Supporting Information for

Generalized Jeffreys' Approximate Objective Bayes Factor: Model Selection Consistency, Finite Sample Accuracy and the Statistical Evidence in 71,126 Clinical Trial Findings

P. Velidi, Z. Wei, S. N. Kalaria, Y. Liu, C. M. Laumont, B. H. Nelson, F. S. Nathoo

Farouk S. Nathoo.
E-mail: nathoo@uvic.ca

This PDF file includes:

1. Preprocessing Steps

CTG Download We downloaded the CTG database for studies that were completed and had results posted as a JSON with all available fields at this URL https://clinicaltrials.gov/search?aggFilters=results:with,status:com on August 7th, 2025. We excluded four entries for which we found errors in sample size upon further investigation of the candidate type I errors at $\alpha = 0.01$.

Record construction We transformed each ClinicalTrials.gov JSON record into flat, analysis-level rows before p-value cleaning and method harmonization: For each study (`nctId`), we iterated outcome measures that contained at least one analysis. For every outcome–analysis pair, we emitted one row per denominator count where the units were "Participants" and the `groupId` belonged to the analysis's declared `groupIds`. We carried forward study metadata (e.g., `briefTitle`, `studyType`, design info fields `interventionModel`, `observationalModel`, `timePerspective`, `allocation`), one–hot encoded phases (`EARLY_PHASE1`, `PHASE1`, `PHASE2`, `PHASE3`, `PHASE4`), and MeSH terms/IDs for conditions and interventions from the `derivedSection`. We also flagged whether the study had any references of type `RESULTS` or `DERIVED`.

Analysis identifier. Each row was keyed by a stable, hash-based analysis identifier

$$\texttt{analysisId} \;=\; \mathrm{MD5}\Big(\texttt{nctId} + \mathrm{JSON}(\texttt{outcome};\ \text{sorted keys}) + \mathrm{JSON}(\texttt{analysis};\ \text{sorted keys})\Big),$$

which ensures reproducible IDs for unique outcome–analysis pairs within a study (independent of JSON key order). This ID is used as the unit of aggregation in subsequent preprocessing (e.g., computing sample sizes and $eJAB_{01}$).

Schema `analysisId`, `nctId`, `briefTitle`, `outcomeType`, `outcomeTitle`, `groupDescription`, `units`, `groupId`, `value`, `pValue`, `statisticalMethod`, MeSH fields (`condition_mesh`, `condition_ids`, `intervention_mesh`, `intervention_ids`), `hasResultsOrDerived`, study design fields, and phase indicators. `value` is the number of participants for a given outcome group/arm.

Standardization of reported p-values Reported p-values often contained textual qualifiers (e.g., "<0.05"). We standardized them as follows:

1. Flagged the presence of a "<" symbol: `has_less_than` ← `grepl("<", pValue)`.
2. Extracted the numeric component using regex: `pValue` ← `str_extract(pValue, "[0-9]*.?[0-9]+")`.
3. Removed rows with missing or blank numeric parts.
4. Converted `pValue` to numeric and retained only valid probabilities: $0 <$ `pValue` < 1.
5. Excluded rows with zero sample size (`value` $\neq 0$), which are incompatible with downstream model mappings.

Harmonization of statistical method names Free-text `statisticalMethod` labels were mapped to canonical families using regular-expression rules with `str_detect`. Observations with no match were excluded. The exact mapping used was:

Listing 1. Regex-based mapping of free-text methods to canonical families

```
statisticalMethod <- case_when(
  # --- T-TESTS ---
  str_detect(originalMethod, regex("\\bt[- ]?test", ignore_case = TRUE)) &
    str_detect(originalMethod, regex("one[- ]?sample", ignore_case = TRUE)) ~ "One-sample t-test",

  str_detect(originalMethod, regex("\\bt[- ]?test", ignore_case = TRUE)) &
    str_detect(originalMethod, regex("paired", ignore_case = TRUE)) ~ "One-sample t-test",

  str_detect(originalMethod, regex("\\bt[- ]?test", ignore_case = TRUE)) ~ "Two-sample t-test",

  # --- CONDITIONAL LOGISTIC REGRESSION ---
  str_detect(originalMethod, regex("logistic", ignore_case = TRUE)) &
    str_detect(originalMethod, regex("regression", ignore_case = TRUE)) &
    str_detect(originalMethod, regex("conditional", ignore_case = TRUE)) ~ "Conditional logistic
      regression",

  # --- LOGISTIC REGRESSION ---
  str_detect(originalMethod, regex("logistic", ignore_case = TRUE)) &
    str_detect(originalMethod, regex("regression", ignore_case = TRUE)) ~ "Logistic regression",

  # --- COX ---
  str_detect(originalMethod, regex("cox", ignore_case = TRUE)) &
    !str_detect(originalMethod, regex("wilcoxon|mantel|signed|rank", ignore_case = TRUE)) ~ "Cox",

  # --- LOGRANK / MANTEL-COX ---
  str_detect(originalMethod, regex("log ?rank|mantel[- ]?cox", ignore_case = TRUE)) &
    !str_detect(originalMethod, regex("wilcoxon|signed", ignore_case = TRUE)) ~ "Logrank",

  # --- MANN-WHITNEY ---
```

```
  str_detect(originalMethod, regex("mann[- ]?whitney", ignore_case = TRUE)) |
    (str_detect(originalMethod, regex("wilcoxon", ignore_case = TRUE)) &
       str_detect(originalMethod, regex("rank", ignore_case = TRUE)) &
       !str_detect(originalMethod, regex("signed", ignore_case = TRUE))) ~ "Mann-Whitney",

  # --- WILCOXON SIGNED-RANK ---
  str_detect(originalMethod, regex("wilcoxon", ignore_case = TRUE)) &
    str_detect(originalMethod, regex("signed", ignore_case = TRUE)) &
    str_detect(originalMethod, regex("rank", ignore_case = TRUE)) ~ "Wilcoxon",

  # --- KRUSKAL-WALLIS ---
  str_detect(originalMethod, regex("kruskal|wallis", ignore_case = TRUE)) ~ "Kruskal-Wallis",

  # --- CHI-SQUARE ---
  str_detect(originalMethod, regex("chi[- ]?square| 2 |chi2|chisq", ignore_case = TRUE)) ~ "Chi-
      square test",

  # --- REPEATED MEASURES ---
  str_detect(originalMethod, regex("repeated[- ]?measures?", ignore_case = TRUE)) ~ "Repeated
      measures analysis",

  # --- LINEAR REGRESSION ---
  str_detect(originalMethod, regex("linear", ignore_case = TRUE)) &
    str_detect(originalMethod, regex("regression", ignore_case = TRUE)) ~ "Linear regression",

  # --- ANOVA ---
  str_detect(originalMethod, regex("anova|analysis of variance", ignore_case = TRUE)) ~ "ANOVA",

  TRUE ~ NA_character_
)
```

