UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions

Dablander, F.; Huth, K.; Gronau, Q.F.; Etz, A.; Wagenmakers, E.-J.

DOI
10.1002/sim.9278

Publication date
2022

Document Version
Final published version

Published in
Statistics in Medicine

License
CC BY-NC

Link to publication

Citation for published version (APA):
Dablander, F., Huth, K., Gronau, Q. F., Etz, A., & Wagenmakers, E.-J. (2022). A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions. *Statistics in Medicine*, *41*(8), 1319-1333. https://doi.org/10.1002/sim.9278

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

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

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

Download date:12 Oct 2026

Check for updates

Received: 11 August 2021 | Revised: 5 November 2021 | Accepted: 17 November 2021

DOI: 10.1002/sim.9278

RESEARCH ARTICLE

Statistics in Medicine WILEY

A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions

Fabian Dablander[1] | **Karoline Huth**[1] | **Quentin F. Gronau**[1] | **Alexander Etz**[2] | **Eric-Jan Wagenmakers**[1]

[1]Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands

[2]Department of Cognitive Sciences, University of California, Irvine, California

Correspondence
Fabian Dablander, Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands.
Email: dablander.fabian@gmail.com

Testing the equality of two proportions is a common procedure in science, especially in medicine and public health. In these domains, it is crucial to be able to quantify evidence for the absence of a treatment effect. Bayesian hypothesis testing by means of the Bayes factor provides one avenue to do so, requiring the specification of prior distributions for parameters. The most popular analysis approach views the comparison of proportions from a contingency table perspective, assigning prior distributions directly to the two proportions. Another, less popular approach views the problem from a logistic regression perspective, assigning prior distributions to logit-transformed parameters. Reanalyzing 39 null results from the *New England Journal of Medicine* with both approaches, we find that they can lead to markedly different conclusions, especially when the observed proportions are at the extremes (ie, very low or very high). We explain these stark differences and provide recommendations for researchers interested in testing the equality of two proportions and users of Bayes factors more generally. The test that assigns prior distributions to logit-transformed parameters creates prior dependence between the two proportions and yields weaker evidence when the observations are at the extremes. When comparing two proportions, we argue that this test should become the new default.

KEYWORDS

Bayesian testing, evidence of absence, equality of proportions, data reanalysis

1 | INTRODUCTION

Researchers frequently wish to test whether two populations differ. In medicine and public health, for example, the resulting statistical analysis frequently concerns testing whether or not two proportions differ. Examples include testing whether a vaccine decreases the number of infections compared to a control,[1] whether sexual minorities are more prone to suicide compared to their heterosexual counterparts,[2] or whether tightly or less-tightly controlling hypertension leads to fewer miscarriages in pregnant women.[3]

In these applications, it is crucial to be able to discriminate between evidence of absence and absence of evidence. For example, Magee et al conducted a trial to investigate the effect of a tight (target diastolic blood pressure, 85 mm Hg)

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2021 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

or a less-tight (target diastolic blood pressure, 100 mm Hg) control of hypertension in pregnant woman on, among other outcomes, pregnancy loss.[3] They found no significant difference between the two conditions, with 15 out of 493 women in the less-tight control condition and 13 out of 488 in the tight control condition having lost their child, yielding an estimated odds ratio of 1.14 (95% CI: [0.53, 2.45]). How confident are we that there is indeed no difference between the two conditions rather than the data being inconclusive? Bayesian statistics provides a principled way of quantifying evidence via the Bayes factor,[4-6] thus providing one avenue to discriminate between evidence of absence and absence of evidence.[7]

The Bayes factor quantifies how well one hypothesis predicts the data compared to another. Using the test of equality between two proportions as an example, let $D = (y_1, y_2, n_1, n_2)$ denote the combined data from the two groups. We have:

$$Y_1 \sim \text{Binomial}(n_1, \theta_1),$$
$$Y_2 \sim \text{Binomial}(n_2, \theta_2),$$

where the sample sizes (n_1, n_2) are assumed fixed and under $\mathcal{H}_0$ we have that $\theta \equiv \theta_1 = \theta_2$ while under $\mathcal{H}_1$ we have that $\theta_1 \neq \theta_2$. By quantifying relative predictive performance, the Bayes factor tells us how we should update our prior beliefs about $\mathcal{H}_0$ relative to $\mathcal{H}_1$ after observing the data:[8]

$$\underbrace{\frac{p(\mathcal{H}_0|D)}{p(\mathcal{H}_1|D)}}_{\text{Posterior odds}} = \underbrace{\frac{p(D|\mathcal{H}_0)}{p(D|\mathcal{H}_1)}}_{\text{Bayes factor}} \times \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\text{Prior odds}}.$$

A Bayes factor of, say, 15 means that the data are 15 times more likely under one hypothesis compared to the other. While there exist verbal guidelines that may aid in the interpretation of the Bayes factor (eg, Bayes factors in the range from 1 to 3 constitute *weak* evidence, those in the range from 3 to 10 constitute *moderate* evidence, and values larger than 10 constitute *strong* evidence)[9-11] the Bayes factor should be understood as a continuous measure of evidence.[6]

While the Bayes factor does not depend on the prior probability of hypotheses, it does depend crucially on the prior over parameters in the models instantiating $\mathcal{H}_0$ and $\mathcal{H}_1$, which becomes apparent when expanding:

$$\frac{p(D|\mathcal{H}_0)}{p(D|\mathcal{H}_1)} = \frac{\int_\theta p(D|\theta, \mathcal{H}_0)\pi_0(\theta|\mathcal{H}_0)\ \mathrm{d}\theta}{\int_{\theta_1}\int_{\theta_2} p(D|\theta_1, \theta_2, \mathcal{H}_1)\pi_1(\theta_1, \theta_2|\mathcal{H}_1)\ \mathrm{d}\theta_1\ \mathrm{d}\theta_2},$$

where π_0 and π_1 indicate the respective prior distributions.

There exist two main Bayes factor approaches for testing the equality of two proportions. The more popular one comes from the analysis of contingency tables, and assigns independent beta distributions directly to (θ_1, θ_2).[12,13] We call this approach the "Independent Beta" (IB) approach. The second approach is less widely used, and assigns a prior to the average log odds β and the log odds ratio ψ.[14,15] We call this approach the "logit transformation" (LT) approach. In this paper, we show that these two approaches can yield markedly different results. This is especially the case when the observed proportions are at the extremes (ie, very low or very high), as is the case for a large number of applications including the three examples mentioned above. Consider the study by Magee et al again.[3] The IB approach yields a Bayes factor of 12.30 in favor of $\mathcal{H}_0$, while the LT approach yields a mere 1.17. In other words, under the IB approach the data are about 12 times more likely under the hypothesis that tightly or less-tightly controlling hypertension have the same effect on miscarriages compared to a hypothesis assuming a difference. Under the LT approach, however, the data are about equally likely under both hypotheses, which constitutes equivocal evidence. The answer to the question "is observing two equally small proportions strong or weak evidence for the null hypothesis?" depends, therefore, crucially—and nontrivially—on the prior setup.

This article is structured as follows. In Section 2, we outline these two ways of testing the equality of two proportions in more detail. In Section 3, we highlight the occasionally stark differences of the two approaches by reanalyzing 39 statistical tests reported in the *New England Journal of Medicine* and explain why these differences occur. In Section 4, we end by reviewing the implications of the prior setup and what users of Bayes factors should be mindful of when testing the equality of two population parameters. We argue that the LT approach should become the default when testing the equality of two proportions because it (a) induces prior dependence between proportions which almost always are, in fact, dependent, and (b) yields a sensibly milder assessment of the evidence compared to the IB approach when the observations are at the extremes.

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2 | TWO WAYS OF TESTING THE EQUALITY OF TWO PROPORTIONS

In this section, we outline two ways of testing the equality of two proportions. In Section 2.1 we describe the Independent Beta approach, and in Section 2.2 we describe the logit transformation approach.

