steady increase in the popularity and interest in Bayesian statistics in the past years [1]. In this paper, we leverage the advantages of the longstanding Bayesian estimation [2], hypothesis testing [3], and model-averaging [4] approaches and apply them to parametric survival analysis. This type of Bayesian survival analysis is possible thanks to the recent development of flexible tools for fitting Bayesian models (such as JAGS [5] and Stan [6]) and efficient techniques for estimating marginal likelihoods (such as bridge-sampling [7–9]).

Survival analysis is a frequently used method with important applications in evaluating lifetime outcomes in clinical trials [10, 11]. The most commonly used version of survival analysis is the non-parametric Cox proportional hazard model, which does not require specification of the baseline hazard [12]. There are,

*Correspondence: f.bartos96@gmail.com

[1] Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

however, good reasons to use parametric survival models instead: (1) generalizability, the models can be easily extended to deal with interval censored data [13], (2) simplicity, the models can be defined using only few parameters, (3) completeness, both hazard and survival functions are fully specified, (4) consistency with theoretical survival functions [11], and (5) the ability to extrapolate survival functions beyond the measurement time frame [14] (predictions for the Cox's proportional hazard model can be obtained by combination with the Breslow's [15] estimate of baseline hazard). For these reasons, we focus on parameteric survival models here. Yet, the advantages of parametric models come at the cost of additional assumptions about the data generating process—assumptions that can result in model missspecification, which can be addressed with the Bayesian approach outlined here.

The Bayesian approaches offer multiple benefits to parametric survival analysis, which we elaborate below: (1) the seamless incorporation of external knowledge, (2) the possibility to monitor the evidence continuously, and (3) the possibility to embrace uncertainty about the data generating process.[1]

Bayesian estimation allows us to seamlessly incorporate external knowledge into statistical models via prior distributions [16–19] (see [20] for frequentist alternatives). Incorporating either historical data or expert opinions is not a novel concept. In medicine, it was proposed more than 45 years ago [21] and repeatedly advocated for [22–25]. Such external knowledge can improve the precision of estimates, lower error rates, grant better small sample properties, and improve the precision of survival estimates [24, 26–31]. While improper incorporation of external knowledge might bias estimates and increase error rate [32], safeguards against adverse effects of external knowledge exist. For example, researchers can use meta-analytic predictive priors [30] that incorporate the information about between-study heterogeneity to adjust for dissimilarities to previous studies [33].

Bayesian hypothesis testing allows us to continuously monitor the evidence in favor of (against) a hypothesis with Bayes factors [34–37]. Bayes factors quantify the relative evidence for two competing hypotheses, which stands in contrast to frequentist hypothesis tests that reference hypothetical error rates under the assumption that the null hypothesis is true [38] (see 'Bayesian Evidence' section for detailed treatment). Moreover, where frequentist alternatives usually examine the data only a limited, pre-specified number of times [39, 40], Bayes factor optional stopping is capable of monitoring the evidence in a truly continuous manner [36, 41] (see Lan-DeMets spending function for an alternative [42]). This is advantageous because continuous monitoring may increases the efficiency of clinical studies that are not only expensive but also costly in terms of life and harm [43–45]. As shown in different settings, the Bayes factor sequential analysis might further increase the benefits of frequentist group sequential designs [34, 46–48]. Note that Bayesians can continuously monitor evidence via Bayes factors but not posterior parameter distributions as sometimes claimed [22, 23, 49] (see Ibrahim and colleagues [50] for details).

Finally, Bayesian model-averaging (BMA) [51–53] allows us to embrace uncertainty about the data generating process by basing inference on multiple models simultaneously[2] (an alternative is using Akaike weights for frequentist model averaging [54–58] or to use smoothing approach [13]). BMA is especially relevant for the parametric survival analysis where models based on different parametric families may not lead to the same conclusions, estimates, and/or predictions [14, 59, 60]. BMA also simplifies the analysis for researchers. Instead of drawing inference informally after inspecting the results from various model specifications and subjectively evaluating their fit [10, 29, 55, 61], BMA allows researchers to automatically combine models based on their posterior model probabilities—that is their suitability to the current application.

Despite all the advantages, Bayesian survival analysis is rarely used in practice. E.g., external information rarely enters the analysis via informed priors on parameters of interest [62] and only about 15% of studies in pediatric medicine use historical information [63]. While sequential analyses are common in medicine, they are rarely based on Bayes factors [50, 64]. And even though the advantages of BMA are recognized [65, 66], Gallacher, Auguste, and Connock [67] found that BMA was not used in any of the recent 14 appraisals performing extrapolation based on survival models.

We suspect that the outlined Bayesian approaches remain under-utilized because researchers are relatively unfamiliar with them and because easily accessible software implementations are currently not available [34, 68]. Moreover, there is a noteworthy lack of official FDA and EMA guidelines for Bayesian analyses [69]. The goal of this paper is therefore three-fold. First, we review the Bayesian approaches for survival

[1] Many arguments raised in this paper apply to the non-and semi-parametric survival models as well (see [50, 102, 132, 133]) and sources therein for Bayesian versions of non-and semi-parametric survival models).

[2] Whereas some empirical Bayesian literature uses Bayesian information criteria for "pseudo" BMA [e.g., 65, 66], we perform proper BMA based on marginal likelihoods which allow us to properly incorporate prior information and test informed hypotheses.

analysis, including Bayesian estimation, Bayesian hypothesis testing, and Bayesian model-averaging. We discuss the critical differences between Bayesian and frequentist frameworks to quantify evidence and how to reconcile them. Second, we apply both frameworks and showcase their respective benefits with an example from a colon cancer trial [70]. We demonstrate that incorporating historical data and specifying an informed hypothesis could shorten the trial by more than a year and possibly increase the participants' progress-free survival by 1,299 years in total. Finally, we support our claims with a simulation study. We show that the Bayesian framework can decrease the duration of sequential trials by almost half, slightly increase power in fixed-n designs, and improve the precision of treatment effect estimates in small samples. The only downside is a minor impact on the false-negative rate under model misspecification. We make the methodology available to the research community by implementing the analyses in the RoBSA R package [71].

Bayesian survival analysis

In this section, we outline a coherent, fully Bayesian framework for survival analysis. We start with Bayesian estimation, move towards the less common Bayesian hypothesis testing, and extend both estimation and testing to multi-model inference with Bayesian model-averaging (BMA; see [51, 72] for in-depth tutorials on BMA). For simplicity, we use a single treatment variable, right censoring, and parametric models with accelerated failure times (AFT) specifications which later allows us to obtain a model-averaged estimate of the acceleration factor (AF) across all specified models.[3] The AFT models assume that the ratio of the survival times between groups, the acceleration factor is constant over time. An AF larger than one indicates a longer than expected survival for a given group (in contrast to proportional hazard models, PH, where a higher hazard ratio, HR, means an increased risk of the event). More specific topics, such as comparing the interpretation of Bayes factors and *p*-values, specifying prior parameter distributions, prior model probabilities, or Bayes Factor Design Analysis are outlined in the 'Bayesian Evidence' and 'Example' sections.

Bayesian estimation

Following the standard notation, we use $S_d(.)$, and $h_d(.)$ to denote the survival and hazard function, respectively, of a parametric family d (e.g., exponential, Weibull, log-normal, log-logistic, or gamma) that describes the observed survival times t_i with the censoring indicator c_i ($c_i = 1$ for observed events) for each participant i. We use β to denote the treatment effect of the dummy coded treatment x_i, and use α_d and γ_d for intercepts and auxiliary parameters (if applicable). The likelihood of the data $= \{t, c\}$ under a survival model M_d given the parameters $\theta_d = \{\beta, \alpha_d, \gamma_d\}$ can be written as:

$$p(\text{data}|\theta_d, \mathcal{M}_d) = \prod h_d(t_i|x_i, \theta_d)^{I(c_i=1)} \times S_d(t_i|x_i, \theta_d). \tag{1}$$

We finish the model specification by assigning prior distributions ($\mathrm{p}(\theta_\mathrm{d} \mid \mathrm{M}_\mathrm{d})$) to each parameter ($\beta \sim \mathrm{f}_\beta(.)$, $\alpha_\mathrm{d} \sim \mathrm{f}_{\alpha,\mathrm{d}}(.)$, and $\gamma_\mathrm{d} \sim \mathrm{f}_{\gamma,\mathrm{d}}(.)$; see 'Model Specification' subsection of the 'Example' section for more details about specifying prior distributions) and obtain the posterior parameter distributions according to the Bayes theorem:

$$p(\theta_d|\text{data}, \mathcal{M}_d) = \frac{p(\text{data}|\theta_d, \mathcal{M}_d) \times p(\theta_d|\mathcal{M}_d)}{p(\text{data}|\mathcal{M}_d)}, \tag{2}$$

where $p(\text{data} \mid \mathrm{M}_d)$ denotes the marginal likelihood, an integral of the likelihood weighted by the prior parameter distributions over the whole parameter space,

$$p(\text{data}|\mathcal{M}_d) = \int_{\theta_d} p(\text{data}|\theta_d, \mathcal{M}_d) \times p(\theta_d|\mathcal{M}_d) d\theta_d, \tag{3}$$

which also quantifies the models' prior predictive performance for the observed data [73].

Bayesian hypothesis testing

Whereas Bayesian estimation allows us to obtain posterior parameter distributions assuming the treatment has an effect, it does not quantify the evidence in favor of presence/absence of the treatment effect. In other words, in order to test the hypothesis that a treatment effect is non-zero, one must compare a model assuming absence of the effect to a model assuming presence of the effect [41, 74]. To do that, we adopt Sir Harold Jeffreys' Bayesian hypothesis testing framework [3, 75] and split the specified models into two variants. Models assuming the absence of the treatment effect ($\beta = 0$), $\mathrm{M}_{0,d}$, and models assuming the presence of the treatment effect ($\beta \sim f_\beta(.)$), $\mathrm{M}_{1,d}$. In the following equations we explicitly, and somewhat unconventionally, condition on the parametric family d to highlight that the results depend on this particular choice. We assign prior model probabilities, $p(\mathrm{M}_d \mid d)$, to each variant of the model, and apply Bayes rule one more time to obtain the posterior model probabilities,

[3] The framework can be easily generalized to multiple covariates, left and interval censoring, and frailties. Furthermore, PH models can also be incorporated into the framework and they can be used to (a) jointly assess the evidence for either the AFT or PH effect of the treatment but also (b) to test which assumption is more plausible.

$$p(\mathcal{M}_{0,d}|\text{data}, d) = \frac{p(\text{data}|\mathcal{M}_{0,d}, d) \times p(\mathcal{M}_{0,d}|d)}{p(\text{data}|d)},$$
$$p(\mathcal{M}_{1,d}|\text{data}, d) = \frac{p(\text{data}|\mathcal{M}_{1,d}, d) \times p(\mathcal{M}_{1,d}|d)}{p(\text{data}|d)}, \quad (4)$$

where $p(\text{data}|d)$ follows from the law of total probability,

$$p(\text{data} \mid \text{d}) = p(\text{data}|\mathcal{M}_{0,d}, d) \times p(\mathcal{M}_{0,d}|d) + p(\text{data}|\mathcal{M}_{1,d}, d) \times p(\mathcal{M}_{1,d}|d). \quad (5)$$

More importantly, Bayesian hypothesis testing allows us to quantify the evidence for either of the models, irrespective of the prior model probabilities with Bayes factors (BF) [4, 76–78],

$$\text{BF}_{10} = \frac{p(\text{data}|\mathcal{M}_{1,d}, d)}{p(\text{data}|\mathcal{M}_{0,d}, d)}, \quad (6)$$

as a ratio of marginal likelihoods. Bayes factors are a continuous measure of evidence and their size can be directly interpreted as the attained support for one of the models over the other model.

As can be seen from Equation 3, the model comparison is determined by both data and the prior distributions for the parameters, which effectively specify the model comparison/hypothesis test. However, since both models contain the same prior distributions for α_d and γ_d, the BF_{10} depends only on the prior distribution for β—the treatment effect we intended to test.

The Bayes factor quantifies the updating rate from prior to posterior model probabilities. Consequently, we can reformulate the Equation 6 as the change from prior model odds to posterior model odds [78, 79],

$$\frac{p(\text{data}|\mathcal{M}_{1,d}, d)}{p(\text{data}|\mathcal{M}_{0,d}, d)} = \frac{p(\mathcal{M}_{1,d}|\text{data}, d)}{p(\mathcal{M}_{0,d}|\text{data}, d)} \Big/ \frac{p(\mathcal{M}_{1,d}|d)}{p(\mathcal{M}_{0,d}|d)}, \quad (7)$$

which is useful when comparing multiple models simultaneously.

Bayesian model-averaging

So far, we summarized how to obtain the posterior parameter distribution with Bayesian *estimation* and how to evaluate evidence for the presence vs. absence of an effect with Bayesian *testing*. Now, we expand both Bayesian estimation and testing with Bayesian *model-averaging* (BMA), which allows us to relax the commitment to a single set of auxiliary assumptions [80]. This assumption is visible in Equation 5, which shows that all inference is conditional on the assumed data generating process, the specific parametric family d (which, of course, also applies to the corresponding frequentist analysis). We relax this assumption by specifying multiple models based on different parametric families and combining them according to their relative predictive performance [51–53]. In this way, our posterior parameter distributions and evidence for the absence vs. presence of the treatment effect are no longer based on the assumption of one particular data generating mechanism. In other words, "not putting all eggs into one basket" protects researchers from idiosyncrasies in the data and leads to more robust inference [81].

