Data-driven Prior Elicitation for Bayes Factors in Cox Regression for Nine Subfields in Biomedicine

Maximilian Linde[1,2], Laura Jochim[3], Jorge N. Tendeiro[4], and Don van Ravenzwaaij[2]

[1]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

[2]University of Groningen, Groningen, The Netherlands

[3]Utrecht University, Utrecht, The Netherlands

[4]Hiroshima University, Hiroshima, Japan

Author Note

Correspondence concerning this article should be addressed to: Maximilian Linde, Department of Computational Social Science, GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Köln, Germany,

E-mail: maximilian.linde@gesis.org.

Maximilian Linde https://orcid.org/0000-0001-8421-090X

Jorge N. Tendeiro https://orcid.org/0000-0003-1660-3642

Don van Ravenzwaaij https://orcid.org/0000-0002-5030-4091

Abstract

Biomedical research often utilizes Cox regression for the analysis of time-to-event data. The pervasive use of frequentist inference for these analyses implicates that the evidence for or against the presence (or absence) of an effect cannot be directly compared and that researchers must adhere to a predefined sampling plan. As an alternative, the use of Bayes factors improves upon these limitations, which is especially important for costly and time-consuming biomedical studies. However, Bayes factors involve their own difficulty of specifying priors for the parameters of the statistical model. In this article, we develop data-driven priors centered around zero for Cox regression tailored to nine subfields in biomedicine. To this end, we extracted hazard ratios and associated $x\%$ confidence intervals from the abstracts of large corpora of already existing studies within the nine biomedical subfields. We used these extracted data to inform priors for the nine subfields. All of our suggested priors are Normal distributions with means of 0 and standard deviations closely scattered around 1. We propose that researchers use these priors as reasonable starting points for their analyses.

Keywords: Bayes factor, Cox regression, hypothesis testing, prior elicitation, survival

Data-driven Prior Elicitation for Bayes Factors in Cox Regression for Nine Subfields in Biomedicine

Introduction

The collection and analysis of time-to-event data forms a central part of modern biomedical research (see, e.g., Boney et al., 2005; D'Agostino et al., 2008; Diener et al., 2004; Kenchaiah et al., 2002; Singh and Mukhopadhyay, 2011; Stupp et al., 2009). In these types of designs, the outcome variable is the time until an event of interest occurs, which is commonly called the survival time. In the medical context, this event of interest could be, for example, death, relapse towards alcoholism, disease/symptom onset, or recovery. Using survival analysis (we refer the interested reader to other sources, which treat survival analysis more thoroughly; e.g., Bradburn et al., 2003a, 2003b; Clark et al., 2003a, 2003b; Collett, 2015; Harrell, 2015; Hosmer et al., 2008; Klein and Moeschberger, 1997; Therneau and Grambsch, 2000), it is possible to estimate differences in event rates between conditions, which makes it particularly appealing for clinical trials. For instance, evidence about the effectiveness of an oncological treatment of cancer patients could be gathered by comparing the survival times of patients receiving the treatment to the survival times of patients receiving a placebo or an active control treatment (for examples see Kantoff et al., 2010; Rinke et al., 2009; Stupp et al., 2009).

The use of frequentist inference for the analysis of survival data has a long tradition in biomedical research and is still very common today Brard et al., 2017. Classical frequentist inference, however, is not well suited to quantify evidence in favor of the absence of an effect. In addition, frequentist inference requires fixing the study sample size in advance to avoid an inflation of the Type I and/or Type II error rates Yu et al., 2014, a downside that can be particularly costly in the realm of resource-intensive biomedical research and clinical trials Berry, 2006; Brard et al., 2017; Moyé, 2008. A prominent way to deal with this issue within the frequentist framework is the use of group sequential designs, whereby sequential testing and optional stopping of data collection is permitted at

the cost of implementing some procedure that controls the Type I and/or Type II error rates. These designs have a long and successful tradition in biomedicine (e.g., Pocock, 1977). As an alternative, Bayesian statistics have gained popularity among researchers van de Schoot et al., 2017. In particular, Bayes factors Jeffreys, 1939, 1948, 1961; Kass and Raftery, 1995 do not suffer from the two limitations mentioned before and are therefore a valuable alternative for conducting inference in biomedical research Goodman, 1999b.

Bayesian modeling with the goal of either hypothesis testing or parameter estimation requires the specification of a prior distribution for the parameters of the model. The prior expresses one's beliefs about the plausibility of parameter values before observing the data Kruschke, 2010, 2015; Kruschke and Liddell, 2018. Oftentimes, it is notoriously difficult to express these beliefs. Moreover, within Bayesian hypothesis testing, different priors sometimes lead to qualitatively different Bayes factors Gallistel, 2009; Kruschke and Liddell, 2018; Liu and Aitkin, 2008; Sinharay and Stern, 2002; Tendeiro and Kiers, 2019; Vanpaemel, 2010. As a result, some researchers lament that the use of Bayesian statistics involves subjectivity (e.g., Efron, 1986) and that proper guidance is missing, possibly resulting in hesitation to use Bayesian inference. Hence, recommendations for well-established default priors in Bayesian survival analysis - in particular Cox proportional hazards regression (henceforth called Cox regression or Cox model; Cox, 1972) - are missing and urgently needed.

In this article, we propose default priors for Bayesian Cox regression tailored to nine subfields within biomedicine. The construction of these priors harnesses historical records consisting of large corpora of hazard ratios and associated $x\%$ confidence intervals from existing studies within the respective subfields. We argue that these proposed priors can be used as reasonable defaults or starting points for biomedical researchers wishing to conduct a Bayesian Cox regression.

The remainder of this article is structured as follows: First, we provide an overview of Cox regression and Bayes factors. Second, we briefly review how priors can be defined.

Third, we explain our process of generating priors for nine subfields in biomedicine and present the corresponding results. Fourth, we reflect on our findings and implications thereof, and conclude with recommendations.

Bayes Factors in Cox Regression

Survival analysis is a statistical method to analyze time-to-event/survival data. Among the many existing forms of survival analysis - for example, Kaplan-Meier product-limit estimator Kaplan and Meier, 1958, parametric survival analyses (e.g., Exponential, Gompertz, and Weibull), and Cox regression Cox, 1972 - the latter is used most frequently within biomedicine (e.g., Bradburn et al., 2003a). Therefore, our treatment of priors for survival analysis is limited to the case of Cox regression.

In Cox regression, the hazard function $\lambda(t)$ presents the risk of an event happening in a small time period around a specific time t within cases for which the event has not yet happened before time t (see, e.g., Clark et al., 2003a; Harrell, 2015). The specific $\lambda(t)$ is allowed to have any shape but must be proportional across all values of an independent variable x. Usually, the main goal is not to estimate $\lambda(t)$ but rather to estimate the β parameter of the Cox model:

$$\lambda(t \mid x) = \lambda(t)\, e^{x\beta}. \tag{1}$$

Here and throughout, we work with the specific case where x is dichotomous and dummy-coded (i.e., there are two conditions, a common situation in biomedical designs). For this scenario, a hazard ratio can be calculated

$$\mathrm{HR} = e^{\beta}, \tag{2}$$

which provides information about the relative hazard rates between conditions.

In clinical trials, HR is often the key indicator regarding the effectiveness of a treatment. $\mathrm{HR} = 1$ (or $\beta = 0$) means that the two conditions have the same risk of the event happening at any t; $\mathrm{HR} > 1$ (or $\beta > 0$) means that the experimental condition has a higher risk of the event happening at any t; $\mathrm{HR} < 1$ (or $\beta < 0$) means that the control

condition has a higher risk of the event happening at any t. Frequentist inference on HR or β is then conducted either in the form of null hypothesis significance testing (i.e., test statistic and p-value) or in the form of estimation (i.e., a point estimate accompanied with a confidence interval).

The reliance on frequentist inference Chavalarias et al., 2016; Goodman, 1999a has some undesirable consequences for biomedical research. Here, we focus on two of these consequences, namely (1) the impossibility to obtain evidence for the null hypothesis and (2) the inability to adjust the sampling plan based on interim results. Concerning (1), it is important to not only determine whether a treatment is working but also whether a treatment is *not* working over and above a placebo effect. The frequentist approach is not suitable for this because it only allows rejecting the hypothesis that there is no effect, but not accepting it (e.g., Bakan, 1966; Gallistel, 2009; Hoekstra et al., 2018; Ioannidis, 2005; Rouder et al., 2009; Wagenmakers, 2007).

Concerning (2), the use of frequentist inference prescribes the diligent adherence to a predefined sampling plan, prohibiting to continue or prematurely stop data collection based on interim data analyses (e.g., Armitage et al., 1969; Rouder, 2014; Schönbrodt et al., 2017; Tendeiro et al., 2022). Further criticism is described elsewhere (see, e.g., Goodman, 1999a; International Committee of Medical Journal Editors, 1997; Rennie, 1978; van Ravenzwaaij et al., 2019; Wagenmakers, 2007).

Bayesian testing in the form of Bayes factors permits a direct comparison between the evidence for the null hypothesis that there is no effect and an alternative hypothesis that operationalizes that there is some effect Rouder et al., 2009. For instance, with $\mathrm{BF}_{10} = 14$, it allows the interpretation that the obtained data is 14 times more likely under the chosen hypothesis that there is some effect compared to the hypothesis that there is no effect; similarly, $\mathrm{BF}_{10} = 0.2$ indicates that the obtained data is $1/0.2 = 5$ times more likely under the hypothesis that there is no effect compared to the chosen hypothesis that there is some effect.