2.1 | The independent beta (IB) approach

In order to nest the null hypothesis under the alternative hypothesis, we introduce the difference parameter $\eta = \theta_2 - \theta_1$ and the grand mean $\zeta = \frac{1}{2}(\theta_1 + \theta_2)$. Using this parameterization, we have that:

$$\theta_1 = \zeta - \frac{\eta}{2},$$
$$\theta_2 = \zeta + \frac{\eta}{2}.$$

The hypotheses are then specified as:

$$\mathcal{H}_0 : \eta = 0,$$
$$\mathcal{H}_1 : \eta \neq 0.$$

Under this approach, we assign independent Beta(a, a) priors to θ_1 and θ_2. Figure 1 visualizes the joint prior distribution (top) under $\mathcal{H}_1$ for $a = 1$ (left) and $a = 2$ (right). Increasing values of a implies that the joint prior mass is more

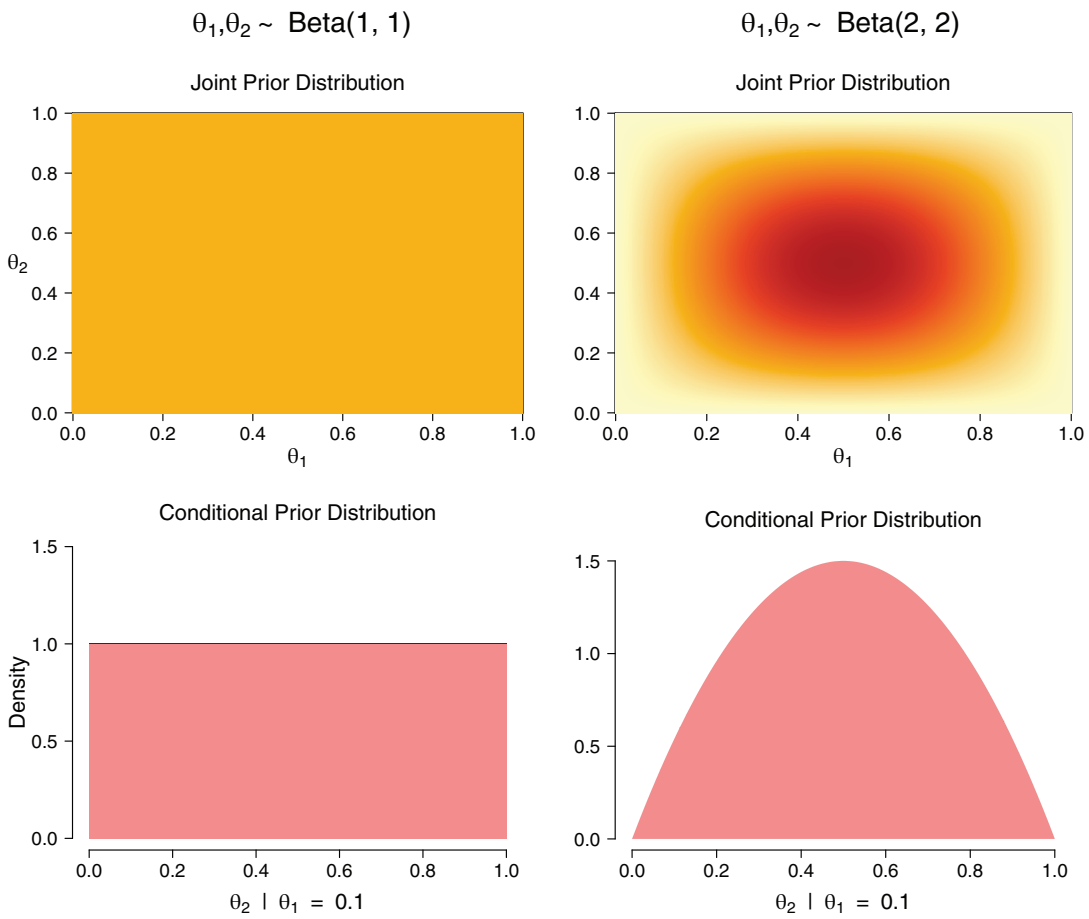

FIGURE 1 Top: Joint prior distribution assigned to (θ_1, θ_2) under $a = 1$ (left) and $a = 2$ (right). Bottom: Conditional prior distribution of θ_2 given that $\theta_1 = 0.10$

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

concentrated around $(\theta_1, \theta_2) = (\frac{1}{2}, \frac{1}{2})$. The bottom panels visualize the conditional prior distribution of θ_2 given that we know that $\theta_1 = 0.10$. Knowing the value of θ_1 does not change our prior about θ_2, which follows from the assumption of prior independence.

The IB Bayes factor is available in analytic form.[9,12,16] In the literature on contingency tables, our setup corresponds to an independent multinomial sampling scheme where the row (or column) sums of the contingency table (here n_1 and n_2) are fixed; for an extension to other sampling schemes and more than two groups, see References 12 and 13. Gunel and Dickey suggest $a = 1$ as a default value.[12] Note that values $a < 1$ lead to an undefined prior density for $\theta = 0$ and $\theta = 1$, which implies model selection inconsistency in case θ is indeed 1 or 0. Hence these values should be avoided on principle grounds. Because the beta prior is conjugate for the binomial likelihood, the posterior distributions of θ_1 and θ_2 are again (independent) beta distributions.

2.2 | The logit transformation (LT) approach

The test proposed by Kass and Vaidyanathan[14] and implemented by Gronau et al[15] does not assign a prior directly to (θ_1, θ_2), but applies a logit transformation and assigns priors to the transformed parameters (β, ψ). Specifically, we write:

$$\log\left(\frac{\theta_1}{1-\theta_1}\right) = \beta - \frac{\psi}{2},$$
$$\log\left(\frac{\theta_2}{1-\theta_2}\right) = \beta + \frac{\psi}{2},$$

where β is a grand mean and ψ is the difference in log odds (ie, the log odds ratio):

$$\beta = \frac{1}{2}\left(\log\left(\frac{\theta_1}{1-\theta_1}\right) + \log\left(\frac{\theta_2}{1-\theta_2}\right)\right),$$
$$\psi = \log\left(\frac{\theta_2}{1-\theta_2}\right) - \log\left(\frac{\theta_1}{1-\theta_1}\right).$$

While this is a more involved reparameterization than in the IB approach above, another way to formulate this setup is by writing:

$$\theta_1 = \frac{e^{\beta - \frac{\psi}{2}}}{1 + e^{\beta - \frac{\psi}{2}}},$$
$$\theta_2 = \frac{e^{\beta + \frac{\psi}{2}}}{1 + e^{\beta + \frac{\psi}{2}}},$$

which readers familiar with logistic regression may recognize. Using this setup, we test the hypotheses:

$$\mathcal{H}_0 : \psi = 0,$$
$$\mathcal{H}_1 : \psi \neq 0.$$

In contrast to above, we now assign priors to β and ψ rather than to θ_1 and θ_2 directly. In particular, under both hypotheses we assume $\beta \sim \mathcal{N}(0, \sigma_\beta)$ with $\sigma_\beta = 1$. Under $\mathcal{H}_1$ we assume $\psi \sim \mathcal{N}(0, \sigma_\psi)$.* The top left panel in Figure 2 visualizes the implied joint prior distribution on (θ_1, θ_2) under $\mathcal{H}_1$ for $\sigma_\psi = 1$. The prior mass is concentrated along the diagonal, which indicates that θ_1 and θ_2 are dependent. The bottom left panel illustrates this fact: if we know that $\theta_1 = 0.10$, then the prior on θ_2 shifts toward this value. Setting $\sigma_\psi = 2$ removes the prior dependency, as the right column in Figure 2 shows. For values of $\sigma_\psi > 2$, θ_1 and θ_2 become *anti-correlated* and hence observing a small value of θ_1 results in a prior that puts more mass on large values for θ_2. Such an inverse relation is undesirable in almost all empirical applications, and so values $\sigma_\psi > 2$ are therefore to be avoided. Gronau et al[15] developed software to compute the Bayes factor using this prior specification, first proposed by Kass and Vaidyanathan,[14] suggesting $\sigma_\psi = 1$ as a default value.

