*Article*

# Causal Confirmation Measures: From Simpson's Paradox to COVID-19

Chenguang Lu [1,2]

1    Intelligence Engineering and Mathematics Institute, Liaoning Technical University, Fuxin 123000, China, survival99@gmail.com
2    School of Computer Engineering and Applied Mathematics, Changsha University, Changsha 410022, China

**Abstract:** When we compare the influences of two causes on an outcome, if the conclusion from every group is against that from the conflation, we think there is Simpson's Paradox. The Existing Causal Inference Theory (ECIT) can make the overall conclusion consistent with the grouping conclusion by removing the confounder's influence to eliminate the paradox. The ECIT uses relative risk difference $P_d = \max(0, (R-1)/R)$ ($R$ denotes the risk ratio) as the probability of causation. In contrast, Philosopher Fitelson uses confirmation measure $D$ (posterior probability minus prior probability) to measure the strength of causation. Fitelson concludes that from the perspective of Bayesian confirmation, we should directly accept the overall conclusion without considering the paradox. The author proposed a Bayesian confirmation measure $b^*$ similar to $P_d$ before. To overcome the contradiction between the ECIT and Bayesian confirmation, the author uses the semantic information method with the minimum cross-entropy criterion to deduce causal confirmation measure $Cc = (R-1)/\max(R, 1)$. $Cc$ is like $P_d$ but has normalizing property (between $-1$ and 1) and cause symmetry. It especially fits cases where a cause restrains an outcome, such as the COVID-19 vaccine controlling the infection. Some examples (about kidney stone treatments and COVID-19) reveal that $P_d$ and $Cc$ are more reasonable than $D$; $Cc$ is more useful than $P_d$.

**Keywords:** causal confirmation; Bayesian confirmation; causal inference; semantic information measure; cross-entropy; Simpson's Paradox; COVID-19; risk measures

## 1. Introduction

Causal confirmation is the expansion of Bayesian confirmation. It is also a task of causal inference. The Existing Causal Inference Theory (ECIT), including Rubin's (or Neyman-Rubin) potential outcomes model [1,2] and Pearl's causal graph [3,4], has achieved great success. However, causal confirmation is rarely mentioned.

Bayesian confirmation theories are also called confirmation theories, which can be divided into incremental and inductive schools. The incremental school affirms that the confirmation measures the supporting strength of evidence $e$ to hypothesis $h$, as explained by Fitelson [5]. Following Carnap [6], the incremental school's researchers often use the increment of a hypothesis' probability or logical probability, $P(h \mid e) - P(h)$, as a confirmation measure. Fitelson [5] discussed causal confirmation with this measure and obtained some conclusions incompatible with the ECIT. On the other hand, the inductive school [7,8] considers confirmation as induction's modern form, whose task is to measure a major premise's creditability supported by a sample or sampling distribution.

We use $e \to h$ to denote a major premise. Variable $e$ takes one of two possible values $e_1$ and its negation $e_0$. Variable $h$ takes one of two possible values $h_1$ and its negation $h_0$. Then a sample includes four examples $(e_1, h_1)$, $(e_1, h_0)$, $(e_0, h_1)$, and $(e_0, h_0)$ with different proportions. The inductive school's researchers often use positive examples and counterexamples' proportions ($P(e_1 \mid h_1)$ and $P(e_1 \mid h_0)$) or likelihood ratio ($P(e_1 \mid h_1)/P(e_1 \mid h_0)$) to express confirmation measures.

A confirmation measure is often denoted by *C(e, h)* or *C(h, e)*. The author (of this paper) agrees with the inductive school and suggests using *C(e→h)* to express a confirmation measure so that the task is evident [8]. In this paper, we use "*x=>y*" to denote "Cause *x* leads to outcome *y*".

Although the two schools understand confirmation differently, both use sampling distribution *P(e, h)* to construct confirmation measures. There have been many confirmation measures [8,9]. Most researchers agree that an ideal confirmation measure should have the following two desired properties:

- normalizing property [9,10], which means *C(e, h)* should change between −1 and 1 so that the difference between a rule *e→h* and the best or the worst rule is clear;
- hypothesis symmetry [11] or consequent symmetry [8], which means $C(e_1 \to h_1) = -C(e_1 \to h_0)$. For example, *C*(raven→black) = −*C*(raven→non-black).

The author in [8] distinguished channels' confirmation and predictions' confirmation and provided channels' confirmation measure *b\*(e→h)* and predictions' confirmation measure *c\*(e→h)*. Both have the above two desired properties and can be used for the probability predictions of *h* according to *e*.

Bayesian confirmation confirms associated relationships, which are different from causal relationships. Association includes causality, but many associated relationships are not causal relationships. One reason is that the existence of association is symmetrical (if $P(h|e) \neq 0$, then $P(e|h) \neq 0$), whereas the existence of causality is asymmetrical. For example, in medical tests, *P*(positive | infected) reflects both association and causality. However, inversely, *P*(infected | positive) only indicates association. Another reason is that two associated events, *A* and *B*, such as electric fans' easy selling and air conditioners' easy selling, are the outcomes caused by the third event (hot weather). Neither *P(A | B)* nor *P(B | A)* indicates causality.

Causal inference only deals with uncertain causal relationships in nature and human society without considering those in mathematics, such as $(x + 1)(x - 1) < x^2$ because $(x + 1)(x - 1) = x^2 - 1$. We know that Kant distinguishes analytic judgments and synthetic judgments. Although causal inference is a mathematical method, it is used for synthetic judgments to obtain uncertain rules in biology, psychology, economics, etc. In addition, causal confirmation only deals with binary causality.

Although causal confirmation was rarely mentioned in the ECIT, the researchers of causal inference and epidemiology have provided many measures (without using the term "confirmation measure") to indicate the strength of causation. These measures include risk difference [12]:

$$RD = P(y_1|x_1) - P(y_1|x_0),\tag{1}$$

relative risk difference or the risk ratio (like the likelihood ratio for medical tests):

$$RR = P(y_1|x_1)/P(y_1|x_0),\tag{2}$$

and the probability of causation $P_d$ (used by Rubin and Greenland [13]) or the probability of necessity *PN* (used by Pearl [3]). There is:

$$P_d = PN = \max(0, \frac{P(y_1|x_1) - P(y_1|x_0)}{P(y_1|x_1)}).\tag{3}$$

$P_d$ is also called Relative Risk Reduction (RRR) [12]. In the above formula, $\max(0, \cdot)$ means its minimum is 0. This function is to make $P_d$ more like a probability. Measure *b\** proposed by the author in [8] is like $P_d$, but *b\** changes between −1 and 1. The above risk measures can measure not only risk or relative risk but also success or relative success raised by the cause.

The risk measures in Equations (1)–(3) are significant; however, they do not possess the two desired properties and hence are improper as causal confirmation measures.

We will encounter Simpson's Paradox if we only use sampling distributions for the above measures. Simpson's Paradox has been accompanying the study of causal inference, as the Raven Paradox has been going with the study of Bayesian confirmation. Simpson proposed the paradox [14] using the following example.
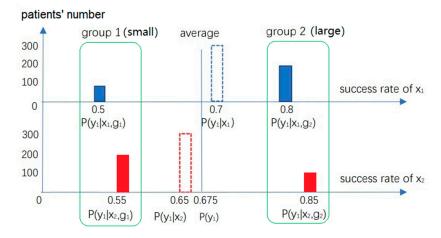
**Example 1** [15]. *The admission data of the graduate school of the University of California, Berkeley (UCB), for the fall of 1973 showed that 44% of male applicants were accepted, whereas only 35% of female applicants were accepted. There was probably gender bias present. However, in most departments, female applicants' acceptance rates were higher than male applicants.*

Was there a gender bias? Should we accept the overall conclusion or the grouping conclusion (i.e., that from every department)? If we take the overall conclusion, we can think that the admission had a bias against the female. On the other hand, if we accept the grouping conclusion, we can say that the female applicants were priorly accepted. Therefore, we say there exists a paradox.

Example 1 is a little complicated and easy to raise arguments against. To simplify the problem, we use Example 2, which the researchers of causal inference often mentioned, to explain Simpson's Paradox quantitatively.

We use $x_1$ to denote a new cause (or treatment) and $x_0$ to denote a default cause or no cause. If we need to compare two causes, we may use $x_1$ and $x_2$, or $x_i$ and $x_j$, to represent them. In these cases, we may assume that one is default like $x_0$.

**Example 2** [16,17]. *Suppose there are two treatments, $x_1$ and $x_2$, for patients with kidney stones. Patients are divided into two groups according to their size of stones. Group $g_1$ includes patients with small stones, and group $g_2$ has large ones. Outcome $y_1$ represents the treatment's success. Success rates shown in Figure 1 are possible. In each group, the success rate of $x_2$ is higher than that of $x_1$; however, the overall conclusion is the opposite.*



**Figure 1.** Illustrating Simpson's Paradox. In each group, the success rate of $x_2$, $P(y_1 \mid x_2, g)$, is higher than that of $x_1$, $P(y_1 \mid x_1, g)$; however, using the method of finding the center of gravity, we can see that the overall success rate of $x_2$, $P(y_1 \mid x_2) = 0.65$, is lower than that of $x_1$, $P(y_1 \mid x_1) = 0.7$.

According to Rubin's potential outcomes model [1], we should accept the grouping conclusion: $x_2$ is better than $x_1$. The reason is that the stones' size is a confounder, and the overall conclusion is affected by the confounder. We should eliminate this influence. The method is to imagine the patients' numbers in each group are unchanged whether we use $x_1$ or $x_2$. Then we replace weighting coefficients $P(g_i \mid x_1)$ and $P(g_i \mid x_2)$ with $P(g_i)$ ($i = 1, 2$) to obtain two new overall success rates. Rubin [1] expresses them as $P(y_1^{x_1})$ and $P(y_1^{x_2})$; whereas Pearl [3] expresses them as $P(y_1 \mid do(x_1))$ and $P(y_1 \mid do(x_2))$. Then, the overall conclusion is consistent with the grouping conclusion.

Should we always accept the grouping conclusion when the two conclusions are inconsistent? It is not sure! Example 3 is a counterexample.

