Causally Sound Priors for Binary Experiments

Nicholas J. Irons* and Carlos Cinelli†

Abstract. We introduce the BREASE framework for the Bayesian analysis of randomized controlled trials with a binary treatment and a binary outcome. Approaching the problem from a causal inference perspective, we propose parameterizing the likelihood in terms of the **b**aseline **r**isk, **e**fficacy, and **a**dverse **s**ide **e**ffects of the treatment, along with a flexible, yet intuitive and tractable jointly independent beta prior distribution on these parameters, which we show to be a generalization of the Dirichlet prior for the joint distribution of potential outcomes. Our approach has a number of desirable characteristics when compared to current mainstream alternatives: (i) it naturally induces prior dependence between expected outcomes in the treatment and control groups; (ii) as the baseline risk, efficacy and risk of adverse side effects are quantities commonly present in the clinicians' vocabulary, the hyperparameters of the prior are directly interpretable, thus facilitating the elicitation of prior knowledge and sensitivity analysis; and (iii) we provide analytical formulae for the marginal likelihood, Bayes factor, and other posterior quantities, as well as an exact posterior sampling algorithm and an accurate and fast data-augmented Gibbs sampler in cases where traditional MCMC fails. Empirical examples demonstrate the utility of our methods for estimation, hypothesis testing, and sensitivity analysis of treatment effects.

Keywords: Binomial Proportions, Potential Outcomes, Generalized Dirichlet.

1 Introduction

Randomized controlled trials (RCTs) form the cornerstone of scientific research across numerous disciplines. In their most basic form, these trials compare the occurrence of an adverse (or favorable) outcome between treatment and control groups. This is particularly evident in a drug or vaccine trial, in which the efficacy of an intervention is established by comparing the number of individuals who die or develop a disease in each arm of the study. We refer to this type of study design as a "binary experiment," wherein each participant is subjected to either a treatment or a control condition (a binary exposure), and we observe either the presence or absence of the adverse effect of interest (a binary outcome).

If participants of the trial are independent draws from a common (super-)population, statistical inference in binary experiments amounts to what is perhaps the simplest of tasks in statistics—the comparison of two binomial proportions. Indeed, from a Bayesian perspective, inference on the parameter of a binomial distribution dates back to at least as early as the origins of Bayesian inference itself, as evidenced by the seminal

*PhD Candidate, Department of Statistics, University of Washington, Seattle, USA.
Email: njirons@uw.edu; url: https://njirons.github.io.

†Assistant Professor, Department of Statistics, University of Washington, Seattle, USA.
Email: cinelli@uw.edu; url: http://carloscinelli.com.

works of Bayes (1763) and Laplace (1774). The task comprises specifying a joint prior distribution for both binomial parameters, and computing the posterior distribution (or Bayes factors) of (relevant contrasts of) such parameters (e.g., the risk difference, or the risk ratio). Yet, despite this long tradition, their widespread occurrence in the sciences, and the apparent simplicity of the inferential task, mainstream approaches for prior specification in the analysis of binary experiments have several shortcomings.

As reviewed in Agresti and Min (2005) and Dablander et al. (2022), and also evident from perusing popular textbooks (e.g., Gelman et al., 1995; Kruschke, 2014; McElreath, 2020), the two predominant approaches for the Bayesian analysis of binary experiments consist of: (i) assigning independent beta priors to each of the binomial proportions, which are conjugate priors to the (also independent) binomials comprising the likelihood; and, (ii) what is essentially a logistic regression, i.e., applying a logit transformation to the binomial proportions, and assigning Gaussian priors to the average log odds and the log odds ratio. For all their popularity, these two approaches are unsatisfactory in several ways. For example, in the first case, the assumption of prior independence of the two proportions is often not credible—e.g., in most settings, one expects that learning about the mortality rate in the control group should inform our beliefs about the mortality rate in the treatment group. Moreover, while the logit approach addresses the problem of prior dependence, it does so at the sacrifice of clarity and interpretation—odds ratios are notoriously difficult to understand (Davies et al., 1998), hindering the utility of this approach for prior elicitation and sensitivity analysis.

In this paper we demonstrate how causal logic can be used to address these challenges. Approaching the problem from a causal inference perspective, we first propose parameterizing the likelihood in terms of three clinically meaningful counterfactual quantities: the baseline risk, efficacy, and risk of adverse side effects (BREASE) of the intervention. We then propose a flexible, yet intuitive and tractable jointly independent beta prior distribution on these parameters, which we show to be a generalization of the Dirichlet prior on the joint distribution of potential outcomes. Our approach has a number of desirable characteristics: (i) it naturally induces prior dependence between the two binomial proportions of the treatment and control arms of the study; (ii) as the baseline risk, efficacy and risk of adverse side effects are quantities familiar to clinicians, the hyperparameters of the prior are directly interpretable, thus facilitating the elicitation of prior knowledge and sensitivity analysis; and (iii) we derive analytical formulae for the marginal likelihood, Bayes factor, and other posterior quantities, as well as an exact posterior sampling algorithm and an accurate and fast data-augmented Gibbs sampler in cases where traditional MCMC fails.

Related literature. The literature on Bayesian causal inference is extensive—see Li et al. (2023) for a recent review. Related to our setup are studies in the analysis of RCTs using a *traditional* Dirichlet prior on response types, such as Chickering and Pearl (1996) and Imbens and Rubin (1997), or studies using a uniform prior on the response type counts, such as Ding and Miratrix (2019). The Dirichlet prior on response types is a special case of our proposal, and our analysis not only extends it, but also clarifies when and how its use can be desirable as a way to induce causally

sound priors on the the two binomial proportions. Our study also relates to a growing body of literature investigating sensitivity and prior specification in Bayesian causal inference and analysis of experiments. In a seminal paper, Spiegelhalter et al. (1994) argued in favor of the Bayesian analysis of randomized trials with a focus on prior specification for normally distributed data. Robins and Wasserman (2012) and Linero (2023a,b) discuss the pitfalls of prior independence between the parameters governing the outcome and selection models that can yield inconsistent causal inference in high dimensional observational studies. In a similar vein, our analysis shows that—even in a low-dimensional experimental setting—causally-inspired priors encoding dependence between potential outcomes can lead to more sensible inferences than the traditional conjugate prior asserting their independence.

More generally, when framed in the language of potential outcomes, causal inference can be seen as a missing data problem. Thus, our analysis is most closely related to the literature on contingency tables with missing or incomplete observations on certain cell counts. In fact, our proposed prior can be shown to induce a *generalized* Dirichlet distribution on the joint distribution of potential outcomes. This distribution has been studied in the 1970s and 1980s (Antelman, 1972; Kaufman and King, 1973; Dickey, 1983; Dickey et al., 1987), though mostly in the context of survey sampling. Similar priors have also appeared in the analysis of diagnostic testing, such as in Branscum et al. (2005). Perhaps due to the intractability of the integrals, the difficulty in interpretation of the original generalized Dirichlet parameterization, and the missing connection to formal causal inference, this prior has received little to no attention in the analysis of binary experiments. Our analysis shows that the generalized Dirichlet distribution emerges naturally from the causal formulation of the problem, that the parameters of the distribution can be cast in intuitive clinical terms, and that statistical inference is manageable, with exact posterior sampling, efficient data-augmentation algorithms, as well as analytical formulae for Bayes factors—all of which we derive in this paper.

Outline of the paper. Section 2 introduces the statistical setup for the analysis of binary experiments and reviews existing methods for Bayesian inference in this setting. Section 3 introduces our proposal. It also derives key results for implementation, such as analytical formulae for the marginal likelihood, algorithms for posterior sampling, and an extension of the model accommodating covariates. Section 4 demonstrates the utility of our method in three empirical examples. Section 5 concludes the paper, and suggests possible extensions for future research. Code to replicate our analysis is available at `https://github.com/njirons/causally-sound`.

2 Preliminaries

In this section we set notation, the statistical setup, and briefly review the two main approaches currently used for the Bayesian analysis of binary experiments—the independent beta and logit transformation approaches. We also briefly introduce the response type parameterization of the joint distribution of potential outcomes, which is an important stepping stone for understanding our proposal.

2.1 Potential outcomes

Our analysis is situated within the potential outcomes framework of causal inference (Neyman, 1990; Rubin, 1974). Let N denote the total number of participants in the study, Z_i a binary treatment indicator and Y_i a binary outcome indicator for subject $i \in \{1, \ldots, N\}$. We denote by $Y_i(z)$ the potential outcome of subject i under the experimental condition $Z_i = z$, where $z = 0$ indicates the control and $z = 1$ the treatment condition. Under the standard consistency assumption, we have that the observed outcome of subject i equals the potential outcome associated to the experimental condition that subject i has actually received, i.e., $Y_i = Y_i(Z_i)$. Throughout the paper, we adopt the convention that $Y_i = 1$ denotes an adverse outcome, such as death or the contraction of a disease. We take a super-population perspective, and assume that subjects are independent and identically distributed (i.i.d.) draws from a common population. We assume complete randomization, which implies ignorability of the treatment assignment, $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp Z_i$.

2.2 Marginal parameterization

When subjects are independently drawn from a common super-population and the treatment is assigned at random, it follows that the observed *counts* of adverse outcomes in each treatment arm,

$$y_0 = \sum_{i=1}^{N} Y_i(1 - Z_i), \qquad y_1 = \sum_{i=1}^{N} Y_i Z_i,$$

follow independent binomial distributions (see Supplement A Section 7 for derivation):

$$y_0 \sim \text{Binomial}(N_0, \theta_0) \quad \perp\!\!\!\perp \quad y_1 \sim \text{Binomial}(N_1, \theta_1),$$

where here, $\theta_1 = \mathbb{P}(Y_i(1) = 1)$, $N_1 = \sum_i Z_i$ denote the probability of an adverse outcome and the sample size of the treatment group, and $\theta_0 = \mathbb{P}(Y_i(0) = 1)$, $N_0 = N - N_1$ are the analogous quantities for the control group. We refer to the probabilities θ_0 and θ_1 as the *baseline risk* and *risk of treatment*, respectively.

This defines the likelihood under the marginal parameterization of a binary experiment, so called because the parameters (θ_0, θ_1) are defined in terms of the marginal distribution of the potential outcomes $Y_i(0)$ and $Y_i(1)$:

$$L(\mathcal{D}|\theta_0, \theta_1) = \binom{N_0}{y_0} \theta_0^{y_0} (1 - \theta_0)^{N_0 - y_0} \times \binom{N_1}{y_1} \theta_1^{y_1} (1 - \theta_1)^{N_1 - y_1}, \tag{2.1}$$

where hereafter we denote the observed data by $\mathcal{D} = (y_0, y_1, N_0, N_1)$. To determine the effect of treatment, if any, Bayesian inference is carried out using the posterior distribution of the parameters (θ_0, θ_1), which requires specification of a prior distribution for (θ_0, θ_1). There are two main parameterizations with accompanying priors currently in use, discussed extensively in Agresti and Min (2005) and Dablander et al. (2022). These are the independent beta (IB) and logit transformation (LT) approaches.

Independent beta (IB) approach

The independent beta (IB) approach (Jeffreys, 1935) assigns the prior

$$\theta_0 \sim \text{Beta}(a_0, b_0) \quad \perp\!\!\!\perp \quad \theta_1 \sim \text{Beta}(a_1, b_1), \tag{2.2}$$

for some hyperparameters $a_0, b_0, a_1, b_1 > 0$. We refer to (2.2) as the IB$(a; b)$ prior, where $a = (a_0, a_1), b = (b_0, b_1)$. A common *default* specification is $a_0 = b_0 = a_1 = b_1 = 1$, which assigns a uniform distribution to (θ_0, θ_1). This choice of flat priors is usually thought to encode ignorance of (θ_0, θ_1) *a priori*, though it makes strong implicit assumptions as we discuss next.