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

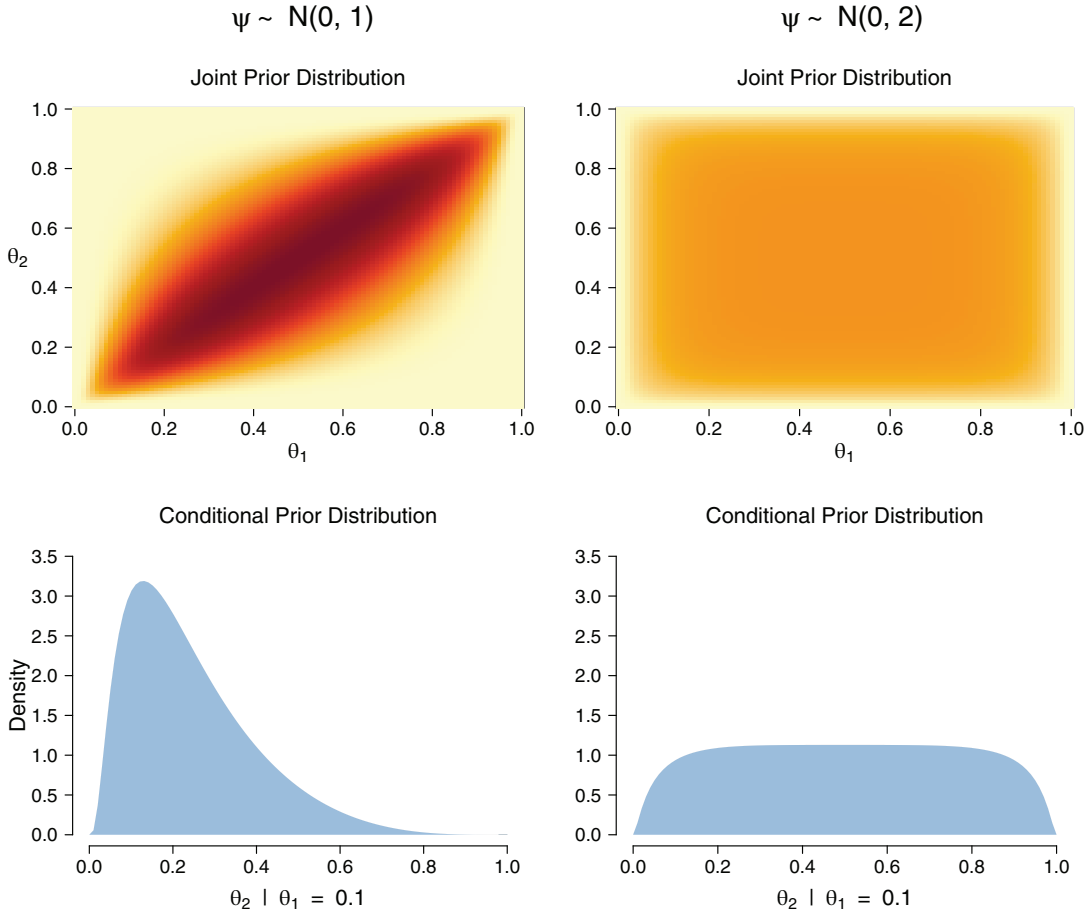

FIGURE 2 Top: Joint prior assigned to (θ_1, θ_2) for $\sigma_\psi = 1$ (left) and $\sigma_\psi = 2$ (right). Bottom: Conditional prior distribution of θ_2 given that $\theta_1 = 0.10$. In both cases we assume $\sigma_\beta = 1$

2.3 | Comparison of priors

A direct comparison of the two prior specifications may be helpful to get further intuition for their differences. While the IB approach does not assign a prior distribution to (β, ψ) or (η, ζ) explicitly, assigning a prior to θ_1 and θ_2 induces a prior distribution on these quantities. Conversely, the LT approach assigns a prior to ψ and β and this induces a prior on (θ_1, θ_2) and (η, ζ). The induced prior distributions under both approaches are nonstandard (see Appendix A), but their densities can be calculated numerically.

The top left panel in Figure 3 shows the prior distribution assigned to η by the LT approach for $\sigma_\psi = 1$ (shaded blue) and $\sigma_\psi = 2$ (striped blue) and the IB approach for $a = 1$ (shaded red) and $a = 2$ (striped red) under $\mathcal{H}_1$. Similarly, the top right panel shows the prior distribution assigned to ψ for the two approaches and prior parameter values. The (default) IB approach assigns comparatively more mass to large values of η and ψ, which in practice means that it expects larger differences between the sample proportions. The bottom panel shows the marginal priors for θ_1 and θ_2, where we find that the LT approach assigns comparatively less mass to extreme values. The LT approach cannot result in a uniform distribution on the proportions under $\mathcal{H}_0$ because of the Gaussian prior on β. If it instead would assign a (standard) logistic prior to β (which has fatter tails), the prior on the proportions would be uniform; see a related discussion in Appendix B. In the next section, we discuss a somewhat surprising difference between these two tests.

3 | PRACTICAL IMPLICATIONS OF THE PRIOR SETUP

To see the implications of the two different prior specifications in practice, in Section 3.1 we reanalyze 39 null results published in the *New England Journal of Medicine*, previously analyzed by Hoekstra et al[17] using the IB approach. We

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

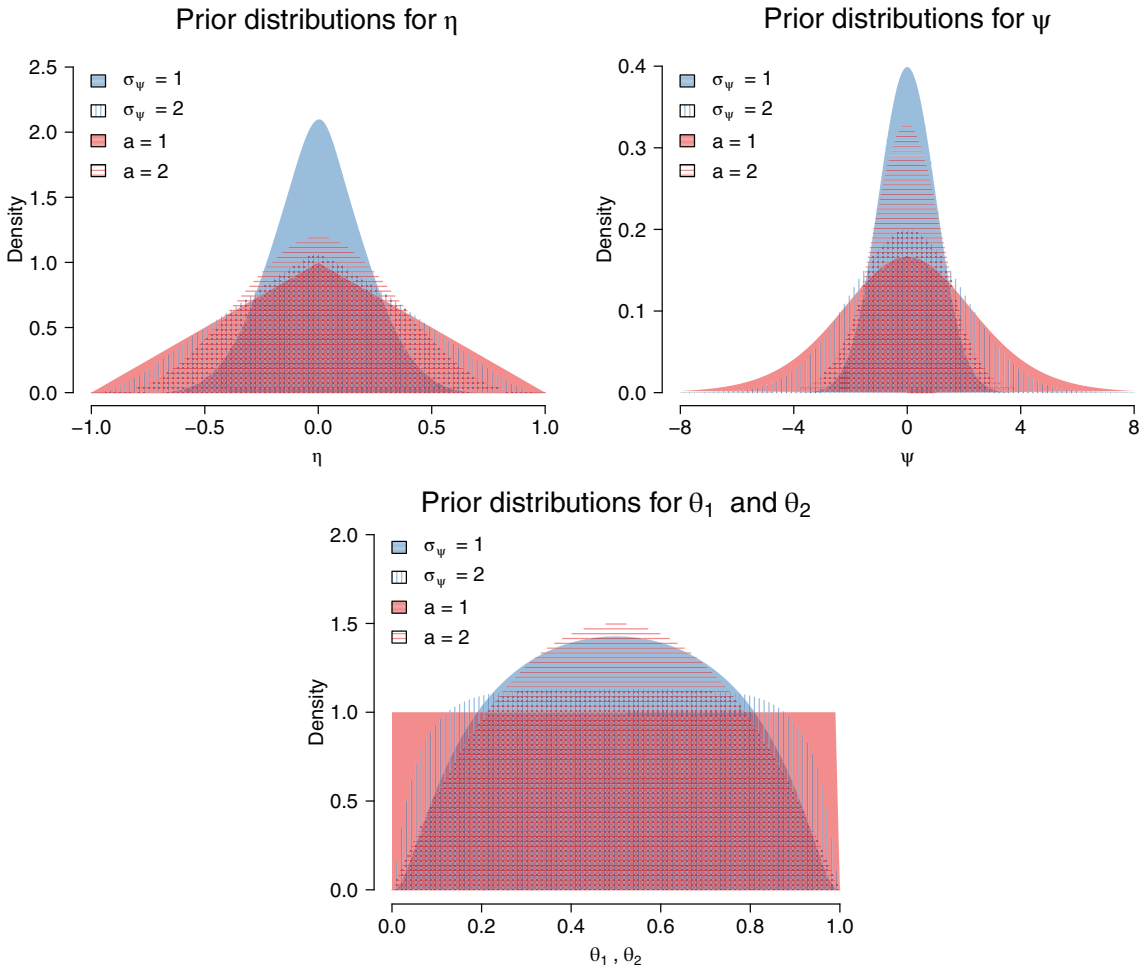

FIGURE 3 Top: Prior distributions assigned to η (left) and ψ (right) under the LT (blue, vertical lines) and IB approach (red, horizontal lines), respectively. Bottom: Marginal prior distribution of θ_1 and θ_2 under the two approaches. The density that is filled and has the highest peak corresponds to $\sigma_\psi = 1$

then explain why these difference occur in Section 3.2. The data and code to reproduce all analyses and figures is available from https://github.com/fdabl/Proportion-Puzzle.

3.1 | Reanalysis of *New England Journal of Medicine* studies

Hoekstra et al considered all 207 articles published in the *New England Journal of Medicine* in 2015.[17] The abstract of 45 of these articles contained a claim about the absence or nonsignificance of an effect for a primary outcome measure, and 37 of those allowed for a comparison of proportions, reporting 43 null results in total. We focus on those results that can be reanalyzed using a test between two proportions, which results in a total of 39 tests from 32 articles. The top left panel in Figure 4 contrasts Bayes factors in favor of $\mathcal{H}_0$ computed using the IB approach (rectangles) across $a \in [1, 5]$ with Bayes factors computed using the LT approach (circles) across $\sigma_\psi \in [1, 2]$. In virtually all cases and across specifications, the Bayes factor in favor of $\mathcal{H}_0$ is higher under the IB approach, and this difference is frequently substantial.[†] As the parameter a is increased under the IB approach, the expected difference between the two groups is smaller (see top left panel in Figure 3). Therefore, the predictions under $\mathcal{H}_1$ become more similar to the predictions under $\mathcal{H}_0$, and the Bayes factor decreases. Conversely, as σ_ψ is increased under the LT approach, the expected difference between the two group increases, and the Bayes factor in favor of $\mathcal{H}_0$ increases.