Moreover, using Bayes factors, it is legitimate to monitor the results and stop data collection once a predetermined evidence threshold is reached Armitage et al., 1969; Rouder, 2014; Schönbrodt and Wagenmakers, 2018; Schönbrodt et al., 2017; Stefan, Schönbrodt, et al., 2022; Tendeiro et al., 2022. Thus, Bayes factors take the evidence for and against both the null and alternative hypotheses into account, yield more substantial interpretations, and empower researchers to sample just the sufficient amount of cases. These characteristics of Bayes factors are critical for biomedical research as studies (especially clinical trials) can be expensive and time-consuming.

In the case where there is a point null hypothesis stating that there is no effect

$$\mathcal{H}_0 : \ \beta = 0 \tag{3}$$

and an interval alternative hypothesis stating that there is an effect

$$\mathcal{H}_1 : \ \beta \sim f\left(\phi\right), \tag{4}$$

the Bayes factor is a ratio of a marginal likelihood and a likelihood evaluated at $\beta = 0$. Here, $f\left(.\right)$ represents any probability density function and ϕ the associated parameters. Let Ω_1 be the parameter space under the alternative hypothesis; then the Bayes factor is:

$$\mathrm{BF}_{10} = \frac{P\left(D \mid \mathcal{H}_1\right)}{P\left(D \mid \mathcal{H}_0\right)} = \frac{\int_{\beta \in \Omega_1} \overbrace{f\left(D \mid \beta\right)}^{\text{Likelihood}} \overbrace{f\left(\beta\right)}^{\text{Prior}} \mathrm{d}\beta}{\underbrace{f\left(D \mid \beta = 0\right)}_{\text{Likelihood at } \beta=0}}. \tag{5}$$

In Eq (5), the integral constitutes a weighted average of the likelihood, with weights supplied by the prior. Depending on the complexity of the underlying statistical model, computing the expression in Eq (5) can be challenging. Through concerted efforts of researchers to develop closed-form solutions and through the explosion of computational power over recent decades that allows applying complex numerical methods (e.g., Monte Carlo sampling and bridge sampling; Betancourt, 2018; Brooks et al., 2011; Gilks et al., 1995; Gronau et al., 2017; van Ravenzwaaij et al., 2018), computing the expression in Eq (5) and variants of it has become feasible. These efforts have led to method

developments and software implementations for calculating Bayes factors for survival analyses Bartoš, 2022; Bartoš et al., 2022; Linde, Tendeiro, et al., 2022; Linde, van Ravenzwaaij, et al., 2022 and many other designs (e.g., Gronau et al., 2020; Gu et al., 2021; Heck et al., 2020; JASP Team, 2023; Morey and Rouder, 2018; Rouder et al., 2012; Rouder et al., 2009; van Lissa et al., 2021; van Ravenzwaaij et al., 2019).

Choosing Priors

Even though it is possible nowadays to calculate Bayes factors for various sorts of designs, it remains difficult to specify an appropriate prior distribution (or prior for short) for the parameters of interest (cf. Eq (5)). The prior is a probability distribution that is placed on the statistical model's parameters of interest and it expresses belief over the plausibility across all possible parameter values before taking into account new data. In the context of null hypothesis Bayes factor calculations, the prior is one important element of the alternative hypothesis. Even among researchers who advocate Bayesian statistics, there is disagreement on how the prior should be specified (see, e.g., Berger, 2006; Goldstein, 2006).

Objective Bayesians strive to define non-informative priors that are as "objective" as possible. Objective Bayesians assert that the results of Bayesian analyses should depend only to a minimal extent on the beliefs of different people. They promote default priors that can be used when no other information is available and often seek to find priors that "behave well" and fulfill certain mathematical properties (see, e.g., Bayarri et al., 2012; Consonni et al., 2018 for more details). Subjective Bayesians, on the other hand, counter that the subjective nature of the prior is an integral part of Bayesian analyses. According to them, the prior allows the incorporation of domain knowledge and results from prior studies into Bayesian analyses and therefore permits tests of theories Dienes, 2011; Vanpaemel, 2010. Further, they state that it is questionable whether a truly "objective" prior even exists.

Recently, the opportunities that well-defined priors open were increasingly

recognized in biomedical research. Prior elicitation procedures, in which informed priors are defined by means of using external sources, gained popularity within biomedical research (e.g., Guo et al., 2019; Johnson et al., 2010; Thall and Cook, 2004; van de Schoot et al., 2018; Zondervan-Zwijnenburg et al., 2017). There are various forms of prior elicitation. For example, through structured interviews, information about prior beliefs can be extracted by one or multiple experts in the respective field, which is subsequently combined into one prior (see, e.g., Johnson et al., 2010; O'Hagan et al., 2006; Stefan, Evans, et al., 2022; van de Schoot et al., 2018). Tools like the MATCH software Morris et al., 2014 have been developed for this purpose. Alternatively, the results of meta-analyses and prior research in general can be used to create a prior (e.g., Rietbergen et al., 2011; van de Schoot et al., 2018). That is, researchers could use the overall effect size combined with a measure of uncertainty from a meta-analysis to construct an informed prior for their own analysis; or they could conduct their own literature search and extract the relevant statistics and use them for developing priors. This approach of using prior study results can also be combined with an empirical Bayes approach, which utilizes the current data to create a prior instead of predefining it (e.g., Casella, 1985). Such a procedure was proposed by van Zwet and Gelman, 2022.

In this article, we follow the approach of using results of prior studies to create priors. For this, we conducted our own literature search instead of relying on meta-analyses. The reason for this is that we aim to suggest priors that are generic, such that they apply to entire medical subfields; most meta-analyses do not offer this generality.

The Current Study

In the present article, we develop priors centered around zero for Bayesian Cox regression for nine subfields that we believe are representative for different areas of research within biomedicine. For the construction of these priors we make use of reported hazard ratios and associated $x\%$ confidence intervals from large corpora of existing studies. These extracted data are then combined through pooling to generate priors. These priors apply

to two-sided alternative hypotheses. Since we are ignorant of the direction of effects, our proposed priors can be truncated at 0 for directional hypotheses.

Methods

We selected the subfields in biomedicine considered for further investigation based on a taxonomy provided by Scimago (available at https://www.scimagojr.com/journalrank.php; SCImago, n.d.). On the Scimago website, we used "Medicine" as a subject area, upon which a list of medical subfields were provided (see Fig 1). Among those, we selected the following nine subfields for further consideration:

1. Anesthesiology and pain medicine
2. Cardiology and cardiovascular medicine
3. Gastroenterology
4. Hematology
5. Immunology and allergy
6. Neurology
7. Oncology
8. Psychiatry and mental health
9. Pulmonary and respiratory medicine

Our selection was based on three criteria: (1) we aimed to obtain a manageable number of subfields (between eight and twelve); (2) we aimed to obtain subfields with limited overlap; and (3) we aimed to obtain subfields that represent relatively large areas of study within biomedicine. For each of the nine subfields, we obtained a list of the top journals from Scimago. We considered only journals (i.e., neither book series, nor conferences and proceedings, nor trade journals); we considered journals from all

regions/countries; and the journal list was based on the year 2022 (see Fig 1 for an example of the settings on the Scimago website for the subfield of "Anesthesiology and pain medicine"). The extraction of the top journals for all nine subfields yielded 2,469 journals in total (see columns 1 and 2 of Table 1). Some subfields shared a set of journals; for example, the journal "Pain" is a top journal for both the subfields of "Anesthesiology and pain medicine" and "Neurology". We found that there was not a lot of overlap of journals between the subfields, with 2,196 of the 2,469 journals being uniquely assigned to only one subfield.

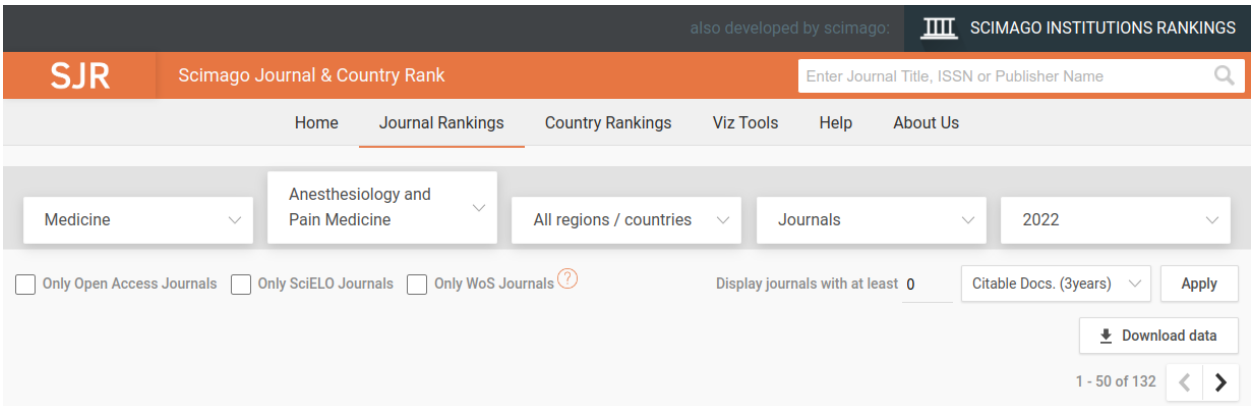

Figure 1

Settings for extracting top journals from Scimago (available at https://www.scimagojr.com/journalrank.php; SCImago, n.d.) for one of the nine considered subfields in biomedicine; in this case "Anesthesiology and pain medicine".

