Scientific Research Publishing

Open Journal of Epidemiology, 2016, 6, 198-232
http://www.scirp.org/journal/ojepi
ISSN Online: 2165-7467
ISSN Print: 2165-7459

Deciding on a Measure of Effect under Indeterminism

Doron J. Shahar

Department of Mathematics, University of Arizona, Tucson, USA
Email: dshahar@math.arizona.edu

How to cite this paper: Shahar, D.J. (2016) Deciding on a Measure of Effect under Indeterminism. *Open Journal of Epidemiology*, **6**, 198-232.
http://dx.doi.org/10.4236/ojepi.2016.64022

Received: August 13, 2016
Accepted: November 5, 2016
Published: November 9, 2016

Abstract

Estimating causal effects is a principal goal in epidemiology and other branches of science. Nonetheless, what constitutes an effect and which measure of effect is preferred are unsettled questions. I argue that, under indeterminism, an effect is a change in the tendency of the outcome variable to take each of its values, and then present a critical analysis of commonly used measures of effect and the measures of frequency from which they are calculated. I conclude that all causal effects should be quantified using a unifying measure of effect called the log likelihood ratio (which is the log probability ratio when the outcome is a discrete variable). Furthermore, I suggest that effects should be estimated for all causal contrasts of the causal variable (*i.e.*, exposure), on all values of the outcome variable, and for all time intervals between the cause and the outcome. This goal should be kept in mind in practical approximations.

Keywords

Measure of Effect, Measure of Frequency, Indeterminism, Causation, Causal Diagram

1. Introduction

A great many disagreements in science originate in the debate between determinism and indeterminism, and that is also the case when considering measures of effect. I shall begin the article with a preliminary section (Section 2) that provides a quick explanation of determinism and indeterminism, but rather than revisiting the debate, I will develop this article from an indeterministic viewpoint and leave a critique of determinism for another day.

Section 3 begins with a trivial question: What is a measure of effect? The answer: A measure of effect is a way to quantify a change in tendency. Much of Section 3 specifies

more precisely which tendencies and which changes in them are of interest. Section 4 contains a thorough discussion of measures of frequency and their use in quantifying tendency. The discussion culminates in one satisfactory measure capable of generically quantifying the tendency of interest.

Section 5 focuses on measures of effect, as derived from measures of frequency. Two common measures will be considered: The ratio and the difference. One key argument decides between the two, which is then applied to all measures of effect. That argument results in many closely related measures, all equally capable of quantifying changes in tendency. Among them one measure has nicer mathematical properties, which lead me to consider it the ideal measure of effect. Lastly, in Section 6, I consider a few complementary points that are not stressed earlier in the article, including philosophical interpretations of probability as they relate to quantifying tendency.

2. Preliminaries

There are several schools of thought regarding how scientific knowledge should be advanced, at the heart of each is a set of axioms, not always clearly articulated. For now, let's consider three major schools of thought: determinism, indeterminism, and individualized-indeterminism. These three viewpoints diverge at the very essence of causation: How do causes influence their effects? Specifically, what happens to an effect (variable G) once all of its causes (variables A, B, and C) take values? (Figure 1).

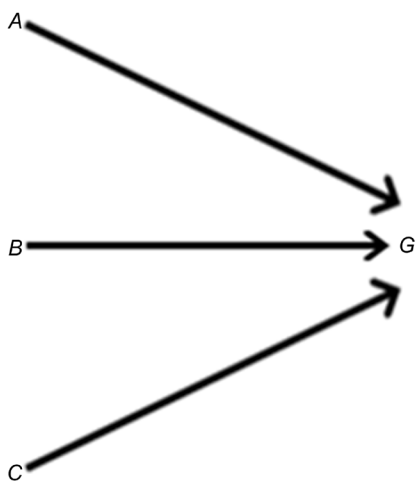

Figure 1. An effect G and all of its causes: A, B, and C.

Let A, B, and C represent the amount of vitamins A, B, and C in the blood, respectively, and let G be blood glucose level. Suppose Mr. White, Ms. Black, Mrs. Brown, and Dr. Green all have identical amounts of those vitamins: $A = a$, $B = b$, and $C = c$.

- According to determinism, once G realizes, it must take the same value, say g_0, for any of these people. Thus, Mr. White will have $G = g_0$, and Ms. Black will have $G = g_0$, as will Mrs. Brown and Dr. Green. In fact, anyone who has $A = a$, $B = b$, and $C = c$ must have G take the value g_0.

- According to indeterminism, everyone who has $A = a$, $B = b$, and $C = c$ shares the same tendency of having $G = g$ for any value g. (I will use the notation $T(G = g)$ to mean the tendency of having $G = g$). Thus, for $G = g_0$, $T(G = g_0 \mid \text{Being Mr. White}) = T(G = g_0 \mid \text{Being Ms. Black})$, and for $G = g_1$, $T(G = g_1 \mid \text{Being Mrs. Brown}) = T(G = g_1 \mid \text{Being Dr. Green})$. Furthermore, it is standard to assume that $T(G = g) \neq 0$ for any g. That is, "anything" can happen.
- Individualized-indeterminism, contrary to the other two viewpoints, does not accept a universal rule of causation. Rather, everyone has his or her own tendency of having $G = g$. That is, $T(G = g_0 \mid \text{Being Mr. White})$ need not equal $T(G = g_0 \mid \text{Being Ms. Black})$. Under an individualized model, effects are person-specific.

3. The Matter at Hand

3.1. What Is a Measure of Effect?

Under indeterminism, there are only two types of measures of effect: Those that deal with tendencies and those that deal with averages of such tendencies. Since indeterminism asserts that causation was built with underlying tendencies, any measure of effect that deals with tendencies describes the building blocks of causation. Other measures, such as arithmetic means, do not quantify the underlying tendencies. They are functions of many such tendencies, which may not even exist, as in the Cauchy distribution.

Even worse, averages can indicate null effects when clearly something has changed. To illustrate the point, let's join Mrs. Brown on her visit to Dr. Green, who recently conducted a study of the effect of A on G, while assuming null effects of B and C (Figure 2). Dr. Green estimated that when $A = a_0$, G has a bimodal distribution with a mean of 85 mg/dL that peaks at 10 mg/dL and at 300 mg/dL. He also estimated that when $A = a_1$ ($a_1 > a_0$), the probability distribution of G is approximately Gaussian with a mean of 85 mg/dL.

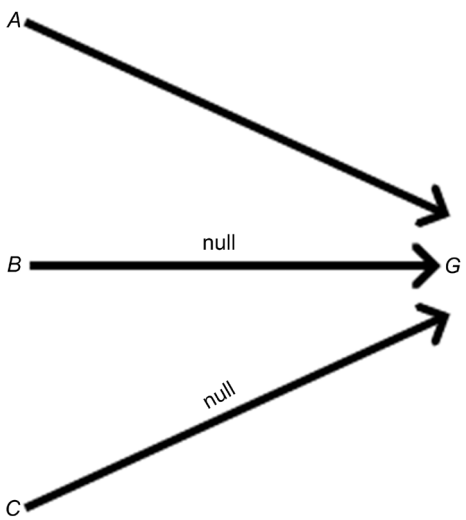

Figure 2. Dr. Green's theories displayed in a causal diagram.

"Mrs. Brown, your current value of A is a_0," starts Dr. Green. "If you were contemplating increasing it to a_1, know the effect I found in my study was null on a mean difference scale. As such, I see no reason for you to add more vitamin A to your diet."

"Null on a mean difference scale?" wonders Mrs. Brown. "I've heard that outside of our hypothetical world some guy is writing a story about why the mean difference is not a good measure of effect. Would you mind telling me the effect on another scale?"

"On other scales the effect is very large, and according to them you should increase your vitamin A level from a_0 to a_1. But don't worry. The mean difference scale is just fine."

"Great, I think I'm going to go eat a giant carrot."

Let's leave Mrs. Brown to her lunch and jump into the framework of measures of effect that deal with tendencies. Under this framework, an effect is a change in tendency, and a measure of effect quantifies that change.

3.2. Why Bother with Effects?

Referring back to Figure 1, the effect of A on G is the change between $T\left(G=g \mid A=a_0, B=b, C=c\right)$ and $T\left(G=g \mid A=a_1, B=b, C=c\right)$ for some values b and c. But indeterminism assumes that fixing all the causes of G fixes its tendency, so why not just estimate $T\left(G=g \mid A=a, B=b, C=c\right)$ for all possible a, b, and c? Why bother estimating an effect, a change in tendency, in the first place? The answer is simple: Whenever we estimate or use $T\left(G=g \mid A=a, B=b, C=c\right)$, we find ourselves far more interested in the change in tendency.

Suppose we hold a theory that A is the only variable among A, B, and C, that has a non-null effect on G. Thus, we proceed to estimate $T\left(G=g \mid A=a\right)$. We finish our multimillion dollar study and arrive at an estimate for that tendency just as Ms. Black receives her doctorate. Upon seeing our study, the now Dr. Black proposes the theory that B's effect on G is non-null and rather large. We proceed to test the theory and find support for it. And what are we supposed to do with the estimate from our multimillion dollar study? We may discard it and proceed to estimate $T\left(G=g \mid A=a, B=b\right)$ in another multimillion dollar study that would go to waste as soon as we learn that C's effect on G is also non-null. This approach to science is black and white.

I would prefer a grey approach. One in which we can salvage $T\left(G=g \mid A=a\right)$ and $T\left(G=g \mid A=a, B=b\right)$ to some extent by claiming that these estimates add bias in return for reduced variance or effort. For example, when estimating $T\left(G=g \mid A=a, B=b\right)$, the magnitude of the bias depends on how greatly the tendency varies with C. If $T\left(G=g \mid A=a, B=b, C=c_0\right) \approx T\left(G=g \mid A=a, B=b, C=c_1\right)$ the bias is small; the more they differ the larger the bias. That is, the magnitude of the bias depends on the effect of C on G. And in all cases, the bias depends on the magnitude of one effect or another. Therefore, if the goal of science were to estimate tendencies, we must constantly worry about bias due to effects (*i.e.*, changes in tendencies).

Moreover, suppose we are supplied with the value of $T\left(G=g \mid A=a, B=b, C=c\right)$ for all a, b, and c—the ultimate knowledge. What are we to do with this information? If

we want G to take the value g, we should find the values of A, B, and C that maximize $T(G = g \mid A = a, B = b, C = c)$. But what if it's difficult to change the value of B? Should we bother trying to change it? Maybe, if B changes the tendency by a lot. But if B hardly changes the tendency, why bother? Translation: Decide whether the effect is large enough to matter, and proceed to ignore the tendency altogether. As such, both during and after a study, we are interested in the change in tendency, not in the tendency itself.

3.3. Tendency of What?

Before discussing how to quantify the change in tendency or even the tendency itself, we must answer a basic question: Tendency of what? As the example in Section 2 illustrates, the tendency is always of some variables to take some values. The tendency of a single variable taking one of its values will be called an individual tendency. If multiple variables (or multiple values) are being considered, the tendency will be referred to as a group tendency.

When the variables in question do not depend on one another, group tendencies are just a function of individual tendencies. As such, they suffer from the same drawbacks as averages. In this case, there is no point in considering group tendencies. When the variables do depend on each other, group tendencies are a function of the individual tendencies along with other tendency-like quantities. Since in most cases, we cannot know *a priori* whether variables depend on one another, we cannot guarantee that we are estimating a legitimate group tendency.

3.4. One Time Point at a Time

I have just argued that the tendency is of a *single* variable having a given value. The important thing to remember is that a variable exists at a time point: G_1 (G at time 1) is not the same variable as G_2 (G at time 2). Yet the tendency of having a value *within* a time interval is often calculated in research. For example, the tendency of having a stroke within the next ten years; that is, the tendency of at least one stroke status in the next ten years having the value "stroke". But there are an infinite number of stroke status variables during that time interval: stroke status one minute into the study, after a year, as well as after 2 years 6 months 6 days 50 minutes and 8 seconds. In the end, the only tendencies that matter are those of individual variables—variables at a time point.

Furthermore, tendencies over time intervals, like averages, can indicate null effects when clearly something has changed. Consider two food items, food A and food B, that lead to heartburn. Suppose people who eat food A are likely to start having heartburn five minutes after eating, whereas people who eat food B are equally likely to get heartburn, but only after three hours. Suppose further that in both cases, heartburn will last for about an hour (**Figure 3** in red). Therefore, the tendency of having heartburn *within* five hours of eating is the same for those eating food A and those eating food B (a null effect of A vs. B). For anyone about to give a lecture, however, it will be preferable to choose food B over food A, knowledge that is found in time point effects.

