Three Arguments for Absolute Outcome Measures

Jan Sprenger and Jacob Stegenga*†

Data from medical research are typically summarized with various types of outcome measures. We present three arguments in favor of absolute over relative outcome measures. The first argument is from cognitive bias: relative measures promote the reference class fallacy and the overestimation of treatment effectiveness. The second argument is decision-theoretic: absolute measures are superior to relative measures for making a decision between interventions. The third argument is causal: interpreted as measures of causal strength, absolute measures satisfy a set of desirable properties, but relative measures do not. Absolute outcome measures outperform relative measures on all counts.

1. Introduction. Clinical trials are performed in order to assess whether an experimental intervention is effective and, if so, to what degree. To make these inferences, data from clinical trials must be quantitatively summarized and analyzed in particular ways. Similar questions arise in epidemiology for assessing the degree to which exposure to a risk changes the probability of developing a disease.

Several classes of such quantitative methods of analysis are available to medical researchers, including 'relative' outcome measures and 'absolute' outcome measures (we precisely define prominent examples of these measures in sec. 2). Relative outcome measures are the most widely employed

*To contact the authors, please write to: Jan Sprenger, Tilburg Center for Logic, Ethics and Philosophy of Science, Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands; e-mail: j.sprenger@uvt.nl. Jacob Stegenga, Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge CB2 3RH, United Kingdom; e-mail: jms303@cam.ac.uk.

†The authors wish to thank Aaron Kenna, Clark Glymour, Felipe Romero, and the audience at PSA 2016 for helpful feedback and discussion. Research on this topic was financially supported by European Research Council Starting Investigator grant 640638 (Sprenger). The authors contributed equally to this article.

Philosophy of Science, 84 (December 2017) pp. 840–852. 0031-8248/2017/8405-0004$10.00
Copyright 2017 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/693930 Published online by Cambridge University Press

class of outcome measures. Here we offer three distinct arguments that absolute outcome measures are superior to relative outcome measures.

Relative measures are widely used in clinical science—our three arguments entail that this widespread practice is epistemologically corrupt. The first argument for the superiority of absolute measures is from *cognitive bias*: the use of relative measures promotes a reasoning fallacy and often leads to overestimation of intervention effectiveness, but absolute measures do not have this property (sec. 3; see also Stegenga 2015). Relative measures are erroneously taken to indicate risk reduction in the population as a whole. We prove, in section 4, that absolute outcome measures are sufficient (given the costs and utilities associated with the interventions) for making a rational choice between interventions. By contrast, relative outcome measures are neither necessary nor sufficient for choosing between two interventions. Thus, from a *decision-theoretic perspective*, the widespread use of relative outcome measures is misguided. In section 5 we present our third argument, the *causal strength* argument, in which we develop principled desiderata for probabilistic measures of causal strength and argue that absolute measures are superior to relative measures with respect to these desiderata. All three arguments employ particular absolute outcome measures (absolute risk reduction and number needed to treat) and relative outcome measures (relative risk and relative risk reduction) as exemplars. We conclude that medical science should more consistently use and report absolute outcome measures.

2. Outcome Measures in Medical Research. Clinical trials often measure outcomes in binary terms, such as the (non)occurrence of a heart attack in a certain time period. Many prominent outcome measures apply to binary events: the odds ratio, relative risk (or risk ratio), relative risk reduction, absolute risk reduction (or risk difference), and number needed to treat. These measures can be defined by constructing a two-by-two table: suppose a trial has one group (E) composed of subjects who receive the experimental intervention and a control group (C) composed of subjects who receive a different intervention (e.g., another intervention, a placebo, or no treatment at all). Suppose further that the binary outcome is measured as present (Y) or absent (~Y) and the number of subjects with each outcome in each group is represented by letters (a–d), as shown in table 1.

TABLE 1. Frequency Table for a Clinical Trial with Binary Outcomes

Group/Outcome	Outcome Present (Y)	Outcome Absent (~Y)	Total Number in Group
Experimental intervention (E)	a	b	a + b
Control (C)	c	d	c + d

https://doi.org/10.1086/693930 Published online by Cambridge University Press

Then the most prominent outcome measures can be defined as follows.

