Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Theoretical Note

Causal analysis of absolute and relative risk reductions☆

Björn Meder [a,*], Charley M. Wu [b,c,d,e,f], Felix G. Rebitschek [g]

[a] Institute for Mind, Brain, and Behavior, Health and Medical University, Potsdam, Germany
[b] Human and Machine Cognition Lab, University of Tübingen, Tübingen, Germany
[c] Max Planck Institute for Human Development, Berlin, Germany
[d] Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[e] Centre for Cognitive Science, Institute of Psychology, Technical University of Darmstadt, Darmstadt, Germany
[f] Hessian.AI, Darmstadt, Germany
[g] Harding Center for Risk Literacy, Faculty of Health Sciences Brandenburg, University of Potsdam, Potsdam, Germany

ARTICLE INFO

Keywords:
Causal models
Relative risk reduction
Absolute risk reduction
Causal power
Causal Bayes net
Causal modeling
Risk communication

ABSTRACT

Any medical innovation must first prove its benefits with reliable evidence from clinical trials. Evidence is commonly expressed using two metrics, summarizing treatment benefits based on either absolute risk reductions (ARRs) or relative risk reductions (RRRs). Both metrics are derived from the same data, but they implement conceptually distinct ideas. Here, we analyze these risk reductions measures from a causal modeling perspective. First, we show that ARR is equivalent to ΔP, while RRR is equivalent to causal power, thus clarifying the implicit causal assumptions. Second, we show how this formal equivalence establishes a relationship with causal Bayes nets theory, offering a basis for incorporating risk reduction metrics into a computational modeling framework. Leveraging these analyses, we demonstrate that under dynamically varying baseline risks, ARRs and RRRs lead to strongly diverging predictions. Specifically, the inherent assumption of a linear parameterization of the underlying causal graph can lead to incorrect conclusions when generalizing treatment benefits (e.g, predicting the effect of a vaccine in new populations with different baseline risks). Our analyses highlight the shared principles underlying risk reduction metrics and measures of causal strength, emphasizing the potential for explicating causal structure and inference in medical research.

1. Causal analysis of relative and absolute risk reductions

During the Covid-19 pandemic, the quantification of risk posed an important challenge. How dangerous was it to go to the grocery store in person vs. paying extra to have them delivered to your doorstep? How much safer was it to go to crowded places once you had one or two vaccines? We all struggled with these dilemmas, while medical professions fought to communicate accurate information to the public. Assessing and communicating the effectiveness of vaccinations and treatments against Covid-19 made this challenge even more complex, further complicating the public's understanding of risk and benefits.

This struggle highlights a broader issue in public health: the quantification and communication of treatment effects. Quantifying the causal impact of treatments and communicating medical information in an effective manner remains a key challenge for public policy and informed decision making (Bonner et al., 2021; Gigerenzer & Edwards, 2003; Gigerenzer et al., 2007; Woloshin et al., 2023). Two widely used metrics for quantifying treatment benefits are relative risk reductions (RRR) and absolute risk reductions (ARR), which convey information about risk changes, such as the reduced risk of disease for vaccinated individuals or the reduced risk of recurrence for patients undergoing chemotherapy. These measures play a crucial role in summarizing study outcomes, in the communication of treatment benefits to healthcare professionals and the public, as well as tools for generalizing medical effects to new contexts and populations.

Here, we highlight central connections between risk reduction measures used in medical research and the theoretical framework of causal Bayes nets used to study causal learning. By analyzing ARR and RRR through the lens of causal modeling, we clarify the assumptions that each measure makes about the underlying data-generating processes, with different structural and inferential implications.

First, we demonstrate that ARR is mathematically equivalent to ΔP (Cheng & Novick, 1992; Waldmann & Holyoak, 1992), while RRR

☆ CMW is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1-390727645.

* Corresponding author.
E-mail address: bjoern.meder@hmu-potsdam.de (B. Meder).

https://doi.org/10.1016/j.jmp.2025.102942
Received 23 October 2024; Received in revised form 17 June 2025; Accepted 14 August 2025

is equivalent to causal power (Cheng, 1997; Novick & Cheng, 2004), which are two central measures for defining causal strength. Thus, metrics from the causal modeling literature are incidentally used under different labels in medical research and risk communication, yet these connections are not well known (but see Sprenger, 2018; Stephan et al., 2021).

Second, we show how this equivalence establishes connections with the formalism of causal Bayes nets (Pearl, 2000; Pearl & Mackenzie, 2018; Spirtes et al., 2000), offering a basis for incorporating risk reductions into a causal modeling framework. Specifically, we formalize how ARR and RRR instantiate distinct parameterizations of the same causal structure within the framework of causal graphical models, offering a unified computational perspective on treatment effect measures. This integration reveals how the choice of a particular risk reduction measure inevitably corresponds to different assumptions about how causes probabilistically influence an outcome.

Third, we highlight the diverging consequences of employing either ARR or RRR to measure treatment effects when generalizing study results to scenarios with different baseline risks. These analyses offer a novel perspective on the normative merits of each measure (Hoefer & Krauss, 2021; Jäntgen, 2023; Sprenger & Stegenga, 2017; Stegenga, 2015), and contribute to discussions about the consequences of using different metrics for quantifying and generalizing treatment effects (Colnet et al., 2024; Huitfeldt et al., 2018; Murad et al., 2024; Xiao et al., 2022).

2. From risk reductions to measures of causal strength

The purpose of risk reduction metrics is to quantitatively summarize treatment effects, such as the impact of a vaccine on disease cases. This problem is both conceptually and mathematically related to the problem of causal induction: How can we induce unobservable causal relations from observed data and quantify their strength? The nature of the inference mechanisms that support the development of both intuitive and scientific theories about the world is a key issue in philosophy (Cartwright, 2007), psychology (Waldmann, 2017), statistics (Pearl et al., 2016), and machine learning (Schölkopf, 2022). Because medical interventions naturally refer to cause–effect relations, researchers have also emphasized the importance of causal inference in the medical domain (Etminan et al., 2020; Greenland et al., 1999; Prosperi et al., 2020; Sanchez et al., 2022; Stovitz & Shrier, 2019). Contributing to this growing body of work (Bareinboim & Pearl, 2016; Digitale et al., 2022; Kyriacou et al., 2023), we offer a causal analysis of risk reduction metrics, highlighting how they map onto models from the literature on human causal induction and clarifying the underlying assumptions and inferential implications.

2.1. Measuring treatment effects: Absolute and relative risk reductions

Absolute and relative risk reductions are measures for quantifying changes in the risk associated with the occurrence of an adverse binary health outcome (e.g., disease present or absent). Applied to an experimental design with two groups (treatment vs. control), one way to quantify treatment effects is in terms of their absolute risk reduction (ARR): the arithmetic difference of the probability of the adverse outcome under the treatment (e.g., vaccine) vs. the control group (e.g., placebo):

$$ARR = P(\text{event}|\text{control}) - P(\text{event}|\text{treatment}) \quad (1)$$

In applied settings, these probabilities and their difference are typically expressed as percentages. Consider the Covid-19 pandemic, where different vaccines were evaluated in randomized controlled trials (e.g., Baden et al., 2021; Polack et al., 2020; Voysey et al., 2021). For instance, in the BNT162b2 (Pfizer-BioNTech) trial, 162 out of 18,325 people (0.88%) in the placebo group developed symptomatic Covid-19, compared to 8 out of 18,198 people (0.04%) in the vaccine group (Polack et al., 2020). The absolute risk reduction is the difference between these event rates: 0.88%–0.04% = 0.84 percentage points (Table 1). This measure provides a quantitative estimate of the vaccine's ability to reduce disease cases.

Table 1
Data from the BNT162b2 vaccine trial against Covid-19 (Polack et al., 2020). Shown are registered (symptomatic) Covid-19 cases in the vaccine and placebo group. Computed are the probability of Covid-19 in each group, absolute risk reduction (ARR), and relative risk reduction (RRR).

	Covid-19	No Covid-19	*P*(Covid-19)	ARR	RRR
Control: Placebo	162	18,163	0.88%	0.84%	95%
Treatment: Vaccine	8	18,190	0.04%		

Another way of expressing treatment effects is the *relative risk reduction* (RRR). The RRR normalizes the arithmetic difference (i.e., ARR) on the probability of the disease in the control group:

$$RRR = \frac{P(\text{event}|\text{control}) - P(\text{event}|\text{treatment})}{P(\text{event}|\text{control})} \quad (2)$$

In the BNT162b2 trial, expressed as percentage, the RRR is 95%, as it reduces the number of infections from 162 cases in the placebo group to 8 cases in the vaccine group. Risk reductions and treatment benefits are frequently expressed in such relative terms. For instance, in vaccine epidemiology the *efficacy* of a vaccine is commonly defined as the RRR estimated from a randomized placebo-controlled study (Greenwood & Yule, 1915; Tentori et al., 2021; Weinberg & Szilagyi, 2010).