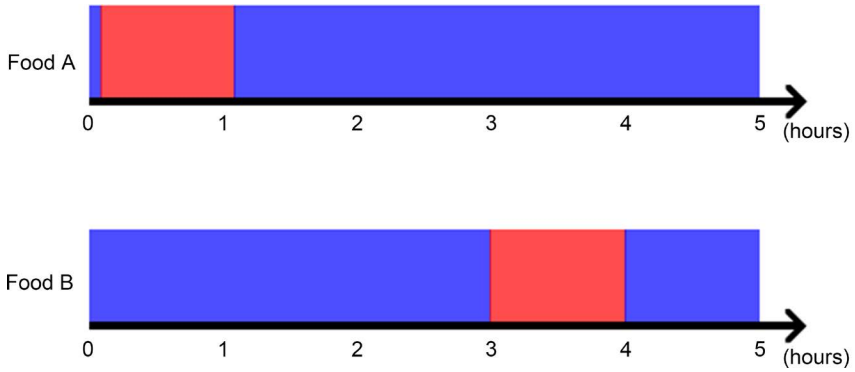

Figure 3. Heartburn over time.

Lastly, recall that the tendency is fixed only when all of the causes take specified values. Furthermore, all of the causes must be concurrent (*i.e.*, at the same time point). Otherwise, we are asking about the tendency of the outcome to occur given the current state of the world *and* a future state that may not be realized. Although such hypothetical tendencies may be calculated, they do not correspond to the building blocks of causation.

Having just argued that the tendency is of a variable at a given time point taking a single value, let's add time indices to the causal diagram in **Figure 1**. The resulting diagram is shown in **Figure 4**.

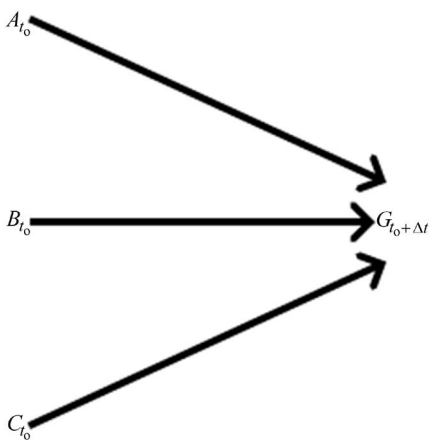

Figure 4. The causal diagram in **Figure 1** with time indices.

3.5. Event-Full Science, Not So Eventful

It is common practice to estimate the tendency of an event [1]-[3]. But I have just explained that tendency refers to the tendency of a variable to take one of its values. Is an event a variable, or at least a value of a variable? To answer that question we must first ask: What is a variable? Are all variables alike?

To begin, we should distinguish between two types of variables: natural variables and derived variables. Natural variables are properties of physical objects; they make up the causal structure of the universe. Derived variables, in contrast, are variables whose values are determined mathematically. When treated as causes or effects of interest, de-

rived variables account for a bias (termed "thought bias") that arises when a "causal parameter" is estimated and no such parameter exists [4]. As such, they often poison estimators in that they render useless the testing of any theory in which they are involved. Such a harsh bias arises when you unknowingly commit the scientific treason of fabricating variables.

An event is the value of a derived variable, not a natural variable, as shown next. Allison ([1], p. 2) defines an event as "a qualitative change that may be situated in time." Specifically, an event is defined as a change from one value of a variable to another [2]. Event-status as a variable can take two values: event and no event. The simple fact that an event is *defined* means that event-status is a derived variable.

Although there are several arguments as to why derived variables are not natural variables [4] [5], I will provide just one argument relevant to event-status. The value taken by a natural variable describes something about the world *at a time point*, whereas event—a value of event-status—does not. To illustrate the point, consider the event time: the time before which a variable X took the value 0 and after which it took the value 1 (**Figure 5**).

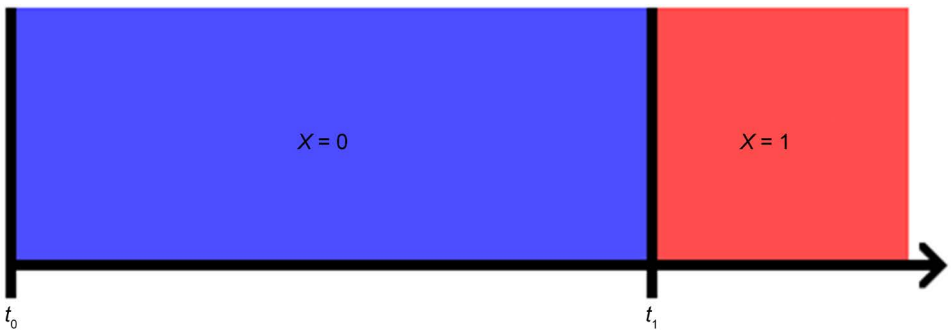

Figure 5. An illustration of the time of an event.

In **Figure 5**, the event is situated at time t_1, but knowing the value of X at time t_1 does not determine whether an event occurred at that time. In fact, as far as event status is concerned, it makes no difference what value X takes at time t_1. Since $X = 0$ before t_1 and $X = 1$ after t_1, the event occurs at t_1 *regardless of the value* X *takes at time* t_1. Therefore, the so-called event at t_1 does not describe the status of the world at time t_1. It follows that an event is not a value of a natural variable.

That completes my reasoning for not considering the tendency of the occurrence of events. Rather, I would like to study effects (changes in tendency) where the tendency is of a natural variable taking one of its values. Even still, Dr. Green and Dr. Black insist that events are the proper way to study the world. (I have reached an impasse with fictional characters!) I should probably stop talking to them and make a declaration.

The Declaration of Values:

When in the course of scientific events it becomes necessary for one scientist to dissolve the bands which have connected him with an inappropriate methodology and to assume another, a decent respect to the opinions of other scientists requires that he

should declare the reasoning which impels him to the separation.

I hold these truths to be self-evident, that all variables are created equal, that they are endowed with certain values, and that among these no value may be demarcated an event.—That the tendency I wish to consider is of a natural variable having a given value at a specified time point, not the occurrence of an event.

Finally, let me formally answer the question, "Tendency of what?", in one brisk sentence using **Figure 4** as an example.

The tendency of interest is the tendency of a natural variable (G) at a given time ($t_0 + \Delta t$) having a value (g) conditional on all of its causes (A, B, C) at a prior time (t_0) having specified values (a, b, c). (Notation: $T\left(G_{t_0+\Delta t} = g \mid A_{t_0} = a, B_{t_0} = b, C_{t_0} = c\right)$).

That tendency can vary with Δt, the time between the causes (A_{t_0}, B_{t_0}, C_{t_0}) and the effect ($G_{t_0+\Delta t}$); it can vary with the values of the causes (a, b, c); and it can vary with the value of the effect (g). Which brings us to the next point: On which value of the outcome should the effect be estimated?

3.6. The Value of Values

My answer is short: on every value of the outcome. But prevailing answers to that question make quite a list. For continuous variables, most researchers estimate the arithmetic mean difference, or rarely, the geometric mean ratio. They should direct their attention to the story of Mrs. Brown's vitamin A deficiency in subsection 3.1.

For binary variables, some propose to estimate the effect on the desired value, whereas others advocate for estimating the effect on the unwanted value [6]. Sheps [7] makes another proposal, although he does not think it is always applicable: "A beneficial factor acts on…'failures'…Conversely, a harmful factor acts on…'successes'." That is, a beneficial effect should be estimated on the undesired value and a harmful effect should be estimated on the desired value. Causation, though, is not assumed to vary with such desires. Therefore, I would prefer a methodology that does not depend on them.

Another solution avoids the issue of desired and undesired values altogether: Use a measure of effect that contains the information for the effect on all values, such as the proportion difference or the odds ratio [8]. That solution, however, works only for binary variables, and I would like a measure of effect that is applicable to all variables.

For categorical (non-binary) variables, Allison and others discuss a modeling method—multinomial regression—that estimates the effect on all values of the outcome [9]. Therein lies the solution: Estimate the effect on every value of the outcome. That way we have included every change in tendency in which we might be interested.

To illustrate why you might be interested in all such effects, consider the effect of a binary variable E on a trinary variable D. Let 0 and 1 be the two values of E and suppose that D has one value which is considered good, another value which is considered bad, and a third value to which we are indifferent. I will creatively name the good value *good* and the bad value *bad*. Suppose that on some scale the effect of E changing from 0 to 1 on D = *good* is very large. That is, D is much more likely to take its good value if $E = 1$ than if $E = 0$. Should you prefer that E takes the value 1 rather than the value 0? Not

necessarily. $E = 1$ might also make it much more likely for D to take the value *bad*, which is mathematically possible for a trinary D. So the preferred value of E depends on the magnitude of the effects on both $D = good$ and $D = bad$.

In general, it is necessary to consider the effect on all values of the outcome when making decisions based upon effects, just as one would consider possible side effects of some treatment along with possible benefits. The last point is usually obscured for binary variables since an increase in the probability of the good value is always accompanied by a decrease in the probability of the bad value. Nonetheless, even for binary variables we should estimate the effect on both values of the outcome. To paraphrase Sheps, since the two effects may give very different impressions, it is always well advised to consider both comparisons [7].

3.7. A Quick Recap

The effect of A_{t_0} on $G_{t_0+\Delta t}$ is a change in the tendency $T\left(G_{t_0+\Delta t} = g \mid A_{t_0} = a, B_{t_0} = b, C_{t_0} = c\right)$ due to the changing values of A_{t_0} (**Figure 4**). On a ratio scale, for instance, the effect of the causal contrast a_1 vs. a_0 is $T\left(G_{t_0+\Delta t} = g \mid A_{t_0} = a_1, B_{t_0} = b, C_{t_0} = c\right) \Big/ T\left(G_{t_0+\Delta t} = g \mid A_{t_0} = a_0, B_{t_0} = b, C_{t_0} = c\right)$. To specify such an effect, we must indicate the causal contrast (a_1 vs. a_0); the value of the outcome on which the effect is calculated (g); and the time interval between the cause and the outcome (Δt). We are ultimately interested in the set of such effects for all causal contrasts, all values of the outcome, and all time intervals between the cause and effect. This goal should be kept in mind in unavoidable, practical approximations.

(Note: The cause A_{t_0} is often called the exposure; for the causal contrast a_1 vs. a_0, a_1 and a_0 are often referred to as exposed and unexposed, respectively.)

4. Measures of Frequency

4.1. Quantifying the Tendency of Interest

How should tendency be quantified? Naturally, by some measures of frequency. I propose, however, that a single measure of frequency should be used to quantify the tendency of interest for any variable having any of its values at any time point. From a causal standpoint, there is no fundamental difference between different variables, different values, and different time points. To clarify, by "a single measure of frequency" I mean a measure of frequency that can be fully described by one mathematical framework as it applies to all variables, values, and time points. Furthermore, that mathematical framework should prescribe only one way of quantifying each tendency of interest; it should not allow for two different ways to quantify the same tendency.

In this section, I consider numerous measures of frequency and argue for or against them based upon whether they are capable of quantifying the tendency of interest *in all cases*. For now, let's consider quantifying the tendency for discrete variables. Later, I will discuss how the arguments may be extended to continuous variables.

For the rest of Section 4, if M is a measure of frequency, $M\left(X_t = x\right)$ will denote $M\left(X_t = x \mid \boldsymbol{C}_{t_0} = \boldsymbol{c}\right)$, where $\boldsymbol{C}_{t_0}$ is a vector of all the causes of X at time $t_0 < t$, and $\boldsymbol{c}$ is

a vector representing the values they take. For example, probability, denoted by *P*, is a measure of frequency. So $P\left(X_t = x\right)$ will denote $P\left(X_t = x \,|\, \boldsymbol{C}_{t_0} = \boldsymbol{c}\right)$.

4.2. Surviving the Cumulative Distribution Function

Several measures of frequency arise in survival analysis. In one way or another, all of them describe the tendency of an event to occur. Since event-status is not a natural variable (Subsection 3.5), none of them can generically quantify the tendency of a natural variable taking one of its values. Fortunately, measures of frequency in survival analysis can be modified slightly so as not to refer to events, and such measures may be able to quantify the tendency of a natural variable taking one of its values.

First, let's consider the cumulative distribution function, $F\left(t\right)$, which is the probability of having the event between t_0, the beginning of the study, and time *t*. That is, $F\left(t\right) = P\left(t_0 < T \le t\right)$, where *T* is the time at which the event occurred. For simplicity, I will define the event as the first occurrence during the study. (If someone has multiple occurrences during the study, only the first occurrence will be called an event.)