Relative risk: $RR = [a/(a + b)]/[c/(c + d)]$.

Relative risk reduction: $RRR = \{[a/(a + b)] - [c/(c + d)]\}/[c/(c + d)]$.

Absolute risk reduction: $ARR = a/(a + b) - c/(c + d)$.

Number needed to treat: $NNT = 1/\{[a/(a + b)] - [c/(c + d)]\}$.

Observe that all these measures are defined in terms of the observed relative frequencies $a/(a + b)$ and $c/(c + d)$.[1] It is convenient to write the outcome measures as a function of conditional probabilities that represent these frequencies. The probability of a subject having a Y outcome given that the subject is in group E, $P(Y|E)$, is $a/(a + b)$, and likewise, the probability of having a Y outcome given that the subject is in group C, $P(Y|C)$, is $c/(c + d)$.[2] Thus, we can define relative risk, relative risk reduction, absolute risk reduction, and number needed to treat as

$$\begin{aligned} RR &= P(Y|E)/P(Y|C). \\ RRR &= [P(Y|E) - P(Y|C)]/P(Y|C) = RR - 1. \\ ARR &= P(Y|E) - P(Y|C). \\ NNT &= 1/[P(Y|E) - P(Y|C)] = 1/ARR. \end{aligned}$$

Thus, RR and RRR are interchangeable and just differ in their scaling properties. Same with ARR and NNT. An intuitive interpretation of RR is the ratio of the frequency of recovery (or risk) in the treatment and the control group. RRR, by contrast, can be interpreted as a measure of causal attribution. For instance, let $P(Y|C)$ (the probability of dying without taking a particular drug) be 4%, and let $P(Y|E)$ (the probability of dying after taking the drug) be 1%. Then RRR is equal to 75%: this is the proportion of deaths in the control group that vanish in the treatment group. For this reason, RRR is

1. Another prominent outcome measure is the odds ratio $OR = (a/b)/(c/d)$. This is the same as the ratio of the relative risk (RR) for Y and for ~Y. OR is often proposed as an alternative to RR and RRR (e.g., Nurminen 1995), especially for case-control studies, and it belongs to the class of relative measures. Also note that ARR is sometimes called 'risk difference'.

2. These probabilities are calculated from observed actual frequencies, and we switch freely between both notations. However, they can also be interpreted as estimates of limiting relative frequencies, causal propensities, or subjective degrees of belief. Our article does not take a stand on this question.

https://doi.org/10.1086/693930 Published online by Cambridge University Press

a particularly popular outcome measure in clinical science and epidemiology (e.g., Walter 1976; Northridge 1995).

The probabilistic notation of outcome measures abstracts away from the sample size of a clinical trial. Here is an example. A large randomized controlled trial called the Heart Protection Study was performed to test the capacity of a cholesterol-lowering drug to mitigate heart attacks and death among men with heart disease (HPSCG 2002). Over 20,000 middle-age and elderly men who had heart disease or were at high risk for heart disease were recruited to the study, and half were randomly allocated to receive simvastatin (the cholesterol-lowering drug) and half to receive a placebo, for 5 years. After these 5 years, death from all causes was 12.9% in the simvastatin group and 14.7% in the placebo group, for an ARR of 1.8% and an RRR of 12.2%.[3] Notably, although the event rates for the study groups were reported in the abstract, the only outcome measures reported were relative measures.

This neglect of absolute outcome measures is a common practice in clinical research. A survey by King, Harper, and Young (2012) took a large a sample of articles published in medical and epidemiology journals and found that 75% reported only relative measures. This is encouraged by numerous proclamations to prefer relative outcome measures over alternatives, such as editorials in influential journals such as the *British Medical Journal*: "Authors and journal editors should ensure that the results of trials and systematic reviews are reported as relative risks unless there is a convincing argument otherwise" (Deeks 1998, 1155). In what follows we challenge this practice by providing three different arguments for the superiority of absolute measures over relative measures.

3. The Argument from Cognitive Bias. The framing of a medical risk often affects the conclusions that are drawn. Physicians and patients overestimate the effectiveness of medical interventions when presented with only relative measures (see also Stegenga 2015). This systematic overestimation occurs because the employment of relative measures, such as RR or RRR, promotes the reference class fallacy, which we will explain below.