As a separate step, we used Scopus to obtain a list of medical articles. We used the following search query:

```
ABS(("hazard ratio" OR {hr}) AND {cox}) AND
SUBJAREA(medi) AND
PUBYEAR > 1999 AND PUBYEAR < 2021 AND
(LIMIT-TO(SRCTYPE,"j")) AND
(LIMIT-TO(PUBSTAGE,"final")) AND
```

Table 1

Number of used journals and studies for each of the nine subfields within biomedicine. The first column indicates the subfield, the second column the number of used (i.e., matched between Scopus data and Scimago data) journals (not necessarily unique) from all Scimago journals, the third column the number of studies allocated, the fourth column the number of studies for which there was a match and data extraction was successful, and the fifth column the number of studies remaining after excluding studies that provide flawed results.

Subfield	N journals	N studies allocated	N studies matched	N studies considered
Anesthesiology and pain medicine	44/132	360	215	211
Cardiology and cardiovascular medicine	230/366	10,718	6,555	6,504
Gastroenterology	88/157	2,088	1,311	1,300
Hematology	66/134	1,601	857	849
Immunology and allergy	82/214	1,360	839	833
Neurology	178/387	2,840	1,651	1,640
Oncology	239/373	13,163	7,741	7,684
Psychiatry and mental health	159/560	1,750	1,038	1,029
Pulmonary and respiratory medicine	84/146	2,551	1,561	1,548
All	1,170/2,469	36,431/59,646	21,768	21,598

```
(LIMIT-TO(DOCTYPE,"ar")) AND
(LIMIT-TO(LANGUAGE,"English"))
```

Only fully published (line 5) articles (line 6) from a journal (line 4) written in English (line 7) were considered. Furthermore, the results had to belong to the field of medicine (line 2) and be published between the years 2000 and 2020, inclusive (line 3). Lastly, the abstracts of the results needed to contain the keywords “hazard ratio” or “HR” and the keyword

"Cox", ignoring case (line 1). Note, however, that this search query was generic such that it did not restrict the results towards one of the nine subfields. The Scopus query yielded $59,669$ results, of which 23 could not be exported, leaving $59,646$ results in total (see column 3 of Table 1).

The $59,646$ Scopus results were allocated to the nine subfields by matching the journal names indicated by Scopus to the Scimago lists of journal names for the nine subfields. Importantly, a Scopus result could be allocated to multiple subfields as some subfields had journals in common. Before the allocation of Scopus results to the nine subfields, both the Scopus journal names and the Scimago lists of journal names were cleaned and standardized in order to accommodate slight differences in their presentation. This included replacing "&" and "&" with "AND", removing all characters that are not alphabetic or white space, repositioning the word "the" (e.g., "Lancet Oncology, The" was turned into "The Lancet Oncology"), and transforming all characters to uppercase. The number of matched journal names relative to the total number of Scimago journal names for the nine subfields are shown in column 2 of Table 1 and the number of allocated results for each of the nine subfields can be seen in column 3 of Table 1.

Once the individual Scopus results were allocated to the nine subfields, we extracted hazard ratios and associated $x\%$ confidence intervals from the abstracts of the results. This was done in an automatic fashion through the use of regular expressions (see Friedl, 2006 for the standard reference on regular expressions). We extracted the following information from the abstracts:

- Hazard ratio (HR)
- Confidence level of the confidence interval (CI) for HR (i.e., $100\,(1-\alpha)$)
- Lower boundary of the CI for HR (HR^l)
- Upper boundary of the CI for HR (HR^u)

There are several important details about our implemented text-mining procedure. First, we exclusively considered the abstracts for data extraction. In an attempt to ascertain that this did not result in bias in the extracted effect sizes, we randomly sampled 100 articles for which we could match results in the abstract with our regular expression and another 100 articles for which we could not match results in the abstract with our regular expression. For these articles, we applied our regular expression on the main text (i.e., not the abstracts). The effect size distributions from this sample looked very similar, indicating no strong evidence for a bias as a result of focusing our search on abstracts only (see Fig A1 in Appendix A). Second, if the regular expression yielded multiple matches for a given abstract, only the first match was considered; any other matches were discarded. The justification for this decision was that we assumed that the main findings are commonly reported first, followed by secondary or exploratory findings.

Third, data extraction was only done when all of the four above-mentioned information were available in the abstract. We disregarded abstracts where results were not complete or were presented in any other form. For instance, presentations of a HR coupled with a p-value and potentially a test statistic were ignored. Although this seems like an overly drastic measure, the number of matches was still very high (see column 4 of Table 1). Fourth, we did not distinguish between variations of Cox regression (e.g., multivariate, stratified, multiple predictors).

Fifth, we allowed various forms of the displayed results. For example, the following variations were all captured by our regular expression: "HR = 2.3" (with or without spaces around =), "hazard ratio = 2.3", "hazards ratio = 2.3", "hazard ratios = 2.3", "hazard ratio (HR) = 2.3", "hazards ratio [HR] = 2.3", "HR : 2.3", "H.R. : 2.3", and many more. Thus, we attempted to make the regular expression as flexible as possible, so that it could capture the maximum amount of valid text, while still maintaining a healthy level of restrictions. For more details, please consult our code, available at https://osf.io/ua4ys/. In total, we were able to extract data for $21,768$ out of $36,431$ results (see column 4 of

Table 1).

We applied additional checks on the extracted data to make sure that both the regular expression worked properly and the information in the abstracts themselves was correct. As a first step, we checked whether the confidence level of the CI was between 0 and 100. Second, we tested whether HR, HR^l, and HR^u were higher than 0 (because the possible range goes from 0 to ∞). Third, we examined whether the $\log \mathrm{HR}$ was approximately (because of rounding) in the middle of HR^l and HR^u. Here, we also excluded results where the $\log \mathrm{HR}$ and at least one of HR^l or HR^u had the same value due to rounding. Any extracted data that did not fulfill all of these criteria was discarded. Column 5 of Table 1 shows that for all nine subfields only a small number of extracted data had to be excluded (170 out of $21,768$ in total), leaving a final number of considered results of $21,598$.

With this step completed, the nine subfields and their hazard ratios and associated $x\%$ confidence intervals were considered separately. For each study i in $i \in \{1,\ 2,\ \ldots,\ N\}$ within one of the nine subfields (where N is the number of studies within one of the nine subfields), the extracted HR_i was log-transformed (b_i). Also, the standard error of b_i was calculated based on the confidence interval (e.g., Higgins et al., 2019) of HR (i.e., HR_i^l and HR_i^u):

$$SE\left(b_i\right) = \frac{\log \mathrm{HR}_i^u - \log \mathrm{HR}_i^l}{2z_i^*}, \tag{6}$$

where $z_i^* = Q\left(1 - \alpha_i/2\right)$ and $Q\left(.\right)$ is the quantile function of the standard Normal distribution. The sign of b_i, however, is meaningless because it depends on how the independent variable is coded. For instance, commonly the control condition is coded with 0 and the experimental condition with 1; occasionally, the opposite is the case, which would reverse the sign of b_i. We take this arbitrariness into account in any further calculations.

The obtained b_i and $SE\left(b_i\right)$ within a subfield could then be used for the construction of a prior for each subfield separately. We decided to use the Normal distribution for the prior. To obtain a prior, we combined the data through pooling (e.g.,

Higgins et al., 2019). Using this procedure, b_i and $SE\left(b_i\right)$ were treated as coming from separate samples that were combined into a single pooled sample. One desirable feature of this pooling method is that b_i values with higher $SE\left(b_i\right)$ are central to (rather than discarded for) the calculation of the pooled standard error. In other words, the $SE\left(b_i\right)$ around the b_i values had a direct influence on the calculation of the pooled standard error: All else being equal, the higher the $SE\left(b_i\right)$ of a sample, the more it would increase the pooled standard error. We deemed this behavior desirable since we wanted the prior to reflect uncertainty. The resulting mean and standard error of a single pooled sample served as the mean and standard deviation of the prior.

We decided against using the inverse-variance weighting procedure that is commonly used in meta-analyses (see, e.g., Borenstein et al., 2021). The reason for this was that the prior would get increasingly narrow as the corpus of considered studies increases. In addition, b_i with high $SE\left(b_i\right)$ values have a relatively low influence (they are less diagnostic in determining the mean), which was undesirable for our purposes of trying to estimate the spread of the prior. Also, we decided to not use the partly empirical Bayes procedure described in van Zwet and Gelman, 2022 because the nature of it prescribes that the current data (i.e., not only the corpus of prior studies) is used to determine the prior.

Assuming a constant n across studies i and taking into consideration the arbitrariness of the sign of b_i by using both b_i and $-b_i$ for each study i, we can compute the mean μ^p of the prior (cf. Higgins et al., 2019). Since we consider both b_i and $-b_i$ for each study i, μ^p must necessarily be 0:

$$\mu^p = \frac{n\sum_{i=1}^{N} b_i + n\sum_{i=1}^{N} -b_i}{2Nn} = \frac{\sum_{i=1}^{N} b_i - b_i}{2N} = 0. \tag{7}$$

Similarly, we can compute the standard deviation of the prior σ^p:

$$\sigma^p = \sqrt{\frac{\left(n-1\right)\sum_{i=1}^{N} SE\left(b_i\right)^2 + n\sum_{i=1}^{N} b_i^2}{Nn-1}}. \tag{8}$$

In essence, Eq (8) sums up $SE\left(b_i\right)^2$ and it sums up b_i^2. Thus, all else being equal, σ^p increases as any of the two quantities increases. Note that our extracted information did

not provide information on the sample size n_i within each study. We used an arbitrary sample size of $n = 200$ for all n_i. However, in order to verify the robustness of the assumption of equal sample size per study, we randomly varied n_i across studies, so that it could take on values sampled from $\mathrm{U}\,(10,\ 10000)$. We repeated this procedure $100,000$ times to accommodate many possible arrangements of n_i.