The main advantage of the IB approach is its simplicity. As the beta prior is conjugate to the binomial likelihood, estimation and posterior simulation can be carried out exactly without resorting to approximate sampling algorithms, such as MCMC. Furthermore, marginal likelihoods and Bayes factors, which are widely used for Bayesian hypothesis testing and can be difficult to calculate in general (usually requiring numerical approximation or estimation via posterior simulation), can be calculated analytically (Kass and Raftery, 1995).

A drawback of the IB approach is the restrictive assumption of independence between θ_0 and θ_1. In most experimental settings, we would expect our knowledge about the risks in the control and treatment groups to be dependent. For example, if we know that the population prevalence of an infectious disease is approximately 1%, we would expect the prevalence of the disease among those receiving a vaccine to be concentrated around 1% or below, reflecting the common prior belief that it is unlikely that the vaccine would cause the disease. The IB prior fails to accommodate this natural dependence between risks in each arm of the trial. Furthermore, since independence in the prior and the likelihood implies independence *a posteriori*, this failure also extends to the posterior.

Logit Transformation (LT) approach

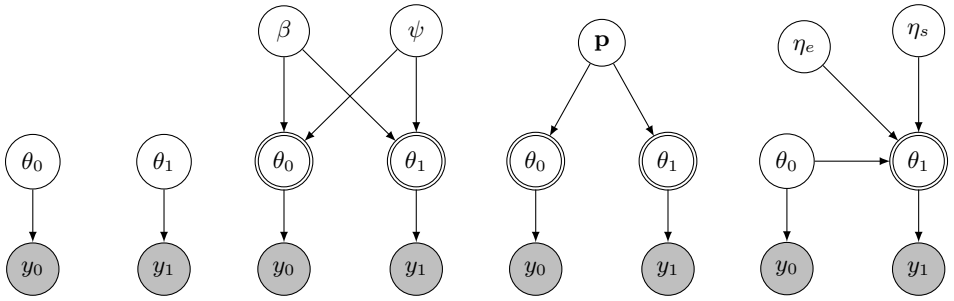
The logit transformation (LT) approach (Kass and Vaidyanathan, 1992; Agresti and Hitchcock, 2005; Dablander et al., 2022) reparameterizes the model in terms of the logit-transformed risks, by defining the parameters (β, ψ) satisfying

$$\log\left(\frac{\theta_0}{1-\theta_0}\right) = \beta - \frac{\psi}{2}, \qquad \log\left(\frac{\theta_1}{1-\theta_1}\right) = \beta + \frac{\psi}{2}.$$

Note this parameterization is equivalent to a logistic regression of the outcome on the treatment with the encoding $Z \in \{-1/2, 1/2\}$ (Gronau et al., 2021). It then assigns an independent normal prior to (β, ψ):

$$\beta \sim \text{Normal}(\mu_\beta, \sigma^2_\beta) \quad \perp\!\!\!\perp \quad \psi \sim \text{Normal}(\mu_\psi, \sigma^2_\psi), \tag{2.3}$$

where $\mu = (\mu_\beta, \mu_\psi)$ and $\sigma = (\sigma_\beta, \sigma_\psi) > 0$ are hyperparameters. A common default choice is $\mu = (0, 0)$ and $\sigma = (1, 1)$. We refer to (2.3) as the LT$(\mu; \sigma)$ prior. This prior encodes correlation between θ_0 and θ_1 through their shared dependence on β and ψ.



(a) Independent Beta (b) Logit Transform (c) Response Type (d) BREASE

Figure 1: Probabilistic graphical models for different parameterizations and prior setups. Gray nodes denote observed variables, white nodes denote latent parameters, and double borders indicate that a node is a deterministic function of its parents. (a) Independent beta priors are placed directly on θ_0 and θ_1; (b) Independent Gaussian priors are placed on the log odds quantities β and ψ; (c) A Dirichlet prior is placed on the response type probabilities $\mathbf{p}$; (d) Our proposal, independent beta priors are placed on θ_0, η_e, and η_s.

Figure 1 depicts probabilistic graphical models comparing the IB and LT parameterizations, as well as the other approaches we will later discuss.

While the LT approach induces prior dependence between θ_0 and θ_1, this comes at the cost of a less intuitive parameterization. Here β is interpreted as the "grand log odds," i.e, the average of the log odds across treatment arms, whereas ψ is the log odds ratio. Odds ratios are notoriously difficult to understand, and thus reasoning about the prior means and variances of log odds—two unbounded hyperparameters—is often challenging in practice. The LT approach also has other computational disadvantages relative to the IB prior. Unlike the IB approach, marginal likelihoods and Bayes factors are not available analytically, and posterior sampling must be carried out approximately.

2.3 Response type (RT) parameterization

The IB and LT approaches focus on the margins of the joint distribution of potential outcomes $(Y_i(0), Y_i(1))$. This focus is natural, as the observed data depends only upon the parameters θ_0 and θ_1. However, thinking in terms of their *joint* distribution reveals alternative ways of inducing prior dependence between these parameters. Specifically, the joint distribution of potential outcomes is fully characterized by four probabilities

$$p_{jk} = \mathbb{P}(Y_i(0) = j, Y_i(1) = k), \qquad j, k \in \{0, 1\}. \tag{2.4}$$

The probabilities $\mathbf{p} = \{p_{jk}\}_{j,k\in\{0,1\}}$ describe the frequencies of the four possible response types in the population (Copas, 1973; Greenland and Robins, 1986). These include: (i) the "doomed" $\{Y_i(0) = 1, Y_i(1) = 1\}$, for whom death occurs regardless of treatment; (ii) the "immune" $\{Y_i(0) = 0, Y_i(1) = 0\}$, for whom death does not occur regardless of treatment; (iii) the "preventive" $\{Y_i(0) = 1, Y_i(1) = 0\}$, for whom treatment

	$Y_i(0) = 0$	$Y_i(0) = 1$	Row Sum
$Y_i(1) = 0$	$p_{00} = (1 - \eta_s)(1 - \theta_0)$	$p_{10} = \eta_e \theta_0$	$1 - \theta_1$
$Y_i(1) = 1$	$p_{01} = \eta_s(1 - \theta_0)$	$p_{11} = (1 - \eta_e)\theta_0$	θ_1
Column Sum	$1 - \theta_0$	θ_0	

Table 1: 2×2 contingency table of potential outcomes for a binary experiment. Only the margins of the table are identified from the observed data.

prevents death; and, (iv) the "causal" $\{Y_i(0) = 0, Y_i(1) = 1\}$, for whom treatment *causes* death. These probabilities are also sometimes referred to as "probabilities of causation" (Tian and Pearl, 2000; Pearl, 2009). Here θ_0 and θ_1, which satisfy $\theta_0 = p_{10} + p_{11}$ and $\theta_1 = p_{01} + p_{11}$, define the margins of Table 1.

Whereas in the marginal parameterization, independence of the likelihood and prior imply that estimation of θ_0 is only informed by data in the control group (and similarly for θ_1), the response type (RT) parameterization intertwines the data from each arm of the study. The shared dependence of θ_0 and θ_1 on the response type proportions reveals the link between outcomes in the control and treated groups.

A Bayesian approach to modeling the response type probabilities $\mathbf{p}$ requires specification of a prior density supported on the probability simplex, making the Dirichlet distribution a natural candidate

$$\mathbf{p} = (p_{00}, p_{10}, p_{01}, p_{11}) \sim \text{Dirichlet}(a_{00}, a_{10}, a_{01}, a_{11}), \qquad a_{00}, a_{10}, a_{01}, a_{11} > 0. \quad (2.5)$$

Indeed, priors of this type have been used in the analysis of partially identified quantities in randomized trials with non-compliance, such as in Chickering and Pearl (1996); see also Imbens and Rubin (1997); Madigan (1999); Hirano et al. (2000). As we show next, the Dirichlet prior is a special case of our proposal, and our analysis not only extends it, but also clarifies its advantages and limitations as a means to induce the desired joint prior distribution on the two binomial proportions (θ_0, θ_1).

3 The BREASE framework

In this section we introduce the BREASE framework for the analysis of binary experiments. We start by parameterizing the likelihood in terms of the baseline risk, efficacy, and risk of adverse side effects of the treatment. We then propose a jointly independent beta prior distributions on these three parameters, which we show to be a generalization of the Dirichlet prior on the response types. Our proposal has a number of advantages. From a statistical perspective, it induces dependence between the risks in the treatment and control groups, while also enabling exact posterior sampling, and marginal likelihood calculations. From a clinical perspective, this parameterization casts the model in terms of natural quantities appearing frequently in the clinician's vocabulary, thereby facilitating interpretability, elicitation of prior knowledge, and sensitivity analyses.

3.1 Baseline risk, efficacy and adverse side effects

To make things concrete, suppose $Y_i = 1$ denotes death. We define the *efficacy* of the treatment, η_e, as the probability that the treatment *prevents* the death of a patient that would have otherwise died without it:

$$\eta_e = \mathbb{P}(Y_i(1) = 0 | Y_i(0) = 1). \tag{3.1}$$

Similarly, we define the risk of *adverse side effects* of the treatment, η_s, as the probability that the treatment *causes* the death of a patient that would have otherwise been healthy:

$$\eta_s = \mathbb{P}(Y_i(1) = 1 | Y_i(0) = 0). \tag{3.2}$$

Note that these are severe adverse side effects that result in an outcome (e.g., death) opposite to the desired outcome of interest (i.e., survival). In the medical literature, this is sometimes called a "paradoxical reaction" (Smith et al., 2012). Such events could be the result not only of severe adverse biological reactions, but also of other forms of iatrogenesis, such as medical errors.

These quantities can be interpreted as probabilities of sufficient causation (Tian and Pearl, 2000; Cinelli and Pearl, 2021), i.e., η_e is the probability that treatment is sufficient to save or cure a patient, while η_s is the probability that treatment is sufficient to kill or hurt a patient. They correspond directly to the counterfactual interpretation of what clinicians colloquially refer to as "efficacy" and "side effects" of a drug or vaccine. Indeed, a commonly used measure in clinical trials called "efficacy", defined as $1-\theta_1/\theta_0$, equals precisely η_e under the assumption that treatment causes no harm ($\eta_s = 0$).

Applying the law of total probability, we can decompose the risk of treatment in terms of the baseline risk, efficacy, and risk of adverse side effects (BREASE), as

$$\theta_1 = (1 - \eta_e)\theta_0 + \eta_s(1 - \theta_0). \tag{3.3}$$

Table 1 shows how the response type probabilities $\mathbf{p}$ can be written as products of θ_0, η_s, and η_e. As with the response type approach, this parameterization highlights the natural dependence between θ_0 and θ_1 that is easy to miss without framing the problem in the language of potential outcomes. For example, note that θ_0 and θ_1 are functionally independent only under the strong assumption that $\eta_e = 1 - \eta_s$, i.e., the probability of treatment saving a patient is equal to the probability that it does not kill one.

Likelihood

Plugging in (3.3), we can rewrite the likelihood (2.1) in terms of $(\theta_0, \eta_e, \eta_s)$.

Theorem 1. *Under (2.1) and (3.1)-(3.3), the likelihood is*

$$L(\mathcal{D}|\theta_0, \eta_e, \eta_s) = \binom{N_0}{y_0}\binom{N_1}{y_1}\sum_{j=0}^{y_1}\sum_{k=0}^{N_1-y_1}\left\{\binom{y_1}{j}\binom{N_1-y_1}{k}\theta_0^{y_0+j+k}(1-\theta_0)^{N-(y_0+j+k)}\right.$$
$$\left.\times\,\eta_e^k(1-\eta_e)^j\eta_s^{y_1-j}(1-\eta_s)^{N_1-y_1-k}\right\}, \qquad (\theta_0, \eta_e, \eta_s) \in [0,1]^3. \tag{3.4}$$

Theorem 1 follows from applying the binomial theorem twice. As the likelihood (3.4) is polynomial in $(\theta_0, \eta_e, \eta_s)$, any prior distribution $\pi(\theta_0, \eta_e, \eta_s)$ for which the moments can be explicitly calculated yields an analytical expression for the marginal likelihood. In particular, if

$$\pi(\theta_0, \eta_e, \eta_s) \propto \theta_0^{\alpha_0 - 1}(1-\theta_0)^{\beta_0 - 1} \times \eta_e^{\alpha_e - 1}(1-\eta_e)^{\beta_e - 1} \times \eta_s^{\alpha_s - 1}(1-\eta_s)^{\beta_s - 1}$$

is a product of independent beta distributions, as we will see in the next section, then the marginal likelihood is a weighted sum of beta function values. Furthermore, the posterior distribution $\pi(\theta_0, \eta_e, \eta_s | \mathcal{D})$ will be a mixture of independent beta distributions, from which we can sample exactly via simulation.

