Generalized Jeffreys's approximate objective Bayes factor: Model-selection consistency, finite-sample accuracy and statistical evidence in 71,126 clinical trial findings

Puneet Velidi[a,1], **Zhengxiao Wei**[a,1], **Shreena Nisha Kalaria**[a,b,c], **Yimeng Liu**[a], **Céline M. Laumont**[c,d], **Brad H. Nelson**[b,c,d], **and Farouk S. Nathoo**[a,2]

This manuscript was compiled on September 21, 2025

Concerns about the misuse and misinterpretation of p-values and statistical significance have motivated alternatives for quantifying evidence. We develop a generalized form of Jeffreys's approximate objective Bayes factor ($eJAB$), a one-line calculation that is a function of the p-value, sample size, and parameter dimension. We establish conditions under which $eJAB$ is model-selection consistent and verify them via simulations for 11 common statistical tests. We assess finite-sample accuracy by comparing $eJAB$ with Markov chain Monte Carlo computed Bayes factors in 12 simulation studies. We then apply $eJAB$ to 71,126 results from ClinicalTrials.gov (CTG) and find that the proportion of findings with p-value $\leqslant \alpha$ yet $eJAB_{01} > 1$ closely tracks the significance level α, suggesting that such contradictions are pointing to the type I errors. We catalog 4,088 such candidate type I errors and provide details for 131 with reported p-value $\leqslant .01$. We also identify 487 instances of the Jeffreys–Lindley paradox. Finally, we estimate that 75% (6%) of clinical trial plans from CTG set $\alpha \geqslant .05$ as the target evidence threshold, and that 35.6% (0.22%) of results significant at $\alpha = .05$ correspond to evidence that is no stronger than anecdotal under $eJAB$.

Jeffreys's approximate objective Bayes factor — null-hypothesis significance testing — p-values — replicability — type I error

Significance Statement

Bayes factors quantify statistical evidence but are often computationally intensive, and simple approximations can be inaccurate. We propose $eJAB$, a generalized Jeffreys's approximate objective Bayes factor that converts the p-value, sample size, and parameter dimension into an evidence measure via a one-line calculation. We establish regularity conditions under which $eJAB$ is model-selection consistent. Applied to 71,126 clinical trial results, $eJAB$ flags 4,088 candidate type I errors. Our studies demonstrate a scalable and transparent method for screening the credibility of reported effects.

Concerns about the replicability of scientific research - stemming in part from the overreliance on null-hypothesis significance testing at traditional p-value thresholds (1) - have prompted calls for more stringent evidence criteria (2, 3), more judicious interpretation (4), and alternative measures of evidence (5–8). The Bayes factor—the ratio of marginal likelihoods under competing hypotheses—can provide evidence for either the null or the alternative (9–14). It represents how data update the prior model odds to form the posterior model odds. The evidential behavior of both p-values and Bayes factors has been analyzed by stochastic ordering in sample size and effect size, with implications for the Jeffreys–Lindley paradox (15). Relatedly, emphasizing effect sizes rather than p-value significance thresholds, a quadratic exponential relationship between the Bayes factor and the separation of credible intervals is described in (16). However, computation remains a practical hurdle: Bayes factors require numerical integration over prior distributions, often via Markov chain Monte Carlo (MCMC) with model-specific tuning (17–19). While these computational challenges are not insurmountable, a simple Bayes factor calculation requiring only the p-value, sample size, and the dimension of the parameter under test has greater applicability to situations where the entire dataset is not available.

Several integral-free methods have been proposed to reduce computational complexity. For example, certain Bayes factors use a point estimate of the parameter under test in conjunction with informed priors (20, 21). The Bayesian information criterion (BIC) is related to Bayes factors for model comparison (22). The minimum Bayes factor provides an upper bound on evidence against the null that can be derived from a test statistic or a p-value (23). Bayes factors based on parametric test statistics are developed in (24), with model-selection consistency under the prior predictive and connections to BIC examined in (25); extensions to nonparametric statistics appear in (26). A proposed $3p\sqrt{n}$ rule, based on a piecewise approximation to the χ^2_1 quantile function is connected to Jeffreys's original work in the scalar-parameter case (27–29). Here, we define a generalization ($eJAB$) that can be applied for testing multidimensional parameters and that includes a finite sample correction

Author affiliations: [a]Department of Mathematics and Statistics, University of Victoria; [b]Department of Biochemistry and Microbiology, University of Victoria; [c]Deeley Research Centre, British Columbia Cancer Agency; [d]Department of Medical Genetics, University of British Columbia

S.N.K., C.M.L., and F.S.N. conceptualized research; C.M.L., B.H.N, and F.S.N. supervised research; Z.W. and F.S.N. contributed methodology; P.V., Z.W., S.N.K., Y.L., C.M.L., and F.S.N. collected data; P.V., S.N.K, and F.S.N. analyzed data; Z.W. and F.S.N. conducted simulations; P.V. and Y.L. built software; Z.W. maintained repository; F.S.N. managed project; P.V., Z.W., and F.S.N. wrote the initial draft; S.N.K., Y.L., B.H.N., C.M.L. reviewed and provided comments on the original draft

The authors declare no competing interest.

[1] P.V. and Z.W. contributed equally to this work.

[2] To whom correspondence may be addressed. E-mail: nathoo@uvic.ca

factor. We present a theoretical justification, and examine its finite sample performance as a Bayes factor approximation in 12 simulation studies examining 11 common statistical tests covering both parametric and nonparametric settings. These results indicate that *eJAB* is a generally applicable, reliable approach to approximate the Bayes factor.

Our work differs from (22) in its consideration of p-values beyond those from the likelihood ratio test and in our explicit statement of conditions for model selection consistency. For the latter, (22) demonstrates asymptotic equivalence to the BIC so that model selection consistency arises under the conditions specified in (30). It differs from (20, 21, 27–29) primarily in our development of a theoretical justification.