The top left panel in Figure 4 shows that, for some studies, the IB and LT Bayes factors cannot be brought into the vicinity of each other by changing the prior parameters in the way specified above, while for other studies the Bayes factors

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

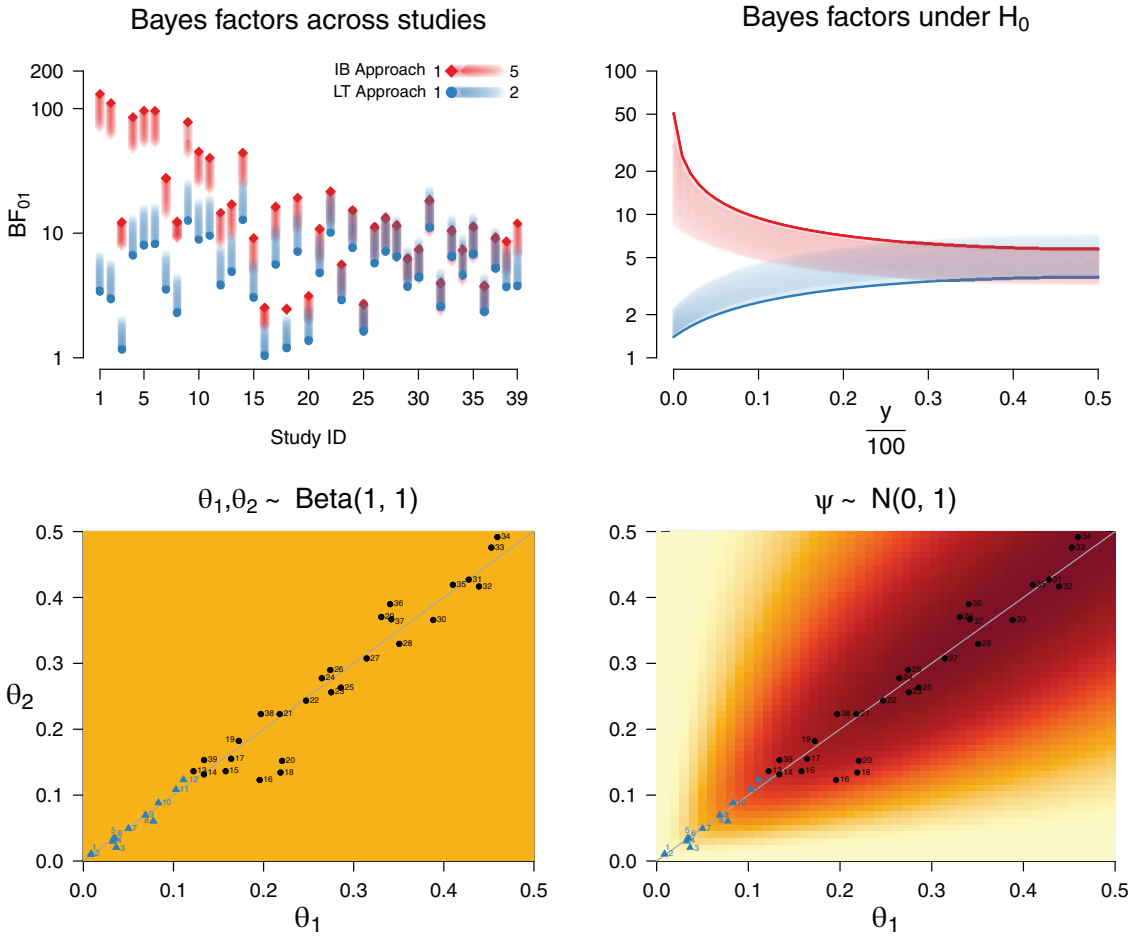

FIGURE 4 Top: Bayes factors using the IB (rectangles) and LT approach (circles) in favor of $\mathcal{H}_0$ across studies reported in Hoekstra et al[17] (left) or for simulated equal proportions with $n = 100$ (right) for values $a \in [1, 5]$ and $\sigma_\psi \in [1, 2]$ with $\sigma_\beta = 1$. Bottom: Joint prior distribution of (θ_1, θ_2) under the IB approach with $a = 1$ (left) and under the LT approach with $\sigma_\psi = 1$ and $\sigma_\beta = 1$ (right). Black dots and blue rectangles indicate the maximum likelihood estimates of the proportions in the studies analyzed by Hoekstra et al.

do overlap substantially. The studies without overlap are those indexed as 1 to 12. They are shown as blue rectangles in the bottom panels in Figure 4, which shows the joint prior density for the IB (left) and the LT approach (right) with the symbols indicating the maximum likelihood estimates for the individual studies. Note that the larger the proportions in the bottom panels, the more likely are the two Bayes factors to overlap in the top left panel.

While Hoekstra et al are reassured by the fact that their Bayesian reanalysis (using the IB approach) yields strong evidence in favor of $\mathcal{H}_0$ on average (given that all studies reported a nonsignificant P-value), using the LT approach yields a more uncertain picture. While the median Bayes factor across the studies under the IB approach is 12.30, the median Bayes factor under the LT approach is only 4.79. This difference is driven by the extremes, where the two approaches suggest substantially different conclusions. We have already seen one example in the introduction concerning hypertension and pregnancy loss (analysis identifier 3). Another example is given by Joura et al,[18] which compared the efficacy of the 9vHPV against the qHPV vaccine for preventing, among others, cervical, vulvar, and varginal disease and persistent human papillomavirus related infections in women. Comparing different outcome measures in women who were HPV-negative and HPV-positive at baseline, the IB approach yields Bayes factors consistently around 100 while the LT approach never shows Bayes factors larger than 10 (see analysis identifiers 1, 2, 5, and 6). The IB approach thus strongly suggests that the two vaccines have a similar efficacy, while the LT approach suggests that more data is needed to reach a firm conclusion.

The initially diverging and then converging pattern of the Bayes factors can be seen neatly in the top right panel in Figure 4, which shows the Bayes factors under the null hypothesis for increasing values of $\frac{y}{100} \equiv \frac{y_1}{100} = \frac{y_2}{100}$. Note that the

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

pattern is symmetric, so that for $\frac{y}{100} \in [\frac{1}{2}, 1]$ the same holds, just mirrored. At the extremes $\frac{y}{n} \in \{0, 1\}$, the IB approach yields a Bayes factor in favor of $\mathcal{H}_0$ of 50.75, while the LT Bayes factor gives a mere 1.40. The Bayes factor *decreases* as $\frac{y}{n} \to \frac{1}{2}$ under the IB approach, but *increases* under the LT approach. Consequently, the difference between the Bayes factors becomes less pronounced as we move to more central proportions, which matches the observation in the empirical analysis shown in the top left panel in Figure 4. For $\frac{y}{n} = \frac{1}{2}$, the IB Bayes factor yields 5.70 and the LT Bayes factor yields 3.67.

In summary, we have found that the two approaches give a very different answer to the question "is observing two equally small proportions strong or weak evidence in favor of the null hypothesis?". The IB Bayes factor suggests that the evidence is often orders of magnitude larger than the LT Bayes factor.

3.2 | Explaining the difference

Why does the IB Bayes factor yield so much stronger evidence than the LT Bayes factor for $\mathcal{H}_0$ when the proportions are small and of roughly equal size? And why does the IB Bayes factor *decrease* as the proportions get closer to $\frac{1}{2}$ while the LT Bayes factor *increases*? To answer these questions, we need to zoom in on the differences in the respective prior specifications. The two approaches differ in two key ways. First, while the IB approach assigns *independent* priors to θ_1 and θ_2, the LT approach assigns *dependent* priors to them. Second, the LT approach employs a logit transformation while the IB approach does not. Appendix D shows that it is not the prior dependence that underlies the difference between the two approaches. Instead, as we will see below, it is the logit transformation.