Adjustment for one-sided p-values When the free text indicated a one-sided analysis (patterns: `one-sided` or `1-sided`), we converted to a two-sided p-value by doubling and truncating at 1:

$$\texttt{pValue} \leftarrow \min\{2 \cdot \texttt{pValue},\ 1\}.$$

This rule is applied only when detected by `str_detect` on `originalMethod`.

analysisId Sample Size QC Certain unit-level tests should have exactly one row per `analysisId` (e.g., One-sample t-test, Wilcoxon signed-rank). To enforce internal consistency, we removed groups where such tests appeared with multiple rows.

Aggregation to the test level and parameter construction We summarized data by `analysisId` (test level). Invariant fields within a test were taken from the first row (e.g., `nctId`, `pValue`, `statisticalMethod`, `studyType`, phase indicators `EARLY_PHASE1`, `PHASE1`, `PHASE2`, `PHASE3`, `PHASE4`, `outcomeTitle`, `outcomeType`, `groupDescription`, `conditionIds`, `briefTitle`, `has_less_than`).

We constructed the effective sample size N per test as:

$$N = \begin{cases} \texttt{first(value)}, & \text{One-sample t-test or Wilcoxon signed-rank;} \\ \max \texttt{value}, & \text{Conditional logistic regression;} \\ 0.5 \sum \texttt{value}, & \text{CoxPH or Logrank;} \\ \sum \texttt{value}, & \text{otherwise.} \end{cases}$$

We also set method-specific parameters required by `JAB01`:

$$\texttt{I} = \begin{cases} n_{\text{rows within test}}, & \text{ANOVA, Kruskal–Wallis, Repeated measures;} \\ \text{NA}, & \text{otherwise,} \end{cases} \qquad (\texttt{R}, \texttt{C}) = \begin{cases} (n_{\text{rows within test}},\ 2), & \text{Chi-square test;} \\ (\text{NA},\ \text{NA}), & \text{otherwise.} \end{cases}$$

Mapping to JAB01 models and computation of eJAB01 We mapped the canonical `statisticalMethod` to the `JAB01` model string required for eJAB01. Here we only perform mapping for tests we intend on including in our analysis, so excluding Chi-square, CoxPH, Logrank, and repeated measures:

Two-sample t-test, Paired t-test, One-sample t-test → `"t-test"`,
Linear regression → `"linear_regression"`,
Logistic regression, Conditional logistic regression → `"logistic_regression"`,
ANOVA → `"anova"`, Kruskal–Wallis → `"kruskal_wallis"`,
Mann–Whitney → `"mann_whitney"`, Wilcoxon → `"wilcoxon"`.

Given $(N, \texttt{pValue}, \texttt{model}, R, C, I)$ per test, we computed eJAB01:

$$\texttt{JAB} = \texttt{JAB01}(N,\ \texttt{pValue},\ \texttt{model};\ R,\ C,\ I).$$

Exclusions and final dataset We excluded rows with missing `model`, missing `JAB` or `pValue`, and tests with $N = 1$. The resulting dataset `study3_summary` contains one record per `analysisId` with standardized `pValue`, computed `JAB`, method family, phase indicators, and associated metadata.