Our contributions are thus threefold. First, we develop an extension of JAB, *eJAB*, which generalizes to multidimensional parameters by incorporating the parameter dimension. Our approach differs from BIC-based approximations by accommodating p-values beyond those from likelihood-ratio tests, and from other methods by establishing a model-selection consistency justification under regularity conditions. Second, we assess the finite-sample performance of *eJAB* across 11 common parametric and nonparametric tests. The results suggest that *eJAB* is a broadly applicable and reliable approach to approximate the Bayes factor. Third, we apply *eJAB* to 71,126 reported findings from https://ClinicalTrials.gov (CTG), downloaded on August 7th, 2025, flagging 4,088 candidate type I errors. We find that the distribution of statistical evidence is uniform across five clinical trial phases for both primary and secondary outcomes in CTG and identify 487 instances of the Jeffreys–Lindley paradox.

1. Methodology

Definition 1.1. To test the hypotheses $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, we define *eJAB* as

$$eJAB_{01} = \sqrt{n}\exp\left\{-\frac{1}{2}\frac{n^{1/q}-1}{n^{1/q}}\,Q_{\chi^2_q}(1-p)\right\},$$

where q is the dimension of the parameter vector $\boldsymbol{\theta}$, n is the sample size, $Q_{\chi^2_q}(\cdot)$ is the quantile function of the chi-squared distribution with q degrees of freedom (df), and p is the p-value from a null-hypothesis significance test. The coherence condition specifies that $eJAB_{10} = 1\,/\,eJAB_{01}$.

Lemma 1.2. For $x \to 1^-$,

$$f(x) = -2\ln(1-x) - 2\ln(-\ln(1-x)) < Q_{\chi^2_q}(x).$$

The proof is provided on the Open Science Framework at https://osf.io/qeprj/ ('Files', 'Theorem').

Theorem 1.3. Assuming the p-value from a null hypothesis significance test satisfies two regularity conditions,

- **R1**: $p \xrightarrow{D} \mathrm{Unif}(0,1)$ as $n \to \infty$ under $\mathcal{H}_0$,
- **R2**: $D_n = -\sqrt{n}\cdot p\ln p \xrightarrow{P} 0$ as $n \to \infty$ under $\mathcal{H}_1$,

$$\operatorname*{plim}_{n\to\infty} eJAB_{01} = \begin{cases} \infty & \text{under } \mathcal{H}_0 \\ 0 & \text{under } \mathcal{H}_1 \end{cases}$$

$eJAB_{01}$ is model-selection consistent.

Proof: Assume $\mathcal{H}_0$ is true. Let $\delta \in (0,1)$ and $K > 0$ be arbitrary. Note that **R1** and $W_n = Q_{\chi^2_q}(1-p) \xrightarrow{D} \chi^2_q$ by the Continuous Mapping Theorem (CTM). For $n \geqslant N^* = \lceil K^2\exp\{Q_{\chi^2_q}(1-\delta)\}\rceil$, we have

$$\begin{aligned}
&\mathbb{P}(eJAB_{01} > K) = \mathbb{P}\left(\sqrt{n}\exp\left\{-\frac{1}{2}\frac{n^{1/q}-1}{n^{1/q}}\,W_n\right\} > K\right)\\
>&\mathbb{P}\left(\sqrt{n}\exp\left\{-W_n\,/\,2\right\} > K\right)\\
=&\mathbb{P}\left(W_n < -2\ln\frac{K}{\sqrt{n}}\right) \geqslant \mathbb{P}\left(W_n < -2\ln\frac{K}{\sqrt{N^*}}\right)\\
\geqslant&\mathbb{P}\left(W_n < -2\ln\frac{K}{\sqrt{K^2\exp\{Q_{\chi^2_q}(1-\delta)\}}}\right)\\
=&\mathbb{P}\left(W_n < Q_{\chi^2_q}(1-\delta)\right) = 1-\delta.
\end{aligned}$$

Thus, $eJAB_{01} \xrightarrow{P} \infty$ as $n \to \infty$ under $\mathcal{H}_0$.

Assume $\mathcal{H}_1$ is true. Let $\delta \in (0,1)$ and $\epsilon > 0$ be arbitrary. Lemma 1.2 gives $f(1-p) < Q_{\chi^2_q}(1-p)$ for $p \to 0$.

$$\begin{aligned}
&\mathbb{P}(eJAB_{01} < \epsilon)\\
=&\mathbb{P}\left(\sqrt{n}\exp\left\{-\frac{1}{2}\frac{n^{1/q}-1}{n^{1/q}}\,Q_{\chi^2_q}(1-p)\right\} < \epsilon\right)\\
>&\mathbb{P}\left(\sqrt{n}\exp\left\{-\frac{1}{2}\frac{n^{1/q}-1}{n^{1/q}}(-2\ln p - 2\ln(-\ln p))\right\} < \epsilon\right)\\
=&\mathbb{P}\left(\sqrt{n}\ \ (p\cdot(-\ln p))^{\frac{n^{1/q}-1}{n^{1/q}}} < \epsilon\right) > 1-\delta.
\end{aligned}$$

The final inequality is true for n sufficiently large by **R2** and the continuous mapping theorem with $1 - \frac{1}{n^{1/q}} \to 1$.

Thus, $eJAB_{01} \xrightarrow{P} 0$ as $n \to \infty$ under $\mathcal{H}_1$. ■

Remark 1.4. We can derive *eJAB* using a generalization of the multivariate normal unit-information prior for the parameters under test,

$$\boldsymbol{\theta} \sim \boldsymbol{\mathcal{N}}_q\left(\hat{\boldsymbol{\theta}},\ n^{1/q}\cdot\mathbf{J}_n^{-1}(\hat{\boldsymbol{\theta}})\right),$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (MLE), and $\mathbf{J}_n(\hat{\boldsymbol{\theta}})$ is the observed information matrix of size $q \times q$.