The fact that the IB Bayes factor is larger than the LT Bayes factor even when the latter approach expects a larger difference in the proportions (eg, a larger η or ψ, compare $a = 2$ with $\sigma_\psi = 2$ in Figure 3) means that focusing only on the difference parameter is not sufficient to explain the difference between the Bayes factors. Instead, we turn to a sequential predictive perspective. From such a perspective, we may first use the prior to predict the data from group one, update the prior to a posterior, and then predict the data from group two. One can rewrite the marginal likelihoods to make this sequential perspective apparent, see Appendix E. This perspective shows that there is a crucial difference in the predictions that the CT and LT approach make under $\mathcal{H}_1$. Under both approaches there is a common parameter θ under $\mathcal{H}_0$, and the prior assigned to θ gets updated by data from the first group, after which predictions about the second group are made. Under the alternative hypothesis, however, the IB approach implies that observing data from group one does not update our beliefs about likely values of θ_2, and so data from group one cannot inform the subsequent prediction about group two. Under the LT approach, such information sharing does take place. This difference is shown in the top panels in Figure 5, which visualizes the joint prior distribution for θ_1 and $\eta \equiv \theta_2 - \theta_1$ under the IB (left) and LT (right) approaches. We see that under the IB approach, learning about θ_1 does not influence our predictions about likely differences between the two groups. In contrast, under the LT approach we find that when we learn that θ_1 is either small or large, we expect small differences between the proportions compared to when we learn that θ_1 is about half. This explains why the LT Bayes factor in favor of $\mathcal{H}_0$ *increases* as we move from extreme values of $\frac{y}{n}$ toward values around $\frac{1}{2}$, as shown in the top right panel in Figure 4: at the extremes, $\mathcal{H}_1$ expects smaller differences between θ_2 and θ_1, and hence it more closely resembles $\mathcal{H}_0$, leading to a more equivocal Bayes factor. Moving toward more central values, $\mathcal{H}_1$ expects larger differences and hence $\mathcal{H}_0$ outpredicts it by a larger margin.

While prior dependence is necessary to change one's beliefs about likely differences between the proportions upon learning θ_1, the fact that the expected difference *decreases* in the LT approach is due to the logit transformation. When the two rates (θ_1, θ_2) are small, say $\theta_1 = 0.05$ and $\theta_2 = 0.10$, the log odds difference ψ is large, in this case $\psi = 0.75$. The prior on ψ in the LT approach renders such large log odds differences to be unlikely, which means that a smaller rate difference is expected. Conversely, if the rates are somewhere in the center, say $\theta_1 = 0.50$ and $\theta_2 = 0.55$, then the log odds difference ψ is small, in this case $\psi = 0.20$. The prior on ψ in the LT approach views such small log odds differences as likely; in fact, much larger ones are possible. Therefore, the LT approach expects larger rate differences when the rates are in the center. On the log odds scale, the expected difference does not change markedly as θ_1 varies, as the bottom right panel in Figure 5 shows.[‡]

In contrast to the LT approach, the IB approach does not automatically reduce the expectations about likely differences as we learn that θ_1 takes on extreme values. Taking slices at particular values for θ in the top left panel in Figure 5 would show uniform distributions ranging from $-\theta_1$ to $1 - \theta_1$. Instead of reducing the size of an expected difference at the extremes, the IB approach actually *amplifies* it. This occurs because of the boundedness of the rate parameters. In particular, if $\theta_1 = 0$, then the only values for which, say, $|\eta| < 0.10$ are $\theta_2 \in [0, 0.10]$. Differences in absolute magnitude

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

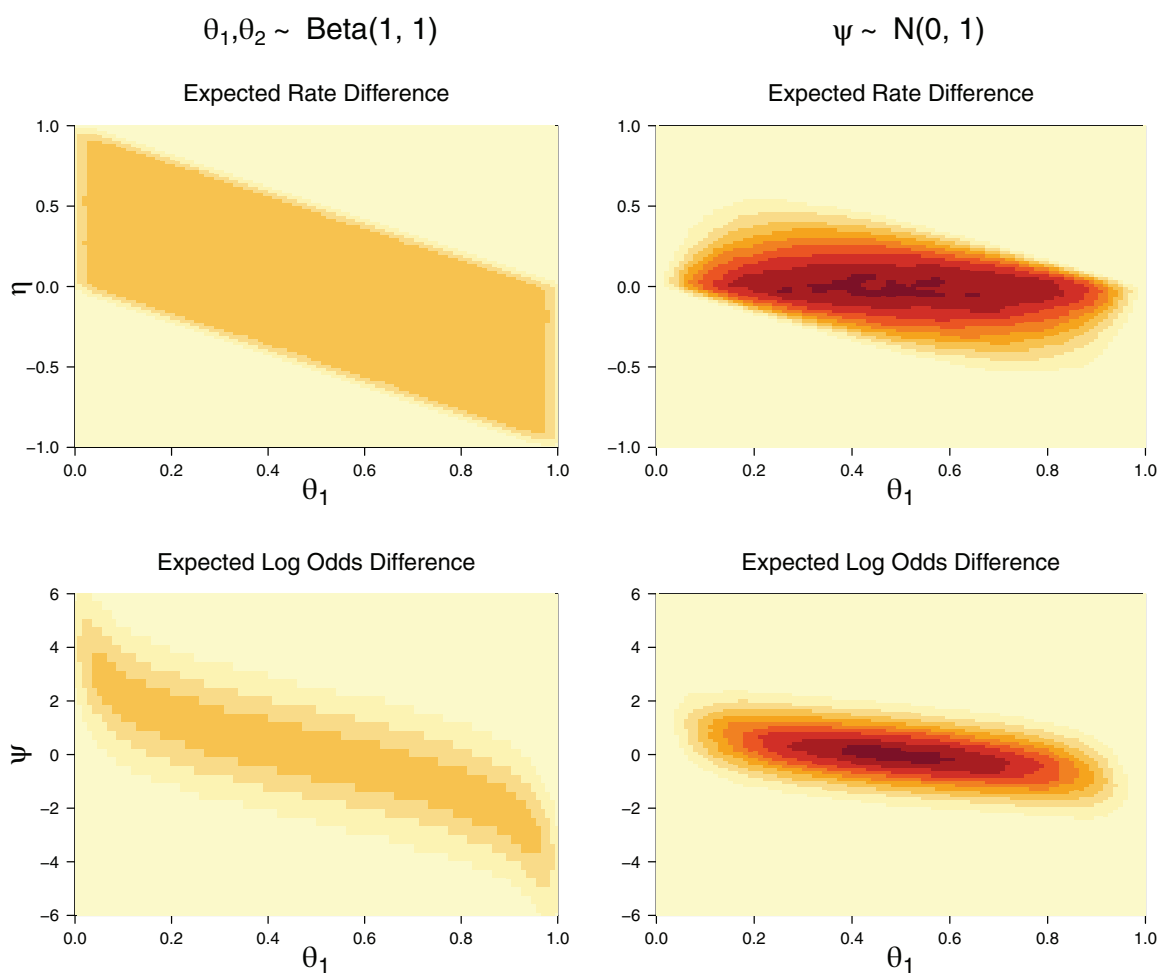

FIGURE 5 Top: Joint prior distribution for (θ_1, η) under the IB approach with $a = 1$ (left) and under the LT approach with $\sigma_\psi = 1$ and $\sigma_\beta = 1$ (right). Bottom: Joint prior distribution for (θ_1, ψ) under the two approaches

between the two rates of up to 0.10 if $\theta_1 = 0$ are therefore assigned a probability of 0.10. If, however, $\theta_1 = 0.50$, the values of θ_2 for which $|\eta| < 0.10$ are $\theta_2 \in [0.40, 0.60]$. This means that differences in absolute magnitude between the two rates of up to 0.10 if $\theta_1 = 0.50$ are assigned a probability of 0.20, or twice as much as in the case that $\theta_1 = 0$. Another way to see this is to note that the expected difference in the case that $\theta_1 = 0$ is $\eta = 0.50$, while the expected difference in the case that $\theta_1 = 0.50$ is $\eta = 0$; more generally, the expected difference increases linearly as we move to the extremes, as can be seen in the top left panel in Figure 5. The bottom left panel in Figure 5 shows an associated and pronounced increase in the expected log odds difference as θ_1 moves toward extreme values.

4 | DISCUSSION

Being able to discriminate between evidence of absence and absence of evidence is key in medicine and public health. The Bayes factor, which pits the predictive performance of two hypotheses against each other, is a method for doing so. Reanalyzing 39 null results published in the *New England Journal of Medicine*, we have found that the strength of evidence depends crucially on the prior specification. Comparing an approach that assigns independent beta distributions directly to the rate parameters (IB) to one that employs a logit transformation (LT), we have found that the former approach suggests evidence that is orders of magnitudes larger than the latter approach when observing small proportions of roughly equal size. Consider again the effect of tightly or less-tightly controlling hypertension in pregnant women.[3] The IB approach suggests a Bayes factor about ten times larger in favor of the null hypothesis than the LT approach. Similarly, the Bayes factor in favor of the hypothesis that two vaccines against the human papillomavirus are equally effective[18] is

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

up to 38 times larger under the IB approach, depending on the outcome measure. While the original statistical analysis by the authors is more involved (eg, adjusting for various covariates), these differences are striking. As we have seen, they occur because of the boundedness of the parameter space, which leads to the expectations of large differences at the extremes under the IB approach. In contrast, the LT approach expects smaller differences at the extremes.