Next, I'll modify $F\left(t\right)$ to remove any reference to the event. The event for discrete variables is always defined as a natural variable, say *X*, taking a value, say x_1, after a period of time during which $X \ne x_1$. Therefore, the condition "$t_0 < T \le t$" is equivalent to the condition "for some $\tau \in \left(t_0, t\right], X_\tau = x_1$ and for some time $t_1 \in \left(t_0, \tau\right), X_s \ne x_1$ for all $s \in \left(t_1, \tau\right)$". Substituting the later condition for the former in the above expression produces

$$F\left(t\right) = P\left(\text{for some } \tau \in \left(t_0, t\right], X_\tau = x_1 \text{ and for some time } t_1 \in \left(t_0, \tau\right), X_s \ne x_1 \text{ for all } s \in \left(t_1, \tau\right)\right) \tag{1}$$

In the previous equation, the only reference to the event is the condition "for some time $t_1 \in \left(t_0, \tau\right), X_s \ne x_1$ for all $s \in \left(t_1, \tau\right)$". After removing that condition, we arrive at the following measure of frequency:

$$\mathcal{F}\left(t\right) = P\left(X_\tau = x_1 \text{ for some } \tau \in \left(t_0, t\right]\right) \tag{2}$$

The measure of frequency, $\mathcal{F}\left(t\right)$, is the modified form of the cumulative distribution function and has no reference to the event. Still, $\mathcal{F}\left(t\right)$ is the probability over a time *interval* and thus, cannot generically be used to quantify the tendency at a time *point*.

There is an exception, however, when events—forgive me for using the word—are nearly irreversible. In that case, almost everyone who has had $X = x_1$ prior to time *t* will have $X_t = x_1$. Therefore, $\mathcal{F}\left(t\right) \approx P\left(X_t = x_1\right)$, which has the potential for quantifying the tendency at a time point as will be discussed later (Subsection 4.5).

The survivor function, $S\left(t\right) = 1 - F\left(t\right)$, is even less fit than the cumulative density function to quantify the tendency of interest. To see that, first remove the survivor function's reference to the event. That leaves us with the following measure of frequency:

$$\mathcal{S}\left(t\right) = 1 - \mathcal{F}\left(t\right) = P\left(X_\tau \ne x_1 \text{ for all } \tau \in \left(t_0, t\right]\right) \tag{3}$$

$\mathcal{S}\left(t\right)$ is the probability of a variable *not having* a value in a time *interval*. As such, it

cannot generically be used to quantify the tendency of a variable *having* a value at a time *point*.

Again there is an exception, although with more constraints than before. If events—cringe!—are nearly irreversible and X is a binary variable, whose other value is denoted x_0, then $\mathcal{S}(t) = P(X_\tau = x_0 \text{ for all } \tau \in (t_0, t]) \approx P(X_t = x_0)$.

Even when the exceptions hold, neither $\mathcal{F}(t)$ nor $\mathcal{S}(t)$ can generically quantify the tendency of interest. Suppose for example, that X is binary and the event is nearly irreversible. Then $\mathcal{F}(t)$ cannot quantify the tendency of X_t having the value x_0, and $\mathcal{S}(t)$ cannot quantify the tendency of X_t having the value x_1. And remember, an appropriate measure of frequency should quantify the tendency of interest for a variable having any one of its values.

4.3. Finding the Nonexistent Probability Density Function

Next in survival analysis is the probability density function, $f(t) = \mathrm{d}F(t)/\mathrm{d}t$. The previous equality is true almost everywhere. Where it's not true, $f(t)$ is arbitrary and thus, unable to quantify the tendency of interest. Therefore, I shall assume $f(t) = \mathrm{d}F(t)/\mathrm{d}t$ everywhere. Under that assumption,

$$f(t) = \lim_{\Delta t \to 0^+} P(t \le T < t + \Delta t)/\Delta t \tag{4}$$

As before, $f(t)$ should be purged of any mention of the event. We have to first deal with the condition "$t \le T < t + \Delta t$" as it is more complicated than "$t_0 < T \le t$". In the latter condition, the event (the first occurrence) is indistinguishable from later occurrences, whereas in the former we must take explicit care to note whether the first occurrence was in the specified interval. In removing the mention of events, the reference to the first occurrence is altogether forgotten. As such, we may replace "$t \le T < t + \Delta t$" with "for some $\tau \in [t, t + \Delta t), X_\tau = x_1$ and for some time $t_1 \in (t_0, \tau), X_s \ne x_1$ for all $s \in (t_1, \tau)$". The event is still referred to in the condition "for some time $t_1 \in (t_0, \tau), X_s \ne x_1$ for all $s \in (t_1, \tau)$". Remove it, and we are left with an event-less measure of frequency:

$$\wr(t) = \lim_{\Delta t \to 0^+} P(X_\tau = x_1 \text{ for some } \tau \in [t, t + \Delta t))/\Delta t \tag{5}$$

The above expression does not consider events; it describes X having a value. And the limit guarantees that X is considered at a time point. Great! Too bad the limit doesn't exist for every value of the outcome.

Proposition: $\wr(t)$ is not defined for every value x of a discrete variable X.

Proof: Let X be a discrete variable with values $x_1, x_2, \cdots, x_n$ ($n \in \mathbb{N}$). Then, for all $i \in \{1, 2, \cdots, n\}$, $P(X_\tau = x_i \text{ for some } \tau \in [t, t + \Delta t)) \ge P(X_t = x_i) \ge 0$.

Since $P(X_\tau = x_i \text{ for some } \tau \in [t, t + \Delta t))$ is bounded below by zero and decreases monotonically as Δt approaches zero from the right,

$$\lim_{\Delta t \to 0^+} P(X_\tau = x_i \text{ for some } \tau \in [t, t + \Delta t)) \tag{6}$$

exists. Furthermore,

$$\lim_{\Delta t \to 0^+} P(X_\tau = x_i \text{ for some } \tau \in [t, t + \Delta t)) \ge \lim_{\Delta t \to 0^+} P(X_t = x_i) = P(X_t = x_i) \ge 0 \tag{7}$$

Still, $\sum_{i=1}^{n} P\left(X_t = x_i\right) = 1$. Therefore, there exists an $x \in \left\{x_1, x_2, \cdots, x_n\right\}$ such that $P\left(X_t = x\right) > 0$ And so,

$$\lim_{\Delta t \to 0^+} P\left(X_\tau = x \text{ for some } \tau \in \left[t, t+\Delta t\right)\right) > 0 \tag{8}$$

Since the numerator of

$$P\left(X_\tau = x \text{ for some } \tau \in \left[t, t+\Delta t\right)\right) / \Delta t \tag{9}$$

doesn't approach zero while the denominator does,

$$\lim_{\Delta t \to 0^+} P\left(X_\tau = x \text{ for some } \tau \in \left[t, t+\Delta t\right)\right) / \Delta t \tag{10}$$

does not exist. QED

The above proof shows that at every time point there is at least one value of a discrete variable for which the tendency of interest cannot be quantified by $\mathfrak{f}(t)$. Moreover, the limit that defines $\mathfrak{f}(t)$ will not exist for any value x_i and at any time t for which $P\left(X_t = x_i\right) > 0$. Therefore, we cannot know for which values (if any) and at what times (if ever) we may use $\mathfrak{f}(t)$ to quantify tendency. So much for the modified probability density function; it cannot quantify the tendency of interest for a discrete variable having any one of its values at any time point.

4.4. Avoiding the Hazard (Function)

The hazard function, $h(t)$, is merely a conditional probability density function. Conditional measures of frequency are not truly new measures of frequency. Rather, they use the old measures of frequency to estimate different frequencies. Thus, the hazard would be expected to contain all of the defects of the probability density function, but that is not the case.

The hazard function is the probability density function conditional on not having had the event by time *t*. That is,

$$h(t) = \lim_{\Delta t \to 0^+} P\left(t \le T < t+\Delta t \middle| T \ge t\right) / \Delta t \tag{11}$$

Upon removing "all" of the references to the event in the hazard function, we are left with the following measure of frequency:

$$\mathfrak{h}(t) = \lim_{\Delta t \to 0^+} P\left(X_\tau = x_1 \text{ for some } \tau \in \left[t, t+\Delta t\right) \middle| X_s \ne x_1 \text{ for all } s \in \left(t_0, t\right)\right) / \Delta t \tag{12}$$

It is debatable whether the condition "$X_s \ne x_1$ for all $s \in \left(t_0, t\right)$" is truly a reference to the event, since it appears as a given condition. If you consider it a reference to the event, remove it and you will be left with $\mathfrak{f}(t)$.

Suppose, instead, that $\mathfrak{h}(t)$ does not have any references to the event. Then, unlike $\mathfrak{f}(t)$, $\mathfrak{h}(t)$ might exist. If you were to repeat the above proof with $\mathfrak{h}(t)$ it would require that $\sum_{i=1}^{n} P\left(X_t = x_i \middle| X_s \ne x_i \text{ for all } s \in \left(t_0, t\right)\right) = 1$, which is not necessarily true. As such, $\mathfrak{h}(t)$ may be defined for all times and for all values of *X*.

Still, $\mathfrak{h}(t)$ cannot quantify the tendency of interest. The tendency of interest is the tendency of the outcome conditional on all of its causes at a prior time having specified values. (See the italics in subsection 3.5). To reiterate, that means conditional on what-

ever is specified and nothing else. Even if X at an earlier time is a cause of X_t, the tendency would include conditioning on $X_{t_0} = x_1$, but most certainly not on "$X_s \neq x_1$ for all $s \in (t_0, t)$". To include such conditioning is to ask for the tendency given a hypothetical future state (Subsection 3.4). Therefore, $\ell\!\!\!h(t)$ is not the desired measure of frequency.

4.5. Probability at a Time Point

So far I have mentioned in passing only one, possibly suitable, measure of frequency: The probability at a time point. $P(X_t = x)$ describes the tendency of a natural variable (X) at a given time (t) having a given value (x) and contains no unwanted conditioning. Furthermore, $P(X_t = x)$ is defined for all variables X, all time points t, and all values x. Still it remains to see whether the probability at a time point can quantify the tendency of interest.

For discrete variables, it most certainly can. We will need, however, to accept axiomatically that $P(X_t = x) \neq 0$ for any x to avoid violating the indeterministic assumption that tendencies are never zero (See Section 2).

For continuous variables, $P(X_t = x) = 0$ for all x, and our indeterministic assumption is automatically violated. Furthermore, the tendency is not assumed to be fixed *a priori*; $P(X_t = x)$ must be able to vary if it is to quantify tendency. Since that is not possible for continuous variables, we will have to look elsewhere for an acceptable measure of frequency. We will see, though, how probability at a time point still plays a crucial role in developing such a measure.

4.6. Meet the Relatives

Two common measures of frequency are related to the probability at a time point: the probability *over* a time interval and the rate. Those measures primarily differ in their consideration of time. Both, however, consider the probability of having a value *within* a time interval and therefore, cannot generically quantify the tendency at a time point. You might think to take their limits as the time interval shrinks to zero; those limits either don't exist or *equal* the probability at a time point. As such, they offer no new potential candidates for quantifying the tendency. Nonetheless, I think it is worthwhile to discuss their relation to the probability at a time point.

The probability over a time interval is $P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t])$. For small enough time intervals, it is assumed that a discrete X is unlikely to change values. So for small Δt,

$$P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t]) \approx P(X_t = x) \tag{13}$$

For larger time intervals, $P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t])$ does not estimate the probability at a time point. Still, it is related to that probability.

To illustrate that relation, let's consider why that measure of frequency is used. It is used when the probability at a time point is very small ($P(X_\tau = x) \approx 0$) such that you are unlikely to be able to estimate it without an incredibly large sample. If, however, we

want to relate the probability over a large time interval to the probability at a time point, we have to assume that $P(X_\tau = x)$ is approximately constant in the time interval $[t, t+\Delta t]$.

For argument's sake, let's consider an idealized model. Assume that every participant in the study is observed for the entire duration of the study; that $P(X_\tau = x) = p$ is constant; and that $p \approx 0$. Furthermore, assume that X cannot change within the following time intervals: $[t, t+D)$, $[t+D, t+2D)$, $\cdots$, $[t+\Delta t-D, t+\Delta t]$, where Δt is an integer multiple of D. Also assume that the value of X in one of those time intervals is independent of X in any of the other time intervals. Thus, the probability of not having $X = x$ in any time interval is $(1-p)^n$, where $n = \Delta t/D$, is the number of specified time intervals. Since p is small,

$$P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t]) = 1-(1-p)^n \approx 1-(1-pn) = pn \qquad (14)$$

The idealized model is similar to a study in which the outcome of interest is rare and the duration (D) during which a study participant has the outcome is more or less identical. For example, consider a study in which X is heart attack status, and x is the acute phase of a heart attack. It is unlikely for a person to be in the *acute* phase at a particular instance (*i.e.*, $P(X_\tau = x) = p \approx 0$), and the acute phase of heart attacks usually last for several hours (*i.e.*, D = several hours). In reality, the relation between the probability during a large time interval and the probability at a time point is more complicated. With more relaxed assumptions than those of the idealized model, it is possible to show that for $p \approx 0$, $P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t])$ is approximately proportional to p.