Under an asymptotic Gaussian approximation, the posterior density takes the form, $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \mathcal{H}_1) \approx$

$$(2\pi)^{-\frac{q}{2}}\cdot|\mathbf{J}_n(\hat{\boldsymbol{\theta}})|^{\frac{1}{2}}\cdot\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\cdot\mathbf{J}_n(\hat{\boldsymbol{\theta}})\cdot(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\}.$$

Assuming that the nuisance-parameter priors under the nested competing models satisfy the usual conditioning constraint, the Bayes factor (BF) can be expressed as the Savage–Dickey density ratio (31),

$$BF_{01} = \frac{\pi(\boldsymbol{\theta}_0 \mid \boldsymbol{y}, \mathcal{H}_1)}{\pi(\boldsymbol{\theta}_0 \mid \mathcal{H}_1)} \approx \sqrt{n}\exp\left\{-\frac{1}{2}\frac{n^{1/q}-1}{n^{1/q}}W\right\},$$

where $\boldsymbol{y}$ represents the data, and W is the Wald statistic. When the p-value is obtained from a Wald test, it yields $eJAB_{01}$ in Definition 1.1 by noting that $W = Q_{\chi^2_q}(1-p)$.

The prior underlying the approximate Bayes factor interpretation is centered around the MLE $\hat{\theta}$ with $|Var[\theta]| =$

$n|\mathbf{J_n^{-1}}(\hat{\theta})|$. When $q = 1$ this is the unit information prior, more generally the volume of prior probability ellipsoids around the MLE is $O(n^{(1-q)/2})$. The prior distribution is asymptotically diffuse relative to the likelihood.

Remark 1.5. When the p-value is obtained from a likelihood-ratio test and $q = 1$, *eJAB* reduces to the evidential BIC method of (22). This approximation is derived by applying a finite-sample correction to the BIC approximation to the Bayes factor, which involves raising the likelihood ratio to the power of $1 - \frac{1}{n}$. Model-selection consistency for the evidential BIC follows from its asymptotic equivalence to BIC together with standard consistency results for BIC (22, 30). By contrast, the conditions of Theorem 1.3 are broadly applicable: they cover nonparametric tests and allow certain forms of model misspecification, provided **R1** and **R2** hold.

2. Simulation Studies

We conduct simulations across 11 parametric and nonparametric tests. For each design, 1,000 data sets are generated under the null hypothesis and the alternative hypothesis with small, medium, and large effect sizes across various sample sizes. The designs include:

1. Two-sample t-test;
2. Simple linear regression;
3. Simple logistic regression;
4. One-way analysis of variance (ANOVA);
5. One-way repeated-measures ANOVA (rANOVA) with high and low intraclass correlations $\rho \in \{.9, .2\}$;
6. Chi-squared tests for independence under multinomial (total count fixed) and product-multinomial (row counts fixed) designs;
7. Cox proportional-hazards regression with right censoring;
8. Conditional logistic regression;
9. Wilcoxon signed-rank test;
10. Mann–Whitney U-test;
11. Kruskal–Wallis H-test.

For nonparametric tests (items 9 through 11), we simulated using right-skewed and heavy-tailed (non-Gaussian) distributions. Detailed simulation settings, additional designs (e.g., two-way ANOVA and large contingency tables), and the `R` code for the subsequent subsections are available on the Open Science Framework at http://osf.io/qeprj/, under 'Files' within the directories 'Theorem', 'Simulations' (parametric simulations), and 'Supplementary Information' (nonparametric simulations).

We used p-values from the t-test for items 1 and 2, the F-test for items 4 and 5, the Wald test for items 3, 7, and 8, and the asymptotic null distribution for item 6. The effective sample size n was taken as the number of uncensored or matched observations for survival analysis, and the number of independent observations (i.e., the product of the number of subjects and one less than the number of conditions) for rANOVA (32, 33). For the other designs, n always refers to the total number of observations or counts. In particular, we tested the harmonic mean of group sizes for item 1, as in JAB (29), but do not demonstrate it here. Among the tests with multiple *df* ($q > 1$) are items 4 through 6, and 11.

A. Verifying the Regularity Conditions. Condition **R1** of Theorem 1.3 is satisfied whenever the p-value is computed from an exact or asymptotic null distribution with a continuous cumulative distribution function.

Next, we computed D_n of R2 using the p-values and sample sizes from the simulated data, generated with medium effect sizes, across the 11 designs. In this subsection, items 4 and 5 ($\rho = .9$) each have $4-1 = 3$ *df*; item 6 has $(3-1)\times(3-1) = 4$ *df* under a multinomial design; item 7 represents a semi-parametric Cox model; item 11 has $5 - 1 = 4$ *df*. Due to randomness, the x-axis has to display the total number of observations or pairs, rather than the number of uncensored observations for item 7, or the number of matched pairs for item 8, as required in D_n and *eJAB*. Condition **R2**, which shrinks to zero as the sample size increases, appears to hold in all cases examined (Fig. 1).

B. Accuracy of the Bayes-Factor Approximation. To assess the finite-sample accuracy of an approximation to the Bayes factor, we compared *eJAB* with alternative Bayes-factor computations across items 1 through 7, fixing representative sample sizes. In this subsection, items 4 and 5 each have $3 - 1 = 2$ *df*; item 6 maintains 4 *df*; item 7 now represents a parametric survival regression model with an exponential baseline and expected censoring fractions of 23% under the null, and 21%, 19%, and 18% under increasing effect sizes.

Except for the two chi-squared tests, the MCMC-based Bayes factors were computed by sampling the marginal posterior density using '*rstan*' (34), applying logspline density estimation (35), and calculating the Savage–Dickey density ratio. In these cases, the model underlying the MCMC sampler assumed the same prior as *eJAB*, so discrepancies reflect errors of approximation rather than differences in prior. For the chi-squared tests, we compared *eJAB* with (i) test-statistic Bayes factors (24) and (ii) Dirichlet Bayes factors from `contingencyTableBF` in the '*BayesFactor*' `R` package (18, 36).