Data Availability

Data used in section "Example Application" was shared with the authors by NIAID (see Beigel et al., 2020) and may not be distributed beyond the first and last author (for more information, see https://accessclinicaldata.niaid.nih.gov/). All other data may be obtained from the OSF page associated with the project: https://osf.io/ua4ys/.

Results

The results for the construction of the Normal priors for the nine subfields through the pooling method Higgins et al., 2019 can be seen in Fig 2. The panels represent the nine subfields. Histograms show the distributions of b_i (and $-b_i$) for the different corpora, independent of $SE\,(b_i)$. The red curves show the priors resulting from the application of the pooling method Higgins et al., 2019. Descriptive statistics can be found in Appendix B.

For all nine subfields, the center of the Normal prior was located at $\mu^p = 0$ as a necessary consequence of our decision to ignore the sign of b_i. The more interesting parameter of the Normal prior is the standard deviation σ^p because it reflects both the effect sizes and the uncertainties around them based on past studies within a subfield. From the histograms and the Normal priors it can be seen that the effect sizes, and therefore also the Normal priors, were similar across the nine subfields. σ^p ranged between 0.915 for "Psychiatry and mental health" and 1.079 for "Gastroenterology".

Since the calculation of σ^p through the pooling method Higgins et al., 2019 was based on an arbitrary choice of $n = 200$ that was the same for all studies within a corpus of a specific subfield, we also investigated the dependence of σ^p on n. Fig 3 shows this dependence, where the panels represent the different subfields. The box plots represent the

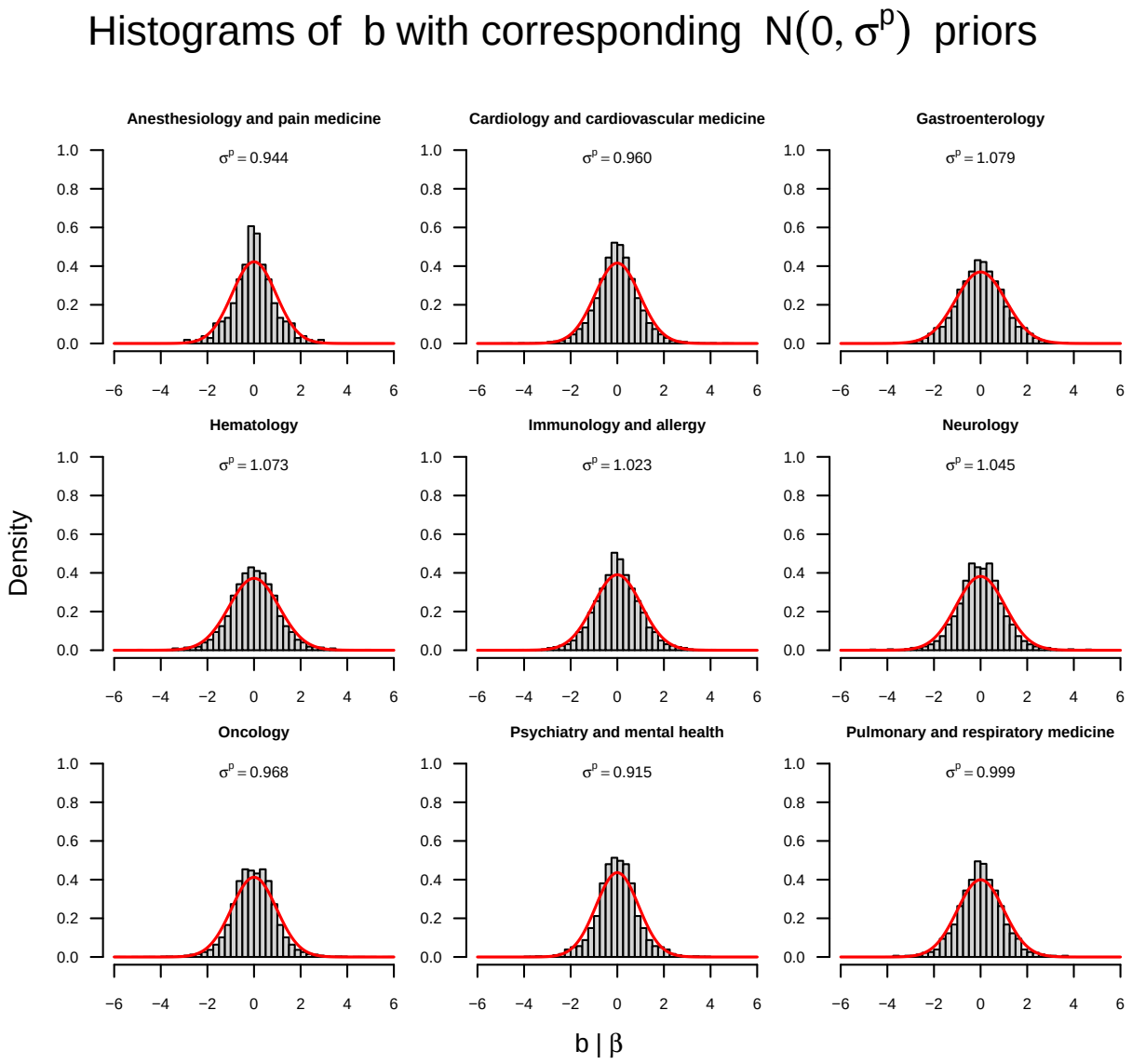

Figure 2

Prior distributions for the nine subfields, one subfield per panel. The histograms show the distribution of b_i (and $-b_i$), ignoring $SE(b_i)$. The red curves display the densities of the Normal priors on β, where σ^p is presented at the top of each panel.

variation of σ^p as a function of n_i, where n_i for each study i was sampled from $\mathrm{U}(10,\ 10000)$ (i.e., different sample sizes were possible across studies), over $100,000$ repetitions.

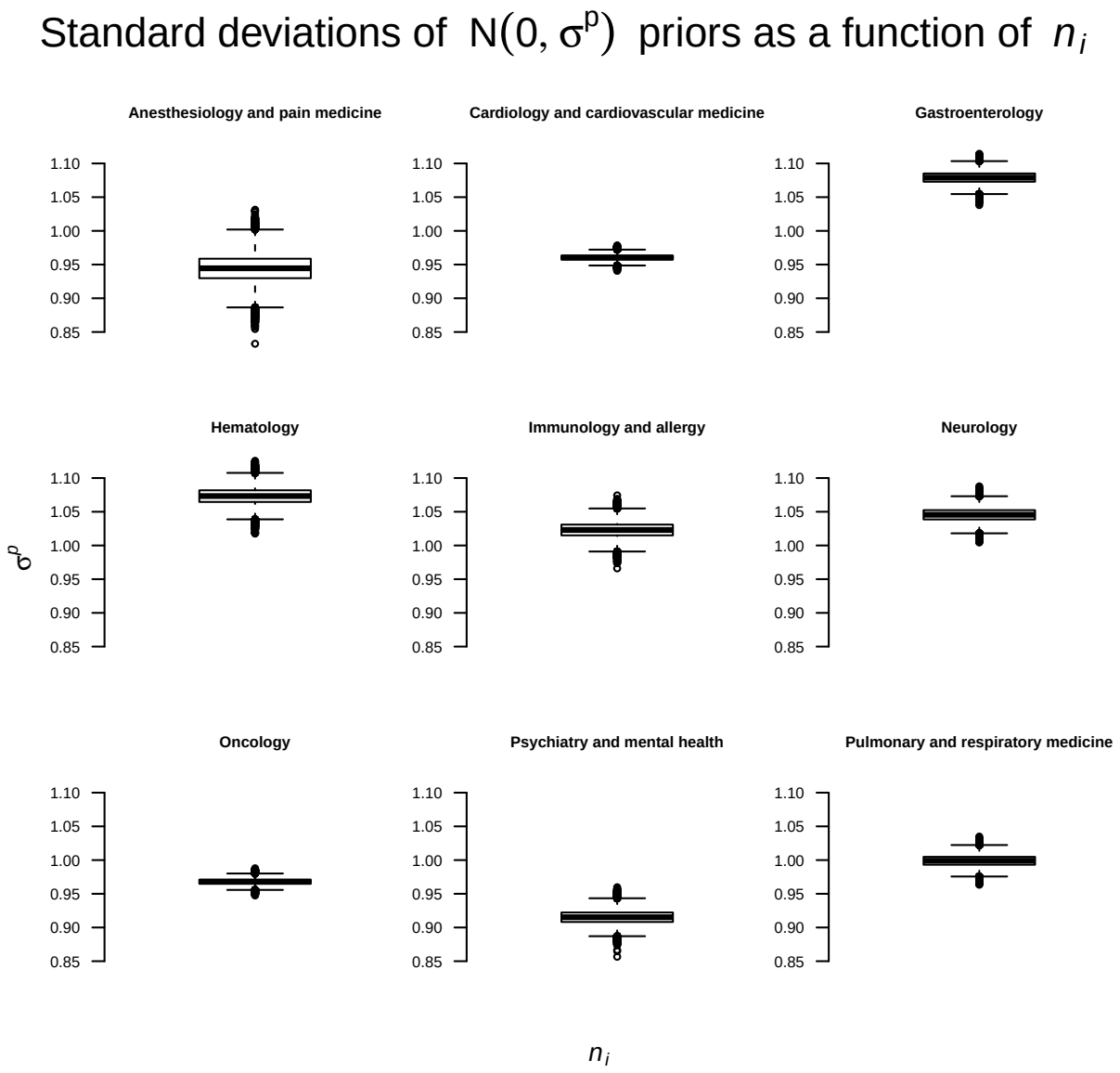

Figure 3

Sensitivity of the pooling method with respect to the choice of n_i for the estimation of σ^p. The nine panels correspond to the nine subfields. The box plots show σ^p of the priors when each n_i is drawn randomly from $\mathbf{U}(10,\ 10000)$ (i.e., studies have different sample sizes); this process was repeated $100,000$ times.