Partial identification and monotonicity

The counterfactual parameters η_e and η_s are only partially identified by the observed data. That is, in the limit of infinite data, even though θ_0 and θ_1 are point identified, (3.3) defines a single equation with two unknowns, η_e and η_s, which cannot be solved uniquely. Without further assumptions, we thus have the bounds

$$\max\left\{0, 1 - \frac{\theta_1}{\theta_0}\right\} \leq \eta_e \leq \min\left\{\frac{1-\theta_1}{\theta_0}, 1\right\}, \quad \max\left\{0, \frac{\theta_1 - \theta_0}{1-\theta_0}\right\} \leq \eta_s \leq \min\left\{\frac{\theta_1}{1-\theta_0}, 1\right\}.$$

As the sample size increases, the posterior distribution of η_s and η_e will not concentrate in a point—rather, it will remain spread over its partially identified region (Richardson et al., 2011; Gustafson, 2015). Notice, however, that this does not affect the behavior of the posterior distribution of (θ_0, θ_1). The BREASE parameterization thus explicitly separates the identified and partially identified parameters—(θ_0, θ_1) and (η_e, η_s), respectively. Even if interest does not lie in the counterfactual probabilities (η_s, η_e) *per se*, assigning a prior to those quantities can be thought of as a causally principled way to specify a joint prior on the identified target parameters (θ_0, θ_1).

Finally, a common assumption in the potential outcomes literature is called *monotonicity*, which states that the treatment does no harm. In our framework, this corresponds to the constraint $\eta_s = 0$. This assumption may be reasonable in many clinical settings. Under monotonicity, the efficacy of the treatment is in fact point identified, and given by $\eta_e = 1 - \theta_1/\theta_0$. The quantity θ_1/θ_0 is known as the risk ratio, and the quantity $1 - \theta_1/\theta_0$ is indeed known as "efficacy" in the clinical trials literature. In cases where side-effects are not expected to be exactly zero, but are expected to be small, the BREASE approach allows one to instead place an informative prior on η_s.

3.2 Prior specification

Bayesian inference with the likelihood (3.4) requires specifying a prior distribution on three separate and variation independent probabilities, i.e, $(\theta_0, \eta_e, \eta_s) \in [0,1]^3$ (Basu, 1977). We propose setting jointly independent beta prior distributions on these parameters:

$$\theta_0 \sim \text{Beta}^*(\mu_0, n_0) \quad \perp\!\!\!\perp \quad \eta_e \sim \text{Beta}^*(\mu_e, n_e) \quad \perp\!\!\!\perp \quad \eta_s \sim \text{Beta}^*(\mu_s, n_s), \tag{3.5}$$

where here $\text{Beta}^*(\mu, n)$ denotes a $\text{Beta}(a, b)$ distribution, with mean $\mu = a/(a+b)$ and prior "sample size" $n = a + b$. We refer to (3.5) as the $\text{BREASE}(\boldsymbol{\mu}; \boldsymbol{n})$ prior, where $\boldsymbol{\mu} = (\mu_0, \mu_e, \mu_s)$, $\boldsymbol{n} = (n_0, n_e, n_s)$.

Since (3.5) defines a jointly independent beta prior on $(\theta_0, \eta_e, \eta_s)$, the discussion in Section 3.1 applies. In particular, the posterior of $(\theta_0, \eta_e, \eta_s)$ is a mixture of independent betas, which permits exact sampling via simulation, and the marginal likelihood is available analytically as a weighted sum of beta functions, as we show in Sections 3.3 and 3.4.

Connections to the (generalized) Dirichlet. The prior (3.5) induces a *generalized* Dirichlet distribution (Dickey, 1983; Dickey et al., 1987; Tian et al., 2003) on the vector of potential outcomes probabilities $\mathbf{p}$—see Supplement A Section 2 for derivation and further discussion. In particular, the *generalized* Dirichlet reduces to the *traditional* Dirichlet distribution (2.5) for the following restricted choice of prior sample sizes

$$n_e = \mu_0 n_0, \quad n_s = (1 - \mu_0) n_0. \tag{3.6}$$

Moreover, since $\theta_1 = p_{01} + p_{11}$, by the aggregation property of the Dirichlet (Ng et al., 2011), marginally we have

$$\theta_1 \sim \text{Beta}^* \left((1 - \mu_e)\mu_0 + \mu_s(1 - \mu_0), n_0 \right), \tag{3.7}$$

which resembles the decomposition (3.3). The BREASE approach thus reveals an implicit "equal confidence" assumption of the *traditional* Dirichlet: the prior spread for θ_0 determines the spread of the distributions of η_e, η_s, and θ_1 *a priori*. Hence, the *traditional* Dirichlet is underparameterized, and unsuitable for cases in which, say, we have ample knowledge of the baseline risk but relatively little information about the possible efficacy or side effects of the treatment (or vice-versa), such as in clinical trials with historical control information (Schmidli et al., 2014). Casting the likelihood in terms of the BREASE parameters makes such choices explicit, by allowing the hyperparameters governing θ_0, η_e and η_s to be set independently.

Induced prior distribution of (θ_0, θ_1)

As mentioned in Section 3.1, our goal with the BREASE approach is primarily to induce causally sound priors on the identified parameters of interest, the two binomial proportions (θ_0, θ_1). Thus we now discuss the induced marginal and conditional distribution of the risk of treatment, θ_1, under the BREASE prior (3.5).

From equation (3.3) we see that θ_1, conditionally on θ_0, is distributed as a convex combination of independent beta random variables *a priori*. This distribution was studied in Pham-Gia and Turkkan (1998) and is given in terms of Appell's first hypergeometric function F_1—in Supplement A Section 1 we derive the explicit formula and provide further discussion. From here, the marginal prior on θ_1 can be obtained as $\pi(\theta_1) = \int_0^1 \pi(\theta_1|\theta_0)\pi(\theta_0)d\theta_0$. While the general formula for $\pi(\theta_1|\theta_0)$ may look unwieldy, and the integration in $\pi(\theta_1)$ prohibitive, there are noteworthy specific cases.

Equal confidence. As noted in the previous discussion, under the equal confidence assumption, $n_e = \mu_0 n_0$, $n_s = (1-\mu_0)n_0$, the marginal prior induced on θ_1 is the beta distribution in (3.7). In particular, to obtain equal marginal priors for the treatment and control groups, i.e., $\theta_z \sim \text{Beta}^*(\mu_0, n_0)$ for $z \in \{0,1\}$, it suffices to set $\mu_s = (\mu_0/(1-\mu_0))\mu_e$, with $0 \leq \mu_e \leq \min(1, (1-\mu_0)/\mu_0)$. Choosing $\mu_0 = 1/2$, $n_0 = 2$, and $\mu_e = \mu_s = \mu$ results in marginal uniform priors with prior correlation $\text{Cor}(\theta_0, \theta_1) = 1 - 2\mu$.

Monotonicity. Under the "no harm" monotonicity assumption, $\eta_s = 0$, we have $\theta_1 = (1-\eta_e)\theta_0$, in which case θ_1 is a product of independent beta random variables *a priori*. Springer and Thompson (1970) derived the form of this distribution, with the density given as a Meijer G-function. In particular, if $n_e = \mu_0 n_0$, we can show that $\theta_1 \sim \text{Beta}((1-\mu_e)n_e, \mu_e n_e + (1-\mu_0)n_0)$. For another example, if $(\theta_0, \eta_e) \sim \text{Uniform}(0,1)^2$, we have $\pi(\theta_1) = -\log\theta_1$. Regarding the conditional prior $\pi(\theta_1|\theta_0)$ under the "no harm" assumption, it is clearly a scaled beta distribution, since $\theta_1 = (1-\eta_e)\theta_0$. If $\eta_e \sim \text{Uniform}(0,1)$, we have $\theta_1|\theta_0 \sim \text{Uniform}(0, \theta_0)$. Similarly, under the "no benefit" assumption $\eta_e = 0$, we have that $\theta_1 = \theta_0 + \eta_s(1-\theta_0)$, which is a scaled and shifted beta random variable conditional on θ_0. If $\eta_s \sim \text{Uniform}(0,1)$, then $\theta_1|\theta_0 \sim \text{Uniform}(\theta_0, 1)$.

Moments. The joint density $\pi(\theta_0, \theta_1)$ induced by the BREASE$(\mu; n)$ prior is generally complicated, but its moments are easily computed in terms of the hyperparameters (μ, n) as θ_1 is a polynomial in $(\theta_0, \eta_e, \eta_s)$, which are beta distributed *a priori*. For example, the prior covariance has a simple form, $\text{Cov}(\theta_0, \theta_1) = \frac{\mu_0(1-\mu_0)}{n_0+1}(1-\mu_e-\mu_s)$. This implies the following directions of the prior correlation,

$$\text{Cor}(\theta_0, \theta_1) \begin{cases} < 0, & \mu_e + \mu_s > 1, \\ = 0, & \mu_e + \mu_s = 1, \\ > 0, & \mu_e + \mu_s < 1. \end{cases} \tag{3.8}$$

In words, θ_0 and θ_1 are positively correlated *a priori* when the expected harm and benefit of treatment are small, and negatively correlated otherwise.

Default prior. While we encourage the use of informative priors, it is useful to have reasonable defaults to start the analysis. If we would like to put θ_0 and θ_1 on equal footing, the BREASE$(1/2, \mu, \mu; 2, 1, 1)$ is thus the natural choice, with the following properties: (i) puts flat uniform priors on θ_0 and θ_1 (as with the IB approach); (ii) induces prior correlation between parameters (as with the LT approach); (iii) assumes no effect of treatment, on average (as with the IB and LT approaches); and, (iv) depends on a single, easily interpretable parameter μ denoting the expected benefits (efficacy) or harm (side effects) of the treatment. When $\mu > 1/2$, θ_1 and θ_0 become anti-correlated, and thus for most cases, $\mu \leq 1/2$ is a reasonable choice. Our preferred specification uses $\mu = 0.3$ as the default. As Figure 1 in Supplement A shows, this (weakly) encodes the expectation of moderate effects and concentrates mass on the diagonal $\theta_0 = \theta_1$. This quality is useful in the context of Bayesian hypothesis testing. When testing a null hypothesis H_0 (e.g., no effect of treatment on average, $H_0 : \theta_0 = \theta_1$) nested within an alternative H_1, it is desirable for the prior under H_1 to concentrate mass around the null model (Jeffreys, 1961; Gunel and Dickey, 1974; Casella and Moreno, 2009).

Algorithm 1 BREASE posterior—exact sampling algorithm

Input: Data $\mathcal{D} = (y_0, y_1, N_0, N_1)$, hyperparameters $(\mu_0, \mu_e, \mu_s, n_0, n_e, n_s)$, and desired number of posterior samples T.
Iterate: For sample $t \in \{1, \dots, T\}$,

(i) Sample $P_1 \in \{0, \dots, N_1 - y_1\}$ conditional on $\mathcal{D}$ with probability, as per (3.11),

$$\pi(P_1|\mathcal{D}) = \sum_{C_1=0}^{y_1} \pi(C_1, P_1|\mathcal{D}).$$

(ii) Sample $C_1 \in \{0, \dots, y_1\}$ conditional on $(P_1, \mathcal{D})$ with probability, as per (3.11),

$$\pi(C_1|P_1, \mathcal{D}) \propto \pi(C_1, P_1|\mathcal{D}).$$

(iii) Sample $(\theta_0, \eta_e, \eta_s)$ conditional on $(C_1, P_1, \mathcal{D})$ from the distribution (3.12).