**Example 3** (from [18]). *Treatment $x_1$ denotes taking a kind of antihypertensive drug, and treatment $x_0$ means taking nothing. Outcome $y_1$ denotes recovering health, and $y_0$ means not. Patients are divided into group $g_1$ (with high blood pressure) and group $g_0$ (with low blood pressure). It is very possible that in each group g, $P(y_1 | g, x_1) < P(y_1 | g, x_0)$ (which means $x_0$ is better than $x_1$); whereas overall result is $P(y_1 | x_1) > P(y_1 | x_0)$ (which means $x_1$ is better than $x_0$).*

The ECIT tells us that we should accept the overall conclusion that $x_1$ is better than $x_0$ because blood pressure is a mediator, which is also affected by $x_1$. We expect that $x_1$ can move a patient from $g_1$ to $g_0$; hence we need not change the weighting coefficients from $P(g | x)$ to $P(g)$. The grouping conclusion, $P(y_1 | g, x_1) < P(y_1 | g, x_0)$, exists because the drug has a side effect.

There are also some examples where the grouping conclusion is acceptable from one perspective, and the overall conclusion is acceptable from another.

**Example 4** [19]. *The United States statistical data about COVID-19 in June 2020 show that COVID-19 led to a higher Case Fatality Rate (CFR) of non-Hispanic whites than others (overall conclusion). We can find that only 35.3% of the infected people were non-Hispanic whites, whereas 49.5% of the infected people who died from COVID-19 were non-Hispanic whites. It seems that COVID-19 is more dangerous to non-Hispanic whites. However, Dana Mackenzie pointed out [19] that we will obtain the opposite conclusion from every age group because the CFR of non-Hispanic whites is lower than that of other people in every age group. So, there exists Simpson's Paradox. The reason is that non-Hispanic whites have longer lifespans and a relatively large proportion of the elderly, while COVID-19 is more dangerous to the elderly.*

Kügelgen et al. [20] also pointed out the existence of Simpson's Paradox after they compared the CFRs of COVID-19 (reported in 2020) in China and Italy. Although the overall conclusion was that the CFR in Italy was higher than in China, the CFR of every age group in China was higher than in Italy. The reason is that the proportion of the elderly in Italy is larger than in China.

According to Rubin's potential outcomes model or Pearl's causal graph, if we think that the reason for non-Hispanic whites' longevity is good medical conditions or other elements instead of their race, then the lifespan is a confounder. Therefore, we should accept the grouping conclusion. On the other hand, if we believe that non-Hispanic whites are longevous because they are whites, then the lifespan is a mediator, so we should accept the overall conclusion.

Example 1 is similar to Example 4, but the former is not easy to understand. The data show that the female applicants tended to choose majors with low admission rates (perhaps because lower thresholds resulted in more intense competition). This tendency is like the lifespan of the white. If we regard lifespan as a confounder, Berkeley University had no gender bias against the female. On the other hand, if we believe the female is the cause of this tendency, the overall conclusion is acceptable, and gender bias should have existed. Deciding which of the two judgments is right depends on one's perspective.

Pearl's causal graph [3] makes it clear that for the same data, if supposed causal relationships are different, conclusions are also different. So, it is not enough to have data only. We also need the structural causal model.

However, the incremental school's philosopher Fitelson argues that from the perspective of Bayesian confirmation, we should accept the overall conclusion according to the data without considering causation; Simpson's Paradox does not exist according to his rational explanation. His reason is that we can use the measure [5]:

$$measure\ _i = P(y_1 | x_1) - P(y_1) \tag{4}$$

to measure causality. Fitelson proves (see Fact 3 of Appendix in [5]) that if there is:

$$P(y_1 \mid x_1, g_i) > P(y_1 \mid x_2, g_i), \ i=1, 2, \tag{5}$$

then there must be $P(y_1 \mid x_1) > P(y_1)$. The result is the same when ">" is replaced with "<". Therefore, Fitelson affirms that, unlike $RD$ and $P_d$, *measure i* does not result in the paradox.

However, Equation (5) expresses a rigorous condition, which excludes all examples with joint distributions $P(y, x, g)$ that cause the paradox, including Fitelson's simplified example about the admissions of the UCB.

One cannot help asking:

- For Example 2 about kidney stones, is it reasonable to accept the overall conclusion without considering the difficulties of treatments?
- Is it necessary to extend or apply a Bayesian confirmation measure incompatible with the ECIT and medical practices to causal confirmation?
- Except for the incompatible confirmation measures, are there no compatible confirmation measures?

In addition to the incremental school's confirmation measures, there are also the inductive school's confirmation measures, such as $F$ proposed by Kemeny and Oppenheim in 1952 and $b^*$ provided by the author in 2020.

This paper mainly aims at:

- combining the ECIT to deduce causal confirmation measure $Cc(x_1 => y_1)$ ("C" stands for confirmation and "c" for the cause), which is similar to $P_d$ but can measure negative causal relationships, such as "vaccine => infection";
- explaining that measures $Cc$ and $P_d$ are more suitable for causal confirmation than *measure $_i$* by using some examples with Simpson's Paradox;
- supporting the inductive school of Bayesian confirmation in turn.

When the author proposed measure $b^*$, he also provided measure $c^*$ for eliminating the Raven Paradox [8]. For extending $c^*$ to causal confirmation, this paper presents measure $Ce(x_1 => y_1)$, which indicates the outcome's inevitability or the cause's sufficiency.

## 2. Background

### 2.1. Bayesian Confirmation: Incremental School and Inductive School

A universal judgment is equivalent to a hypothetical judgment or a rule, such as "All ravens are black" is equivalent to "For every $x$, if $x$ is a raven, then $x$ is black". Both can be used as a major premise for a syllogism. Due to the criticism of Hume and Popper, most philosophers no longer expect to obtain absolutely correct universal judgments or major premises by induction but hope to obtain their degrees of belief. A degree of belief supported by a sample or sampling distribution is the degree of confirmation.

It is worth noting that a proposition does not need confirmation. Its truth value comes from its usage or definition [8]. For example, "People over 18 are adults" does not need confirmation; whether it is correct depends on the government's definition. Only major premises (such as "All ravens are black" and "If a person's Nucleic Acid Test is positive, he is likely to be infected with COVID-19") need confirmation.

A natural idea is to use conditional probability $P(h \mid e)$ to confirm a major premise or rule denoted with $e \rightarrow h$. This measure is also recommended by Fitelson [5], and called *confirm $_f$*. There is [5,6]:

$$confirm\ _f = f(e,h) = P(h \mid e). \text{ (Carnap, 1962, Fitelson, 2017)}$$

However, $P(h \mid e)$ depends very much on the prior probability $P(h)$ of $h$. For example, where COVID-19 is prevalent, $P(h)$ is large, and $P(h \mid e)$ is also large. Therefore, $P(h \mid e)$ cannot reflect the necessity of $e$. An extreme example is that $h$ and $e$ are independent of each other, but if $P(h)$ is large, $P(h \mid e) = P(h, e)/P(e) = P(h)$ is also large. At this time, $P(h \mid e)$ does not reflect the creditability of the causal relationship. For example, $h =$ "There will be

no earthquake tomorrow", $P(h) = 0.999$, and $e$ = "Grapes are ripe". Although $e$ and $h$ are irrelative, $P(h \mid e) = P(h) = 0.999$ is very large. However, we cannot say that the ripe grape supports no earthquake happening.

For this reason, the incremental school's researchers use posterior (or conditional) probability minus prior probability to express the degree of confirmation. These confirmation measures include [6,10,21,22]:

$$D(e_1, h_1) = P(h_1 \mid e_1) - P(h_1) \qquad \text{(Carnap, 1962)},$$

$$M(e_1, h_1) = P(e_1 \mid h_1) - P(e_1) \qquad \text{(Mortimer, 1988)},$$

$$R(e_1, h_1) = \log[P(h_1 \mid e_1)/P(h_1)] \qquad \text{(Horwich, 1982)},$$

$$C(e_1, h_1) = P(h_1, e_1) - P(e_1)\, P(h_1) \qquad \text{(Carnap, 1962)},$$

$$Z(h_1, e_1) = \begin{cases} [P(h_1 \mid e_1) - P(h_1)]/P(h_0), & \text{as } P(h_1 \mid e_1) \ge P(h_1), \\ [P(h_1 \mid e_1) - P(h_1)]/P(h_1), & \text{otherwise,} \end{cases} \qquad \text{(Crupi et al., 2007)}.$$

In the above measures, $D(e_1, h_1)$ is *measure* $_i$ recommended by Fitelson in [5]. $R(e_1, h_1)$ is an information measure. It can be written as $\log P(h_1 \mid e_1) - \log P(h_1)$. Since $\log P(h_1 \mid e_1) - \log P(h_1) = \log P(e_1 \mid h_1) - \log P(e_1) = \log P(h_1, e_1) - \log[P(h_1)P(e_1)]$, $D$, $M$, and $C$ increase with $R$ and hence can be replaced with each other. $Z$ is the normalization of $D$ for having the two desired properties [10]. Therefore, we can also call the incremental school the information school.

On the other hand, the inductive school's researchers use the difference (or likelihood ratio) between two conditional probabilities representing the proportions of positive and negative examples to express confirmation measures. These measures include [7,8,23–25]:

$$S(e_1, h_1) = P(h_1 \mid e_1) - P(h_1 \mid e_0) \qquad \text{(Christensen, 1999)},$$

$$N(e_1, h_1) = P(e_1 \mid h_1) - P(e_1 \mid h_0) \qquad \text{(Nozick, 1981)},$$

$$L(e_1, h_1) = \log[\, P(e_1 \mid h_1)/P(e_1 \mid h_0)] \qquad \text{(Good, 1984)},$$

$$F(e_1, h_1) = [P(e_1 \mid h_1) - P(e_1 \mid h_0)]/[\, P(e_1 \mid h_1) + P(e_1 \mid h_0)] \quad \text{(Kemeny and Oppenheim, 1952)},$$

$$b^*(e_1, h_1) = [P(e_1 \mid h_1) - P(e_1 \mid h_0)]/\max(P(e_1 \mid h_1), P(e_1 \mid h_0)) \quad \text{(Lu, 2020)}.$$

They are all positively related to the Likelihood Ratio ($LR^+ = P(e_1 \mid h_1)/P(e_1 \mid h_0)$). For example, $L = \log LR^+$ and $F = (LR^+ - 1)/(LR^+ + 1)$ [7]. Therefore, these measures are compatible with risk (or reliability) measures, such as $P_d$, used in medical tests and disease control. Although the author has studied semantic information theory for a long time [26–28] and believe both schools have made important contributions to Bayesian confirmation, he is on the side of the inductive school. The reason is that information evaluation occurs before classification, whereas confirmation is needed after classification [8].