While ARRs and RRRs are calculated from the same data, their values can vary greatly. For instance, a 50% RRR could correspond to an ARR of 30% if the event rate decreases from 60% to 30%. But, it could also correspond to an absolute decrease of 1%, if the event rate decreases from 2% to 1%. The former would indicate a substantial effect, while the latter might be negligible in practice because the absolute risk of experiencing the undesirable event is low even without the treatment. This distinction highlights the importance of absolute measures and baseline risks in decision-making, as emphasized in the medical decision-making literature (Bodemer et al., 2014; Gigerenzer et al., 2007; Jäntgen, 2023; Natter & Berry, 2005; Sheridan et al., 2003; Sprenger & Stegenga, 2017). Conversely, the same ARR can imply very different RRRs, depending on the baseline risk (e.g., Bruynesteyn et al., 2004).

Because RRRs are typically numerically larger than ARRs, especially when the baseline risk is low, they may create a disproportionate impression of the treatment benefits, which has lead several researchers to argue that RRRs can be misleading when communicating health information (Ancker et al., 2006; Gigerenzer & Edwards, 2003; Gigerenzer et al., 2007; Hembroff et al., 2004; Madrigal et al., 2024; Schwartz et al., 1997; Sprenger & Stegenga, 2017). Guidelines for reporting treatment effects in RCTs and systematic reviews therefore recommend communicating both ARRs and RRRs (Guyatt, Oxman, Akl et al., 2011; Guyatt, Oxman, Kunz et al., 2011; Higgins et al., 2023; Moher et al., 2010), reflecting ongoing debates about the relative merits of each measure, such as their stability (portability) across studies and populations (Deeks, 2002; Doi et al., 2022; Furukawa et al., 2002; Murad et al., 2024; Schmid et al., 1998; Xiao et al., 2022). Parallel concerns have been raised in philosophy of science, where several authors have advanced normative arguments in favor of absolute measures (Colnet et al., 2024; Jäntgen, 2023; Sprenger & Stegenga, 2017; Stegenga, 2015). These debates about effect measures reflect a foundational issue central to many disciplines: how to appropriately quantify causal influence.

2.2. Measuring causal strength: Δp and causal power

ARR and RRR represent two distinct ways of expressing the impact of a treatment on an outcome. Many other fields have engaged in

similar debates about how to quantify and interpret causal relationships, leading to different proposals on measuring causal strength. Two central metrics for quantifying causal relations are ΔP and *causal power*. Applied to assessing the strength of a causal relation between a candidate cause C and an effect E, ΔP formalizes causal strength as the difference of the likelihood of the effect given the presence and absence of the cause, respectively:

$$\Delta P = P(e|c) - P(e|\neg c) \tag{3}$$

where $P(e|c)$ is the probability of the effect given that the cause is present, and $P(e|\neg c)$ is the probability of the effect given that the cause is absent. If the cause increases the likelihood of the effect, ΔP is positive (e.g., smoking increases the likelihood of lung cancer). Conversely, if the cause decreases the likelihood of the effect, ΔP is negative (e.g., a vaccination decreases the probability of disease). If the presence of the putative cause does not change the likelihood of the effect, ΔP is zero (i.e., an ineffective treatment).

Causal power (Cheng, 1997; Glymour, 2003; Novick & Cheng, 2004), denoted as q_c, provides an alternative measure of causal strength corresponding to the probability that C generates or prevents E in the absence of alternative influences on the effect. Typically, this quantity does not coincide with the probability of the effect in the presence of C, because this empirical probability also includes the influence of other (unobserved) causes. For preventive causes ($\Delta P < 0$; see Appendix A.3 for the generative case), causal power is defined as:

$$q_c = \frac{P(e|c) - P(e|\neg c)}{P(e|\neg c)} = \frac{-\Delta P}{P(e|\neg c)} \tag{4}$$

Because causal power denotes a theoretical probability — that is, the probability that a cause would generate or prevent the effect in the absence of other influencing factors — it is always strictly non-negative and ranges between 0 and 1. Like ΔP, causal power is zero if the cause neither reduces nor raises the probability of the effect. For other scenarios, the two measures usually give different values (see Appendix A.1 in the Appendix).

2.3. Equivalence of measures of causal strength and risk reductions

Measures of risk reduction and measures of causal strength have been proposed in different fields, with distinct problems and applications in mind. In the medical literature, risk reductions are used to measure treatment benefits. In the cognitive science and causal modeling literature, measures of causal strength quantify the magnitude of causal relations and provide normative benchmarks for human causal learning (Buehner et al., 2003; Griffiths & Tenenbaum, 2005; Lober & Shanks, 2000; Meder et al., 2014; Waldmann & Holyoak, 1992). The terminology differs across fields, even though the employed measures are formally equivalent.

In medical research, common terms include "treatment and control group" "event rates", "baseline risk", and "efficacy" of treatments. Conversely, the causal modeling literature uses the generic terms "cause" and "effect" to refer to different conditions and explain how people infer causal relations from observed frequencies, with recurring debates on how the "strength" of the relations can be expressed formally.

Despite these differences in nomenclature, the mapping of concepts is straightforward, especially in experimental designs. The treatment group corresponds to instances in which the candidate cause C is present (e.g., vaccine), whereas the control group corresponds to instances in which the candidate cause is absent (e.g., placebo). The effect E is the undesirable event (e.g., disease) that the treatment is supposed to prevent. Accordingly, the definition of ARR aligns with the ΔP measure (Eqs. (1) and (3)):

$$\begin{aligned} ARR &= P(\text{event}|\text{control}) - P(\text{event}|\text{treatment}) \\ &= P(e|\neg c) - P(e|c) \\ &= |\Delta P| \end{aligned} \tag{5}$$

In the causal literature, ΔP is consistently defined as $P(e|c) - P(e|\neg c)$, yielding negative values for ΔP in case of a preventive cause that reduces the likelihood of the effect, and positive values for generative causes that increase the likelihood of the effect. By contrast, in the medical literature, both reductions and increases are commonly expressed as non-negative numbers (often percentages), with the direction indicated by the verbal label: $P(e|c) < P(e|\neg c)$ indicates a risk reduction, and $P(e|c) > P(e|\neg c)$ signifies a risk increase. Otherwise, however, the metrics are defined identically.

Analogously, RRR is mathematically equivalent to causal power in preventive scenarios (Eq. (2) and (4)):

$$\begin{aligned} RRR &= \frac{P(\text{event}|\text{control}) - P(\text{event}|\text{treatment})}{P(\text{event}|\text{control})} \\ &= \frac{P(e|\neg c) - P(e|c)}{P(e|\neg c)} \\ &= \frac{|\Delta P|}{P(e|\neg c)} \\ &= q_c \end{aligned} \tag{6}$$

Thus, while each field employs these measures for distinct purposes, using unique labels and terminology, they are mathematically identical. Of course, formal equivalence does not imply that all risk reductions are causal. In experimental studies, randomization establishes independence of the cause from other factors influencing the effect, whereas non-experimental studies are more susceptible to confounding and other sources of bias. Accordingly, whether a causal interpretation is warranted depends not on the chosen measure, but on the conditions under which it was derived.

Finally, we note that interpreting ARR and RRR as measures of causal strength through their equivalence to ΔP and causal power entails a focus on average effects across groups. This interpretation is consistent with medical research, where risk reduction measures quantify group-level effects based on observed event frequencies, as well as psychological research, where ΔP and causal power are used as normative or descriptive models of how people infer general causal tendencies from summary data or frequency information (Buehner et al., 2003; Griffiths & Tenenbaum, 2005; Meder et al., 2014). This interpretation is also consistent with the causal Bayes net framework adopted in the following section, where causal relations are defined over probability distributions and interventional quantities are typically interpreted as population-level expectations (Pearl, 2000).

3. Counterfactual inference through causal modeling

What are the implications of the equivalence between measures of causal strength on the one hand, and measures of risk reductions on the other hand? Besides highlighting fundamental parallels in the metrics used in different disciplines, the equivalence provides pathways for applying causal modeling techniques to make predictions for novel contexts. For instance, during the Covid-19 pandemic, human challenge trials (Adams-Phipps et al., 2023) deliberately exposed healthy volunteers to the virus (Killingley et al., 2022), representing the extreme of maximal exposure to the pathogen. In contrast, diseases like smallpox, declared eradicated in 1980 (Smith & McFadden, 2002), exemplify the other extreme—contexts with effectively zero baseline risk. These two situations represent the endpoints of a spectrum with varying baseline risk, with some medical problems characterized by fairly stable baseline risks (e.g., hypertension or type 2 diabetes) and others, like Covid-19, being characterized by high volatility, with baseline risk strongly varying across time, populations, geography, and personal behavior.

How can we formally model different scenarios and derive empirically testable predictions if the baseline risk in the novel context is different from the one in which the treatment effect was assessed? We address this question using causal Bayes nets theory, a computational

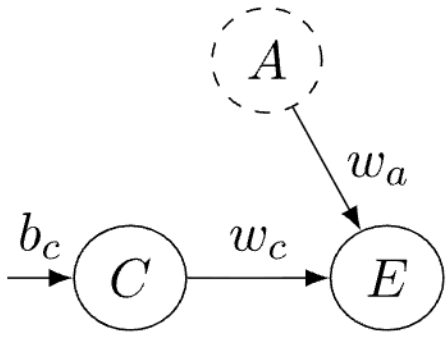

Fig. 1. Basic causal model with a candidate cause C, an effect E, and the composite of unmeasured background causes A. Nodes represent domain variables and directed edges denote causal relations. Parameters b_c, w_a, and w_c quantify the strength of the relations.

framework that combines graphical causal models and probability calculus to represent generative causal models. The formalism supports counterfactual generalizations in a systematic way, even when the baseline risk varies. We utilize the framework to reinterpret absolute and relative risk reductions as distinct parameterizations of a causal graph, illustrating how they can lead to diverging conclusions when generalizing treatment effects.