When n_i were allowed to vary across studies, some variation in σ^p was observed, as shown by the box plots. Still, the variation in σ^p was not radical enough to question our

heuristic of using $n = 200$ for all studies within a corpus of a subfield.

Example Application

To illustrate how the obtained priors can be used, we conducted a Bayesian Cox regression on a medical data set, which is described in Beigel et al., 2020. In this study, the authors aimed to examine a drug called Remdesivir, which promised to be beneficial for treating patients with Covid-19. The study included $n = 1,062$ participants, of which $n_c = 521$ were randomly placed in the placebo condition and $n_e = 541$ in the Remdesivir condition. All participants had a Covid-19 infection. The primary analysis focused on the amount of time until participants recovered.

It can be argued that this study falls both in the realm of "immunology and allergy" as well as "pulmonary and respiratory medicine" (i.e., $\sigma^p = 1.023$ and $\sigma^p = 0.999$, respectively). Therefore, we conducted our analyses with both priors. Furthermore, we think that it is always good practice to conduct sensitivity analyses. So, we further extended our used priors to entail Normal priors centered on $\mu^p = 0$ with $\sigma^p \in \{0.3, 0.4, ..., 1.7\}$. We argued that a major advantage of Bayesian statistics is to monitor and optionally stop data collection based on interim results. We demonstrate this procedure on the basis of the data set by Beigel et al., 2020. Unfortunately, the data file does not record the order in which participants entered the study. However, for the purposes of this demonstration, we deemed it appropriate to simply take a random order. In essence, we pretend that the study is still ongoing and we stop data collection once the resulting Bayes factor falls outside the decision threshold of 1/20 or 20. In general, the choice for a specific decision threshold should be tailored to the research at hand. There are scenarios where the priority lies with more certainty and scenarios where the priority lies with more efficiency. However, researchers should be aware of the tradeoff between certainty and resource demands; the more certain we want to be, the more cases must be sampled. Lastly, for our analysis we use a one-sided positive alternative hypothesis because it is hypothesized that patients receiving Remdesivir recover quicker than patients

receiving a placebo.

Fig 4 shows the development of the Bayes factors as new data enter the study for various priors. The blue and red lines correspond to the "immunology and allergy" and "pulmonary and respiratory medicine" priors and the gray lines to the priors used in the sensitivity analyses. It can be seen that all Bayes factor trajectories cross the upper threshold of 20 at some point, demonstrating the robustness of the findings. For some extreme priors, the upper threshold is already crossed at about 300 observations.

Discussion

Survival analysis, and in particular Cox regression Cox, 1972, is an indispensable statistical tool for biomedical research. The ubiquitous frequentist framework is limited in that it cannot quantify evidence in favor of the null hypothesis that there is no effect Rouder et al., 2009 and in that it does not allow continuing or stopping data collection based on interim analyses Armitage et al., 1969; Rouder, 2014; Schönbrodt and Wagenmakers, 2018; Schönbrodt et al., 2017; Stefan, Schönbrodt, et al., 2022; Tendeiro et al., 2022. Bayes factors remedy these shortcomings and permit intuitive interpretations. Nevertheless, the specification of priors that are required for Bayesian analyses can be difficult.

In this paper, we have developed priors for the β parameter in Cox regression for nine subfields in biomedicine (see column 1 in Table 1). These priors were informed by large corpora of already existing studies within the respective subfields and therefore provide reasonable approximations to the to-be-expected effect sizes and uncertainties thereof. For all nine subfields, we decided to use a Normal prior, which is centered on $\mu^p = 0$. We found very similar standard deviations for the Normal priors across the nine subfields, ranging from $\sigma^p = 0.915$ for "Psychiatry and mental health" to $\sigma^p = 1.079$ for "Gastroenterology", suggesting considerable similarities across subfields. Since our developed priors differ only slightly across the nine subfields, we believe that it is reasonable to use a standard Normal prior (i.e., $\mathrm{N}(0,\ 1)$) for all nine subfields, forming a

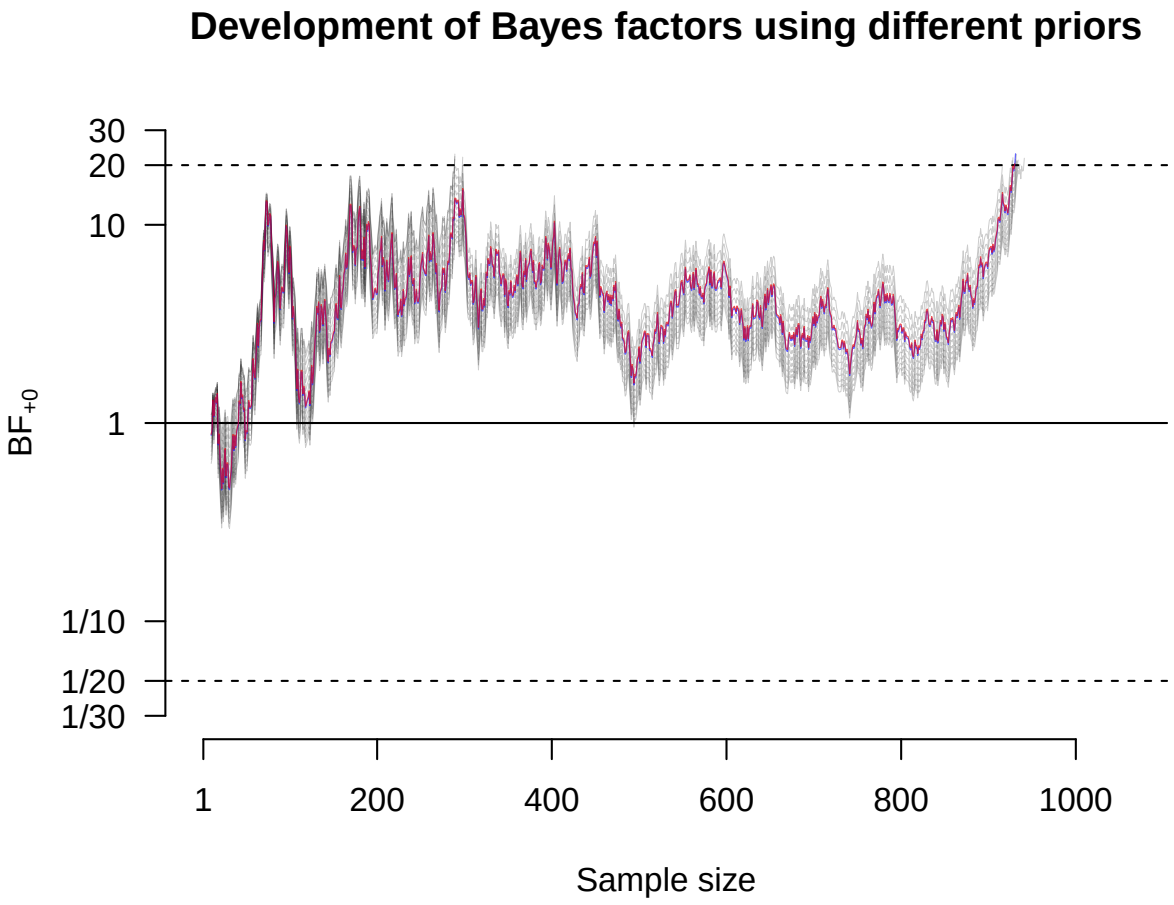

Figure 4

Results of a Bayesian Cox regression applied to the data set provided by Beigel et al., 2020. The blue and red lines show the development of the Bayes factor as new data enter the study for the "immunology and allergy" and "pulmonary and respiratory medicine" priors. The blue and red lines heavily overlap, so the blue line is hardly visible. The gray lines show the development of the Bayes factor for the sensitivity analysis (i.e., various Normal priors). Data collection is stopped once the Bayes factor exceeds $1/20$ **or** 20**.**

compromise among the nine individual priors, as a starting point. Still, any choice of prior is always arbitrary to some degree. Therefore, we urge researchers to complement their analysis using a specific prior with sensitivity analyses (e.g., Berger et al., 1994; Depaoli

and van de Schoot, 2017; Du et al., 2019; Kruschke, 2015), in which parameters of the prior are systematically varied and even entirely different (reasonable) priors are chosen, in order to examine the robustness of the resulting Bayes factors.

We caution the reader to not take our proposed priors to be set in stone. The choice of prior always depends to some extent on the goals of the researcher. For example, it might not always be desirable for the prior to accurately reflect expected effects. Sometimes, the focus might be on ensuring sufficient shrinkage of the parameter estimate (e.g., Park and Casella, 2008; van Erp et al., 2019; van Zwet and Gelman, 2022). Furthermore, we strongly advice researchers to only apply these priors in settings where they are appropriate. That is, these priors should only be applied in Cox regressions. We discourage the transfer of these priors to other methods, such as meta analysis or linear regression. In addition, our developed priors represent expected effect sizes of entire medical fields at a very global level. It is likely that researchers often have access to information that is more tailored to their research question. In that case, we encourage researchers to utilize this more informed and problem-specific information for the construction of a prior. We also want to highlight that researchers should define their prior, including the priors for any sensitivity analyses, before they inspect the data. Trying out multiple priors and then reporting those few that yield favorable results, is bad scientific practice.