For starters, relative and absolute outcome measures can appear very different when the control event rate (i.e., $P(Y|C)$) is low. Consider the example of the Helsinki Heart Study, which tested the capacity of a drug (gemfibrozil) to decrease cardiac disease and death (Frick et al. 1987). After 5 years of tak-

3. Strictly speaking, both ARR and RRR deliver negative values, but we follow the convention in much of the medical literature to only report absolute values and to suppress the sign.

https://doi.org/10.1086/693930 Published online by Cambridge University Press

ing the drug, the subjects in the experimental group had a reduced relative risk of cardiac disease of 34% (RRR). Because of the low base probability of cardiac disease, this amounted to an absolute risk reduction of 1.4% (ARR). These are different orders of magnitude.

It is a robust empirical finding that physicians are more likely to prescribe a drug when the risk is expressed in relative than in absolute terms (Forrow, Taylor, and Arnold 1992; Bobbio, Demichelis, and Giustetto 1994; Nexøe et al. 2002). In the experiment by Bobbio et al., which drew on data from the Helsinki Heart Study, physicians had to choose between various drugs on the basis of reported outcome measures. The effect of drug A was quantified with a relative outcome measure (RRR = 34%), and the effect of drug B was quantified with an absolute outcome measure (ARR = 1.4%). Physicians were much more likely to prescribe drug A than drug B, although both outcome measures were quantifications of the same data about the same drug. Patients show a similar pattern when asked for their acceptance of a medical treatment (Malenka et al. 1993; Hux and Naylor 1995; Sorensen et al. 2008).

This behavior represents a substantive overestimation of treatment effects. In many cases of common preventive care (e.g., lowering blood pressure or cholesterol levels), the rates of the risk (e.g., a cardiac event) are low in both the treatment and the control group. The above levels of RRR and ARRs correspond to a control event rate of $P(Y|C) = 4.1\%$ and a treatment event rate of $P(Y|E) = 2.7\%$. The relative outcome measure RRR = 34% suggests a strong effect when actually the treatment only helps a small number of patients exposed to the risk (1.4% of patients, to be precise).

Overestimation of intervention effectiveness is due to the *reference class fallacy*. That is, the sentence "a 34% cardiac event reduction was demonstrated" is taken to imply that 34% of all patients benefit from the treatment when in reality this number only refers to a small subset of that population: the patients in the control group that develop Y. The reference class fallacy explains why framing risk in relative terms leads to more optimistic estimates of effectiveness.

However, do physicians and patients really commit a fallacy? In the above study by Bobbio et al., the control event rate $P(Y|C)$ was not revealed to the participants. But without such information, one cannot meaningfully compare the values of ARR and RRR and realize that they have been computed from the same data set. Therefore, one cannot infer that participants in the above study are committing a proper reasoning fallacy.

This objection is sound, but unfortunately experiments reveal that cognitive bias persists in the face of full information. Malenka et al. (1993) observed that patients are, for the most part, unable to translate relative outcome measures into absolute outcome measures, even if the control event

https://doi.org/10.1086/693930 Published online by Cambridge University Press

rate is known. In their experiment, participants were presented with a control event rate P(Y|C) = 50% (the risk of dying in the next year without treatment) and a medication that would decrease this risk by 50%. Only 28.2% of the participants of the study drew the correct conclusion that this medication would prevent 25 deaths if 100 people were treated; 47.7% of participants claimed that 50 deaths would be prevented. Most other participants said they did not know the answer.

This experiment reveals that we are dealing with a proper reasoning fallacy and that this fallacy is due to a misidentification of the relevant reference class. Hence, inferring effectiveness of a medical treatment on the basis of relative outcome measures is indeed prone to cognitive bias. Since relative outcome measures trigger cognitive biases in both physicians and patients, such measures should be avoided. We will now argue that absolute outcome measures are excellent alternatives.

4. The Decision-Theoretic Argument. Some commentators have suggested that the absolute risk reduction measure is superior to relative measures in a decision context. This view is occasionally expressed in the clinical literature and sometimes by philosophers, such as Worrall (2010) and Stegenga (2015). Here we prove that this is indeed the case.