Output: Posterior samples $\{(\theta_0^{(t)}, \eta_e^{(t)}, \eta_s^{(t)})\}_{t\in\{1,\dots,T\}}$.

3.3 Posterior sampling

Exact sampling

The posterior under (3.5) is given by the following mixture of independent betas

$$\pi(\theta_0, \eta_e, \eta_s|\mathcal{D}) \propto \sum_{j=0}^{y_1} \sum_{k=0}^{N_1-y_1} \left\{ \binom{y_1}{j} \binom{N_1 - y_1}{k} \theta_0^{y_0+j+k+\mu_0 n_0} (1-\theta_0)^{N-(y_0+j+k)+(1-\mu_0)n_0} \right.$$
$$\left. \times \eta_e^{k+\mu_e n_e} (1-\eta_e)^{j+(1-\mu_e)n_e} \eta_s^{y_1-j+\mu_s n_s} (1-\eta_s)^{N_1-y_1-k+(1-\mu_s)n_s} \right\}. \tag{3.9}$$

As with the prior, this posterior falls into the family of generalized Dirichlet distributions on the vector of potential outcomes probabilities $\mathbf{p}$. While some posterior quantities can be obtained analytically (see Supplement A Section 4), working with the posterior density can be cumbersome; we now describe how to sample exactly from the posterior via simulation. See Supplement A Section 3.1 for a full derivation of Theorem 2.

Theorem 2. *Let $(\theta_0, \eta_e, \eta_s)$ be random variables drawn according to Algorithm 1. Then $(\theta_0, \eta_e, \eta_s)$ are distributed according to the BREASE posterior (3.9).*

Sketch of proof. We define the counterfactual counts

$$C_1 = \sum_{i=1}^{N} I(Z_i = 1, Y_i(1) = 1, Y_i(0) = 0), \qquad P_1 = \sum_{i=1}^{N} I(Z_i = 1, Y_i(1) = 0, Y_i(0) = 1),$$

which are unobserved quantities. Here, C_1 is the number of "causal" subjects in the treatment group, i.e., those who died under treatment but would have survived if untreated. Similarly, P_1 is the number of "preventive" subjects in the treatment group, i.e.,

those who survived under treatment but would have died if untreated. The BREASE posterior can then be expressed as a mixture distribution:

$$\pi(\theta_0, \eta_e, \eta_s|\mathcal{D}) = \sum_{C_1=0}^{y_1} \sum_{P_1=0}^{N_1-y_1} \pi(\theta_0, \eta_e, \eta_s|C_1, P_1, \mathcal{D}) \times \pi(C_1, P_1|\mathcal{D}). \tag{3.10}$$

Hence, we can sample from the posterior by first drawing from the distribution of unobserved counts (C_1, P_1) conditional on the observed data $\mathcal{D}$. This distribution has probability mass function

$$\begin{aligned}\pi(C_1, P_1|\mathcal{D}) \propto &\binom{y_1}{C_1}\binom{N_1 - y_1}{P_1}\mathrm{B}(P_1 + \mu_e n_e, y_1 - C_1 + (1-\mu_e)n_e) \\ &\times \mathrm{B}(y_0 + y_1 - C_1 + P_1 + \mu_0 n_0, N - (y_0 + y_1 - C_1 + P_1) + (1-\mu_0)n_0) \\ &\times \mathrm{B}(C_1 + \mu_s n_s, N_1 - y_1 - P_1 + (1-\mu_s)n_s). \end{aligned} \tag{3.11}$$

We then sample the parameters $(\theta_0, \eta_e, \eta_s)$, which have an independent beta distribution conditional on the augmented data $(C_1, P_1, \mathcal{D})$:

$$\begin{aligned}\pi(\theta_0, \eta_e, \eta_s|C_1, P_1, \mathcal{D}) = &\mathrm{Beta}(\eta_e; P_1 + \mu_e n_e, y_1 - C_1 + (1-\mu_e)n_e) \\ &\times \mathrm{Beta}(\theta_0; y_0 + y_1 - C_1 + P_1 + \mu_0 n_0, N - (y_0 + y_1 - C_1 + P_1) + (1-\mu_0)n_0) \\ &\times \mathrm{Beta}(\eta_s; C_1 + \mu_s n_s, N_1 - y_1 - P_1 + (1-\mu_s)n_s). \end{aligned} \tag{3.12}$$

Note that this derivation of the distribution (3.11) provides a counterfactual interpretation of the mixture weights that result from directly normalizing the kernels in (3.9). □

Data augmentation (DA) algorithm

Algorithm 2 BREASE posterior—data augmentation algorithm

Input: Data $\mathcal{D} = (y_0, y_1, N_0, N_1)$, hyperparameters $(\mu_0, \mu_e, \mu_s, n_0, n_e, n_s)$, desired number of posterior samples T, number of burn-in iterations B, and BREASE parameter initialization $(\theta_0^{(0)}, \eta_e^{(0)}, \eta_s^{(0)}) \in (0,1)^3$.
Iterate: For sample $t \in \{1, \ldots, T\}$,

(i) Sample $(C_1^{(t)}, P_1^{(t)})$ conditional on $(\theta_0^{(t-1)}, \eta_e^{(t-1)}, \eta_s^{(t-1)}, \mathcal{D})$ from the independent binomial distributions

$$C_1^{(t)} \sim \text{Binomial}\left(y_1, \tfrac{(1-\theta_0^{(t-1)})\eta_s^{(t-1)}}{\theta_1^{(t-1)}}\right), P_1^{(t)} \sim \text{Binomial}\left(N_1 - y_1, \tfrac{\theta_0^{(t-1)}\eta_e^{(t-1)}}{1-\theta_1^{(t-1)}}\right),$$

where $\theta_1^{(t-1)} = \theta_0^{(t-1)}(1-\eta_e^{(t-1)}) + (1-\theta_0^{(t-1)})\eta_s^{(t-1)}$.

(ii) Sample $(\theta_0^{(t)}, \eta_e^{(t)}, \eta_s^{(t)})$ conditional on $(C_1^{(t)}, P_1^{(t)}, \mathcal{D})$ from the independent beta distributions (3.12).

Output: Posterior samples after burn-in $\{(\theta_0^{(t)}, \eta_e^{(t)}, \eta_s^{(t)})\}_{t \in \{B+1, \ldots, T\}}$.

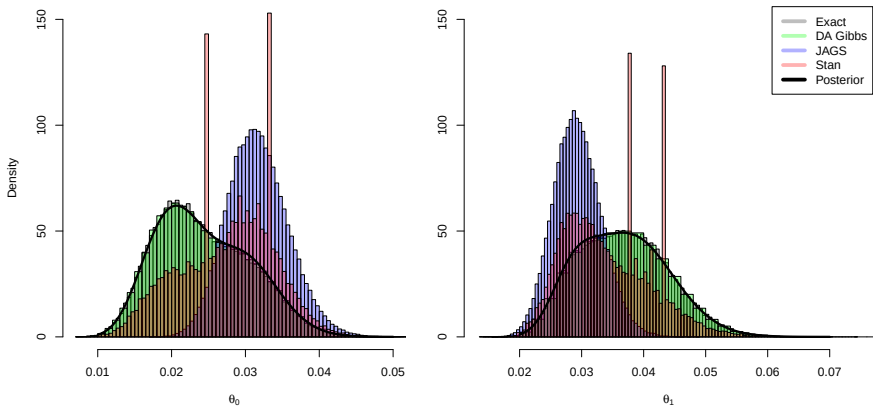


Figure 2: Pathological MCMC posterior sampling exhibited in posterior histograms of the baseline risk θ_0 (left) and treatment risk θ_1 (right). The marginal posterior of θ_1 (black curve) was approximated using numerical integration.

We now derive a Gibbs sampler targeting the BREASE posterior (3.9) based on the data augmentation scheme introduced for Algorithm 1. Algorithm 2 defines the Gibbs sampler. It consists of two steps: (i) first, we sample the counterfactual counts C_1 and P_1 conditional on the BREASE parameters; and, (ii) we sample θ_0, η_e, η_s conditional on the augmented data. In numerical experiments, we find that the algorithm converges to the BREASE posterior quickly, often mixing within a few hundred iterations, and the sampling is also quite fast. The conditional distribution of the unobserved counts $(C_1, P_1)|(\theta_0, \eta_e, \eta_s, \mathcal{D})$ is derived in Supplement A Section 3.1.

Pathological sampling

To demonstrate the utility of our posterior sampling algorithms, we now turn to an example for which RJAGS (Plummer, 2023) and RStan (Stan Development Team, 2023), two popular MCMC software packages, fail to sample from the BREASE posterior. We use the data $y_0 = 20$, $N_0 = 1000$, $y_1 = 40$, $N_1 = 1000$, and the hyperparameters $\mu_0 = 0.5$, $n_0 = 2$, $\mu_e = 0.5$, $n_e = 2, \mu_s = 0.01$, $n_s = 1$. The prior distributions for θ_0 and η_e are vague independent Uniform$(0, 1)$ distributions. On the other hand, the prior on the risk of side effects η_s is concentrated near 0 with mean $\mu_s = 0.01$. This prior encodes a quasi-monotonicity assumption on the treatment that is clearly in conflict with the data.

Prior-data conflict, which arises when the prior is concentrated on parameter values that are unlikely given the data, is a common culprit when diagnosing pathological MCMC sampling (Evans and Moshonov, 2006). It is also a salient issue in the Bayesian analysis of clinical trials, particularly when historical information or clinical expertise are brought to bear on the design and analysis of the study (Schmidli et al., 2014). This example is no exception. Figure 2 shows histograms of 100,000 posterior samples of θ_0

and θ_1 drawn using Algorithm 1 (grey), Algorithm 2 (green), JAGS (blue), and Stan (red). The marginal posterior density is plotted in black for reference. The posterior of θ_0, which is a mixture of beta distributions, is exhibited in the left panel of Figure 2. While Algorithms 1 and 2 produce posterior samples that fully capture the distribution, JAGS and Stan fail to adequately explore the left half of the distribution. Although Stan manages to deviate from the right half as compared to JAGS, its chains get stuck at $\theta_0 \approx 0.024$ and $\theta_0 \approx 0.033$ when the sampler rejects numerous proposal draws. The story is much the same for θ_1.

This example demonstrates that it is useful to have bespoke algorithms that perform well, even in adversarial settings. In particular, the algorithms we provide here may prove useful for future extensions of the model, as we will later discuss. Nevertheless, we note that JAGS and Stan do work well for this model in most cases—indeed, this is a pathological example designed to be challenging. Furthermore, in the case of prior-data conflict (or more generally when a sampler is struggling), a reassessment of the prior may be warranted, perhaps in favor of a more robust approach (Schmidli et al., 2014). In Section 9 of Supplement A, we further investigate the numerical issues causing the sampling difficulties in JAGS and Stan and discuss solutions.

Monotonicity. Posterior sampling under monotonicity constraints can be obtained with similar procedures. See Theorems 3.1-3.2 of Supplement A, Section 3.