3.1. Causal Bayes nets: Strength and structure

Causal Bayes nets theory provides a formal approach for representing causal relations and modeling probabilistic inferences across causal networks (Pearl, 2000; Pearl et al., 2016; Pearl & Mackenzie, 2018; Spirtes et al., 2000). This approach combines qualitative assumptions about the causal structure of the domain with quantitative estimates specifying the magnitude of the causal relations.

Importantly, the causal Bayes nets framework also offers a causal reinterpretation of relative and absolute risk reductions, explicating their diverging implications when employed to predict treatment outcomes across varying baseline risks. To illustrate, we use the vaccine trial data (Table 1). Fig. 1 shows the basic causal model for representing situations involving a single binary cause C and a single binary effect E, both of which can be present or absent (Griffiths & Tenenbaum, 2005; Meder et al., 2014). The arrow connecting C and E represents the presumed causal dependency — the ability of the vaccine to reduce the probability of disease. In addition, the model contains a node A representing the (unobserved) background causes that generate the effect, that is, different ways of contracting Covid-19. (For mathematical convenience, A is assumed to be constantly present; Griffiths and Tenenbaum (2005).)

The graphical model is complemented by a set of parameters, b_c, w_c. Parameter b_c denotes the marginal probability of cause C. In an RCT with two conditions, b_c will usually be about 0.5, as half of the people are randomly assigned to each condition. In other settings, like non-experimental epidemiological studies, this parameter could vary strongly and would typically be estimated from data. Parameter w_a represents the strength of the background cause A, which captures exogenous influences not explicitly represented in the model. An estimate for this influence is provided by the likelihood of the effect in the absence of the cause, $P(e|\neg c)$ — the baseline risk. For instance, in the BNT162b2 trial, 162 out of 18,325 participants in the placebo group developed Covid, hence $w_a = 0.0088$ (0.88%; Table 1). This value reflects the comparably low rate of Covid cases in this particular study. But, of course, over the course of a pandemic, this probability typically fluctuates significantly across time and populations.

Particularly relevant for the present analysis is the parameter w_c, which denotes the strength of cause C — the ability of the vaccine to reduce disease cases. While b_c and w_a correspond to marginal or conditional probabilities estimated from observable frequencies, w_c is not a probability but a parameter reflecting causal strength. Formally, both ΔP and causal power provide maximum-likelihood estimates for w_c in the causal graph in Fig. 1, but under distinct assumptions about how C and A influence the probability of E (Glymour, 2003; Griffiths & Tenenbaum, 2005). Specifically, ΔP corresponds to the assumption of a linear-additive combination, whereas causal power assumes independent causal influences following a noisy-OR rule (for generative causes) or a noisy-AND-NOT integration rule (for preventive causes) (Pearl, 1988; Yuille & Lu, 2007). Thus, each measure reflects different theoretical assumptions about how causes interact to produce or prevent effects. Because ARR and RRR are equivalent to ΔP and causal power, respectively, it follows that they also provide maximum-likelihood estimates for the causal influence of C.

Using ΔP and, equivalently, ARR as an estimate for w_c corresponds to the assumption that the probability of E is a linear combination of C and A, such that the cause decreases the baseline risk by a constant amount. Accordingly, the probability of E in the presence of C (e.g., probability of disease in vaccinated people) is given by

$$\begin{aligned} P(e|c,a;w_a,w_c) &= a \cdot w_a + c \cdot w_c \\ &= a \cdot w_a + c \cdot \Delta P \\ &= a \cdot w_a + c \cdot ARR \end{aligned} \tag{7}$$

where $a, c \in \{0,1\}$ denote the presence and absence of candidate cause C and background cause A, while w_a and w_c denote their causal influence. According to this linear parameterization, if C is present the probability of effect E occurring is an additive function of w_a and w_c, and reduces to w_a if C is absent. In other words, the probability of the effect (e.g., Covid-19) is determined by subtracting a fixed amount from the baseline risk.

In contrast, using causal power q_c and, equivalently, RRR as an estimate for w_c corresponds to a different parameterization of the graphical model. This approach represents a different structural assumption about how the treatment influences the outcome. Rather than subtracting a fixed amount from the baseline risk, as in the linear-additive model, this parameterization assumes that the treatment reduces the effectiveness of the background cause A proportionally. Mathematically, this corresponds to a noisy-AND-NOT model, a probabilistic generalization of a logical rule in which effect E occurs if A is active and C fails to prevent it (e.g., a vaccine failing to prevent disease upon exposure). Conceptually, this parameterization aligns with the interpretation of RRR as a proportional reduction in risk, and belongs to a broader class of probabilistic functions that capture graded causal influence and interactions in the causal Bayes nets framework (Griffiths & Tenenbaum, 2005; Novick & Cheng, 2004; Yuille & Lu, 2007). Under this assumption, the conditional probability of the effect given the cause is given by

$$\begin{aligned} P(e|c,a;w_a,w_c) &= a \cdot w_a(1 - c \cdot w_c) \\ &= a \cdot w_a(1 - c \cdot q_c) \\ &= a \cdot w_a(1 - c \cdot RRR) \end{aligned} \tag{8}$$

which reflects that C probabilistically inhibits the ability of A to produce the effect, and reduces to w_a for the probability of E when C is absent.

Regardless of whether a linear or noisy-logical parameterization of the graph is chosen, the causal-based factorization of the joint probability distribution $P(C,E)$ recovers the empirical probabilities. For instance, in the BNT162b2 trial, 8 out of 18,198 people in the vaccine group developed Covid-19 (Table 1), thus $P(\text{disease}|\text{vaccine}) = P(e|c) = 0.0004$. This probability remains consistent whether computed directly from the empirical data or derived from the parameterized causal model. Specifically, it remains unchanged when using ΔP (absolute risk reduction) or causal power (relative risk reduction) to estimate w_c in Eqs. (7) and (8), respectively (see Appendix A.2 for a numerical example).

To summarize, given the mathematical equivalence of ARR to ΔP on one hand, and RRR to causal power on the other, this implies that both risk measures are maximum-likelihood estimates for the causal strength parameter w_c of the graph show in Fig. 1, but under different parameterizations. Thus, when viewed through a causal modeling lens, selecting either of these metrics to quantify treatment effects incidentally entails the adoption of distinct assumptions on how causes influence the effect. Importantly, while both parameterizations accurately capture empirical probabilities, they imply strongly diverging predictions for counterfactual inferences about scenarios with baseline risks different from the one under which the treatment effect was originally estimated.

3.2. Generalizing causal strength and treatment effects to new contexts

The parameterized graph defines a generative model that supports causal inferences in a principled way, including predictions about the effects of interventions, that is, reasoning about scenarios for which no data is (yet) available. Formally, such counterfactuals can be calculated by adjusting the model's parameters to reflect a specific context of interest and deriving the desired probability distribution. For example, the parameterized causal model in Fig. 1 supports estimating a vaccine's effect in populations where the baseline risks differs from the original trial. This can be modeled by adjusting the strength of the background cause, w_a, to the desired level and deriving the conditional probability distribution of effect E given the presence and absence of the cause (i.e., predicted disease cases with and without vaccine). For instance, a human challenge trial (Adams-Phipps et al., 2023) where all subjects are exposed to the pathogen could be modeled by setting $w_a = 1$. Conversely, a context with a virus-free population, akin to the conditions following smallpox eradication in the 1970s would be simulated by letting $w_a = 0$. These scenarios represent the two extremes of a spectrum ranging from a theoretical baseline risk of 0% to 100%.

To derive counterfactual predictions, such as the probability of disease under different baseline risks (i.e., different values of w_a), the strength w_c of cause C also needs to be quantified. As noted, both ΔP and causal power provide estimates for w_c, but since they instantiate different assumptions of how A and C interact to generate E, they can yield strongly diverging predictions. In the case of the BNT162b2 vaccine trial (Table 1), the RRR (equivalent to causal power) was approximately 95%, while the ARR (corresponding to ΔP) was around 0.84%. These estimates were obtained under a specific baseline risk (e.g., in the BNT162b2 trial the baseline risk, as estimated from the control group cases, was 0.88%; Table 1).

However, baseline risks can vary substantially across individuals, populations, and contexts — especially, but not exclusively, in the case of rapidly evolving conditions like a pandemic. Consider the scenario where the vaccine is administered to another group of people under conditions where the baseline risk is, say, 10% (e.g., healthcare professionals with greater exposure). Given the original trial, how many cases of disease should we expect? For instance, if we had 1000 healthcare professionals and a 10% baseline risk of getting infected, we would expect about 100 to get infected without vaccines. But what if they were vaccinated? Answering this question requires a causal inference — we need to generalize the original treatment effect to a new context with a different baseline risk.

If we used the ARR, which assumes that the baseline risk is decreased by a constant amount, the vaccine would be expected to decrease the number of disease cases by only 0.84 percentage points, resulting in around 92 cases (9.16%; Eq. (7)). In stark contrast, when using a RRR we would predict a 95% reduction from the new baseline, yielding an expected 5 disease cases among the vaccinated people (0.5%; Eq. (8)).[1] In these scenarios, where RRR correctly captures the preventive power of the vaccines, ARR strongly underestimates its causal impact in terms of reducing disease cases in novel settings.