Let A mean that a patient consumes treatment A; let B mean that the patient consumes treatment B (this could be a competitor intervention, placebo, or nothing at all); let a be the cost of consuming A (where cost is construed broadly, to include all harmful effects of A); let b be the cost of consuming B (again construed broadly). Let Y mean that the outcome of interest occurs (e.g., recovery); finally, let the utility of Y be u and the utility of ~Y be u′.[4] The expected utility of consuming A is EU[A], and the expected utility of consuming B is EU[B].

The principle of maximizing expected utility holds that a patient should consume A rather than B if and only if (iff) the expected utility of consuming A is greater than that of consuming B. The corresponding decision rule is

> (#) For any u, u′, a, and b (without loss of generality: $a > b$ and $u > u'$), consume A rather than B if and only if $EU[A] > EU[B]$.

An outcome measure is *EU-sufficient* if and only if the outcome measure is sufficient to compare EU[A] and EU[B], for given a, b, u, and u′. An outcome measure is *EU-insufficient* if and only if it is not EU-sufficient (i.e., if and only if the outcome measure is insufficient to compare EU[A] and EU[B],

4. These a's and b's, which denote the costs of a treatment, should not be conflated with those from the introduction.

https://doi.org/10.1086/693930 Published online by Cambridge University Press

for given a, b, u, and u′). If an outcome measure is EU-sufficient then there is a strong pro tanto reason for requiring its use in measuring the effectiveness of medical interventions, and conversely, if an outcome measure is EU-insufficient then there is a strong pro tanto reason against its use in measuring the effectiveness of medical interventions. We now prove that ARR and NNT are EU-sufficient and RR and RRR are EU-insufficient.[5]

4.1. ARR and NNT Are EU-Sufficient. With the above approach, we can calculate the expected utility of treatment A and B as

$$\begin{aligned} \mathrm{EU[A]} &= \mathrm{P(Y|A)u} + \mathrm{P(\sim Y|A)u'} - \mathrm{a} \\ &= \mathrm{P(Y|A)u} + [1 - \mathrm{P(Y|A)}]\mathrm{u'} - \mathrm{a} \\ &= \mathrm{P(Y|A)}\left(\mathrm{u} - \mathrm{u'}\right) + \mathrm{u'} - \mathrm{a}. \end{aligned}$$

$$\begin{aligned} \mathrm{EU[B]} &= \mathrm{P(Y|B)u} + [1 - \mathrm{P(Y|B)}]\mathrm{u'} - \mathrm{b} \\ &= \mathrm{P(Y|B)}\left(\mathrm{u} - \mathrm{u'}\right) + \mathrm{u'} - \mathrm{b}. \end{aligned}$$

The expected utility of consuming A rather than consuming B is

$$\begin{aligned} \mathrm{EU[A]} - \mathrm{EU[B]} &= \mathrm{P(Y|A)}\left(\mathrm{u} - \mathrm{u'}\right) + \mathrm{u'} - \mathrm{a} - \left[\mathrm{P(Y|B)}\left(\mathrm{u} - \mathrm{u'}\right) + \mathrm{u'} - \mathrm{b}\right] \\ &= [\mathrm{P(Y|A)} - \mathrm{P(Y|B)}]\left(\mathrm{u} - \mathrm{u'}\right) - (\mathrm{a} - \mathrm{b}). \end{aligned}$$

Note that ARR appears as the leftmost multiplicand in this term.

Thus,

$$\mathrm{EU[A]} - \mathrm{EU[B]} > 0 \text{ iff } [\mathrm{P(Y|A)} - \mathrm{P(Y|B)}]\left(\mathrm{u} - \mathrm{u'}\right) - (\mathrm{a} - \mathrm{b}) > 0,$$

which, assuming $\mathrm{u} \neq \mathrm{u'}$, is equivalent to

$$\mathrm{EU[A]} > \mathrm{EU[B]} \text{ iff } [\mathrm{P(Y|A)} - \mathrm{P(Y|B)}] > \frac{\mathrm{a} - \mathrm{b}}{\mathrm{u} - \mathrm{u'}}.$$