Although the two Bayes factors frequently differed markedly in size, they always *both* provided evidence in favor of $\mathcal{H}_0$ in the cases we studied in this paper. This need not be the case, however. For example, a large-scale investigation into the effect of aspirin found that 26 out of $n_1 = 11{,}034$ who received a placebo suffered a fatal heart attack, but only 10 out of $n_2 = 11{,}037$ who received aspirin did.[19] The IB Bayes factor yields strong evidence in favor of $\mathcal{H}_0$ ($\text{BF}_{01} = 19.78$) while the LT Bayes factor finds moderate evidence in favor of $\mathcal{H}_1$ ($\text{BF}_{10} = 5.36$). This is again owed to the fact that the IB approach expects larger differences than the LR approach; a small effect size is more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$ in the IB approach simply because under $\mathcal{H}_1$ unrealistically large effect sizes are assumed to be plausible.

Our results have a number of implications. First, researchers should think carefully about their prior setup. This not only implies thinking about the parameter values for the priors, but also the nature of the priors and what they imply for the data. As we have seen, one may be fooled by assessing how sensitive the Bayes factor is by varying only prior parameters conditional on the model. For all studies we reanalyzed that had proportions at the extremes, varying the prior parameters of the IB approach still resulted in substantial evidence in favor of the null hypothesis. However, changing the prior setup to the LT approach, we found markedly reduced evidence. If there are sensible alternative parameterizations, it may be prudent to explore how sensitive the results are to these. In particular, our exposition suggests that assessing the predictions that follow from the model can help get a better intuition of its assumptions. In our case, this made clear that the IB approach expects larger rate differences at the boundaries of the parameter space compared to the center, while the LT approach expects the reverse. We were struck by the initially puzzling contrast between these two tests, given that the task—comparing two proportions—seemed so simple. This suggests that even more caution needs to be applied when using Bayes factors for comparing models that are invariably more complex. In the same spirit of sensitivity analysis, it is prudent to compare inferences based on the Bayes factor with inferences based on the posterior distributions of the parameters. While testing precise hypotheses based on credible intervals is arguably inappropriate since it makes inference conditional on $\mathcal{H}_1$, ignoring $\mathcal{H}_0$,[20,21] alarm bells should go off when they show stark differences to the Bayes factor results, as in our case (compare Figure 4 and C1). Researchers who rely only on the posterior distribution for inference or for trial design are unaffected by our results.

The second implication of our work is that the LT approach appears to be better suited for testing the equality of two proportions than the currently more popular IB approach. First, assuming prior dependence strikes us as a more sensible approach especially for medicine and public health. This is because it is generally unlikely that gaining information about one group does not influence our beliefs about the other group. For example, if a particular treatment for cancer yields a 30% success rate, surely this should inform our expectations for the success rate of a comparison treatment. Incorporating this dependency results in more cautious predictions, which reduces the eagerness with which the standard IB approach suggests strong evidence in favor of $\mathcal{H}_0$ (see also Appendix D). While we have focused on testing the sharp null hypothesis $\theta_1 = \theta_2$, the difference between dependent and independent priors is important also when testing other hypotheses. Howard[22] discusses the difference between a dependent and independent prior in the context of comparing $\theta_1 < \theta_2$ against $\theta_1 > \theta_2$, and similarly recommends dependent priors, as independent priors lead to a test that is not "sufficiently cautious".

Most importantly, however, is the fact that the LT approach avoids the issues associated with a bounded parameter space. In particular, the IB approach is not sensitive to the fact that the implications of a rate difference $\theta_2 - \theta_1$ depends on the actual values of the rates. For example, $\theta_2 = 0.10$ and $\theta_1 = 0.05$ and $\theta_2 = 0.55$ and $\theta_1 = 0.50$ imply the same rate difference, but while the former represents a 100% increase, the latter represents only a 10% increase. Yet the IB approach assigns more prior mass to large rate differences at the extremes of the parameter space compared to its center. In contrast, the reverse holds for the LT approach because of the logit transformation. This strikes us as more sensible. With the recent development of accessible software that makes the LT approach easily available,[15,23] we hope that the use of the LT approach for testing the equality of two proportions becomes more widespread.

To sum up, we have seen an initially puzzling divergence in evidence when reanalyzing 39 null results from the *New England Journal of New Medicine* with two different Bayes factor tests. We have explained this divergence and sided with the approach that employs a logit transformation and assumes prior dependence between the two proportions. We suggest that this approach, rather than the one motivated from a contingency table perspective, should be routinely applied when testing the equality of two proportions. Our journey should also act as a cautionary tale for users of Bayes factors, stressing the importance of assessing the predictions of one's models, conducting thorough sensitivity analyses, and never relying

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

on just a single quantity. We have demonstrated the usefulness of this in the simple case of testing the equality of two proportions. Most applications will arguably be more complex, however, increasing the potential for puzzles and stressing the need for a holistic evidence assessment.

AUTHOR CONTRIBUTIONS

Eric-Jan Wagenmakers proposed the study. Karoline Huth, Quentin F. Gronau, Alexander Etz, and Eric-Jan Wagenmakers worked on the problem initially. After an interim period, Fabian Dablander and Eric-Jan Wagenmakers resumed the work and solved the puzzle. Fabian Dablander and Karoline Huth analyzed the data. Fabian Dablander wrote the manuscript. All authors read, provided comments, and approved the submitted version of the paper.

ACKNOWLEDGEMENT

The authors thank Rink Hoekstra, Rei Monden, and Don van Ravenzwaaij for sharing the *New England Journal of Medicine* data with us.

CONFLICTS OF INTEREST

The authors declare that there were no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data and code to reproduce all analyses and figures is available from https://github.com/fdabl/Proportion-Puzzle.[24]

ENDNOTES

*While this prior specification assigns the log odds ratio ψ a Gaussian distribution, the IB prior specification results in a nonstandard distribution on ψ. For more details, see Appendix A.

†In contrast, the conclusions one would draw based on posterior distributions are very similar, see Appendix C.

‡The fact that there is barely any prior mass at extreme values of θ_1 under the IB approach is due to the Gaussian prior on β, as discussed in Section 2.3.

ORCID

Fabian Dablander https://orcid.org/0000-0003-2650-6491
Quentin F. Gronau https://orcid.org/0000-0001-5510-6943

REFERENCES

1. Polack FP, Thomas SJ, Kitchin N, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med*. 2020;383(27):2603-2615.
2. Plöderl M, Wagenmakers EJ, Tremblay P, et al. Suicide risk and sexual orientation: a critical review. *Arch Sex Behav*. 2013;42(5):715-727.
3. Magee LA, Von Dadelszen P, Rey E, et al. Less-tight versus tight control of hypertension in pregnancy. *N Engl J Med*. 2015;372(5):407-417.
4. Jeffreys H. Some tests of significance, treated by the theory of probability. *Proc Camb Philos Soc*. 1935;31(2):203-222.
5. Ly A, Verhagen J, Wagenmakers EJ. Harold Jeffreys's default Bayes factor hypothesis tests: explanation, extension, and application in psychology. *J Math Psychol*. 2016;72:19-32.
6. Morey RD, Romeijn JW, Rouder JN. The philosophy of Bayes factors and the quantification of statistical evidence. *J Math Psychol*. 2016;72:6-18.
7. Keysers C, Gazzola V, Wagenmakers EJ. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nat Neurosci*. 2020;23(7):788-799.
8. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995;90(430):773-795.
9. Jeffreys H. *Theory of Probability*. 1st ed. Oxford, UK: Oxford University Press; 1939.
10. Wasserman L. Bayesian model selection and model averaging. *J Math Psychol*. 2000;44(1):92-107.
11. Lee MD, Wagenmakers EJ. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, UK: Cambridge University Press; 2013.
12. Gunel E, Dickey J. Bayes factors for independence in contingency tables. *Biometrika*. 1974;61(3):545-557.
13. Jamil T, Ly A, Morey RD, Love J, Marsman M, Wagenmakers EJ. Default "Gunel and Dickey" Bayes factors for contingency tables. *Behav Res Methods*. 2017;49(2):638-652.
14. Kass RE, Vaidyanathan SK. Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J Royal Stat Soc Ser B (Methodol)*. 1992;54(1):129-144.
15. Gronau QF, Raj A, Wagenmakers EJ. Informed Bayesian inference for the A/B test. *J Stat Softw*. 2021;100(17):1-39.
16. Dickey JM, Lientz B. The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Ann Math Stat*. 1970;41(1):214-226.