This divergence between the two measures only becomes larger if the baseline risk further increases. For instance, if the baseline risk in a human challenge study was 100%, using ARR as a measure of vaccine efficacy would predict 99.16% disease cases, whereas a RRR predicts 5% cases. Fig. 2 illustrates this divergence across varying baseline risks, showing the predicted probability of disease for three vaccines against Covid-19, using the absolute or relative risk reduction obtained in the corresponding clinical trials (Baden et al., 2021; Polack et al., 2020; Voysey et al., 2021) as causal strength estimates.

Another problem with using ARR (or equivalently, ΔP) to quantify treatment efficacy is illustrated by cases in which the baseline risk is extremely low or effectively zero, as seen in scenarios like the eradication of smallpox. In such circumstances, the probability of disease under the linear parameterization inherent in the ΔP model (Eq. (7)) erroneously yields a negative probability (Feynman, 1987), although by definition probabilities are limited to the interval [0,1]. Generally, this happens if the baseline risk in the novel context is below the estimate of the absolute reduction. To avoid such cases and to ensure a proper probability distribution, it is necessary to manually constrain the sum of w_c and w_a to the interval [0, 1]. In contrast, the functional form inherent in RRR (and equivalently, in the causal power model) correctly indicates that the probability of the effect is zero. Generally, using causal power and Eq. (8) always yields a valid probability. Thus, form a normative perspective causal power and RRRs are better-behaved measures than the ΔP model and ARRs when making predictions for scenarios with varying baseline risks.

Does using ARR and the ΔP model always lead to an underestimation of the treatment effect under a new baseline risk, compared to using RRR? The answer is no. In fact, we can precisely characterize the circumstances under which ΔP and, equivalently, an ARR, yields lower, higher, or equal probabilities compared to using causal power. Fig. 3 illustrates the diverging implications for three hypothetical scenarios. In each case, the strength of C is estimated under a specific baseline risk (i.e., probability of the effect in the absence of the cause), and then used to make inferences about the likelihood of the effect in situations with new baseline probabilities. The figure highlights several important points.

First, there is always a cross-over of the predicted $P(e|c)$ based on ΔP vs. caus al power. Specifically, the predictions intersect (red dot in Fig. 3) at the originally observed $P(e|c)$ and $P(e|\neg c)$. This is necessarily the case because if the new baseline probability is identical to the original one, both metrics should yield the same $P(e|c)$ — namely recover the original probability (Appendix A.2). Second, left of this point (i.e., when then new baseline probability is *lower* than the original one), the predicted probability of the effect given the cause based on ΔP is lower than the probability predicted by causal power. In this case, ΔP and ARR *overestimate* the treatment effect. Moreover, negative probabilities will result when the new baseline probability is lower than ΔP. Third, right of this point, when the new baseline probability is *higher* than the one under which the strength of C was estimated, the predicted probability of the effect given the cause based on ΔP is higher than the probability predicted by causal power. In this case, using ΔP (ARR) *underestimates* the treatment effect, relative to using RRR.

These analyses demonstrate the application of causal modeling techniques to formally represent and assess hypothetical scenarios through parameter adjustments in the corresponding causal graph. Note that our analyses and predictions based on Eqs. (7) and (8) assume that the treatment effect, quantified using either the ARR or the RRR, remains constant when generalizing to new contexts. While such stability may not always hold empirically (e.g., the efficacy of a vaccine may vary across age groups), our analyses highlight the implications of applying each metric under the assumption that they are stable across contexts. Additionally, the results emphasize that generalizing causal effects to

[1] Note that in practice, we would not expect these exact numbers, due to uncertainty about the statistical estimates and other sources of noise. These considerations, while important in practical applications, are omitted here as they do not touch upon the conceptual argument.

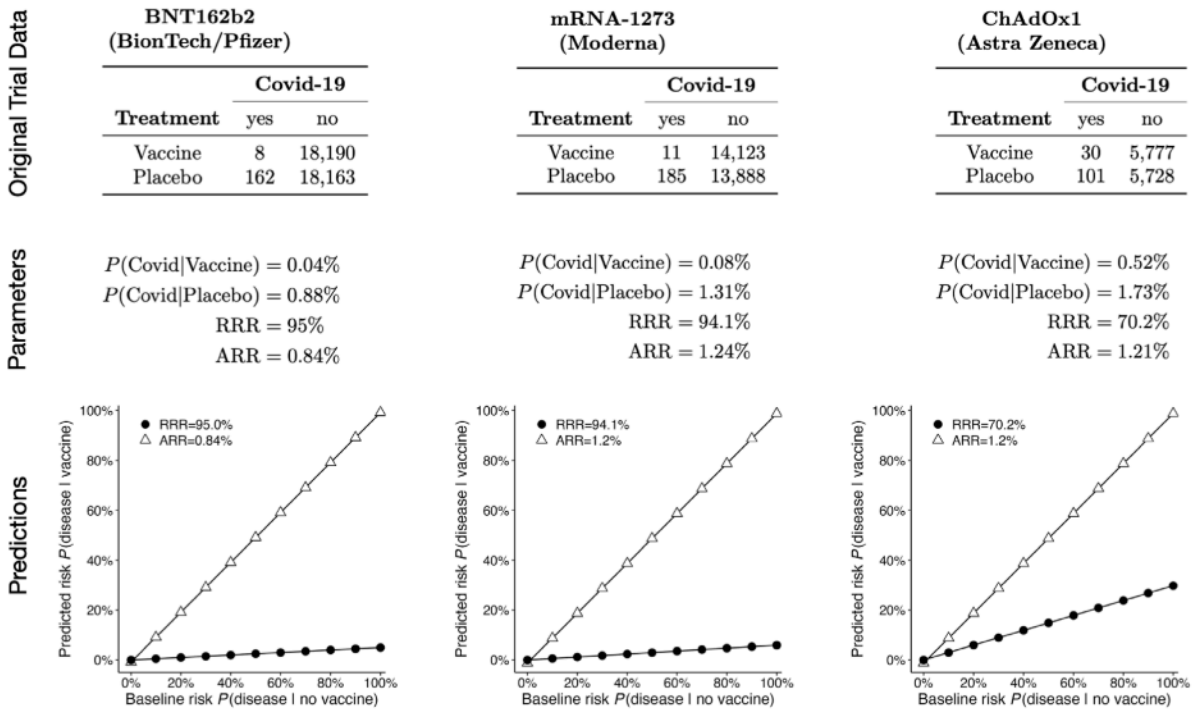

Fig. 2. Counterfactual predictions under varying baseline risks. Top row: Data from three randomized control trials of Covid-19 vaccines (Baden et al., 2021; Polack et al., 2020; Voysey et al., 2021). Middle row: Corresponding probabilities of Covid given vaccine and placebo, and entailed relative (RRR) and absolute risk reduction (ARR). Bottom row: Predicted probability of disease under varying new baseline risks, calculated from a causal model with the treatment effect quantified using the ARR (ΔP) or RRR (causal power) (Eqs. (7) and (8)).

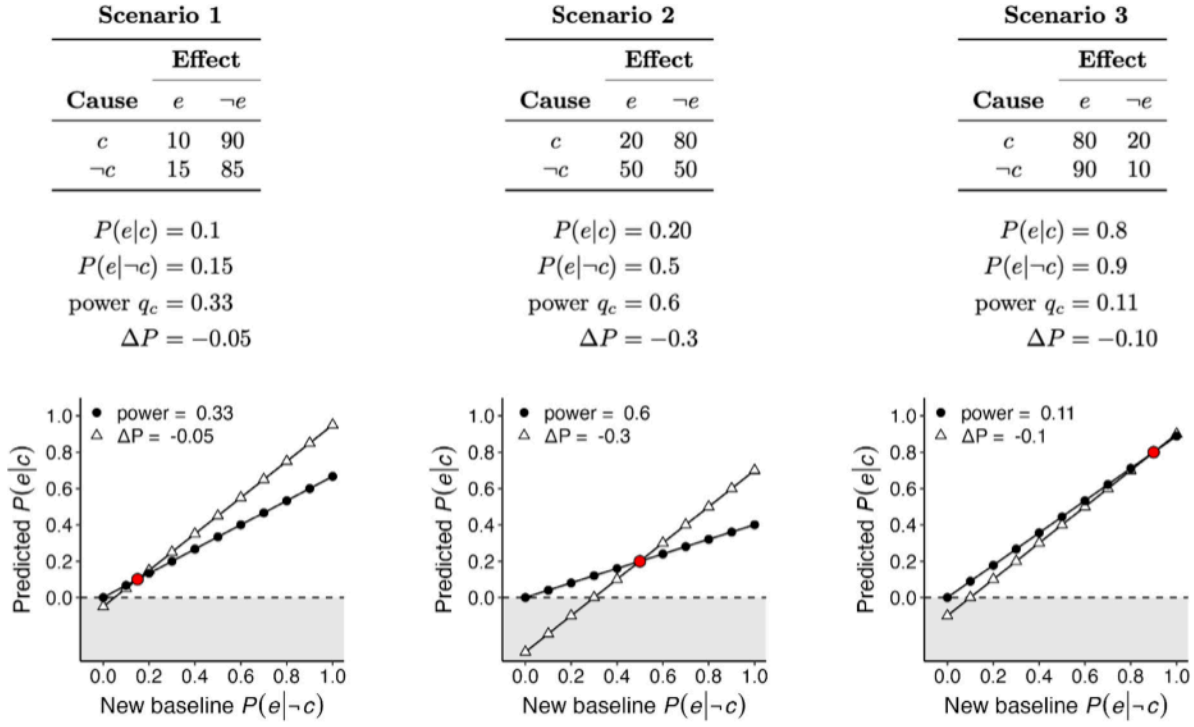

Fig. 3. Divergence of ΔP and causal power q_c when generalizing treatment effects to new baseline risks. Top row: Three different scenarios, where the baseline risk (probability of effect in the absence of the preventive cause) is rather low (left), intermediate (middle), or high (right). Middle row: Corresponding probabilities and values of ΔP (ARR) and causal power (RRR). Bottom row: Predicted probability of the effect given the cause under new baseline probabilities, calculated from a causal model where the causal strength of C is quantified using ΔP or causal power q_c (Eqs. (7) and (8)). The red dot indicates the combination of $P(e|c)$ and $P(e|\neg c)$ from which the strength of C was originally estimated. Shaded area indicate negative predicted probabilities.