Bayesian model-averaged estimation To use BMA for estimation, we rely on models assuming the presence of the treatment effect from competing parametric families ($H_{1,d}$). We aim to obtain a posterior distribution of the treatment effect (assuming the treatment effect exists) that takes the uncertainty about the competing parametric families into account. Here, we limit ourselves to models that share an AFT parameterization and quantify the treatment effect as accelerated failure. The AF has the same interpretation in all parametric families, which allows us to directly combine the treatment effect estimates across all models into a single pooled estimate. In general, we could add models with different parameterizations, such as PH models which quantify the treatment effect as hazard ratio. In this case, we would pool the different measures of the treatment effect separately (and determine the posterior probabilities of each parameterization), however, we could still obtain survival and hazard functions pooled across all models.

We start by specifying prior model probabilities for each model and expanding Equation 5 to accommodate models of all parametric families assuming presence of the treatment effect,

$$p(\text{data}|\mathcal{M}_1) = \sum_{d=1}^{5} p(\text{data}|\mathcal{M}_{1,d}) \times p(\mathcal{M}_{1,d}). \quad (8)$$

Then we obtain posterior model probabilities for each model assuming presence of the treatment effect via Bayes theorem (analogously to Equation 4).

Because we focus only on models with AFT parameterization, the treatment effect β, and its prior distribution $f_\beta(.)$, can be specified interchangeably as log(AF) across models from all parametric families. The posterior model probabilities, based on the marginal likelihoods, are therefore independent of the common prior distribution on the treatment effect. Any difference in posterior probabilities among the models $\text{M}_{1,d}$, assuming the presence of an effect, reflect differences in their

prior predictive performance due to scaling and shapes of their survival time distributions. These differences flow from the parametric assumptions and the associated model-specific prior distributions for intercepts and auxiliary parameters.

Now, we can combine the posterior distributions of the treatment effect from the competing parametric families by weighting them according to the posterior model probabilities [3, 4]. The resulting model-averaged posterior distribution of the treatment effect then corresponds to a mixture distribution,

$$p(\beta|\text{data}, \mathcal{M}_1) = \sum_{d=1}^{5} p(\beta|\mathcal{M}_{1,d}, \text{data}) \times p(\mathcal{M}_{1,d}|\text{data}). \tag{9}$$

In the same manner, we also obtain the posterior model-averaged survival and hazard functions assuming the presence of the treatment effect,

$$\begin{aligned} S(.) &= \textstyle\sum_{d=1}^{5} S_{1,d}(.) \times p(\mathcal{M}_{1,d}|\text{data}) \\ h(.) &= \textstyle\sum_{d=1}^{5} h_{1,d}(.) \times p(\mathcal{M}_{1,d}|\text{data}). \end{aligned} \tag{10}$$

Bayesian model-averaged hypothesis testing To apply BMA to Bayesian hypothesis testing we compare models assuming the absence of a treatment effect to models assuming its presence for all distributional families. Our aim is to quantify the evidence for or against a treatment effect that takes the uncertainty about the competing parametric families into account.

We, again, start by specifying prior model probabilities for each model and by further expanding Equation 8 to also accommodate models assuming absence of the treatment effect,

$$p(\text{data}) = \sum_{m=0}^{1} \sum_{d=1}^{5} p(\text{data}|\theta_d, \mathcal{M}_{m,d}) \times p(\mathcal{M}_{m,d}), \tag{11}$$

where $m = 0$ indicates models assuming the absence of a treatment effect and $m = 1$ models assuming the presence of a treatment effect. We, again, obtain posterior model probabilities for each model via Bayes theorem (analogously to Equation 4).

In contrast to Bayesian model-averaged estimation, the set of models now also includes those that assume the absence of a treatment effect. That is, in addition to their parametric assumptions the models now differ with respect to the prior distribution for the treatment effect. This critical difference separates the two sets of models compared by the model-averaged inclusion Bayes factor for the treatment effect: (1) all models assuming the presence of a treatment effect (in the nominator) and (2) all models assuming the absence of the treatment effect (in the denominator) [51, 81],

$$\underbrace{\text{BF}_{10}}_{\text{Inclusion Bayes factor for effect}} = \underbrace{\frac{\sum_{d=1}^{5} p(\mathcal{M}_{1,d}|\text{data})}{\sum_{d=1}^{5} p(\mathcal{M}_{0,d}|\text{data})}}_{\text{Posterior inclusion odds for models assuming effect}} \Big/ \underbrace{\frac{\sum_{d=1}^{5} p(\mathcal{M}_{1,d})}{\sum_{d=1}^{5} p(\mathcal{M}_{0,d})}}_{\text{Prior inclusion odds for models assuming effect}}. \tag{12}$$

Similarly to the Bayesian model-averaged estimation, the posterior model probabilities are influenced by the prior predictive accuracy of each parametric family. Nevertheless, since each parametric family is represented in both the nominator and denominator, possible miss-specification of a parametric family results in downweighting its contribution into the model-averaged evidence.

We can also evaluate the evidence supporting employment of one parametric family over the remaining families. To do so, we compare the predictive performance of models from a given parametric family to the rest of the model ensemble. Suppose $d = 1$ denotes the exponential family, then the inclusion Bayes factor in support of the exponential family over all other specified parametric families is

$$\underbrace{\text{BF}_{\text{exp}}}_{\text{Inclusion Bayes factor for exponential family}} = \underbrace{\frac{\sum_{m=0}^{1} p(\mathcal{M}_{m,1}|\text{data})}{\sum_{m=0}^{1}\sum_{d=2}^{5} p(\mathcal{M}_{m,d}|\text{data})}}_{\text{Posterior inclusion odds for exponential family}} \Big/ \underbrace{\frac{\sum_{m=0}^{1} p(\mathcal{M}_{m,1})}{\sum_{m=0}^{1}\sum_{d=2}^{5} p(\mathcal{M}_{m,d})}}_{\text{Prior inclusion odds for exponential family}}. \tag{13}$$

Since we use models assuming the absence and presence of the treatment effect, the resulting inclusion Bayes factor in support of the parametric family accounts for the uncertainty about the presence of the treatment effect.

Lastly, we can evaluate the evidence in favor of or against the inclusion of any single model into the model ensemble. For example, if $M_{0,1}$ denotes the exponential family model assuming the absence of a treatment effect, the inclusion Bayes factor in favor of adding this single model to the ensemble is defined as,

$$\underbrace{\mathrm{BF}_{\mathcal{M}_{0,1}}}_{\text{Inclusion Bayes factor for } \mathcal{M}_{0,1}} = \underbrace{\frac{p(\mathcal{M}_{0,1}|\text{data})}{1-p(\mathcal{M}_{0,1}|\text{data})}}_{\text{Posterior inclusion odds for } \mathcal{M}_{0,1}} \Big/ \underbrace{\frac{p(\mathcal{M}_{0,1})}{1-p(\mathcal{M}_{0,1})}}_{\text{Prior inclusion odds for } \mathcal{M}_{0,1}}. \tag{14}$$

In contrast to Equation 12, the comparison of different parametric families as well as single models is dependent on the prior distributions for the intercept and auxiliary parameters that are not shared across the competing parametric families and thus do not cancel out.

Bayesian evidence

Although Bayes factors are based on sound statistical methodology [82] and despite repeated calls for their usage (e.g., [34, 38, 83, 84]), they are rarely used in medicine. In this section, we explain the appeal of Bayes factors, by highlighting two notable differences from the currently dominant p-value based Neyman-Pearson approach [85] and conceptually similar approaches based on posterior parameter distributions (e.g., [86, 87]): The Bayes factor's interpretation and its behavior under sequential analysis. The Bayes factor's interpretation, and its behavior under sequential analysis. An additional desirable property of Bayes factors in comparison to basing inference on posterior parameter distributions, is the contrasting dependency on prior parameter distributions. Bayes factors provide the most evidence in favor of the alternative hypothesis when specifying the alternative hypothesis as the maximum likelihood estimate (turning the Bayes factor test to a likelihood ratio test [88]) and any other specification results in less evidence in favor of the alternative hypothesis; posterior parameter distribution and inference based on posterior credible intervals can be shifted to provide an inflated rate of false positives [34], an even higher rate than we would see with frequentist p-values. This makes the Bayes factor a conservative measure of evidence, and its benefits are particularly pronounced when informed hypotheses are specified, for example on the basis of historical data.

Also note the difference between the information provided by the Bayesian estimation and Bayesian hypothesis testing (described in the previous section). Posterior parameter distributions obtained by Bayesian estimation inform us about the degree of the effect assuming it is present whereas Bayes factors inform us about the evidence in favor of the presence of the effect. Consequently, the 95% credible interval might contain the number zero while a Bayes factor shows evidence for the presence of the effect (or vice versa). These two pieces of information are however not at odds since each of them answers a different question.

Interpretation of Bayes factors

The strength of evidence measured by Bayes factors corresponds to the relative prior predictive performance of one model compared to another model [4, 76–78]. In other words, obtaining $\mathrm{BF}_{10} = 5$ means that the data are 5 times more likely under the alternative hypothesis than under the null hypothesis [89]. This interpretation is notably different from that of p-values; it does not mean that we would reject a true null hypothesis in 1/5 cases or that we would observe such or more extreme data in 1/5 cases if the null hypothesis were to be true.[4]

To illustrate, consider a simple binomial example where we attempt to treat ten patients. Let us assume that the patients spontaneously recover in 50% of the cases, so we set our null hypothesis of no effect to $\theta_0 = 0.5$. Furthermore, we define two alternative hypotheses, the first specifies $\theta_1 = 0.6$ and the second one specifies $\theta_2 = 0.7$, corresponding to a recovery rate of 60% or 70% after treatment. Let us say that we observe 8/10 patients recover. That leads to two different Bayes factors, $\mathrm{BF}_{10} = 2.75$ and $\mathrm{BF}_{20} = 5.31$ quantifying the evidence in favor of the first and second efficacy hypothesis over the null hypothesis respectively.[5]

[4] Although Bayes factors are a truly continuous measure of evidence, some researchers suggested general rules of thumb to provide intuition about the strength of evidence. E.g., Bayes factors between 1 and 3 (between 1 and $^1/3$) are regarded as anecdotal evidence, Bayes factors between 3 and 10 (between $^1/3$ and $^1/10$) are regarded as moderate evidence, and Bayes factors larger than 10 (smaller than $^1/10$) are regarded as strong evidence in favor of (against) a hypothesis ([75], Appendix I; [112], p. 105).

[5] These settings simplify the outlined methodology such that all prior parameter distributions are reduced to a point, i.e., all probability mass is concentrated to a single point. These simplified parameter priors result in a standard likelihood ratio test, which is a special case of Bayes factors. The Bayes factors can be computed as a ratio of binomial distributions of given data (8 successes out of 10 trials) under the probability parameter θ corresponding to θ specified by the alternative and null hypothesis.

Unsurprisingly, comparing different hypotheses about the treatment effect yields different evidence in favor of presence or absence of the treatment effect. Consequently, setting the same criterion for a decision, e.g., BF = 5, to make a binary choice between the null and each alternative hypothesis would lead to difference in choices and subsequently different rates of misleading evidence if the null hypothesis were to be true. In the example, the evidence would mislead us to incorrectly choose the alternative hypothesis in 5.5% when comparing the null to the first alternative hypothesis and in < 0.1% of the cases when comparing the null to the second alternative hypothesis. We deliberately use the term "misleading evidence" in order to highlight the fact that while the obtained evidence corresponds to the data, in other words, observing 8/10 successes is truly 5.31 more likely under the second alternative hypothesis, the sampling variability of the data itself mislead us into a wrong decision [90].

This is a starring difference to the currently dominant *p*-value based Neyman-Pearson approach [91–93] that builds statistical inference around a binary decision to either accept or reject the null hypothesis with given Type I and Type II error rates. Whereas controlling for the rate of mislead decisions is not the objective of the Bayes factor, the expected proportion of mislead decisions can be evaluated via the means of the Bayes factor design analysis [94, 95]. The Bayes factor design analysis allows researchers to assess how likely a decision at a given Bayes factor leads to false-positive and false-negative evidence. We illustrate how to obtain the frequentist characteristics of Bayesian hypothesis testing with both the fixed-n and sequential design in the 'Bayes Factor Design Analysis' subsection (located in the 'Example' section). Alternatively, decision makers can specify a full decision function based on the costs and benefits of the competing hypotheses and their posterior model probabilities, coherently following from the Bayesian framework. This however goes beyond the scope of the current paper and is discussed elsewhere [96].

Bayes factor sequential analysis

Another crucial difference between Bayes factors and the *p*-value based approach lies in sequential analysis. Bayes factors follow the likelihood principle and are therefore independent of the sampling plan [35–37, 41]. In other words, researchers can decide to collect data until reaching a satisfactory evidence level without affecting the interpretation of Bayes factors. In contrast, the probability of incorrectly rejecting a true null hypothesis approaches unity under continuous monitoring of *p*-values [39, 40]. This crucial difference is due to the fact that *p*-values have a uniform distribution under the null hypothesis and freely "drift" between 0 and 1, whereas Bayes factors approach either 0 or ∞ with increasing sample size, dependent on whether the null or alternative hypothesis is true [97].