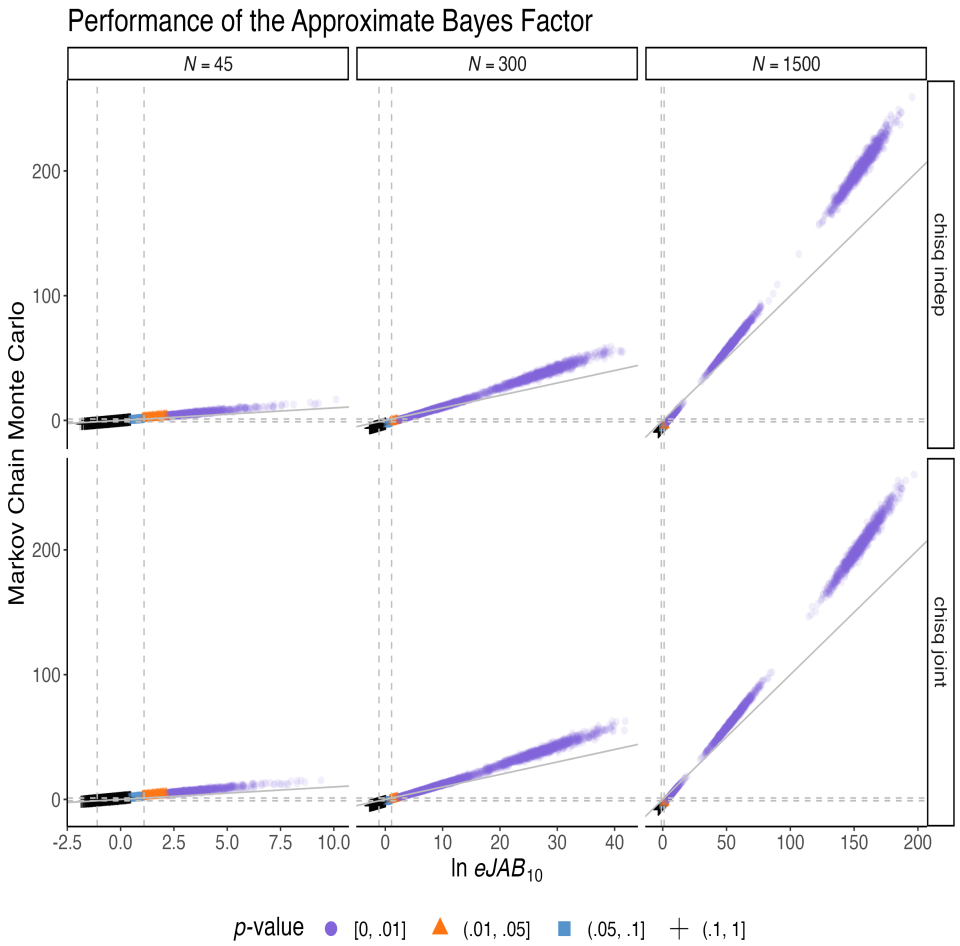

Fig. S1. Comparison of $\ln eJAB_{10}$ with $\ln BF_{10}$ computed using Markov chain Monte Carlo through the contingencyTableBF function from the BayesFactor R package. The sample size N is the number of items in the contingency table.

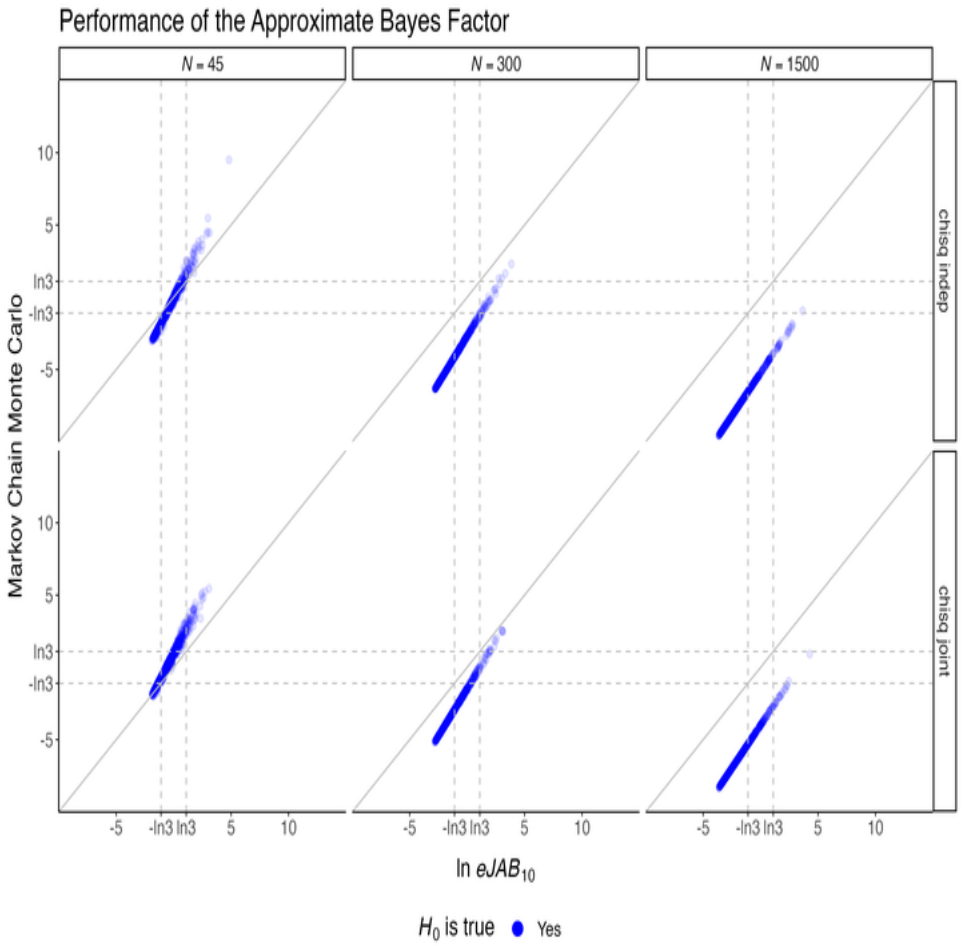

Fig. S2. The same plots depicted in Figure S1 focusing on a subregion closer to the origin and showing cases simulated from the null hypothesis.