In general, $P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t])$ cannot be used to estimate the probability at a time point unless the constant of proportionality is known. In some measures of effect, such as the ratio, the constant of proportionality will cancel so long as it is the same in both exposed and unexposed. So although $P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t])$ cannot quantify tendencies, it may under certain conditions be used to estimate an effect.

If the two probabilities, $P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t])$ and $P(X_t = x)$, may be considered reasonably behaved siblings, the incidence rate is the problem child. The incidence rate is defined as the number of people observed to have the event in a given interval divided by the total time at risk. That being said, it has a few unresolved issues when it comes to quantifying tendency. First, the incidence rate deals with events, and as discussed earlier, the event should be replaced with "$X_\tau = x$ for some $\tau \in [t, t+\Delta t]$". Second, "time at risk" is a misnomer once the event is forgone. If we are to disregard the event, "time at risk" really means "time observed". Third, the incidence rate is defined in terms of counts which may vary between populations even when the underlying tendency is assumed fixed. As such, the incidence rate does not distinguish estimates from the parameter which is being estimated.

The incidence rate can be altered to fix the above problems. I will call the resulting measure of frequency the event-less rate. First, the number of people who have the event needs to be replaced by N_x, the number of people for which $X_\tau = x$ for some

$\tau \in [t, t+\Delta t]$. Second, the total time at risk should be replaced by $N\Delta t$, where N is the number of people in the study. After these alterations, the incidence rate becomes $N_x/(N\Delta t)$. Third, N_x/N, which is calculated in terms of counts, may be replaced by $P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t])$, which is the underlying parameter that N_x/N estimates. Therefore, the event-less rate may be defined as
$P(X_\tau = x \text{ for some } \tau \in [t, t+\Delta t])/\Delta t$.

In the idealized model discussed above, the event-less rate would be approximately equal to p/D. In more realistic models, the event-less rate is approximately proportional to p for small p. As such, the event-less rate ratio may be used to estimate the probability ratio when the constant of proportionality is assumed to be identical in exposed and unexposed.

4.7. Continuing Discreetly from Discrete to Continuous

Most of the measures of frequency discussed so far cannot quantify the tendency of interest at all, because they do not consider the outcome at a single time point. Only the hazard function applies to a time point, but it introduces irrelevant conditioning. Whatever continuous analogs those measures may have, they may at best serve to approximate a relevant measure of frequency.

So far only two measures are applicable to quantifying tendency: the event-less probability density function, $\not{f}(t)$ (See Subsection 4.3), and the probability at a time point, $P(X_\tau = x)$. The event-less probability density function may exist for continuous variables—intuitively, I think it does—but its nonexistence for discrete variables prevents it from generically quantifying the tendency of interest. The probability at a time point schleps around the opposite problem; it can quantify the tendency for discrete variables but not for continuous ones. While there is no way to have the probability density function describe the tendency for discrete variables, the probability at a time point may be extended to continuous variables.

Before discussing that extension, I will mention a probability density function of a different sort. Instead of the probability density function $f(t) = \mathrm{d}F(t)/\mathrm{d}t$, let's consider the probability density function $f_{X_t}(x) = \mathrm{d}F_{X_t}(x)/\mathrm{d}x$, where X_t takes real values and $F_{X_t}(x) = P(X_t \le x)$. As with $f(t)$, the equality $f_{X_t}(x) = \mathrm{d}F_{X_t}(x)/\mathrm{d}x$ holds almost everywhere. Where it's not true, $f_{X_t}(x)$ is arbitrary and thus, unable to quantify the tendency of interest. Therefore, I shall assume $f_{X_t}(x) = \mathrm{d}F_{X_t}(x)/\mathrm{d}x$ everywhere. Under that assumption,

$$f_{X_t}(x) = \lim_{\Delta x \to 0^+} P(x < X_t \le x + \Delta x)/\Delta x \tag{15}$$

After a quick inspection, you will see that $f_{X_t}(x)$ is perfectly capable of quantifying the tendency of interest for continuous variables. Furthermore, $f_{X_t}(x)$ may be viewed as $P(X_t = x)$ extended to continuous variables.

That extension is best explained with some new concepts. Consider a generic variable, X_t, and let $\mathcal{X}$ denote the set of all its values. Let d be a metric on $\mathcal{X}$. That is, $d(x_1, x_2)$ describes how similar two values $x_1, x_2 \in \mathcal{X}$ are to one another. If x_1 and

x_2 are very different, then $d(x_1, x_2)$ will be a large positive number. If x_1 and x_2 are similar, then $d(x_1, x_2)$ will be a small positive number. And if $x_1 = x_2$ then $d(x_1, x_2) = 0$. We will also consider a measure, μ, on $\mathcal{X}$. That is, $\mu(A)$ is a function that quantifies the size of a subset A of $\mathcal{X}$ (e.g., the number of elements in A). For those familiar with measure theory, μ will be defined on the Borel σ-algebra. Lastly, we will define the notion of a ball. The open ball of radius $r > 0$ centered at the value $x_0 \in \mathcal{X}$ is the set $B_r(x_0) = \{x \in \mathcal{X} : d(x_0, x) < r\}$.

We might then consider the following measure of frequency: $P(X_t \in B_r(x_0))$, where $X_t \in B_r(x_0)$ means X_t takes a value belonging to the set $B_r(x_0)$. $P(X_t \in B_r(x_0))$, however, quantifies the tendency of X_t to take one of many values, not a single value. Taking the limit as r approaches zero would ensure that only a single value is considered, but the limit is zero for continuous variables:

$$\lim_{r \to 0^+} P(X_t \in B_r(x_0)) = 0 \tag{16}$$

That problem can be avoided if we first divide $P(X_t \in B_r(x_0))$ by $\mu(B_r(x_0))$, the limit of which need not equal zero for continuous variables. The resulting measure of frequency will be called "the likelihood at a time point" to be denoted

$$L(X_t = x_0) = \lim_{r \to 0^+} P(X_t \in B_r(x_0)) / \mu(B_r(x_0)) \tag{17}$$

For those familiar with measure theory, the likelihood will be a representative of the Radon-Nikodym derivative, dP/dμ, under mild restrictions (*i.e.*, $L(X_t = x_0) = \mathrm{d}P/\mathrm{d}\mu(x_0)$ almost everywhere) [10].

At first glance, the likelihood at a time point suffers three problems. First, the limit below need not exist:

$$\lim_{r \to 0^+} P(X_t \in B_r(x_0)) / \mu(B_r(x_0)) \tag{18}$$

We can, however, accept axiomatically that it exists just as we accepted that $f_{X_t}(x) = \mathrm{d}F_{X_t}(x)/\mathrm{d}x$ everywhere. Even still certain restrictions will need to be placed on the functions d and μ. Otherwise, we could take μ to be the zero measure, for instance, in which case the limit never exists. Second, the likelihood might equal zero violating indeterminism (See Section 2). We may accept axiomatically that the likelihood is never zero, just as we did regarding probabilities for discrete variables in subsection 4.5. Third, the likelihood at a time point need not be unique; it can depend on the choice of the functions d and μ. The functions d and μ correspond to the choice of units for continuous variables. As such, "different" likelihoods essentially correspond to different units. Once the functions d and μ are specified, it is straightforward to change from one likelihood to another. In that respect, the likelihood is essentially unique.

It remains to show how $L(X_t = x_0)$ extends $P(X_t = x_0)$ to continuous variables. In particular, we will show that $L(X_t = x_0) = P(X_t = x_0)$ for discrete X_t, so that $L(X_t = x_0)$ may indeed be considered an extension of $P(X_t = x_0)$. Then we shall prove that $L(X_t = x_0) = f_{X_t}(x_0)$ for continuous X_t taking real values.

For discrete X_t, let d be the discrete metric:

$$d\left(x_1, x_2\right) = \begin{cases} 1 & \text{if } x_1 \neq x_2 \\ 0 & \text{if } x_1 = x_2 \end{cases} \tag{19}$$

and let μ be the counting measure. That is, $\mu(A)$ is the number of elements in A. Then, $B_r\left(x_0\right) = \left\{x_0\right\}$ for small enough r. Therefore,

$$L\left(X_t = x_0\right) = \lim_{r \to 0^+} \frac{P\left(X_t \in B_r\left(x_0\right)\right)}{\mu\left(B_r\left(x_0\right)\right)} = \frac{P\left(X_t \in \left\{x_0\right\}\right)}{\mu\left(\left\{x_0\right\}\right)} = \frac{P\left(X_t = x_0\right)}{1} = P\left(X_t = x_0\right) \tag{20}$$

for discrete X_t.

For continuous X_t taking real values, let d be the Euclidean distance on $\mathbb{R}$ (*i.e.*, $d\left(x_1, x_2\right) = \left|x_2 - x_1\right|$), and let μ be the Lebesgue measure on $\mathbb{R}$ (*i.e.*, the size of an interval is its length). Therefore,

$$L\left(X_t = x_0\right) = \lim_{r \to 0^+} \frac{P\left(X_t \in B_r\left(x_0\right)\right)}{\mu\left(B_r\left(x_0\right)\right)} = \lim_{\Delta x \to 0^+} \frac{P\left(x - \Delta x/2 < X_t < x + \Delta x/2\right)}{\Delta x} \tag{21}$$

where we viewed r as being $\Delta x/2$. It then follows that $L\left(X_t = x_0\right) = f_{X_t}\left(x_0\right)$ if we assume as before that $f_{X_t}\left(x\right) = \mathrm{d}F_{X_t}\left(x\right)/\mathrm{d}x$ everywhere.

In summary, the likelihood combines $P\left(X_t = x_0\right)$ for discrete variables and $f_{X_t}\left(x_0\right)$ for continuous variables under one mathematical framework. Furthermore, it applies to more general variables (e.g., random vectors) by generalizing the notions of similarity (d) and size (μ). As such, the likelihood is capable of describing the tendency of interest for all variables. I will discuss the last point further at the end of this section.

4.8. Beating the Odds

Among the measures of frequency discussed so far, only the likelihood at a time point has been shown to possibly quantify the tendency of interest. All other measures of frequency either cannot quantify the tendency of interest, or are merely approximations of the likelihood at a time point under certain conditions. And although there may be an infinite number of ways to quantify tendency, I will consider only one more common measure of frequency: the odds.

Many articles purport that the odds is a bad measure of frequency because it approximates the probability only in certain cases, or more specifically that the odds ratio is a bad measure of effect because it approximates the probability ratio only in certain cases [6] [11] [12]. Those arguments amount to nothing more than noting that the odds and probability are different measures of frequencies [8]. Their being different does not imply that the odds is an improper way to quantify tendency. Even when the odds ratio has been given proper consideration [13], little has been mentioned about the odds itself. As such, it's about time the odds receives a fair trial.

The odds at a time point is defined as $O\left(X_t = x\right) = P\left(X_t = x\right)/P\left(X_t \neq x\right)$. Odds calculated from probabilities over time intervals or by other methods cannot quantify the tendency at a time point. And as with probability, those other odds merely approximate the odds at a time point under certain conditions. Due to the close relation between odds and probability, the odds at a time point can quantify the tendency for

discrete variables just as aptly as the probability at a time point. For continuous variables, however, the odds fails to quantify time point tendency, because $O(X_t = x) = 0$ for all values *x*. As such, the odds at a time point cannot generically quantify tendency. Perhaps though, the odds at a time point can be extended to quantify the tendency for continuous variables by replacing probabilities with likelihoods. Unfortunately, the denominator of the odds would have to be replaced by $L(X_t \neq x)$, which is not defined.

We can instead try to extend an odds-like measure that has been used for non-binary outcomes [1] [9]. I will call this measure the partial odds. To illustrate the partial odds consider a binary variable X_t whose values are x_0 and x_1.
$O(X_t = x_1) = P(X_t = x_1)/P(X_t \neq x_1) = P(X_t = x_1)/P(X_t = x_0)$. That is, the odds for a binary variable is simply the probability of taking one value divided by the probability of taking the other value. For discrete variables, the partial odds is defined analogously. That is, the partial odds for a discrete variable is the probability of taking one value divided by the probability of taking another value. This definition can now be easily extended to continuous variables by replacing probabilities with likelihoods.