Sequential analysis is, of course, still possible with the *p*-value based Neyman-Pearson approach. However, it requires adjustment of the alpha level accordingly to a pre-specified analysis plan that outlines how often, when, and what decisions are going to be made [98]. In contrast to many alpha spending functions for the frequentist sequential analysis that usually result in either rejecting or accepting the null hypothesis at the end of the pre-specified sampling plan, Bayes factor sequential analysis does not necessarily yield decisive evidence in favor of either of the hypotheses. Frequentist properties of the Bayes factor sequential analysis (i.e., the rate of false-positive and false-negative evidence) can, again, be assessed with a Bayes factor design analysis if necessary. Bayes factors sequential stopping rules calibrated for frequentist properties (in contrast to their evidence interpretation) must be adjusted to a more stringent evidence criterion than Bayes factor stopping rules for fixed-n analyses. That is, like *p*-values Bayes factors are influenced by the sampling variability of the data. However, unlike *p*-values the evidence level for Bayes factor sequential analysis is adjusted to account for the sampling variability of the data under *both* hypotheses. This is fundamentally different from the multiple testing adjustment made to *p*-values due to their "free drift" behavior when a null hypothesis is true. As a consequence, unlike *p*-values Bayes factors are consistent under both hypotheses and on average provide increasing support for the true hypothesis as sample size increases [47, 94, 95].

Example

In this section, we apply the outlined modeling framework retrospectively to a real data set and discuss further details, such as the specification of prior parameter distributions and prior model probabilities. All analysis scripts, using the newly developed RoBSA R package [71], are available at https://osf.io/ybw9v/. All data sets can be obtained from the Project Data Sphere [99] following a simple registration at https://www.projectdatasphere.org/.

Data

We use data provided by Project Data Sphere [99] which contains $n = 2{,}968$ patients [100] attending a randomized phase III trial of adjuvant therapy for resected stage III colon cancer [101]. The data set is an extended version of the Alberts and colleagues [70] published study ($n = 2{,}686$) and yields essentially the same results despite

having a slightly larger sample size. The published study was run from 2004 to 2009 and primarily evaluated the effect of adding Cetuximab to the standard sixth version of FOLFOX regimen on disease-free survival. The trial was halted after a planned interim analysis which did not show an improved outcome in patients in the FOLFOX + Cetuximab condition. The authors of the original study adjusted the analyses for several covariates. For the sake of simplicity, we analyse the data without adjusting for the covariates. Omission of the covariates had no significant impact on the main comparison of interest: the disease-free survival in the FOLFOX regime (comparator arm, $n = 1247$, 22.9% events) vs. the FOLFOX + Cetuximab regime (experimental arm, $n = 1251$, 25.1% events) in patients with metastatic wild-type KRAS.

Model specification

To use the outlined Bayesian model-averaging framework, we need to specify three components: (1) the model space, including the parametric families d that specify the plausible data generating mechanisms, their prior model probabilities, and prior model probabilities of models assuming the presence and absence of the treatment effect, (2) the prior distribution for the treatment effect β, separately for both the Bayesian model-averaged estimation and the Bayesian model-averaged testing (because we focus on AFT models here, in the following we will refer to the treatment effect β more specifically as log(AF)), and (3) the prior distributions for the supporting parameters, including the prior distributions for the intercept α_d and auxiliary parameters γ_d.

Model space We define the model space by focusing on five AFT survival parametric families: exponential, Weibull, log-normal, log-logistic, and gamma. Since we do not have a strong a priori preference for any single parametric family, we follow a common convention in BMA and spread the prior model probabilities equally across all parametric families and models assuming the presence and absence of the effect (i.e., we set prior model probability for each model in Bayesian model-averaged estimation to $1/5$ and to $1/10$ in Bayesian model-averaged testing) [19, 77, 81, 102–107].

Treatment effect In Bayesian model-averaged estimation, the prior distribution for the treatment effect $f_{\log(\text{AFT})}$ does not play a large role because it is shared by all models that assume the presence of the effect. Consequently, the prior distribution on the treatment effect does not influence the posterior model probabilities of models assuming the presence of the effect (Equation 8) which in turn determine weighting of the posterior distribution of the model-averaged treatment effect (Equation 9). We therefore aim to specify a weakly informative prior distribution for the treatment effect log(AF)—a distribution across all plausible values, without range constraints, which allows us to incorporate as much information from the data as possible while excluding a priori unrealistic values (e.g., AF > 10) [108]. One possible candidate for such a prior distribution is a standard normal distribution (that places 95% of the prior probability mass within the range of acceleration factors from ~ 0.14 to ~ 7.10).

In contrast to Bayesian model-averaged estimation, the prior distribution on the treatment effect plays a crucial role in Bayesian model-averaged testing. As outlined in Equation 12, the prior distribution for the treatment effect defines the alternative hypothesis, which posits the presence of an effect, and subsequently determines the computed model-averaged inclusion Bayes factor for the effect. Therefore, the prior distributions correspond to the effect sizes that we would expect to observe if the treatment was effective. In our example, we specify $\log(\text{AFT}) \sim \text{Normal}(0.30,\ 0.15)_{[0,\infty]}$ on the log(AF) scale. This normal distribution bounded to positive numbers is centered at the effect size of interest as specified in the preregistration protocol (90% power for a hazard ratio of 1.3),[6] and the small standard deviation quantifies our interest in effect sizes slightly smaller or larger (see Johnson and Cook [34] for more complex alternatives).

Supporting parameters In contrast to simple Bayesian estimation and Bayesian hypothesis testing, the prior distributions on the intercepts α_d and auxiliary parameters γ_d that are specific to each parametric family play an essential role in determining the posterior model probabilities. The posterior model probabilities weight (1) the posterior distributions of the treatment effect across the parametric families in Bayesian model-averaged estimation (Equation 9) and (2) the evidence from the individual parametric families in the inclusion Bayes factor for the treatment effect in Bayesian model-averaged testing (Equation 12). Therefore, a gross miss-specification of supporting parameters for any single parametric family (e.g., survival times distribution on different time scales) will decrease its predictive performance and down-weight the influence of the parametric family on the model-averaged results.

In our example, we used historical information about previous colon cancer trials on disease-free survival and combined them into meta-analytic predictive prior distributions. Meta-analytic predictive prior distributions

[6] Should the treatment be harmful and result in a negative acceleration factor, the result would be much better predicted by the models assuming the absence of a treatment effect. Hence, the trial would be quickly terminated (see upper left panel of Fig. 5).

incorporate information about the between-study heterogeneity of the past studies that down-weights influence of the past data in accordance to their dissimilarities [30, 33].[7] This allows us to not only calibrate the prior distributions for the supporting parameters but also to use already present information about the scaling and shapes of the survival time distributions, making our analysis more efficient. To obtain the historical data, we searched the remainder of the Project Data Sphere database and identified another 24 studies under the 'Colorectal' tumor type. We successfully extracted relevant participant-level data from $k = 3$ data sets corresponding to studies assessing disease-free survival (combined $n = 2{,}860$) [109–111]. While three data sets provide only limited information, especially with regard to the between study heterogeneity, we used Bayesian meta-analysis with weakly informative prior distributions to estimate the meta-analytic predictive prior distributions (see Additional file 1: Appendix A for details).[8]

The resulting prior distributions are summarized in Table 1 show that all meta-analytic prior distributions for the intercepts are fairly similar, with means slightly below 9 and similar standard deviations around 2. The same is true for the auxiliary parameters where the mean-log parameters are close to 0 and the standard deviations around 0.30. The left panel of Fig. 1 visualizes similarities in the scaling and shapes of the different parametric families, with the full lines corresponding to the predicted mean survival function (in the comparator arm) and the shaded areas to 95% prior predictive intervals. The right panel of Fig. 1 visualizes the prior model-averaged survival function for the comparator arm (in light green) vs. the model-averaged survival function for the experimental arm (in deep purple; predicted by the models assuming the presence of the treatment effect specified by the Bayesian model-averaged testing). The predicted model-averaged survival function for the experimental arm is slightly above the predicted model-averaged survival function for the comparator arm since the specified alternative hypothesis describes a positive treatment effect, in other words, longer survival times (note that in the estimation ensemble there is no difference between the model-averaged prior predicted survival functions because we do not constrain the estimated effect to be positive a priori). The shaded 95% prior predictive intervals show a considerable uncertainty based on the historical data, which warrants enough flexibility for the Bayesian updating process.

Bayes factor design analysis

We use a Bayes factor design analysis [47, 94, 95] to evaluate the frequentist properties of the specified Bayesian model-averaged testing model. First, we evaluate the expected rate of the false-positive and false-negative evidence when using symmetrical decision criteria to make a decision about the evidence in favor of presence/absence of the treatment effect either under a fixed-n analysis based on the whole sample or a sequential analysis. Second, we calibrate the decision criteria to match the expected rate of the false-positive and false-negative evidence to the frequentist Type I ($\alpha = 0.05$) and Type II ($\beta = 0.10$) errors. The calibrated decision criteria allow us to analyze the example data in the same way as Alberts and colleagues intended [101].

Settings To evaluate the properties of the specified model, we simulate data from the specified prior distributions (Table 1) under two scenarios. In the first scenario, we simulate data under the assumption that the models assuming the absence of the effect are true. In the second scenario, we simulate data under the assumption that the models assuming the presence of the effect are true. We repeat the simulations 1,000 times and divide them equally amongst the true data generating parametric families (i.e., we simulate data 200 times from the exponential parametric family assuming absence of the treatment effect, 200 times from the exponential family assuming presence of the treatment effect...).[9] For the fixed-n design, we analyze all expected 2070 participants [101] after a 5 years period. For the sequential design, we simplified the trial by assuming that all 2070 participants start at the same time and analyze their data every month until reaching a 5 year period (or if the Bayes factor drifts outside the range of 1/15 to 15 to speed up the computation).

Evaluating misleading evidence The left panel of Fig. 2 visualizes the distribution of inclusion Bayes factors

[7] In the case that we do not have access to historical information or participant level data, we can still utilize more easily obtainable information, such as the expected median and interquartile survival times. The expected information about survival can be used to solve for the means of prior distributions of α_d and γ_d parameters, so they produce survival time distributions with the appropriate summary statistics. Then, we set standard deviations of the prior distributions in a way that produces a suitable amount of flexibility.

[8] Ideally, more informed prior distributions for the usual treatment effects and their heterogeneities would be available for different sub-fields of medicine, such as the ones developed by Bartoš and colleagues [19]. These suggestions would provide better starting points for obtaining meta-analytic predictive prior distributions in case of few primary studies.

[9] We used the Weibull parametric family to simulate the censoring times since (1) the censoring process itself is not modeled by the survival models, (2) estimating meta-analytic predictive prior distribution for a flexible parametric spline model is not a straightforward task, and (3) it was the best fitting distribution to the censoring times according to AIC and BIC across all 3 historical data sets.

Table 1 Overview of the prior distributions for the treatment effect β, intercepts α_d, and auxiliary parameters γ_d across the competing parametric families for both Bayesian model-averaged testing (upper Table) and estimation (lower Table). "Pr. prob." denotes the prior model probabilities, "Post. prob." the posterior model probabilities, "log(marglik)" the log of marginal likelihood, and "Incl. BF" the inclusion Bayes factor for including each model into the model ensemble

Bayesian Model-Averaged Testing

Distribution	Prior β	Prior α	Prior γ	Pr. prob	Post. prob	log(marglik)	Incl. BF
Exponential	S(0)	N(8.70, 2.04)		0.10	0.00	5158.39	0.00
Weibull	S(0)	N(8.80, 2.20)	LogN(−0.07, 0.22)	0.10	0.00	-5151.96	0.00
LogN	S(0)	N(8.70, 1.95)	LogN(0.62, 0.25)	0.10	0.95	-5138.23	182.44
Log-logistic	S(0)	N(8.54, 2.37)	LogN(0.02, 0.27)	0.10	0.03	-5141.65	0.29
Gamma	S(0)	N(8.88, 2.05)	LogN(−0.10, 0.39)	0.10	0.00	-5149.33	0.00
Exponential	$N(0.3, 0.15)_{[0,\infty]}$	N(8.70, 2.04)		0.10	0.00	-5162.05	0.00
Weibull	$N(0.3, 0.15)_{[0,\infty]}$	N(8.80, 2.20)	LogN(−0.07, 0.22)	0.10	0.00	-5155.86	0.00
LogN	$N(0.3, 0.15)_{[0,\infty]}$	N(8.70, 1.95)	LogN(0.62, 0.25)	0.10	0.02	-5142.37	0.14
Log-logistic	$N(0.3, 0.15)_{[0,\infty]}$	N(8.54, 2.37)	LogN(0.02, 0.27)	0.10	0.00	-5145.66	0.01
Gamma	$N(0.3, 0.15)_{[0,\infty]}$	N(8.88, 2.05)	LogN(−0.10, 0.39)	0.10	0.00	-5153.30	0.00

Bayesian Model-Averaged Estimation

Distribution	Prior β	Prior α	Prior γ	Pr. prob	Post. prob	log(marglik)	Incl. BF
Exponential	N(0, 1)	N(8.70, 2.04)		0.20	0.00	-5159.70	0.00
Weibull	N(0, 1)	N(8.80, 2.20)	LogN(0.07, 0.22)	-0.20	0.00	-5153.36	0.00
LogN	N(0, 1)	N(8.70, 1.95)	LogN(0.62, 0.25)	0.20	0.99	-5137.99	363.90
Log-logistic	N(0, 1)	N(8.54, 2.37)	LogN(0.02, 0.27)	0.20	0.01	-5142.50	0.04
Gamma	N(0, 1)	N(8.88, 2.05)	LogN(-0.10, 0.39)	0.20	0.00	-5150.61	0.00

S denotes a Spike prior distribution, N denotes a Normal prior distribution, and LogN denotes a Log-Normal prior distribution

for the presence of the treatment effect in the fixed-n design. The light green density corresponds to the inclusion Bayes factors for the presence of the treatment effect when computed on the data simulated from models assuming the absence of the treatment effect. The deep purple density corresponds to the inclusion Bayes factors for the presence of the treatment effect when computed on the data simulated from models assuming the presence of the treatment effect (32.5% Bayes factors were larger than 1000 and are omitted from the Figure). We can see a visible separation of the densities, 87.2% of Bayes factors under the null hypothesis are lower than 1, correctly favoring the null hypothesis, and 78.1% Bayes factors under the alternative hypothesis are higher than 1, correctly favoring the alternative hypothesis. We can also compute the proportion of misleading evidence if we were to apply a decision at a symmetric boundary $BF_{10} = 10/BF_{10} = 1/10$ corresponding to strong evidence [75, 112]. The strong evidence boundary would lead us to wrongly accept the alternative hypotheses in 0.3% of the cases (assuming the null hypotheses were true) and wrongly accept the null hypotheses in 3.7% of cases (assuming the alternative hypotheses were true). This is a much lower percentage of errors than we would obtain with the commonly recommended frequentist settings of $\alpha = 0.05$ and $\beta = 0.10$. Note that the error rates for a any given evidence level depend on the sample size, which is why Bayes factor design analyses are needed.