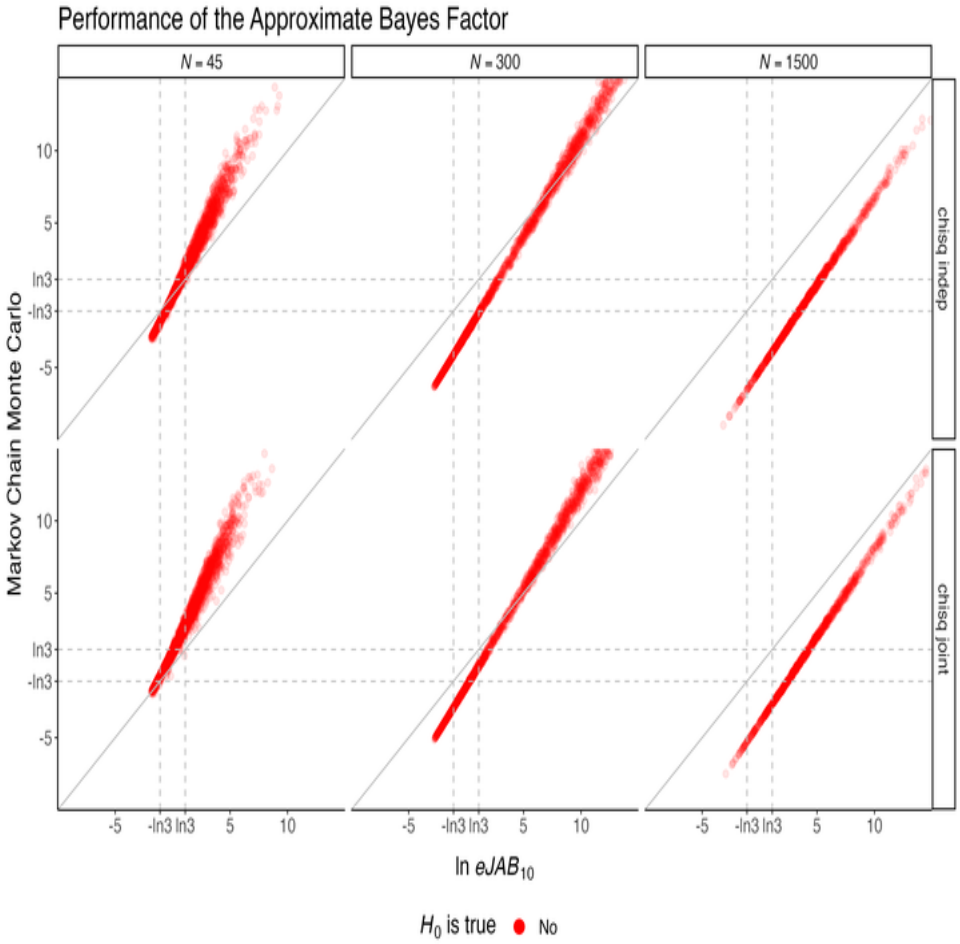

Fig. S3. The same plots depicted in Figure S1 focusing on a subregion closer to the origin and showing cases simulated from the alternative hypothesis.

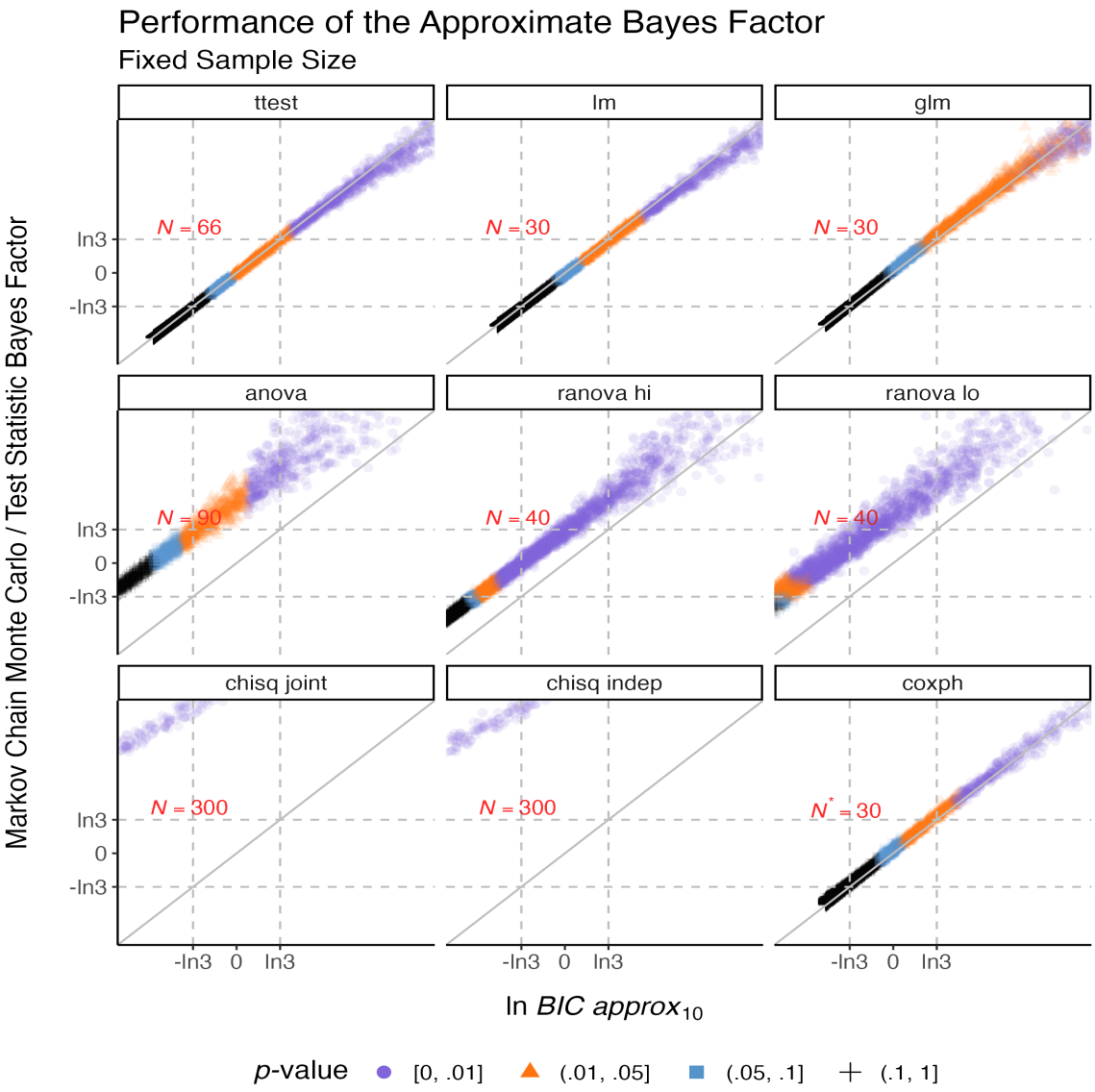

Fig. S4. Comparison of the BIC approximation to the Bayes factor with alternative approaches for computing the Bayes factor. The y-axis is the $\ln BF_{10}$ computed from simulated data with MCMC sampling, logspline density estimation and the Savage-Dickey density ratio for all panels with the exception of the two panels corresponding to chi-square tests. For chi-square tests, the y-axis is $\ln TSBF_{10}$, the test statistic Bayes factor. In all cases the x-axis is $\ln BF_{10}^{(BIC)}$. In computing the BIC, the sample size for survival analysis is taken to be the number of uncensored observation while we use (1) for repeated measures designs.