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

17. Hoekstra R, Monden R, van Ravenzwaaij D, Wagenmakers EJ. Bayesian reanalysis of null results reported in medicine: strong yet variable evidence for the absence of treatment effects. *PLoS One*. 2018;13(4):e0195474.
18. Joura EA, Giuliano AR, Iversen OE, et al. A 9-valent HPV vaccine against infection and intraepithelial neoplasia in women. *N Engl J Med*. 2015;372(8):711-723.
19. Physicians' Health Study Research Group SCot. Final report on the aspirin component of the ongoing Physicians' health study. *N Engl J Med*. 1989;321(3):129-135.
20. Berger JO. Bayes factors. In: Kotz S, Balakrishnan N, Read C, Vidakovic B, Johnson NL, eds. *Encyclopedia of Statistical Sciences*. Vol 1. 2nd ed. Hoboken, NJ: Wiley; 2006:378-386.
21. Wagenmakers EJ, Lee MD, Rouder JN, Morey RD. The principle of predictive irrelevance or why intervals should not be used for model comparison featuring a point null hypothesis. In: Gruber CW, ed. *The Theory of Statistics in Psychology – Applications, Use and Misunderstandings*. Cham, Switzerland: Springer; 2020:111-129.
22. Howard J. The 2 × 2 table: a discussion from a Bayesian viewpoint. *Stat Sci*. 1998;13(4):351-367.
23. Hoffmann T, Wagenmakers EJ. Bayesian inference for the A/B test: example applications with R and JASP; 2021. doi: 10.31234/osf.io/z64th
24. Dablander F. A puzzle of proportions: data and code; 2021. 10.17605/OSF.IO/3VUX5
25. Pham-Gia T, Turkkan N, Eng P. Bayesian analysis of the difference of two proportions. *Commun Stat Theory Methods*. 1993;22(6):1755-1771.
26. Ojo MO. A remark on the convolution of the generalized logistic random variables. *ASSET Serves A*. 2003;1(2):1-5.
27. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. *Stat Sci*. 1999;14(4):382-401.
28. Hinne M, Gronau QF, van den Bergh D, Wagenmakers EJ. A conceptual introduction to Bayesian model averaging. *Adv Methods Pract Psychol Sci*. 2020;3(2):200-215.

How to cite this article: Dablander F, Huth K, Gronau QF, Etz A, Wagenmakers E-J. A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions. *Statistics in Medicine*. 2022;41(8):1319-1333. doi: 10.1002/sim.9278

APPENDIX A. INDUCED PRIORS

Under the LT approach, the marginal priors for θ_1 and θ_2 as well as η are given in integral representation by Gronau et al.[15] Under the IB approach, the density of η is given by:

$$f(\eta; a) = \begin{cases} \frac{1}{B(a,a)} \eta^{2a-1}(1-\eta)^{2a-1} F_1\left(a, 4a-2, 1-a; 2a; 1-\eta, 1-\eta^2\right) & 0 < \eta \leq 1 \\ \frac{1}{B(a,a)} (-\eta)^{2a-1}(1+\eta)^{2a-1} F_1\left(a, 1-a, 4a-2; 2a; 1-\eta^2, 1+\eta\right) & -1 \leq \eta < 0 \end{cases},$$

see Pham et al.[25] In our case, we always have that $2a > 1$ and so $f(0) = \frac{B(2a-1,2a-1)}{B(a,a)^2}$.[25] We can derive the density for ψ under prior independence for the special case of $a = 1$. Note that, because $\theta \sim \text{Beta}(1, 1)$, it follows that:

$$\log\left(\frac{\theta}{1-\theta}\right) \sim \text{Logistic}(0, 1).$$

Recall that:

$$\psi = \log\left(\frac{\theta_2}{1-\theta_2}\right) - \log\left(\frac{\theta_1}{1-\theta_1}\right).$$

Ojo (2003)[26] gives the distribution function of the sum of n i.i.d. logistic random variables. Since the logistic distribution is symmetric, we have that if $X \sim \text{Logistic}$ then $-X \sim \text{Logistic}$. Hence we can use the results by Ojo (2003) and write the density function of ψ as:

$$f(\psi; a = 1) = \frac{e^{\psi}(e^{\psi}(\psi - 2) + \psi + 2)}{(e^{\psi} - 1)^3}.$$

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX B. COMBINING APPROACHES THROUGH MODEL-AVERAGING

If one is uncertain as to which test one should employ, an approach would be to combine both through model-averaging.[27,28] Let $\mathcal{M}_0^{\text{IB}}$ and $\mathcal{M}_0^{\text{LT}}$ denote the null models under the IB and the LT approach, respectively, and let $\mathcal{M}_1^{\text{IB}}$ and $\mathcal{M}_1^{\text{LT}}$ denote the respective alternative models. The model-averaged Bayes factor is given by:

$$\text{BF}_{01}^{\text{AVG}} = \frac{p(\mathcal{D}|\mathcal{M}_0^{\text{IB}})\pi(\mathcal{M}_0^{\text{IB}}) + p(\mathcal{D}|\mathcal{M}_0^{\text{LT}})\pi(\mathcal{M}_0^{\text{LT}})}{p(\mathcal{D}|\mathcal{M}_1^{\text{IB}})\pi(\mathcal{M}_1^{\text{IB}}) + p(\mathcal{D}|\mathcal{M}_1^{\text{LT}}), \pi(\mathcal{M}_1^{\text{LT}})},$$

where $\pi()$ gives the prior probability of the particular model. One has to be careful when model-averaging. While the nuisance parameters generally do not matter when testing nested models,[5,14] they do here. In particular, the IB approach assigns β a logistic distribution, while the LT approach assigns β a Gaussian distribution. The logistic distribution has fatter tails, and thus puts more mass on extreme values of θ than the Gaussian distribution. In our case, this can result in somewhat surprising results. Let $\mathcal{D}_1 = (y_1, n_1)$ and $\mathcal{D}_2 = (y_2, n_2)$ denote data in the two groups, respectively. For data $\mathcal{D}_1 = (0, 50)$ and $\mathcal{D}_2 = (0, 50)$, for example, the IB model that assumes a difference in the population ($\mathcal{M}_1^{\text{IB}}$) outpredicts the LT null model ($\mathcal{M}_0^{\text{LT}}$) by a factor of seven! Since we want to focus on differences under $\mathcal{H}_1$ between the IB and LT approach, we would need to make the models identical under $\mathcal{H}_0$, for example by assigning β a logistic prior not only under the IB approach, but also under the LT approach.

APPENDIX C. INFERENCE BASED ON POSTERIOR DISTRIBUTIONS

The contrasting results discussed in Section 3 do not occur when inferences are based on the posterior distribution of the log odds difference ψ or the rate difference η. Figure C1 shows the posterior mean and 95% credible interval of ψ (top)

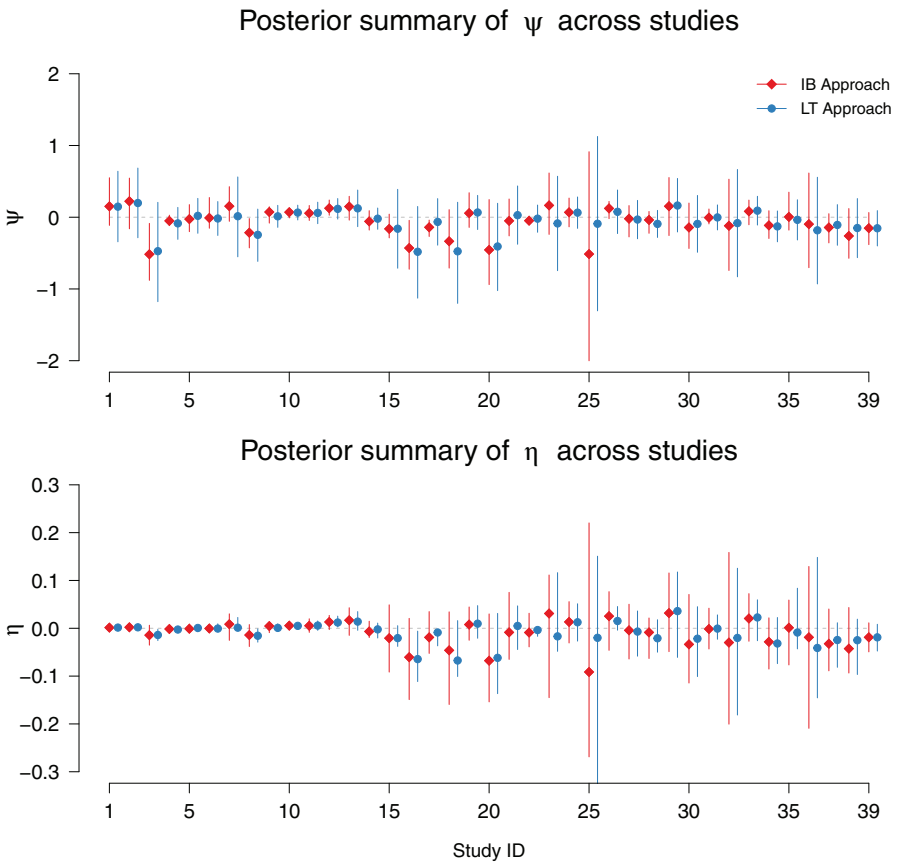

FIGURE C1 Shows the posterior mean and 95% credible interval of the log odds difference ψ (top) and the rate difference η (bottom) under the IB approach (rectangles, $a = 1$) and the LT approach (circles, $\sigma_\psi = 1$)