Pairs of ($\ln eJAB_{10}$, $\ln BF_{10}$) are color- and shape-coded by the p-value. Overall, *eJAB* tracks the MCMC-based Bayes factors closely (Fig. 2). Some inaccuracy, $\ln eJAB_{10} < \ln BF_{10}$, appears for (binary) logistic regression when evidence favors $\mathcal{H}_1$. For chi-squared tests, agreement with test-statistic Bayes factors is excellent when evidence favors $\mathcal{H}_1$. As evidence shifts toward $\mathcal{H}_0$, $eJAB_{10}$ decreases toward zero, as expected from model-selection consistency, whereas the test-statistic Bayes factor is inherently bounded below by 1. Additional comparisons with Dirichlet Bayes factors for chi-squared tests are provided in the *SI Appendix*. The two Bayes factors reach broad agreement, although differences persist at larger n (Fig. S1). At $n = 1,500$, under $\mathcal{H}_0$, the Dirichlet $BF_{10} < 1/3$ is almost always true, whereas $eJAB_{10} < 3$ is mostly observed; under $\mathcal{H}_1$, $eJAB_{10}$ more readily identifies an effect. Replacing *eJAB* with the BIC approximation to the Bayes factor performs well when $q = 1$, but marked severe discrepancies for $q > 1$, indicating that *eJAB* better

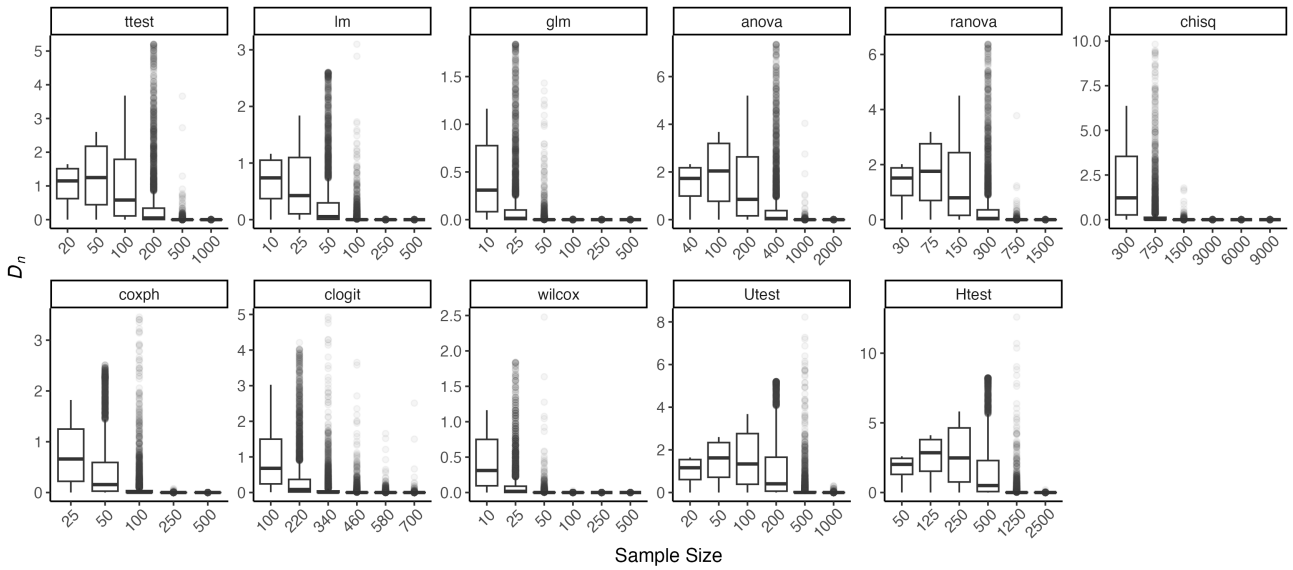

Fig. 1. Sampling distribution of $D_n = -\sqrt{n} \cdot p \ln p$ under $\mathcal{H}_1$ for 11 statistical tests. Panel labels: ttest (t-test), lm (linear regression), glm (generalized linear regression), anova (analysis of variance), ranova (repeated-measures ANOVA), chisq (chi-squared test), coxph (Cox proportional-hazards regression), clogit (conditional logistic regression), wilcox (Wilcoxon signed-rank test), Utest (Mann–Whitney test), and Htest (Kruskal–Wallis test).

approximates Bayes factors for ANOVA, rANOVA, and chi-squared tests (Fig. S2).

In sum, *eJAB* provides an accurate approximation to Bayes factors across a range of designs and effect sizes for parametric tests.

C. Bayesian Hypothesis Tests Using Nonparametric Statistics. Classical nonparametric tests do not specify a sampling distribution for the data, so marginal likelihoods are not defined. As such, the Bayes factor can be defined based on the marginal likelihood of the test statistic rather than on the data (26). We compared *eJAB* with Bayes factors constructed from nonparametric statistics (*BFNP*) across items 9 through 11. The supplementary material reports sequences of box plots depicting how the sampling distributions of *eJAB* and *BFNP* vary with effect size and sample size. Figures depicting these results and a detailed description of the experiments along with their associated R Markdown files is available on the Open Science Framework at http://osf.io/qeprj/, under 'Files' within the directories 'Supplementary Information', 'Nonparametric'.

Under $\mathcal{H}_1$, when effects and samples are both relatively small, *BFNP* tends to provide stronger evidence for $\mathcal{H}_1$ than *eJAB*. With larger effects or larger samples, the two methods produce comparable results. Under $\mathcal{H}_0$, *eJAB* effectively quantifies evidence for $\mathcal{H}_0$, as with chi-square tests.

3. Reevaluating Statistical Evidence in 71,126 Clinical Trial Findings with eJAB

Having evaluated *eJAB* across a wide range of tests, we apply it to reevaluate the evidence in a large corpus of CTG results. Prior work has raised concerns about the strength of statistical evidence in clinical research. For example, a 2012 study reported replication for only 11% of landmark preclinical cancer findings (37). A 2016 study found that most late-stage trials fail, largely for lack of efficacy (38). And, a 2019 study estimated that 90% of drug-development trials fail for similar reasons (39). Failures at later phases may reflect type I errors, where the null hypothesis was incorrectly rejected at an earlier phase.

Our objective is to characterize the overall strength of evidence in 71,126 CTG findings and to identify potential type I errors. These results span 43 MeSH categories (e.g., neoplasms, cardiovascular diseases, and autoimmune diseases) and encompass nine statistical tests: one-sample t-test, two-sample t-test, linear regression, logistic regression, ANOVA, conditional logistic regression, Wilcoxon signed-rank test, Mann–Whitney test, and Kruskal–Wallis test. For each, we extracted the p-value, sample size, and parameter dimension, and, thus, computed $eJAB_{01}$. We excluded survival analyses and chi-squared tests because CTG usually omits the number of uncensored observations and full contingency-table dimensions. Preprocessing to address misreported and left-censored p-values, sample sizes, and test labels (40) is described in *SI Appendix*.