Moreover, our process of arriving at the priors contained decisions, assumptions, and heuristics that might be questioned. First, the allocation of the articles to the nine subfields based alone on the journal is a drastic heuristic. A proportion of the articles is therefore probably classified into the wrong subfield. Second, the regular expression that we created to extract hazard ratios and associated $x\%$ confidence intervals from the abstracts of the articles might have been biased to some extent. Some journals have very specific reporting guidelines for the abstracts, which might not have been captured by our regular expression. Thus, it is possible that certain journals were systematically

underrepresented in our results.[1]. Third, we only matched abstracts where a hazard ratio is combined with a confidence interval but not abstracts where a hazard ratio is combined with a p-value. We made this decision because p-values often do not map directly onto the confidence intervals. Additional information on how the p-value was calculated would be needed, which is rarely available in abstracts. Fourth, for the Cox regression analyses, we did not differentiate between different types of predictors (e.g., categorical, continuous) and types of analyses (e.g., stratified, multivariate), leaving open the possibility that certain nuances are ignored by our calculations. Fifth, our results are predicated on articles that have come up in the Scopus search engine. Any articles that were published in predatory journals, but that were somehow not screened out as such (see e.g., https://www.elsevier.com/connect/the-guardians-of-scopus), may have introduced bias in the estimate of effect size variance, because of lack of proper peer review. Sixth, we assumed that HR refers to hazard ratio in the abstracts. However, HR could also be an abbreviation for other terms like heart rate. To mitigate this concern, we manually screened a random sample of 88 articles for which we could obtain a match and found that in all cases HR corresponded to hazard ratio.

It is clear that our proposed priors for Bayesian Cox regressions are very generic, such that one individual prior accommodates an entire subfield (or even all nine subfields if we are willing to accept the standard Normal prior). These priors might still be appropriate approximations for smaller specializations within subfields. However, in these cases it might be worthwhile to obtain more informed and precise priors that are tailored to these smaller subfields.

[1] We investigated this potential bias by examining the number of articles that remain after conducting the processing steps shown in Table 1 for each journal in each field. We did not find convincing evidence for a systematic underrepresentation of certain journals (see Appendix C)

Conclusion

The analysis of time-to-event data with Cox regression Cox, 1972 is pervasive in biomedical research. Cox regression combined with Bayes factors has much to offer over traditional frequentist inference because it allows researchers to directly contrast the evidence for the null and alternative hypotheses Rouder et al., 2009 and because it allows monitoring results during data collection and continue or stop at any time Armitage et al., 1969; Rouder, 2014; Schönbrodt and Wagenmakers, 2018; Schönbrodt et al., 2017; Stefan, Schönbrodt, et al., 2022; Tendeiro et al., 2022. These characteristics of Bayes factors have the potential to reduce the waste of scarce resources in biomedical research and especially clinical trials Macleod et al., 2014; van Ravenzwaaij et al., 2019. However, the specification of priors for these Bayesian analyses can be challenging and be perceived as overly subjective. We propose default priors in Cox regression that are informed by large corpora of already existing studies for nine subfields. These priors are all Normal distributions centered on 0 with standard deviations that are close to 1. They can be used as a default or starting point for medical researchers and can be augmented with sensitivity analyses.

References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, *132*(2), 235–244. https://doi.org/10.2307/2343787

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Bartoš, F. (2022). RoBSA: An R package for robust Bayesian survival analyses [R package version 1.0.0]. https://CRAN.R-project.org/package=RoBSA

Bartoš, F., Aust, F., & Haaf, J. M. (2022). Informed Bayesian survival analysis. *BMC Medical Research Methodology*, *22*(1), 238. https://doi.org/10.1186/s12874-022-01676-9

Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, *40*(3), 1550–1577. https://doi.org/10.1214/12-AOS1013

Beigel, J. H., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., Hohmann, E., Chu, H. Y., Luetkemeyer, A., Kline, S., Lopez de Castilla, D., Finberg, R. W., Dierberg, K., Tapson, V., Hsieh, L., Patterson, T. F., Paredes, R., Sweeney, D. A., Short, W. R., ... Lane, C. (2020). Remdesivir for the treatment of Covid-19 - final report. *The New England Journal of Medicine*, *383*(19), 1813–1826. https://doi.org/10.1056/NEJMoa2007764

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402. https://doi.org/10.1214/06-BA115

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., Dasgupta, A., Gustafson, P., Wasserman, L., Kadane, J. B., Srinivasan, C., Lavine, M., O'Hagan, A., Polasek, W., Robert, C. P., ... Sivaganesan, S. (1994). An overview of robust Bayesian analysis. *Test*, *3*(1), 5–124. https://doi.org/10.1007/BF02562676

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, *5*(1), 27–36. https://doi.org/10.1038/nrd1927

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo.

Boney, C. M., Verma, A., Tucker, R., & Vohr, B. R. (2005). Metabolic syndrome in childhood: Association with birth weight, maternal obesity, and gestational diabetes mellitus. *Pediatrics*, *115*(3), 290–296. https://doi.org/10.1542/peds.2004-1808

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd). John Wiley & Sons.

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003a). Survival analysis part II: Multivariate data analysis - an introduction to concepts and methods. *British Journal of Cancer*, *89*(3), 431–436. https://doi.org/10.1038/sj.bjc.6601119

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003b). Survival analysis part III: Multivariate data analysis - choosing a model and assessing its adequacy and fit. *British Journal of Cancer*, *89*(4), 605–611. https://doi.org/10.1038/sj.bjc.6601120

Brard, C., Le Teuff, G., Le Deley, M.-C., & Hampson, L. V. (2017). Bayesian survival analaysis in clinical trials: What methods are used in practice? *Clinical Trials*, *14*(1), 78–87. https://doi.org/10.1177/1740774516673362

Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*(2), 83–87. https://doi.org/10.1080/00031305.1985.10479400

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of reporting p values in the biomedical literature, 1990-2015. *Journal of the American Medical Association*, *315*(11), 1141–1148. https://doi.org/10.1001/jama.2016.1952

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003a). Survival analysis part I: Basic concepts and first analyses. *British Journal of Cancer*, *89*(2), 232–238. https://doi.org/10.1038/sj.bjc.6601118

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003b). Survival analysis part IV: Further concepts and methods in survival analysis. *British Journal of Cancer*, *89*(5), 781–786. https://doi.org/10.1038/sj.bjc.6601117

Collett, D. (2015). *Modelling survival data in medical research* (3rd). CRC Press.

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, *13*(2), 627–679. https://doi.org/10.1214/18-BA1103

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*, *117*(6), 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065

Diener, H.-C., Bogousslavsky, J., Brass, L. M., Cimminiello, C., Csiba, L., Kaste, M., Leys, D., Matias-Guiu, J., & Rupprecht, H.-J. (2004). Aspirin and clopidogrel compared with clopidogrel alone after recent ischaemic stroke or transient ischaemic attack in high-risk patients (MATCH): Randomised, double-blind, placebo-controlled trial. *The Lancet*, *364*(9431), 331–337. https://doi.org/10.1016/S0140-6736(04)16721-4

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Du, H., Edwards, M. C., & Zhang, Z. (2019). Bayes factor in one-sample tests of means with a sensitivity analysis: A discussion of separate prior distributions. *Behavior Research Methods*, *51*(5), 1998–2021. https://doi.org/10.3758/s13428-019-01262-w

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, *40*(1), 1–5. https://doi.org/10.2307/2683105

Friedl, J. E. F. (2006). *Mastering regular expressions* (4th). O'Reilly Media.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453. https://doi.org/10.1037/a0015251

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*(3), 403–420. https://doi.org/10.1214/06-BA116

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, *130*(12), 995–1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, *130*(12), 1005–1013. https://doi.org/10.7326/0003-4819-130-12-199906150-00019

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

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C. J. (2021). *bain: Bayes factors for informative hypotheses* [R package version 0.2.8]. https://CRAN.R-project.org/package=bain

Guo, B., Park, Y., & Liu, S. (2019). A utility-based Bayesian phase I-II design for immunotherapy trials with progression-free survival end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(2), 411–425. https://doi.org/10.1111/rssc.12288

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (2nd). Springer.

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. L., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., … Hoijtink, H. (2020). A review of applications of the Bayes factor in psychological research. https://doi.org/10.31234/osf.io/cu43g

Higgins, J. P. T., Li, T., & Deeks, J. J. (2019). Choosing effect measures and computing estimates of effect. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, J. Page Matthew, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* (2nd, pp. 143–176). John Wiley & Sons.

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. *PLoS ONE*, *13*(4), e0195474. https://doi.org/10.1371/journal.pone.0195474

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time-to-event data* (2nd). John Wiley & Sons.

International Committee of Medical Journal Editors. (1997). Uniform requirements for manuscripts submitted to biomedical journals. *Pathology*, *29*, 441–447. https://doi.org/10.1080/00313029700169515

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

JASP Team. (2023). JASP (Version 0.17.2)[Computer software]. https://jasp-stats.org/

Jeffreys, H. (1939). *Theory of probability*. The Clarendon Press.

Jeffreys, H. (1948). *Theory of probability* (2nd). The Clarendon Press.