The right panel of Fig. 2 visualizes the trajectories of inclusion Bayes factors for the presence of the treatment effect in the sequential design. The light green density corresponds to 95% of the most supportive inclusion Bayes factors trajectories for the presence of the treatment effect when computed on the data simulated from models assuming the absence of the treatment effect. The deep purple density corresponds to 90% of the most supportive inclusion Bayes factors trajectories for the presence of the treatment effect when computed on the data simulated from models assuming the presence of the treatment effect. The right panel of Fig. 2 also shows ten example trajectories of the Bayes factors. As discussed in the 'Bayesian Evidence' section, Bayes factors tend to converge towards the evidence in favor of the true hypothesis. Nevertheless, the sampling variance of the data can introduce oscillations in the trajectories, which is the reason behind a higher rate of positive and negative evidence in the Bayes factor sequential analysis [47, 94, 95]. In our case, using the same decision criteria corresponding to strong evidence (a symmetric boundary $BF_{10} = 10/BF_{10} = 1/10$),

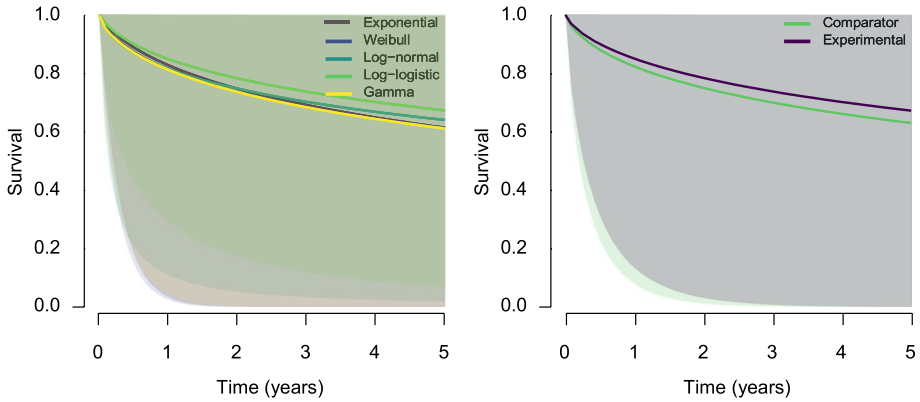

Fig. 1 Left: The prior predicted survival function for the comparator arm in each parametric family. The shaded areas correspond to 95% prior predictive intervals. Right: Model-averaged prior predicted survival function in the comparator arm (light green) and experimental arm (deep purple) assuming the presence of the treatment effect specified by the Bayesian model-averaged hypothesis testing approach. The shaded areas correspond to 95% prior predictive intervals

we would wrongly accept the null hypothesis in 3.3% of the cases and wrongly accept the alternative hypothesis in 3.1% of cases.[10] Again, a lower percentage of errors than in the common frequentist settings and with the possibility to continuously monitor the evidence without adjusting the boundaries.

Calibration for frequentist properties To make the results of our example directly comparable to the frequentist analysis, we calibrate the decision boundaries on Bayes factors to lead to a rate of false-positive and false-negative evidence corresponding to the expected Type I and Type II error rate ($\alpha = 0.05$, $\beta = 0.10$). We calibrate the Bayes factors by computing the 95%$^{\text{th}}$ and 10%$^{\text{th}}$ quantile of the Bayes factors under the null and alternative hypotheses respectively for the fixed-n design, and by finding upper and lower bounds that are not crossed by more then 5% and 10% of the trajectories for the sequential design.

The left panel of Fig. 2 visualizes the calibrated decision criteria for the fixed-n design as two dashed vertical lines corresponding to $\text{BF}_{01} = 2.72$ and $\text{BF}_{10} = 1.72$. These boundaries, calibrated to the common frequentist error rates, correspond to much weaker evidence. Similarly, the right panel of Fig. 2 visualizes the calibrated decision criteria for the sequential design as two dashed horizontal lines corresponding to $\text{BF}_{01} = 4.4$ and $\text{BF}_{10} = 6.9$. These boundaries are considerably wider than the boundaries for the fixed-n design due to the sampling variance of the data that would lead to misleading decisions when crossing a tighter boundary. Nevertheless, the calibrated boundaries are still noticeably tighter than boundaries corresponding to strong Bayesian evidence. It is worth considering whether such weak evidence, in both the fixed-n and sequential designs warrants permission to draw strong conclusions.

Implementation

While analytical solutions for certain combinations of prior distributions and parametric families exist [50], we use MCMC sampling to estimate the posterior distributions (Equation 4; implemented in the runjags R package [113] accessing the JAGS statistical programming language on the background [5]). We use bridge-sampling [8, 9] to estimate each model's marginal likelihoods (Equation 3; implemented in the bridgesampling R package [7]). We combined all of the required functionality to fit, interpret, and plot Bayesian model-averaged survival analyses into the RoBSA R package [71].

Results

The upper part of Table 1 summarizes the results of the Bayesian model-averaged testing ensemble. It contains five models assuming the absence of a treatment effect and five models assuming the presence of a positive treatment effect. We find strong evidence against the models assuming the presence of the positive treatment effect $\text{BF}_{01} = 62.5$, which crosses the Bayesian strong evidence threshold as well as the threshold calibrated for frequentist properties. The obtained evidence decreases the prior probability of the positive treatment

[10] The fact that we observe a slight decrease of false-positive evidence is due to the sampling variance of the data which can lead to an earlier correct acceptance of the null hypothesis, eliminating the chance of crossing the wrong boundary, and due to the variance in the Bayes factor design analysis itself.

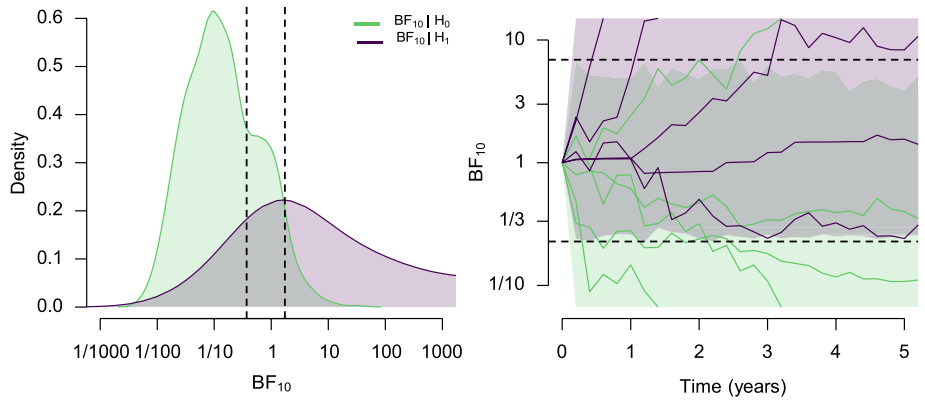

Fig. 2 Left: Distribution of the inclusion Bayes factors for the presence of the treatment effect in the fixed-n design under the null hypothesis (assuming the absence of the treatment effect; light green) and the alternative hypothesis (assuming the presence of the treatment effect; deep purple). 32.5% Bayes factors under the alternative hypothesis are larger than 1000 and not shown. The vertical dashed lines visualize boundaries for obtaining 10% false-negative evidence and 5% false-positive evidence. Right: Trajectories of the inclusion Bayes factors for the presence of the treatment effect in the sequential design under the null hypothesis (assuming the absence of the treatment effect; light green) and the alternative hypothesis (assuming the presence of the treatment effect; deep purple). Ten example trajectories are visualized in the full colored lines. The bounds are truncated in the range of 1/15 and 15. The horizontal dashed lines visualize boundaries for obtaining 10% false-negative evidence and 5% false-positive evidence

effect from 0.50 to 0.02. We inspect the inclusion Bayes factors for the competing parametric families and find strong evidence supporting the models based on the log-normal parametric family, $BF_{\text{log-normal}} = 122.0$ (averaged across models assuming both presence and absence of the effect, i.e., Equation 13). We find more fine-grained results in the upper part of Table 1, which shows that the data were most consistent with the log-normal model assuming the absence of the positive treatment effect, increasing its posterior model probability from 0.10 to 0.95.

The Bayesian model-averaged estimation ensemble contains only models assuming the presence of the treatment effect, with a wider, unbounded, prior distribution over the treatment effect. The fact that we found strong evidence for models assuming the absence of the treatment effect over the models assuming the presence of the positive treatment effect does not mean that the effect is necessarily zero— negative values of the treatment effect would also provide evidence against models assuming the presence of the positive treatment effect. Indeed, we find a mostly negative model-averaged estimate of the treatment effect, log(AF) = -0.188, 95% CI [-0.346, -0.034]. The left panel of Fig. 3 visualizes the prior (dashed grey line) and posterior (full black line) distribution for the treatment effect. The peaked posterior distribution signifies the amount of information learned from the observed data. Another example of the learning process is visualized in the right panel of Fig. 3, where the posterior credible intervals of the survival functions is noticeably tightened. As shown in the lower part of Table 1 we find that most of the posterior model probability, 0.99, is, again, ascribed to the log-normal parametric family.

These results lead to qualitatively similar conclusions as the results obtained by the Cox proportional hazard model presented by Alberts and colleagues (HR = 1.21, 95% CI [0.98, 1.49], $p = .08$; [70]).

Sequential analysis

We can take advantage of the ability to update the evidence in a truly sequential manner and inspect how it accumulates throughout the trial. Since the data provided at Project Data Sphere do not contain the time of enrollment into the study, we simplify our settings by assuming that all participants start the study at the same time. We re-estimate the model ensemble for the Bayesian model-averaged testing (upper part of Table 1) every month (30 days) and evaluate the evidence for the presence vs. absence of the specified treatment effect (with observations with survival/censoring times beyond the current time scope being censored at the current evaluation time).

The left panel of Fig. 4 visualizes the flow of evidence with time towards the models assuming the absence of the effect. We find that the evidence against the alternative hypothesis of positive treatment effect accumulates rather quickly. The Bayes factor for the presence of the treatment effect falls below the Bayesian strong evidence threshold at 6 months from the start of the trial. Furthermore, the Bayes factor crosses the calibrated sequential threshold for frequentist properties (BF_{01}=4.4) at 3 months from the start of the trial.

