Causal Inference in the Bayesian Framework: Principles and Applications

Abstract

Causality is a fundamental concept in science, technology, and our general understanding of the world. Based on the advocacy of the U.S. Food and Drug Administration (FDA) for Bayesian methods in drug development, this study introduces the Bayesian framework into causal inference. This method provides a unified framework for causal inference based on the axioms of Bayesian statistics. It treats causal statements as hypotheses or models about the world, and its central task is to compute the posterior distribution of these hypotheses conditional on the observed data and background knowledge. Therefore, this method has wide applicability and can be applied to the research of various causal problems. This paper investigates causal inference methodologies under the Bayesian framework. It first delineates the theoretical underpinnings and subsequently employs a numerical example to elucidate how Bayesian methods quantify causal effects, highlighting their comparative advantages.

Key words: Bayesian statistical, causal inference, potential outcomes, Bayes factor.

1 Introduction

The fundamental question of causal inference is to answer: "Is it because of A that B occurred?" The core challenge lies in the fact that we can never observe the outcome of the same individual both receiving and not receiving the treatment at the same time-a dilemma known as the "fundamental problem of causal inference." Randomized Controlled Trials

(RCTs) address this issue through their randomized design, which effectively eliminates confounding factors, making RCTs the gold standard in causal inference [1].

The literature on causal effect estimation tends to focus on the population mean estimand, which is less informative as medical treatments are becoming more personalized and there is increasing awareness that subpopulations of individuals may experience a group-specific effect that differs from the population average [2]. In fact, it is possible that there is underlying systematic effect heterogeneity that is obscured by focusing on the population mean estimand. In this context, understanding which covariates contribute to this treatment effect heterogeneity (TEH) and how these covariates determine the differential treatment effect (TE) is an important consideration.

Within the framework of causal inference, researchers often rely on the potential outcomes framework to define treatment effects. As a result, causal inference can be naturally formulated as a special type of missing data problem: for each individual, some of their potential outcomes are missing [3, 4]. To address this missingness structure, conventional approaches often employ strategies such as imputation, matching, propensity score weighting, or multiple imputation to recover the missing data and thereby estimate causal effects [5]. Although these methods are widely used in practice, they have certain limitations in incorporating external knowledge, adequately representing parameter uncertainty, and providing intuitive probabilistic interpretations. In contrast, Bayesian methods offer a unified and flexible framework for handling missing data problems in causal inference [6]. As causal studies increasingly involve real-world big data, there has been a recent surge of research in Bayesian inference of causal effects [7, 8, 9, 10].

In causal inference research, one of the central challenges lies in comparing and selecting among competing models. Traditional frequentist methods typically rely on hypothesis testing and p-values to assess the adequacy of models. However, such approaches suffer from important limitations: their results only indicate the probability of observing the data under the null hypothesis and do not directly quantify the relative evidential support for different models [11]. Furthermore, issues such as multiple testing and sensitivity to sample size within the frequentist framework often undermine the robustness of conclusions, particularly in subgroup analyses. In contrast, the Bayes Factor (BF) offers a more natural and coherent framework for model comparison. It quantifies the evidence provided by the data in favor of one model over another by comparing their marginal likelihoods [12].

In conclusion, the core challenge of causal inference lies in the "absence" of potential outcomes, and the Bayesian method provides a flexible and consistent framework for address-

ing this difficult problem. Consequently, the Bayes factor, as a tool for model comparison and hypothesis testing, can effectively evaluate the relative rationality of different causal models and provide a more robust evidence basis for the heterogeneity test of therapeutic effects in clinical trials.

To achieve the above goals, the structure of this article is organized as follows: Section 2 elaborates on the theoretical basis of causal inference; Section 3 introduces the basic framework of Bayesian causal inference; Section 4 systematically expounds the methodological framework of Bayes factors in causal inference. And Section 5 summarizes the research results, explores their theoretical significance and practical value, analyzes the advantages and limitations of the current research, and suggests directions for future research.

2 Causal Estimands and Identification in the Potential Outcomes Framework

Here, we only focus on the situation of binary treatment, which can be easily extended to multiple treatments. In randomized clinical trials, consider the samples drawn from the target population, where these individuals are indexed by $i \in \{1, ..., N\}$. Each individual may be assigned to one of two treatment levels T_i: $T_i = 1$ indicates active treatment, and $T_i = 0$ indicates the control group. Let $T_i(= t)$ be the binary variable for observing the treatment status in individual i. For each individual i, there is a p dimensional covariate X_i that was observed before the treatment, and Y_i represents the potential outcome after the treatment. Lowercase t_i and x_i are implementations of their uppercase equivalents.

We maintain the standard stable unit treatment value assumption (SUTVA) [13], namely, there is (i) only a single version of each treatment level, and it is administered uniformly to all units, and (ii) the potential outcome of any unit is independent of the treatments assigned to all other units. Under SUTVA, each individual i has two potential outcomes: $Y_i(1)$ and $Y_i(0)$. We can observe only one or the other of $Y_i(1)$ and $Y_i(0)$ as indicated by T_i. Therefore, we have:

$$Y_i = Y_i(1)T_i - Y_i(0)(1 - T_i), \tag{2.1}$$

where T_i is the treatment indicator.