Jeffreys, H. (1961). *Theory of probability* (3rd). Oxford University Press.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., & Feldman, B. M. (2010). Methods to elicit beliefs for Bayesian priors: A systematic review. *Journal of Clinical Epidemiology*, *63*, 355–369. https://doi.org/10.1016/j.jclinepi.2009.06.003

Kantoff, P. W., Higano, C. S., Shore, N. D., Berger, E. R., Small, E. J., Penson, D. F., Redfern, C. H., Ferrari, A. C., Dreicer, R., Sims, R. B., Xu, Y., Frohlich, M. W., & Schellhammer, P. F. (2010). Sipuleucel-t immunotherapy for castration-resistant prostate cancer. *New England Journal of Medicine*, *363*(5), 411–422. https://doi.org/10.1056/NEJMoa1001294

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.2307/2291091

Kenchaiah, S., Evans, J. C., Levy, D., Wilson, P. W. F., Benjamin, E. J., Larson, M. G., Kannel, W. B., & Vasan, R. S. (2002). Obesity and the risk of heart failure. *New England Journal of Medicine*, *347*(5), 305–313. https://doi.org/10.1056/NEJMoa020245

Klein, J. P., & Moeschberger, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*. Springer.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676. https://doi.org/10.1002/wcs.72

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd). Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review*, *25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Linde, M., Tendeiro, J. N., & van Ravenzwaaij, D. (2022). Bayes factors for two-group comparisons in Cox regression. https://doi.org/10.1101/2022.11.02.22281762

Linde, M., van Ravenzwaaij, D., & Tendeiro, J. N. (2022). *baymedr: Computation of Bayes factors for common biomedical designs* [R package version 0.1.1.9000]. https://github.com/maxlinde/baymedr

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. https://doi.org/10.1016/j.jmp.2008.03.002

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A., Salman, R. A.-S., Chan, A.-W., & Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *The Lancet*, *383*(9912), 101–104. https://doi.org/10.1016/S0140-6736(13)62329-6

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs* [R package version 0.9.12-4.2]. https://CRAN.R-project.org/package=BayesFactor

Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, *52*, 1–4. https://doi.org/10.1016/j.envsoft.2013.10.010

Moyé, L. A. (2008). Bayesians in clinical trials: Asleep at the switch. *Statistics in Medicine*, *27*(4), 469–482. https://doi.org/10.1002/sim.2928

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199. https://doi.org/10.1093/biomet/64.2.191

Rennie, D. (1978). Vive la différence (p<0.05). *New England Journal of Medicine*, *299*, 828–829. https://doi.org/10.1056/NEJM197810122991509

Rietbergen, C., Klugkist, I., Janssen, K. J. M., Moons, K. G. M., & Hoijtink, H. J. A. (2011). Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemporary Clinical Trials*, *32*(6), 848–855. https://doi.org/10.1016/j.cct.2011.06.002

Rinke, A., Müller, H.-H., Schade-Brittinger, C., Klose, K.-J., Barth, P., Wied, M., Mayer, C., Aminossadati, B., Pape, U.-F., Bläker, M., Harder, J., Arnold, C., Gress, T., & Arnold, R. (2009). Placebo-controlled, double-blind, prospective, randomized study on the effect of octreotide LAR in the control of tumor growth in patients with metastatic neuroendocrine midgut tumors: A report from the PROMID study group. *Journal of Clinical Oncology*, *27*(28), 4656–4663. https://doi.org/10.1200/JCO.2009.22.8510

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(2), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

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

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

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

SCImago. (n.d.). *SJR - SCImago journal & country rank [portal]*. Retrieved March 3, 2023, from http://www.scimagojr.com

Singh, R., & Mukhopadhyay, K. (2011). Survival analysis in clinical trials: Basics and must know areas. *Perspectives in Clinical Research*, *2*(4), 145–148. https://doi.org/10.4103/2229-3485.86872

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, *56*(3), 196–201. https://doi.org/10.1198/000313002137

Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (2022). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*, *27*(2), 177–197. https://doi.org/10.1037/met0000354

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

Stupp, R., Hegi, M. E., Mason, W. P., van den Bent, M. J., Taphoorn, M. J. B., Janzer, R. C., Ludwin, S. K., Allgeier, A., Fisher, B., Belanger, K., Hau, P., Brandes, A. A., Gijtenbeek, J., Marosi, C., Vecht, C. J., Mokhtari, K.,

Wesseling, P., Villa, S., Eisenhauer, E., … Mirimanoff, R.-O. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *The Lancet Oncology*, *10*(5), 459–466. https://doi.org/10.1016/S1470-2045(09)70025-7

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

Tendeiro, J. N., Kiers, H. A. L., & Van Ravenzwaaij, D. (2022). Worked-out examples of the adequacy of Bayesian optional stopping. *Psychonomic Bulletin & Review*, *29*(1), 70–87. https://doi.org/10.3758/s13423-021-01962-5

Thall, P. F., & Cook, J. D. (2004). Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*, *60*(3), 684–693. https://doi.org/10.1111/j.0006-341X.2004.00218.x

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. Springer.

van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S. D., Olff, M., & van Loey, N. E. (2018). Bayesian PTSD-trajectory analysis with informed priors based on a systematic literature search and expert elicitation. *Multivariate Behavioral Research*, *53*(2), 267–291. https://doi.org/10.1080/00273171.2017.1412293

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. https://doi.org/10.1037/met0000100.supp

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penelized regression. *Journal of Mathematical Psychology*, *89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

van Lissa, C. J., Gu, X., Mulder, J., Rosseel, Y., van Zundert, C., & Hoijtink, H. (2021). Teacher's corner: Evaluating informative hypotheses using the Bayes factor in

structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 292–301. https://doi.org/10.1080/10705511.2020.1745644

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154. https://doi.org/10.3758/s13423-016-1015-8

van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*(1), 71. https://doi.org/10.1186/s12874-019-0699-7

van Zwet, E., & Gelman, A. (2022). A proposal for informative default priors scaled by the standard error of estimates. *The American Statistician*, *76*(1), 1–9. https://doi.org/10.1080/00031305.2021.1938225

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. https://doi.org/10.1016/j.jmp.2010.07.003

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin and Review*, *21*, 268–282. https://doi.org/10.3758/s13423-013-0495-z

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & van de Schoot, R. (2017). Where do priors come from? applying guidelines to construct informative priors in small sample research. *Research in Human Development*, *14*(4), 305–320. https://doi.org/10.1080/15427609.2017.1370966

Appendix A

Comparison of Abstract and Main Text Matches

To mitigate the possibility that the effect size distributions are biased, we took our database of articles found on Scopus that matched any of the journals listed in Scimago, yielding $36,431$ articles (see Table 1 in the manuscript). Of those articles, we randomly sampled 100 articles for which we could not match results in the abstract with our regular expression (no-match group); and we randomly sampled 100 articles for which we could match results in the abstract with our regular expression (match group). Then, we tried to obtain the full texts of these 200 articles. We were able to obtain 86 full texts for the no-match group and 88 full texts for the match group (some articles were not accessible to us). We transformed the PDF files into txt files using the pdftotext command from the Xpdfreader tool https://www.xpdfreader.com/. Then we manually removed the abstracts from the txt files. Subsequently, we used the same regular expression to match results in the main text of the txt files. Fig A1 shows a plot that compares the effect size distributions for extracted results from the abstracts and the main text. As can be seen, the effect size distributions look similar across the different panels, providing evidence that there is probably no strong bias.

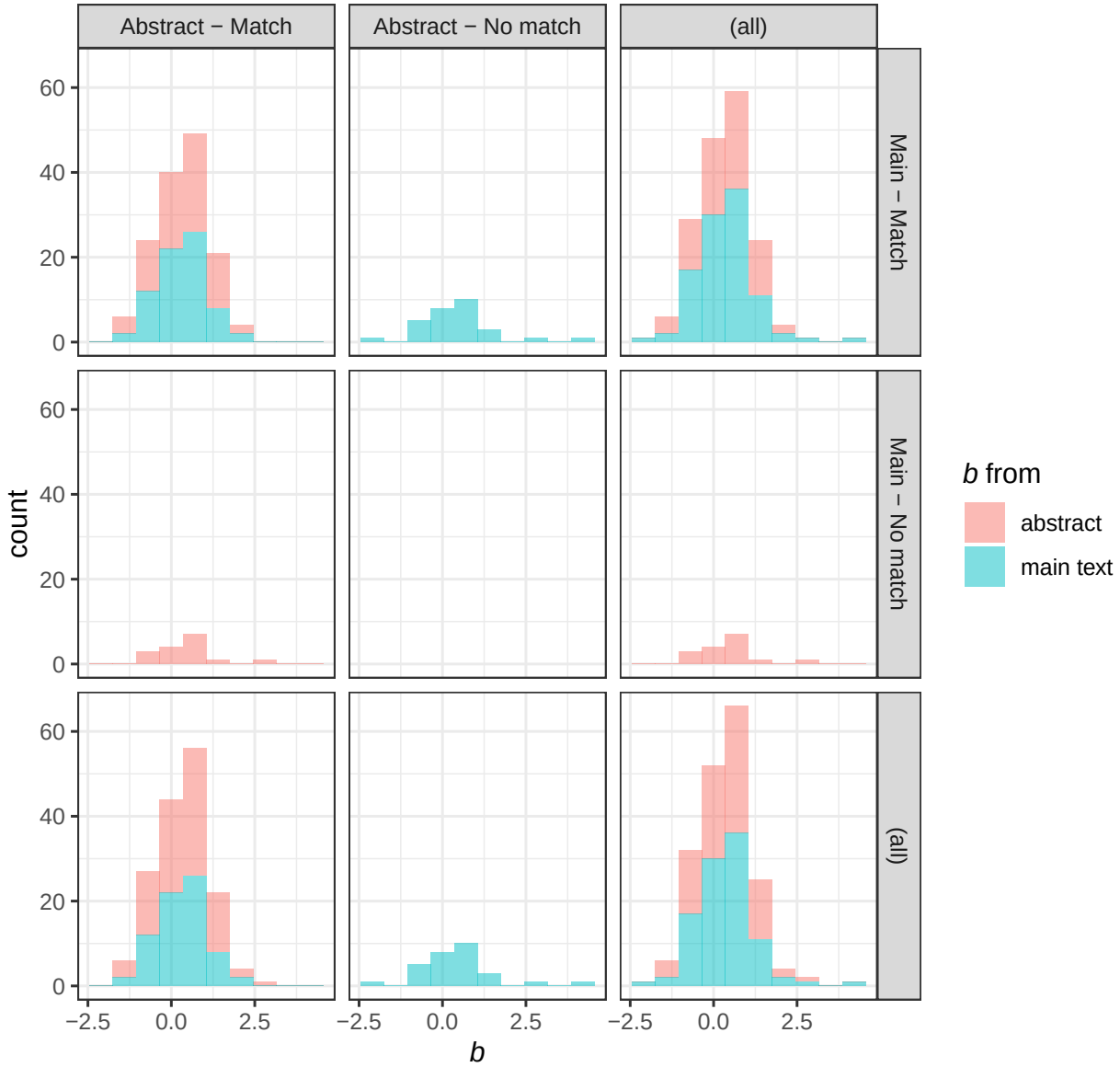

Figure A1

Effect size distributions for effect sizes matched in the abstract or main text of a sample of 200 articles.