different baseline risks requires careful consideration of the underlying causal mechanisms and how they influence the probability of the outcome. For the vaccine example, we adopted the standard medical assumption that RRRs capture the preventive power, mathematically

aligning with a noisy-NOT-AND parameterization where the cause probabilistically blocks the effect. Under this assumption, ARRs do not adequately generalize due to their assumption of a constant linear decrease in risk.

However, we emphasize that this need not be universally true. Whether ARR, RRR, or any other effect measure provides a more appropriate basis for generalization critically depends on the underlying causal structure of the system and involved mechanisms. In domains where treatment effects are better characterized by linear-additive shifts in outcome probability (e.g., Bruynesteyn et al., 2004), ARR (ΔP) may provide a more appropriate basis for generalization. On this view, generalizability, i.e.transportability, is not an intrinsic property of the measure alone (Colnet et al., 2024), but fundamentally depends on how well the measure's mathematical assumptions align with the causal mechanisms and biological processes at play.

4. Conclusions

We have integrated concepts from three distinct fields: (i) health and medical research, which seeks to quantify and generalize treatment effects, (ii) risk communication, which is concerned with the understanding of health information in patients and health professionals, and (iii) cognitive psychology, which investigates how people infer causal relations from covariation data. Conceptually, our analyses provide a theoretical foundation for interpreting risk reduction measures through the lens of causal inference, making causal considerations explicit that are otherwise left implicit. From an applied standpoint, the findings demonstrate how the causal Bayes nets framework can be leveraged to clarify the diverging implications of using ARR and RRR when generalizing treatment effects to contexts with varying baseline risks.

Notably, there are further relations between probabilistic theories of causality proposed in philosophy and psychology, and metrics routinely applied in medical research and practice (Sprenger, 2018; Stephan et al., 2021). Variants of the ΔP model have also featured prominently in philosophy of science and confirmation theory (Crupi & Tentori, 2014; Tentori et al., 2007) and models of the value of information (Nelson, 2005; Nelson et al., 2022; Wu et al., 2017). These relationships provide further traction for theory integration across disciplinary boundaries, while also highlighting differences among concepts, such as disparities between the definition of generative causal power and relative risk increases (see Appendix A.3 for details). They also point to further normative distinctions, as absolute measures have been argued to better support rational decision-making than relative measures (Colnet et al., 2024; Jäntgen, 2023; Sprenger & Stegenga, 2017; Stegenga, 2015). Our analyses contribute to these debates by specifying the structural assumptions underlying each measure and by offering a formal framework that helps to clarify when and why each measure is appropriate to use. This perspective also connects to recent work by Colnet et al. (2024) who examine the generalizability of a broader set of causal effect measures using the potential outcomes framework (Imbens & Rubin, 2015), with a focus on covariate shifts and treatment effects that vary across subgroups. Notably, their work highlights conditions under which linear metrics like ARR may be advantageous, providing complementary insights into when and why different effect measures support valid generalization.

The analyses also hold specific implications for the different fields and central debates that have engaged them. First, our analyses underscore the potential of real-world data for the theoretical and experimental analysis of human causal induction (Bramley et al., 2017; Gerstenberg et al., 2021; Goddu & Gopnik, 2024; Griffiths & Tenenbaum, 2009; Holyoak & Cheng, 2011; Meder et al., 2014; Waldmann, 2017). Psychologists have focused on the inference processes that transform raw data (i.e., described or experienced frequencies) into estimates of causal strength, with recurring debates about the normative and descriptive validity of different metrics (Cheng, 1997; Ghomi & Stegenga, 2022; Griffiths & Tenenbaum, 2005; Hoefer & Krauss, 2021; Holyoak & Cheng, 2011; Lober & Shanks, 2000; Meder et al., 2014; Sprenger, 2018; Sprenger & Stegenga, 2017; Stegenga, 2015). Typically, causal learning experiments present participants with stylized scenarios akin to real-world experiments (e.g., hypothetical study data on the joint occurrences of a virus and disease cases), typically with equal base rates for the presence and absence of the candidate cause (Buehner et al., 2003; Griffiths & Tenenbaum, 2005; Meder et al., 2014). While this design facilitates the assessment of the relevant probabilities assumed to enter the computations, it also incurs a mismatch with real-world environments characterized by low probability events, such as the comparably low exposure rate to a pathogen like the Coronavirus. To enhance ecological validity, empirical studies could use real-world data from medical studies (e.g., vaccine trials) as stimuli, where base rates are typically much lower and therefore ARR (ΔP) and RRR (causal power) strongly diverge. Participants could then be prompted to make inferences about contexts with varying baseline risks, where different models imply distinct predictions. Such experiments would provide additional insights into how people generalize their rich causal knowledge to novel situations to inform their judgment and decision making processes.

Second, our analyses contribute to a causal inference perspective on medical problems. Causal modeling techniques support the explicit specification of causal structure, which despite recent advances in formal methodology remains rare in clinical research. Here, we analyzed common measures of risk reduction from a causal inference perspective, highlighting that in contexts where baseline risks vary across populations and time, ARRs and RRR make strongly diverging predictions. These analyses make explicit the causal assumptions underlying different measures and emphasize that their utility for generalization depends on the alignment between those assumptions and the causal mechanisms linking treatment and outcome. By offering a unified account of ARRs and RRRs in the language of graphical causal models, our findings complement empirical analyses on the stability of different metrics (Deeks, 2002; Doi et al., 2022; Furukawa et al., 2002; Murad et al., 2024; Schmid et al., 1998; Xiao et al., 2022), and clarify the structural conditions under which a given measure should be stable (portable) across contexts.

Third, the analyses contribute to the ongoing debate on the interpretation of ARRs and RRRs in health communication. A central concern in this discourse is how laypeople (Carling et al., 2009; Hembroff et al., 2004) and healthcare professionals (Akl et al., 2011; Bucher et al., 1994; Marcatto et al., 2013) interpret and respond to treatment benefits expressed in relative or absolute terms. We argue that both absolute and relative risk reductions in isolation are inadequate for communicating medical risks and supporting individual decision-making. For instance, while both measures quantify the causal impact of a treatment by a single number derived from the likelihood of an outcome in one group compared to another, research shows that laypeople actually prefer having two numbers: one indicating how many people are likely to develop a disease with treatment and another without treatment (Carling et al., 2009; Trevena et al., 2006). Thus, instead of being presented with a single summary number, people often find it more helpful to see the actual risks in both the treatment and control group — a preference that most scientists would arguably share. Moreover, informed health decisions also require consideration of additional elements, including the magnitude of the baseline risk (Bodemer et al., 2014; Marcatto et al., 2013; Natter & Berry, 2005; Sheridan et al., 2003), heterogeneity of treatment effects across subpopulations (Bruynesteyn et al., 2004; Kent et al., 2010), the applicability of summary findings to individual patients (Kent & Hayward, 2007; Rothwell, 1995), as well as the quality of evidence on which estimates and recommendations are based (Guyatt, Oxman, Akl et al., 2011; Guyatt, Oxman, Kunz et al., 2011; Higgins et al., 2023; Moher et al., 2010). Accordingly, standards of evidence-based health communication consider a broader array of factors to support informed and shared decision-making (Bonner et al., 2021; Gigerenzer et al., 2007; Woloshin et al., 2023). Promising pathways