Causal effects are contrasts of potential outcomes under different treatment conditions for the same set of individual [14]. The individual treatment effect (ITE) for unit i is $\tau_i = Y_i(1) - Y_i(0)$. Averaging τ_i over a sample we obtain the sample average treatment effect (SATE): $\tau^S \equiv N^{-1} \sum_{i=1}^{N} \tau_i$. Furthermore, the conditional average treatment effect (CATE)

is the average of the individual treatment effect of all units with the covariate value x:

$$\tau(x) \equiv E\{Y_i(1) - Y_i(0) \mid X_i = x\} = \mu_1(x) - \mu_0(x), \tag{2.2}$$

where $\mu_t(x) \equiv E\{Y_i(t) \mid X_i = x\}$ for $t = 0, 1$. Averaging τ_i or $\tau(X_i)$ over a target population gives the population average treatment effect (PATE):

$$\begin{aligned}\tau^P \equiv E[Y_i(1) - Y_i(0)] &= E[Y_i(1)] - E[Y_i(0)] \\ &= E_{F_x}\{E[Y \mid T_i = 1, X_i = x] - E[Y \mid T_i = 0, X_i = x]\}.\end{aligned} \tag{2.3}$$

Inference for causal effects is a missing-data problem [15] - the "other" value is missing. For each unit, only the realized outcome $Y_{\text{obs},i}$ is observed, while the counterfactual outcome $Y_{\text{mis},i}$ remains unobserved. This necessitates explicit assumptions about the *assignment mechanism*—the process governing treatment allocation and consequent outcome observability [16]. Most analyses rely on some form of an *ignorable assignment mechanism*, requiring that treatment assignment T_i is conditionally independent of potential outcomes given observed covariates (i.e., $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$). Specifically, in the simple scenario of binary treatment, the ignorability mechanism contains two sub-assumptions [16, 17].

Assumption 2.1 (Ignorability)

(a) Unconfoundedness. $\Pr(T_i \mid Y_i(0), Y_i(1), X_i) = \Pr(T_i \mid X_i)$, or equivalently $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\} \mid X_i$.

(b) Overlap. $0 < e(X_i) < 1$ for all i, where $e(x) \equiv \Pr(T_i = 1 \mid X_i = x)$ is the propensity score [17].

The unconfoundedness assumption and the overlap assumption together ensure that the conditional distribution of the potential outcomes is identifiable from observed data as:

$$\mu_t(x) \equiv E\{Y_i(t) \mid X_i = x\} = E(Y_i \mid T_i = t, X_i = x), \quad \text{for all } t, x. \tag{2.4}$$

Therefore, the CATE is identified as $\tau(x) = \mu_1(x) - \mu_0(x)$, and the PATE is identified as $\tau^P = E\{\mu_1(X_i) - \mu_0(X_i)\}$.

This statement emphasizes an estimation strategy based on outcome modeling: by modeling the potential outcome function $\mu_t(x)$, we can estimate the CATE using $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$, and further derive an estimate of the PATE by $\hat{\tau}^{\text{PATE}} = N^{-1} \sum_{i=1}^{N} \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}$, where $\hat{\mu}_t(x)$ is the estimated outcome model from the observed data.

From the above, we can assume a linear regression model for equation (2.3), where the coefficients reflect treatment-covariate interactions [18]:

$$Y_i = T_i \left(\mu_1 + \sum_{j=1}^{p} \beta_{1j} X_{ij} \right) + (1 - T_i) \left(\mu_0 + \sum_{j=1}^{p} \beta_{0j} X_{ij} \right) + \sigma_{T_i} \varepsilon_i. \tag{2.5}$$

In Equation (2.5), μ_0 and μ_1 represent the main effects of the treatment modalities, respectively. $\beta_{k1}, \beta_{k2}, \cdots, \beta_{kp}$ are the corresponding values of the treatment-covariate interactions for $k = 0, 1$. ε_i is an independent and uniformly distributed random error with zero mean and unit variance, independent of X_{ij}, where $j = 1, 2, \cdots, p$. The standard deviation of heterogeneity is modeled by σ_{T_i}, where:

$$\sigma_{T_i} = \begin{cases} \sigma_1, & \text{if } T_i = 1, \\ \sigma_2, & \text{if } T_i = 0. \end{cases}$$

To calculate the expectation of Equation (2.5), we derive the model for the PATE as follows:

$$\begin{aligned} E[Y_i(1) - Y_i(0)] &= E[Y_i(1)] - E[Y_i(0)] \\ &= E_{F_X} \left\{ T_i \left(\mu_1 + \sum_{j=1}^{p} \beta_{1j} X_{ij} \right) - (1 - T_i) \left(\mu_0 + \sum_{j=1}^{p} \beta_{0j} X_{ij} \right) \right\}. \end{aligned} \tag{2.6}$$

At this point, the model of the CATE can be expressed as:

$$\tau(x) = E[Y_i(1) - Y_i(0) \mid \boldsymbol{X_i} = \boldsymbol{x_i}] = E\{T_i(\boldsymbol{\beta}_1^\top \boldsymbol{x_i}) - (1 - T_i)(\boldsymbol{\beta}_0^\top \boldsymbol{x_i}\}, \tag{2.7}$$

where $\boldsymbol{\beta}_k^\top = \{\beta_{k1}, \beta_{k2}, \cdots, \beta_{kp}\}$, $k = 0, 1$. In Equation (2.7), the CATE $\tau(x)$ is specified by a linear model, and the treatment effect within the target subgroup is captured by the corresponding coefficients. A significant discrepancy between $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ suggests that the included covariate $\boldsymbol{X_i} = \boldsymbol{x_i}$ contributes meaningfully to the heterogeneity in treatment effects.