Appendix B

Descriptive Statistics for medical fields

Table B1

Descriptive statistics for Anesthesiology and pain medicine.

	b	$SE(b)$	b^2	$SE(b)^2$
Min	-2.120	0.002	0.000	0.000
1st Qu.	-0.010	0.069	0.043	0.005
Median	0.336	0.164	0.261	0.027
Mean	0.388	0.219	0.804	0.089
3rd Qu.	0.782	0.308	0.781	0.095
Max	2.896	1.211	8.386	1.468

Table B2

Descriptive statistics for Cardiology and cardiovascular medicine.

	b	$SE(b)$	b^2	$SE(b)^2$
Min	-4.247	0.000	0.000	0.000
1st Qu.	-0.041	0.086	0.061	0.007
Median	0.365	0.184	0.278	0.034
Mean	0.407	0.238	0.824	0.099
3rd Qu.	0.854	0.332	0.935	0.110
Max	4.800	1.883	23.039	3.547

Table B3

Descriptive statistics for Gastroenterology.

	b	$SE(b)$	b^2	$SE(b)^2$
Min	-3.507	0.000	0.000	0.000
1st Qu.	-0.174	0.110	0.090	0.012
Median	0.389	0.219	0.426	0.048
Mean	0.393	0.285	1.027	0.137
3rd Qu.	0.968	0.389	1.264	0.152
Max	3.459	1.274	12.296	1.623

Table B4

Descriptive statistics for Hematology.

	b	$SE(b)$	b^2	$SE(b)^2$
Min	-3.219	0.001	0.000	0.000
1st Qu.	-0.174	0.119	0.083	0.014
Median	0.378	0.225	0.380	0.051
Mean	0.398	0.275	1.029	0.123
3rd Qu.	0.916	0.377	1.148	0.142
Max	4.806	1.332	23.102	1.773

Table B5

Descriptive statistics for Immunology and allergy.

	b	$SE(b)$	b^2	$SE(b)^2$
Min	-2.996	0.001	0.000	0.000
1st Qu.	-0.128	0.105	0.069	0.011
Median	0.351	0.207	0.345	0.043
Mean	0.361	0.259	0.937	0.110
3rd Qu.	0.880	0.365	1.156	0.133
Max	4.248	1.145	18.050	1.311

Table B6

Descriptive statistics for Neurology.

	b	$SE(b)$	b^2	$SE(b)^2$
Min	-4.605	0.001	0.000	0.000
1st Qu.	0.010	0.101	0.090	0.010
Median	0.425	0.202	0.339	0.041
Mean	0.446	0.255	0.978	0.115
3rd Qu.	0.916	0.348	1.060	0.121
Max	4.726	1.799	22.331	3.236

Table B7

Descriptive statistics for Oncology.

	b	$SE(b)$	b^2	$SE(b)^2$
Min	-5.809	0.001	0.000	0.000
1st Qu.	-0.288	0.106	0.081	0.011
Median	0.285	0.206	0.316	0.042
Mean	0.263	0.252	0.830	0.107
3rd Qu.	0.765	0.338	0.856	0.115
Max	5.955	2.784	35.466	7.749

Table B8

Descriptive statistics for Psychiatry and mental health.

	b	$SE(b)$	b^2	$SE(b)^2$
Min	-2.659	0.001	0.000	0.000
1st Qu.	0.068	0.069	0.062	0.005
Median	0.392	0.151	0.261	0.023
Mean	0.454	0.194	0.768	0.070
3rd Qu.	0.815	0.259	0.840	0.067
Max	3.886	1.428	15.099	2.040

Table B9

Descriptive statistics for Pulmonary and respiratory medicine.

	b	$SE(b)$	b^2	$SE(b)^2$
Min	-3.507	0.000	0.000	0.000
1st Qu.	-0.136	0.093	0.068	0.009
Median	0.351	0.203	0.336	0.041
Mean	0.378	0.255	0.886	0.112
3rd Qu.	0.884	0.357	1.012	0.127
Max	3.728	1.499	13.899	2.246

Appendix C

Article retention rate across processing steps

We further investigated the possibility that certain journals are underrepresented in our results by examining the article retention rates over processing steps across fields and journals. In particular, looking at Table 1, it can be seen that there are three processing steps involved: First, Scopus articles are selected based on whether they match the Scimago journal names. Second, articles are selected for which we could obtain a match for the regular expression. Third, of the matched articles, we filtered out articles for which our regular expression yielded ambiguous/wrong results. Fig C1 shows the proportion of articles remaining in a specific field (panels) and journal (lines) when transitioning from the first to the second step (S1) and from the second to the third step (S2). However, journals for which the number of articles is 0 at some processing step are filtered out because the division with 0 is problematic. It can be seen that there is some variation between journals at S1 for all fields. This variation decreases drastically at S2 for all fields.

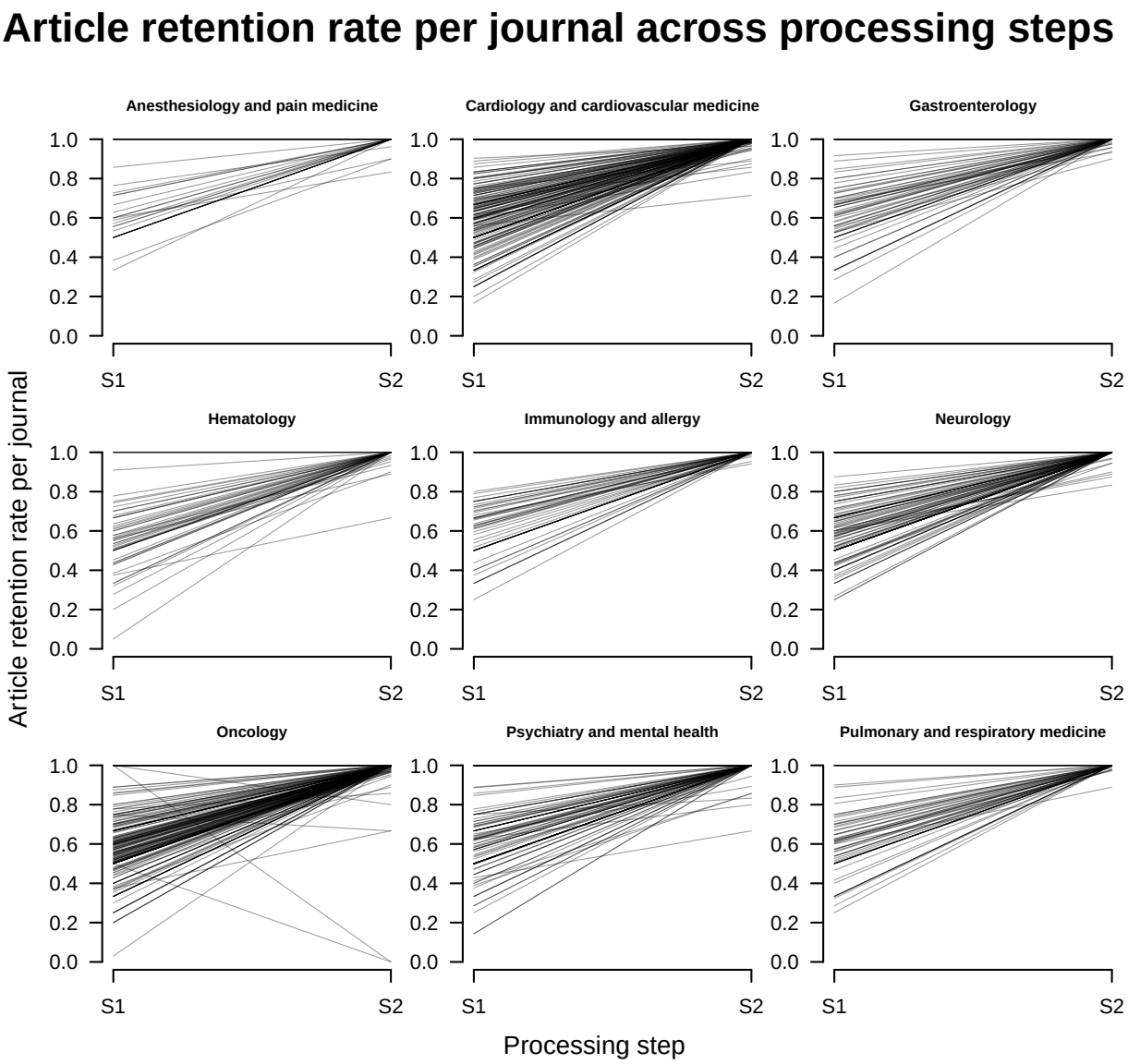

Figure C1

Proportion of articles remaining in a specific field (panels) and journal (lines) when transitioning from the first to the second step (S1) and from the second to the third step (S2). Note that journals for which the number of articles is 0 **at some processing step were not considered here.**