The partial odds, however, introduces a new problem not present in the odds. Consider a discrete variable X_t with values $x_1, x_2, \ldots, x_n$ $(n > 2)$. Then there are $n - 1$ possibly different partial odds of X_t taking the value x_1. All of them take the form $P(X_t = x_1)/P(X_t = x_i)$ where $i \in \{2, 3, \ldots, n\}$. In that respect, the partial odds seem strange because they quantify the tendency of X_t taking the value x_1 by considering other values of X_t. Furthermore, for all non-binary variables the partial odds provides multiple ways of quantifying each tendency of interest. As there is only one tendency to quantify, I would like a measure of frequency that does not prescribe multiple ways of quantifying the same tendency. As such, the partial odds is yet another measure that fails to quantify the tendency of interest.

4.9. A Likely Suspect

As promised I now return to discuss the likelihood at a time point, the only contender still quantifying tendency. As I have not done so before, let's explicitly verify that the likelihood, $L(X_t = x_0)$, can quantify the tendency of interest.

$L(X_t = x_0)$ is a measure of frequency that describes a natural variable (X) at a given time (t) having a given value (x_0) conditional on all of its causes at a prior time having specified values. Recall that the conditioning on all causes is implicit in the notation $L(X_t = x_0)$ as was mentioned near the beginning of Section 4: "if M is a measure of frequency, $M(X_t = x)$ will denote $M(X_t = x \mid \boldsymbol{C}_{t_0} = \boldsymbol{c})$, where $\boldsymbol{C}_{t_0}$ is a vector of all the causes of X at time $t_0 < t$, and $\boldsymbol{c}$ is a vector representing the values they take."

Being a measure of frequency, $L(X_t = x_0)$ can quantify tendencies. Furthermore, the tendency of interest is precisely what is described by the likelihood at a time point. As specified in subsection 3.5: "The tendency of interest is the tendency of a natural variable (X) at a given time (t) having a given value (x_0) conditional on all of its causes ($\boldsymbol{C}$) at a prior time (t_0) having specified values ($\boldsymbol{c}$)."

In order to quantify the tendency, the likelihood must satisfy a few more properties. First, it must be defined for all variables, all values, and all time points. I have axiomatically accepted in subsection 4.7 that the likelihood is defined in all such cases. Second, unlike the time point probability of continuous variables, the likelihood must never be fixed *a priori*. That is, it should vary in all cases with the tendency of the outcome. In constructing the likelihood, $\mu\left(B_r\left(x_0\right)\right)$ was placed in the denominator to guarantee just that. Third, unlike the partial odds, there must be only one likelihood to quantify each tendency of interest, which follows from the definition of the likelihood. Fourth, the likelihood must be described by one mathematical framework for every value of every variable. By viewing the likelihood as a Radon-Nikodym derivative, we see that it can indeed be described by one mathematical framework. Having satisfied all the above properties the likelihood at a time point can indeed quantify the tendency of interest. Moreover, of the measures discussed, it is the *only* one that can quantify the tendency of interest.

One issue remains. There might be yet another measure of frequency capable of quantifying the tendency of interest; the notions of tendency and measure of frequency are not explicit enough to ensure that no such measure exists. If there was such a measure, would it be preferred to the likelihood or not, and how are we to decide? To solve such a problem, perhaps we should not only accept that the likelihood *can* quantify the tendency of interest, but rather accept axiomatically that the likelihood at a time point *is*, in fact, the tendency of interest.

5. Measures of Effect

5.1. Ratio vs. Difference

Having concluded that the likelihood is the "ideal" way to quantify tendency, let's move on to deciding which measure of effect to use. Numerous functions may be proposed to quantify the change in tendency. I will consider just two—the ratio and the difference—and then argue generically against almost all others.

To begin, let me explicitly define the ratio and the difference. Consider the effect of E_0 on D_1. Specifically, consider the effect of E_0 (e_1 vs. e_0) on D_1 taking the value d, and let $\boldsymbol{C}_0$ be a vector of all the causes of D_1 at time 0 except E_0, and let $\boldsymbol{c}$ be a vector representing the values they take. On a ratio scale that effect is $L\left(D_1=d \mid E_0=e_1,\boldsymbol{C}_0=\boldsymbol{c}\right)/L\left(D_1=d \mid E_0=e_0,\boldsymbol{C}_0=\boldsymbol{c}\right)$; on a difference scale that effect is $L\left(D_1=d \mid E_0=e_1,\boldsymbol{C}_0=\boldsymbol{c}\right)-L\left(D_1=d \mid E_0=e_0,\boldsymbol{C}_0=\boldsymbol{c}\right)$. Fairly simple.

The debate between proponents of the ratio and proponents of the difference is filled with many comments, which I find insignificant. I will note previous arguments and explain their lack of importance in deciding between the ratio and the difference. Finally, I will present the only substantive argument I have found capable of deciding between the two measures.

5.2. Irrelevant Arguments

The choice of a measure of effect depends, in part, on the goal in mind, and mine is

simply to quantify the change in tendency. The arguments I will discuss in this section are irrelevant in the sense that they arise from other goals—namely, estimating averages in a population, and enhancing people's understanding of the data.

One branch of determinism occupies itself with target populations (*i.e.*, estimating effects in finite populations). To that end, the difference is thought to be more applicable than the ratio. Greenland, for instance, notes that the difference of the average probability in a population equals the average probability difference, whereas the ratio of the average probability is not the average probability ratio [13]. The probability difference also equals the proportion causative minus the proportion preventive in a target [14], which under determinism tells you how many more people have been helped than harmed. Although those arguments may be applicable to studying target populations under determinism, they are irrelevant to the goal of quantifying the change in tendency under indeterminism.

Some people prefer a measure of effect that "enhances people's understanding" or is "easily interpretable" [8]. After gleaning the literature, those vague phrases can be rephrased precisely: a measure of effect should quantify effects in a way that corresponds to intuitive notions of large and small effects. Intuition, however, is subjective. And so, there is a disagreement as to whether the difference or the ratio is more easily interpreted. According to Sheps' intuition [7], "Many people would consider that…the difference… could not be adequately appreciated…without reference to the level of the [probabilities] themselves". For example, if the difference were $P(D=1|E=1)-P(D=1|E=0)=0.01$, that could be considered a large effect if $P(D=1|E=0)=0.000001$ and a small effect if $P(D=1|E=0)=0.5$. On the other hand, I have heard complaints that the ratio is difficult to interpret since it can purport large effects when the probability in both exposed and unexposed is small. For example, if $P(D=1|E=1)=0.000008$ and $P(D=1|E=0)=0.000001$, the effect on a ratio scale will be 8, a rather large effect. Yet some people intuitively consider the effect to be minuscule, as is the case on a difference scale [3] [15] [16]. Both viewpoints are based on intuition, and although intuition may spark debates, it cannot settle them. Intuition, after all, may be mistaken. The arguments are further weakened given that the above intuitions contradict each other.

Cook and Sackett present another related argument against the ratio [15]. They cite a stroke study in which the effect of treatment on stroke on a ratio scale is the same in two groups. They interpret that as "suggesting that both should be treated with equal vigour". On the difference scale the effects differ, however. Therefore, they conclude, "The clinical recommendation is…likely to be different between the two groups." But even on the ratio scale, the clinical recommendation would differ between the groups. They should not be "treated with equal vigour", for the reason presented in subsection 3.6: the effect should be estimated on all values of the outcome. In the example cited by Cook and Sackett, we should estimate the effect on both values of stroke status: stroke and no stroke. On the ratio scale, the effect on stroke status taking the value "no stroke" is different for the two groups, and the clinical recommendation should differ accordingly.

5.3. Math before Reason

The ratio and the difference have different mathematical properties that are relevant to the debate, but those mathematical properties are of little value without substantiating their necessity on philosophical grounds. I have listed below the mathematical arguments of which I am aware. All of them, except for the first, have no philosophical grounds; the first is based on philosophical grounds against which I have already argued.

1. The effect on a difference scale is symmetric for binary variables. That is, if you know the effect on one value of the outcome you know the effect on both. For binary variables, the symmetry of the difference scale solved the dilemma people had: On which value to estimate the effect [7] [8]. I have already offered a solution in subsection 3.6. Even so, this property of the difference scale has no value once we consider non-binary variables.
2. Some people prefer that effects approach extreme values as they approach deterministic limits [16]. For discrete variables, that means that effects should become very large as probabilities approach one or zero. The difference does not meet this criterion. It does not approach an extreme value when either the probability in exposed or unexposed approaches zero or one. The difference only approaches the extremes of 1 or −1 when the probability approaches one in one group and zero in the other.

   The ratio, on the other hand, does not approach any extreme when the probability in exposed or unexposed approaches one. It does, however, approach infinity as the probability nears zero in unexposed, and it approaches zero when the probability in exposed nears zero. Thus, the ratio has the advantage of approaching an extreme in more cases than the difference.

   For continuous variables the story is a little different. Since the likelihood for continuous variables is not constrained between zero and one but between zero and infinity, there are more instances in which the ratio and difference approach extremes. The ratio, however, still has the advantage of approaching an extreme in more cases than the difference. In fact, for continuous variables, the ratio approaches an extreme in all deterministic limits.
3. Earlier in the article, I reviewed a few measures of frequency related to the probability at a time point. Under certain conditions, those measures are approximately proportional to the probability at a time point. Furthermore, the constant of proportionality may be the same in exposed and unexposed, in which case it will cancel on the ratio scale. Thus, those measures of effects may be used to estimate the probability ratio under certain conditions.

   On a difference scale, however, the constants of proportionality will not cancel. As such, we cannot estimate the probability difference from related measures of frequency unless the constants of proportionality are known. Nonetheless, if the difference was found to be the preferred measure of effect, it wouldn't matter that the ratio can be estimated by those measures.
4. For continuous variables, the difference scale has units while the ratio scale does not. The units can be said to make it difficult to appreciate the magnitude of an ef-

fect. Once the units are understood, however, the difference scale for continuous variables is not much more difficult to interpret than the difference scale for discrete variables. (Note: different units correspond to different functions d and μ in the definition of the likelihood. See Subsection 4.7.)

5.4. A Reason to Decide between the Ratio and the Difference

The arguments noted so far are in no way definitive. There is only one argument that I consider to be important in deciding between the ratio and difference. To explain that argument we first need to consider the way in which effects are estimated.

To estimate an effect, auxiliary causal theories must always be invoked. Those theories can be represented in a causal diagram. For example, Figure 6 shows a causal diagram applicable to estimating the effect $E_0 \rightarrow D_1$; the auxiliary theories are represented by the arrows $C_{-1} \rightarrow I_0$, $I_0 \rightarrow D_1$, and $C_{-1} \rightarrow E_0$. Those theories are essential for estimating effects unbiasedly. For example, according to Figure 6, we would need to condition on C_{-1} or I_0 to remove confounding bias. In general, the causal diagram to which we hold indicates on which variables we must condition to remove bias.

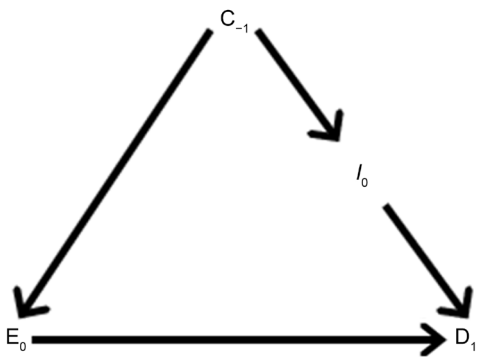

Figure 6. A causal diagram depicting confounding bias.

Removing bias, however, is not the only way to improve a study. In general, a study is better the more likely it is to produce estimates closer to the true effect. As such, we must strike a balance between minimizing bias and minimizing variance. Therefore, it is not necessary to condition on all of the variables mandated by a causal diagram. If we don't condition on all the relevant variables, then the magnitude of bias may be described by a bias term, which depends, in part, on the effects specified in auxiliary theories.

In general, the bias term is a function of effects, tendencies, and other likelihoods. The bias depends on the effects in auxiliary theories, and occasionally it depends on the effect being estimated. It may also depend on various tendencies. Lastly, the bias term may contain marginal likelihoods. These are likelihoods which do not include conditioning on the causes of the outcome. As such, they are not assumed fixed between studies.

Taking Figure 6 as an example, we can illustrate how the bias term depends on ef-

fects, tendencies, and marginal likelihoods. For simplicity assume all of the variables in question are discrete. Suppose we were to estimate the effect $E_0 \rightarrow D_1$ without conditioning on C_{-1} or I_0 as needed to remove bias. Then, the bias term depends on the effects $C_{-1} \rightarrow I_0$, $I_0 \rightarrow D_1$, and $C_{-1} \rightarrow E_0$; on various tendencies; and on the marginal likelihood $L\left(C_{-1} = c\right) = P\left(C_{-1} = c\right)$, which gives the marginal distribution of C_{-1}.