Although the researchers understand confirmation differently, they all agree to use a sample including four types of examples $(e_1, h_1)$, $(e_0, h_1)$, $(e_1, h_0)$, and $(e_0, h_0)$ with different proportions as the evidence to construct confirmation measures [8,10]. The main problem with the incremental school is that they do not distinguish the evidence of a major premise and that of the consequent of the major premise well. When they use the four examples' proportions to construct confirmation measures, $e$ is regarded as the major premise's antecedent, whose negation $e_0$ is meaningful. However, when they say "to evaluate the supporting strength of $e$ to $h$", $e$ is understood as a sample, whose negation $e_0$ is meaningless. It is more meaningless to put a sample $e$ or $e_0$ in an example $(e_1, h_1)$ or $(e_0, h_1)$.

We compare $D$ (i.e., *measure* $_i$) and $S$ to show the main difference between the two schools' measures. Since:

$$\begin{aligned} D(e_1, h_1) &= P(h_1 \mid e_1) - P(h_1) = P(h_1 \mid e_1) - [P(e_1)P(h_1 \mid e_1) + P(e_0)P(h_1 \mid e_0)] \\ &= [1 - P(e_1)]P(h_1 \mid e_1) - P(e_0)P(h_1 \mid e_0) = P(e_0)S(e_1, h_1), \end{aligned} \qquad (6)$$

we can find that $D$ changes with $P(e_0)$ or $P(e_1)$, but $S$ does not. $P(e)$ means the source and $P(h\,|\,e)$ means the channel. $D$ is related to the source and the channel, but $S$ is only related to the channel. Measures $F$ and $b^*$ are also only related to channel $P(e\,|\,h)$. Therefore, the author calls $b^*$ the channels' confirmation measure.

*2.2. The P-T Probability Framework and the Methods of Semantic Information and Cross-Entropy for Channels' Confirmation Measure $b^*(e \rightarrow h)$*

In the P-T probability framework [28] proposed by the author, there are both statistical probability $P$ and logical probability (or truth value) $T$; the truth function of a predicate is also a membership function of a fuzzy set [29]. Therefore, the truth function also changes between 0 and 1. The purpose of proposing this probability framework is to set up the bridge between statistics and fuzzy logic.

Let $X$ be a random variable representing an instance, taking a value $x \in A = \{x_0, x_1, \dots \}$, and $Y$ be a random variable representing a label or hypothesis, taking a value $y \in B = \{y_0, y_1, \dots \}$. The Shannon channel is a conditional probability matrix $P(y_j\,|\,x_i)$ $(i = 1, 2, \dots; j = 1, 2, \dots )$ or a set of transition probability functions $P(y_j\,|\,x)$ $(j = 1, 2, \dots )$. The semantic channel is a truth value matrix $T(y_j\,|\,x_i)$ $(i = 1, 2, \dots ; j = 1, 2, \dots )$ or a set of truth functions $T(y_j\,|\,x)$ $(j = 0, 1, \dots )$. Let the elements in $A$ that make $y_j$ true form a fuzzy subset $\theta_j$. The membership function $T(\theta_j\,|\,x)$ of $\theta_j$ is also the truth function $T(y_j\,|\,x)$ of $y_j$, i.e., $T(\theta_j\,|\,x) = T(y_j\,|\,x)$.

The logical probability of $y_j$ is:

$$T(y_j) = T(\theta_j) = \sum_i P(x_i) T(\theta_j\,|\,x_i). \tag{7}$$

Zadeh calls it the fuzzy event's probability [30]. When $y_j$ is true, the conditional probability of $x$ is:

$$P(x\,|\,\theta_j) = P(x) T(\theta_j\,|\,x) / T(\theta_j). \tag{8}$$

Fuzzy set $\theta_j$ can also be understood as a model parameter; hence $P(x\,|\,\theta_j)$ is a likelihood function.

The differences between logical probability and statistical probability are:

- The statistical probability is normalized (the sum is 1), whereas the logical probability is not. Generally, we have $T(\theta_0) + T(\theta_1) + \dots > 1$.
- The maximum value of $T(\theta_j\,|\,x)$ is 1 for different $x$, whereas $P(y_0\,|\,x) + P(y_1\,|\,x) + \dots = 1$ for a given $x$.

We can use the sample distribution to optimize the model parameters. For example, we use $x$ to represent the age, use a logistic function as the truth function of the elderly: $T(\text{``elderly''}\,|\,x) = 1/[1 + \exp(-bx + a)]$, and use a sampling distribution to optimize $a$ and $b$.

The (amount of) semantic information about $x_i$ conveyed by $y_j$ is:

$$I(x_i; \theta_j) = \log \frac{P(x_i\,|\,\theta_j)}{P(x_i)} = \log \frac{T(\theta_j\,|\,x_i)}{T(\theta_j)}. \tag{9}$$

For different $x$, the average semantic information conveyed by $y_j$ is:

$$
\begin{aligned}
I(X; \theta_j) \ &= \sum_i P(x_i\,|\,y_j) \log \frac{T(\theta_j\,|\,x_i)}{T(\theta_j)} = \sum_i P(x_i\,|\,y_j) \log \frac{P(x_i\,|\,\theta_j)}{P(x_i)} \\
&= -\sum_i P(x_i\,|\,y_j) \log P(x_i) - H(X\,|\,\theta_j).
\end{aligned}
\tag{10}
$$

In the above formula, $H(X\,|\,\theta_j)$ is a cross-entropy:

$$H(X\,|\,\theta_j) = -\sum_i P(x_i\,|\,y_j) \log P(x_i\,|\,\theta_j). \tag{11}$$

The cross-entropy has an important property: when we change $P(x\,|\,\theta_j)$ so that $P(x\,|\,\theta_j) = P(x\,|\,y_j)$, $H(X\,|\,\theta_j)$ reaches its minimum. It is easy to find from Equation (10) that $I(X;$

$\theta_j$) reaches its maximum as $H(X|\theta_j)$ reaches its minimum. The author has proved that if $P(x|\theta_j) = P(x|y_j)$, then $T(\theta_j|x) \propto P(y_j|x)$ [27]. If for all $j$, $T(\theta_j|x) \propto P(y_j|x)$, we say that the semantic channel matches the Shannon channel.

We use the medical test as an example to deduce the channels' conformation measure $b^*$. We define $h \in \{h_0, h_1\}$ = {infected, uninfected} and $e \in \{e_0, e_1\}$ = {positive, negative}. The Shannon channel is $P(e|h)$, and the semantic channel is $T(e|h)$. The major premise to be confirmed is $e_1 \to h_1$, which means "If one's test is positive, then he is infected".

We regard a fuzzy predicate $e_1(h)$ as the linear combination of a clear predicate (whose truth value is 0 or 1) and a tautology (whose truth value is always 1). Let the tautology's proportion be $b_1'$ and the clear predicate's proportion be $1 - b_1'$. Then we have:

$$T(e_1|h_0) = b_1'; \; T(e_1|h_1) = b_1' + b_1 = b_1' + (1 - b_1') = 1. \tag{12}$$

The $b_1'$ is also called the degree of disbelief of rule $e_1 \to h_1$. The degree of disbelief optimized by a sample, denoted by $b_1'^*$, is the degree of disconfirmation. Let $b_1^*$ denote the degree of confirmation; we have $b_1'^* = 1 - |b_1^*|$. By maximizing average semantic information $I(H; \theta_1)$ or minimizing cross-entropy $H(H|\theta_j)$, we can deduce (see Section 3.2 in [8]):

$$b_1^* = b^*(e_1 \to h_1) = \frac{P(e_1|h_1) - P(e_1|h_0)}{\max(P(e_1|h_1), P(e_1|h_0))} = \frac{LR^+ - 1}{\max(LR^+, 1)}. \tag{13}$$

Suppose that likelihood function $P(h|e_1)$ is decomposed into an equiprobable part and a part with 0 and 1. Then, we can deduce the predictions' confirmation measure $c^*$:

$$c_1^* = c^*(e_1 \to h_1) = \frac{P(h_1|e_1) - P(h_0|e_1)}{\max(P(h_1|e_1), P(h_0|e_1))} = \frac{2P(h_1|e_1) - 1}{\max(P(h_1|e_1), 1 - P(h_1|e_1))}. \tag{14}$$

Measure $b^*$ is compatible with the likelihood ratio and suitable for evaluating medical tests. In contrast, measure $c^*$ is appropriate to assess the consequent inevitability of a rule and can be used to clarify the Raven Paradox [8]. Moreover, both measures have the normalizing property and symmetry mentioned above.

### 2.3. Causal Inference: Talking from Simpson's Paradox

According to the ECIT, the grouping conclusion is acceptable for Example 2 (about kidney stones), whereas the overall conclusion is acceptable for Example 3 (about blood pressure). The reason is that $P(y_1|x_1)$ and $P(y_1|x_0)$ may not reflect causality well; in addition to the observed data or joint probability distribution $P(y, x, g)$, we also need to suppose the causal structure behind the data [3].