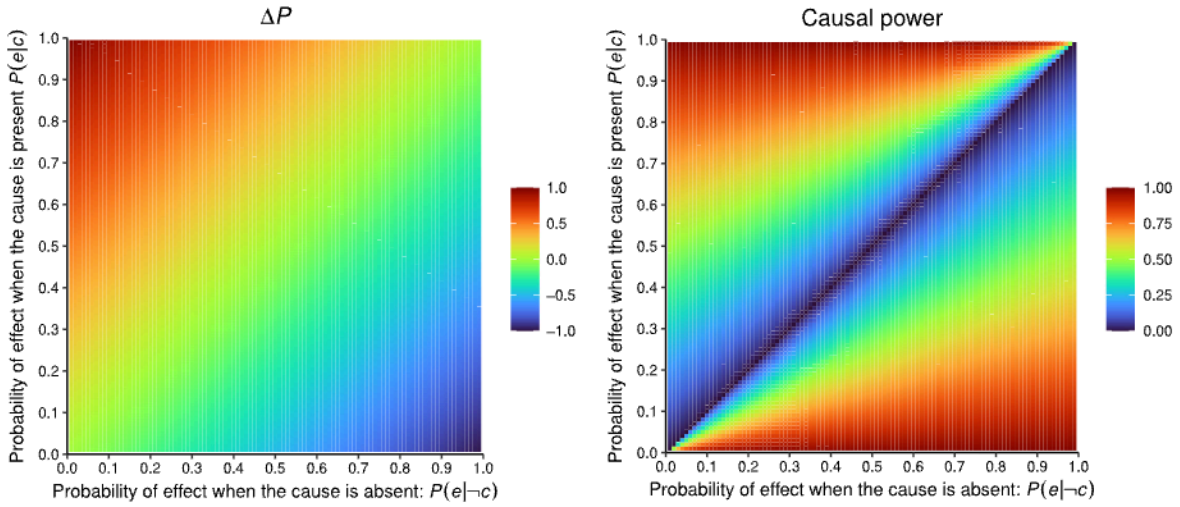

Fig. A.4. ΔP and causal power as a function of the probability of the effect in the presence and absence of the candidate cause, $P(e|c)$ and $P(e|\neg c)$. ΔP is always in the range $[-1, 1]$, whereas causal power is strictly non-negative in the range $[0, 1]$.

include so-called Fact Boxes (McDowell et al., 2019, 2016; Schwartz & Woloshin, 2013) that summarize the best available evidence on the benefits and harms associated with medical interventions, as well as carefully designed visual information (Woloshin et al., 2023). Thus, while our analyses demonstrate that relative reductions provide a more accurate assessment of treatment effects when generalizing to varying levels of baseline risk, effective risk communication requires a more comprehensive approach than providing healthcare professionals and the general public with a single number (or two).

CRediT authorship contribution statement

Björn Meder: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Charley M. Wu:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Felix G. Rebitschek:** Writing – review & editing, Methodology, Formal analysis, Conceptualization.

Appendix

A.1. ΔP and causal power: Differences in model behavior

Analogously to absolute and relative risk reductions, ΔP and causal power can yield very different values when applied to the same data (Fig. A.4). Consider a disease with a baseline risk of $P(e|\neg c) = 0.00002$, i.e. 2 in 100,000 people have the disease. Assume a vaccine reduces this baseline risk to 1 in 100,000, such that $P(e|c) = 0.00001$. In this case, $\Delta P = 0.00001$, but causal power $q_c = 0.5$, as the vaccine reduces the number of cases by 50%.

Generally, the same level of ΔP can entail different levels of causal power, depending on the probability of the effect in the absence of the cause (i.e., $P(e|\neg c)$; or the baseline risk). Conversely, low values of ΔP can imply maximum values of causal power, namely in those cases where the adverse event never occurs in the presence of the cause (i.e., $P(e|c) = 0$). In this case, while the level of ΔP depends on and in fact is equal to the baseline risk $P(e|\neg c)$, causal power is always 1 as the presence of the cause eliminates the presence of the effect (e.g., if a vaccine provides perfect protection against a disease).

A.2. Causal inference with graphical models: Numerical example

To provide a numerical example for modeling causal inference with the graph shown in Fig. 1, we use the data from the BioNTech SARS-CoV-2 vaccine trial (Polack et al., 2020, ; Table 1). First, we show how the parameterized causal model recovers the empirical probabilities. Then, we demonstrate how it can be used to generalize to different baseline risks.

The first step is to estimate parameters w_a and w_c from the observed frequencies (The base rate of the cause, b_c, can be estimated analogously, but is not relevant for the present purposes.) If we use maximum-likelihood point estimates derived from the observed frequencies for estimating w_a and w_c, the conditional probability of the effect given the cause computed via Eqs. (7) and (8) corresponds exactly to the empirical probability (i.e., proportion of observed disease cases in the vaccine group). We do not consider alternative approaches based on Bayesian methods (for details, see Griffiths & Tenenbaum, 2005; Lu et al., 2008; Meder et al., 2014) here for the sake of simplicity. While Bayesian methods involve more intricate mathematical computations, the conceptual implications do not change.

In the BNT162b2 trial, 8 out of 18,198 people in the vaccine group developed Covid-19 (Table 1). Hence, $P(disease|vaccine) = P(e|c) = 8/181988 = 0.0004$. This probability remains consistent when computing $P(e|c)$ from the parameterized causal model according to Eq. (7) and using ΔP (absolute risk reduction) as an estimate for w_c, or when using Eq. (8) and using causal power (relative risk reduction) as an estimate for w_c. An estimate for the strength of the background cause, or baseline risk, is provided by the number of disease cases in the placebo group, where 162 out of 18,325 participants developed Covid-19 (Table 1). Thus, $w_a = \frac{162}{18,325} = 0.0088$.

For estimating causal strength, represented by w_c, we can either use ΔP, which is equal to an absolute risk reduction, or causal power, which corresponds to a relative risk reduction. First, we compute ΔP:

$$\begin{aligned} \Delta P &= P(e|c) - P(e|\neg c) \\ &= \frac{8}{18198} - \frac{162}{18325} \\ &= 0.0004 - 0.0088 \\ &= -0.0084 \end{aligned} \tag{A.1}$$

Next, we use Eq. (7) to compute the conditional probability of the effect given the cause, that is, the probability of Covid-19 given vaccination:

$$\begin{aligned} P(e|c,a;w_a,w_c) &= a \cdot w_a + c \cdot w_c \\ &= a \cdot w_a + c \cdot \Delta P \\ &= a \cdot w_a + c \cdot ARR \\ &= 0.0088 - 0.0084 \\ &= 0.0004 \end{aligned} \tag{A.2}$$

This probability is identical to the probability obtained from the observed frequencies as computed above.

Next, we show that the same probability obtains when using causal power q_c instead of ΔP as estimate for w_c (Eq. (4)), and compute the probability of the effect given the cause according to Eq. (8). Using causal power as estimate for w_c corresponds to an alternative parameterization of the graph, instantiating a probabilistic generalization of logical functions (Pearl, 1988; Yuille & Lu, 2007). Incidentally, this aligns with using a relative risk reduction as a measure of vaccine efficacy. First, we compute the causal power of the vaccine according to the observed data (Table 1):

$$\begin{aligned} q_c &= \frac{P(e|c) - P(e|\neg c)}{P(e|\neg c)} \\ &= \frac{-\Delta P}{P(e|\neg c)} \\ &= \frac{0.0084}{0.0088} \\ &= 0.95 \end{aligned} \tag{A.3}$$

Now we can compute the probability of the effect (Covid-19) given the cause (vaccine) in accordance with Eq. (8):

$$\begin{aligned} P(e|c,a;w_a,w_c) &= a \cdot w_a(1 - c \cdot w_c) \\ &= a \cdot w_a(1 - c \cdot q_c) \\ &= a \cdot w_a(1 - c \cdot RRR) \\ &= 0.0088(1 - 0.95) \\ &= 0.0004 \end{aligned} \tag{A.4}$$

Again, this yields the same probability as calculated directly from the empirical frequencies.

These calculations illustrate how one can perform probabilistic causal inferences with a parameterized causal model. For the present purpose, we have used maximum-likelihood estimates for the graph's parameters that are directly derived from the empirical data. Accordingly, the recalculated probability of effect given cause corresponds exactly to the empirical proportion, both under a linear-additive parameterization based on ΔP and a noisy-logical parameterization based on causal power.

This approach corresponds to a causality-based factorization of the joint probability distribution over C and E, which enables inferences about scenarios that differ from the observed data. In particular, we can perform counterfactual reasoning by making inferences about novel situations for which no data is (yet) available. For instance, we might be interested in making predictions about the expected number of disease cases under a baseline risk different from the one under which the treatment effect was estimated (e.g., for a group of health professionals who have a much higher risk of being exposed to the virus). This inference can be modeled by setting the parameter w_a to the desired baseline risk, where a baseline risk of 1% would correspond to $w_a = 0.01$, a baseline risk of 30% would correspond to $w_a = 0.3$, etc.

Importantly, in this case it does matter whether a linear or noisy-logical parameterization of the graph is adopted, hence whether an absolute or relative risk reduction is used to quantify vaccine efficacy. An absolute risk reduction amounts to a linear-additive parameterization; accordingly, the predicted probability of the effect given the cause when the baseline risk is 30% (i.e., $w_a = 0.3$) would be

$$\begin{aligned} P(e|c,a;w_a,w_c) &= a \cdot w_a + c \cdot w_c \\ &= a \cdot w_a + c \cdot \Delta P \\ &= a \cdot w_a + c \cdot ARR \\ &= 0.3 - 0.0084 \\ &= 0.2916 \end{aligned} \tag{A.5}$$

Thus, when making a generalization using an absolute risk reduction, the inherent linearity assumption yields an (erroneous) estimate of about 29%.