Note that ARR appears on the left side of this inequality, and the right side of the inequality is fully determined by a, b, u, and u′. So, given a, b, u, u′, and ARR, one can determine whether EU[A] > EU[B]. Thus, ARR is EU-sufficient, and one should consume A if and only if ARR > (a − b)/(u − u′). The same result holds for NNT, which is just the inverse of ARR:

$$\mathrm{EU[A]} > \mathrm{EU[B]} \text{ iff } \frac{1}{\mathrm{NNT}} > \frac{\mathrm{a} - \mathrm{b}}{\mathrm{u} - \mathrm{u'}}.$$

5. In our derivation, we estimate the conditional probabilities by the observed relative frequencies (see table 1). This is a significant idealization, but it does not affect the argument that follows. For discussion, see Stegenga (2015).

https://doi.org/10.1086/693930 Published online by Cambridge University Press

4.2. RR and RRR Are EU-Insufficient. Assume without loss of generality that RR > 1. Above we showed that

$$\mathrm{EU[A]} > \mathrm{EU[B]} \text{ iff } [\mathrm{P(Y|A)} - \mathrm{P(Y|B)}] > \frac{a - b}{u - u'}.$$

Note that

$$\mathrm{P(Y|A)} - \mathrm{P(Y|B)} = \mathrm{P(Y|B)}\left(\frac{\mathrm{P(Y|A)}}{\mathrm{P(Y|B)}} - 1\right),$$

which is equivalent to

$$\mathrm{P(Y|A)} - \mathrm{P(Y|B)} = \mathrm{P(Y|B)}(\mathrm{RR} - 1),$$

and so

$$\mathrm{EU[A]} > \mathrm{EU[B]} \text{ iff } \mathrm{P(Y|B)}(\mathrm{RR} - 1) > \frac{a - b}{u - u'},$$

which is, given RR > 1, equivalent to

$$\mathrm{EU[A]} > \mathrm{EU[B]} \text{ iff } \mathrm{P(Y|B)} > \frac{a - b}{(u - u')(\mathrm{RR} - 1)}. \tag{1}$$

Thus, # holds that one should consume A rather than B if and only if $\mathrm{P(Y|B)} > (a - b)/(u - u')(\mathrm{RR} - 1)$. Note that a given RR does not constrain the values that P(Y|B) can take in the interval [0,1], nor do the values of a, b, u, or u′. So, for any particular value of RR we consider two cases:

i) $\mathrm{P(Y|B)} = (a - b)/[(u - u')(\mathrm{RR} - 1)] - \varepsilon$
ii) $\mathrm{P(Y|B)} = (a - b)/[(u - u')(\mathrm{RR} - 1)] + \varepsilon$

for some ε that is suitably small such that P(Y|B) remains bounded between 0 and 1. Now consider both cases separately:

Case i: $\mathrm{P(Y|B)} < (a - b)/[(u - u')(\mathrm{RR} - 1)]$, and thus $\mathrm{EU[A]} < \mathrm{EU[B]}$.

Case ii: $\mathrm{P(Y|B)} > (a - b)/[(u - u')(\mathrm{RR} - 1)]$, and thus $\mathrm{EU[A]} > \mathrm{EU[B]}$.

So, if given a, b, u, u′, and RR, one cannot determine whether $\mathrm{EU[A]} > \mathrm{EU[B]}$. Thus, RR is EU-insufficient, which means that decisions based on RR may not have maximal expected utility, depending on the values of P(Y|B). The same result can be shown for RRR = RR − 1. We simply rewrite (1) as

$$\mathrm{EU[A]} > \mathrm{EU[B]} \text{ iff } \mathrm{P(Y|B)} > \frac{a - b}{(u - u') \times \mathrm{RRR}}.$$

https://doi.org/10.1086/693930 Published online by Cambridge University Press

Again, for any particular RRR, we can consider two cases:

i) $P(Y|B) = [(a - b)/[(u - u') \times RRR] - \varepsilon$
ii) $P(Y|B) = [(a - b)/[(u - u') \times RRR] + \varepsilon$

for some ε that is suitably small such that $P(Y|B)$ remains bounded between 0 and 1. Now consider both cases separately:

Case i: $P(Y|B) < (a - b)/[(u - u') \times RRR]$, and thus $EU[A] < EU[B]$.