3.4 Marginal likelihoods and Bayes factors

From a Bayesian perspective, hypothesis testing is essentially a model comparison exercise (Jeffreys, 1961; Dickey and Lientz, 1970; Kass and Raftery, 1995). Consider two competing hypotheses, H_0 and H_1. For each hypothesis H_k, $k \in \{0, 1\}$, the Bayesian approach requires postulating a fully specified model M_k, with likelihood $L_k(\mathcal{D}|\theta)$ and prior $\pi_k(\theta)$, respecting the constraints of the hypothesis the model is intended to represent. Evidence in favor of H_1 relative to H_0 is then quantified using the Bayes factor BF_{10}, given by the ratio of the marginal likelihoods of the observed data under each model, $\text{BF}_{10} = L_1(\mathcal{D})/L_0(\mathcal{D})$, where $L_k(\mathcal{D}) = \int L_k(\mathcal{D}|\theta)\pi_k(\theta)d\theta$. Given prior model probabilities $\mathbb{P}(M_0)$, $\mathbb{P}(M_1)$, the posterior odds of M_1 and M_0 are then $\mathbb{P}(M_1|\mathcal{D})/\mathbb{P}(M_0|\mathcal{D}) = \text{BF}_{10} \times \mathbb{P}(M_1)/\mathbb{P}(M_0)$. In this section we show how to formulate such models instantiating a number of relevant statistical hypotheses with the BREASE approach, and provide analytical formulae for the marginal likelihoods. For all models considered here the likelihood is the same, so we focus the discussion on the formulation of the prior.

Let us first consider testing the null hypothesis $H_0 : \theta_1 = \theta_0$ against the alternative hypothesis $H_1 : \theta_1 \neq \theta_0$. For H_1, we propose using the unconstrained model M_1, with the BREASE prior in (3.5) and equation (3.3),

$$M_1 : (\theta_0, \eta_e, \eta_s) \sim \text{BREASE}(\mu; n), \qquad \theta_1 = (1 - \eta_e)\theta_0 + \eta_s(1 - \theta_0). \tag{3.13}$$

As for the null hypothesis $H_0 : \theta_1 = \theta_0$, we instantiate it with the null model,

$$M_0 : \theta_0 \sim \text{Beta}^*(\mu_0, n_0), \qquad \theta_1 = \theta_0. \tag{3.14}$$

One benefit of M_0 is that its prior is logically consistent with the marginal distribution of θ_0 under M_1, both implying $\theta_0 \sim \text{Beta}^*(\mu_0, n_0)$ *a priori*. Note that the prior (3.14) emerges naturally from M_1 in at least two ways: (i) when postulating that the treatment does not work at all, by setting $\eta_s = \eta_s = 0$; or, (ii) by noting that, if the treatment has no effect on average (i.e, the efficacy of the treatment precisely offsets its side effects), one can side-step thinking about η_s and η_e altogether. In both cases, we borrow the prior of θ_0 from M_1, and simply set θ_1 equal to θ_0. We discuss alternative prior formulations for H_0 in Supplement A Section 5.1.

Other relevant hypothesis one may wish to test are that the treatment is beneficial $H_- : \theta_1 < \theta_0$ or that the treatment is harmful $H_+ : \theta_1 > \theta_0$, on average. A straightforward approach to specify models for such hypotheses is to note that M_1 already induces positive probabilities to the events postulated in H_- and H_+. Thus, we can borrow this knowledge, already elicited when forming M_1, to define the priors π_- and π_+,

$$\pi_-(\theta_0, \eta_e, \eta_s) := \pi_1(\theta_0, \eta_e, \eta_s | \theta_1 < \theta_0), \quad \pi_+(\theta_0, \eta_e, \eta_s) := \pi_1(\theta_0, \eta_e, \eta_s | \theta_1 > \theta_0). \tag{3.15}$$

The priors π_- and π_+ result in the models M_- and M_+, for H_- and H_+ respectively. Similarly to M_0, one benefit of these models is that the induced priors on $(\theta_0, \eta_e, \eta_s)$ are logically consistent with the beliefs expressed in M_1, under the constraints H_- and H_+. The same strategy employed here can be used for interval hypotheses of the type $H_0^\delta : |\theta_1 - \theta_0| \leq \delta$, with $\delta > 0$ (or, more generally, for any event with nonzero probability under M_1). Alternative models for H_- and H_+, leveraging instead monotonicity constraints, such as $\eta_s = 0$, are discussed in Supplement A Section 5.2.

In all cases above, the marginal likelihood can be obtained using analytical formulae and simple Monte Carlo approximation, facilitating the computation of Bayes factors.

Theorem 3. *The marginal likelihood of the data under M_0 is given by a beta-binomial distribution. Under M_1, it is given by a weighted sum of beta functions:*

$$\begin{aligned} L_1(\mathcal{D}) = \binom{N_0}{y_0}\binom{N_1}{y_1} \sum_{j=0}^{y_1} \sum_{k=0}^{N_1 - y_1} \binom{y_1}{j}\binom{N_1 - y_1}{k} &\times \frac{B(k + \mu_e n_e, j + (1-\mu_e)n_e)}{B(\mu_e n_e, (1-\mu_e)n_e)} \\ &\times \frac{B(y_0 + j + k + \mu_0 n_0, N - (y_0 + j + k) + (1-\mu_0)n_0)}{B(\mu_0 n_0, (1-\mu_0)n_0)} \\ &\times \frac{B(y_1 - j + \mu_s n_s, N_1 - y_1 - k + (1-\mu_s)n_s)}{B(\mu_s n_s, (1-\mu_s)n_s)}. \end{aligned} \tag{3.16}$$

Under M_- and M_+, it can be obtained from $L_1(\mathcal{D})$ as follows,

$$L_-(\mathcal{D}) = L_1(\mathcal{D}) \times \frac{\pi_1(\theta_1 < \theta_0 | \mathcal{D})}{\pi_1(\theta_1 < \theta_0)}, \qquad L_+(\mathcal{D}) = L_1(\mathcal{D}) \times \frac{\pi_1(\theta_1 > \theta_0 | \mathcal{D})}{\pi_1(\theta_1 > \theta_0)}. \tag{3.17}$$

Proof. The result for M_0 is well-known. $L_1(\mathcal{D})$ in (3.16) follows directly from integration of (3.4) under the prior (3.5). $L_-(\mathcal{D})$ and $L_+(\mathcal{D})$ in (3.17) follow from Bayes' rule. □

Remark 1. The prior and posterior probabilities $\pi_1(\theta_1 < \theta_0)$ and $\pi_1(\theta_1 < \theta_0 | \mathcal{D})$ can be approximated using Monte Carlo integration with exact samples, as per Section 3.3.

Remark 2. As per (3.17), if one postulates prior model probabilities $\mathbb{P}(M_-|M_1) = \pi_1(\theta_1 < \theta_0)$ and $\mathbb{P}(M_+|M_1) = \pi_1(\theta_1 > \theta_0)$, the Bayes factor testing $H_0 : \theta_1 = \theta_0$ against $H_1 : \theta_1 \neq \theta_0$ (using M_1) conveniently decomposes into the weighted average of the Bayes factors testing H_0 against H_- (using M_-) and H_0 against H_+ (using M_+)—though, of course, users can postulate prior probabilities for the models M_- and M_+ as they wish.

3.5 Extension to covariates

We conclude this section by demonstrating how the BREASE approach can be extended to accommodate discrete covariates. By extending the method in this way, we can address a number of important applications, which include: estimating conditional average treatment effects in randomized experiments; accounting for stratification in randomized experiments, or measured confounding in observational studies; and pooling evidence across multiple trials. We leave extensions to continuous covariates to future work.

Likelihood

Suppose we observe i.i.d. samples $(Y_i, Z_i, X_i),\ i \in \{1, \dots, N\}$, where, as before, Y_i and Z_i denote the binary outcome and treatment indicators for subject i and X_i is a discrete pre-treatment covariate taking values in $\mathcal{X}$. We allow for the possibility of selection into treatment based on X_i. Hence, we now assume that randomization of the treatment holds only within strata of X_i (also known as *conditional* ignorability) $Y_i(z) \perp\!\!\!\perp Z_i|X_i$.

Let $y_{z,x}$ denote the observed death count and $N_{z,x}$ the corresponding sample size for each stratum $x \in \mathcal{X}$ and study arm $z \in \{0, 1\}$. Further define the total count for stratum x as $N_x = N_{0,x} + N_{1,x}$ and the total population size $N = \sum_{x\in\mathcal{X}} N_x$. We use boldface to indicate vectors, $\boldsymbol{N} = \{N_{z,x}\}_{z\in\{0,1\},x\in\mathcal{X}}$ and $\boldsymbol{y} = \{y_{z,x}\}_{z\in\{0,1\},x\in\mathcal{X}}$. Finally, let $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{N})$ denote the full data and $(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\delta}, \boldsymbol{p_X})$ parameters,

$$\boldsymbol{\theta} = \{\theta_{z,x}\}_{z\in\{0,1\},x\in\mathcal{X}}, \quad \boldsymbol{\eta} = \{\eta_{e,x}, \eta_{s,x}\}_{x\in\mathcal{X}}, \quad \boldsymbol{\delta} = \{\delta_x\}_{x\in\mathcal{X}}, \quad \boldsymbol{p_X} = \{p_x\}_{x\in\mathcal{X}},$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ collect the risks, efficacy and side effects for each stratum; $\delta_x := P(Z_i = 1|X_i = x)$ denotes the propensity score for each stratum x; and $p_x := P(X_i = x)$ denotes the marginal probability of $X_i = x$.

The full likelihood is then given by (see Supplement A Section 7 for derivation)

$$\begin{aligned} L(\mathcal{D}|\boldsymbol{\theta}_0, \boldsymbol{\eta}, \boldsymbol{\delta}, \boldsymbol{p_X}) = \prod_{x\in\mathcal{X}} \Bigg[\binom{N_{0,x}}{y_{0,x}} \binom{N_{1,x}}{y_{1,x}} \sum_{j=0}^{y_{1,x}} \sum_{k=0}^{N_{1,x}-y_{1,x}} \Bigg\{ \binom{y_{1,x}}{j} \binom{N_{1,x}-y_{1,x}}{k} \theta_{0,x}^{y_{0,x}+j+k} \\ \times (1-\theta_{0,x})^{N_x-(y_{0,x}+j+k)} \eta_{e,x}^k (1-\eta_{e,x})^j \eta_{s,x}^{y_{1,x}-j} (1-\eta_{s,x})^{N_{1,x}-y_{1,x}-k} \Bigg\} \Bigg] \\ \times \prod_{x\in\mathcal{X}} \binom{N_x}{N_{1,x}} \delta_x^{N_{1,x}} (1-\delta_x)^{N_{0,x}} \times \frac{N!}{\prod_{x\in\mathcal{X}} N_x!} \prod_{x\in\mathcal{X}} p_x^{N_x}. \end{aligned}$$

The first component above corresponds to the BREASE likelihood (3.4) for each stratum $x \in \mathcal{X}$; the second component corresponds to the binomial likelihood for the treatment

assignment, again for each stratum $x \in \mathcal{X}$; the final component is the marginal likelihood of X, which is a multinomial distribution.

Priors and posterior sampling

The likelihood factorizes into three independent components, corresponding to the BREASE parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$, to the propensity score parameters $\boldsymbol{\delta}$, and finally to the parameters of the marginal distribution of the observed covariates $\boldsymbol{p_X}$. Thus, if the priors for these components are also mutually independent, this independence extends to the posterior, allowing the parameters of each component to be sampled independently. We make this assumption going forward in our discussion of prior specification. We propose two priors for $(\boldsymbol{\theta}, \boldsymbol{\eta})$: (i) an independent BREASE prior for each stratum $x \in \mathcal{X}$; and, (ii) a hierarchical prior that pools information across strata.

Independent BREASE prior. The simplest prior for this setup is to assign independent BREASE priors to the within-stratum parameters $(\theta_{0,x}, \eta_{e,x}, \eta_{s,x})$. Given that the strata are also independent in the likelihood, posterior samples can be drawn independently for each stratum using either the exact sampler (Algorithm 1) or the data-augmented Gibbs sampler (Algorithm 2).