In stark contrast, using a relative risk reduction yields a very different estimate, which aligns with the assumption that the vaccine should prevent the disease in 95% of the cases:

$$\begin{aligned} P(e|c,a;w_a,w_c) &= a \cdot w_a(1 - c \cdot w_c) \\ &= a \cdot w_a(1 - c \cdot q_c) \\ &= a \cdot w_a(1 - c \cdot RRR) \\ &= 0.3(1 - 0.95) \\ &= 0.015 \end{aligned} \tag{A.6}$$

For instance, if we had 1000 health care professionals with a 30% base line risk, without vaccination we would expect about 300 to get infected. If they were vaccinated, we would expect that 95% of these cases would be prevented, leaving 15 expected disease cases among the vaccinated people.

A.3. Relative risk increases and generative causal power

Analogously to absolute and relative risk reductions, the medical literature uses quantitative estimates to assess *increases* in health-related risks, such as the increased risk of skin cancer from repeated sunlight exposure or the increased risk of lung cancer for smokers. Experimental study designs that regularly examine potentially beneficial interventions for risk reduction also reveal risk increases for adverse events due to the interventions (e.g. allergic responses to medications). Similar to the preventive case there are fundamental mathematical relationships among the concepts used in the medical and causal modeling literature.

The *absolute risk increase* (ARI) is defined analogously to the preventive case as the arithmetic difference between the event rate in a control condition compared to a condition where participants have been exposed to the risk factor or treatment(Eq. (1)). Thus, both absolute risk increases and absolute risk reductions align with the ΔP model, quantifying the (beneficial or harmful) change in risk as an absolute difference in the probability of the effect in the presence and absence of the cause.

In contrast, a *relative risk increase* (RRI) as used in the medical literature and generative causal power are not defined identically. The relative increase normalizes the absolute risk increase by the event rate in the control group:

$$RRI = \frac{P(event|treatment) - P(event|control)}{P(event|control)} \tag{A.7}$$

However, for generative causes ($\Delta P > 0$), causal power is defined as

$$q_c = \frac{P(e|c) - P(e|\neg c)}{1 - P(e|\neg c)} = \frac{\Delta P}{1 - P(e|\neg c)} \tag{A.8}$$

where $1 - P(e|\neg c)$ serves as denominator. This normalization ensures that generative causal power is always in the range [0,1], whereas the

Scenario 1

Cause	Effect e	Effect $\neg e$
c	15	85
$\neg c$	10	90

$P(e|c) = 0.15$
$P(e|\neg c) = 0.10$
power $q_c = 0.06$
$\Delta P = 0.05$

Scenario 2

Cause	Effect e	Effect $\neg e$
c	80	20
$\neg c$	50	50

$P(e|c) = 0.8$
$P(e|\neg c) = 0.5$
power $q_c = 0.6$
$\Delta P = 0.3$

Scenario 3

Cause	Effect e	Effect $\neg e$
c	90	10
$\neg c$	80	20

$P(e|c) = 0.9$
$P(e|\neg c) = 0.8$
power $q_c = 0.5$
$\Delta P = 0.1$

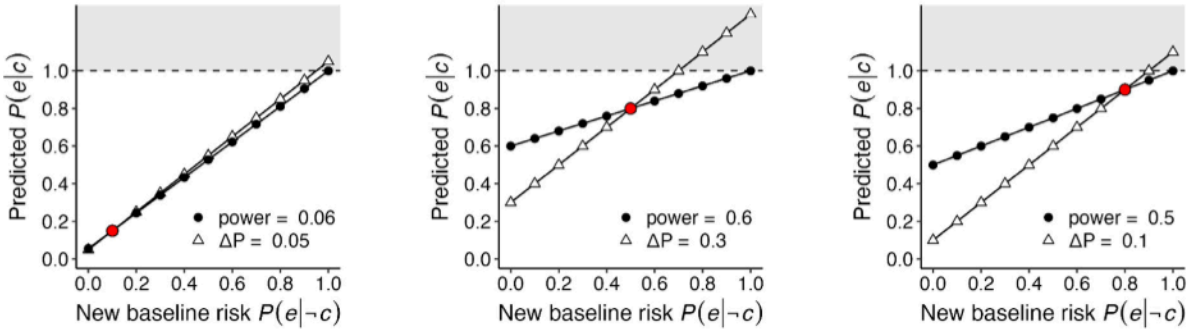

Fig. A.5. Generalizing generative causes to situations with a new baseline risk. Top row: Three different scenarios, where the baseline risk (i.e., probability of effect in the absence of the generative cause) is low (left), intermediate (middle), or high (right). Middle row: Corresponding probabilities and values of ΔP (ARR) and causal power (RRR). Bottom row: Predicted probability of the effect given the cause under new baseline probabilities, calculated from a causal model where the causal strength of C is quantified using ΔP or causal power q_c (Eqs. (7) and (A.10), respectively). The red dot indicates the combination of $P(e|c)$ and $P(e|\neg c)$ from which the strength of C was originally estimated. The shaded area indicates predicted probabilities that exceed unity.

calculation of a relative risk increase (Eq. (A.7)) does not impose any such constraints, rather making the RRI a multiple of the baseline risk. Generative causal power is zero when the cause does not change the probability of the effect (i.e., $\Delta P = 0$), and obtains the maximum value of 1 if $P(e|c) = 1$, regardless of the value of the baseline risk, $P(e|\neg c)$.

Both relative risk increases and generative causal power quantify increases in the probability of the effect given the cause event. However, normatively generative causal power has several advantages over relative risk increases as defined in the medical literature. First, generative power is constrained to the same range as relative risk reductions, namely [0,1] (or, when expressed as percentages, 0%–100%). This symmetry facilitates a causal interpretation analogously to decreases in risk. Second, mathematically causal power is better behaved: because the denominator in the definition of relative risk increases is $P(e|\neg c)$, the measure is undefined if the effect never occurs in the absence of the cause (Eq. (A.8)). To illustrate, assume 10 people eat a fish dish (cause C) and get food poisonings afterwards (effect E), whereas 10 other people eat no fish and develop no food poisoning. Accordingly, $P(e|c) = 1$ and $P(e|\neg c) = 0$. Using Eq. (A.7) entails division by zero, leaving the risk increase undefined. By contrast, generative causal power is defined and takes value 1 (Eq. (A.8)), suggesting that the fish caused the food poisoning. Third, relative increases can loom large even if the overall risk remains very low. For instance, a factor that increases the risk from 0.001% (1 in 100,000) to 0.002% (2 in 100,000) increases the relative risk by 100%, although the absolute risk is still fairly low. Causal power, being bounded between 0 and 1, does not suffer from this shortcoming:

$$q_c = \frac{P(e|c) - P(e|\neg c)}{1 - P(e|\neg c)} = \frac{\Delta P}{1 - P(e|\neg c)}$$
$$= \frac{(0.0002 - 0.0001)}{(1 - 0.00001)} = \frac{0.0001}{(1 - 0.0001)} \quad \text{(A.9)}$$
$$\approx 0.00001$$

This low value better aligns with the intuition that the candidate cause has only a weak influence on the likelihood of the effect. Fourth, using generative causal power ensures that the estimate can be incorporated into a causal modeling framework and be used to represent w_c, thereby supporting counterfactual inferences and the evaluation of hypothetical scenarios. Thus, from a normative perspective generative causal power is a better-behaved metric that relative risk increases as defined in the medical literature. Pragmatically though, it seems unlikely to replace an established measure like relative risk increase.

Like in the preventive case, using an absolute risk increase or, equivalently, ΔP, as estimate for w_c in the causal graph (Fig. 1) corresponds to a linear parameterization. In contrast, using generative causal power (but not an RRI) instantiates a so-called noisy-OR parameterization, which extends the logical OR by allowing the cause events A and C to influence E probabilistically (Cheng, 1997; Glymour, 2003; Pearl, 1988; Yuille & Lu, 2007). Given a noisy-OR parameterization, the conditional probability of the effect given the cause (e.g., probability of infection given vaccine) can be computed from the causal model's parameters as follows:

$$P(e|c, a; w_c, w_a) = c \cdot w_c + a \cdot w_a - c \cdot w_c c \cdot a \cdot w_a \quad \text{(A.10)}$$

where $a, c \in \{0, 1\}$ denote the presence and absence of candidate cause C and background cause A, and w_a and w_c denote their causal strength.

Like in the preventive case, using causal power has several advantages compared to using a linear parameterization where ΔP provides the estimate for w_c. Fig. A.5 illustrates this using three scenarios analogous to the analysis of preventive causes and risk reductions (Fig. 3). In each instance, the causal impact of the cause is assessed under a specific baseline risk $P(e|\neg c)$, and this estimate is then utilized to draw inferences about the likelihood of the effect in scenarios with new baseline probabilities, that is, $P(e|c)$. Analogous to the preventive case, there is always a crossover of the predicted $P(e|c)$ based on ΔP versus causal power. The intersection is always located at values of $P(e|c)$ and $P(e|\neg c)$ that were originally used to estimate causal strength, because for this

particular combination both measures recover the original $P(e|c)$. To the left of this point, i.e., when the new baseline probability is *lower* than the original one, the predicted conditional probability of the effect given the cause based on ΔP (or, equivalently, an ARR) is lower than the probability predicted by the causal power model. Conversely, to the right of this point, i.e., when the new baseline probability is higher than the one under which the strength of C was estimated, the predicted probability of the effect given the cause based on ΔP is higher than the probability predicted by causal power. Moreover, as in the preventive scenarios a linear parameterization based on ΔP does not always yield valid probability estimates. Specifically, while in the preventive case negative probabilities result if the new baseline probability is lower than the (absolute) value of ΔP, in generative scenarios the probability estimates exceed unity if the baseline risk is higher than $1 - \Delta P$.