The relationship between α and the share of reported findings with $p \leqslant \alpha$ and $eJAB_{01} > 1$ is strongly linear for $\alpha \in (0, .25)$, with a visible jump at $\alpha = .05$ due to results reported exactly as $p = .05$ (Fig. 3A). We attribute this to left-censoring at the target threshold. Restricting to complete cases removes this artifact and yields an almost perfectly linear trend for $\alpha \in (0, .4)$, strongly suggesting that findings with $p \leqslant \alpha$ and $eJAB_{01} > 1$ are type I errors at level α (Fig. 3B).

We therefore classify findings with $p \leqslant \alpha$ and $eJAB_{01} > 1$ as candidate type I errors at the significance level α. Among 30,790 results significant at $\alpha = .05$, we catalog 4,088 candidates. Although lower thresholds have been recomended in the literature (2, 3), $\alpha = .05$ remains relevant in CTG: in a random sample of 52 trial protocols, 75% (SE 6%) explicitly set $\alpha \geqslant .05$ as the target evidential threshold. We list all 4,088 candidates on the Open Science Framework at http://osf.io/qeprj/ ('Files', 'Supplementary Information') and

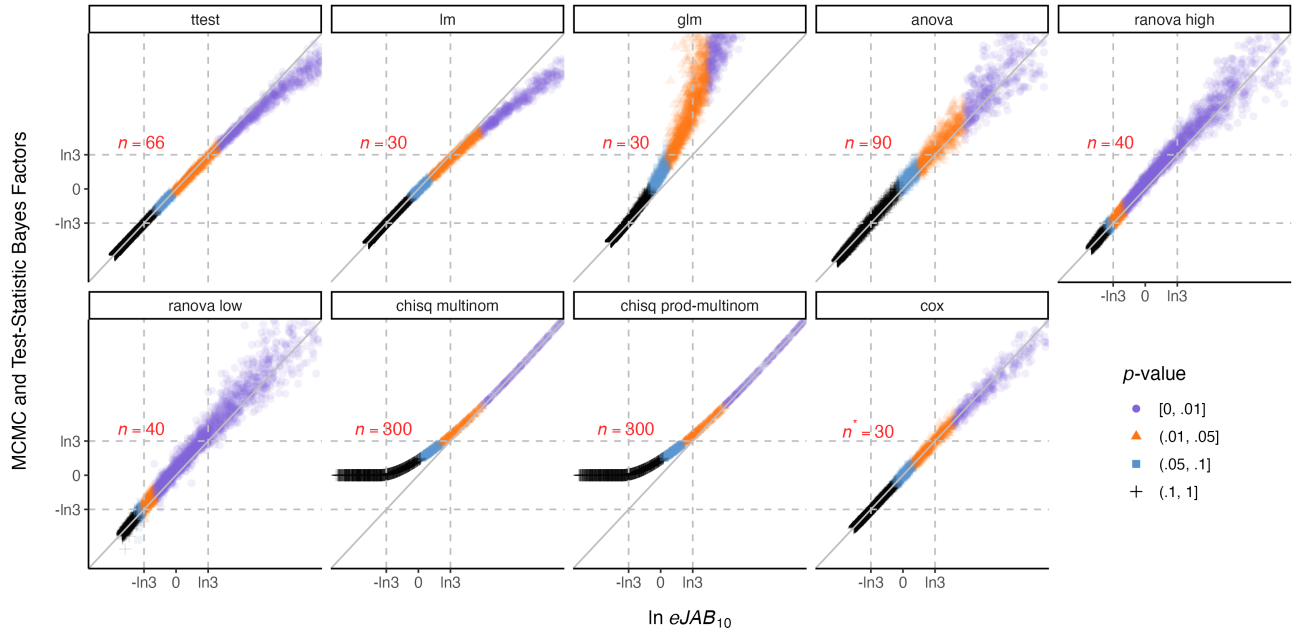

Fig. 2. Comparison of the natural logarithmic $eJAB_{10}$ with alternative Bayes-factor computations at a fixed sample size. The y-axis shows the MCMC-based Bayes factor for all panels, except for the two chi-squared tests, which use the test-statistic Bayes factor instead. The red text indicates the effective sample size n, except for the Cox model, which shows the number of observations. The dashed lines indicate $eJAB_{10}$ of 1/3 and 3, representing moderate evidence against and for an effect, respectively.

provide an annotated subset of 131 cases with $p \leqslant .01$ and $eJAB_{01} > 1$, cross-referenced to CTG and PubMed.

We fit a hierarchical Dirichlet-multinomial model with a flat prior on study-specific proportions of $p \leqslant .05$ and $eJAB_{01} \geqslant 1/3$, aggregating draws to estimate the overall proportion. We estimate that 35.6% (posterior SD 0.22%) of $p \leqslant .05$ results correspond to $eJAB_{01} \geqslant 1/3$, i.e., evidence for the alternative no stronger than anecdotal (Fig. 3C).

In the plot of $\ln eJAB_{01}$ versus p, with points colored by test type and shaped by sample size, the region to the left of $p = .05$ and above $eJAB_{01} = 3$ indicates instances of the Jeffreys–Lindley paradox—that is, cases where the p-value leads to rejection of $\mathcal{H}_0$, while the Bayes factor favors $\mathcal{H}_0$ over $\mathcal{H}_1$. Notably, $\ln eJAB_{01}$ shows substantial variability at the smallest p-values (Fig. 3D). Contradictions with $p \leqslant .05$ yet $eJAB_{01} > 3$ occur predominantly at larger sample sizes, where observed effect sizes can be negligible (Fig. 3E).

We also examine the proportion of all $p \leqslant .05$ and $eJAB_{01} \geqslant K$ for $K \in (0, 300)$; at $K = 1$, this proportion recovers the candidate type I error at $\alpha = .05$. The inset highlights Jeffreys–Lindley paradox cases. We identify 487 such instances in the CTG data (Fig. 3F). Additional results stratified by MeSH category appear in *SI Appendix*.