Hierarchical BREASE prior. One drawback of independent priors is that they prevent information from being shared across strata. For example, learning about the efficacy of a vaccine in males would have no impact on our inferences about its efficacy in females. To overcome this, hierarchical priors can be introduced to partially pool information across different categories of X_i. This approach also supports meta-analyses across studies, with X_i representing a study indicator. A natural hierarchical prior would be

$$\begin{aligned}
\theta_{0,x} &\sim \text{Beta}^*(\mu_0, n_0), \quad \mu_0 \sim \text{Beta}^*(\lambda_0, \nu_0), \quad n_0 \sim \text{Gamma}(\alpha_0, \beta_0), \\
\eta_{e,x} &\sim \text{Beta}^*(\mu_e, n_e), \quad \mu_e \sim \text{Beta}^*(\lambda_e, \nu_e), \quad n_e \sim \text{Gamma}(\alpha_e, \beta_e), \\
\eta_{s,x} &\sim \text{Beta}^*(\mu_s, n_s), \quad \mu_s \sim \text{Beta}^*(\lambda_s, \nu_s), \quad n_s \sim \text{Gamma}(\alpha_s, \beta_s).
\end{aligned}$$

Hence, we specify a BREASE$(\boldsymbol{\lambda}, \boldsymbol{\nu})$ prior on the hierarchical BREASE parameters (μ_0, μ_e, μ_s) and Gamma priors on the random effects precision parameters (n_0, n_e, n_s). Posterior sampling can proceed in two stages: (i) conditional on the hierarchical parameters, an independent BREASE update for the BREASE parameters; and (ii) conditional on the BREASE parameters, a Metropolis-Hastings update for the hierarchical parameters. We leave for future work the study of other priors and sampling algorithms.

Population effects. The two procedures described above give us posterior samples of the within-stratum parameters $(\theta_{0,x}, \eta_{e,x}, \eta_{s,x})$, which allow us to obtain posterior samples of conditional treatment effects, such as the conditional risk ratio, $\tau_x := \theta_{1,x}/\theta_{0,x}$, as well as any contrasts of such effects (e.g., $\tau_x - \tau_{x'}$). To recover population (marginal) effects, we need to average over the marginal distribution of X, e.g., $\theta_0 = \sum_{x \in \mathcal{X}} \theta_{0,x} p_x$ and $\theta_1 = \sum_{x \in \mathcal{X}} \theta_{1,x} p_x$. Since $\boldsymbol{p_X}$ and $(\boldsymbol{\theta}, \boldsymbol{\eta})$ are independent *a posteriori*, this averaging can be done at any point in the analysis by simply generating independent posterior samples of $\boldsymbol{p_X}$ (e.g, using a conjugate Dirichlet prior for $\boldsymbol{p_X}$).

4 Empirical Examples

We now demonstrate the utility of our approach in three empirical examples. We show how the BREASE framework can be used to facilitate Bayesian estimation, hypothesis testing, and sensitivity analysis of the results of binary experiments. Concretely, the examples illustrate how our proposal can: (i) help analysts distinguish robust from fragile findings; (ii) clarify what one needs to believe in order to claim that a treatment is effective; and (iii) reconcile disparate results obtained from different methods. See Supplement A Section 6 for details of the calculation of Bayes factors for the IB and LT approaches.

4.1 The effect of aspirin on fatal myocardial infarction

Cardiovascular disease is the leading cause of death in the United States, responsible for more than one in four deaths (Davidson et al., 2022). The Physicians' Health Study (PHS), a large-scale, randomized, placebo-controlled trial conducted in the 1980s, was designed in part to investigate whether low-dose aspirin reduces the risk of cardiovascular mortality (Steering Committee of the Physicians' Health Study Research Group, 1989). This landmark study reported significant reductions in both fatal and nonfatal myocardial infarctions in the treatment group, findings that played a crucial role in the widespread adoption of aspirin for heart attack prevention. Here, we revisit the aspirin component of the PHS, applying the BREASE framework to assess the sensitivity of its results to prior specification.

During the study, $y_0 = 26$ out of $N_0 = 11{,}034$ subjects in the placebo group experienced fatal myocardial infarction compared to $y_1 = 10$ out of $N_1 = 11{,}037$ prescribed aspirin. Using maximum likelihood estimation, the estimated risk ratio θ_1/θ_0 is 0.38, with 95% confidence interval (based on inverting Fisher's exact test) $\text{CI}(95\%) = [0.17, 0.82]$. Consequently, we reject the null hypothesis of zero effect, $H_0 : \theta_1 = \theta_0$, with p-value 0.008. Results based on asymptotic Wald and Pearson tests are nearly identical. Hence, a frequentist would confidently conclude that low-dose aspirin significantly reduces cardiovascular mortality in this population.

Bayesian estimation using *default* priors under the alternative hypothesis (i.e, with a prior that gives zero probability to the null hypothesis of zero effect) yields qualitatively similar, though more conservative answers. The $\text{BREASE}(1/2, \mu, \mu; 2, 1, 1)$ prior with $\mu = 0.3$ yields a posterior median of the risk ratio of 0.44 with a wider 95% credible interval of $\text{CrI}(95\%) = [0.2, 0.96]$. Results for the default IB and LT priors are qualitatively similar, though less conservative: the $\text{LT}(0, 0; 1, \sigma_\psi)$ with $\sigma_\psi = 1$ results in a posterior median of 0.48 and $\text{CrI}(95\%) = [0.25, 0.87]$; the $\text{IB}(a, a; a, a)$ with $a = 1$ returns posterior median 0.4 and $\text{CrI}(95\%) = [0.18, 0.79]$. But how sensitive are these results to the prior?

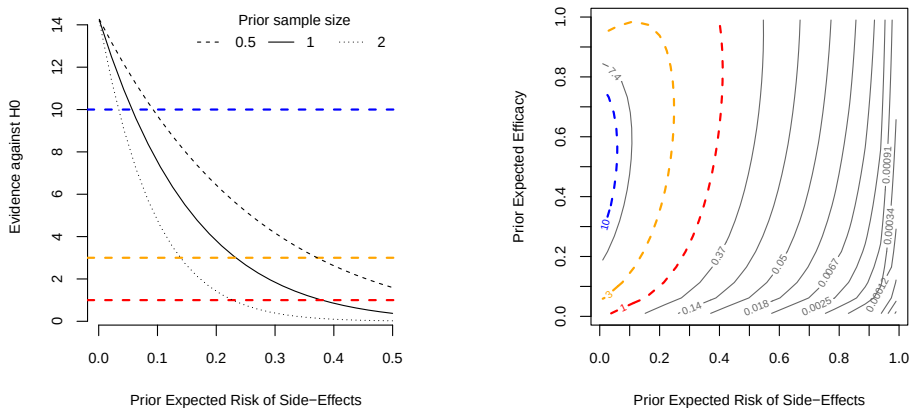
Varying the prior hyperparameter μ of the default BREASE prior (keeping prior sample sizes fixed at $n_e = n_s = 1$) shows that the results are indeed very sensitive to the prior. Credible intervals include the null of no effect as soon as $\mu \leq 0.2$. That is, unless *a priori* we weakly expect efficacy or side-effects to be about 20% or more,

credible intervals would not exclude the null hypothesis of zero effect. This sensitivity also shows up, though it is less apparent, with the IB and LT parameterizations. For the LT prior, this happens when $\sigma_\psi \leq 0.4$. However, notice how the variance of the log odds ratio is harder to be directly interpreted than μ. For the IB, this happens only when $a \geq 17$; this prior specifies 17 deaths in the control and treatment groups, which is on par with the number of deaths observed in the data. Also notice that, in this example, inferences under an independent prior are less conservative than those under dependent priors. This is to be expected, because the LT and BREASE priors shrink estimates toward the null of no effect whereas the IB does not.

One may also be interested in performing a Bayesian hypothesis test based on the Bayes factor, which assigns nonzero prior probability to H_0. As we will see, prior sensitivity is even more pronounced in this case. Here we focus on the exact null, but we note that researchers can also specify an interval null hypothesis, such as $|\theta_1 - \theta_0| < \delta$, as per the discussion in Section 3.4. Perhaps surprisingly, a test based on the IB approach yields a Bayes factor $\text{BF}_{01} = 20.27$, now suggesting that the data provide strong evidence *in favor* of H_0. On the other hand, the Bayes factor under the LT approach is $\text{BF}_{10} = 5.24$, which suggests moderate evidence in favor of $H_1 : \theta_1 \neq \theta_0$. Finally, the default BREASE prior results in $\text{BF}_{10} = 1.2$ providing essentially little evidence in favor of one hypothesis or the other. Hence, when considering Bayes factors, unlike in the previous case, the IB prior results in more conservative inferences compared to the BREASE and LT priors. This occurs, however, for the same reason: under H_1, the IB prior assigns a substantial amount of mass to unreasonably large effect sizes.

How can we make sense of these disparate results? One benefit of the BREASE approach is that it allows one to clearly encode prior assumptions in terms of the expected efficacy and side effects of aspirin, and to easily examine how sensitive the BF is to those assumptions, over the whole range possible values. For example, starting with μ_s, aspirin is an over-the-counter medicine, with ample usage, and it would thus be *unreasonable* to expect that aspirin would *cause* myocardial infarction in a large fraction of otherwise healthy patients. Figure 3a inspects how the Bayes factor is affected as we vary the prior expectation of side effects, ranging from 0.01% (reasonable) to 50% (unreasonable), while still keeping relatively vague priors on the baseline risk and efficacy. The dashed red, orange, and blue lines denote (slightly modified) Jeffreys' thresholds for weak ($1 \leq \text{BF}_{10} \leq 3$), moderate ($3 \leq \text{BF}_{10} \leq 10$), and strong ($\text{BF}_{10} \geq 10$) evidence against H_0, respectively (Jeffreys, 1961; Kass and Raftery, 1995). Indeed, as the plot shows, the results are sensitive to the choice of μ_s. Setting the expected value of side effects to 1% results in $\text{BF}_{10} = 13.45$, yielding strong evidence in favor of H_1, while setting it to 50% results in $\text{BF}_{01} = 2.66$, yielding weak evidence in favor of H_0.

We now conduct a sensitivity analysis with respect to both hyperparameters simultaneously. Figure 3b shows the contour lines of BF_{10} as a function of $(\mu_e, \mu_s) \in (0, 1)^2$ over their full range of possible values, while keeping $n_e = n_s = 1$ fixed. Overall, only when (i) side effects are expected to be small ($< 1\%$), and (ii) the efficacy is expected to be relatively large (between 30% and 70%), does the Bayes factor provide strong evidence against the null of no effect. For all other combinations of prior hyperparameters, the evidence is either moderate, weak, or favors the null. In this light, the results of the



(a) Sensitivity plot varying μ_s and n_s. (b) Sensitivity contours varying μ_s and μ_e.

Figure 3: Sensitivity analysis of BF_{10} for the aspirin trial.

trial are ambiguous, and the conclusion that aspirin is effective for primary prevention of fatal heart attack strongly depends on the prior. Note that this need not always be the case, as we show in our reanalysis of the Pfizer-BioNTech COVID-19 vaccine trial.

Combining data from multiple trials. Following the PHS, numerous subsequent trials in different study populations have later found mixed evidence for a reduction in cardiovascular events due to aspirin, along with increased risk of major hemorrhage (Ridker et al., 2005; Gaziano et al., 2018; ASCEND Study Collaborative Group, 2018) and, in older age groups, increased all-cause mortality (McNeil et al., 2018). Consequently, several organizations recommended against aspirin therapy for primary prevention of cardiovascular disease in elderly patients (Arnett et al., 2019; Davidson et al., 2022). In light of these findings, we now demonstrate how to pool evidence across multiple trials using the BREASE approach. Specifically, we focus on the risk of myocardial infarction (both fatal and non fatal), combining data from thirteen trials as analyzed in Zheng and Roddick (2019), encompassing a total of 161,680 participants.