We were not able to say, however, on which various tendencies the bias term depends because the answer depends on the measure of effect. In general, different measures of effect have different bias terms. On the difference scale, for instance, the bias term is another term added onto the effect. That is, every study that estimates an effect on the difference scale actually estimates the effect plus a bias term. And when that term equals zero, the study is said to be unbiased. On the ratio scale, the bias term is actually a bias factor. That is, every study that estimates an effect on the ratio scale actually estimates the effect multiplied by a bias factor. And when that factor equals one, the study is said to be unbiased.

Note that I am not using the term "bias" to mean that the expected value of the estimator differs from the parameter being estimated. Rather, I use it to mean that the parameter being estimated differs from the parameter we wish to estimate [14]. For example, in **Figure 6** the effect on the difference scale is $P\left(D_1 = d \mid E_0 = e_1, I_0 = i\right) - P\left(D_1 = d \mid E_0 = e_0, I_0 = i\right)$. If we don't condition on I_0, then we are estimating $P\left(D_1 = d \mid E_0 = e_1\right) - P\left(D_1 = d \mid E_0 = e_0\right)$. The difference between those terms is the bias term on the difference scale. Similarly, on the ratio scale the effect is $P\left(D_1 = d \mid E_0 = e_1, I_0 = i\right) / P\left(D_1 = d \mid E_0 = e_0, I_0 = i\right)$. If we don't condition on I_0, then we are estimating $P\left(D_1 = d \mid E_0 = e_1\right) / P\left(D_1 = d \mid E_0 = e_0\right)$. The ratio of the two is the bias factor on the ratio scale.

Now that we have described the bias term, we return to our earlier comment about striking a balance between minimizing bias and minimizing variance. Since the bias term depends on effects that are not known *a priori*, the balancing act between bias and variance is qualitative. We are not interested in the exact amount of bias, but whether the bias is large enough to be of concern, or small enough to be ignored. Such a description of bias requires that the auxiliary theories be qualitative, not quantitative. It also requires that we know the relation between the bias term and different effects. I will refer to those relations as the rules of causal diagrams.

In many cases, the rules of causal diagrams can be described as follows. When certain relevant effects are null, the bias term indicates that no bias will be present. For example, if any of the effects $C_{-1} \rightarrow I_0$, $I_0 \rightarrow D_1$, or $C_{-1} \rightarrow E_0$ were null, then no bias will be present when estimating $E_0 \rightarrow D_1$ without conditioning. As the magnitude of these effects increases, the bias term usually indicates that more and more bias is present. We say "usually" because the various tendencies and marginal likelihoods could lead to less or even no bias. When precisely no bias is present because of the tendencies and likelihoods, we say that a lucky cancellation has occurred.

Since auxiliary theories are qualitative, such precise cancellations are usually not of importance. Instead, we need a general description of the rules of causal diagrams as il-

lustrated above. There should be some magnitudes of effects for which no bias is present; the more the effects differ from those special cases, the more bias should usually be present. Furthermore, the measure of effect used in these rules should be the measure of effect we decide upon. That is, we should not write the bias term for the difference in terms of the ratio nor should we write the bias factor for the ratio in terms of the difference. Under that constraint, it is possible that what one measure considers to be a lucky cancellation could be explained by another measure of effect. It seems strange that we could know about bias on one scale *purely* from the effects on another scale. Therefore, I would prefer a measure of effect for which lucky cancellations cannot be explained by another measure.

In many cases, there is no problem because on both the difference scale and the ratio scale, bias is often absent when certain effects are null, and the bias usually increases as the effects strengthen. Since a null effect on the two scales is equivalent, both the difference and the ratio are often capable of explaining their respective bias. There is, however, an exception pertaining to colliding bias, specifically bi-path colliding bias.

Consider Diagram A of Figure 7 where Q_0 and R_0 are two marginally independent causes of K_1. Q_0 and R_0 are dependent conditional on $K_1 = k$ if and only if they modify each other's effects on $K_1 = k$ on the probability ratio scale [17]. This has been proven when Q_0, R_0, and K_1 are all discrete, and is suspected to hold for any Q_0, R_0, and K_1. The result of conditioning on $K_1 = k$ when Q_0 and R_0 modify each other's effects on a probability ratio scale is depicted in Diagram B of Figure 7. Effect modification is denoted by lower case q_0 and r_0 on the arrows; conditioning on $K_1 = k$ is denoted by a box around K_1; the two lines over the adjacent arrows denote that the arrows no longer contribute to associations; and the dashed line between Q_0 and R_0 indicates that the association between them has changed after conditioning.

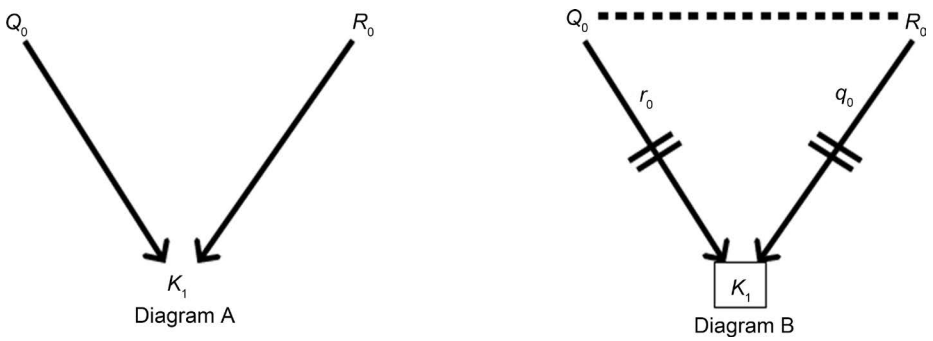

Figure 7. Conditioning on a collider.

If the diagram in Figure 7 is embedded in another causal diagram, then the association created between Q_0 and R_0 can produce bias. Consider the M-structure depicted in Figure 8. If we condition on just $K_1 = k$ when estimating the effect $E_1 \rightarrow D_2$, then no bias is present on either the difference scale or the ratio scale if Q_0 and R_0 do not modify each other's effect on $K_1 = k$ on the ratio scale (Figure 8 Diagram A) . The stronger the effect modification between Q_0 and R_0 on the ratio scale, the larger the bias usually is on

both the difference scale and the ratio scale (Figure 8 Diagram B).

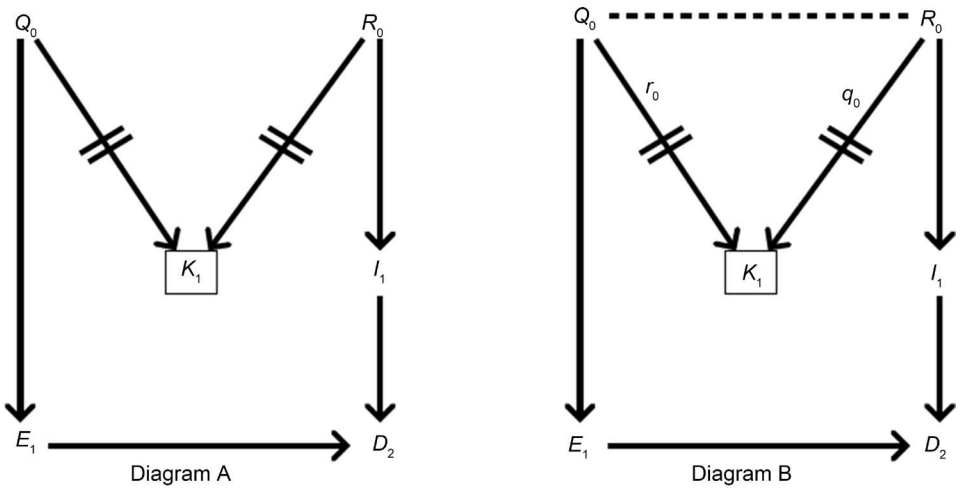

Figure 8. The M-structure.

Therefore, the rule governing bias for either scale in Figure 8 depends on effect modification on the ratio scale. That in itself is not a problem so long as effects on the difference can inform as to whether or not effect modification exists on the ratio scale. But that is not possible as will be illustrated in the next section. And so the difference is not an acceptable measure of effect.

Nonetheless, it is not clear from the above reasoning that the ratio *is* an acceptable measure of effect. In the next subsection, however, I show that by the above reasoning the ratio is essentially the only possible, appropriate measure of effect. Still it might be the case that a lucky cancellation on the ratio scale could be explained by another measure of effect. If so, the above reasoning does not give an absolute preference for the ratio. To verify that that is not the case, we would need to check all the rules of causal diagrams as they apply to the ratio.

5.5. Math after Reason

So far I have concluded that the ratio is preferred to the difference. Nonetheless, there are many other possible measures of effect. In general, a measure of effect is a function, f, of two tendencies: one in exposed and one in unexposed, which will be denoted by $u \in (0,\infty)$ and $w \in (0,\infty)$, respectively. After requiring f to satisfy reasonable properties, only certain measures of effect will be possible.

First, I would like effects to be quantified by real number (or possibly a real number with units). It may be denoted by other things, but it seems strange to quantify a change between real numbers (possibly with units) with a complex number, a 19-dimensional vector, or anything else.

Second, effects are intuitively considered to be similar if the relevant tendencies are similar. As such, I would like such effects to be denoted by similar real numbers. That is, I would like f to be jointly continuous in its two arguments. Informally, if Δu and

Δw are small enough changes in the tendency, then

$$f(u,w) \approx f(u+\Delta u, w+\Delta w) \tag{22}$$

Note that both the difference and the ratio scale satisfy that property. (For the ratio, this holds so long as the likelihood is never zero. See Subsection 4.7.).

Third, there should be a real number representing a null effect on the new scale. Why? Because the null is unique! While we may argue what is considered a large effect or a small effect, the null represents the special case when the outcome is indifferent to the value of the cause. Even the ratio and the difference agree as to what is considered a null effect. I would like the new measure of effect to share in that consensus. To make the previous statement mathematically rigorous, there should be a real number C representing a null effect such that

$$f(u,w) = C \text{ if and only if } u = w \tag{23}$$

Lastly, we consider the rules of causal diagrams. As shown in Figure 7, the association between Q_0 and R_0 as quantified on the ratio scale may change upon conditioning on the collider K_1. Whether such a change will occur depends on effect modification on the ratio scale. Yet a non-null association on the ratio scale is accompanied by a non-null association on the new scale, because associations and effects are quantified in the same way. It will follow that whether bias will be present on the new scale in Figure 8 depends on effect modification on the ratio scale. That is, the rule governing bias in Figure 8 for the new scale depends on effect modification on the ratio scale. If the new scale is to overcome the deficiency of the difference, we should be able to tell whether there is effect modification on the ratio scale based solely upon effects on the new scale.

To make the last statement mathematically rigorous, consider the causal diagram in Figure 7. Let $u_r = L(K_1 = k \mid Q_0 = 2, R_0 = r)$ and $w_r = L(K_1 = k \mid Q_0 = 1, R_0 = r)$. We will consider two effects: the effect of Q_0 when $R_0 = 1$ and the effect of Q_0 when $R_0 = 2$. On the ratio scale, these two effects are u_1/w_1 and u_2/w_2, respectively, and on the new scale, they are $f(u_1,w_1)$ and $f(u_2,w_2)$, respectively.

To be able to tell whether effect modification is present on the ratio scale based solely upon effects on the new scale, there must be some relation, R, such that $f(u_1,w_1)$ and $f(u_2,w_2)$ are related by R ($f(u_1,w_1)\mathrm{R}\, f(u_2,w_2)$) if and only if there is no effect modification on a ratio scale ($u_1/w_1 = u_2/w_2$). That is,

$$u_1/w_1 = u_2/w_2 \Leftrightarrow f(u_1,w_1)\mathrm{R}\, f(u_2,w_2) \tag{24}$$

A quick check will show that R is an equivalence relation. In fact, we will prove that R is just equality on the range of f. (*i.e.*, $x\mathrm{R}y \Leftrightarrow x = y$ where both x and y are in the range of f.) It then follows that $f(u,w) = g(u/w)$, for some injective function *g*. That is, any acceptable measure of effect is an injective function of the ratio. As such, both the ratio scale and the new scale contain exactly the same information.

Next, we prove the claims made in the last paragraph.

Proposition: $f(u,w) = g(u/w)$, for some injective function *g*.

Proof: First, we will prove by contradiction that R is equality on the range of f. Suppose that is not the case, and consider the functions $f_\lambda(w) = f(\lambda w, w)$ and

$f^u(w) = f(u,w)$.

For each λ, $f_\lambda(w)$ is a continuous function of w. Moreover, there is a λ_0 such that $f_{\lambda_0}(w)$ is a non-constant function, otherwise R would just be equality on the range of f equality by Equation (24). Therefore, there are two non-zero numbers (possibly with units) w_0 and w_1 such that $f_{\lambda_0}(w_0) < f_{\lambda_0}(w_1)$.