Suppose there is the third variable, $u$. Figure 2 shows the causal relationships in Examples 2, 3, and 4. Figure 2a shows the causal structure of Example 2, where $u$ (kidney stones' size) is a confounder that affects both $x$ and $y$. Figure 2b describes the causal structure of Example 3, where $u$ (blood pressure) is a mediator that affects $y$ but is affected by $x$. In Figure 2c, $u$ can be interpreted as either a confounder or a mediator. The causality will differ from different perspectives, and $P(y_1|\text{do}(x))$ will also differ. In all cases, we should replace $P(y|x)$ with $P(y|\text{do}(x))$ (if they are different) to get $RD$, $RR$, and $P_d$.
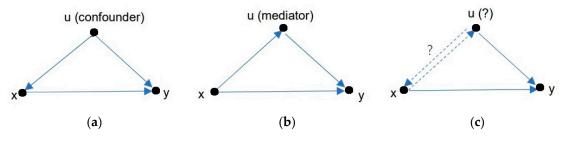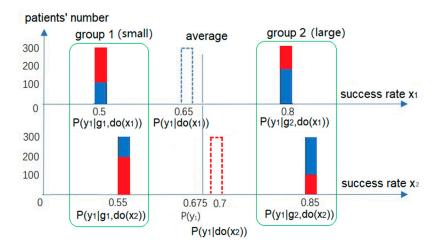


**Figure 2.** Three causal graphs: (**a**) for Example 2; (**b**) for Example 3; (**c**) for Examples 4 and 1.

We should accept the overall conclusion for the example where $u$ is a mediator. However, for the example where $u$ is a confounder, how do we obtain a suitable $P(y \mid do(x))$? According to Rubin's potential outcomes model, we use Figure 3 to explain the difference between $P(y \mid do(x))$ and $P(y \mid x)$.



**Figure 3.** Eliminating Simpson's Paradox as the confounder exists by modifying the weighting coefficients. After replacing $P(g_k \mid x_i)$ with $P(g_k)$ ($k = 1,2$; $i = 1,2$), the overall conclusion is consistent with the grouping conclusion; the average success rate of $x_2$, $P(y_1 \mid do(x_2)) = 0.7$, is higher than that of $x_1$, $P(y_1 \mid do(x_1)) = 0.65$.

To find the difference in the outcomes caused by $x_1$ and $x_2$, we should compare the two outcomes in the same background. However, there is often no situation where other conditions remain unchanged except for the cause. For this reason, we need to replace $x_1$ with $x_2$ in our imagination and see the shift in $y_1$ or its probability. If $u$ is a confounder and not affected by $x$, the number of members in $g_1$ and $g_2$ should be unchanged with $x$, as shown in Figure 3. The solution is to use $P(g)$ instead of $P(g \mid x)$ for the weighting operation so that the overall conclusion is consistent with the grouping conclusion. Hence, the paradox no longer exists.

Although $P(x_0) + P(x_1) = 1$ is tenable, $P(do\ (x_1)) + P(do\ (x_0)) = 1$ is meaningless. That is why Rubin emphasizes that $P(y^x)$, i.e., $P(y \mid do(x))$, is still a marginal probability instead of a conditional probability, in essence.

Rubin's reason [2] for replacing $P(g \mid x)$ with $P(g)$ is that for each group, such as $g_1$, the two subgroups' members (patients) treated by $x_1$ and $x_2$ are interchangeable (i.e., Pearl's causal independence assumption mentioned in [5]). If a member is divided into the subgroup with $x_1$, its success rate should be $P(y_1 \mid g, x_1)$; if it is divided into the subgroup with $x_2$, the success rate should be $P(y_1 \mid g, x_2)$. $P(g \mid x_1)$ and $P(g \mid x_2)$ are different only because half of the data are missing. However, we can fill in the missing data using our imagination.

If $u$ is a mediator, as shown in Figure 2b, a member in $g_1$ may enter $g_2$ because of $x$, and vice versa. $P(g \mid x_0)$ and $P(g \mid x_1)$ are hence different without needing to be replaced with $P(g)$. We can let $P(y_1 \mid do\ (x)) = P(y_1 \mid x)$ directly and accept the overall conclusion.

*2.4. Probability Measures for Causation*

In Rubin and Greenland's article [13]:

$$P(t) = [R(t) - 1]/R(t) \tag{15}$$

is explained as the probability of causation, where $t$ is one's age of exposure to some harmful environment. $R(t)$ is the age-specific infection rate (infected population divided by uninfected population). Let $y_1$ stand for the infection, $x_1$ for the exposure, and $x_0$ for no exposure. Then there is $R(t) = P(y_1 \mid do(x_1), t)/P(y_1 \mid do(x_0), t)$. Its lower limit is 0 because

the probability cannot be negative. When the change of $t$ is neglected, considering the lower limit, we can write the probability of causation as:

$$P_d = \max(0, \frac{R-1}{R}) = \max(0, \frac{P(y_1|\text{do}(x_1)) - P(y_1|\text{do}(x_0))}{P(y_1|\text{do}(x_1))}). \tag{16}$$

Pearl uses *PN* to represent $P_d$ and explains *PN* as the probability of necessity [3]. $P_d$ is very similar to confirmation measure $b^*$ [8]. The main difference is that $b^*$ changes between $-1$ and 1.

Robert van Rooij and Katrin Schulz [31] argue that conditionals of the form "If $x$, then $y$" are assertable only if:

$$\Delta^* P_x^y = \frac{P(y_1|x_1) - P(y_1|x_0)}{1 - P(y_1|x_0)} \tag{17}$$

is high. This measure is similar to confirmation measure $Z$. The difference between $P_d$ and $\Delta^* P_x^y$ is that $P_d$, like $b^*$, is sensitive to counterexamples' proportion $P(y_1|x_0)$, whereas $\Delta^* P_x^y$ is not. Table 1 shows their differences.

**Table 1.** Comparing $P_d$ and $\Delta^* P_x^y$.

|  | $P(y_1|x_1)$ | $P(y_1|x_0)$ | $P_d$ | $\Delta^* P_x^y$ | Comparison |
|---|---|---|---|---|---|
| No big difference | 0.9 | 0.8 | 0.11 | 0.5 | $P_d \ll \Delta^* P_x^y$ |
| No counterexample | 0.2 | 0 | 1 | 0.2 | $P_d \gg \Delta^* P_x^y$ |

David E. Over et al. [32] support the Ramsey test hypothesis, implying that the subjective probability of a natural language conditional, $P(\text{if } p \text{ then } q)$, is the conditional subjective probability, $P(q|p)$. This measure is *confirm* $_f$ in [5].

The author [8] suggests that we should distinguish two types of confirmation measures for $x \Rightarrow y$ or $e \rightarrow h$. One is to stand for the necessity of $x$ compared with $x_0$; the other is for the inevitability of $y$. $P(y|x)$ may be good for the latter but not for the former. The former should be independent of $P(x)$ and $P(y)$. $P_d$ is such a one.

However, there is a problem with $P_d$. If $P_d$ is 0 when $y$ is uncorrelated to $x$, then $P_d$ should be negative instead of 0 when $x$ inversely affects $y$ (e.g., vaccine affects infection). Therefore, we need a confirmation measure between $-1$ and 1 instead of a probability measure between 0 and 1.

### 3. Methods

#### 3.1. Defining Causal Posterior Probability

To avoid treating association as causality, we first explain what kind of posterior probabilities indicate causality. Posterior probability and conditional probability are often regarded as the same. However, Rubin emphasizes that probability $P(y^x)$ is not conditional; it is still marginal. To distinguish $P(y^x)$ and marginal probability $P(y)$, we call $P(y^x)$, i.e., $P(y|\text{do}(x))$, the Causal Posterior Probability (CPP). What posterior probability is the CPP? We use the following example to explain.

About the population age distribution, let $z$ be age and the population age distribution be $p(z)$. We may define that a person with $z \geq 60$ is called an elderly; that is, $P(y_1|z) = 1$ for $z \geq z_0$. The label of an elderly is $y_1$, and the label of a non-elderly is $y_0$. The probability of the elderly is:

$$P(y_1) = \sum_{\text{all } z} p(z)P(y_1|z) = \sum_{z \geq 60} p(z) \tag{18}$$

Let $x_1$ denote the improved medical condition. After a period, $p(z)$ becomes $p(z^{x_1}) = p(z|\text{do}(x_1))$ and $P(y_1)$ becomes:

$$P(y_1^{x_1}) = P(y_1|\text{do}(x_1)) = \sum_{z \geq 60} p(z|\text{do}(x_1)), \tag{19}$$

Let $x_0$ be the medical condition existing already. We have:

$$P(y_1|\text{do}(x_0)) = \sum_{z \geq 60} p(z|\text{do}(x_0)). \tag{20}$$

There are similar examples:

- About whether a drug ($x_1$) can lower blood pressure, blood sugar, blood lipid, or uric acid ($z$) or not, if $z$ drops to a certain level $z_0$, we say that the drug is effective ($y_1$).
- About whether a fertilizer ($x_1$) can increase grain yield ($z$), if $z$ increases to a certain extent $z_0$, the grain yield is regarded as a bumper harvest ($y_1$).
- Can a process $x_1$ reduce the deviation $z$ of a product's size? If the deviation is smaller than the tolerance ($z_0$), we consider the product qualified ($y_1$).

From the above examples, we can find that the action $x$ can be the cause of a causal relationship because it can cause the change of probability distribution $p(z)$ of objective result $z$, rather than the change of probability distribution $P(y \mid \cdot)$ of outcome $y$. The reason is that $P(y \mid \cdot)$ also changes with the dividing boundary $z_0$. For example, if the dividing boundary of the elderly changes from $z_0 = 60$ to $z_0' = 65$, the posterior probability $P(y_1 \mid z_0')$ of $y_1$ will become smaller. This change seemingly also reflects causality. However, the author thinks this change is due to a mathematical cause, which does not reflect the causal relationship we want to study. Therefore, we need to define the CPP more specifically.

**Definition 1.** *Random variable Z takes a value $z \in \{z_1, z_2, \ldots \}$ and $p(z)$ is the probability distribution of the objective result. Random variable Y takes a value $y \in \{y_0, y_1\}$ and represents the outcome, i.e., the classification label of z. The cause or treatment is $x \in \{x_0, x_1\}$ or $\{x_1, x_2\}$. If replacing $x_0$ with $x_1$ (or $x_1$ with $x_2$) can cause the change of probability distribution $p(z)$, we call x the cause, $p(z \mid x)$ or $p(z^x)$ the CPP distribution, and $P(y^x) = P(y \mid do(x))$ the CPP.*

According to the above definition, given $y_1$, the conditional probability distribution $p(z \mid y_1)$ is not the CPP distribution because the probability distribution of $z$ does not change with $y$.