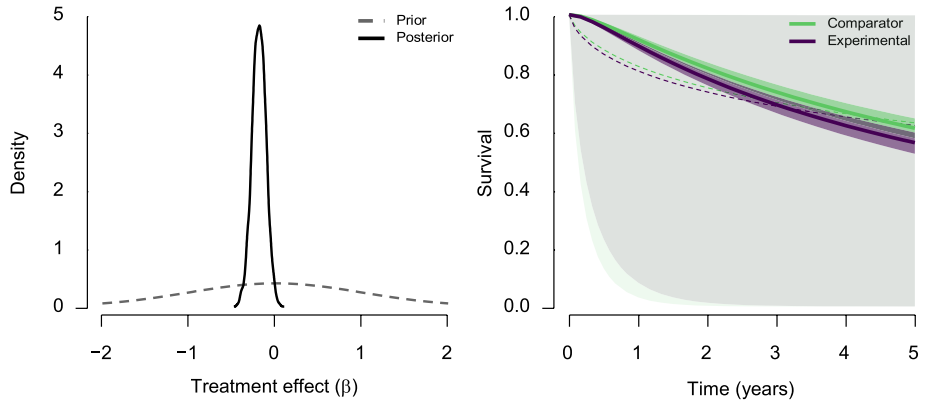

Fig. 3 Left: Prior (grey dashed line) and posterior (full black line) distribution of the treatment effect obtained with the Bayesian model-averaged estimation. Right: Model-averaged prior (dashed lines) and posterior (full lines) survival function in the comparator arm (light green) and experimental arm (deep purple) obtained with the Bayesian model-averaged estimation. The shaded areas correspond to a 95% prior (less saturated) and posterior (more saturated) credible intervals. The mean of model-averaged prior distribution in the experimental arm for Bayesian model-averaged estimation is slightly bellow the mean in the control arm due to the non-linear transformations involved in computing survival, despite the prior distribution being centered at zero treatment effect

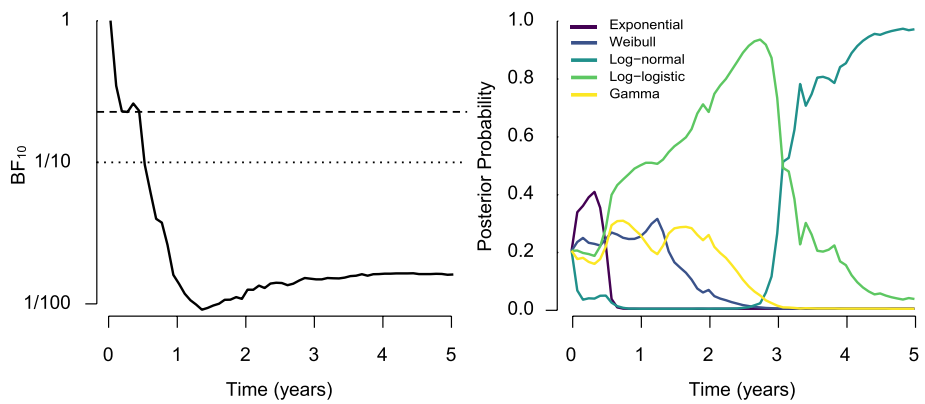

Fig. 4 Left: Trajectory of the Bayes factors for the presence of the treatment effect in the sequential analysis. The horizontal dashed line visualizes the lower decision boundary calibrated for 10% false-negative evidence ($BF_{01} = 4.4$ crossed at 3 months since the start of the trial) and the horizontal dotted line visualizes the strong evidence in favor of the null hypothesis boundary ($BF_{01} = 10$ crossed at 6 months since the start of the trial). Right: Trajectories of the posterior model probabilities of the individual parametric families

The right panel of Fig. 4 visualizes the flow of evidence with time towards the competing parametric families as a posterior probability of a given parametric family. We can see the advantages of model-averaging, especially early in the data collection when there is a lot of uncertainty about the most likely parametric family.

We can compare the results to the analysis plan of the original study that specified interim analysis after reaching 25%, 50%, and 75% of the planned number of 515 events using an O'Brien-Fleming stopping boundary [43], truncated at ±3.5, resulting in boundaries at ±3.5, ±2.996, ±2.361, and ±2.015 for each step [101]. Using our simplified version of the trial, the registered analysis plan would lead to early stopping at the second interim analysis (at 50% of the expected observed events) after 13.3 months. That is 10.3 months later in comparison to stopping at the calibrated sequential threshold or 7.3 months later than stopping upon reaching strong evidence. Using the full data set and Bayesian model-averaged estimation, we find that the mean progression-free survival is 19.1 years in the experimental arm vs. 23.1 years in the comparator arm. Ending the trial 10.3 months earlier and switching the patients from the experimental to the comparator arm

would increase their mean progress-free survival time by 12.5 months.[11] With 1251 patients in the experimental arm, the difference makes 1,299 progress-free survival years in total.

Simulation Study

We designed a simulation study closely based on the example data set to evaluate the described methodology in real-life like settings while controlling for potential confounds and unknown factors specific to the example data set. The simulation code is available at https://osf.io/ybw9v/.

We evaluated estimation and testing performance of Bayesian model-averaging and compared it to model selection over the parametric families with either Bayes factor or AIC/BIC implemented in the flexsurv R package.[12] For the Bayesian approaches we used the historical data to specify prior distributions (as in the example, c.f., Table 1).[13] We evaluated the performance of the methods in a fixed-n design and a sequential design. To assess performance under realistic conditions, i.e., when the true data generating process is unknown and may not match any parametric family, we omitted the family used to simulate the data from the set of models used in the model selection/model-averaging analyses. For example, if the data were simulated from the exponential parametric family, the results for all methods were computed without considering the exponential parametric family models (see Supplementary Materials for more details and similar results when including all parametric families).

We based the data generating process for the simulation study on the example data set from Alberts and colleagues [111]. We considered five parametric families (exponential, Weibull, log-normal, log-logistic, and gamma) for the fixed-n design and one parametric family (Weibull) for the sequential design as the true data generating mechanisms for the survival times. We used a parametric spline model [114] as the true data generating mechanism for the censoring times. This allowed us to compare the performance of the methods across different, controlled, data generating processes while leaving the censoring process flexible. We censored all survival times larger than 5 years. In the sequential design, we started with all participants being censored and revealing their true or censoring times as the time of the trial progressed (as in the example). We estimated the parameters for simulating survival and censoring times by fitting the corresponding maximum-likelihood parametric models to the Alberts and colleagues' data set [111]. Furthermore, we manipulated the true acceleration factor (log(AF) = -0.2, 0, 0.2, 0.4) and considered different sample sizes (N = 50, 200, 1000 for the fixed-n design, and N = 2070 for the sequential design). That resulted in 5 (data generating mechanisms) × 4 (AF) 3 (samples sizes) = 60 simulation conditions in the fixed-n design and 1 (data generating mechanisms) × 4 (AF) × 1 (samples sizes) = 4 simulation conditions in the sequential design. We repeated each simulation condition 500 times in both designs. The number of repetitions was limited by the computational resources required for estimating the Bayesian model-averaged methods.

Results: fixed-n design

We evaluated the performance of the methods according to the bias (the difference between the true value and the estimate), root mean square error (RMSE, the square root of the mean square difference between the true value and the estimate), and confidence interval coverage of the log(AF) estimate. Ideally, we would like to observe as low RMSE as possibly, indicating high precision of the estimates, no or with sample size decreasing bias, indicating that our estimates are converging to the true values, and a nominal confidence interval coverage, indicating properly calibrated confidence intervals. We evaluated the error rate and power when making decisions about the presence of the treatment effect (with $\alpha = 0.05$, one-sided, for the frequentist methods, and Bayes factors thresholds calibrated for the corresponding frequentist properties with the historical data, as in the 'Example' section).[14] Ideally, we would like to observe an error rate around the nominal α level, indicating proper calibration of p-values and Bayes factors, and as high power as possible, indicating high efficiency of test. We also evaluated bias and RMSE for the mean predicted survival at 20 years period. We used formulas provided by Morris and colleagues [115] to compute MCMC errors (SE) of the bias, confidence interval coverage, error rate, and power. We used the jackknife estimate of variance [116] to compute the standard error of the RMSE.

RMSE, bias, and confidence interval coverage of the estimated mean log(AF) as well as the power and error rate were comparable across the different data generating mechanisms. Therefore, we present aggregated results across the different data generating mechanisms (tables with detailed results for each parametric family are available in Supplementary Materials).

[11] Under the assumption that spending 3 vs. 13.3 months in the experimental comparator would "take" 1.3 vs. 5.8% of their mean progress-free survival time and the rest of their progress-free survival time would be based on the mean progress-free survival in the comparator arm.

[12] BIC model selection corresponds to Bayes factor model selection when a unit information prior is used (e.g., [134]).

[13] Each Bayesian model was estimated with two chains, each run for 1000 burnin and 5000 sampling iterations.

[14] That led to $BF_{10} = 1.9, 2.3, 2.2$ and 1.9, 2.4, 2.2, and $BF_{01} = 1.6, 2.2, 2.6$ and 1.6, 2.2, 2.6 for $N = 50, 200, 1000$ for Bayesian model-averaging and Bayes factor model selection respectively.

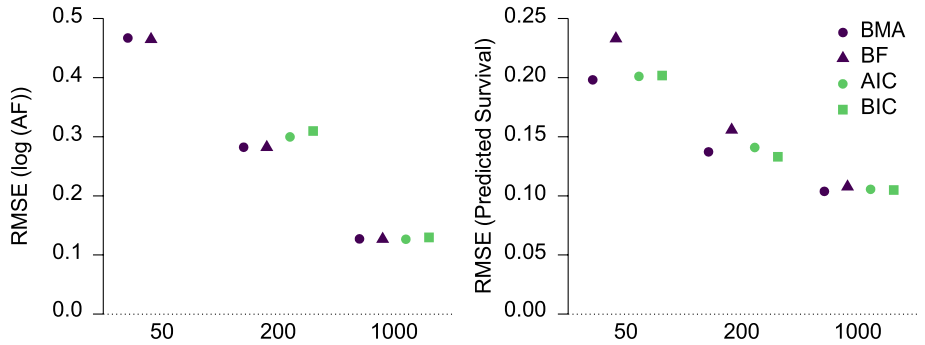

Fig. 5 Left: Root mean square error (RMSE, *y*-axis; 95% confidence intervals are not shown as they are shorter than the symbols) of the mean log acceleration factor estimates for different sample sizes (*x*-axis) and methods (colors/shapes) averaged across all simulation conditions. RMSE of AIC and BIC model selection with n = 50 estimates is out of the plotting range (AIC = 1.78, BIC = 1.91). Right: Root mean square error and 95% confidence intervals (RMSE, *y*-axis) of the predicted mean survival at 20 years estimates for different sample sizes (*x*-axis) and methods (colors/shapes) averaged across all simulation conditions. Methods: Bayesian model-averaging (BMA, deep purple circles) and model selection over parametric families with: Bayes factors (BF = light green triangles), AIC (light green circles), and BIC (deep purple squares)

The left panel of Fig. 5 visualizes the RMSE of the mean log(AF) estimates aggregated across true log(AF), all of which led to comparable RMSEs. We see that both Bayesian model-averaging and Bayesian model selection with Bayes factors outperformed the frequentist model selection with AIC/BIC in small to medium sample sizes. This benefit is a result of the regularizing properties of the prior distributions that reduce the otherwise large variability of the log(AF) estimates under small sample sizes. Results of bias showed a similar pattern as the RMSE (see Additional file 1: Appendix B), however, whereas the frequentist methods lead to overestimation of the log(AF) in small sample sizes (i.e., more extreme estimates of the log(AF) estimates regardless of the direction) the Bayesian methods lead to underestimation of the log(AF) (i.e., more conservative estimates regardless of the direction due to shrinkage introduced by the prior distributions). Regardless of the differences in RMSE and bias, the confidence interval coverage did not seem to differ by methods— all achieving proper confidence interval coverage (see Additional file 1 Appendix B).

The right panel of Fig. 5 visualizes the RMSE of the predicted survival at twenty years. We see that Bayesian model-averaging and AIC/BIC model selection outperform Bayesian model selection with Bayes factors in all but the largest sample sizes, where all models converge to similar results. Results of bias favored the AIC/BIC model selection over the Bayesian methods. Results of bias of the predicted survival at twenty years showed a similar pattern as the RMSE (see Additional file 1: Appendix B).

Figure 6 visualizes the error rate (first row) and power (second row) for the test of negative or null and positive log(AF) respectively. We see that all methods showed similar error rate in case of the negative log(AF) (upper left), however, the Bayesian model-averaging and model selection with Bayes factors exhibited an inflated error rate for $n = 200$ participants in the case of no treatment effect (upper right). The elevated error rate was balanced by increased power in settings with presence of the positive treatment effect (log(AF) = 0.2 in bottom left and log(AF) = 0.4 in bottom right).[15]

Results: sequential design

We evaluated the performance of the methods according to the error rate and power when making decisions about the presence of the treatment effect and time to make the decision. For the Bayesian model-averaging and Bayes factor model selection we used the Bayes factor thresholds calibrated for the corresponding frequentist properties with the historical data, as in the 'Example' section.[16] Ideally, we would like to observe an error rate around the nominal α level, indicating proper calibration of p-values and Bayes factors, as high power as possible, indicating high efficiency of test, and as short times to make decisions as possible, indicating high efficiency of the sequential testing procedure. Similarly to the example, we re-estimated the models to monitor the evidence every month. For the frequentist methods, we used varying numbers of steps ($k = 2, 4, 5, 10, 20$), to assess the different degrees of sequential efficiency, for sequential analysis with binding asymmetric boundaries,

[15] Power of the different methods in conditions with log(AF) = 0.2; BMA = .09, .22, .48 and BF = .09, .21, .49, vs. AIC = .08, .17, .48 and BIC = .07, .16, .48, and log(AF) = 0.4, BMA = .15, .46, .92 and BF = .15, .44, .92 vs. AIC = .12, .38, .91 and BIC = .12, .37, .91, for 50, 200, and 1000 observations respectively.

[16] That led to $BF_{10} = 4.4$ and 4.7, and $BF_{01} = 6.9$ and 7.1 for Bayesian model-averaging and Bayes factor model selection, respectively.