Case ii: $P(Y|B) > (a - b)/[(u - u') \times RRR]$, and thus $EU[A] > EU[B]$.

Thus, RRR cannot determine alone, given a, b, u, and u′, whether A has a higher expected utility than B. This concludes the proof of the EU-insufficiency of RR and RRR.[6] But, decisions based on ARR and NNT will always pick the intervention with the higher expected utility.

Hence, relative outcome measures may be useful for attributing outcomes to causal factors, but they are not suitable for making choices that are supposed to maximize the expected utility of a future patient. This demonstrates once more the special status of absolute outcome measures such as ARR and NNT. In practice, a, b, u, and u′ may be unknown or a matter of contention, but it is important that we are in principle able to base a rational decision on the value of an outcome measure.

5. The Causal Strength Argument. The various outcome measures can also be regarded as a quantification of statistical effect size or as measures of the causal strength of the link between a treatment and an effect. Indeed, the literature on probabilistic causation often quantifies the strength of a causal link by comparing two conditional probabilities: the probability of an effect given the putative cause, $P(Y|E)$, and the probability of the same effect given the absence of the cause, $P(Y|C)$ (Suppes 1970; Cartwright 1979; Eells 1991; Fitelson and Hitchcock 2011). We can interpret outcome measures in medicine as measures of the causal strength between treatment and recovery. After all, medical trials try to answer questions about the causal effectiveness of interventions.

Our argument in this section draws on two observations: (1) ARR, NNT, and derived absolute outcome measures combine assessments of causal strength in an intuitive way, and (2) RR, RRR, and derived relative measures misrepresent the causal strength of an intervention for a conjunction of unrelated effects. For a detailed axiomatic investigation of probabilistic causal strength measures for binary outcomes, see Sprenger (forthcoming).

6. EU-insufficiency can also be demonstrated for another relative outcome measure, the odds ratio OR (proof omitted; see n. 1 for a definition).

https://doi.org/10.1086/693930 Published online by Cambridge University Press

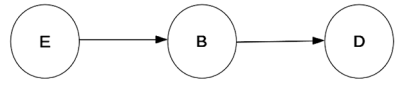


Figure 1. Causal relationship between three variables represented as a directed acyclical graph.

The first observation deals with combining assessments of causal strength along a single path. In causal inference, we often have to deal with *mediators*: variables that transfer an effect from an intervention to an observed effect. For instance, frequent exercise (E) has an effect on the occurrence of cardiovascular diseases (D) via various intermediate properties such as one's blood pressure (B). (Assume for the sake of simplicity that this is the only link between exercise and cardiovascular diseases; see fig. 1.) Now, it is often desirable to combine assessments of causal strength along a causal graph such as the one in figure 1 in a natural manner. That is, the strength of the two causal links between exercise and blood pressure, and between blood pressure and cardiovascular diseases, should be sufficient to determine the overall causal strength of the relationship between exercise and cardiovascular diseases. In other words, there is a function f such that for any measure c of causal strength $c(E, D) = f(c(E, B), c(B, D))$. One may also demand that $c(E, D) \leq c(E, B)$ and $c(E, D) \leq c(B, D)$: the presence of intermediate variables does not increase causal strength. A natural function that satisfies this requirement and several other ones on combining causal strength is $ARR(E, D) = P(D|E) - P(D| \sim E)$ (see also Good 1961). In fact, $ARR(E, D) = ARR(E, B) \times ARR(B, D)$, which allows for a particularly simple calculation of overall causal strength as a function of the strength of individual links. Similarly, $NNT(E, D) = NNT(E, B) \times NNT(B, D)$.