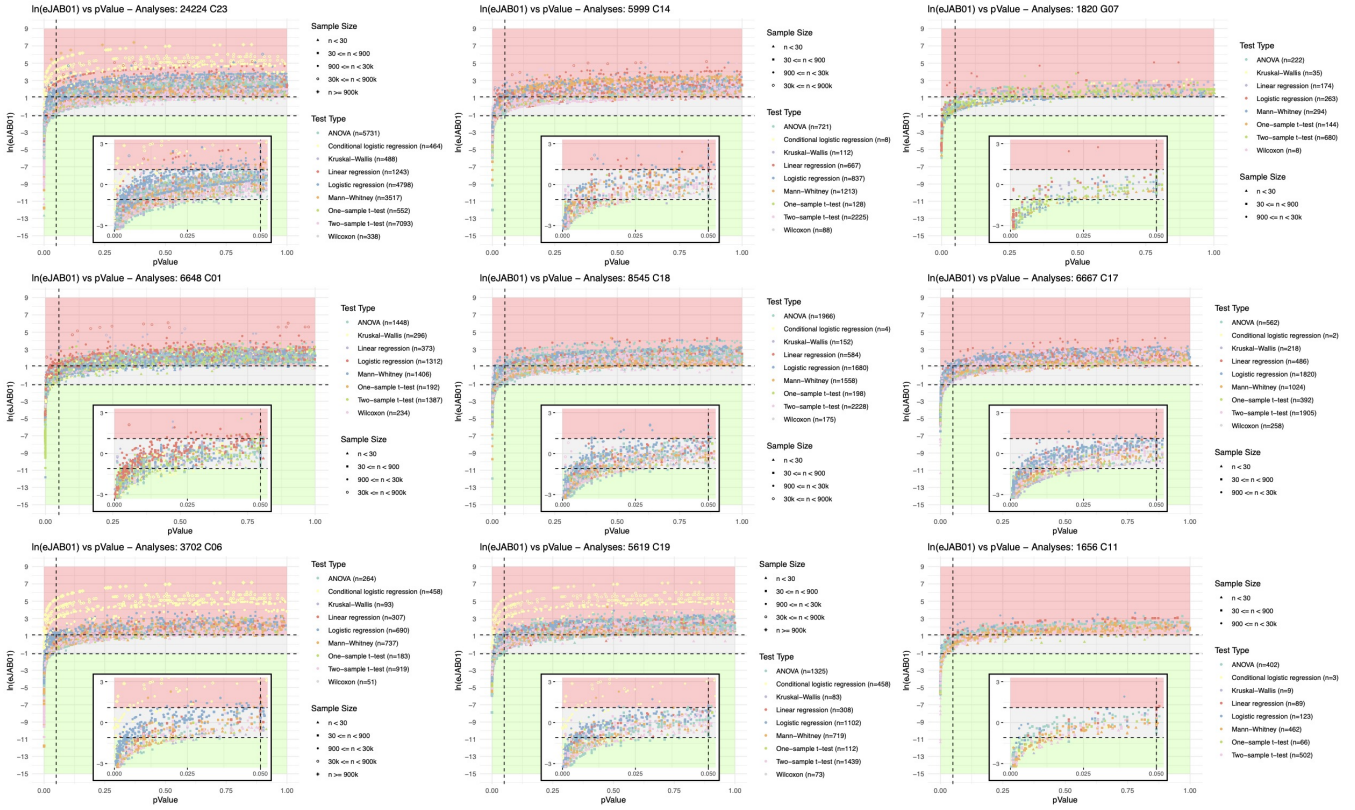

Fig. S5. Scatterplot of $\ln(eJAB_{01})$ by pValue for MeSH condition with an inset plot for closer examination of significant findings. The red band is the range of evidence for $\mathcal{H}_0$ by $eJAB_{01}$, the grey band is the range of inconclusive evidence for either $\mathcal{H}_0$ or $\mathcal{H}_1$ by $eJAB_{01}$, the green band is the range of evidence for $\mathcal{H}_1$ by $eJAB_{01}$. These Bayes factor categories are from a widely accepted standard of strength of evidence.