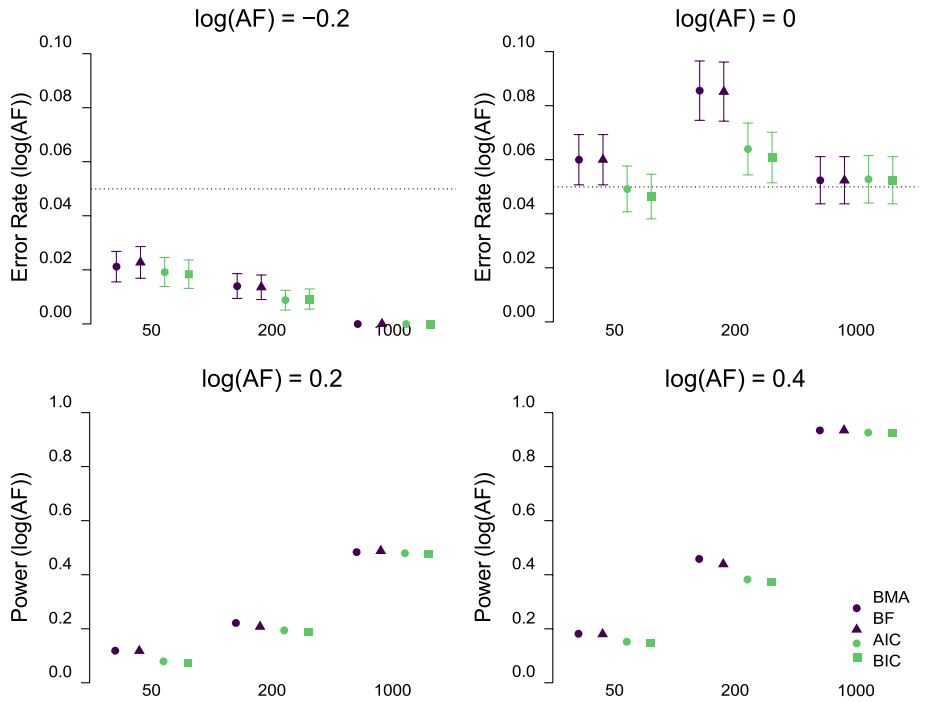

Fig. 6 First row: Error rate and 95% confidence intervals (*y*-axis; 95% confidence intervals are not shown in cases where they are shorter than the symbols) for the test of the positive acceleration factor for different sample sizes (*x*-axis), methods (colors/shapes), and true acceleration factors (columns) averaged across all simulation conditions. Second row: Power and 95% confidence intervals of the test of the positive acceleration factor (*y*-axis) for different sample sizes (*x*-axis), methods (colors/shapes), and true acceleration factors (columns) averaged across conditions with different parametric families. Methods: Bayesian model-averaging (BMA, deep purple circles) and model selection over parametric families with: Bayes factors (BF = deep purple triangles), AIC (light green circles), and BIC (light green squares). Note the different scaling of the *y*-axis for the error rate (first row) and power (second row)

Hwang-Shih-DeCani spending function [98, 117], and $\alpha = 0.05$ for one-sided test. The Hwang-Shih-DeCani spending function allows stopping both for efficacy and futility while leading to optimal sample size [118]. We used formulas provided by Morris and colleagues [115] to compute MCMC errors (SE) of the error rate, and power, and conventional standard errors for the required time to reach the decision.

Figure 7 visualizes the distribution of times to reach either the correct decision (deep purple), the incorrect decision (light green), or not reaching a decision at all (grey) in the sequential analysis with different true log(AF) depicted in different panels (see Table A1 in Appendix B for numerical summaries of the times to and probabilities of making a decision). Different rows compare different methods (with AIC/BIC corresponding to a sequential analysis with the most optimal $k = 20$ steps; tables with detailed results for each number of steps are available in Table 15 of Supplementary Materials). We see that both Bayesian model-averaging and Bayesian model selection with Bayes factors outperformed the frequentist model selection with AIC/BIC in terms of time to reach either the correct or incorrect decisions regardless of the true log(AF). Table A1 in Appendix B shows that the time to reach either the correct or incorrect decisions was almost half for the Bayesian methods in comparison to the frequentist alternatives. The error rate was either lower or about equal to the set significance threshold for all the methods in conditions with negative or no log(AF), and the power was essentially the same for all methods in conditions with positive log(AF). However, while the frequentist methods had a higher proportion of undecided trials in the log(AF) = 0.2 condition (AIC = 0.212, BIC = 0.208) in contrast to the Bayesian methods (BMA = 0.136, BF = 0.152), conversely, the Bayesian methods had a higher proportion of incorrect decisions (BMA = 0.184, BF = 0.164) in contrast to the frequentist methods (AIC= 0.112, BIC = 0.116).

Discussion

We described benefits of the Bayesian framework consisting of estimation, hypothesis testing, and model-averaging when applied to survival analysis. Specifically, we highlighted how to: (1) include historical data into the analysis, (2) specify and test informed hypotheses,

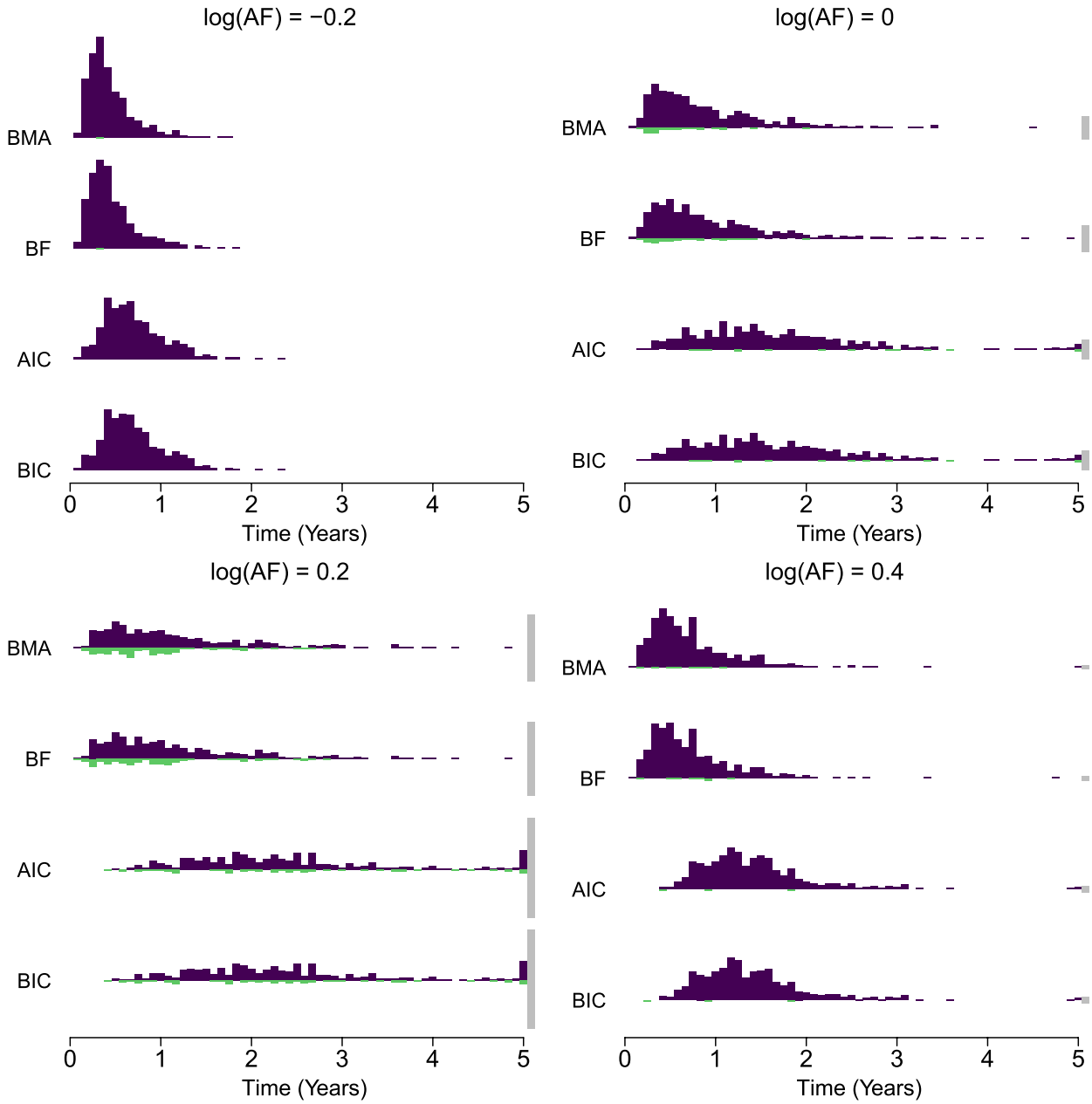

Fig. 7 Time until reaching a conclusion in sequential analysis for different true acceleration factors (panels). The upper part of each histogram shows distribution of times until reaching the correct conclusion (deep purple), the lower part of the histogram shows distribution of times until reaching the wrong conclusion (light green), and the double sided bar at the end shows the proportion of undecided sequential analyses at the end of the trial (grey). Methods: Bayesian model-averaging (monitored every month, first row) and model selection over parametric families with: Bayes factors (monitored every month, second row), AIC (with 20 interim analyses, third row), and BIC (with 20 interim analyses, fourth row)

and (3) incorporate uncertainty about the true data generating process into the analysis. Furthermore, we discussed the differences between the frequentist and Bayesian frameworks towards evidence, showed how to calibrate the Bayesian analyses for frequentist properties (if needed) with Bayes factor design analysis, and demonstrated efficiency of the Bayesian framework in an example and a simulation study. In this simulation study we found the Bayesian approach had (1) shorter times required for sequential designs, (2) slightly higher statistical power and false-positive rate in fixed-n designs, and

(3) more precise estimates of the treatment effect in small and medium sample sizes.

Including historical data into current studies can greatly improve efficiency of the analyses, especially when including more participants is costly. As other researchers repeatedly stressed: there is plenty of historical data, and not utilizing them is a waste of resources [21–25]—resources that could be used to provide better treatment to the current patients and develop new treatments [43–45]. Specifying hypotheses in accordance with the expectations of the treatment efficiency, i.e., performing an informed Bayesian hypotheses test, further builds on this idea. Informed Bayesian hypothesis testing allows researchers to evaluate the evidence in favor or against specific claims—making the most of the data and allowing for richer inference [34, 95, 119]. Finally, incorporating uncertainty about the true data generating mechanism dispenses with the need to commit to assuming a single parametric family, making the analyses more robust to model misspecifications.

Bayesian analysis requires the full specification of priors for all parameters. While researchers may have good intuitions for priors on the treatment effect, auxiliary parameters may be more challenging to reason about—particularly, when historical data on auxiliary parameters are lacking. In this case, the appropriate prior distributions can be constructed by eliciting expert knowledge [120–123]. Nevertheless, projections about the approximate mean, median, or interquartile range of survival are often important part of planning and registering clinical trials and can be used to calibrate prior distributions for the auxiliary parameters.

Importantly, Bayesian hypothesis tests are defined by the prior distribution on the parameter of interest, usually the treatment effect. Specifying a different prior distribution on the treatment effect parameter corresponds to defining a different hypothesis about the treatment. Different questions necessarily lead to different answers, and similar questions lead to similar answers. This concept is not that dissimilar from frequentist hypothesis testing either. E.g., a two-sided frequentist test might give a different answer than a one-sided test which might, in turn, give a different answer than a frequentist test for a minimal effect size of clinical relevance.

The Bayes factor design analysis also highlighted one important fact: the routinely used Type I error rate of 5% corresponds to weak evidence in favor/against the informed alternative hypothesis. This is especially true with increasing sample size, potentially resulting in the Jeffreys-Lindley paradox (see [124] for an excellent overview). Similar findings have already been described elsewhere, with authors arguing for adopting a more stringent significance level [125] or increasing the significance level with increasing sample size (e.g., [126–129]). Alternatively, researchers may specify a utility function and perform a full decision analysis based on the posterior parameter distributions and posterior model probabilities [96].

Nonetheless, all of these advantages come at a cost. Setting up and executing the outlined analyses is, without a doubt, more demanding than the standard frequentist approach. It requires more computational resources and more of the researchers' time to execute the analysis. However, there are also significant tangible costs of keeping the status quo. In our example, we showed that sequential analysis with Bayesian hypothesis testing can decrease the trial duration by over a year—that is a whole year in which half of the patients could be provided with a treatment with less severe side effects, leading to longer progress-free survival [70].

We verified this result in a simulation study, showing that, more generally, the Bayesian sequential analysis leads to faster decisions in clinical trials. While the analyses showed a higher proportion of false-negatives under model misspecification (i.e., lower treatment efficiency than expected), it kept the same power and proper false-positive rate as the frequentist sequential analyses. The fixed-n design revealed a considerable decrease in bias and root mean square error in small to medium sample sizes with slightly elevated error rates and power. Furthermore, we observed a decrease in power to accept the null hypothesis in case of a negative treatment effect. Surprisingly, we observed little or no benefits of Bayesian model-averaging over Bayesian model selection with Bayes factors in our simulation. However, this finding might be limited to the specific settings derived from a single data set, which may not be representative for other diseases or treatments.

There are multiple avenues for further development of the outlined methodology. For example, different assumptions about the data generating mechanism might be incorporated by performing model-averaging over proportional hazard models or smoothing splines [13, 130, 131]. Frailties, left and interval censoring can be also incorporated into the models and combined with longitudinal models. Furthermore, the performance of Bayesian testing when evaluating more complex hypotheses, e.g., non inferiority can be assessed.

In the end, the Bayesian framework is simply a coherent application of the laws of probability [2–4, 74] and the likelihood principle [35, 41] which allows researchers to draw a richer and more specific set of inferences. These ideas were steadily developed over centuries, but only the

recent boom in computing power and improvements in computational tools enabled their application to complex problems. We believe that now is the time for researchers to utilize technological advancements, further develop easy-to-use statistical software, and fully take advantage of the offerings of Bayesian statistics.