Suppose that $x_1$ is the vaccine for COVID-19, $y_1$ is the infection, and $e_1$ is the test-positive. Then $P(y_1 \mid x_1)$ or $P(y_1 \mid \text{do}(x_1))$ is the CPP, whereas $P(y_1 \mid e_1)$ is not. We may regard $y_1$ as the conclusion obtained by the best test, $e_1$ is from a common test, and $P(y_1 \mid e_1)$ is the probability prediction of $y_1$. $P(y_1 \mid e_1)$ is not a CPP because $e_1$ does not change $p(z)$ and the conclusion from the best test.

### 3.2. Using $x_2/x_1 => y_1$ to Compare the Influences of Two Causes on an Outcome

In associated relationships, $x_0$ is the negation of $x_1$; they are complementary. However, in causal relationships, $x_1$ is the substitute for $x_0$. For example, consider taking medicines to cure the disease. Let $x_0$ denote taking nothing, and $x_1$ and $x_2$ represent taking two different medicines. Each of $x_1$ and $x_2$ is a possible alternative to $x_0$ instead of the negation of $x_0$. Furthermore, in some cases, $x_1$ may include $x_0$ (see Section 4.3).

When we compare the effects of $x_2$ and $x_1$, it is unclear to use "$x_2 => y_1$" to indicate the causal relationship. Therefore, the author suggests that we had better replace "$x_2 => y_1$" with "$x_2/x_1 => y_1$", which means "replacing $x_1$ with $x_2$ will arise or increase $y_1$".

There are two reasons for using "$x_2/x_1$":

- One is to express symmetry ($Cc(x_2/x_1 => y_1) = - Cc(x_1/x_2 => y_1)$) conveniently.
- Another is to emphasize that $x_1$ and $x_2$ are not complementary but alternatives for eliminating Simpson's Paradox easily.

To compare $x_1$ with $x_0$, we may selectively use "$x_1/x_0 => y_1$" or "$x_1 => y_1$".

For Example 2 with a confounder, if we consider the treatment as replacing $x_2$ with $x_1$ in our imagination, we can easily understand why the number of patients in each group should be unchanged, that is, $P(g \mid x_1) = P(g \mid x_2) = P(g)$. The reason is that the replacement will not change everyone's kidney stone size.

In Example 3, $u$ is a mediator, and the number of people in each group (with high or low blood pressure) is also affected by taking an antihypertensive drug $x_1$. When we replace $x_0$ with $x_1$, $P(g \mid x_1) \neq P(g \mid x_0) \neq P(g)$ is reasonable, and hence the weighting coefficients need not be adjusted. In this case, we can directly let $P(y_1 \mid do(x)) = P(y_1 \mid x)$.

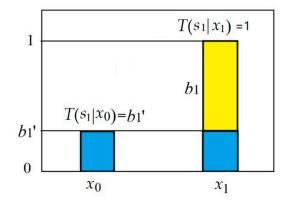*3.3. Deducing Causal Confirmation Measure Cc by the Methods of Semantic Information and Cross-Entropy*

We use $x_1 => y_1$ as an example to deduce the causal confirmation measure $Cc$. If we need to compare any two causes, $x_i$ and $x_k$, we may assume that one is default as $x_0$.
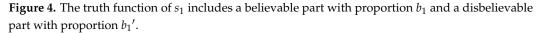
Let $s_1 =$ "$x_1 => y_1$" and $s_0 =$ "$x_0 => y_0$". We suppose that $s_1$ includes a believable part with proportion $b_1$ and a disbelievable part with proportion $b_1'$. Their relation is $b_1' + |b_1| = 1$. First, we assume $b_1 > 0$; hence $b_1 = 1 - b_1'$. The two truth values of $s_1$ are $T(s_1 \mid x_1)$ and $T(s_1 \mid x_0)$, as shown in the last row of Table 2.

**Table 2.** The truth values of $s_0 =$ "$x_0 => y_0$" and $s_1 =$ "$x_1 => y_1$".

|  | $T(s \mid x_0)$ | $T(s \mid x_1)$ |
|---|---|---|
| $s_0 =$ "$x_0 => y_0$" | 1 | $b_0'$ |
| $s_1 =$ "$x_1 => y_1$" | $b_1'$ | 1 |

Figure 4 shows how truth function $T(s_1 \mid x)$ is related to $b_1$ and $b_1'$ for $b_1 > 0$. $T(s_1 \mid x_1) = 1$ means that example $(x_1, y_1)$ makes $s_1$ fully true; $T(s_1 \mid x_0) = b_1'$ is the truth value and the degree of disbelief of $s_1$ for given counterexample $(x_0, y_1)$.



**Figure 4.** The truth function of $s_1$ includes a believable part with proportion $b_1$ and a disbelievable part with proportion $b_1'$.

The degree of belief optimized by a sampling distribution with the maximum semantic information or minimum cross-entropy criterion is the degree of causal confirmation, denoted by $Cc_1 = Cc(x_1/x_0 => y_1) = b_1{}^*$.

The logical probability of $s_1$ is (see Equation (7)):

$$T(s_1) = P(x_1) + P(x_0) b_1', \tag{21}$$

The predicted probability of $x_1$ by $y_1$ and $s_1$ is:

$$P(x_1|\theta_1) = \frac{T(s_1|x_1)P(x_1)}{T(s_1)} = \frac{P(x_1)}{P(x_1) + P(x_0)b_1'}, \tag{22}$$

where $\theta_j$ can be regarded as the parameter of truth function $T(s_j \mid x)$.

The average semantic information conveyed by $y_1$ and $s_1$ about $x$ is:

$$I(X;\theta_1) = \sum_i P(x_i|y_1) \log \frac{P(x_i|\theta_1)}{P(x_i)} = -\sum_i P(x_i|y_1) \log P(x_i) - H(X|\theta_1), \tag{23}$$

where $H(X | \theta_1)$ is a cross-entropy. We suppose that sampling distribution $P(x, y)$ has be modified so that $P(y | x) = P(y | do(x))$. According to the property of cross-entropy, $H(X | \theta_1)$ reaches its minimum so that $I(X; \theta_j)$ reaches its maximum as $P(x | \theta_1) = P(x | y_1)$, i.e.,

$$
\begin{aligned}
P(x_0|\theta_1) &= \frac{P(x_0)b_1'}{P(x_1) + P(x_0)b_1'} = P(x_0|y_1), \\
P(x_1|\theta_1) &= \frac{P(x_1)}{P(x_1) + P(x_0)b_1'} = P(x_1|y_1).
\end{aligned}
\tag{24}
$$

From the above two equations, we obtain:

$$
\frac{P(x_0)}{P(x_1)}b_1' = \frac{P(x_0|y_1)}{P(x_1|y_1)}.
\tag{25}
$$

Order:

$$
m(x_i, y_j) = \frac{P(y_j|x_i)}{P(y_j)} = \frac{P(x_i, y_j)}{P(x_i)P(y_j)}, i = 0,1; j = 0,1,
\tag{26}
$$

which represents the degree of correlation between $x_i$ and $y_j$ and may be independent of $P(x)$ and $P(y)$, unlike $P(x_i, y_j)$. From Equations (25) and (26), we obtain the optimized degree of disbelief, i.e., the degree of disconfirmation:

$$
b_1'^* = m(x_0, y_1) / m(x_1, y_1).
\tag{27}
$$

Then we have the degree of confirmation of $s_1$:

$$
b_1^* = 1 - b_1'^* = 1 - \frac{m(x_0, y_1)}{m(x_1, y_1)} = \frac{m(x_1, y_1) - m(x_0, y_1)}{m(x_1, y_1)}.
\tag{28}
$$

In the above formulas, we assume $b_1^* > 0$ and hence $m(x_1, y_1) \geq m(x_0, y_1)$. If $m(x_1, y_1) < m(x_0, y_1)$, $b_1^*$ should be negative, and $b_1'^*$ should be $m(x_1, y_1) / m(x_0, y_0)$. Then we have:

$$
b_1^* = -(1 - b_1'^*) = -(1 - \frac{m(x_1, y_1)}{m(x_0, y_1)}) = \frac{m(x_1, y_1) - m(x_0, y_1)}{m(x_0, y_1)}.
\tag{29}
$$

Combining the above two equations, we derive the confirmation measure:

$$
Cc(x_1 => y_1) = b_1^* = \frac{m(x_1, y_1) - m(x_0, y_1)}{\max(m(x_1, y_1), m(x_0, y_1))}.
\tag{30}
$$

Since $P(y_j|x_i) = m(x_i, y_j)P(y_j)$, we also have:

$$
b_1'^* = P(x_0 | y_1) / P(x_1 | y_1),
\tag{31}
$$

$$
Cc(x_1 => y_1) = b_1^* = \frac{P(y_1|x_1) - P(y_1|x_0)}{\max(P(y_1|x_1), P(y_1|x_0))} = \frac{R - 1}{\max(R, 1)},
\tag{32}
$$

where $R = P(y_1 | x_1) / P(y_1 | x_0)$ is the relative risk or the likelihood ratio used for $P_d$.

Measure $Cc$ has the normalizing property since its maximum is 1 as $m(x_0, y_1) = 0$ and the minimum is $-1$ as $m(x_1, y_1) = 0$. It has cause symmetry since:

$$
\begin{aligned}
Cc(x_0/x_1 => y_1) &= \frac{m(x_0, y_1) - m(x_1, y_1)}{\max(m(x_0, y_1), m(x_1, y_1))} \\
&= -\frac{m(x_1, y_1) - m(x_0, y_1)}{\max(m(x_1, y_1), m(x_0, |y_1))} = -Cc(x_1/x_0 => y_1).
\end{aligned}
\tag{33}
$$

Similarly, letting probability distribution $P(y | x_1)$ be the linear combination of a uniform probability distribution and a 0–1 distribution, we can obtain another causal confirmation measure:

$$
Ce(x_1 => y_1) = \frac{P(y_1|x_1) - P(y_0|x_1)}{\max(P(y_1|x_1), P(y_0|x_1))} = \frac{2P(y_1|x_1) - 1}{\max(P(y_1|x_1), 1 - P(y_1|x_1))}.
\tag{34}
$$

This measure can be regarded as the direct extension of Bayesian confirmation measure $c^*(e_1 \rightarrow h_1)$ [8]. It increases monotonically with the Bayesian confirmation measure $f(h_1, e_1) = P(h_1 | e_1)$, which is used by Fitelson et al. [5,32]. However, $Ce$ has the normalizing property and the outcome symmetry:

$$Ce(x_1 => y_1) = - Ce(x_1 => y_0). \tag{35}$$

*3.4. Causal Confirmation Measures Cc and Ce for Probability Predictions*

From $y_1$, $b_1^*$, and $P(x)$, we can make the probability prediction about $x$:

$$P(x_1|\theta_1) = \frac{P(x_i)}{P(x_1) + b_1'^* P(x_0)}, \ P(x_0|\theta_1) = \frac{P(x_0)b_1'^*}{P(x_1) + b_1'^* P(x_0)}, \tag{36}$$

where $b_1^* > 0$, $\theta_1$ represents $y_1$ with $b_1'^*$, and $\theta_0$ means $y_0$ with $b_0'^*$. If $b_1^* < 0$, we let $T(s_1 | x_1) = b_1'$ and $T(s_1 | x_0) = 1$, and then use the above formula.