In sum, the near-perfect linearity indicates that findings with $p \leqslant \alpha$ and $eJAB_{01} > 1$ behave as type I errors in the CTG data when tested at $\alpha \in (0, .2]$. To elucidate how such candidates arise, consider a Wald test with $q = 1$ and independent and identically distributed data for fixed $n < \infty$ and $\alpha > 1 - F_{\chi^2_1}\left(\frac{n \ln n}{n-1}\right)$. A candidate type I error occurs when the observed effect size falls within the interval,

$$\sqrt{\frac{1}{n} Q_{\chi^2_1}(1-\alpha)} \cdot I_1^{-\frac{1}{2}}(\hat{\theta}) \leqslant |\hat{\theta} - \theta_0| < \sqrt{\frac{\ln n}{n-1}} \cdot I_1^{-\frac{1}{2}}(\hat{\theta}).$$

where I_1 is the Fisher information for one observation. In this regime, the observed effect size is large enough to achieve standard statistical significance at sample size n. Small enough, $eJAB_{01}$ favors the null. To favor the alternative, the latter requires the observed effect size to exceed a bound of order $\mathcal{O}_p\left(\sqrt{\frac{\ln n}{n}}\right)$.

We review several primary outcomes from interventional clinical trials designated as candidate type I errors at $\alpha = .01$.

1. **#NCT02605174. Three doses of lasmiditan (50, 100, and 200 mg) compared with placebo for the acute treatment of migraine:** $p = .009$ and $eJAB_{01} = 1.06$. Phase 3 trial of lasmiditan for acute migraine. We flag as a candidate type I error the reported effect that 50 mg increased the odds of being free of the most bothersome symptom versus placebo.

2. **#NCT00337727. Aprepitant for the prevention of chemotherapy-induced nausea and vomiting:** $p = .01$ and $eJAB_{01} = 1.05$. Phase 3 trial in patients starting moderately emetogenic chemotherapy. We flag as a candidate type I error the reported effect that aprepitant increased the odds of no vomiting versus standard therapy.

3. **#NCT01087541. Evaluation of safety in patients with diabetes:** $p = .01$ and $eJAB_{01} = 3.00$. Trial of a primary-care professional training program on diabetes endpoints. We flag as a candidate type I error the reported effect that the intervention lowered glycated hemoglobin (long-term blood glucose) relative to usual care.

4. **#NCT02896400. Optimizing tobacco-dependence treatment in the emergency department:** $p = .01$

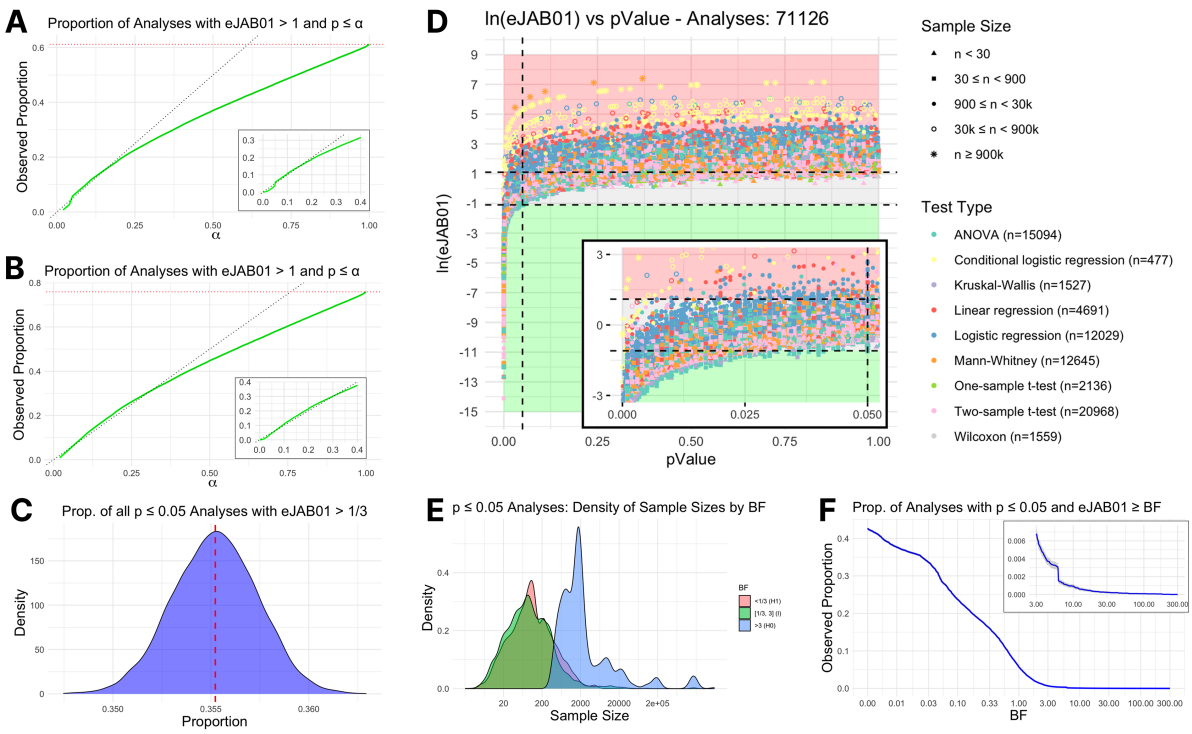

Fig. 3. Evidence in 71,126 results. (A-B) Proportion of analyses with $p \leqslant \alpha$ and $eJAB_{01} > 1$ as candidates type I errors, with $p = .05$ excluded in (B). The red dashed line marks the overall share with $eJAB_{01} > 1$ as $\alpha \to 1$. (C) Posterior distribution of the proportion of $p \leqslant .05$ and $eJAB_{01} > 1/3$. The red dashed line indicates the posterior mean. (D) Scatterplot of $\ln eJAB_{01}$ versus p-value, with an inset zoom for $p \leqslant .05$. Shaded bands denote evidence categories.(E) Density of sample sizes for all $p \leqslant .05$, stratified by Bayes-factor category. (F) Proportion of $p \leqslant .05$ with Bayes-factor threshold. The inset highlights the Jeffreys–Lindley paradox region.

and $eJAB_{01} = 1.18$. Trial of common interventions for adult smokers presenting to the emergency department. We flag as a candidate type I error the reported effect that a brief negotiated interview increased abstinence versus no such interview.