For each u, $f^u(w)$ is a continuous function of w. Furthermore, $f^u(w)$ must be injective, otherwise $f(u,w)$ would be the same for two different values of w, but u/w would be different contradicting Equation (24). Since $f^u(w)$ is both continuous and injective, it is strictly monotonic. Let $u_0 = \lambda_0 w_0 \neq 0$. Since $f^u(w)$ is strictly monotonic and $w_0 \neq 0$, there exists a w_2 such that $f^{u_0}(w_0) < f^{u_0}(w_2)$.

Next, let M be the minimum of $f_{\lambda_0}(w_1)$ and $f^{u_0}(w_2)$, and let S be a number less than M and greater than $f_{\lambda_0}(w_0) = f(\lambda_0 w_0, w_0) = f^{u_0}(w_0)$. Then $f_{\lambda_0}(w_0) < S < f_{\lambda_0}(w_1)$ and $f^{u_0}(w_0) < S < f^{u_0}(w_2)$. Since both $f_{\lambda_0}(w)$ and $f^{u_0}(w)$ are continuous, it follows by the intermediate value theorem that there exist values w_3 and w_4, not equal w_0, such that $f_{\lambda_0}(w_3) = S$ and $f^{u_0}(w_4) = S$. Therefore,

$$f(\lambda_0 w_3, w_3) = f_{\lambda_0}(w_3) = S = f^{u_0}(w_4) = f(\lambda_0 w_0, w_4) \tag{25}$$

Then $f(\lambda_0 w_3, w_3) \mathrm{R} f(\lambda_0 w_0, w_4)$, because R is an equivalence relation. Yet $w_4 \neq w_0$, so $\lambda_0 w_3 / w_3 = \lambda_0 \neq \lambda_0 w_0 / w_4$, which contradicts Equation (24). Therefore, R is in fact equality on the range of f, and

$$u_1/w_1 = u_2/w_2 \Leftrightarrow f(u_1, w_1) = f(u_2, w_2) \tag{26}$$

Next, we will prove the statement of the proposition. By Equation (26), $f_\lambda(w)$ is a constant function for each λ. Let $g(\lambda)$ be that constant. Then $f(u,w) = f_{u/w}(w) = g(u/w)$. Lastly, g is injective: If $\lambda_1 \neq \lambda_2$, then $f_{\lambda_1}(w) \neq f_{\lambda_2}(w)$, so $g(\lambda_1) \neq g(\lambda_2)$. QED

5.6. Math Mediates among Measures

Until now, I have listed some of the properties I want in a measure of effect. The result is that only continuous, injective functions of the ratio have all of those properties. Among them I have yet to give a reason to prefer one over the rest. Since they all contain the same information, the only reason to prefer one over the others has to do with their mathematical properties. In particular, their mathematical properties that pertain to bias and variance—the two integral parts of a study that are related to the chosen measure of effect.

I would like the new measure of effect, $g(\lambda)$, to describe bias in a simple manner. On the ratio scale, bias is described by a bias factor, B, which is a term multiplied by the effect, λ. Describing bias for the new measure of effect, g, will involve a comparison of between $g(B\lambda)$ and $g(\lambda)$. If g is to describe bias in a simple manner, it should make it simple to deal with multiplication. In particular, I would like g to satisfy one of two properties: $g(B\lambda) = g(B)g(\lambda)$ or $g(B\lambda) = g(B) + g(\lambda)$. Since g is continuous and injective, only specific functions satisfy those properties [18]. The first property,

$g(B\lambda)=g(B)g(\lambda)$, is satisfied only by functions of the form $g(x)=x^c$ where $c\neq 0$ is a constant. The functions that satisfy the second property, $g(B\lambda)=g(B)+g(\lambda)$, are logarithms. They can all be written as $g(x)=c\ln(x)$ where $c\neq 0$ is a constant. We thus have two families of measures of effect that describe bias in a simple manner.

Next, let's consider the issue of the variance. To begin, let Λ be a random variable of the estimates of the effect on a ratio scale of E_0 (e_1 vs. e_0) on $D_1 = d$. That is, Λ is a random variable of the estimates of $\lambda = L(D_1 = d \mid E_0 = e_1, \boldsymbol{C}_0 = \boldsymbol{c}) / L(D_1 = d \mid E_0 = e_0, \boldsymbol{C}_0 = \boldsymbol{c})$. $1/\Lambda$ will be a random variable of the estimates of $1/\lambda = L(D_1 = d \mid E_0 = e_0, \boldsymbol{C}_0 = \boldsymbol{c}) / L(D_1 = d \mid E_0 = e_1, \boldsymbol{C}_0 = \boldsymbol{c})$. Both Λ and $1/\Lambda$ estimate what is essentially the same effect using the same data. Yet $\mathrm{Var}(\Lambda)$ may not equal $\mathrm{Var}(1/\Lambda)$, which makes it difficult to appreciate the variance without reference to the estimated effect. For example, a variance of 0.3 with an estimate of λ = 10 might be considered small, while the same variance with an effect of $1/\lambda$ = 0.1 would be quite large. To remedy that problem, g should satisfy $\mathrm{Var}(g(\Lambda)) = \mathrm{Var}(g(1/\Lambda))$, where $g(\Lambda)$ and $g(1/\Lambda)$ are random variables of the estimates of $g(\lambda)$ and $g(1/\lambda)$, respectively. Functions of the form $g(x) = x^c$ do not generally satisfy $\mathrm{Var}(g(\Lambda)) = \mathrm{Var}(g(1/\Lambda))$, but logarithms do:

$$\mathrm{Var}(c\ln(\Lambda)) = \mathrm{Var}(-c\ln(\Lambda)) = \mathrm{Var}(c\ln(1/\Lambda)) \tag{27}$$

Furthermore, the same variance on the ratio scale is intuitively interpreted as a smaller spread for larger effects. For example, a variance of 1 for an effect of 24 is considered a smaller variance than a variance of 1 for an effect of 1.5. That is illustrated in the practice of quantifying the spread on a ratio scale by dividing the variance by the estimated effect. Logarithms, which squish larger values closer together, would naturally have the variance exhibit such a property.

There is one bizarre practice regarding the variance that I haven't mentioned. When estimating the effect on the ratio scale ($g(\lambda)=\lambda^c=\lambda^1$), the variance on the log scale is usually calculated to inform about the distribution of the ratio scale. The same practice may be used when estimating the effect on any other scale of the form $g(\lambda)=\lambda^c$. Nevertheless, to use $\mathrm{Var}(c\ln(\Lambda))$ as a measure of spread of Λ^c makes it even more difficult to appreciate the variance. Therefore, only measures of effect of the form $g(\lambda)=c\ln(\lambda)$ should be used.

Lastly, how do we choose the constant c? Once again, consider the variance. I mentioned before that the variance on a ratio scale is sometimes divided by the estimated effect as a measure of spread. If the effect is estimated to be null that would just be the variance. Furthermore, if Λ has a very tight distribution about the null, then $\mathrm{Var}(\Lambda)\approx\mathrm{Var}(1/\Lambda)$. In that case, the variance on the ratio scale satisfies all the properties fitting of the variance. Therefore, I propose using the variance in such a case as a standard. That is, we should choose the constant c so that $\mathrm{Var}(\Lambda)\approx\mathrm{Var}(c\ln(\Lambda))$ when Λ has a very tight distribution about the null. In that case,

$$\mathrm{Var}(c\ln(\Lambda)) = \mathrm{Var}(c\ln(1+\Lambda-1)) \approx \mathrm{Var}(c(\Lambda-1)) = c^2\mathrm{Var}(\Lambda-1) = c^2\mathrm{Var}(\Lambda) \tag{28}$$

Therefore, $c = \pm 1$. If we add the restriction that $g(x) = c\ln(x)$ be monotonically increasing, then $c = 1$. The function g is already guaranteed to be monotonic, since it is both continuous and injective; the restriction that it be increasing prevents effects from decreasing on the new scale as the ratio increases.

Therefore, the natural logarithm of the ratio (or log ratio, for short) is the only measure of effect that satisfies all the nice mathematical properties I have outlined. It will take time to get used to describing effects on the log ratio scale as opposed to the ratio or any other scale. But for all of its advantages I think the effort is worthwhile.

And so we have reached the climax of the article. The change in tendency should be quantified by the natural logarithm of the ratio of likelihood. For example, the effect of E_{t_0} (e_1 vs. e_0) on $D_{t_0+\Delta t}$ taking the value d would be quantified by $\ln\left(L\left(D_{t_0+\Delta t} = d \mid E_{t_0} = e_1, \boldsymbol{C}_{t_0} = \boldsymbol{c}\right) \Big/ L\left(D_{t_0+\Delta t} = d \mid E_{t_0} = e_0, \boldsymbol{C}_{t_0} = \boldsymbol{c}\right)\right)$. To fully describe what is generally considered the effect of E on D, we would have to estimate the previous quantity for all causal contrasts e_1 vs e_0, all values d, and all time intervals Δt. That is the ultimate meaning of causal knowledge.

In practice, we can estimate the log likelihood ratio by relying on the assumption that the likelihood $L\left(D_{t_0+\Delta t} = d \mid E_{t_0} = e, \boldsymbol{C}_{t_0} = \boldsymbol{c}\right)$ is a jointly continuous function of e, d, and Δt. That is, we can estimate the likelihood by considering people whose value of E_{t_0} is similar to e and whose value of $D_{t_0+\Delta t'}$ is about d where $\Delta t' \approx \Delta t$. Alternatively, we could try fitting a model of the form

$$\ln\left(L\left(D = d\right)\right) = \beta_0\left(d\right) + \beta_1\left(d\right)E \tag{29}$$

where $\beta_1(d)$ is the effect of E increasing by one on $D = d$. Since the parameters $\beta_0(d)$ and $\beta_1(d)$ depend on d, there may be too many parameters in the model depending on how many values D has. Therefore, it might be necessary to force restrictions on how the β depend on d to allow for proper estimation.

5.7. An Example of an Effect on the Log Ratio Scale

To illustrate the log ratio scale, consider the effect of a binary variable E_0 on all fives values of a discrete variable D_1. The top of Table 1 shows $L\left(D_1 = d \mid E_0 = e, \boldsymbol{C}_0 = \boldsymbol{c}\right)$ for all the combinations of values of d and e, which equals $P\left(D_1 = d \mid E_0 = e, \boldsymbol{C}_0 = \boldsymbol{c}\right)$ since D_1 is discrete. The bottom of Table 1 shows the effect on the log ratio scale of E_0 (e_1 vs. e_0) on D_1 taking each of its values. I have included in the bottom of Table 1 the effect on the ratio scale so as to help compare the effects as quantified by the two measures.

Table 1. An example of effects on the log ratio scale (vs. the ratio scale).

	$D_1 = d_1$	$D_1 = d_2$	$D_1 = d_3$	$D_1 = d_4$	$D_1 = d_5$
$E_0 = e_1$	0.16	0.29	0.35	0.12	0.08
$E_0 = e_0$	0.16	0.35	0.29	0.01	0.19
Log ratio	0	−0.19	0.19	2.485	−0.86
Ratio	1	0.83	1.21	12	0.42

A few things are worth noting about the effects in Table 1. The null effect on $D_1 = d_1$ is denoted by the number 0 on the log ratio scale. The effect on $D_1 = d_2$ and $D_1 = d_3$ are reciprocals of each other on the ratio scale and additive inverses of each other on the log ratio scale. Furthermore, for effects close to the null the effect on the log ratio scale is approximately the effect on the ratio scale minus one. For example, the effect on $D_1 = d_3$ on the log ratio scale is $0.19 \approx 0.21 = 1.21 - 1$, which is the effect on the ratio scale minus one. Finally, very large effects on the ratio scale are denoted by much smaller numbers on the log ratio scale. For example, the effect on $D_1 = d_4$ is 12 on the ratio scale, but only 2.485 on the log ratio scale.

6. Discussion

6.1. Likes and Dislikes

"I would like..." and "I would prefer..." are how I began some of my sentences in this article. But why do my preferences matter? It's not that my preferences matter, so much as that they detail how I propose to study science. To explain further, we must first answer: what is science?

I have yet to encounter a good definition of science, scientific knowledge, or scientific theories. A rudimentary "definition" tells us that science is the study of how the universe works, a statement that is clearer than most definitions but otherwise incredibly vague. What is the universe? What is meant by "how it works"? And what about it are we to study exactly?

The answers to the first two questions arise from axioms of science. I partially answered the third at the beginning of this article: I argued in favor of estimating an effect (a change in tendency). That is the knowledge in which we are interested. Well, at least I'm interested in it; you may not be. Still, the answer I supplied is vague. So I have filled this article with statements explaining more thoroughly what I want to study about the universe and how I think it should be studied. Those statements were indicated with the phrases "I would like...", "I would prefer...", and "should". Now I would like to move on to the next subsection.