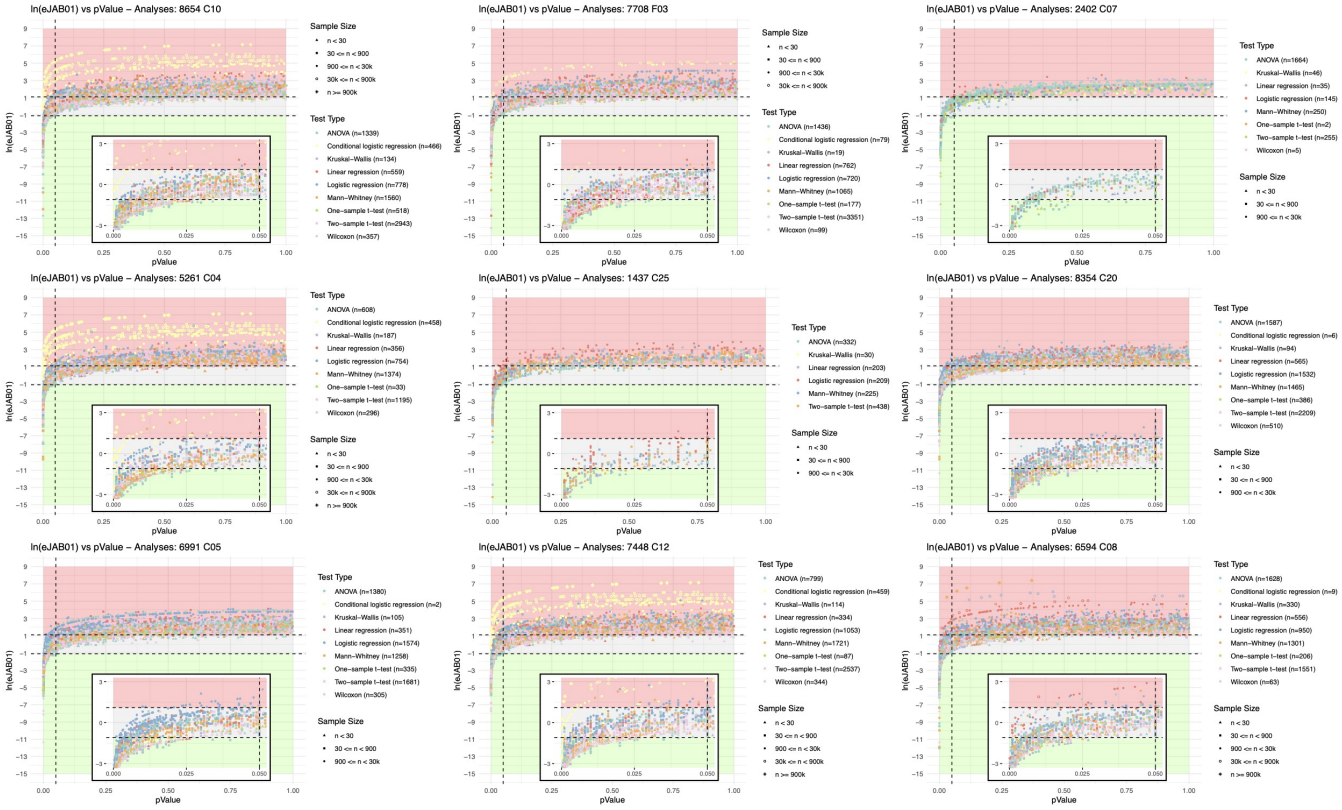

Fig. S5. Scatterplot of $\ln(eJAB_{01})$ by pValue for MeSH condition with an inset plot for closer examination of significant findings. The red band is the range of evidence for $\mathcal{H}_0$ by $eJAB_{01}$, the grey band is the range of inconclusive evidence for either $\mathcal{H}_0$ or $\mathcal{H}_1$ by $eJAB_{01}$, the green band is the range of evidence for $\mathcal{H}_1$ by $eJAB_{01}$. These Bayes factor categories are from a widely accepted standard of strength of evidence.

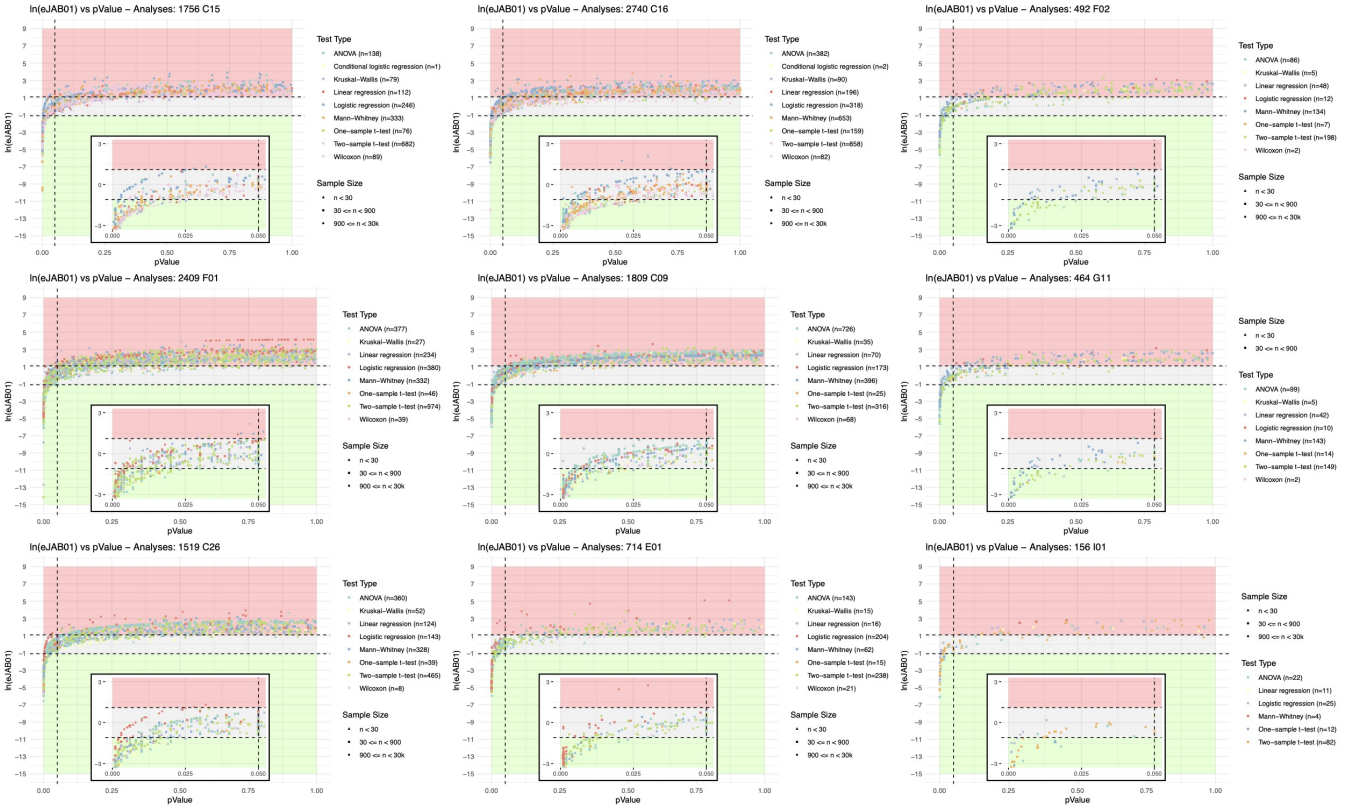

Fig. S5. Scatterplot of $\ln(eJAB_{01})$ by pValue for MeSH condition with an inset plot for closer examination of significant findings. The red band is the range of evidence for $\mathcal{H}_0$ by $eJAB_{01}$, the grey band is the range of inconclusive evidence for either $\mathcal{H}_0$ or $\mathcal{H}_1$ by $eJAB_{01}$, the green band is the range of evidence for $\mathcal{H}_1$ by $eJAB_{01}$. These Bayes factor categories are from a widely accepted standard of strength of evidence.