Starting with a *complete pooling* analysis, the default BREASE prior yields a posterior median for the risk ratio of 0.90, with CrI(95%)=[0.84, 0.97]. The Bayes factor is 2.43, indicating only weak evidence against the null hypothesis. Despite the large sample size, results are still very sensitive to the prior. For example, the 95% credible interval includes the null of 1 as soon as $\mu_s < 0.1$. Next we apply a hierarchical BREASE prior, as discussed in Section 3.5, to *partially pool* information across studies. We set a BREASE(λ, ν) prior on the hierarchical proportions (μ_0, μ_e, μ_s), with $\lambda = (.5, .5, .5)$, $\nu = (10, 10, 10)$, and independent Gamma(10,.1) priors on (n_0, n_e, n_s). As Table 1 of Supplement A shows, there is considerable effect heterogeneity across trials. The posterior median for the average effect is 0.9, with CrI(95%)=[0.78, 1.13].

4.2 The Pfizer-BioNTech COVID-19 vaccine trial

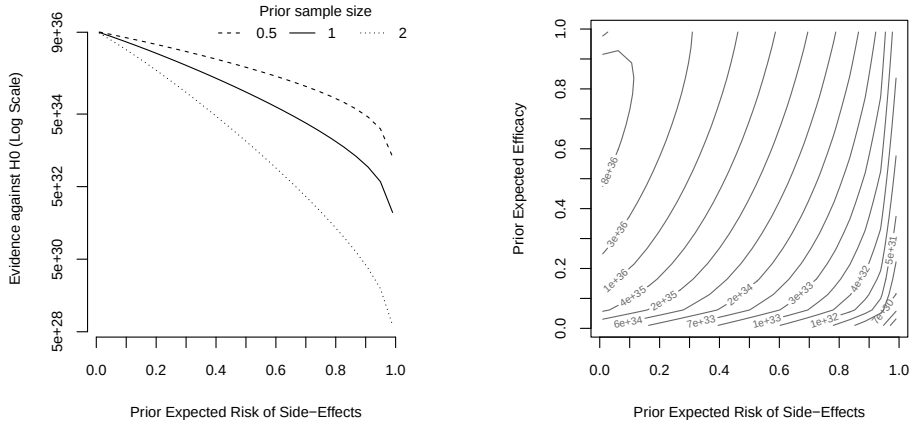
We now reexamine the results of the Pfizer-BioNTech mRNA COVID-19 vaccine study (Polack et al., 2020). The experiment was a global multi-phase randomized placebo-controlled trial designed, in part, to evaluate the efficacy of the BNT162b2 vaccine candidate in preventing COVID-19. Vaccine development and evaluation were carried out in rapid response to the emerging SARS-CoV-2 pandemic. The results of the trial were definitive and precipitated the U.S. Food and Drug Administration's emergency use authorization for widespread dissemination of the vaccine (U.S. Food and Drug Administration, 2020).

During the study, $y_1 = 9$ out of $N_1 = 19{,}965$ subjects contracted COVID-19 subsequent to the second dose of the vaccine, while there were $y_0 = 169$ cases out of $N_0 = 20{,}172$ subjects receiving placebo injections. In their paper, Polack et al. adopted a Bayesian approach, focusing particularly on evaluating the vaccine efficacy (VE), defined in the study as the estimand $\text{VE} := 1 - \theta_1/\theta_0$. The efficacy of the vaccine was estimated at 0.95, with credible interval $\text{CrI}(95\%) = [0.90, 0.97]$. Frequentist estimates are similar, with a point estimate of 0.95, confidence interval $\text{CI}(95\%) = [0.90, 0.97]$, and a p-value for testing the null hypothesis of zero effect of the order 6×10^{-33}.

Polack et al. (2020) estimate VE as the efficacy of the vaccine, but, as per Section 3.1, this only has the counterfactual interpretation of efficacy (i.e., $\eta_e = 1 - \theta_1/\theta_0$) under the assumption of monotonicity. Using the BREASE approach we can easily encode the monotonicity assumption by setting $\eta_s = 0$ and then proceed with estimation. The default BREASE prior, with the monotonicity constraint, results in posterior median and 95% credible interval for $\eta_e = 1 - \theta_1/\theta_0$ that are essentially the same as the previous results, namely, 0.94 and $\text{CrI}(95\%) = [0.90, 0.97]$. In the absence of the monotonicity assumption, we have that VE is in fact a lower bound on η_e. Again using the default BREASE prior, results are virtually unchanged, with posterior median and 95% credible interval for VE of 0.94 and $\text{CrI}(95\%) = [0.90, 0.97]$. Conclusions from the IB(1,1;1,1) prior are practically equivalent: the posterior median of VE is 0.94 with $\text{CrI}(95\%) = [0.90, 0.97]$. Under the LT(0,0;1,1) prior, however, we obtain posterior median 0.91 and $\text{CrI}(95\%) = [0.86, 0.95]$, owing to the fact that it not only shrinks θ_0 and θ_1 toward each other, but also toward 0.5—see Figure 3 of Dablander et al. (2022).

Turning to hypothesis testing, differently from the aspirin study, here all approaches point to the same direction, with overwhelming evidence against H_0. The Bayes factors against the null hypothesis of zero effect are 9×10^{33}, 5×10^{34} and 4×10^{35} for the IB, LT and BREASE default priors, respectively. Further, sensitivity analyses reveal the Bayes factor is in fact robust to variations in the hyperparameters across the whole range of prior expected efficacy and side effects of the vaccine, i.e., $(\mu_e, \mu_s) \in (0, 1)^2$. Figure 4 replicates the same sensitivity plots of the aspirin study for the COVID-19 trial. Notice that, in all scenarios, the posterior probability of the null hypothesis is essentially zero even if we posit equal prior odds for H_0 and H_1.

Conditional vaccine efficacy. In addition to overall VE for their sample, Polack et al. (2020) report estimates of VE across subgroups stratified by age, sex, race, ethnicity,



(a) Sensitivity plot varying μ_s and n_s.

(b) Sensitivity contours varying μ_s and μ_e.

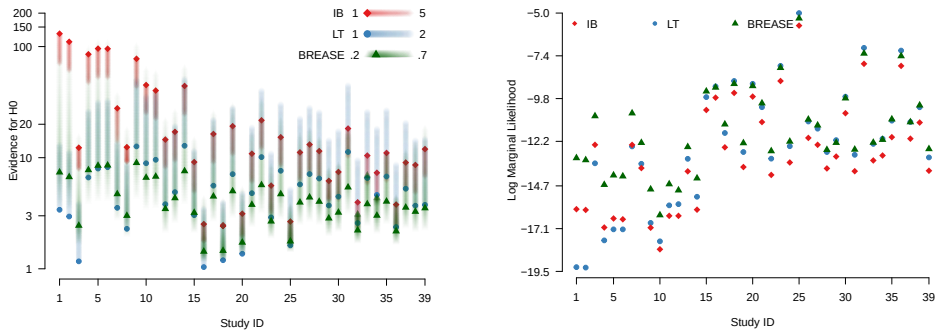
Figure 4: Sensitivity analysis of BF_{10} for the COVID-19 vaccine trial.

and country. In many subgroups, sample sizes were too small to establish efficacy of the vaccine at the 30% threshold prespecified by Polack et al. (2020). For example, in the oldest age group of individuals 75 years or older—who face the greatest risk of death from COVID-19—the 95% CrI for VE reported by Polack et al. (2020) ranges from -13.1% to 100.0%, which allows for the possibility that vaccination increases the risk of infection. Similarly, an age-stratified analysis using the independent BREASE prior discussed in Section 3.5 with our choice of default hyperparameters yields a 95% credible interval ranging from -10.8% to 99.4% for this age group.

The situation improves if we allow for some pooling of information across age groups using a hierarchical prior, as described in Section 3.5. Here we use the same hyperparameters as discussed in the aspirin example. Table 2 in Supplement A reports estimates of the Pfizer-BioNTech COVID-19 vaccine efficacy stratified by age, race, and country using the independent and hierarchical BREASE priors. With partial pooling, VE in the 75 and older age group now ranges from 45.0% to 97%, surpassing the 30% threshold.

4.3 Null results in the *New England Journal of Medicine*

Dablander et al. (2022) conducted a Bayesian reanalysis of 39 binary experiments reporting null results (claiming absence or nonsignificance of an effect of treatment) in the *New England Journal of Medicine* (NEJM). They were particularly concerned with distinguishing between *absence of evidence* and *evidence of absence* of an effect when outcomes in the treatment and control groups are similar. Finding that Bayes factors calculated using the IB approach often strongly favored the null hypothesis (leaning heavily toward *evidence of absence*) whereas LT Bayes factors were generally equivocal, Dablander et al. concluded that the LT approach should be preferred for Bayesian tests for an equality of proportions. In our final empirical example, we expand their reanalysis



(a) Bayes factors.

(b) Log marginal likelihoods.

Figure 5: Comparisons of log marginal likelihoods and Bayes factors across 39 NEJM studies, for the IB, LT and BREASE priors.

to include the BREASE approach, and we show how it can easily address the concerns of Dablander et al. while also providing a better fit to the data in most cases.

Figure 5a contrasts the Bayes factors in favor of the null hypothesis using: (i) the $\text{IB}(a, a; a, a)$ prior varying $a \in [1, 5]$ (red diamonds); (ii) the $\text{LT}(0, 0; 1, \sigma_\psi)$ prior varying $\sigma_\psi \in [1, 2]$ (blue circles); and, the $\text{BREASE}(1/2, \mu, \mu; 2, 1, 1)$ prior varying $\mu \in [.2, .7]$ (green triangles). The solid color stands for the proposed default values of each method, namely $a = 1$ for the IB, $\sigma_\psi = 1$ for the LT and $\mu = .3$ for the BREASE. Note that the Bayes factors of the BREASE and LT default priors (solid triangle and circles) are similar across studies. Moreover, Dablander et al. (2022) noted that, in many examples, the Bayes factors of the IB and LT approaches could not be easily reconciled, even when reasonably varying their hyperparameters. The BREASE approach shows that this behavior is a mere artifact of those parameterizations. Indeed, for all studies, the BREASE prior easily interpolates between the two regimes, thus solving the apparent contradiction between the results of the LT and IB approaches. Finally, Figure 5b compares the predictive performance of the default IB, LT, and BREASE priors via the log marginal likelihood. The BREASE prior exhibits superior performance in *every study* when compared to the IB prior, and in more than 74% of the studies when compared to the LT prior. Thus, in this setting, our default prior provides both a more sensible parameterization and a better fit to the data.

5 Conclusion

We have introduced the BREASE framework for the Bayesian analysis of randomized controlled trials with a binary treatment and outcome. Framing the problem in the language of potential outcomes, we reparameterized the likelihood in terms of clinically meaningful quantities—the baseline risk, efficacy, and risk of adverse side effects

of the treatment—and proposed a simple, yet flexible jointly independent beta prior distribution on these parameters. We provided algorithms for exact posterior sampling, an accurate and fast data-augmented Gibbs sampler, as well as analytical formulae for marginal likelihoods, Bayes factors, and other quantities. Finally, we showed with three empirical examples how our proposal facilitates estimation, hypothesis testing, and sensitivity analysis of treatment effects in binary experiments.

Many interesting extensions of our framework are possible. One interesting direction is to incorporate continous covariates in the model. For example, one possibility is to model BREASE parameters as functions of covariates on the logit scale, and use a Gibbs sampler that alternates between our data-augmentation algorithm for the BREASE parameters, and a specialized algorithm for logistic models, such as the Pólya-Gamma augmentation of Polson et al. (2013). Another important avenue for future work is handling noncompliance in clinical trials. In Section 8 of Supplement A, we lay the groundwork for such extension, and show how the joint distribution of compliance and response types is naturally amenable to the BREASE parameterization and prior.