6.2. Approximations of Effects for Discrete Outcomes

Several measures of frequency may be used, under certain circumstances, to approximate the probability ratio (*i.e.*, the likelihood ratio for discrete outcomes), and therefore the log likelihood ratio as well [19] [20]. To the extent that those measures may be used to estimate the probability ratio, we need not view them as different measures of effect. Instead, we may view them as estimators that increase bias in return for reduced variance.

For example, the odds ratio from a cohort study need not be viewed as a separate measure of effect, but rather as a probability ratio with a bias factor. Specifically, the odds ratio in a cohort study equals the probability ratio times a bias factor, which is the effect of the exposure on not having the disease. If E and D are binary variables with values denoted 0 and 1, then the effect of E (1 vs. 0) on having the disease ($D = 1$) on

the odds ratio scale (OR) is related to the probability ratio (PR) of having the disease as follows:

$$\begin{aligned} OR &= \frac{P(D=1|E=1)/P(D=0|E=1)}{P(D=1|E=0)/P(D=0|E=0)} \\ &= \frac{P(D=1|E=1)}{P(D=1|E=0)} \times \frac{P(D=0|E=0)}{P(D=0|E=1)} \\ &= PR \times \frac{P(D=0|E=0)}{P(D=0|E=1)} \end{aligned} \tag{30}$$

Therefore, the log of the odds ratio equals the log likelihood ratio plus a bias term as follows:

$$\ln(OR) = \ln(PR) + \ln\left(\frac{P(D=0|E=0)}{P(D=0|E=1)}\right) \tag{31}$$

The same bias term can be developed for the log odds ratio from a case-control study in which the exposure (E) effects selection (S) only through the disease (D). That is, under the causal structure $E \rightarrow D \rightarrow S$.

6.3. Redefining Cause

The tendency of interest includes conditioning on all causes of the outcome. But without an explicit definition of the word "cause", that may be confusing. The word "cause" has multiple meanings. In Section 2, for example, I used "cause" to mean one thing in determinism and another in indeterminism. Within indeterminism, there may be two definitions of the word:

Definition 1: A cause of a natural variable Y_{t_1} is a natural variable X_{t_0} $(t_0 < t_1)$ with at least two values x_0 and x_1 such that the effect of X_{t_0} (x_1 vs. x_0) on Y_{t_1} taking the value y is not null for at least one value y of Y_{t_1}.

Definition 2: A cause of a natural variable Y_{t_1} is a natural variable X_{t_0} for which $t_0 < t_1$.

Definition 1 specifies that a cause must have a non-null effect on the outcome, whereas Definition 2 omits that requirement. I prefer Definition 2 since it defines cause so as to differentiate between "having a null effect" and "having no effect". Although the two phrases are often considered synonymous, they need not be. Note that "null" is the German word for zero, and in math a distinction is often made between "zero" and "no". "No" generally refers to something not existing, whereas "zero" refers to something that exists and takes the value zero.

Consider the two variables X_{t_0} and Y_{t_1} with $t_0 < t_1$. The effect of X_{t_0} on Y_{t_1} may be null, but the effect of Y_{t_1} on X_{t_0} does not exist. The effect of one variable on another exists only when the former is prior to the latter and both are natural variables (Subsection 3.5). Therefore, I propose that the phrase "no effect" be reserved for instances when an effect does not exist, and that the phrase "null effect" be used to indicate that an effect exists but its magnitude is null.

Variables that have null effects are qualitatively different from variables that have no effect. We may estimate null effects, and in fact, they are very similar to effects that are near the null. As such, I do not want to use Definition 1 that lumps together variables with null effects and variables with no effects under the title "not a cause". I suggest that within indeterminism Definition 2 be taken as the definition of the word "cause".

6.4. What Is Probability

Just like the word "cause", probability has several definitions, which are related to its philosophical interpretations [21]. Since probability is involved in quantifying tendency (*i.e.*, in the definition of likelihood), it requires a philosophical interpretation suitable for such a description.

In indeterminism, tendencies are objective. Furthermore, they apply only to single cases. For example, consider tossing a coin repeatedly. The tendency of having the coin land on heads applies to each and every coin toss. That is, we may speak of the tendency of the first toss landing on heads, the third toss landing on heads, and the seventh toss landing on tails. Thus, our interpretation of probability must be objective and apply to single cases.

For such a philosophical interpretation, I prefer a view similar to Miller's formulation of Popper's propensity theory [22] [23]. Miller ([23], p. 182) writes, "Probability of an outcome is…a measure of the inclination of current state of affairs to realize that outcome." I disagree with Miller but slightly; I would replace the probability with likelihood, since probability cannot quantify tendency for continuous variables. Miller ([23], p. 185) also notes, "Strictly, every propensity…must be referred to…the complete situation of the universe…at [a] time." That is equivalent to my noting that the tendency includes conditioning on all of the causes (Definition 2) at a time (including causes with null effects).

Gillies claims that such a philosophical interpretation does not allow for empirical estimation of probability, which is necessary to test scientific theories [21]. His reasoning is that the "complete situation of the world" is generally not a repeatable condition, and he deems it necessary for conditions to be repeatable in order to estimate probabilities. Perhaps there is a solution. If we are interested in quantifying effects (changes in tendency) and not the tendency itself, we only need to estimate probabilities conditional on all of the variables that would otherwise produce bias (e.g., confounders); we do *not* need to condition on all of the causes of the outcome.

Even with such an approach, it may not be feasible to estimate effects from empirical data. That, however, is the consequence of a bias-centered approach. Conditioning on everything that produces even the smallest amount of bias may prevent the estimation of an effect, but removing bias is not the only matter of importance in a study. The goal of any study is to be as likely as possible to produce estimates that are as close as possible to the true causal parameter. As such, it is beneficial in some cases to allow bias in exchange for a "small enough" variance. In every study, researchers need to decide how to balance bias and variance. It is then possible to estimate effects while holding to a

single-case propensity theory.

The above discussion clarifies that probability (or likelihood in general) may be interpreted so as to quantify the tendency of interest. The discussion, however, is incomplete without a philosophical definition of probability. In all of its definitions, probability is somehow related to proportions. In an objective interpretation, probability is related to the proportions in a sample that must be infinite in some sense. A few methods introduce an infinite sample as a finite sample becoming larger and larger. As far as I know, none successfully describes single-case probabilities.

I propose that probability be defined as *a proportion in an infinite sample*. In such a definition, I do not consider a finite sample getting larger but the infinite sample in and of itself. For example, a single-case probability of a particular coin toss landing heads is the proportion of heads in an infinite number of theoretical replications of that *exact* coin toss.

The proportion in an infinite sample (*i.e.*, over infinite replications) cannot be calculated by simple division as is the case in finite samples. Rather, it will be taken as a primitive notion, a notion that is not defined but understood fundamentally. (Primitive notions are common and necessary. For example, "time point" is a primitive notion.)

Finally, the axioms of probability can be derived from some of the definitions of probability [21], and in those cases the axioms are actually theorems that can be proven. My definition of probability cannot be used to prove the axioms of probability. Shockingly enough, they will truly be taken as axioms.

7. Conclusions

This article began with a thorough discussion of measures of effect from an indeterministic viewpoint. Under indeterminism, a measure of effect quantifies a change in tendency; that tendency is of a natural variable at a given time having a given value—conditional on all of its causes at a prior time having specified values. I then argued in favor of estimating effects on all values of the outcome.

Starting with Section 4, the majority of the article was devoted to finding an appropriate measure of effect based upon properties I wanted such a measure to possess. After considering all of those properties, I found only one acceptable measure of effect: the natural logarithm of the ratio of the likelihood at a time point. Although best among contenders, it is still just a helpful human-made computation on likelihoods, which I proposed to equate with tendencies. Under indeterminism, the causal structure contains only tendencies, not contrasts between tendencies (Subsection 3.2).

The question remains, though, whether there are other possible measures of effect besides the log likelihood ratio. The short answer is no; I think that one generic measure of effect should be used to quantify effects in all cases. The long answer is that there is still work to be done on the matter. If we do not accept axiomatically that the likelihood at a time point equals the tendency, then a rigorous definition of a measure of frequency must be given in order to consider all other measures of frequency. Lastly, all the rules of causal diagrams for the log likelihood ratio need to be developed in order to

verify that a lucky cancellation on the log ratio scale cannot be explained by another measure of effect.

References

[1] Allison, P.D. (1995) Survival Analysis Using the SAS® System: A Practical Guide. SAS Institute Inc., Cary.

[2] Tuma, N.B. (1982) Nonparametric and Partially Parametric Approaches to Event-History Analysis. *Sociological Methodology*, **13**, 1-60. http://dx.doi.org/10.2307/270717

[3] Pocock, S.J., Calyton, T.C. and Altman, D.G. (2002) Survival Plots of Time-to-Event Outcomes in Clinical Trials: Good Practice and Pitfalls. *The Lancet*, **359**, 1686-1689. http://dx.doi.org/10.1016/S0140-6736(02)08594-X

[4] Shahar, E. and Shahar, D.J. (2010) Causal Diagrams, Information Bias, and Thought Bias. *Pragmatic and Observational Research*, **1**, 33-47. http://dx.doi.org/10.2147/POR.S13335

[5] Shahar, E. and Shahar, D.J. (2012) Causal Diagrams and Change Variables. *Journal of Evaluation in Clinical Practice*, **18**, 143-148. http://dx.doi.org/10.1111/j.1365-2753.2010.01540.x

[6] Sinclair, J.C. and Bracken, M.B. (1994) Clinically Useful Measure of Effect in Binary Analyses of Randomized Trials. *Journal of Clinical Epidemiology*, **47**, 881-889. http://dx.doi.org/10.1016/0895-4356(94)90191-0

[7] Sheps, M.C. (1958) Shall We Count the Living or the Dead? *The New England Journal of Medicine*, **259**, 1210-1214. http://dx.doi.org/10.1056/NEJM195812182592505

[8] Walter, S.D. (2000) Choice of Effect Measure for Epidemiological Data. *Journal of Clinical Epidemiology*, **53**, 931-939. http://dx.doi.org/10.1016/S0895-4356(00)00210-9

[9] Allison, P.D. (1999) Logistic Regression Using the SAS® System: Theory and Application. SAS Institute Inc., Cary.

[10] Folland, G.B. (1999) Real Analysis: Modern Techniques and Their Applications. 2nd Edition, John Wiley & Sons, Inc., New York.

[11] Zhang, J. and Yu, K.F. (1998) What's the Relative Risk? A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. *JAMA*, **280**, 1690-1691. http://dx.doi.org/10.1001/jama.280.19.1690

[12] Rothwell, P.M. (1995) Can Overall Results of Clinical Trials be Applied to All Patients? *Lancet*, **345**, 1616-1619. http://dx.doi.org/10.1016/S0140-6736(95)90120-5

[13] Greenland, S. (1987) Interpretation and Choice of Effect Measures in Epidemiologic Analyses? *American Journal of Epidemiology*, **125**, 761-768.

[14] Maldonado, G. and Greenland, S. (2002) Estimating Causal Effects. *International Journal of Epidemiology*, **31**, 422-429. http://dx.doi.org/10.1093/ije/31.2.422

[15] Cook, R.J. and Sackett, D.L. (1995) The Number Needed to Treat: A Clinically Useful Measure of Treatment Effect. *BMJ*, **310**, 452-454. http://dx.doi.org/10.1136/bmj.310.6977.452

[16] Elbarbary, M. (2010) Understanding and Expressing "Risk". *Journal of the Saudi Heart Association*, **22**, 159-164. http://dx.doi.org/10.1016/j.jsha.2010.04.002

[17] Shahar, D.J. and Shahar, E. (Unpublished) A Theorem at the Core of Colliding Bias.

[18] Speck, G.P. (1997) On Elementary Functions Characterized by Functional Relationships. *International Journal of Mathematical Education in Science and Technology*, **28**, 609-612. http://dx.doi.org/10.1080/0020739970280416

[19] Symons, M.J. and Moore, D.T. (2002) Hazard Rate Ratio and Prospective Epidemiological Studies. *Journal of Clinical Epidemiology*, **55**, 893-899. http://dx.doi.org/10.1016/S0895-4356(02)00443-2

[20] Pearce, N. (2004) Effect Measures in Prevalence Studies. *Environmental Health Perspectives*, **112**, 1047-1050. http://dx.doi.org/10.1289/ehp.6927

[21] Gillies, D. (2000) Philosophical Theories of Probability. Routledge, New York.

[22] Popper, K. (1982) The Open Universe: An Argument for Indeterminism. Routledge, New York.

[23] Miller, D.W. (1994) Critical Rationalism: A Restatement and Defense. Open Court, Chicago.