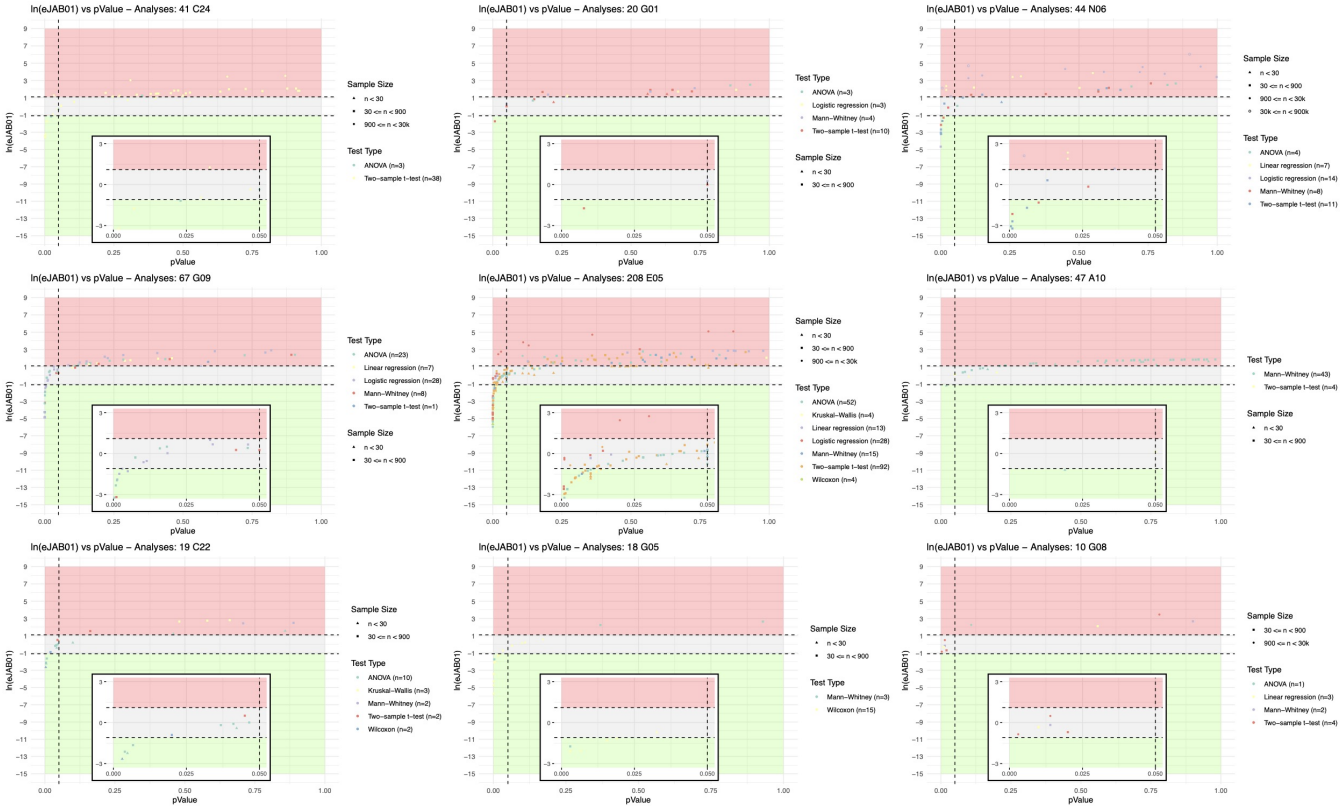

Fig. S5. Scatterplot of $\ln(eJAB_{01})$ by pValue for MeSH condition with an inset plot for closer examination of significant findings. The red band is the range of evidence for $\mathcal{H}_0$ by $eJAB_{01}$, the grey band is the range of inconclusive evidence for either $\mathcal{H}_0$ or $\mathcal{H}_1$ by $eJAB_{01}$, the green band is the range of evidence for $\mathcal{H}_1$ by $eJAB_{01}$. These Bayes factor categories are from a widely accepted standard of strength of evidence.

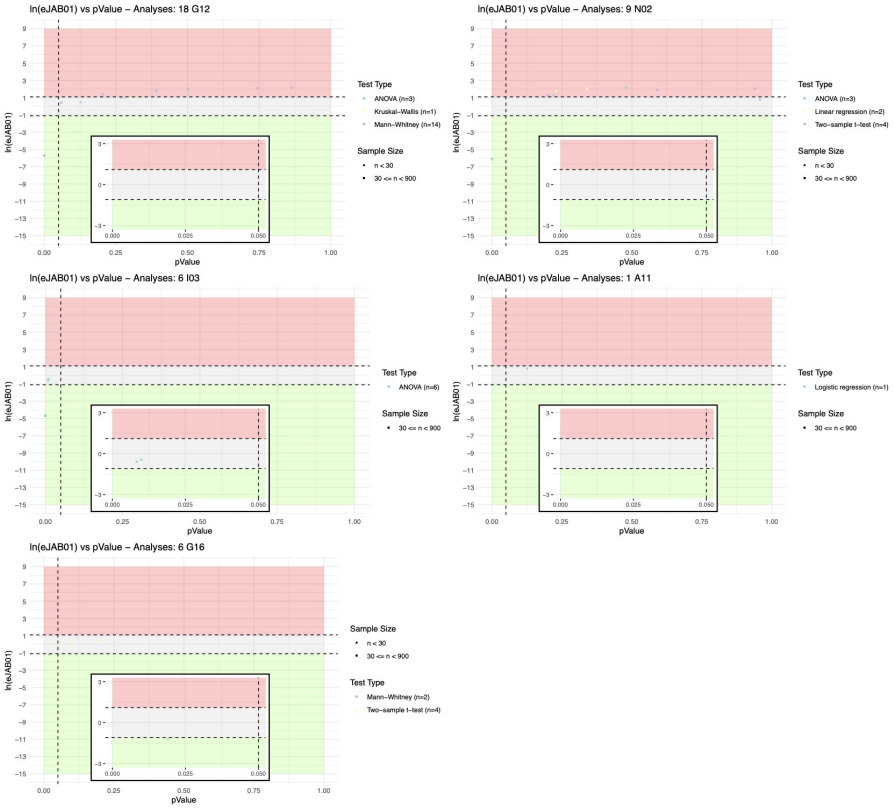

Fig. S5. Scatterplot of ln($eJAB_{01}$) by pValue for MeSH condition with an inset plot for closer examination of significant findings. The red band is the range of evidence for $\mathcal{H}_0$ by $eJAB_{01}$, the grey band is the range of inconclusive evidence for either $\mathcal{H}_0$ or $\mathcal{H}_1$ by $eJAB_{01}$, the green band is the range of evidence for $\mathcal{H}_1$ by $eJAB_{01}$. These Bayes factor categories are from a widely accepted standard of strength of evidence.