Following the probability prediction with Bayesian confirmation measure $c^*$ [8], we can also make probability prediction for given $x_1$ and $Ce_1$. For example, when $Ce_1$ is greater than 0, there is:

$$P(y_1|\theta_{x1}) = 1/(2 - Ce_1), \tag{37}$$

where $\theta_{x1}$ denotes $x_1$ and $Ce_1$.

Given the semantic channel ascertained by $b_1 > 0$ and $b_0 > 0$, as shown in Table 2, we can obtain the corresponding Shannon channel $P(y | x)$. According to Equation (32), we can deduce:

$$P(y_1|x_1) = \frac{1-b_0'}{1-b_1'b_0'}, \ P(y_0|x_0) = \frac{1-b_1'}{1-b_1'b_0'}, \\ P(y_0|x_1) = 1 - P(y_1|x_1), \ P(y_1|x_0) = 1 - P(y_0|x_0). \tag{38}$$

## 4. Results

### 4.1. A Real Example of Kidney Stone Treatments

Table 3 shows Example 2 with detailed data about kidney stone treatments [15]. The data were initially provided in [16]. In Table 3, *% means a success rate, and the number behind it is the number of patients. The stone size is a confounder. The conclusion from every group (with small or large stones) is that treatment $x_2$ (i.e., treatment A in [15]) is better than treatment $x_1$ (i.e., treatment B in [15]); whereas the conclusion according to average success rates, $P(y_1 | x_2) = 0.78$ and $P(y_1 | x_1) = 0.83$, treatment $x_1$ is better than treatment $x_2$. There seems to be a paradox.

**Table 3.** Comparing two treatments' success rates ($y_1$ means the success).

| | Treat. $x_1$ | Treat. $x_2$ | Number | P(g) or Cc |
|---|---|---|---|---|
| Small stones ($g_1$) | 87%/270 | 93%/87 * | 357 | 0.51 |
| Large stones ($g_2$) | 69%/80 | 73%/263 | 343 | 0.49 |
| Overall | 83%/350 | 78%/350 | 700 | |
| $P(y_1 | x)$ | 0.83 | 0.78 | | $[(P(y_1 | x_2) - P(y_1 | x_1)]/P(y_1 | x_2) = -0.064$ |
| $P(y_1 | \text{do}(x))$ | 0.78 | 0.83 | | $Cc_1 = Cc(x_2/x_1 => y_1) = 0.06$ |
| $P(y_0 | \text{do}(x))$ | 0.22 | 0.17 | | $Cc_0 = Cc(x_1/x_2 => y_0) = 0.23$ |

* "87%/270" means that the success rate is 87%, and the number in this subgroup is 270.

We used $P(g)$ instead of $P(g | x_1)$ or $P(g | x_2)$ as the weighting coefficient for $P(y_1 | \text{do}(x_1))$ and $P(y_1 | \text{do}(x_2))$. After replacing $P(y_1 | x_1)$ with $P(y_1 | \text{do}(x_1))$ and $P(y_1 | x_2)$ with $P(y_1 | \text{do}(x_2))$, we derived $Cc_1 = Cc(x_2/x_1 => y_1) = 0.06$ (see Table 3), which means that the overall conclusion is that treatment $x_2$ is better than treatment $x_1$.

For $Cc_1$ in Table 3, we used treatment $x_1$ as the default; the degree of causal confirmation $Cc_1 = Cc(x_2/x_1 => y_1)$ is 0.06. If we used $x_2$ as the default, $Cc_1 = Cc(x_1/x_2 => y_1) = -0.06$.

Using measure $Cc$, we need not worry about which of $P(y_1 \mid do(x_1))$ and $P(y_1 \mid do(x_2))$ is larger, whereas, using $P_d$, we have to consider that before calculating $P_d$.

We used the incremental school's confirmation measure $D(x_1, y_1)$ to compare $x_1$ and $x_2$. We obtained:

- $P(y_1) = P(x_1)P(y_1 \mid x_1) + P(x_2)P(y_1 \mid x_2) = 0.805$,
- $P(y_1 \mid x_2, g_1) - P(y_1) = 0.93 - 0.805 > 0$,
- $P(y_1 \mid x_2, g_2) - P(y_1) = 0.73 - 0.805 < 0$, and
- $D(x_1, y_1) = P(y_1 \mid x_1) - P(y_1) = 0.83 - 0.805 > 0$
- $D(x_2, y_2) = P(y_1 \mid x_2) - P(y_1) = 0.78 - 0.805 < 0$.

The results mean that $x_1$ is better than $x_2$. There seems to be no paradox only because the paradox is avoided rather than eliminated when we use $D(x_1, y_1)$.

We tested Equation (38) by the aforementioned example. The Shannon channel $P(y \mid x)$ derived from the two degrees of disconfirmation $b_1'^*$ and $b_0'^*$ is the same as $P(y \mid do(x))$ shown in the last two rows of Table 3.

### 4.2. An Example of Eliminating Simpson's Paradox with COVID-19

Table 4 shows Example 2 with detailed data about the CFRs of COVID-19. The original data were obtained from the website of the Centers for Disease Control and Prevention (CDC) in the United States up until 2 July 2022 [33]. The data only include reported cases; otherwise, the CFRs should be lower. The $x_1$ represents the non-Hispanic white and $x_2$ means the other races. $P(y_1 \mid x_1, g)$ and $P(y_1 \mid x_2, g)$ are the CFRs of $x_1$ and $x_2$ in an age group $g$. See Appendix A for the original data and median results.

**Table 4.** The CFRs of COVID-19 of non-Hispanic white ($x_1$) and other people ($x_2$) from different age groups.

| Age Group ($g$) | $P(x_1 \mid g)$ | $P(g)$ | $P(y_1 \mid x_1, g)$ | $P(g \mid x_1)$ | $P(y_1 \mid x_2, g)$ | $P(g \mid x_2)$ |
|---|---|---|---|---|---|---|
| 0–4 Years | 44.200 | 0.041 | 0.0002 | 0.0349 | 0.0002 | 0.0480 |
| 5–11 Years | 44.200 | 0.078 | 0.0001 | 0.0659 | 0.0001 | 0.0907 |
| 12–15 Years | 46.300 | 0.052 | 0.0001 | 0.0458 | 0.0001 | 0.0578 |
| 16–17 Years | 48.700 | 0.029 | 0.0001 | 0.0268 | 0.0002 | 0.0307 |
| 18–29 Years | 48.700 | 0.223 | 0.0004 | 0.2081 | 0.0006 | 0.2388 |
| 30–39 Years | 49.300 | 0.178 | 0.0011 | 0.1681 | 0.0019 | 0.1883 |
| 40–49 Years | 51.000 | 0.146 | 0.0030 | 0.1427 | 0.0048 | 0.1493 |
| 50–64 Years | 59.100 | 0.163 | 0.0102 | 0.1843 | 0.0144 | 0.1389 |
| 65–74 Years | 67.300 | 0.055 | 0.0333 | 0.0704 | 0.0457 | 0.0373 |
| 75–84 Years | 72.900 | 0.025 | 0.0762 | 0.0356 | 0.0938 | 0.0144 |
| 85+ Years | 76.300 | 0.012 | 0.1606 | 0.0173 | 0.1751 | 0.0059 |
| sum | | 1 | | 1 | | 1 |

Table 5 shows that the overall (average) CFRs vary before and after we change the weighting coefficient from $P(g \mid x)$ to $P(g)$.

**Table 5.** Comparing the CFRs of non-Hispanic whites and the other people.

| | The CFR of Non-Hispanic Whites ($x_1$) | The CFR of of Other People ($x_2$) | Risk Measure * |
|---|---|---|---|
| $P(y_1 \mid x)$ | 1.04 | 0.73 | $P_d = (R - 1)/R = 0.30$ |
| $P(y_1 \mid do(x))$ | 0.80 | 1.05 | $P_d = 0$; $Cc(x_1/x_2 => y_1) = -0.28$ |

* $R = P(y_1 \mid x_1)/P(y_1 \mid x_2)$.

From Table 4, we can find that for different age groups, the CFR of the non-Hispanic whites is lower than or close to that of the other races. However, for all age groups (see Table 5), the overall (average) CFR (1.04) of the non-Hispanic whites is higher than the CFR (0.73) of the other races. After replacing $P(g \mid x)$ with $P(g)$, the overall CFR (0.80) of the non-Hispanic whites is also lower than that (1.05) of the other races.

We followed Fitelson to use $D(x_1, y_1)$ to assess the risk. The average CFR is 0.97 (found on the same website [33]). We obtained:

$$D(\text{non-Hispanic whites, death}) = P(y_1 \mid x_1) - P(y_1) = 1.04 - 0.97 = 0.07,$$

$$D(\text{other people, death}) = P(y_1 \mid x_2) - P(y_1) = 0.73 - 0.97 = -0.14,$$

which means that non-Hispanic whites are at higher risk.

### 4.3. COVID-19: Vaccine's Negative Influences on the CFR and Mortality

Using causal probability measure $P_d$ is not convenient to measure the "probability" of "vaccine => infection" or "vaccine => death", since $P_d$ is regarded as the probability, whose minimum value is 0, while the vaccine's influence is negative. However, there is no problem using $Cc$ because $Cc$ can be negative.

Table 6 shows data obtained from the website of the US CDC [34] and the two degrees of causal confirmation. The numbers of cases and deaths are among 100,000 people (ages over 5) in a week (from June 20 to 26, 2022).