5. **#NCT02266277. System alignment for vaccine delivery:** $p = .0026$ and $eJAB_{01} = 1.07$. Trial of nonmedical strategies to increase influenza and pneumococcal vaccination. We flag as a candidate type I error the reported effect that a patient-portal vaccine reminder increased completion versus no message.

We examine the distributions of $\ln eJAB_{01}$ and $\ln p$ by trial phase for primary and secondary outcomes (Fig. 4A-B). For primary outcomes, there is no discernible trend toward stronger evidence from early phase to phase 4. The distributions of $\ln p$ shift downward at phase 3, consistent with the markedly larger average sample sizes at that phase. For secondary outcomes, we again observe no phase-wise increase in evidence, though variability in both measures is higher in phase 3. Plotting p and sample size by phase shows that changes in p track changes in sample size (Fig. S3).

To better visualize trends in evidence across phases, we fit a linear mixed model with a fixed intercept and trial-specific random effects to each of $\ln eJAB_{01}$ and $\ln p$, and then examine residuals re-centered by the estimated intercept to remove between-trial heterogeneity. The adjusted measures are essentially uniform across phases (Fig. 4C-D). For primary outcomes, most adjusted p-values lie near .05 and above .01. Using $eJAB$, this interpretation corresponds to 'anecdotal evidence' or 'evidence that is barely worth mentioning' for both primary and secondary outcomes (23). Comparing primary with secondary outcomes, we note a slight and phase-invariant shift toward greater evidence for the alternative in primary outcomes.

4. Conclusions

$eJAB$ can be interpreted as an approximate objective Bayes factor when p-values are obtained from Wald or likelihood-ratio tests. The same interpretation should hold for all tests whose statistics have an $o_p(1)$ difference with the Wald statistic. If this is not the case, $eJAB$ is model-selection consistent for p-values from any test satisfying **R1** and **R2** of Theorem 1.3.

In simulations, we compared the finite-sample performance of $eJAB$ with Bayes factors computed from MCMC and the Savage–Dickey density ratio for the t-test, linear regression, logistic regression, ANOVA, rANOVA, and parametric survival analysis. Under the same generalized unit-information prior, $eJAB$ closely approximates the MCMC-based Bayes factor. For a chi-squared test under multinomial or product-multinomial sampling, it agrees with the test-statistic Bayes factor under the alternative, unlike the latter, also registers evidence for the null when true. For nonparametric tests, namely Wilcoxon signed-rank, Mann–Whitney, and Kruskal–Wallis, $eJAB$ performs comparably to Bayes factors based on nonparametric test statistics. Our simulations also show improved performance compared to the BIC

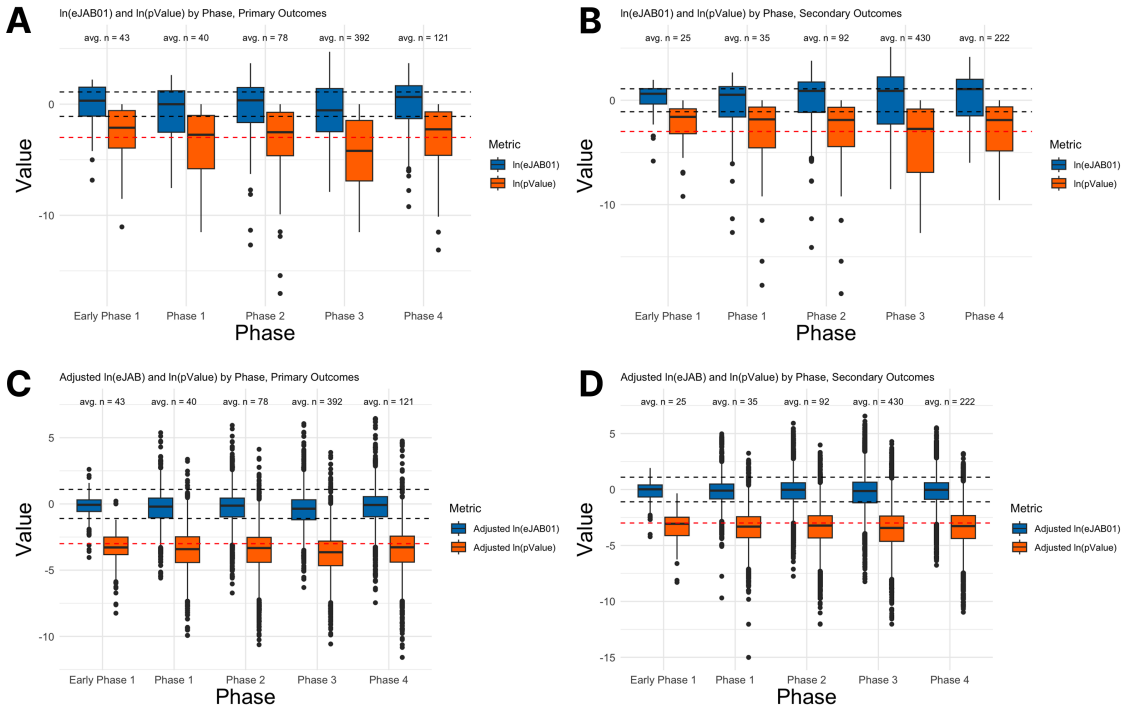

Fig. 4. Measures of evidence by clinical trial phase. (A-B) Distributions of $\ln eJAB_{01}$ and $\ln p$ by phase for primary (A) and secondary (B) outcomes. (C-D) Distributions of adjusted (A-B). Adjusted measures are obtained from linear mixed models. The average analysis sample size is shown above each facet. Black dashed lines mark $-\ln 3$ and $\ln 3$. The red dashed line marks $\ln 0.05$.

approximation to the Bayes factor when $q > 1$. Overall, the $eJAB$ calculation is satisfactory and broadly applicable.