3 General structure of Bayesian causal inference

Because of the unavoidable missing potential outcomes, causal inference under the potential outcomes framework is inherently a missing data problem [19, 20]. The Bayesian paradigm provides a unified probabilistic framework for statistical inference under missing data, thereby establishing a systematic and rigorous methodological foundation for causal reasoning [21].

3.1 The theoretical basis of Bayes

Suppose parameter $\boldsymbol{\theta}$ is an unknown quantity of interest, and let $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ denote a set of n independent random observations. In contrast to the frequentist paradigm, Bayesian methodology treats $\boldsymbol{\theta}$ as a random variable rather than a fixed unknown value. Within this framework, Bayes' rule may be expressed as follows:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \tag{3.1}$$

which is often simplified to: $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$. The term $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ represents a conditional probability, where the probability of the model parameters ($\boldsymbol{\theta}$) is computed conditional on the data ($\boldsymbol{y}$), representing the posterior distribution. The term $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ represents the conditional probability of the data given the model parameters, and this term represents the likelihood function. Finally, the term $\pi(\boldsymbol{\theta})$ represents the probability of particular model parameter values existing in the population, also known as the prior distribution. The term denominator is a normalizing factor and can be dropped from the equation as it does not depend on $\boldsymbol{\theta}$. Thus, the posterior distribution is proportional to the likelihood function multiplied by the prior distribution.

In practical situations, the elegance and flexibility of the Bayesian framework is often faced with two well-known challenges: (1) the choice of the prior degree of belief $\pi(\theta)$; and (2) the actual computation of the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, and any related expectations, which may not always be solvable analytically and may require Monte Carlo approaches [22].

3.2 A Bayesian framework for causal inference

Causal inference is fundamentally characterized as a problem of missing data, and common processing methods include interpolation and other filling strategies. In such problems, the Bayesian method demonstrates its unique advantages: it can not only flexibly integrate prior knowledge, but also naturally handle uncertainties and directly provide probabilistic explanations of parameters, thereby providing more interpretable and evidentially robust results for causal identification.

3.2.1 The posterior distribution of causal effects

The fundamental task of Bayesian causal analysis is to calculate the posterior probability of causal hypothesis given the corresponding data and background information [23]. Bayesian

inference for causal effects directly confronts the explicit missing potential outcomes,

$$\text{where } Y_{\text{mis}} = \{Y_{\text{mis},i}\}, i = 1, 2 \cdots N$$
$$\text{with } Y_{\text{mis},i} = T_i Y_i(1) + (1 - T_i) Y_i(0).$$

This method only requires the parameters of the assignment mechanism and the base data as input to derive the posterior predictive distribution of Y_{mis}, which represents the distribution of Y_{mis} given all observed values,

$$\Pr(Y_{\text{mis}} \mid X, Y_{\text{obs}}, T), \tag{3.2}$$

where $Y_{\text{obs}} = \{Y_{\text{obs},i}\}$, with $Y_{\text{obs},i} = T_i Y_i(1) + (1 - T_i) Y_i(0)$. Based on this distribution and the observed values of the potential outcomes Y_{obs} and covariate data, the posterior distribution of any causal effect can, in principle, be calculated.

If the posterior prediction distribution in the formula (3.2) is regarded as the definition of the random draw process of Y_{mis}, then this conclusion is obvious. After obtaining the sampling value of Y_{mis}, any causal effect can be directly calculated based on this sampling value as well as the observed X and Y_{obs}. By repeatedly drawing the value of Y_{mis} and calculating the corresponding causal effect for each draw, the posterior distribution of the target causal effect can be generated. Therefore, causal inference can be fully regarded as a problem of missing data, in which we multiply-impute [24] the missing potential outcomes to generate the posterior distribution of causal effects.

3.2.2 The posterior predictive distribution of missing potential outcomes under ignorability

Establishing the posterior predictive distribution of Y_{mis} under an ignorable treatment assignment mechanism. In general [25]:

$$\Pr(Y_{\text{mis}} \mid X, Y_{\text{obs}}, T) \;=\; \frac{\Pr(X, Y(0), Y(1)) \; \Pr(T \mid X, Y(0), Y(1))}{\int \Pr(X, Y(0), Y(1)) \; \Pr(T \mid X, Y(0), Y(1)) \, \mathrm{d}Y_{\text{mis}}}, \tag{3.3}$$

where $Y(1) = Y_i(1)$ and $Y(0) = Y_i(0)$. With ignorable treatment assignment, Equation (3.3) becomes:

$$\Pr(Y_{\text{mis}} \mid X, Y_{\text{obs}}, T) \;=\; \frac{\Pr(X, Y(0), Y(1))}{\int \Pr(X, Y(0), Y(1)) \, \mathrm{d}Y_{\text{mis}}}.$$

Since all information is contained within the underlying data, the unit labels can essentially be regarded as random identifiers. Consequently, the data matrix $(X, Y(0), Y(1))$ is row exchangeable. Thus, without any substantial loss of generality, according to de Finetti's

[26] theorem, we may treat the distribution of $(X, Y(0), Y(1))$ as independent and identically distributed (i.i.d.) given some parameter θ:

$$\Pr(X, Y(0), Y(1)) = \int \left[\prod_{i=1}^{N} f(X_i, Y_i(0), Y_i(1) \mid \theta) \right] p(\theta) \mathrm{d}\theta, \tag{3.4}$$

where $p(\theta)$ is some prior distribution of θ. Equation (3.4) establishes a bridge between foundational theory and the practical use of i.i.d. models.

3.2.3 Analytical solution for causal effects in a normal model with no covariates

Consider a completely randomized experiment with no covariates and a scalar outcome variable. The causal estimand of interest is the mean difference between $Y(1)$ and $Y(0)$ across all individuals, denoted as $\bar{Y}_1 - \bar{Y}_0$. Under this setting, we have:

$$\Pr(Y) = \int \prod_{i=1}^{N} f\big(Y_i(0), Y_i(1) \mid \theta\big) p(\theta) \mathrm{d}\theta, \tag{3.5}$$

where $f(\cdot|\theta)$ is a bivariate density function indexed by a parameter θ, and $p(\theta)$ denotes the prior distribution of θ. If we further assume that $f(\cdot|\theta)$ follows a bivariate normal distribution with means $\mu = (\mu_1, \mu_0)$, variances (σ_1^2, σ_0^2) and correlation coefficient ρ.

Under the conditions of given parameter θ, observed values of Y (i.e., Y_{obs}), and observed treatment assignment variable T, where the number of units with treatment level $T_i = k$ is n_k $(k = 0, 1)$, and provided that $n_0 + n_1 = N$, the joint distribution of $(\bar{Y}_1, \bar{Y}_0)$ follows a bivariate normal distribution with means

$$\frac{1}{2}\left[\bar{y}_1 + \mu_1 + \rho\frac{\sigma_1}{\sigma_0}(\bar{y}_0 - \mu_0)\right],$$

$$\frac{1}{2}\left[\bar{y}_0 + \mu_0 + \rho\frac{\sigma_0}{\sigma_1}(\bar{y}_1 - \mu_1)\right],$$

variances $\sigma_1^2(1-\rho^2)/4n_0, \sigma_0^2(1-\rho^2)/4n_1$, and zero correlation, where $\bar{y}_1$ and $\bar{y}_0$ are the observed sample means of Y in the two treatment groups. To facilitate comparison with conventional results, assume that N is sufficiently large and the prior distribution for $(\mu_1, \mu_0, \sigma_1^2, \sigma_0^2)$ given ρ is relatively diffuse. Under these conditions, the conditional posterior distribution of $\bar{Y}_1 - \bar{Y}_0$ given ρ is normal, with expectation:

$$E[\bar{Y}_1 - \bar{Y}_0 | Y_{\text{obs}}, T, \rho] = \bar{y}_1 - \bar{y}_0 \tag{3.6}$$

and variance

$$V[\bar{Y}_1 - \bar{Y}_0 | Y_{\text{obs}}, T, \rho] = \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0} - \frac{1}{N}\sigma^2_{(1-0)}, \tag{3.7}$$

where $\sigma^2_{(1-0)}$ denotes the prior variance of the individual treatment effect $Y_i(1) - Y_i(0)$, $\sigma_1^2 + \sigma_0^2 - 2\sigma_1\sigma_0\rho$. Although Equations (3.6) and (3.7) provide explicit analytical solutions, simulation methods are often employed in practice. Simulation serves as a more versatile tool, capable of handling more complex model specifications and offering broader applicability than closed-form analytical approaches.

Since there is no information about ρ in the observed data and the correlations among the potential results have never been observed simultaneously, to conservatively infer $\bar{Y}_1 - \bar{Y}_0$, we take $\sigma^2_{(1-0)} = 0$.

3.2.4 Analytical solution for causal effects in a normal model with covariates

Consider a completely randomized experiment with covariates X. Assume the true model for potential outcomes is

$$\begin{pmatrix} Y_i(1) \\ Y_i(0) \end{pmatrix} |(X_i, \beta_1, \beta_0, \sigma_1^2, \sigma_0^2, \rho) \sim N \begin{pmatrix} \beta_1' X_i \\ \beta_0' X_i \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix}, \quad i = 1, \ldots, N.$$

This model implies two univariate normal marginal models: $Y_i(t)|X_i, \beta_t, \sigma_t^2 \sim N(\beta_t' X_i, \sigma_t^2)$ for $t = 0, 1$. In this example, suppose the causal estimation we are interested in is still the mean difference between $Y(1)$and $Y(0)$ among all individuals, expressed as $\bar{Y}_1 - \bar{Y}_0$. In a completely randomized experiment, where treatment assignment is independent of potential outcomes, and under the model assumption of linear conditional means, we have:

$$E[\bar{Y}_1 - \bar{Y}_0] = (\beta_1 - \beta_0)' E(X_i),$$

Given covariates X and model parameters, the conditional variance of $\bar{Y}_1 - \bar{Y}_0$ is:

$$Var(\bar{Y}_1 - \bar{Y}_0 | X, \beta_1, \beta_0, \sigma_1^2, \sigma_0^2, \rho) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} - 2\rho\sigma_1\sigma_0 \cdot \frac{1}{N}$$

where n_1, n_0 are the sample sizes of the treatment and control groups respectively ($n_1 + n_0 = N$), ρ is the correlation coefficient between potential outcomes $Y_i(1)$ and $Y_i(0)$. If we ignore the finite population correction (when N is large), the variance can be simplified as:

$$Var(\bar{Y}_1 - \bar{Y}_0) \approx \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$$

In fact, the unique value of Bayesian methods in causal inference is usually fully demonstrated when dealing with complex data structures or models. In the standard and simple context, the analysis results are often highly consistent with the conclusions drawn by the frequentism method.

4 Bayesian hypothesis testing method and causal effect analysis

A primary concern in clinical trials is assessing the evidence for a treatment's effectiveness. The traditional frequentist approach, which relies on randomization and significance testing to establish causality [1]. In contrast, the Bayesian framework offers a fundamentally different paradigm. It utilizes the Bayes factor to evaluate and compare causal models, thereby enabling quantitative inference based on the strength of evidence [27].

4.1 Conceptual and theoretical foundations of Bayes factor hypothesis testing

Mathematically, the Bayes factor represents the ratio of posterior to prior odds for H_0 versus H_1, precisely measuring how much the observed data should shift our belief between the competing hypotheses [12]. Importantly, this evidentiary measure maintains complete symmetry - the evidence may favor either H_0 or H_1 equally, as neither hypothesis holds a privileged position in the analysis.

Consider the general case of two competing hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \bigcup \Theta_1 = \Theta$ and $\Theta_0 \bigcap \Theta_1 = \Phi$. Bayesian hypothesis testing requires prior probabilities $\pi_0 = \boldsymbol{P}(H_0)$ and $\pi_1 = \boldsymbol{P}(H_1)$ that sum to 1. Most people just use $\pi_0 = \pi_1 = 1/2$. For observable data $\boldsymbol{y}$, denote the likelihood functions under the two hypotheses by $p(\boldsymbol{y} \mid \theta, H_0)$ and $p(\boldsymbol{y} \mid \theta, H_1)$, and $p(\theta \mid H_k)$ is the prior of the parameter θ if H_k is true. For each $k = 0, 1$, the marginal likelihood is

$$p(\boldsymbol{y} \mid H_k) = \int_{\Theta_k} p(\boldsymbol{y} \mid \theta, H_k) p(\theta \mid H_k) d\theta,$$

which is called the prior predictive distribution of $\boldsymbol{y}$ conditional on H_k. According to Bayes' rule, the posterior probability of H_k given the observable data $\boldsymbol{y}$ is

$$p(H_k \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid H_k)\pi_k}{p(\boldsymbol{y} \mid H_0)\pi_0 + p(\boldsymbol{y} \mid H_1)\pi_1} := p_k.$$

Computing the posterior odds on H_1 against H_0 gives the equation

$$\frac{p_1}{p_0} = \frac{p(\boldsymbol{y} \mid H_1)}{p(\boldsymbol{y} \mid H_0)} \times \frac{\pi_1}{\pi_0}.$$

Then the Bayes factor in favour of H_1 against H_0, denoted BF_{10}, is defined by

$$BF_{10} = \frac{p(\boldsymbol{y} \mid H_1)}{p(\boldsymbol{y} \mid H_0)} = \frac{p_1/p_0}{\pi_1/\pi_0} = \frac{p_1\pi_0}{p_0\pi_1}. \tag{4.1}$$

It follows that Bayes factor BF_{10} represents the likelihood ratio of H_1 relative to H_0, as determined by the observed data. If the priors of the hypotheses are set to $\pi_0 = \pi_1 = 1/2$, then BF_{10} equals the posterior odds of the hypotheses.

The Bayes factor offers a direct quantitative evaluation of competing hypotheses and serves as a robust measure of evidential strength. Different magnitudes of the Bayes factor indicate varying degrees of support: (1) if $BF_{10} > 1$, the data support H_1 over H_0, (2) if $BF_{10} < 1$, the data favor H_0 over H_1, and (3) if $BF_{10} = 1$, the evidence is equivocal, providing equal support for both hypotheses. Beyond merely indicating the relative strength of evidence, the Bayes factor explicitly quantifies the degree of support. Following established conventions [28], Table 1 presents a classification framework for interpreting Bayes factor values.

Table 1: Jeffreys' scale of evidence for interpreting Bayes factor BF_{10}.

Bayes factor BF_{10}	Interpretation
> 100	Extreme evidence for H_1
30 - 100	Very strong evidence for H_1
10 - 30	Strong evidence for H_1
3 - 10	Moderate evidence for H_1
1 - 3	Anecdotal evidence for H_1
1	No evidence
1/3 - 1	Anecdotal evidence for H_0
1/10 - 1/3	Moderate evidence for H_0
1/30 - 1/10	Strong evidence for H_0
1/100 - 1/30	Very strong evidence for H_0
$< 1/100$	Extreme evidence for H_0

Jeffreys' scale of evidence enables researchers to qualitatively assess the strength of evidence for either the null or alternative hypothesis, using predefined thresholds that reflect varying degrees of evidential support. For example, a Bayes factor of 5 indicates that the data support H_1 five times more strongly than H_0, while a Bayes factor of 0.2 suggests that the data favor H_0 five times more than H_1.

4.2 Bayesian hypothesis testing versus null hypothesis significance testing

In clinical trials, Bayes factors and Null Hypothesis Significance Testing (NHST) are two widely utilized statistical inference tools, each possessing distinct applications and interpretations in the context of hypothesis testing. The conventional method of statistical inference is the NHST, which utilizes the P value to determine whether to reject the null hypothesis (H_0) or accept the alternative hypothesis (H_1). The P value represents a specific observed value or a more extreme value that occurs under the assumption that the null hypothesis is true. It quantifies how anomalous the data are under this null hypothesis, reflecting evidence against it. A smaller P value provides stronger evidence against the null hypothesis, as shown in Table 2.

Table 2: Fisher's scale of evidence against null hypothesis H_0 and in favor of H_1, as a function of coverage level (1 minus the P value).

Coverage	(P-value)	Evidence for H_1
.80	(.20)	null
.90	(.10)	borderline
.95	(.05)	moderate
.975	(.025)	substantial
.99	(.01)	strong
.995	(.005)	very strong
.999	(.001)	overwhelming

Each method has unique advantages and disadvantages [29]. The widespread misuse and misinterpretation of P value in NHST have drawn increasing criticism in recent years [11, 30, 31]. This has prompted a shift toward more comprehensive analytical approaches incorporating effect sizes, confidence intervals, and Bayesian methods.

An increasing number of studies highlight the practical advantages of Bayesian methods, responding to the overly simplistic criticism of the frequentist methodology [32]. Nonetheless, it would be unwise to completely overlook the P value, as it requires thorough analysis. A low P value does not definitively disprove the null hypothesis; instead, it could be explained by random variations or additional influencing factors. In certain scenarios, Bayes factors provide deeper and more enlightening perspectives, especially in the area of model evaluation, where Bayesian approaches exhibit significant competence in systematically and effectively

combining evidence [33, 6].

4.3 Apply Bayesian inference to quantify causal effects

For example, consider the 2×2 contingency table (Table3) reporting the results of a fictitious double-blind randomized experiment involving 200 individuals, with 100 given aspirin tablets (the treatment), and 100 chalk tablets (the control) [34]. D_1 and D_2 represent control and treatment group data, respectively. Participants took the assigned formulation upon headache onset, and headache relief time was recorded. Recovery is interpreted as (for instance) "headache disappears within 30 min", was used to evaluate the causal relationship between aspirin and headache relief, possibly in the forms:

- **Null hypothesis (H_0)**: There is no causal relationship between taking aspirin and headache relief.
- **Alternative hypothesis (H_1)**: There is a causal relationship between taking aspirin and headache relief.

Table 3: Contingency table.

	No recovery	**Recovery**	**Total**
Chalk (D_1)	85	15	100
Aspirin (D_2)	63	37	100

To establish a Bayesian probabilistic model for the data in Table 3, parameterized by two key variables: p denotes the conditional probability of headache recovery due to other factors (combining natural recovery and placebo effects), while q represents the conditional probability of headache recovery attributable to Aspirin.

Given that both p and q follow binomial distributions, we naturally assign Beta priors to these parameters, so that

$$\begin{aligned} \pi(p) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, \\ \pi(q) &= \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} q^{c-1}(1-q)^{d-1}, \end{aligned} \tag{3.9}$$

with $a > 0$, $b > 0$, $c > 0$ and $d > 0$. The data D_1 does not contain information about q but it immediately leads to a marginal Beta posterior distribution for p

$$P(p|D_1) = \frac{\Gamma(a+b+100)}{\Gamma(a+15)\Gamma(b+85)} p^{a+14}(1-p)^{b+84}, \tag{3.10}$$

which we can take as the new prior on p before seeing D_2. Assuming independence of the priors, prior to seeing D_2 the complete prior is given by

$$\begin{aligned}\pi(p,q) &= P(p|D_1)\pi(q) \\ &= \frac{\Gamma(a+b+100)}{\Gamma(a+15)\Gamma(b+85)}\, p^{a+14}(1-p)^{b+84}\, \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)}\, q^{c-1}(1-q)^{d-1}. \end{aligned} \tag{3.11}$$

The model specification is completed by defining the likelihood $P(D_2|p,q)$, which accounts for recovery mechanism overlap. Under the assumption of independence between aspirin effects and other factors, we derive a four-category multinomial distribution with probabilities: $p(1-q)$ (other factors only), $q(1-p)$ (aspirin only), pq (both), and $(1-p)(1-q)$ (none). Consequently, the probability of observing count data $N = n_1+n_2+n_3+n_4$ is expressed as:

$$P(n_1,n_2,n_3,n_4|p,q) = \frac{N!}{n_1!n_2!n_3!n_4!}\,(p(1-q))^{n_1}(q(1-p))^{n_2}(pq)^{n_3}((1-p)(1-q))^{n_4}.$$

In Table3, $N = 100$ and n_1, n_2, n_3 are lumped together such that $n_1+n_2+n_3 = 37$. Thus the likelihood $P(D_2|p,q)$ is given by

$$\sum_{n_1=0}^{37}\sum_{n_2=0}^{37-n_1}\frac{100!}{n_1!n_2!n_3!63!}(p(1-q))^{n_1}(q(1-p))^{n_2}(pq)^{37-n_1-n_2}((1-p)(1-q))^{63},$$

or

$$\sum_{n_1=0}^{37}\sum_{n_2=0}^{37-n_1}\binom{100}{n_1}[p(1-q)]^{n_1}\binom{100-n_1}{n_2}[q(1-p)]^{n_2}\binom{100-n_1-n_2}{37-n_1-n_2}[pq]^{37-n_1-n_2}[(1-p)(1-q)]^{63},$$

where n_1 denotes recoveries attributable exclusively to other factors (including natural recovery and placebo effects); n_2 denotes recoveries caused solely by aspirin intervention; n_3 represents recoveries resulting from the synergistic effect of aspirin and other factors; and n_4 corresponds to non-recoveries where neither mechanism was effective.

Under fixed parameters p and q, the the null hypothesis can be expressed as q when including confounding effects, or $q-pq$ for the pure aspirin effect. The the alternative hypothesis is correspondingly $q/(p+q-pq)$ (with confounding) or $(q-pq)/(p+q-pq)$ (pure effect). This probabilistic framework effectively resolves semantic ambiguities, with final inferences based on the joint posterior distribution of (p,q) through MCMC methods.

Here we assume assuming a uniform prior on p and q ($a=b=c=d=1$), and the posterior mean, standard deviation, and 95% confidence interval (CI) of each parameter are shown in Table 4. We also include the estimation of the alternative hypothesis, $q/(q+p-pq)$. The results indicate that aspirin has a significant effect on headache treatment, with over

Table 4: Posterior summaries for the aspirin example.

	95% CI	**Posterior mean**	**Posterior standard deviation**
p	(0.09, 0.23)	0.16	0.04
q	(0.18, 0.30)	0.25	0.03
$\frac{q}{q+p-pq}$	(0.49, 0.82)	0.68	0.09

2/3 of symptom improvement attributable to its pharmacological action rather than chance or other confounding factors. For comparison, Pearson's χ^2 test of independence yields a p-value of 0.0007, indicating that we can reject the null hypothesis at the 0.001 significance level and conclude that aspirin has a significant therapeutic effect on headache treatment. Furthermore, the Bayes factor is 86.99 (substantially exceeding the threshold of 30), which is generally considered as "very strong evidence for H_1" [23].

5 Discussion

This paper reviews Bayesian causal inference within the framework of potential outcomes and directly quantifies the strength of causal relationships using Bayesian factors. In the Bayesian framework for causality analysis, the fundamental problem of Bayesian causality analysis is to compute the posterior of causal hypotheses, given the corresponding data and background information. The Bayesian approach offers distinct advantages for causal inference [14]. First, it enables imputation of all missing potential outcomes, creating a unified framework for estimating any causal quantity-including complex targets like conditional or individual treatment effects, as well as partially identifiable estimands such as principal strata effects. Incontrast, frequentist methods often require case-specific solutions and frequently depend on uninformative bounds or asymptotic approximations. Second, Bayesian inference automatically quantifies uncertainty for any estimand, facilitating integration with decision theory in contexts such as personalized medicine. Third, it naturally incorporates prior knowledge, which is valuable in settings like spatially correlated treatments or outcomes. Fourth, Bayesian modeling provides flexible tools for complex data structures-such as spatial, temporal, functional, or interference settings where SUTVA fails-where frequentist alternatives are limited.

We do not claim that the proposed Bayesian approach is superior to others, nor do we revisit the Bayesian-frequentist debate. Rather, we present it as a flexible and unified

framework that complements existing methods for causal analysis. As a general view, we believe whether to choose a Bayesian approach should be dictated by its practical utility in a specific context rather than an unconditional commitment to the Bayesian doctrine. For causal inference and perhaps everything in statistics, being Bayesian should be a tool, not a goal.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Hernán MA, Robins JM. *Causal Inference*. Boca Raton: Chapman and Hall/CRC; 2018.

[2] Jain KK. (2006) Textbook of Personalized Medicine. *New York, NY: Springer*. https://doi.org/10.1007/978-1-4419-0769-1

[3] Rubin DB. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies[J]. *Journal of educational Psychology*, **66(5)**: 688. https://doi.org/10.1037/h0037350

[4] Little RJA, Rubin DB. Statistical analysis with missing data. *John Wiley & Sons*, 2019.

[5] Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. *Cambridge university press*, 2015. https://doi.org/10.1017/CBO9781139025751

[6] Gelman A, Carlin JB, Stern HS, et al. (2013) Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

[7] Hill JL. (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*. **20**: 217-240. http://doi:10.1198/jcgs.2010.08162

[8] Zigler CM, Dominici F. (2014) Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. *Journal of the American Statistical Association*. **109**: 95-107.http://doi:10.1080/01621459.2013.869498

[9] Hahn PR, Carvalho CM, Puelz F, He J. (2018) Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*. **13**: 163-182. http://doi:10.1214/16-BA1044

[10] Hahn PR, Murray JS, Carvalho CM. (2020) Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*. **15**: 965-1056. http://doi:10.1214/19-BA1195

[11] Marden, JI. (2000) Hypothesis testing: from p values to Bayes factors. *Journal of the American Statistical Association*, **95(452)**, 1316–1320. https://doi.org/10.1080/01621459.2000.10474339

[12] Kass, R. E., & Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, 90(430), 773-795. https://doi.org/10.1080/01621459.1995.10476572

[13] Rubin DB. (1980) Comment on 'Randomization analysis of experimental data: the Fisher randomization test' by D. Basu. *Journal of the American Statistical Association*, **75**:591-593.

[14] Li F, Ding P, Mealli F. (2023) Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, **381(2247)**: 20220153.

[15] Holland PW. (1986) Statistics and causal inference. *Journal of the American Statistical Association*, **81**:945-960. http://doi:10.1080/01621459.1986.10478354

[16] Rubin DB. (1978) Bayesian inference for causal effects: the role of randomization. *The Annals of statistics*, **6**:34-58. http://doi:10.1214/aos/1176344064

[17] Rosenbaum PR, Rubin DB. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**: 41-55. http://doi:10.1093/biomet/70.1.41

[18] Liu Z, Wang X. (2023) Model-based adaptive randomization procedures for heteroscedasticity of treatment responses. *Statistical Methods in Medical Research*, **32(7)**: 1361-1376. http://doi: 10.1177/09622802231173050

[19] Ding P, Li F. (2018) Causal inference: a missing data perspective. *Statistical Science*, **33(2)**: 214-237. http://doi:10.1214/18-STS645

[20] Rubin DB. (1978) Bayesian inference for causal effects: the role of randomization. *The Annals of statistics*, **6**: 34-58. http://doi:10.1214/aos/1176344064

[21] Rubin DB. (1976) Inference and missing data. *Biometrika*, **63**: 581-592. http://doi:10.1093/biomet/63.3.581

[22] Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995) Markov Chain Monte Carlo in Practice. Boca Raton, FL: CRC Press. https://doi.org/10.1201/b14835

[23] Baldi, P., & Shahbaba, B. (2019) Bayesian Causality. *The American Statistician*, **74(3)**: 249-257. https://doi.org/10.1080/00031305.2019.1647876

[24] Rubin, DB. (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York. http://dx.doi.org/10.1002/9780470316696

[25] Rubin, Donald B. (2005) Bayesian inference for causal effects. *Handbook of statistics*, **25**: 1-16. https://doi.org/10.1016/S0169-7161(05)25001-0

[26] de Finetti, B. (1963) Foresight: Its logical laws, its subjective sources. In: Kyburg, H.E., Smokler, H.E. (Eds.), Studies in Subjective Probability. Wiley, New York.

[27] Pierre Baldi, Babak Shahbaba (2020) Bayesian Causality. *The American Statistician*, **74:3**, 249-257. https://doi.org/10.1080/00031305.2019.1647876

[28] Lee, MD, & Wagenmakers, EJ. (2014) Bayesian cognitive modeling: A practical course. Cambridge university press. https://doi.org/10.1017/CBO9781139087759

[29] Efron, B., & Hastie, T. (2021) Computer age statistical inference, student edition: algorithms, evidence, and data science (Vol. 6). Cambridge University Press. http://doi:10.1017/9781108914062

[30] Nuzzo, R. (2014) Scientific method: Statistical errors. *Nature*, 506, 150-152. https://doi.org/10.1038/506150a

[31] Wasserstein, R. L., & Lazar, N. A. (2016) The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

[32] Raftery, AE. (1995) Bayesian model selection in social research. *Sociological methodology*, 111-163. http://doi:10.2307/271063

[33] Box, G. E. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Ser.A*, 143(4), 383-404. https://doi.org/10.2307/2982063

[34] Dawid AP, Faigman DL and Fienberg SE. (2014) Fitting science into legal contexts: assessing effects of causes or causes of effects?. *Sociological Methods & Research*, **43(3)**: 359-390. http://doi:10.1177/0049124113515188