Proof for ARR

$$
\begin{aligned}
ARR(E, D) &= P(D|E) - P(D| \sim E) \\
&= P(D|BE)P(B|E) + P(D| \sim BE)P(\sim B|E) - P(D|B \sim E)P(B| \sim E) \\
&\quad - P(D| \sim B \sim E)P(\sim B| \sim E) \text{ [by the law of total probability]} \\
&= P(D|B)P(B|E) + P(D| \sim B)P(\sim B|E) - P(D|B)P(B| \sim E) \\
&\quad - P(D| \sim B)P(\sim B| \sim E) \text{ [by the causal structure of the graph]} \\
&= P(D|B)[P(B|E) - P(B| \sim E)] + P(D| \sim B)[P(\sim B|E) - P(\sim B| \sim E)] \\
&= P(D|B)[P(B|E) - P(B| \sim E)] - P(D| \sim B)[P(B|E) - P(B| \sim E)] \\
&= [P(D|B) - P(D| \sim B)] \times [P(B|E) - P(B| \sim E)] \\
&= ARR(B, D) \times ARR(E, B).
\end{aligned}
$$

https://doi.org/10.1086/693930 Published online by Cambridge University Press

The analogous result for NNT follows easily from the equality NNT = 1/ARR: NNT(E, D) = 1/ARR(E, D) = 1/(ARR(B, D) × ARR(E, B)) = NNT(B, D) × NNT(E, D). QED

Other measures, such as RR and RRR, do not have this property: for these measures, the overall causal strength is not a function of the measures of the individual causal links. One can demonstrate that derivatives of ARR are the only measures of causal strength that satisfy this property together with the relatively uncontroversial constraint that causal strength is a function on P(Y|E) and P(Y|C), where E and C denote an experimental intervention and a control intervention, respectively (Sprenger, forthcoming, theorem 2).

The second observation deals with composite effects. Imagine that an intervention E (e.g., blood pressure lowering medication) has a certain effect on the occurrence of a binary event Y (e.g., a heart attack). Now suppose that we want to quantify the effect of E on the conjunction of Y and an event Z that is independent of both E and Y (e.g., frequent migraine attacks). Although E does nothing to reduce the risk of Z, the causal effect of E on Y&Z is as great as the causal effect of E for Y, according to RR and RRR. It can be shown that RR, RRR, and their derivatives are the only outcome measures that have this property (Sprenger, forthcoming, theorem 3).

Proof for RR. Suppose that Z is an effect that is independent of the intervention E. Suppose also that Y and Z are independent conditional on E. Then P(YZ|E) = P(Y|E) × p(Z|E) = P(Y|E) × p(Z). The same holds for C = ~ E: P(YZ|C) = P(Y|C) × P(Z|C) = P(Y|C) × p(Z). Hence, RR(E, YZ) = P(Y|E)/P(Y|C) = RR(E, Y). Since RRR = RR − 1, the same results hold for RRR, too. QED

This property is likely to mislead physicians and patients because a nonexistent causal relationship is suggested. It also opens the door to the manipulation of the presentation of trial outcomes. Therefore, ARR and NNT should be preferred to RR, RRR, and other relative outcome measures.

6. Conclusion. We have argued for the superiority of absolute over relative outcome measures. Unfortunately, relative measures are widely employed in clinical research, and absolute measures are underused. Our arguments show this to be a mistake and call for a change of this practice. Some clinical scientists, statisticians, and philosophers have claimed that absolute measures are superior to relative measures, and in this article we provide a principled justification for this view.

We have made a cumulative case for this conclusion. The argument from cognitive bias contends that using the absolute risk reduction ARR instead of the relative risk reduction RRR or other relative outcome measures de-

https://doi.org/10.1086/693930 Published online by Cambridge University Press

creases the chance of overestimating treatment effects and committing the reference class fallacy. The decision-theoretic argument demonstrates that absolute measures are necessary and sufficient (when given pertinent costs and utilities) to choose between intervention options according to dictates of decision theory, while relative measures are insufficient in this regard. Finally, the causal strength argument shows that ARR possesses a natural interpretation as a measure of causal strength between an intervention and an observed result and that it has several properties that distinguish it as such a measure. By contrast, relative outcome measures fail to combine causal strength assessments satisfactorily, and they fail to detect when interventions only affect one instead of several outcomes of interest.

While each single argument may be sufficient to establish the superiority of ARR and its derivatives over relative measures, we consider the cumulative case to be particularly compelling. Medical science, whether in clinical trials or in epidemiology, should always use and report absolute outcome measures.

REFERENCES

Bobbio, M., B. Demichelis, and G. Giustetto. 1994. "Completeness of Reporting Trial Results: Effect on Physicians' Willingness to Prescribe." *Lancet* 343 (8907): 1209–11.

Cartwright, N. 1979. "Causal Laws and Effective Strategies." *Noûs* 13:419–37.

Deeks, J. 1998. "When Can Odds Ratios Mislead?" *British Medical Journal* 317:1155–56.

Eells, E. 1991. *Probabilistic Causality*. Cambridge: Cambridge University Press.

Fitelson, B., and C. Hitchcock. 2011. "Probabilistic Measures of Causal Strength." In *Causality in the Sciences*, ed. P. M. Illari, F. Russo, and J. Williamson, 600–627. Oxford: Oxford University Press.

Forrow, L., W. C. Taylor, and R. M. Arnold. 1992. "Absolutely Relative: How Research Results Are Summarized Can Affect Treatment Decisions." *American Journal of Medicine* 92 (2): 121–24.

Frick, M. H., et al. 1987. "Helsinki Heart Study: Primary-Prevention Trial with Gemfibrozil in Middle-Aged Men with Dyslipidemia; Safety of Treatment, Changes in Risk Factors, and Incidence of Coronary Heart Disease." *New England Journal of Medicine* 317 (20): 1237–45.

Good, I. J. 1961. "A Causal Calculus." Pt. 1. *British Journal for the Philosophy of Science* 11:305–18.

HPSCG (Heart Protection Study Collaborative Group). 2002. "MRC/BHF Heart Protection Study of Cholesterol Lowering with Simvastatin in 20,536 High-Risk Individuals: A Randomised Placebo-Controlled Trial." *Lancet* 360 (9326): 7–22.

Hux, Janet E., and C. David Naylor. 1995. "Communicating the Benefits of Chronic Preventive Therapy: Does the Format of Efficacy Data Determine Patients' Acceptance of Treatment?" *Medical Decision Making* 15:152–57.

King, Nicholas B., Sam Harper, and Meredith E. Young. 2012. "Use of Relative and Absolute Effect Measures in Reporting Health Inequalities: Structured Review." *BMJ* 345. doi:10.1136/bmj.e5774.

Malenka, David, John Baron, Sarah Johansen, Jon Wahrenberger, and Jonathan Ross. 1993. "The Framing Effect of Relative and Absolute Risk." *Journal of General Internal Medicine* 8 (10): 543–48.

Nexøe, Jørgen, Dorte Gyrd-Hansen, Jakob Kragstrup, Ivar Sønbø Kristiansen, and Jesper Bo Nielsen. 2002. "Danish GPs' Perception of Disease Risk and Benefit of Prevention." *Family Practice* 19 (1): 3–6.

https://doi.org/10.1086/693930 Published online by Cambridge University Press

Northridge, M. E. 1995. "Public Health Methods-Attributable Risk as a Link between Causality and Public Health Action." *American Journal of Public Health* 85:1202–4.
Nurminen, Markku. 1995. "To Use or Not to Use the Odds Ratio in Epidemiologic Analyses." *European Journal of Epidemiology* 11:365–71.
Sorensen, L., D. Gyrd-Hansen, I. S. Kristiansen, J. Nexoe, and J. B. Nielsen. 2008. "Laypersons' Understanding of Relative Risk Reductions: Randomised Cross-Sectional Study." *BMC Medical Informatics and Decision Making* 8 (31). doi:10.1186/1472-6947-8-31.
Sprenger, J. Forthcoming. "Foundations for a Probabilistic Theory of Causal Strength." *Philosophical Review*.
Stegenga, Jacob. 2015. "Measuring Effectiveness." *Studies in History and Philosophy of Biological and Biomedical Sciences* 54:62–71.
Suppes, P. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.
Walter, S. D. 1976. "The Estimation and Interpretation of Attributable Risk in Health Research." *Biometrics* 32:829–49.
Worrall, John. 2010. "Do We Need Some Large, Simple Randomized Trials in Medicine?" In *EPSA Philosophical Issues in the Sciences*, ed. Mauricio Suarez, Mauro Dorato, and Miklos Redei. Dordrecht: Springer.

https://doi.org/10.1086/693930 Published online by Cambridge University Press