Beyond binary experiments, we may also consider trials with nonbinary outcomes or more than two arms. For example, with ordinal outcomes, one option is to replace η_e and η_s with the probability that treatment improves the outcome by one step and worsens the outcome by one step, respectively. In trials with more than two arms, we may again define the baseline risk θ_0 in the control or standard of care group. Then, for each treatment arm z, we can introduce treatment-specific efficacy and side effect parameters, η_e^z and η_s^z, respectively, and again place independent beta priors on each parameter to yield a tractable mixture posterior. If the treatments share some feature—e.g., they derive from a common family of therapeutics—and we have reason to believe that treatment effects are similar, we could instead place hierarchical priors on η_e^z and η_s^z to partially pool information across treatment arms, as described in Section 3.5.

Finally, while we have demonstrated how to apply our framework to pool evidence across multiple trials, many interesting questions remain open in that area. For example, under certain assumptions, data from multiple sites may allow one to point identify, or at least narrow the bounds on the fraction of people who benefit from or are harmed by the intervention. These counterfactual probabilities play an important role in public health and legal contexts. In a similar vein, another possibility is to study our framework in the context of crossover trials. Under certain assumptions of temporal homogeneity, the efficacy and side effects of the treatment may again be identifiable, making our parameterization and prior proposal natural candidates for the study of treatment effects in such designs.

Acknowledgments. Irons's research was supported by a Shanahan Endowment Fellowship and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, T32 HD101442-01, to the Center for Studies in Demography & Ecology at the University of Washington. Cinelli's research was supported in part by the Royalty Research Fund at the University of Washington, and by the National Science Foundation under Grant No. MMS-2417955.

References

Agresti, A. and Hitchcock, D. B. (2005). "Bayesian inference for categorical data analysis." *Statistical Methods and Applications*, 14(3): 297–330. 5

Agresti, A. and Min, Y. (2005). "Frequentist Performance of Bayesian Confidence Intervals for Comparing Proportions in 2 × 2 Contingency Tables." *Biometrics*, 61(2): 515–523. 2, 4

Antelman, G. R. (1972). "Interrelated Bernoulli Processes." *Journal of the American Statistical Association*, 67(340): 831–841. 3

Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., Himmelfarb, C. D., Khera, A., Lloyd-Jones, D., McEvoy, J. W., et al. (2019). "2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines." *Circulation*, 140(11): e596–e646. 21

ASCEND Study Collaborative Group (2018). "Effects of aspirin for primary prevention in persons with diabetes mellitus." *New England Journal of Medicine*, 379(16): 1529–1539. 21

Basu, D. (1977). "On the Elimination of Nuisance Parameters." *Journal of the American Statistical Association*, 72(358): 355–366.
URL http://www.jstor.org/stable/2286800 9

Bayes, T. (1763). "An essay toward solving a problem in the doctrine of chances, with Richard Price's foreword and discussion." *Philos. Trans. R. Soc. London*, 53: 370–418. 2

Branscum, A., Gardner, I., and Johnson, W. (2005). "Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling." *Preventive veterinary medicine*, 68(2-4): 145–163. 3

Casella, G. and Moreno, E. (2009). "Assessing Robustness of Intrinsic Tests of Independence in Two-Way Contingency Tables." *Journal of the American Statistical Association*, 104(487): 1261–1271. 11

Chickering, D. M. and Pearl, J. (1996). "A Clinician's Tool for Analyzing Noncompliance." *Proceedings of the AAAI Conference on Artificial Intelligence, 13*. 2, 7

Cinelli, C. and Pearl, J. (2021). "Generalizing experimental results by leveraging knowledge of mechanisms." *European Journal of Epidemiology*, 36: 149–164. 8

Copas, J. B. (1973). "Randomization models for the Matched and Unmatched 2 × 2 Tables." *Biometrika*, 60(3): 467–476.
URL http://www.jstor.org/stable/2334995 6

Dablander, F., Huth, K., Gronau, Q. F., Etz, A., and Wagenmakers, E.-J. (2022). "A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions." *Statistics in Medicine*, 41(8): 1319–1333. 2, 4, 5, 22, 23, 24

Davidson, K. W., Barry, M. J., Mangione, C. M., Cabana, M., Chelmow, D., Coker, T. R., Davis, E. M., Donahue, K. E., Jaén, C. R., Krist, A. H., et al. (2022). "Aspirin use to prevent cardiovascular disease: US Preventive Services Task Force recommendation statement." *JAMA*, 327(16): 1577–1584. 19, 21

Davies, H., Crombie, I., and Tavakoli, M. (1998). "When can odds ratios mislead?" *BMJ*, 316(7136): 989–991. 2

Dickey, J. M. (1983). "Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses." *Journal of the American Statistical Association*, 78(383): 628–637. 3, 10

Dickey, J. M., Jiang, J. M., and Kadane, J. B. (1987). "Bayesian methods for censored categorical data." *Journal of the American Statistical Association*, 82: 773–781. 3, 10

Dickey, J. M. and Lientz, B. P. (1970). "The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain." *The Annals of Mathematical Statistics*, 41(1): 214–226. 15

Ding, P. and Miratrix, L. W. (2019). "Model-free causal inference of binary experimental data." *Scandinavian Journal of Statistics*, 46(1): 200–214. 2

Evans, M. and Moshonov, H. (2006). "Checking for Prior-Data Conflict." *Bayesian Analysis*, 1(4): 893–914. 14

Gaziano, J. M., Brotons, C., Coppolecchia, R., Cricelli, C., Darius, H., Gorelick, P. B., Howard, G., Pearson, T. A., Rothwell, P. M., Ruilope, L. M., et al. (2018). "Use of aspirin to reduce risk of initial vascular events in patients at moderate risk of cardiovascular disease (ARRIVE): a randomised, double-blind, placebo-controlled trial." *The Lancet*, 392(10152): 1036–1046. 21

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC. 2

Greenland, S. and Robins, J. (1986). "Identifiability, Exchangeability, and Epidemiological Confounding." *International Journal of Epidemiology*, 15(3): 413–419. 6

Gronau, Q. F., Raj, K. N. A., and Wagenmakers, E.-J. (2021). "Informed Bayesian Inference for the A/B Test." *Journal of Statistical Software*, 100(17): 1–39. 5

Gunel, E. and Dickey, J. (1974). "Bayes Factors for Independence in Contingency Tables." *Biometrika*, 61(3): 545–557. 11

Gustafson, P. (2015). *Bayesian inference for partially identified models: Exploring the limits of limited data*, volume 140. CRC Press. 9

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X.-H. (2000). "Assessing the effect of an influenza vaccine in an encouragement design." *Biostatistics*, 1(1): 69–88. 7

Imbens, G. W. and Rubin, D. B. (1997). "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *The Annals of Statistics*, 25(1): 305–327. 2, 7

Jeffreys, H. (1935). "Some Tests of Significance, Treated by the Theory of Probability." *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2): 203–222. 5

— (1961). *Theory of Probability*. Oxford, UK: Oxford University Press, 3rd edition. 11, 15, 20

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90(430): 773–795. 5, 15, 20

Kass, R. E. and Vaidyanathan, S. K. (1992). "Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions." *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1): 129–144. 5

Kaufman, G. M. and King, B. (1973). "A Bayesian Analysis of Nonresponse in Dichotomous Processes." *Journal of the American Statistical Association*, 68(343): 670–678. 3

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Elsevier Science & Technology. 2

Laplace, P. S. (1774). "Mémoire sur la probabilité de causes par les évenements." *Mémoire de l'académie royale des sciences*. 2

Li, F., Ding, P., and Mealli, F. (2023). "Bayesian causal inference: a critical review." *Phil. Trans. R. Soc. A.*, 381(20220153). 2

Linero, A. R. (2023a). "In nonparametric and high-dimensional models, Bayesian ignorability is an informative prior." *Journal of the American Statistical Association*, 1–14. 3

— (2023b). "Prior and posterior checking of implicit causal assumptions." *Biometrics*, 79(4): 3153–3164. 3

Madigan, D. (1999). "Bayesian graphical models, intention-to-treat, and the Rubin causal model." In *Seventh International Workshop on Artificial Intelligence and Statistics*. PMLR. 7

McElreath, R. (2020). *Statistical rethinking : a Bayesian course with examples in R and Stan*. Texts in statistical science. Chapman & Hall/CRC, second edition. edition. 2

McNeil, J. J., Nelson, M. R., Woods, R. L., Lockery, J. E., Wolfe, R., Reid, C. M., Kirpach, B., Shah, R. C., Ives, D. G., Storey, E., et al. (2018). "Effect of aspirin on all-cause mortality in the healthy elderly." *New England Journal of Medicine*, 379(16): 1519–1528. 21

Neyman, J. (1990). "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5(4): 465–472. Translated from the 1923 Polish original and edited by D. M. Dabrowska and T. P. Speed. 4

Ng, K. W., Tian, G.-L., and Tang, M.-L. (2011). *Dirichlet and related distributions: Theory, methods and applications*. Wiley. 10

Pearl, J. (2009). *Causality*. Cambridge University Press. 7

Pham-Gia, T. and Turkkan, N. (1998). "Distribution of the linear combination of two general beta variables and applications." *Communications in Statistics - Theory and Methods*, 27(7): 1851–1869. 10

Plummer, M. (2023). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-14. 14

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., Türeci, O., Nell, H., Schaefer, A., Ünal, S., Tresnan, D. B., Mather, S., Dormitzer, P. R., Şahin, U., Jansen, K. U., and Gruber, W. C. (2020). "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine." *New England Journal of Medicine*, 383(27): 2603–2615. 22, 23

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American statistical Association*, 108(504): 1339–1349. 25

Richardson, T. S., Evans, R. J., and Robins, J. M. (2011). "Transparent Parametrizations of Models for Potential Outcomes." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics*, volume 9. Oxford, UK: Oxford University Press. 9

Ridker, P. M., Cook, N. R., Lee, I.-M., Gordon, D., Gaziano, J. M., Manson, J. E., Hennekens, C. H., and Buring, J. E. (2005). "A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women." *New England Journal of Medicine*, 352(13): 1293–1304. 21

Robins, J. and Wasserman, L. (2012). "Robins and Wasserman Respond to a Nobel Prize Winner." *Normal Deviate Blog*. Accessed September 3, 2024. URL https://normaldeviate.wordpress.com/2012/08/28/robins-and-wasserman-respond-to-a-nobel-prize-winner/ 3

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, 66(5): 688. 4

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). "Robust meta-analytic-predictive priors in clinical trials with historical control information." *Biometrics*, 70(4): 1023–1032. 10, 14, 15

Smith, S. W., Hauben, M., and Aronson, J. K. (2012). "Paradoxical and bidirectional drug effects." *Drug safety*, 35: 173–189. 8

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. (1994). "Bayesian approaches to randomized trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 157(3): 357–387. 3

Springer, M. D. and Thompson, W. E. (1970). "The Distribution of Products of Beta, Gamma and Gaussian Random Variables." *SIAM Journal on Applied Mathematics*, 18(4): 721–737. 11

Stan Development Team (2023). "RStan: the R interface to Stan." R package version 2.21.8. 14

Steering Committee of the Physicians' Health Study Research Group (1989). "Final Report on the Aspirin Component of the Ongoing Physicians' Health Study." *New England Journal of Medicine*, 321(3): 129–135. 19

Tian, G.-L., Ng, K. W., and Geng, Z. (2003). "Bayesian computation for contingency tables with incomplete cell-counts." *Statistica Sinica*, 13(1): 189–206. 10

Tian, J. and Pearl, J. (2000). "Probabilities of causation: Bounds and identification." *Annals of Mathematics and Artificial Intelligence*, 28(1): 287 – 313. 7, 8

U.S. Food and Drug Administration (2020). "Guidance for industry: emergency use authorization for vaccines to prevent COVID-19."
URL https://www.fda.gov/media/142749/download 22

Zheng, S. L. and Roddick, A. J. (2019). "Association of aspirin use for primary prevention with cardiovascular events and bleeding events: a systematic review and meta-analysis." *Jama*, 321(3): 277–287. 21