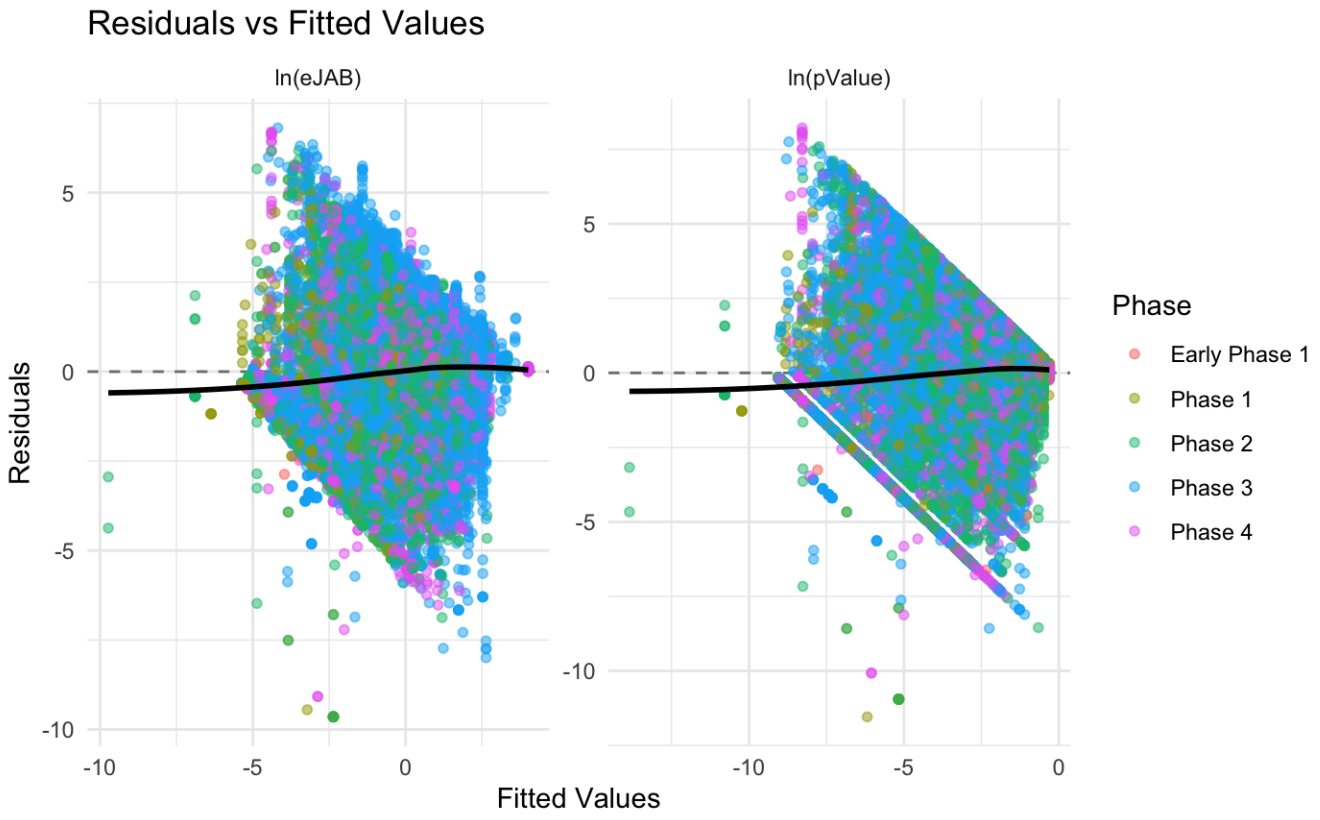

Fig. S6. Residuals versus fitted values from random–intercept linear mixed–effects models fit by REML. The model is $MOE_{i,j} = \mu + b_i + e_{i,j}$ for study i and analysis j, where $e_{i,j}$ is the residual, b_i is the study random effect, and μ is the intercept. The adjusted measure of evidence is $aMOE_{i,j} = MOE_{i,j} - b_i$, with estimator $a\hat{MO}E_{i,j} = \hat{e}_{i,j} + \hat{\mu}$. Facets show fitted values vs residuals for the measures of evidence $\ln(\mathrm{eJAB}_{01})$ and $\ln(p)$ in this model. The parallel lines correspond to discrete values at which the pValue is reported.

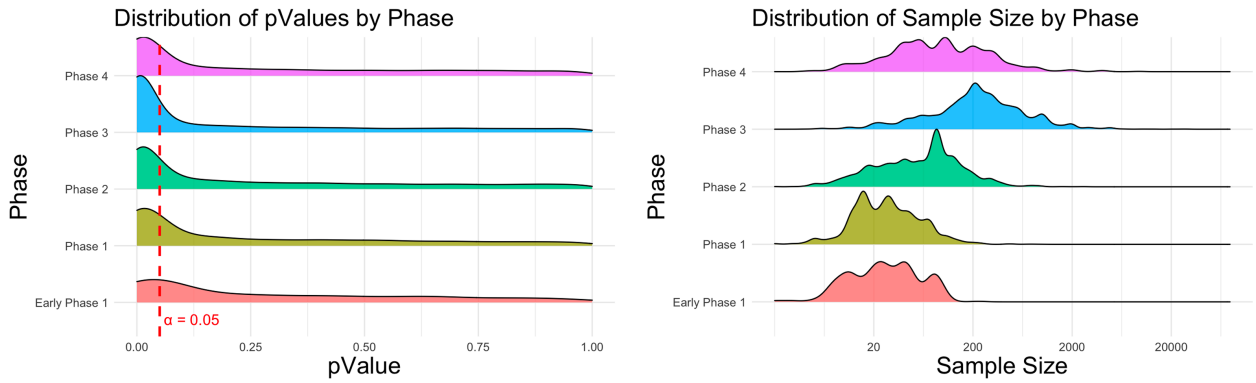

Fig. S7. Distributions of pValues and sample size by clinical trial phase.

References

1. FS Nathoo, MEJ Masson, Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *J. Math. Psychol.* **72**, 144–157 (2016).