Conclusion

In this paper, we outlined the theoretical framework and showed the application of the informed Bayesian estimation, testing, and model-averaged approaches to parametric survival analyses. We evaluated the methodology against the currently used techniques and found that continuously monitoring the evidence, employing more specific hypothesis, incorporating historical data, and basing the inference on multiple models leads to: (1) shorter times required for sequential designs, (2) slightly higher statistical power and false-positive rate in fixed-n designs, and (3) more precise estimates of the treatment effect in small and medium sample sizes. We did not find a clear advantage in predicting survival nor an advantage of Bayesian model selection with Bayes factors against Bayesian model-averaging in our simulation.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01676-9.

Additional file 1.

Additional file 2.

Acknowledgements
This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere. Neither Project Data Sphere nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this publication. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures. We thank Eric-Jan Wagenmakers for many helpful comments and support.

Authors' contributions
F A designed the simulation study and implemented the methodology. J.M.H. and F.A. supervised the research. F.B. analysed the data. All authors interpreted the results, read, and approved the final manuscript.

Funding
F.B. and J.M.H. were supported by a Vici grant from the NWO to Eric-Jan Wagenmakers (016.Vici.170.083).
F A was supported by an Advanced ERC grant to Eric-Jan Wagenmakers (743086 UNIFY). J.M.H. was supported by a Veni grant from the NWO (VI.Veni.201G.019).

Availability of data and materials
The method is implemented in RoBSA R package accessible at https://CRAN.R-project.org/package=RoBSA. The analysis and simulation scripts are available at https://osf.io/ybw9v/. The example data sets are available from www.projectdatasphere.org upon a simple registration. The supplementary materials are available at https://osf.io/8skf7/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Not applicable.

Author details
[1]Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands. [2]Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic.

Received: 16 December 2021 Accepted: 30 May 2022
Published online: 10 September 2022

References

1. van de Schoot R, Depaoli S, King R, Kramer B, Märtens K, Tadesse MG, et al. Bayesian statistics and modelling. Nat Rev Methods Primers. 2021;1(1):1–26. https://doi.org/10.1038/s43586-020-00001-2.
2. Bayes T. An essay toward solving a problem in the doctrine of chances. By the late rev. mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S., 1763. Philos Trans R Soc Lond. 1997;53:370–418.
3. Jeffreys H. Some Tests of Significance, Treated by the Theory of Probability. Proceedings of the Cambridge Philosophy Society. 1935;31:203–222. Available from: https://doi.org/10.1017/S030500410001330X.
4. Wrinch D, Jeffreys H. On certain fundamental principles of scientific inquiry. Philosophical Magazine. 1921;42:369–90. https://doi.org/10.1080/14786442108633773.
5. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing. vol. 124. Vienna, Austria; 2003. p. 1–10.
6. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. JStat Softwe. 2017;76(1):1–32. https://doi.org/10.18637/jss.v076.i01.
7. Gronau QF, Singmann H, Wagenmakers EJ. bridgesampling: An R package for estimating normalizing constants. J Stat Softw. 2020;92(10):1–29. https://doi.org/10.18637/jss.v092.i10.
8. Gronau QF, Sarafoglou A, Matzke D, Ly A, Boehm U, Marsman M, et al. A tutorial on bridge sampling. J Math Psychol. 2017;81:80–97. https://doi.org/10.1016/j.jmp.2017.09.005.
9. Meng XL, Wong WH. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. Statistica Sinica. 1996;6:831–60. Available from: https://www.jstor.org/stable/24306045.
10. Mills M. Introducing survival and event history analysis. London: Sage; 2010.
11. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB Jr, Alper S, et al. Many Labs 2: Investigating variation in replicability across samples and settings. Adv Methods Pract Psychol Sci. 2018;1(4):443–90. https://doi.org/10.1177/515245918810225.
12. Cox DR. Regression models and life-tables. J Royal Stat Soc Series B (Methodological). 1972;34(2):187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.
13. Bogaerts K, Komárek A, Lesaffre E. Survival analysis with interval-censored data: A practical approach with examples in R, SAS, and BUGS. Boca Raton: Chapman and Hall/CRC; 2017.
14. Latimer NR. Survival analysis for economic evaluations alongside clinical trials–extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Med Decis Making: Int J Soc Medl Decis Making. 2013;33(6):743–54. https://doi.org/10.1177/0272989x12472398.
15. Breslow NE. Contribution to discussion of paper by DR Cox. J Royal Statist Soc Ser B. 1972;34:216–7.
16. Gronau QF, Ly A, Wagenmakers EJ. Informed Bayesian *t*-tests. Am Stat. 2020;74:137–43. https://doi.org/10.1080/00031305.2018.1562983.

17. Rhodes KM, Turner RM, Higgins JP. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. J Clin Epidemiol. 2015;68(1):52–60. https://doi.org/10.1016/%2Fj.jclinepi.2014.08.012.
18. Stefan AM, Evans NJ, Wagenmakers EJ. Practical challenges and methodological flexibility in prior elicitation. Psychol Methods. 2020; Available from: https://doi.org/10.1037/met0000354.
19. Bartoš F, Gronau QF, Timmers B, Otte WM, Ly A, Wagenmakers EJ. Bayesian model-averaged meta-analysis in medicine. Stat Med. 2021;40(30):6743–61. https://doi.org/10.1002/sim.9170.
20. Parmar MKB, Sydes MR, Morris TP. How do you design randomised trials for smaller populations? A framework. BMC Med. 2016;14(1):183. https://doi.org/10.1186/s12916-016-0722-3.
21. Pocock SJ. The combination of randomized and historical controls in clinical trials. J Chronic Dis. 1976;29(3):175–88. https://doi.org/10.1016/0021-9681(76)90044-8.
22. Berry DA. Bayesian clinical trials. Nature Reviews Drug Discovery. 2006;5(1):27–36. https://doi.org/10.1038/nrd1927.
23. Hobbs BP, Carlin BP. Practical Bayesian design and analysis for drug and device clinical trials. J Biopharm Stat. 2008;18(1):54–80. https://doi.org/10.1080/10543400701668266.
24. Cope S, Ayers D, Zhang J, Batt K, Jansen JP. Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of CAR-T therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia. BMC Med Res Methodol. 2019;19(1):1–11. https://doi.org/10.1186/s12874-019-0823-8.
25. Thirard R, Ascione R, Blazeby JM, Rogers CA. Integrating expert opinions with clinical trial data to analyse low-powered subgroup analyses: a Bayesian analysis of the VeRDiCT trial. BMC Medical Res Methodol. 2020;20(1):1–10. https://doi.org/10.1186/s12874-020-01178-6.
26. Brard C, Hampson LV, Gaspar N, Le Deley MC, Le Teuff G. Incorporating individual historical controls and aggregate treatment effect estimates into a Bayesian survival trial: a simulation study. BMC Med Res Methodol. 2019;19(1):85. https://doi.org/10.1186/s12874-019-0714-z.
27. Hampson LV, Whitehead J, Eleftheriou D, Brogan P. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. Stat Med. 2014;33(24):4186–201. https://doi.org/10.1002/sim.6225.
28. Molinares C. Parametric and Bayesian modeling of reliability and survival analysis. Graduate Theses and Dissertations. 2011;Available from: https://scholarcommons.usf.edu/etd/3252.
29. Omurlu IK, Ture M, Ozdamar K. Bayesian analysis of parametric survival models: A computer simulation study based informative priors. J Stat Manag Syst. 2015;18(5):405–23. https://doi.org/10.1080/09720510.2014.961763.
30. van Rosmalen J, Dejardin D, van Norden Y, Löwenberg B, Lesaffre E. Including historical data in the analysis of clinical trials: Is it worth the effort? Stat Methods Med Res. 2018;27(10):3167–82. https://doi.org/10.1177/0962280217694506.
31. Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharmaceutical Statistics. 2014;13(1):41–54. https://doi.org/10.1002/%2Fpst.1589.
32. Cuffe RL. The inclusion of historical control data may reduce the power of a confirmatory study. Statistics in Medicine. 2011;30(12):1329–38. https://doi.org/10.1002/sim.4212 PMID: 21432893.
33. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. Clinical Trials (London, England). 2010;7(1):5–18. Available from: https://doi.org/10.1177/1740774509356002.
34. Johnson VE, Cook JD. Bayesian design of single-arm phase II clinical trials with continuous monitoring. Clin Trials. 2009;6(3):217–26. https://doi.org/10.1177/1740774509105221.
35. Berger JO, Wolpert RL. The likelihood principle. 2nd ed. Institute of Mathematical Statistics; 1988.
36. Rouder JN. Optional stopping: No problem for Bayesians. Psych Bulletin Rev. 2014;21(2):301–8. https://doi.org/10.3758/s13423-014-0595-4.
37. Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HLJ, Kievit RA. An agenda for purely confirmatory research. Perspect Psychol Sci. 2012;7(6):632–8. https://doi.org/10.1177/1745691612463078.
38. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. Sage Publications Sage CA: Thousand Oaks, CA; 2005. https://doi.org/10.1191/1740774505cn098oa.
39. Robbins H. Some aspects of the sequential design of experiments. Bulletin of the Ame Math Soc. 1952;58(5):527–35. https://doi.org/10.1007/978-1-4612-5110-1_13.
40. Anscombe FJ. Fixed-sample-size analysis of sequential observations. Biometrics. 1954;10(1):89–100. https://doi.org/10.2307/3001665.
41. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. Am Stat. 1966;20(2):18–23. https://doi.org/10.1080/00031305.1966.10479786.
42. Lan GKK, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983;70(3):659–63. https://doi.org/10.1093/biomet/70.3.659.
43. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979;p. 549–556. https://doi.org/10.2307/2530245.
44. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977;64(2):191–199. Publisher: [Oxford University Press, Biometrika Trust]. https://doi.org/10.1093/biomet/64.2.191.
45. Burnett T, Mozgunov P, Pallmann P, Villar SS, Wheeler GM, Jaki T. Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. BMC Med. 2020;18(1):1–21. https://doi.org/10.1186/s12916-020-01808-2.
46. Schnuerch M, Erdfelder E. Controlling decision errors with minimal costs: The sequential probability ratio *t* test. Psychol Methods. 2020;25(2):206. https://doi.org/10.1037/met0000234.
47. Schönbrodt FD, Wagenmakers EJ, Zehetleitner M, Perugini M. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. Psychol Methods. 2017;22(2):322. https://doi.org/10.1037/met0000061.
48. Wald A. Sequential tests of statistical hypotheses. Ann Math Stat. 1945;16(2):117–86. https://doi.org/10.1007/978-1-4612-0919-5_18.
49. Berry DA. Interim analyses in clinical trials: Classical vs Bayesian approaches. Stat Med. 1985;4(4):521–6. https://doi.org/10.1002/sim.4780040412.
50. Ibrahim JG, Chen MH, Sinha D. Bayesian survival analysis. Springer Series in Statistics. New York: Springer-Verlag; 2001. Available from: https://doi.org/10.1007/978-1-4757-3447-8.
51. Hinne M, Gronau QF, van den Bergh D, Wagenmakers EJ. A conceptual introduction to Bayesian model averaging. Adv Methods Pract Psychol Sci. 2020;3(2):200–15. https://doi.org/10.1177/2515245919898657.
52. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Stat Sci. 1999;14(4):382–401. https://doi.org/10.1214/SS\%2F1009212519.
53. Leamer EE. Specification searches: Ad hoc inference with nonexperimental data, vol. 53. New York: Wiley; 1978.
54. Kleinbaum DG, Klein M. Survival analysis: A self-learning text, third edition. 3rd ed. Statistics for Biology and Health. New York: Springer-Verlag; 2012. https://doi.org/10.1007/978-1-4419-6646-9.
55. Latimer NR. NICE DSU technical support document 14: Survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data. Report by the Decision Support Unit. 2011;Available from: https://www.ncbi.nlm.nih.gov/books/NBK395885/pdf/Bookshelf_NBK395885.pdf.
56. Buckland ST, Burnham KP, Augustin NH. Model selection: An integral part of inference. Biometrics. 1997;53(2):603–618. Publisher: [Wiley, International Biometric Society]. https://doi.org/10.2307/2533961.
57. Hjort NL, Claeskens G. Frequentist model average estimators. J Am Stat Assoc. 2003;98(464):879–99. https://doi.org/10.1198/016214503000000828.
58. Wagenmakers EJ, Farrell S. AIC model selection using Akaike weights. Psychonomic Bulletin & Rev. 2004;11(1):192–6. https://doi.org/10.3758/BF03206482.
59. Bell Gorrod H, Kearns B, Stevens J, Thokala P, Labeit A, Latimer N, et al. A review of survival analysis methods used in NICE technology appraisals of cancer treatments: Consistency, limitations, and areas for improvement. Med Decis Making. 2019;39(8):899–909. https://doi.org/10.1177/0272989x19881967.
60. Kearns B, Stevens J, Ren S, Brennan A. How uncertain is the survival extrapolation? A study of the impact of different parametric survival models on extrapolated uncertainty about hazard functions, lifetime