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and η (bottom) for the IB (rectangles) and the LT approach (circles) with the respective default parameterizations $a = 1$ and $\sigma_\psi = 1$ across all studies. The conclusions one would draw from these posterior distributions are very similar (and changing the prior parameters does not affect them). This is as expected, since the effect of the prior generally washes out as more and more data are observed, which is not the case for Bayes factors.

APPENDIX D. A DEPENDENT IB APPROACH

If prior dependence were responsible for the stark differences between the Bayes factors at the extremes, we would find that a IB approach which assumes prior dependence shows a pattern similar to the LT approach. Here, we show that this is not the case. Recall that $\eta = \theta_2 - \theta_1$ and $\zeta = \frac{1}{2}(\theta_1 + \theta_2)$. We define:

$$\theta_1 = \min\left(\max\left(\zeta - \frac{\eta}{2}, 0\right), 1\right),$$
$$\theta_2 = \min\left(\max\left(\zeta + \frac{\eta}{2}, 0\right), 1\right),$$

and assign truncated Gaussian priors to (η, ζ), that is, $\eta \sim \mathcal{N}(0, \sigma_\eta)_{\mathbb{I}(-1,1)}$ and $\zeta \sim \mathcal{N}(0, \sigma_\zeta)_{\mathbb{I}(0,1)}$. Using $\sigma_\eta = \frac{1}{5}$ and $\sigma_\zeta = \frac{1}{2}$, the left panel in Figure D1 shows the joint prior distribution for (θ_1, θ_2) from which a strong prior dependence between θ_1 and θ_2 is apparent. The right panel, however, shows the characteristic pattern of the IB approach as discussed in the main text, rather than the characteristic pattern of the LT approach. Since the rates are dependent a priori, it is not the prior dependence that is responsible for the pattern. However, employing a prior that has a stronger dependency between the two rates naturally reduces the size of their expected difference. Increasing σ_η results in a lower correlation (going from about 0.77 at $\sigma_\eta = \frac{1}{5}$ to zero at $\sigma_\eta = 1$) and a reduced Bayes factor in favor of $\mathcal{H}_0$, as the right panel in Figure D1 shows.

APPENDIX E. SEQUENTIAL PREDICTIVE PERSPECTIVE

Under $\mathcal{H}_0$ we have for both the IB (ie, independent) and the LT (ie, dependent) setup that (suppressing conditioning on $\mathcal{H}_0$):

$$p(\mathcal{D}_2, \mathcal{D}_1 | \mathcal{H}_0) = \int_\theta p(\mathcal{D}_2, \mathcal{D}_1 | \theta) \pi_0(\theta) \ \mathrm{d}\theta \tag{E1}$$
$$= \int_\theta p(\mathcal{D}_2 | \mathcal{D}_1, \theta) p(\mathcal{D}_1 | \theta) \pi_0(\theta) \ \mathrm{d}\theta \tag{E2}$$
$$= \mathcal{Z} \int_\theta p(\mathcal{D}_2 | \theta) p(\theta | \mathcal{D}_1) \ \mathrm{d}\theta, \tag{E3}$$

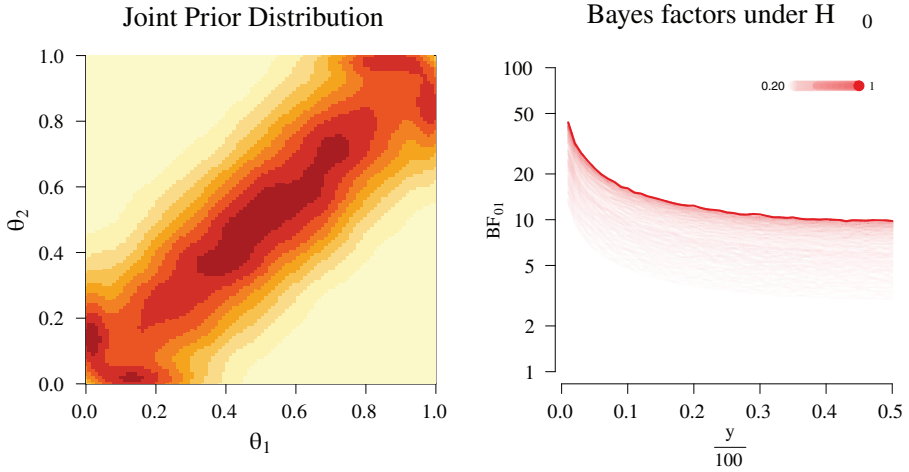

FIGURE D1 Left: Joint distribution on (θ_1, θ_2) using $\sigma_\eta = \frac{1}{5}$ and $\sigma_\zeta = \frac{1}{2}$. Right: Bayes factors in favor of $\mathcal{H}_0$ for simulated proportions with $n = 100$ for values $\sigma_\eta = \left[\frac{1}{5}, 1\right]$ and $\sigma_\zeta = \frac{1}{2}$

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $\mathcal{Z} = \int_\theta p(\mathcal{D}_1|\theta)\pi_0(\theta)\ \mathrm{d}\theta$ is the marginal likelihood for $\mathcal{D}_1$, $p(\theta|\mathcal{D}_1)$ is the posterior of θ after observing $\mathcal{D}_1$, and we can remove the conditioning on $\mathcal{D}_1$ in (2) because all relevant information is in θ. Similarly, under $\mathcal{H}_1$ for LT setup we have that:

$$p(\mathcal{D}_2, \mathcal{D}_1|\mathcal{H}_1) = \int_{\theta_2}\int_{\theta_1} p(\mathcal{D}_2, \mathcal{D}_1|\theta_1, \theta_2)\pi_1(\theta_1, \theta_2)\ \mathrm{d}\theta_1\ \mathrm{d}\theta_2 \tag{E4}$$

$$= \int_{\theta_2}\int_{\theta_1} p(\mathcal{D}_2|\mathcal{D}_1, \theta_1, \theta_2)p(\mathcal{D}_1|\theta_1, \theta_2)\pi_1(\theta_1, \theta_2)\ \mathrm{d}\theta_1\ \mathrm{d}\theta_2 \tag{E5}$$

$$= \mathcal{Z}\int_{\theta_2}\int_{\theta_1} p(\mathcal{D}_2|\theta_1, \theta_2)p(\theta_1, \theta_2|\mathcal{D}_1)\ \mathrm{d}\theta_1\ \mathrm{d}\theta_2, \tag{E6}$$

where $\mathcal{Z}$ is the marginal likelihood for $\mathcal{D}_1$, $p(\theta_1, \theta_2|\mathcal{D}_1)$ is the posterior after observing $\mathcal{D}_1$, and we can remove the conditioning on $\mathcal{D}_1$ in (2) because all relevant information is in (θ_1, θ_2). This is in contrast to the IB setup, where:

$$p(\mathcal{D}_2, \mathcal{D}_1|\mathcal{H}_1) = \int_{\theta_2}\int_{\theta_1} p(\mathcal{D}_2, \mathcal{D}_1|\theta_1, \theta_2)\pi_1(\theta_1, \theta_2)\ \mathrm{d}\theta_1\ \mathrm{d}\theta_2 \tag{E7}$$

$$= \int_{\theta_2} p(\mathcal{D}_2|\theta_2)\pi_1(\theta_2)\ \mathrm{d}\theta_2 \int_{\theta_1} p(\mathcal{D}_2|\theta_1)\pi_1(\theta_1)\ \mathrm{d}\theta_1, \tag{E8}$$

and thus no sharing of information across the two groups takes place.

10970258, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.9278 by Uva Universiteitsbibliotheek, Wiley Online Library on [02/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License