**Table 6.** The negative degrees of causal confirmation for accessing that the vaccine affects infections and deaths.

|  | Unvaccinated ($x_0$) | Vaccinated ($x_1$) | $Cc$ |
|---|---|---|---|
| Cases | 512.6 | 189.5 | $Cc(x_1/x_0 => y_1) = -0.63$ |
| Deaths | 1.89 | 0.34 | $Cc(x_1/x_0) => y_1) = -0.79$ |
| Mortality rate | 0.001 | 0.00018 | |

The negative degree of causal confirmation $-0.63$ means that the vaccine reduced the infection rate by 63%. The $-0.79$ means that the vaccine reduced the CFR by 79%.

To know the impact of COVID-19 on population mortality, we need to compare the regular mortality rate due to common reasons ($x_0$) with the new mortality rate due to common reasons plus COVID-19 ($x_1$) during the same period (such as one year). Since the average lifespan of people in the United States is 79 years old, the annual mortality rate is about $1/79 = 0.013$. From Table 6, we can derive that the yearly mortality rate caused by COVID-19 is 0.001 (for unvaccinated people) or 0.00018 (for vaccinated people).

People who died due to COVID-19 may also die in the same year for common reasons. Therefore, the new mortality rate should be less than the sum of the two mortality rates. Assume that $x_0$ and $x_1$ are independent of each other. Then new mortality rate $P(y_1 \mid x_1)$ should be $0.013 + 0.001 - 0.013 \times 0.001 \approx 0.014$ (for unvaccinated people) or $0.013 + 0.00018 - 0.013 \times 0.00018 \approx 0.01318$ (for vaccinated people). Table 7 shows the degree of causal confirmation of COVID-19 leading to mortality, for which we assume $P(y_1 \mid x) = P(y_1 \mid \mathrm{do}(x))$.

**Table 7.** Using $Cc$ to measure the impact of COVID-19 on the mortality rates.

|  | Mortality Rate $P(y_1 \mid x)$ | Unvaccinated | Vaccinated |
|---|---|---|---|
| $x_0$: common reasons | $P(y_1 \mid x_0)$ | 0.013 | 0.013 |
| $x_1$: $x_0$ plus COVID-19 | $P(y_1 \mid x_1)$ | 0.014 | 0.01318 |
| $Cc_1 = Cc(x_1/x_0 => y_1)$ |  | 0.07 | 0.014 |

In the last line, $Cc_1 = 0.07$ means that among unvaccinated people who die, 7% are due to COVID-19. Moreover, $Cc = 0.014$ means that among the vaccinated people who die, 1.4% are due to COVID-19.

If we used $x_1 = $ COVID-19 instead of $x_1 = x_0 + $ COVID-19, we would get a strange conclusion that COVID-19 could reduce deaths.

We obtained the above results without considering the vaccine's side effects, possibly resulting in chronic deaths.

## 5. Discussion

*5.1. Why Can $P_d$ and Cc Better Indicate the Strength of Causation Than D in Theory?*

We call $m(x_i, y_j)$ $(i = 0,1; j = 0,1)$ the probability correlation matrix, which is not symmetrical. Although there exists $P(x, y)$ first and then $m(x, y)$ from the perspective of calculation, there exists $m(x, y)$ first and then $P(x, y)$ from the perspective of existence. That is, given $P(x)$, $m(x, y)$ only allows specific $P(y)$ to happen.

We can also make probability predictions with $m(x, y)$ (like using Bayes' formula):

$$P(y|x_1) = P(y)m(x_1, y)/m(x_1), m(x_1) = \sum_y P(y)m(x_1, y),$$
$$P(x|y_1) = P(x)m(x, y_1)/m(y_1), m(y_1) = \sum_x P(x)m(x, y_1). \tag{39}$$

From Equations (27)–(30), we can find that $P_d$ and $Cc$ only depend on $m(x, y)$ and are independent of $P(x)$ and $P(y)$. The two degrees of disconfirmation, $b_1'^*$ and $b_0'^*$, ascertain a semantic channel and a Shannon channel. Therefore, the two degrees of causal confirmation, $Cc_1 = b_1^*$ and $Cc_0 = b_0^*$, indicate the strength of the constraint relationship (causality) from $x$ to $y$. Like $Cc$, measure $P_d$ is also only related to $m(x, y)$. $D$ and $\Delta^* P_x{}^y$ are different; they are related to $P(x)$, so they do not indicate the strength of causation well.

For example, considering the vaccine's effect on the CFR of COVID-19 (see Table 7), $P_d$ or $Cc$ are irrelated to vaccination coverage rate $P(x_1)$, whereas measure $\Delta^* P_x{}^y$ is related to $P(x_1)$. Measure $D$ is associated with $P(y)$ and is also related to $P(x_1)$. $P_d$ and $Cc_1$ obtained from one region also fits other areas for the same variant of COVID-19. In contrast, $\Delta^* P_x{}^y$ and $D$ are not universal because the vaccination coverage rate $P(x_1)$ differs in different areas.

According to the incremental school's view of Bayesian confirmation, $P(y_1)$ is a prior probability, and $P(y_1 | x) - P(y_1)$ is its increment. However, when measure $D$ is used for causal confirmation, $P(y_1)$ is obtained from $P(x)$ and $P(y_1 | x)$ after the treatment, so $P(y)$ is no longer a priori probability, which is also a fatal problem with the incremental school.

In addition, as the result of induction, $Cc$ and $P_d$ can indicate the degree of belief of a fuzzy major premise and can be used for probability predictions, whereas $D$ and $\Delta^* P_x{}^y$ cannot.

*5.2. Why Are $P_d$ and Cc Better than D in Practice?*

Two calculation examples in Sections 4.1 and 4.2 support the conclusion that measures $P_d$ and $Cc$ are better than $D$ in practice. The reasons are as follows.

### 5.2.1. $P_d$ and Cc Have Precise Meanings in Comparison with D

$Cc_1 = Cc (x_1/x_0 => y_1)$ indicates what percentage of the result $y$ is due to $x_1$ instead of $x_0$. For example, Table 6 shows that according to the virulence of the virus, COVID-19 will increase the mortality rate of vaccinated people from 1.3% to 1.318%. Therefore, the degree of causal confirmation is $Cc_1 = P_d = 0.014$, which means that 1.4% of the deaths will be due to COVID-19. However, the meanings of $D$ and $\Delta^* P_x{}^y$ are not precise.

Different from measure $RD$ (see Equation (1)), $P_d$ and $Cc$ indicate relative risk or the relative change of the outcome. Many people think COVID-19 is very dangerous because it can kill millions in a country. However, the mortality rate it brings is much lower than that caused by common reasons. $P_d$ and $Cc$ can reveal the relative change in the mortality rate (see Table 7). Although it is essential to reduce or delay deaths, it is also vital to decrease the economic loss due to the fierce fight against the pandemic. Therefore, $P_d$ and

*Cc* can help decision-makers balance between reducing or delaying deaths and reducing financial losses.

### 5.2.2. The Confounder's Effect Is Removed from $P_d$ and *Cc*

When there is a confounder, as shown in Section 4.1, using $P_d$ or *Cc*, we can eliminate Simpson's Paradox and make the overall conclusion consistent with the grouping conclusion: treatment $x_2$ is better than treatment $x_1$. For example, if we use $D$ to compare the success rates of two treatments, although we can avoid Simpson's Paradox, the conclusion is unreasonable. The reason is that we neglect the difficulties of treatments for different sizes of kidney stones. If a hospital only accepts patients who are easy to treat, its overall success rate must be high; however, such a hospital may not be a good one.

### 5.2.3. $P_d$ and *Cc* Allow Us to View the Third Factor, *u*, from Different Perspectives

For the example in Section 4.2, if we think that one's longevity is related to one's race, we can take the lifespan as a mediator and then directly accept the overall conclusion (non-Hispanic whites have a higher CFR than other people). On the other hand, if we believe that one's longevity is not due to one's race, then the lifespan is a confounder. Therefore, we can make the overall conclusion consistent with the grouping conclusion, and then use $P_d$ and *Cc*.

It is worth noting that it is concluded that the CFR of non-Hispanic whites is lower than that of other people, probably because medical conditions affect the CFRs. However, existing data do not contain information about the medical conditions of different races. Otherwise, the CFRs of different races might be similar if we use the medical condition as a confounder. This issue is worth researching further.

### *5.3. Why Is It Better to Replace $P_d$ with Cc?*

Section 4.3 provides the calculation of the two negative degrees of causal confirmation that reflect the impacts of the vaccine on infection and mortality. The negative degrees of confirmation mean that the vaccine can reduce illnesses and deaths. However, if we use $P_d$ as the probability of causation, $P_d$ can only take its lower limit 0. Although we can replace $P_d$(vaccinated => death) with $P_d$(unvaccinated => death) to ensure $P_d > 0$, it does not conform to our thinking habits to take being vaccinated as the default cause. In addition, *Cc* has cause symmetry, whereas $P_d$ does not.

When we used $P_d$ to compare two causes $x_1$ and $x_2$, such as two treatments for kidney stones (see Section 4.1), we had to consider which of $P(y_1 | x_2)$ and $P(y_1 | x_1)$ was larger. However, using *Cc*, we needed to not consider that because it is unnecessary to worry about if $(R - 1)/R < 0$.

The correlation coefficient in mathematics is between 1 and −1. *Cc* can be understood as the probability correlation coefficient. The difference is that the former has only one coefficient between *x* and *y*, whereas the latter has two coefficients: $Cc_1 = Cc(x_1 => y_1)$ and $Cc_0 = Cc(x_0 => y_0)$.

### *5.4. Necessity and Sufficiency in Causality*

Measures $P_d$ and *Cc* only indicate the necessity of cause *x* to outcome *y*; they do not reflect the sufficiency of *x* or the inevitability of *y*. On the other hand, measure $f = P(y | x)$ and *Ce* can indicate the outcome's inevitability.

The medical industry uses the odds ratio to indicate both the necessity and sufficiency of the cause to the outcome. The odds rate [2] is:

$$OR = \frac{P(y_1|x_1)}{P(y_1|x_0)} \times \frac{P(y_0|x_0)}{P(y_0|x_1)}. \tag{40}$$

It is the product of two likelihood ratios. We can use:

$$OR_N = \frac{OR - 1}{\max(OR, 1)} \tag{41}$$

as the confirmation measure of both $x_0 => y_0$ and $x_1 => y_1$ for the same purpose. However, $OR_N$ has the normalizing property and symmetry.

*5.5. The Relationship between Bayesian Confirmation Measures b\* and c\*, and Causal Confirmation Measures Cc and Ce*

Suppose that $P(y_1 \mid x)$ has been modified for $P(y_1 \mid x) = P(y_1 \mid do(x))$. Causal confirmation measure $Cc$ is equal to channels' confirmation measure $b^*$ [8] in value, i.e.,

$$Cc(x_1 => y_1) = [P(y_1 \mid x_1) - P(y_1 \mid x_0)]/\max(P(y_1 \mid x_1), P(y_1 \mid x_0)) = b^*(y_1 \rightarrow x_1). \tag{42}$$

However, their antecedents and consequents are inverted, which means that if $x_1$ is the cause of $y_1$, then $y_1$ is the evidence of $x_1$. For example, if COVID-19 infection is the cause of the test-positive, then the test-positive is the evidence of the infection.

Causal confirmation measure $Ce$ indicating the inevitability of the outcome is equal to prediction confirmation measure $c^*(x_1 \rightarrow y_1)$ in value, i.e.,

$$Ce(x_1 => y_1) = [P(y_1 \mid x_1) - P(y_0 \mid x_1)]/\max(P(y_1 \mid x_1), P(y_0 \mid x_1)) = c^*(x_1 \rightarrow y_1). \tag{43}$$

Their antecedents and consequents are the same.

However, from the right sides' values of the above two equations, we may not be able to obtain the left sides' values because an associated relationship may not be a causal relationship.

## 6. Conclusions

Fitelson, a representative of the incremental school of Bayesian confirmation, used $D(x_1, y_1) = P(y_1 \mid x_1) - P(y_1)$ to denote the supporting strength of the evidence to the consequence and extended this measure for causal confirmation without considering the confounder. This paper has shown that measure $D$ is incompatible with the ECIT and popular risk measures, such as $P_d = \max(0, (R - 1)/R)$. Using $D$, one can only avoid Simpson's Paradox but he cannot eliminate it or provide a reasonable explanation as the ECIT does.

On the other hand, Rubin et al. used $P_d$ as the probability of causation. $P_d$ is better than $D$, but it is improper to call $P_d$ a probability measure and use the probability measure to measure causation. If we use $P_d$ as a causal confirmation measure, it lacks the normalizing property and symmetry that an ideal confirmation measure should have.

This paper has deduced causal confirmation measure $Cc(x_1 => y_1) = (R - 1) / \max(R, 1)$ by the semantic information method with the minimum cross-entropy criterion. $Cc$ is similar to the inductive school's confirmation measure $b^*$ proposed by the author earlier. However, the positive examples' proportion $P(y_1 \mid x_1)$ and the counterexamples' proportion $P(y_1 \mid x_0)$ are replaced with $P(y_1 \mid do(x_1))$ and $P(y_1 \mid do(x_0))$ so that $Cc$ is an improved $P_d$. Compared with $P_d$, $Cc$ has the normalizing property (it changes between –1 and 1) and the cause symmetry ($Cc(x_0/x_1 => y_1) = -Cc (x_1/x_0 => y_1)$). Since $Cc$ may be negative, it is also suitable for evaluating the inhibition relationship between cause and outcome, such as between vaccine and infection.

This paper has provided some examples with Simpson's Paradox for calculating the degrees of causal confirmation. The calculation results show that $P_d$ and $Cc$ are more reasonable and meaningful than $D$, and $Cc$ is better than $P_d$ mainly because $Cc$ may be less than zero. In addition, this paper has also provided a causal confirmation measure $Ce(x_1 => y_1)$ that indicates the inevitability of the outcome $y_1$.

Since measure $Cc$ and the ECIT support each other, the inductive school of Bayesian confirmation are also supported by the ECIT and the epidemical risk theory.

However, like all Bayesian confirmation measures, causal confirmation measure *Cc* and *Ce* also use size-limited samples, hence, the degrees of causal confirmation are not strictly reliable. Therefore, replacing a degree of causal confirmation with a degree interval is necessary to retain the inevitable uncertainty. This work needs further studies by combining existing theories.

## Appendix A. Data and Calculations for Comparing the CFRs of Non-Hispanic Whites and Other People in the USA

The original data were obtained from the USA CDC (Centers for Disease Control and Prevention) website [33]. The Excel file with the original data and computing results can be downloaded from http://survivor99.com/lcg/cm/CFR.zip (accessed on 8 December 2022).

## References

1. Rubin, D. Causal inference using potential outcomes. *J. Amer. Statist. Assoc.* **2005**, *100*, 322–331. [CrossRef]
2. Hernán, M.A.; Robins, J.M. *Causal Inference: What If*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2020.
3. Pearl, J. Causal inference in statistics: An overview. *Stat. Surv.* **2009**, *3*, 96–146. [CrossRef]
4. Geffner, H.; Rina Dechter, R.; Halpern, J.Y. (Eds.) *Probabilistic and Causal Inference: The Works of Judea Pearl, Association for Computing Machinery*; Association for Computing Machinery: New York, NY, USA, 2021.
5. Fitelson, B. Confirmation, Causation, and Simpson's Paradox. *Episteme* **2017**, *14*, 297–309. [CrossRef]
6. Carnap, R. *Logical Foundations of Probability*, 2nd ed.; University of Chicago Press: Chicago, IL, USA, 1962.
7. Kemeny, J.; Oppenheim, P. Degrees of factual support. *Philos. Sci.* **1952**, *19*, 307–324. [CrossRef]
8. Lu, C. Channels' Confirmation and Predictions' Confirmation: From the Medical Test to the Raven Paradox. *Entropy* **2020**, *22*, 384. [CrossRef]
9. Greco, S.; Slowiński, R.; Szczęch, I. Properties of rule interestingness measures and alternative approaches to normalization of measures. *Inf. Sci.* **2012**, *216*, 1–16. [CrossRef]
10. Crupi, V.; Tentori, K.; Gonzalez, M. On Bayesian measures of evidential support: Theoretical and empirical issues. *Philos. Sci.* **2007**, *74*, 229–252. [CrossRef]
11. Eells, E.; Fitelson, B. Symmetries and asymmetries in evidential support. *Philos. Stud.* **2002**, *107*, 129–142. [CrossRef]
12. Relative Risk, Wikipedia the Free Encyclopedia. Available online: https://en.wikipedia.org/wiki/Relative_risk (accessed on 15 August 2022).
13. Robins, J.; Greenland, S. The probability of causation under a stochastic model for individual risk. *Biometrics* **1989**, *45*, 1125–1138. [CrossRef]
14. Simpson, E.H. The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Ser. B* **1951**, *13*, 238–241. [CrossRef]
15. Simpson's Paradox. Wikipedia the Free Encyclopedia. Available online: https://en.wikipedia.org/wiki/Simpson%27s_paradox (accessed on 20 August 2022).
16. Charig, C.R.; Webb, D.R.; Payne, S.R.; Wickham, J.E. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br. Med. J. (Clin. Res. Ed.)* **1986**, *292*, 879–882. [CrossRef]
17. Julious, S.A.; Mullee, M.A. Confounding and Simpson's paradox. *BMJ* **1994**, *309*, 1480–1481. [CrossRef]
18. Pedagogy, W. Simpson's Paradox. Available online: https://weapedagogy.wordpress.com/2020/01/15/5-simpsons-paradox/ (accessed on 21 August 2022).
19. Mackenzie, D. Race, COVID Mortality, and Simpson's Paradox. Available online: http://causality.cs.ucla.edu/blog/index.php/category/simpsons-paradox/ (accessed on 22 August 2022).
20. Kügelgen, J.V.; Gresele, L.; Schölkopf, B. Simpson's Paradox in COVID-19 case fatality rates: A mediation analysis of age-related causal effects. *IEEE Trans. Artif. Intell.* **2021**, *2*, 18–27. [CrossRef]
21. Mortimer, H. *The Logic of Induction*; Prentice Hall: Paramus, NJ, USA, 1988.
22. Horwich, P. *Probability and Evidence*; Cambridge University Press: Cambridge, UK, 1982.

23. Christensen, D. Measuring confirmation. *J. Philos.* **1999**, *96*, 437–461. [CrossRef]
24. Nozick, R. *Philosophical Explanations*; Clarendon: Oxford, UK, 1981.
25. Good, I.J. The best explicatum for weight of evidence. *J. Stat. Comput. Simul.* **1984**, *19*, 294–299. [CrossRef]
26. Lu, C. A generalization of Shannon's information theory. *Int. J. Gen. Syst.* **1999**, *28*, 453–490. [CrossRef]
27. Lu, C. Semantic Information G Theory and Logical Bayesian Inference for Machine Learning. *Information* **2019**, *10*, 261. [CrossRef]
28. Lu, C. The P–T Probability Framework for Semantic Communication, Falsification, Confirmation, and Bayesian Reasoning. *Philosophies* **2020**, *5*, 25. [CrossRef]
29. Zadeh, L.A. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
30. Zadeh, L.A. Probability measures of fuzzy events. *J. Math. Anal. Appl.* **1986**, *23*, 421–427. [CrossRef]
31. Rooij, R.V.; Schulz, K. Conditionals, causality and conditional probability. *J. Log. Lang. Inf.* **2019**, *28*, 55–71. [CrossRef]
32. Over, D.E.; Hadjichristidis, C.; Jonathan St, B.T.; Evans, J.S.B.T.; Handley, D.J.; Sloman, S.A. The probability of causal conditionals. *Cogn. Psychol.* **2007**, *54*, 62–97. [CrossRef] [PubMed]
33. Demographic Trends of COVID-19 Cases and Deaths in the US Reported to CDC. The Website of the US CDC. Available online: https://covid.cdc.gov/covid-data-tracker/#demographics (accessed on 10 September 2022).
34. Rates of COVID-19 Cases and Deaths by Vaccination Status The Website of US CDC. Available online: https://covid.cdc.gov/covid-data-tracker/#rates-by-vaccine-status (accessed on 8 September 2022).