mean survival and cost effectiveness. Pharmacol Econ. 2020;38(2):193–204. https://doi.org/10.1007/s40273-019-00853-x.
61. Ishak KJ, Kreif N, Benedict A, Muszbek N. Overview of parametric survival analysis for health-economic applications. PharmacoEconomics. 2013;31(8):663–75. https://doi.org/10.1007/s40273-013-0064-3.
62. Brard C, Le Teuff G, Le Deley MC, Hampson LV. Bayesian survival analysis in clinical trials: What methods are used in practice? Clin Trials. 2017;14(1):78–87. https://doi.org/10.1177/1740774516673362.
63. Wadsworth I, Hampson LV, Jaki T. Extrapolation of efficacy and other data to support the development of new medicines for children: A systematic review of methods. Statist Methods Med Res. 2018;27(2):398–413. https://doi.org/10.1177/0962280216631359.
64. Stallard N, Todd S, Ryan EG, Gates S. Comparison of Bayesian and frequentist group-sequential clinical trial designs. BMC Med Res Methodol. 2020;20(1):1–14. https://doi.org/10.1186/s12874-019-0892-8.
65. Negrín MA, Nam J, Briggs AH. Bayesian solutions for handling uncertainty in survival extrapolation. Med Decis Making. 2017;37(4):367–76. https://doi.org/10.1177/0272989x16650669.
66. Thamrin SA, McGree JM, Mengersen KL. Modelling survival data to account for model uncertainty: a single model or model averaging? SpringerPlus. 2013;2(1):665. https://doi.org/10.1186/2193-1801-2-665.
67. Gallacher D, Auguste P, Connock M. How do pharmaceutical companies model survival of cancer patients? A review of NICE single technology appraisals in 2017. Int J Technol Assess Health Care. 2019;35(2):160–7. https://doi.org/10.1017/s0266462319000175.
68. O'Hagan A. Science, subjectivity and software (comment on articles by Berger and by Goldstein). Bayesian Analysis. 2006;1(3):445–50. https://doi.org/10.1214/06-BA116G.
69. Giovagnoli A. The Bayesian design of adaptive clinical trials. Int J Environ Res Public Health. 2021;18(2):1–15. https://doi.org/10.3390/IJERPH18020530.
70. Alberts SR, Sargent DJ, Nair S, Mahoney MR, Mooney M, Thibodeau SN, et al. Effect of Oxaliplatin, Fluorouracil, and Leucovorin with or without Cetuximab on survival among patients with resected stage III colon cancer: A randomized trial. JAMA. 2012;307(13):1383–93. https://doi.org/10.1001/jama.2012.385.
71. Bartoš F. RoBSA: An R package for robust Bayesian survival-analyses; 2022. CRAN. Available from: https://CRAN.R-project.org/package=RoBSA.
72. Fragoso TM, Bertoli W, Louzada F. Bayesian model averaging: A systematic review and conceptual classification. Int Stat Rev. 2018;86(1):1–28. https://doi.org/10.1111/insr.12243.
73. Jefferys WH, Berger JO. Ockham's razor and Bayesian analysis. American Scientist. 1992;80:64–72. Available from: http://www.jstor.org/stable/29774559.
74. Jeffreys H. Scientific inference. Cambridge: Cambridge University Press; 1931.
75. Jeffreys H. Theory of probability. 1st ed. Oxford, UK: Oxford University Press; 1939.
76. Etz A, Wagenmakers EJ. JBS Haldane's contribution to the Bayes factor hypothesis Test. Stat Sci. 2017;32:313–29. https://doi.org/10.1214/16-STS599.
77. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995;90(430):773–95. https://doi.org/10.1080/01621459.1995.10476572.
78. Rouder JN, Morey RD. Teaching Bayes' theorem: Strength of evidence as predictive accuracy. Am Stat. 2019;73(2):186–90. https://doi.org/10.1080/00031305.2017.1341334.
79. Wagenmakers EJ, Morey RD, Lee MD. Bayesian benefits for the pragmatic researcher. Curr Dir Psychol Sci. 2016;25(3):169–76. https://doi.org/10.1177/0963721416643289.
80. Box GE. Science and statistics. J Am Stat Assoc. 1976;71(356):791–9. https://doi.org/10.1080/01621459.1976.10480949.
81. Gronau QF, Heck DW, Berkhout SW, Haaf JM, Wagenmakers EJ. A primer on Bayesian model-averaged meta-analysis. Advances in Methods and Practices in Psychological Science. 2021;4(3). Available from: https://doi.org/10.1177/%2F25152459211031256.
82. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester: John Wiley & Sons; 2004.
83. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130(12):1005–13. https://doi.org/10.7326/0003-4819-130-12-199906150-00019.
84. Goodman SN. Of *p*-values and Bayes: a modest proposal. Epidemiol. 2001;12(3):295–7. https://doi.org/10.1097/00001648-200105000-00006.
85. Neyman J, Pearson ES. Contributions to the theory of testing statistical hypotheses. Statistical Research Memoirs. 1936(1):1–37.
86. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. J R Stat Soc: Series A (Statistics in Society). 1994;157(3):357–87. https://doi.org/10.2307/2983527.
87. Kruschke JK. Bayesian estimation supersedes the t test. J Exp Psychol Gen. 2013;142(2):573. https://doi.org/10.1037/a0029146.
88. Edwards W, Lindman H, Savage LJ. Bayesian statistical inference for psychological research. Psychol Rev. 1963;70(3):193–242. https://doi.org/10.1037/h0044139.
89. Rouder JN, Haaf JM, Aust F. From theories to models to predictions: A Bayesian model comparison approach. Commun Monogr. 2018;85(1):41–56. https://doi.org/10.1080/03637751.2017.1394581.
90. Royall R. On the probability of observing misleading statistical evidence. J Am Stat Assoc. 2000;95(451):760–8. https://doi.org/10.1080/01621459.2000.10474264.
91. Altman DG. Statistics in medical journals. Stat Med. 1982;1(1):59–71. https://doi.org/10.1002/sim.4780010109.
92. Altman DG. Statistics in medical journals: developments in the 1980s. Stat Med. 1991;10(12):1897–913. https://doi.org/10.1002/sim.4780101206.
93. Altman DG. Statistics in medical journals: some recent trends. Stat Med. 2000;19(23):3275–89. https://doi.org/10.1002/1097-0258(20001215)19:23\%3C3275::aid-sim626\%3E3.0.co;2-m.
94. Schönbrodt FD, Wagenmakers EJ. Bayes factor design analysis: Planning for compelling evidence. Psychon Bull Rev. 2018;25(1):128–42. https://doi.org/10.3758/s13423-017-1230-y.
95. Stefan AM, Gronau QF, Schönbrodt FD, Wagenmakers EJ. A tutorial on Bayes factor design analysis using an informed prior. Behav Res Methods. 2019;51(3):1042–58. https://doi.org/10.3758/s13428-018-01189-8.
96. Berger JO. Statistical decision theory and Bayesian analysis. New York: Springer Science & Business Media; 2013.
97. Ly A, Verhagen J, Wagenmakers EJ. Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. J Math Psychol. 2016;72:19–32. https://doi.org/10.1016/j.jmp.2015.06.004.
98. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Boca Raton: CRC Press; 1999.
99. Org R. Project Data Sphere; 2019. Publisher: re3data.org - Registry of Research Data Repositories. Available from: https://www.re3data.org/repository/r3d100013015.
100. Alliance for Clinical Trials in Oncology. A randomized phase III trial of Oxaliplatin (OXAL) plus 5-Fluorouracil (5-FU)/Leucovorin (CF) with or without Cetuximab (C225) after curative resection for patients with stage III colon cancer; 2012. https://doi.org/10.34949/zywx-9253.
101. Alberts SR, Sinicrope FA, Grothey A. N0147: a randomized phase III trial of oxaliplatin plus 5-fluorouracil/leucovorin with or without cetuximab after curative resection of stage III colon cancer. Clin Colorectal Cancer. 2005;5(3):211–3. https://doi.org/10.3816/ccc.2005.n.033.
102. Raftery AE, Madigan D, Volinsky CT. Accounting for model uncertainty in survival analysis improves predictive performance. In: In Bayesian Statistics 5. University Press; 1995. p. 323–349.
103. Bartlett M. A comment on D V Lindleys statistical paradox. Biometrika. 1957;44(34):533–533. https://doi.org/10.1093/biomet/44.3-4.533.
104. Madigan D, Raftery AE, York JC, Bradshaw JM, Almond RG. Strategies for graphical model selection. Lecture Notes in Statistics. New York: Springer; 1994. p. 91–100.
105. Gronau QF, van Erp S, Heck DW, Cesario J, Jonas KJ, Wagenmakers EJ. A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. Compr Results Social Psychol. 2017;2(1):123–38. https://doi.org/10.1080/23743603.2017.1326760.
106. Maier M, Bartoš F, Wagenmakers EJ. Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. Psychological Methods. 2022. https://doi.org/10.1037/met0000405.
107. Bartoš F, Maier M, Wagenmakers EJ, Doucouliagos H, Stanley TD. Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. Research Synthesis Methods; in press. https://doi.org/10.31234/osf.io/kvsp7.

108. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press; 2006.
109. Alliance for Clinical Trials in Oncology. Phase III randomized study of adjuvant immunotherapy with monoclonal antibody 17–1A versus no adjuvant therapy following resection for state II (modified Astler-Coller B2) adenocarcinoma of the colon (comparator arm); 2003. https://doi.org/10.34949/y9s0-nz36.
110. Sanofi. Multicenter international study of Oxaliplatin/ 5FU-LV in the adjuvant treatment of colon cancer; 2003. https://doi.org/10.34949/fm7n-cw30.
111. Alliance for Clinical Trials in Oncology. Phase III randomized study of adjuvant immunotherapy with monoclonal antibody 17–1A versus no adjuvant therapy following resection for state II (modified Astler-Coller B2) adenocarcinoma of the colon (experimental arm); 2003. https://doi.org/10.34949/57rt-nr42.
112. Lee MD, Wagenmakers EJ. Bayesian cognitive modeling: A practical course. New York: Cambridge University Press; 2013.
113. Denwood MJ. runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. Journal of Statistical Software. 2016;71(9):1–25. Available from: https://doi.org/10.18637/jss.v071.i09.
114. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175–97. https://doi.org/10.1002/sim.1203.
115. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102. https://doi.org/10.1002/sim.8086.
116. Efron B, Stein C. The jackknife estimate of variance. Ann Stat. 1981;9(3):586–96. https://doi.org/10.1214/aos/1176345462.
117. Hwang IK, Shih WJ, De Cani JS. Group sequential designs using a family of type I error probability spending functions. Stat Med. 1990;9(12):1439–45. https://doi.org/10.1002/sim.4780091207.
118. Anderson KM. Optimal spending functions for asymmetric group sequential designs. Biom J: J Math Methods in Biosci. 2007;49(3):337–45. https://doi.org/10.1002/bimj.200510205.
119. Rouder JN, Morey RD, Verhagen J, Province JM, Wagenmakers EJ. Is there a free lunch in inference? Top Cogn Sci. 2016;8(3):520–47. https://doi.org/10.1111/tops.12214.
120. Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Feldman BM. Methods to elicit beliefs for Bayesian priors: a systematic review. J Clin Epidemiol. 2010;63(4):355–69. https://doi.org/10.1016/j.jclinepi.2009.06.003.
121. Chaloner K. The elicitation of prior distributions. In Case Studies in Bayesian Biostatistics. 1996; 141-156.
122. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, et al. Uncertain judgements: eliciting experts' probabilities. Chichester: Wiley; 2006.
123. Mikkola P, Martin OA, Chandramouli S, Hartmann M, Pla OA, Thomas O, et al. Prior knowledge elicitation: The past, present, and future; 2021. Available from: https://arxiv.org/abs/2112.01380.
124. Wagenmakers EJ, Ly A. History and nature of the Jeffreys-Lindley paradox; 2021. Available from:https://arxiv.org/abs/2111.10191.
125. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2(1):6–10. https://doi.org/10.1038/s41562-017-0189-z.
126. Maier M, Lakens D. Justify your alpha: A primer on two practical approaches; 2021. Available from: https://doi.org/10.31234/osf.io/ts4r6.
127. Wagenmakers EJ. Approximate objective Bayes factors from *p*-values and sample size: The $3p\sqrt{n}$ rule; 2022. Available from: https://doi.org/10.31234/osf.io/egydq.
128. Good I. Standardized tail-area probabilities. J Stat Comput Simul. 1982;16(1):65–6. https://doi.org/10.1080/00949658208810607.
129. Lindley DV. Statistical inference. J R Stat Soc: Series B (Methodological). 1953;15(1):30–65. https://doi.org/10.1111/j.2517-6161.1953.tb00123.x.
130. Ramsay JO. Monotone regression splines in action. Stat Sci. 1988;3(4):425–41. https://doi.org/10.1214/ss/1177012761.
131. Brilleman SL, Elci EM, Novik JB, Wolfe R. Bayesian survival analysis using the rstanarm R package. arXiv:200209633 [stat]. 2020 2;ArXiv: 2002.09633. Available from: http://arxiv.org/abs/2002.09633.
132. King AJ, Weiss RE. A general semiparametric Bayesian discrete-time recurrent events model. Biostatistics. 2019;(kxz029). Available from: https://doi.org/10.1093/biostatistics/kxz029.
133. Sinha D, Dey DK. Semiparametric Bayesian analysis of survival data. J Am Stat Assoc. 1997;92(439):1195–212. https://doi.org/10.1080/01621459.1997.10474077.
134. Held L, Sabanés Bové D. Applied statistical inference. New York: Springer; 2014.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions