Published in final edited form as:
Am Stat. 2020 ; 74(3): 249–257. doi:10.1080/00031305.2019.1647876.

Bayesian Causality

Pierre Baldi,
Department of Computer Science, University of California, Irvine

Babak Shahbaba
Department of Statistics, University of California, Irvine

Abstract

Although no universally accepted definition of causality exists, in practice one is often faced with the question of statistically assessing causal relationships in different settings. We present a uniform general approach to causality problems derived from the axiomatic foundations of the Bayesian statistical framework. In this approach, causality statements are viewed as hypotheses, or models, about the world and the fundamental object to be computed is the posterior distribution of the causal hypotheses, given the data and the background knowledge. Computation of the posterior, illustrated here in simple examples, may involve complex probabilistic modeling but this is no different than in any other Bayesian modeling situation. The main advantage of the approach is its connection to the axiomatic foundations of the Bayesian framework, and the general uniformity with which it can be applied to a variety of causality settings, ranging from specific to general cases, or from causes of effects to effects of causes.

Keywords

Bayesian statistics; foundations; causal inference; hypothesis testing

1 Introduction

Causality is a fundamental concept in science, technology, and our general understanding of the world. However its precise definition is problematic: no universally accepted definition of causality exists even to the point that some have advocated avoiding precise definitions of the concept [“The law of causality...like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm” (B. Russell); “Causality is not to be strongly defined; thus theory cannot be created.” (L. Zadeh)].

While it is clear that correlation is not causation, the complexity of defining “A causes B” is compounded by several factors. In complex, but typical, situations one is confronted with multiple causes and multiple effects interacting in complex ways through mechanisms that are often only partially understood, in both individual and general cases (e.g. relationships between drugs, age, and cancer in a specific individual or in the population). Furthermore, causal inference may proceed in the forward direction associated with the effects of causes (what might happen if we do A) or in the reverse direction associated with the causes of effects (what causes B).

In spite of these difficulties, in practice, especially in legal or scientific settings, one is often faced with the problem of rationally evaluating possible causal relationships and teasing out whether something classified as being statistically significant is associated with a causal relationship or not. With the ongoing data deluge, this has become an area of intense research and several statistical frameworks have been developed. Among the most widely used or debated ones are the counterfactual approach usually associated with Rubin (Rubin, 1974, 1977, 1990), and Pearl's (Directed Acyclic Graph) approach (Pearl, 2000, 2009) (see also Illari et al. (2011); Peters et al. (2011), among many others).

While some statistical frameworks may depend on a precise definition of causality, or on the characteristics of the statistical problem being considered–such as causality in a specific case as opposed to causality in the more general setting, or the important distinction between reverse causal inference (causes of effects) versus forward causal inference (effects of causes) (Dawid et al., 2014; Pearl, 2015), it is conceivable that some frameworks may be more general and apply in some sense uniformly to different situations and different definitions of causality.

The Bayesian statistical framework, in particular, provides a general framework for statistical questions, and thus it is natural to ask what a Bayesian framework for the study of causality should look like? Thus, in short, the goal here is to go back to the foundations of the Bayesian approach and try to derive a complementary, general, framework for causality analysis derived from Bayesian first principles.

2 The Bayesian Statistical Framework and its Axioms

To begin with, it is useful to review the axiomatic foundations of the Bayesian statistical framework (Cox, 1964; Savage, 1972; Jaynes, 2003; Bernardo and Smith, 2001). Broadly speaking there are at least two related kinds of axiomatic frameworks. The first kind is based on "monetary" notions, such as utility and betting. As we do not think that monetary considerations are essential to science and statistics, we choose the second framework, based on the notion of "degrees of belief".

The starting point for this framework is to consider an observer who has some background knowledge $\mathscr{B}$ and observes some data D. The observer is capable of generating hypotheses or models, from a certain class of hypotheses or models. The process by which these models are generated, or by which the class of models is changed, are outside of the scope of the framework.

Bayesian analysis is concerned with the assessment of the quality of the hypotheses given the data and the background model. Given an hypothesis H, the fundamental object of Bayesian statistics is the "observer's degree of belief" $\pi(H \mid D, \mathscr{B})$ in H, given D and $\mathscr{B}$. If the background knowledge $\mathscr{B}$ changes, the degree of belief may change, one of the main reasons this approach is sometimes called "subjective" (incidentally, a major marketing error of the Bayesian camp and a main argument for the detractors of this approach). Although "subjective", the approach aims to be rational in the sense that the fundamental quantity

$\pi(H \mid D, \mathscr{B})$ should satisfy a set of reasonable axioms. Usually three axioms are imposed on $\pi(H \mid D, \mathscr{B})$.

First $\boldsymbol{\pi}$ should be transitive, in the sense that given three hypotheses H_1, H_2, and H_3 if $\pi(H_1 \mid D, \mathscr{B}) \quad \pi(H_2 \mid D, \mathscr{B})$ and $\pi(H_2 \mid D, \mathscr{B}) \quad \pi(H_3 \mid D, \mathscr{B})$, then:

$$\pi(H_1 \mid D, \mathscr{B}) \quad \pi(H_3 \mid D, \mathscr{B}) \tag{1}$$

Here $X \; Y$ is meant to represent that hypothesis Y is preferable to hypothesis X. The transitivity hypothesis essentially allows mapping degrees of beliefs to real numbers and replacing with ≤.

The second axiom states that there exists a function $f(x)$ establishing a systematic relationship between the degree of belief in an hypothesis H and the degree of belief in its negation $\neg H$. Intuitively, the greater the belief in H, the smaller the belief in $\neg H$ should be. In other words:

$$\pi(H \mid D, \mathscr{B}) = f(\pi(\neg H \mid D, \mathscr{B}))\,. \tag{2}$$

Finally, the third axiom, states that given two hypotheses H_1 and H_2, there exists a function $F(x, y)$ establishing a systematic relationship such that:

$$\pi(H_1, H_2 \mid D, \mathscr{B}) = F[\pi(H_1 \mid D, \mathscr{B}), \pi(H_2 \mid H_1, D, \mathscr{B})]\,. \tag{3}$$

The fundamental theorem that results from these axioms is that degrees of beliefs can be represented by real numbers and that if these numbers are re-scaled to the [0,1] interval, then $\pi(H \mid D, \mathscr{B})$ must obey all the rules of probability. In particular, $f(x) = 1-x$ and $F(x, y) = xy$. As a result, in what follows we will use the notation $P(H \mid D, \mathscr{B})$ and call it the probability of H given D and $\mathscr{B}$.

In particular, when re-scaled to [0,1], degrees of belief must satisfy Bayes theorem:

$$P(H \mid D, \mathscr{B}) = \frac{P(D \mid H, \mathscr{B}) P(H \mid \mathscr{B})}{P(D \mid \mathscr{B})}\,. \tag{4}$$

which is the fundamental tool for inversion, where $P(H \mid \mathscr{B})$ is called the prior of the hypothesis, $P(D \mid H, \mathscr{B})$ the likelihood of the data, and $P(D \mid \mathscr{B})$ the evidence.

Thus fundamentally, in this framework, probabilities are viewed very broadly as degrees of beliefs assigned to statements (or hypothesis or models) about the world, rather than the special case of frequencies associated with repeatable events.

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 $P(H \mid \mathscr{B})$; and (2) the actual computation of the posterior $P(H \mid D, \mathscr{B})$, and any related expectations, which may not always be solvable analytically and may require Monte Carlo approaches (Gelman et al., 1995; Gilks et al., 1995).

3 Causal Relationships as Hypotheses about the World

Given this foundational framework, how then should causality be viewed from a Bayesian standpoint? It is best to consider a simple example.

Suppose that Mr. Johnson is driving his car at 80 miles per hour and frontally collides with another stationary car. After the collision, we observe that Mr. Johnson is dead and wish to ask the simple causality question: was the death of Mr. Johnson caused by the high-speed collision? From our past experiences and background knowledge, we immediately get a feeling that this is likely to be the case. However, upon reflection, we also realize that one cannot be absolutely certain of this causal relationship. What if Mr. Johnson suffered a deadly heart attack two seconds before the collision? With additional data, such as the results of an autopsy, these two possibilities could perhaps be disentangled, but one may, or may not, have access to this additional data.

In short, it should be clear that the statement “the death of Mr. Johnson was caused by the high-speed collision” should itself be *regarded as a hypothesis about the world*. Our degree of belief in this hypothesis may vary depending on the data available to support it and our background knowledge. Thus, from a Bayesian standpoint, causality relationships ought to be treated like any other hypothesis or model. And therefore one must conclude that the fundamental object of Bayesian causality analysis ought to be the computation of the probability of the causality hypothesis given the available data and background knowledge.

Although this conclusion comes straight out of the foundations of the Bayesian statistical framework, it is perhaps surprising that, to the best of our knowledge, it has not been clearly articulated as such. In fact, the first reaction of several statisticians we consulted with is to say something like “I would not know where to start in order to compute such probability”.

Note that this general framework applies uniformly both to specific (e.g. Mr. Johnson) or general (e.g. drivers) situations. It applies also uniformly to causes of effects and effects of causes analyses, as these primarily correspond to a change in the data. In one case, the effects are observed, in the other case the causes are observed. This is an important point, as there has been a general tendency to fragment causality analysis into a growing number of “sub-areas” by categorizing different kinds of data (e.g. observational, experimental) and different kinds of analyses (e.g. effect of causes, causes of effects) and studying each combination separately, while in reality there is a continuum of situations (see also Rubin et al. (2008)).

One concern one may raise is that the framework described applies only to binary propositions and may not be able to handle continuous hypothesis or causal effect sizes. However this is not the case. First, any continuous hypothesis can be reduced to a sequence of binary propositions. For instance, if we are interested in a hypothesis about the size of an effect, we can always reduce it to a sequence of binary hypothesis of the form “the size is greater than a_i” for some sequence of numbers a_i. Second, it is natural in a Bayesian framework to consider probabilities as parameters and to compute expectations by integrating over their posterior distributions, and thus producing continuous values. For instance, in the case of a drug treatment, we may have a parameter p that represents the probability of a drug curing an individual randomly drawn from a certain population. We may have a prior $P(p)$ and derive a posterior distribution $P(p / D)$ from some data D. By integrating over $P(p / D)$, one can easily compute effect sizes in the form of the expected mean and standard deviation of the number of people to be cured by the drug in a similar population of size N. Specific examples are given below.

4 Related Work

This is of course not the first time that Bayesian ideas are applied to issues of causality. Following the work of Rosenbaum and Rubin (1983) on propensity score adjustment (based on a model for treatment assignment mechanism) to control for the effect of unknown and non-ignorable confounding variables, there have been many attempts to develop a comparable Bayesian method (Hoshino, 2008; McCandless et al., 2009; An, 2010; Kaplan and Chen, 2012; Matthew Zigler and Dominici, 2014; Chen and Kaplan, 2015; Graham et al., 2015). Saarela et al. (2016) propose an alternative approach based on posterior predictive inference, by using the inverse probability of treatment to decouple the outcome regression model from the treatment assignment model. This alternative approach has been previously discussed by several authors in the context of change of probability measures (Rysland, 2011; Chakraborty and E.M. Moodie, 2013). Such methods use the concept of potential outcomes to define causal effects explicitly, so that Bayesian methods can be used for the joint distribution of the potential outcomes, and probability statements can be provided on the effect size using the resulting posterior distribution.

Rubin (1978) argues that, while Bayesian inference is complicated for nonignorable treatment assignment, principled Bayesian methods for causal inference could still be used when the treatment assignment is either unconfounded or confounded but ignorable. In fact, the Bayesian perspective could be quite useful for handling complex causal inference problems. For example, Imbens and Rubin (1997) propose a novel framework for Bayesian causal inference in randomized experiments with noncompliance. They argue that while Bayesian methods never account for design based standard errors, the compliance behavior of subjects should be taken into account. Schwartz et al. (2011) address the issue of intermediate variables in causal inference by using a Dirichlet process mixture model. Daniels et al. (2012) also propose a Bayesian approach for the causal effect of mediation. Dawid et al. (2016) use a personalist Bayesian perspective to discuss the distinction between the “ effect of causes” (EoC) and the “causes of effects” (CoE), especially when investigating a case in a Court of Law. They argue that while statisticians are mainly concerned with EoC, it is unclear how statistical methods could be used for CoE, which is typically what a Court of Law needs in order to assign responsibility for a given undesirable outcome. Although they admit that this problem might not have a well-determined answer, they show that it is possible to find bounds for the “probability of causation” using a Bayesian approach.

Here, we advocate using the foundations of Bayesian statistics as a general and unifying framework for causality. This framework, as described in Section 3, focuses on computing posterior probabilities, but it does not prescribe how to use these probabilities, for instance in hypothesis testing. In Bayesian statistics, hypothesis testing is commonly discussed within the framework of decision theory, where the goal becomes the minimization of the posterior risk with respect to a loss function. When using a convenient, but not necessarily optimal, 0–1 loss function, the minimization of the posterior risk reduces to choosing the hypothesis with the higher posterior odds. This is of course sensitive to the choice of the prior distribution, which is often assumed to be uniform leading to the notion of Bayes factor, originally developed by Jeffreys (1935, 1961). Thus the Bayes factor is simply the posterior odds for one hypothesis (e.g., the null hypothesis) when we decide not to express preference for either hypothesis *a priori*. In other words: Posterior Odds = Bayes Factor × Prior Odds:

$$\frac{P(H1|D)}{P(H_2|D)} = \frac{P(D|H1)}{P(D|H2)} \frac{P(H1)}{P(H2)}$$

As we can see, the Bayes factor $B_{12} = P(D|H1)/P(D|H2)$ has the form of a likelihood ratio (Kass and Raftery, 1995). Gûnel and Dickey (1974) discuss different Bayes factors, under different sampling models, for two-way contingency tables.

5 Bayesian Causality Calculations

To see how causality may be assessed in a Bayesian fashion more precisely, let us first proceed with the example above.

5.1 Example of Bayesian Causality Calculation (Car Collision)

We let D be the observation that Mr. Johnson is dead. We let A denote the fact that Mr. Johnson was involved in a high speed collision and $B = D$ the fact that Mr. Johnson is dead. We denote by $B \leftarrow A$ the hypothesis that the high speed collision was the cause of Mr. Johnson's death. Thus, within the Bayesian foundational framework, the fundamental task is to compute $P(B \leftarrow A | D, \mathscr{B})$.

As a side note, this is very different from computing $P(B \mid A)$ which is used in the area of probabilistic causation (Anderson and Vastag, 2004; Cardenas et al., 2017).

From Bayes theorem we immediately have:

$$P(B \leftarrow A|D, \mathscr{B}) \\ = \frac{P(D|B \leftarrow A, \mathscr{B})P(B \leftarrow A|\mathscr{B})}{P(D|B \leftarrow A, \mathscr{B})P(B \leftarrow A|\mathscr{B}) + P(D|\neg(B \leftarrow A), \mathscr{B})P(\neg(B \leftarrow A)|\mathscr{B})}. \tag{5}$$

There are rational ways for assigning a numerical value to each term in this equation. The numerical value may vary as a function of, for instance, the background knowledge but this is a good thing, rather than a bad thing. It forces one to examine exactly which assumptions go into the calculation. Furthermore, one can proceed with a robustness analysis to show, for instance, that the posterior does not vary too much under different assumptions, as shown below.

In this particular case, we first obviously have $P(D \mid B \leftarrow A, \mathscr{B}) = 1$. For the prior, $P(B \leftarrow A \mid \mathscr{B})$ we could look at general statistics on how often collisions at 80 miles an hour result in the death of the driver. Let us say that this occurs 95% of the time, then we could use the prior: $P(B \leftarrow A \mid \mathscr{B}) = 0.95$. Then obviously $P(\neg(B \leftarrow A) \mid \mathscr{B}) = 0.05$. Thus the only term left to estimate, and the most interesting one, is: $P(D \mid \neg(B \leftarrow A), \mathscr{B})$.

To estimate $P(D \mid \neg(B \leftarrow A), \mathscr{B})$ requires some analysis and modeling, but the same would be true in the other existing approaches to causality. Only for the purpose of simplifying the discussion, let us assume that the only other possible cause of death for Mr. Johnson is a heart attack, leaving out strokes, stray bullets, etc. Then basically we need to estimate the probability that Mr. Johnson suffered a heart attack in the few seconds, again to simplify let us say two seconds, before the collision. Again this probability could be estimated rationally from background statistical data on the population. Here, only for the sake of simplicity, we assume the model that at every second of human life there is an independent probability ~ 10^{-9} of suffering a heart attack. As a result, under these assumptions, we have: $P(D \mid \neg(B \leftarrow A), \mathscr{B}) \approx 2 \times 10^{-9}$. In short, putting all these elements together, leads us to the following degree of belief in the causal relationship:

$$P(B \leftarrow A \,|\, D, \mathscr{B}) = \frac{0.95}{0.95 + 0.05 \times 2 \times 10^{-9}} = \frac{0.95}{0.9500000001} \approx 0.999999999895 \quad (6)$$

As expected, one has a strong belief in the collision being the cause of Mr. Johnson's death both because high speed collisions have a high probability of resulting in death and because alternative causes of death have low probability. The Bayesian framework allows one to quantify and combine these two elements in a precise way. While in this case, the long decimal expansion may seem like an overkill, it must be noted that there are legal settings where very small probabilities matter, for instance in the case of establishing identity using genetic profiling (Kaye, 1993; Waits et al., 2001; Buckleton et al., 2016).

It is easy to see in this case that the belief in a causal relationship is robust in that most of the decimals would remain unchanged if the probability of dying in a high speed collision was $x = 0.9$ instead of 0.95 or the instantaneous probability of heart attack $y = 10^{-8}$ rather than 10^{-9}. In fact, for sensitivity analysis purposes, the derivative of the posterior with respect to these quantities can easily be computed and analyzed as we have:

$$P(B \leftarrow A \,|\, D, \mathscr{B}) = \frac{x}{x + (1 - x)2y} \quad (7)$$

Likewise, one can modify the prior and analyze the effects of such a change. For instance under an uninformative, maximum entropy, prior $P(B \leftarrow A \mid \mathscr{B}) = P(\neg(B \leftarrow A) \mid \mathscr{B}) = 0.5$. In this case,

$$P(B \leftarrow A \,|\, D, \mathscr{B}) = \frac{0.5}{0.5 + 0.5 \times 2 \times 10^{-9}} = \frac{0.5}{0.5000000001} \approx 0.9999999998 \tag{8}$$

again resulting only in a minor change over the value obtained in Equation 6.

The analysis above was done for a specific individual (Mr. Johnson) and in a causes of effects setting. It should be clear however that with the proper adjustments similar calculations can be done for generic cases, or for the effects of causes setting. Once A and B have been defined with sufficient precision, all these variations formally require:

1. Defining a prior distribution $P(B \leftarrow A | \mathscr{B})$ on the causal hypothesis, regardless of its form or definition; and
2. Computing likelihoods expressions of the form: $P(D | (B \leftarrow A), \mathscr{B})$ and $P(D | \neg(B \leftarrow A), \mathscr{B})$.

The important work lies in these two steps. In particular, these steps may require building a complex Bayesian hierarchical model. *But this is no different from Bayesian modeling in any other situation.*

5.2 Example of Experimental Study (Aspirin)

As another example, consider the 2 × 2 contingency table (Table 1) (Dawid et al., 2014) reporting the results of a fictitious double-blind randomized experiment involving 200 individuals, with 100 given aspirin tablets, and 100 chalk tablets (the control). Let us denote by $D1$ the data corresponding to the first row of the Table (control), and by $D2$ the data corresponding to the second row (experiment). Here one is interested in possible causal relationships between taking aspirin and recovery from headaches, possibly in the forms:

- Effects of Causes: I have a headache. I am wondering whether to take aspirin. Will taking aspirin cause my headache to disappear?
- Causes of Effects: I had a headache and took aspirin. My headache went away. Was that caused by the aspirin only?

The Effects of Cause problem can be written as estimating $P(B \leftarrow A | D, \mathscr{B})$ where A is “I take aspirin”, B is “my headache will disappear within 30 minutes ” and $\mathscr{B}$ is “I have a headache”. The Cause of Effects problem can be written as estimating $P([B \leftarrow A] \wedge \neg[B \leftarrow \text{Other}] | D, \mathscr{B})$ where A is “I took aspirin”, B is “my headache disappeared within 30 minutes”, $\mathscr{B}$ is “I had a headache”, and ¬[$B \leftarrow$ Other] is meant to capture the “only” portion of the statement, i.e. to rule out other causes, as described below.

It should first be obvious from the data that none of these questions has a simple yes/no answer and that these questions can only be addressed in probabilistic terms. There are two factors that complicate the analysis somewhat (Figure 1). The first factor is that $D1$ shows that 12 out of 100 people recovered from headache when given chalk. The observed recovery could be due to “natural causes” or to placebo effects. No data about recovery from headache in controls who have not taken a pill is available to try to disentangle these two effects. Thus to slightly simplify the analysis, we will lump natural causes and placebo effects into a single cause called “Other”. The second factor is that regardless of whether there is a placebo effect or not, there is potentially some overlap between recovery due to aspirin and recovery due to other causes (e.g. through different biochemical pathways) and no direct observation of this overlap is available to us.

To build a Bayesian probabilistic model of the data in Table 1, we assume an overall model characterized by two probabilities p and q, which are the parameters of this model: p is the conditional probability of recovery from a headache due to Other causes, and q is the conditional probability of recovery from a headache due to Aspirin.

It is natural to put Beta priors on the probabilities p and q, so that:

$$P(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \quad \text{and} \quad P(q) = \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} q^{c-1}(1-q)^{d-1} \tag{9}$$

with $a > 0$, $b > 0$, $c > 0$ and $d > 0$. The data $D1$ does not contain information about q but it immediately leads to a marginal Beta posterior distribution for p:

$$P(p \mid D1) = \frac{\Gamma(a+b+100)}{\Gamma(a+12)\Gamma(b+88)} p^{a+11}(1-p)^{b+87} \tag{10}$$

which we can take as the new prior on p before seeing $D2$. Assuming independence of the priors, prior to seeing $D2$ the complete prior is given by:

$$\begin{aligned} P(p,q) &= P(p \mid D1)P(q) \\ &= \frac{\Gamma(a+b+100)}{\Gamma(a+12)\Gamma(b+88)} p^{a+11}(1-p)^{b+87} \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} q^{c-1}(1-q)^{d-1} \end{aligned} \tag{11}$$

To complete the description of the model, we need to define the likelihood $P(D_2 \mid p, q)$, which requires dealing with the overlap issue. We assume that the action of Aspirin and the Other causes are independent. Thus for fixed p and q, we have a four-dimensional multinomial distribution with probabilities: $p\,(1-q)$, $q\,(1-p)$, pq and $(1-p\,)(1-q)$ (Figure 2). Thus, in general, in this model the probability of observing a corresponding set of counts $N = n_1 + n_2 + n_3 + n_4$ is given by:

$$\begin{aligned} &P(n_1, n_2, n_3, n_4 \mid p, q) \\ &= \frac{N!}{n_1!n_2!n_3!n_4!}(p(1-q))^{n_1}(q(1-p))^{n_2}(pq)^{n_3}((1-p)(1-q))^{n_4} \end{aligned} \tag{12}$$

In Table 1, $N = 100$ and n_1, n_2, and n_3 are lumped together such that: $n_1 + n_2 + n_3 = 30$. Thus the likelihood $P(D2 \mid p, q)$ is given by:

$$\sum_{n_1=0}^{30} \sum_{n_2=0}^{30-n_1} \frac{100!}{n_1!n_2!n_3!70!}(p(1-q))^{n_1}(q(1-p)^{n_2}(pq)^{30-n_1-n_2n_3}((1-p)(1-q))^{70} \tag{13}$$

or:

$$\sum_{n_1=0}^{30}\sum_{n_2=0}^{30-n_1}\binom{100}{n_1}[p(1-q)]^{n_1}\binom{100-n_1}{n_2}[q(1-p)]^{n_2}[pq]^{k_2}\binom{100-n_1-n_2}{30-n_1-n_2}[pq]^{30-n_1-n_2}[(1-p)(1-q)]^{70} \tag{14}$$

where n_1 is the number of people recovering from Other only, and n_2 is the number of people recovering from Aspirin only. Note that this could be regarded as a form of missing data imputation.

The model is now complete and, as in any other Bayesian modeling situation, one can compute the posterior distribution on p and q, their modal or mean a posteriori point estimates, as well as any other expectations or integrals with respect to this posterior distribution as a function of the parameters (a,b,c,d) of the prior distribution. For instance, for fixed p and q, the probability associated with the Effects of Causes is: q if we interpret it as including the possibility of a coexisting Other effect (or $q - pq$ if we interpret it as a pure aspirin effect). Likewise, the probability associated with the Causes of Effects question is obtained by noting that the total probability of recovery from headache is $p + q - pq$. Thus it is given by: $q / (p + q - pq)$ in same broad sense as above(or $(q - pq) / (p + q - pq)$ if we disallow any coexisting Other effect). It is worth noting how the probabilistic framework can precisely disambiguate the semantic ambiguity hidden behind common, but vague, language. Finally, inferences for these probabilities are obtained based on the joint posterior for (p, q); this is easily accomplished using standard MCMC methods.

More precisely, assuming a uniform prior on p and q $(a = b = c = d = 1)$, Figure 3 shows the joint (left) and marginal (right) posterior distributions on the parameters p and q. The marginal density curves (in the right panel of Figure 3) are based on approximate Beta distributions fitted to posterior samples. The posterior mean, standard deviation, and 95% credible interval (CI) for each parameter are provided in Table 2. We have also included the estimates for the Causes of Effects, $q / (q + p - pq)$. As we can see, we expect that on average 20% of people with headache feel better after taking aspirin. Also, if a person's headache disappears after taking aspirin, the estimated probability that this is in fact due to taking aspirin is 0.65. For comparison, the p-value using Pearson's χ^2 test of independence is 0.003. That is, we can reject the null hypothesis at 0.01 significance level and conclude that aspirin helps cure headache. (Within the frequentist framework, we assume that the study is designed as a randomized experiment.) We can also compute the Bayes factor against the null hypothesis (independence). The Bayes factor is 19.34 which is generally considered as "strong evidence".

Figure 4 shows how our estimates of p and q change if we use other priors instead of Uniform(0, 1) for p and q. In this case, we allow the prior means for p and q, i.e., $a / (a + b)$ and $c / c + d$ respectively, to change from 0.1 to 0.9. To this end, we set a and c to 0.1, 0.2, …, 0.9, and assume $b = 1 - a$ and $d = 1 - c$. As we can see, our estimates are not very sensitive to the choice of priors in this case.

5.3 Example of Observational Study (Birthweight)

We now analyze a real study based on examining the relationship between low birthweight (defined as birthweight less than 2.5 kg) and maternal smoking during pregnancy. The data include a sample of 189 cases collected at Baystate Medical Center, Springfield, MA during 1986 (Hosmer and Lemeshow, 1989). Table 3 provides the summary data, and Figure 5 (left panel) shows the marginal posterior distributions of p and q. Here, p is the conditional probability of low birthweight due to other causes, and q is the conditional probability of low birthweight caused by smoking. The corresponding 95% CI, posterior means, and posterior standard deviations are presented in Table 4. As we can see, the estimated probability that smoking would cause low birthweight is 0.21. For a smoking mother with a low birthweight baby, there is a 0.48 probability that this has been caused by smoking. For this example, the Bayes factor against the null hypothesis is 1.94, which is considered as "weak evidence" (not worth more than a bare mention according to Jeffreys, 1961), and the p-value based on Pearson's χ^2 test is 0.04, which is generally considered as marginally significant. Note that since this is not a randomized experiment, within the frequentist framework we cannot conclude that the relationship is causal even though we can reject the null hypothesis (no association) at 0.05 significance level. However, we can use the propensity score adjustment method. To this end, we estimated the propensity score using a logistic regression model with maternal age and race as covariates and maternal smoking status as the outcome variable. We then stratified the subjects into five groups based on their estimated propensity scores and used a Cochran–Mantel–Haenszel χ^2 test of the null hypothesis that smoking and low birthweight are conditionally independent in each stratum. Using this approach, the p-value reduces to 0.005. However, whether we can use this result to conclude that smoking during pregnancy can cause low birthweight remains a controversial topic (Chapter 11, Pearl (2009)).

In a more recent study conducted during 1998–2000, (Wang et al., 2002) examined 741 mothers, who delivered singleton live births at Boston Medical Center. Among these women, 174 were smokers and 567 were non-smokers. Table 5 shows the frequencies of babies with normal birthweights and low birthweights for each group. The marginal posterior probability distributions of p and q are shown in Figure 5 (right), and the corresponding summaries (95% CI, mean, and standard deviation) are presented in Table 6. Here, we used the approximate Beta distributions fitted to the marginal posterior distributions using the data from Baystate Medical Center as the priors for p and q. More specifically, we assumed $p \sim \text{Beta}(30, 90)$ and $q \sim \text{Beta}(4, 17)$. Note that compared to the previous example, the estimated probabilities are relatively smaller and have narrower credible intervals. The Bayes factor in this case is 8.84 indicating "substantial evidence" against the null hypothesis (independence). Using Pearson's χ^2 test, the p-value is 0.003. While there is stronger evidence to reject the null hypothesis compared to the previous example, within the frequentist framework we cannot still conclude that smoking causes low birthweight since the study is not designed as a randomized experiment.

6 Discussion

We have presented what we think ought to be the Bayesian framework for causality analysis. In this framework, causality statements are viewed as hypotheses or models about the world, and thus the fundamental problem of Bayesian causality analysis is to compute the posterior of causal hypotheses, given the corresponding data and background information. This framework comes straight out of the axiomatic foundation of Bayesian statistics. In addition to its foundational consistency, one of its advantages is the uniformity with which it treats different causal inference situations, including causal inference in specific or more general cases, as well as forward (effects of causes) and reverse (causes of effects) causal inference.

Treating causality relationships in terms of probabilities that can evolve in time as new data are gathered seems natural to us, and consonant with scientific methodology. Consider for instance the question of whether humans are the main cause of global warming or not. The corresponding probability may have evolved from close to 0 in 1940, to 0.5 in the 1980s, to close to 1 in the 2020s.

It is not our goal to claim that the proposed Bayesian approach is better than any other one, or to try to revisit the Bayesian versus frequentist dispute, although we note that to the best of our knowledge there is no uniform, universally accepted, treatment of causality problems within a frequentist framework. Rather, we view the proposed Bayesian approach as a flexible framework that provides a complementary alternative to other statistical approaches for the analysis of causality relationships in a unified manner.

Within the Bayesian framework, we believe that causality is only one example of many other areas of statistical inference and data science where historical developmental accidents have led to unnecessary fragmentation, providing opportunities for creating more unified views. An example of an area where unification has occurred to some extent is the area of regularization, where it is well recognized that the addition of a regularizing term to an objective function is often equivalent to the choice of a corresponding prior. An example of area where the unification has not occurred in a systematic way is the area of variational methods (Murphy, 2012).

The typical description of variational methods begins with a target probability distribution P (possibly a Bayesian posterior distribution) of interest. The distribution P is considered intractable and thus one wishes to approximate P with some distribution $\mathcal{Q}$ taken from a simpler (e.g. parameterized or factored) family of distributions Q. The variational approach prescription consists in choosing the distribution Q^* that minimizes the relative entropy between Q and P. Using a discrete formalism ($P = (P_i)$ and $Q = (Q_i)$):

$$Q^* = \min_{Q \subset \mathcal{Q}} \sum_i Q_i \log \frac{Q_i}{P_i} = \min\left[-H(Q) - \sum_i Q_i \log P_i\right] \quad (15)$$

where $H(Q)$ is the entropy of Q. The typical description then proceeds with proving some of the nice properties of this approach, how it can be applied to an unnormalized distribution P, and so forth.

However, from a Bayesian standpoint, this description may seem slightly ad hoc and unprincipled since it involves an approximation to a degree of belief, which should itself be Bayesian (incidentally approximation problems in general can often be cast in Bayesian terms). In most variational cases, it is easy to correct this situation by noting that Equation 15 is equivalent to:

$$Q^* = \max_{Q \subset \mathcal{Q}} \prod_i P_i^{Q_i} \times \frac{e^{-H(Q)}}{Z} \tag{16}$$

where Z is the normalizing constant of the entropic prior. In other words, variational methods can viewed in a Bayesian perspective as maximum a posteriori maximization (MAP) methods under a likelihood proportional to $\prod_i P_i^{Q_i}$, with an entropic prior over the approximating family of distributions $\mathcal{Q}$. As a side note, this immediately suggests new ideas such as contrasting the MAP approach with a mean posterior (MP) approach. In short, we believe that a more systematic approach of Bayesian ideas may be one viable approach, both for the purposes of theoretical developments and pedagogy, towards a more unified treatment of currently fragmented areas of statistical inference and data science.

Acknowledgement

We would like to thank the anonymous reviewers for helping us improve the manuscript. PB would like to thank Dr. Ingeborg Waernbaum and Dr. Chris Cole for useful discussions.

Work in part supported by grants NIH GM123558 and DARPA D17AP00002 to PB.

References

An W (2010). Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. Sociological Methodology, 40(1):151–189.

Anderson RD and Vastag G (2004). Causal modeling alternatives in operations research: Overview and application. European Journal of Operational Research, 156(1):92–109.

Bernardo JM and Smith AF (2001). Bayesian theory. IOP Publishing.

Buckleton JS, Bright J-A, and Taylor D (2016). Forensic DNA evidence interpretation. CRC press.

Cardenas IC, Voordijk H, and Dewulf G (2017). Beyond theory: Towards a probabilistic causation model to support project governance in infrastructure projects. International Journal of Project Management, 35(3):432–450.

Chakraborty B and E.M. Moodie E (2013). Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine. Springer, New York.

Chen J and Kaplan D (2015). Covariate balance in bayesian propensity score approaches for observational studies. Journal of Research on Educational Effectiveness, 8(2):280–302.

Cox R (1964). Probability, frequency and reasonable expectation. American Journal of Physics, 14:1–13.

Daniels MJ, Roy JA, Kim C, Hogan JW, and Perri MG (2012). Bayesian inference for the causal effect of mediation. Biometrics, 68(4):1028–1036. [PubMed: 23005030]

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.

Dawid AP, Musio M, and Fienberg SE (2016). From statistical evidence to evidence of causality. Bayesian Analysis, 11(3):725–752.

Gelman A, Carlin JB, Stern HS, and Rubin DB (1995). Bayesian Data Analysis. Chapman and Hall, London.

Gilks WR, Richardson S, and Spiegelhalter D (1995). Markov chain Monte Carlo in practice. CRC press.

Graham D, McCoy JE, and Stephens DA (2015). Approximate bayesian inference for doubly robust estimation. Bayesian Analysis, 11.

Gûnel E and Dickey J (1974). Bayes factors for independence in contingency tables. Biometrika, 61(3):545–557.

Hoshino T (2008). A bayesian propensity score adjustment for latent variable modeling and mcmc algorithm. Computational Statistics & Data Analysis, 52(3):1413–1429.

Hosmer DW and Lemeshow S (1989). Applied logistic regression. John Wiley and Sons.

Illari PM, Russo F, and Williamson J (2011). Causality in the Sciences. Oxford University Press.

Imbens GW and Rubin DB (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. The Annals of Statistics, 25(1):305–327.

Jaynes ET (2003). Probability Theory. The Logic of Science. Cambridge University Press.

Jeffreys H (1935). Some Tests of Significance, Treated by the Theory of Probability. Mathematical Proceedings of the Cambridge Philosophical Society, 31(02):203–222.

Jeffreys H (1961). Theory of Probability. Oxford.

Kaplan D and Chen J (2012). A two-step bayesian approach for propensity score analysis: Simulations and case study. Psychometrika, 77.

Kass RE and Raftery AE (1995). Bayes factors. Journal of the American Statistical Association, 90:773–795.

Kaye DH (1993). Dna evidence: Probability, population genetics, and the courts. Harv. JL & Tech., 7:101.

Matthew Zigler C and 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. [PubMed: 24696528]

McCandless LC, Gustafson P, and Austin PC (2009). Bayesian propensity score analysis for observational data. Statistics in Medicine, 28(1):94–112. [PubMed: 19012268]

Murphy KP (2012). Machine learning: a probabilistic perspective (adaptive computation and machine learning series). MIT Press.

Pearl J (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press Cambridge, UK:.

Pearl J (2009). Causality. Cambridge university press.

Pearl J (2015). Causes of effects and effects of causes. Sociological Methods & Research, 44(1):149–164.

Peters J, Janzing D, and Scholkopf B (2011). Causal inference on discrete data using additive noise models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(12):2436–2450. [PubMed: 21464504]

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70:41–55.

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688.

Rubin DB (1977). Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics, 2(1):1–26.

Rubin DB (1978). Bayesian inference for causal effects: The role of randomization. The Annals of Statistics, 6(1):34–58.

Rubin DB (1990). Formal mode of statistical inference for causal effects. Journal of Statistical Planning and Inference, 25(3):279–292.

Rubin DB et al. (2008). For objective causal inference, design trumps analysis. The Annals of Applied Statistics, 2(3):808–840.

Rysland K (2011). A martingale approach to continuous-time marginal structural models. Bernoulli, 17(3):895–915.

Saarela O, Belzile LR, and Stephens DA (2016). A bayesian view of doubly robust causal inference. Biometrika, 103(3):667–681.

Savage LJ (1972). The foundations of statistics. Dover, New York (First Edition in 1954).

Schwartz SL, Li F, and Mealli F (2011). A bayesian semiparametric approach to intermediate variables in causal inference. Journal of the American Statistical Association, 106(496):1331–1344.

Waits LP, Luikart G, and Taberlet P (2001). Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. Molecular Ecology, 10(1):249–256. [PubMed: 11251803]

Wang X, Zuckerman B, Pearson C, and et al. (2002). Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. JAMA, 287(2):195–202. [PubMed: 11779261]

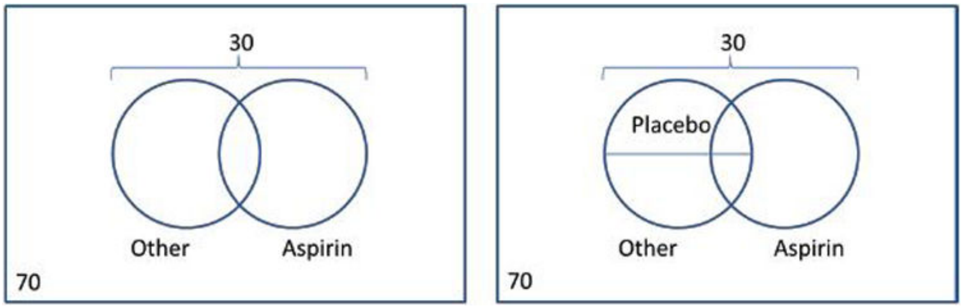

Fig. 1.

Venn diagrams and modeling assumptions associated with the data set $D2$.

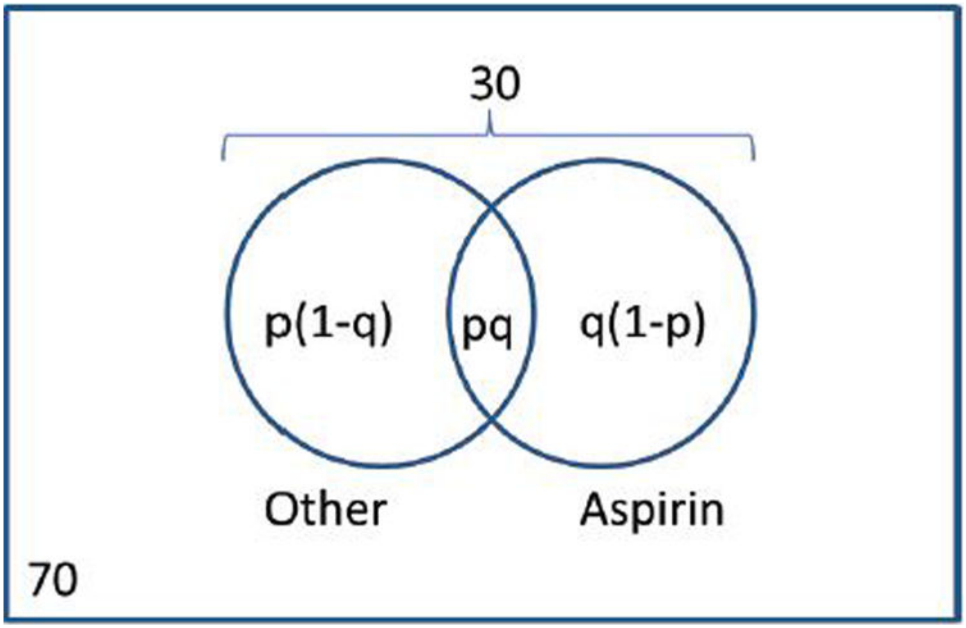

Fig. 2.

Venn diagrams and modeling assumptions associated with the data set *D*2. We assume that, for fixed *p* and *q*, Other and Aspirin act independently.

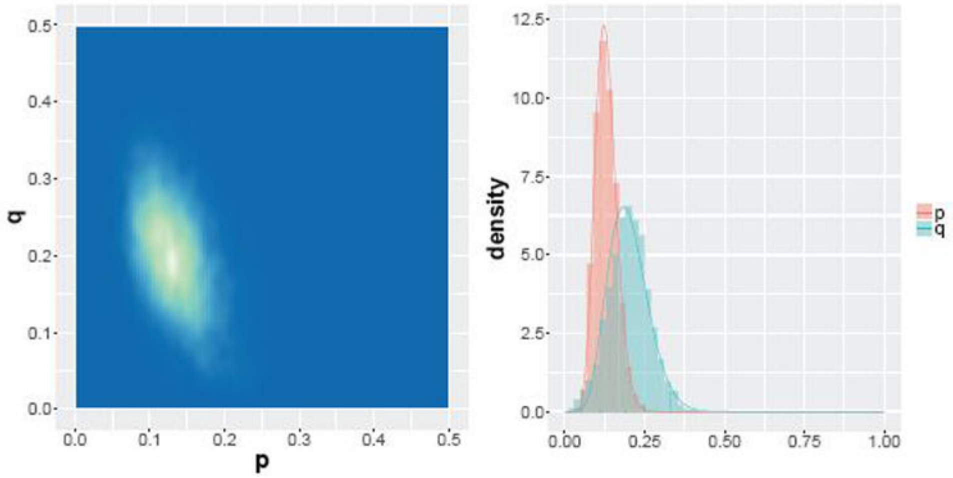

Fig. 3.

Joint (left) and marginal (right) distributions of p and q for the aspirin example.

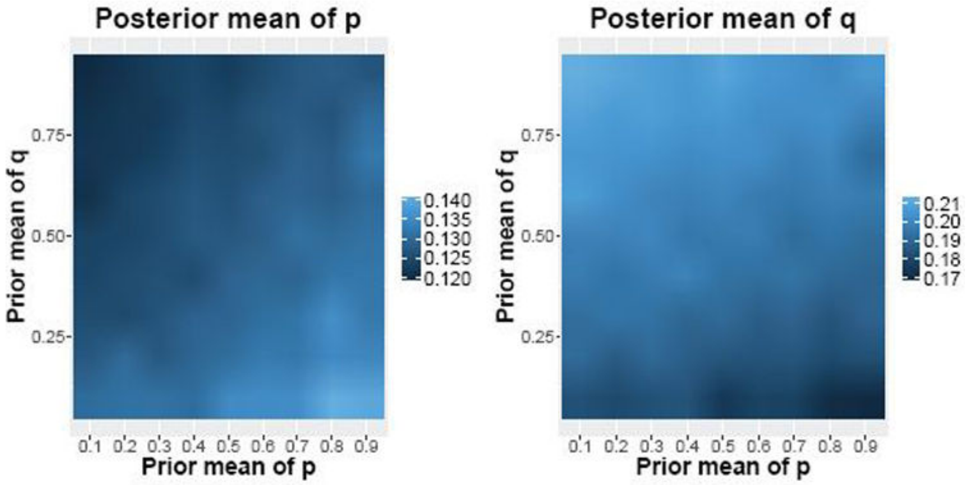

Fig. 4.

Sensitivity analysis: Posterior mean of *p* (left) and posterior mean of *q* (right) for various prior means.

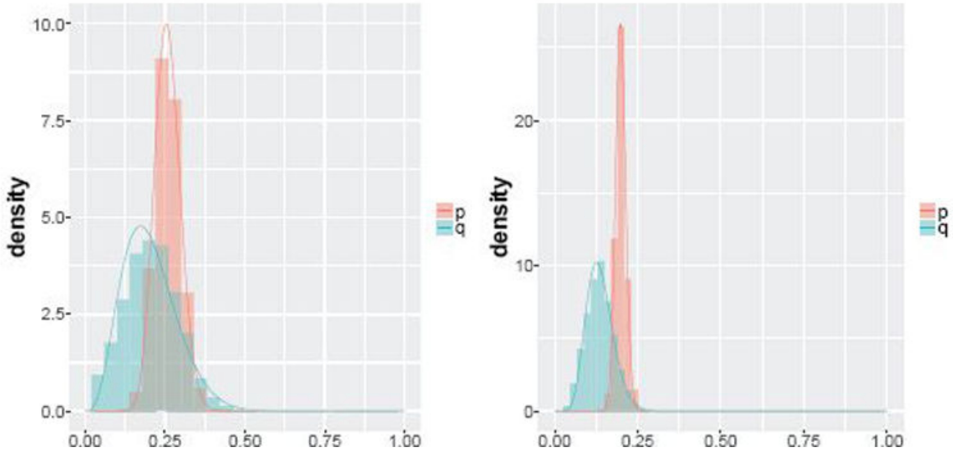

Fig. 5.

Marginal posterior distributions of p and q for infant birthweight datasets (left: Baystate Medical Center during 1986, right: Boston Medical Center during 1998–2000)

Table 1

Contigency Table. The table reporting the data $D = D_1 \cup D_2$ resulting from a fictitious double-blind randomized experiment involving 200 individuals, with 100 given chalk tablets (the control) (D_1), and 100 aspirin tablets (D_2). The patients take their assigned tablets the next time they get a headache, and record how long it is until the headache has gone. Recovery is interpreted as (for instance) "headache disappears within 30 minutes."

	No Recovery	**Recovery**	**Total**
Chalk (D_1)	88	12	100
Aspirin (D_2)	70	30	100

Table 2

Posterior summaries for the aspirin example.

	95% CI	Posterior Mean	Posterior Standard Deviation
p	(0.07, 0.20)	0.13	0.03
q	(0.08, 0.32)	0.20	0.06
$\frac{q}{q+p-pq}$	(0.34, 0.85)	0.65	0.12

Table 3

Maternal smoking status and infant birthweight based on data collected from Baystate Medical Center.

	Normal birthweight	**Low birthweight**
Non-smoking	86	29
Smoking during pregnancy	44	30

Table 4

Posterior summaries for the birthweight example based on data collected from Baystate Medical Center.

	95% CI	Posterior Mean	Posterior Standard Deviation
p	(0.18, 0.34)	0.26	0.04
q	(0.04, 0.37)	0.21	0.08
$\frac{q}{q+p-pq}$	(0.12, 0.73)	0.48	0.16

Table 5

Maternal smoking status and infant birthweight based on data collected from Boston Medical Center.

	Normal birthweight	Low birthweight
Non-smoking	460	107
Smoking during pregnancy	122	52

Table 6

Posterior summaries for the birthweight example based on data collected from Boston Medical Center.

	95% CI	Posterior Mean	Posterior Standard Deviation
p	(0.17, 0.23)	0.20	0.02
q	(0.06, 0.22)	0.14	0.04
$\frac{q}{q+p-pq}$	(0.24, 0.60)	0.43	0.09