In our reevaluation of 71,126 clinical trials, findings with $p \leqslant \alpha$ and $eJAB_{01} > 1$ exhibit characteristics of type I errors in the CTG data when tested at significance level $\alpha \in (0, .2]$. The uniform strength of evidence across phases is consistent with a roughly constant proportion of hypotheses with true effects. The 4,088 CTG candidate type I errors warrant targeted follow-up to investigate replicability.

Materials and Methods

All `R` code and the data used in the analyses are available on the Open Science Framework at https://osf.io/qeprj/. An $eJAB$ calculator is available at https://flxp0x-puneet-velidi.shinyapps.io/bayesfactor/.

ACKNOWLEDGMENTS. This work was supported by a discovery grant to F.S.N. (RGPIN-04044-2020) from the Natural Sciences and Engineering Research Council of Canada and research accelerator awards to P.V. and Y.L. from the Maud Menten Institute.

1. JPA Ioannidis, Why most published research findings are false. *PLOS Medicine* **2**, 696–701 (2005).
2. DJ Benjamin, et al., Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).
3. VE Johnson, Revised standards for statistical evidence. *Proc. Natl. Acad. Sci.* **110**, 19313–19317 (2013).
4. RL Wasserstein, NA Lazar, The ASA statement on p-values: Context, process, and purpose. *The Am. Stat.* **70**, 129–133 (2016).
5. JO Berger, T Sellke, Testing a point null hypothesis: The irreconcilability of p values and evidence. *J. Am. Stat. Assoc.* **82**, 112–122 (1987).
6. PD Grünwald, Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proc. Natl. Acad. Sci.* **121**, e2302098121 (2024).
7. M Evans, Measuring statistical evidence using relative belief. *Comput. Struct. Biotechnol. J.* **14**, 91–96 (2016).
8. LG Halsey, The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum? *Biol. Lett.* **15**, 1–8 (2019).
9. RE Kass, AE Raftery, Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
10. A Ly, J Verhagen, EJ Wagenmakers, An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *J. Math. Psychol.* **72**, 43–55 (2016).
11. AE Raftery, Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–163 (1995).
12. RD Morey, JW Romeijn, JN Rouder, The philosophy of Bayes factors and the quantification of statistical evidence. *J. Math. Psychol.* **72**, 6–18 (2016).
13. EJ Wagenmakers, A practical solution to the pervasive problems of p values. *Psychon. Bull. & Rev.* **14**, 779–804 (2007).
14. RE Kass, L Wasserman, A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* **90**, 928–934 (1995).
15. L Huisman, Are p-values and Bayes factors valid measures of evidential strength? *Psychon. Bull. & Rev.* **30**, 932–941 (2023).
16. Z Wei, FS Nathoo, ME Masson, Investigating the relationship between the Bayes factor and the separation of credible intervals. *Psychon. Bull. & Rev.* **30**, 1759–1781 (2023).
17. JASP Team, JASP (Version 0.95.2)[Computer software] (2025).
18. RD Morey, JN Rouder, BayesFactor: Computation of Bayes factors for common designs (2024) R package version 0.9.12-4.7.
19. S Sinharay, HS Stern, An empirical comparison of methods for computing Bayes factors in generalized linear mixed models. *J. Comput. Graph. Stat.* **14**, 415–435 (2005).
20. F Bartoš, EJ Wagenmakers, A general approximation to nested Bayes factors with informed priors. *Stat* **12**, e600 (2023).
21. K Rostgaard, Simple nested Bayesian hypothesis testing for meta-analysis, Cox, Poisson and logistic regression models. *Sci. Reports* **13**, 1–11 (2023).
22. DR Bickel, A small-sample Bayesian information criterion that does not overstate the evidence, with an application to calibrating p-values from likelihood-ratio tests. *Stat. Pap.* **66**, 1–17 (2025).
23. L Held, M Ott, On p-values and Bayes factors. *Annu. Rev. Stat. Its Appl.* **5**, 393–419 (2018).
24. VE Johnson, Bayes factors based on test statistics. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **67**, 689–701 (2005).
25. VE Johnson, Properties of Bayes factors based on test statistics. *Scand. J. Stat.* **35**, 354–368 (2008).
26. Y Yuan, VE Johnson, Bayesian hypothesis tests using nonparametric statistics. *Stat. Sinica* **18**, 1185–1200 (2008).
27. H Jeffreys, Some tests of significance, treated by the theory of probability. *Math. Proc. Camb. Philos. Soc.* **31**, 203–222 (1935).
28. H Jeffreys, Further significance tests. *Math. Proc. Camb. Philos. Soc.* **32**, 416–445 (1936).
29. EJ Wagenmakers, Approximate objective Bayes factors from p-values and sample size: The $3p\sqrt{n}$ rule. *PsyArXiv* **1**, 1–50 (2022).

30. AA Neath, JE Cavanaugh, The Bayesian information criterion: Background, derivation, and applications. *WIREs Comput. Stat.* **4**, 199–203 (2012).
31. J Mulder, EJ Wagenmakers, M Marsman, A generalization of the Savage–Dickey density ratio for testing equality and order constrained hypotheses. *The Am. Stat.* **76**, 102–109 (2022).
32. ME Masson, A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behav. Res. Methods* **43**, 679–690 (2011).
33. FS Nathoo, MEJ Masson, Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *J. Math. Psychol.* **72**, 144–157 (2016).
34. Stan Development Team, RStan: the R interface to Stan (2025) R package version 2.32.7.
35. C Kooperberg, CJ Stone, Logspline density estimation for censored data. *J. Comput. Graph. Stat.* **1**, 301–328 (1992).
36. E Gunel, J Dickey, Bayes factors for independence in contingency tables. *Biometrika* **61**, 545–557 (1974).
37. CG Begley, LM Ellis, Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
38. TJ Hwang, et al., Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Intern. Medicine* **176**, 1826–1833 (2016).
39. D Sun, W Gao, H Hu, S Zhou, Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sinica B* **12**, 3049–3062 (2022).
40. L Miron, RS Gonçalves, MA Musen, Obstacles to the reuse of study metadata in ClinicalTrials.gov. *Sci. Data* **7**, 1–14 (2020).