References

Adams-Phipps, J., Toomey, D., Więcek, W., Schmit, V., Wilkinson, J., Scholl, K., Jamrozik, E., Osowicki, J., Roestenberg, M., & Manheim, D. (2023). A systematic review of human challenge trials, designs, and safety. *Clinical Infectious Diseases*, *76*(4), 609–619. http://dx.doi.org/10.1093/cid/ciac820.

Akl, E. A., Oxman, A. D., Herrin, J., Vist, G. E., Terrenato, I., Sperati, F., Costiniuk, C., Blank, D., & Schünemann, H. (2011). Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database of Systematic Reviews*, (3).

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: a systematic review. *Journal of the American Medical Informatics Association*, *13*(6), 608–618. http://dx.doi.org/10.1197/jamia.m2115.

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., et al. (2021). Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine*, *384*(5), 403–416. http://dx.doi.org/10.1056/NEJMoa2035389.

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, *113*(27), 7345–7352. http://dx.doi.org/10.1073/pnas.1510507113.

Bodemer, N., Meder, B., & Gigerenzer, G. (2014). Communicating relative risk changes with baseline risk: presentation format and numeracy matter. *Medical Decision Making*, *34*(5), 615–626. http://dx.doi.org/10.1177/0272989X14526305.

Bonner, C., Trevena, L. J., Gaissmaier, W., Han, P. K., Okan, Y., Ozanne, E., Peters, E., Timmermans, D., & Zikmund-Fisher, B. J. (2021). Current best practice for presenting probabilities in patient decision aids: fundamental principles. *Medical Decision Making*, *41*(7), 821–833. http://dx.doi.org/10.1177/0272989X21996328.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning.. *Psychological Review*, *124*(3), 301–338. http://dx.doi.org/10.1037/rev0000061.

Bruynesteyn, K., Wanders, A., Landewé, R., & van der Heijde, D. (2004). How the type of risk reduction influences required sample sizes in randomised clinical trials. *Annals of the Rheumatic Diseases*, *63*(11), 1368–1371.

Bucher, H., Weinbacher, M., & Gyr, K. (1994). Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration. *British Medical Journal*, *309*(6957), 761–764. http://dx.doi.org/10.1136/bmj.309.6957.761, URL: https://europepmc.org/articles/pmc2541000?pdf=render.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power.. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(6), 1119–1140. http://dx.doi.org/10.1037/0278-7393.29.6.1119.

Carling, C. L., Kristoffersen, D. T., Montori, V. M., Herrin, J., Schünemann, H. J., Treweek, S., Akl, E. A., & Oxman, A. D. (2009). The effect of alternative summary statistics for communicating risk reduction on decisions about taking statins: a randomized trial. *PLoS Medicine*, *6*(8), Article e1000134. http://dx.doi.org/10.1371/journal.pmed.1000134.

Cartwright, N. (2007). Hunting causes and using them. *Cambridge Books*, http://dx.doi.org/10.1017/cbo9780511618758.

Cheng, P. W. (1997). From covariation to causation: A causal power theory.. *Psychological Review*, *104*(2), 367–405. http://dx.doi.org/10.1037/0033-295X.104.2.367.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–382. http://dx.doi.org/10.1037/0033-295x.99.2.365.

Colnet, C., Jackson, C., Höhna, S., Rohde, A., & Sachs, R. (2024). Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize?. ArXiv, URL: https://arxiv.org/abs/2401.03645.

Crupi, V., & Tentori, K. (2014). State of the field: Measuring information and confirmation. *Studies in History and Philosophy of Science Part A*, *47*, 81–90. http://dx.doi.org/10.1016/j.shpsa.2014.05.002.

Deeks, J. J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, *21*(11), 1575–1600. http://dx.doi.org/10.1002/sim.1188.

Digitale, J. C., Martin, J. N., & Glymour, M. M. (2022). Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*, *142*, 264–267.

Doi, S. A., Furuya-Kanamori, L., Xu, C., Lin, L., Chivese, T., & Thalib, L. (2022). Controversy and debate: questionable utility of the relative risk in clinical research: paper 1: a call for change to practice. *Journal of Clinical Epidemiology*, *142*, 271–279.

Etminan, M., Collins, G. S., & Mansournia, M. A. (2020). Using causal diagrams to improve the design and interpretation of medical research. *Chest*, *158*(1), S21–S28. http://dx.doi.org/10.1016/j.chest.2020.03.011.

Feynman, R. P. (1987). Negative probability. In B. Hiley, & F. Peat (Eds.), *Quantum implications: essays in honour of david bohm* (pp. 235–248). London and New York: Routledge.

Furukawa, T. A., Guyatt, G. H., & Griffith, L. E. (2002). Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. *International Journal of Epidemiology*, *31*(1), 72–76. http://dx.doi.org/10.1093/ije/31.1.72.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936–975. http://dx.doi.org/10.1037/rev0000281.

Ghomi, H. T., & Stegenga, J. (2022). Conventional choices in outcome measures influence meta-analytic results. *Philosophy of Science*, *89*(5), 949–959.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *BMJ*, *327*(7417), 741–744. http://dx.doi.org/10.1136/bmj.327.7417.741.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96. http://dx.doi.org/10.1111/j.1539-6053.2008.00033.x.

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences*, *7*(1), 43–48. http://dx.doi.org/10.1016/s1364-6613(02)00009-8.

Goddu, M. K., & Gopnik, A. (2024). The development of human causal learning and reasoning. *Nature Reviews Psychology*, *3*(5), 319–339.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*(1), 37–48. http://dx.doi.org/10.1097/00001648-199901000-00008.

Greenwood, M., & Yule, G. U. (1915). The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general. *Proceedings of the Royal Society of Medicine*, *8*, 113–194.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. http://dx.doi.org/10.1016/j.cogpsych.2005.05.004.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716. http://dx.doi.org/10.1037/a0017201.

Guyatt, G. H., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., DeBeer, H., et al. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, *64*(4), 383–394. http://dx.doi.org/10.1016/j.jclinepi.2010.04.026.

Guyatt, G. H., Oxman, A. D., Kunz, R., Atkins, D., Brozek, J., Vist, G., Alderson, P., Glasziou, P., Falck-Ytter, Y., & Schünemann, H. J. (2011). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology*, *64*(4), 395–400.

Hembroff, L. A., Holmes-Rovner, M., & Wills, C. E. (2004). Treatment decision-making and the form of risk communication: results of a factorial survey. *BMC Medical Informatics and Decision Making*, *4*(1), 1–9. http://dx.doi.org/10.1186/1472-6947-4-20.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2023). Cochrane handbook for systematic reviews of interventions version 6.4 (updated august 2023). URL: https://training.cochrane.org/handbook,

Hoefer, C., & Krauss, A. (2021). Measures of effectiveness in medical research: Reporting both absolute and relative measures. *Studies in History and Philosophy of Science*, *88*, 280–283.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*(1), 135–163. http://dx.doi.org/10.1146/annurev.psych.121208.131634.

Huitfeldt, A., Goldstein, A., & Swanson, S. A. (2018). The choice of effect measure for binary outcomes: Introducing counterfactual outcome state transition parameters. *Epidemiologic Methods*, *7*(1), Article 20160014.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jäntgen, I. (2023). How to measure effect sizes for rational decision making. *Philosophy of Science*, *90*(5), 1183–1193.

Kent, D. M., & Hayward, R. A. (2007). Limitations of Applying Summary Results of Clinical Trials to Individual PatientsThe Need for Risk Stratification. *JAMA*, *298*(10), 1209–1212. http://dx.doi.org/10.1001/jama.298.10.1209.

Kent, D. M., Rothwell, P. M., Ioannidis, J. P., Altman, D. G., & Hayward, R. A. (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*, *11*(1), 1–11. http://dx.doi.org/10.1186/1745-6215-11-85.

Killingley, B., Mann, A. J., Kalinova, M., Boyers, A., Goonawardane, N., Zhou, J., Lindsell, K., Hare, S. S., Brown, J., Frise, R., et al. (2022). Safety, tolerability and viral kinetics during SARS-CoV-2 human challenge in young adults. *Nature Medicine*, *28*(5), 1031–1041. http://dx.doi.org/10.1038/s41591-022-01780-9.

Kyriacou, D. N., Greenland, P., & Mansournia, M. A. (2023). Using causal diagrams for biomedical research. *Annals of Emergency Medicine*, *81*(5), 606–613.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of cheng (1997). *Psychological Review*, *107*(1), 195–212. http://dx.doi.org/10.1037/0033-295x.107.1.195.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning.. *Psychological Review*, *115*(4), 955–984. http://dx.doi.org/10.1037/a0013256.

Madrigal, R., Armstrong Soule, C. A., & King, J. (2024). Manipulating consumers with the truth: Relative-difference claims in advertising and inferences of manipulative intent. *Journal of Advertising*, 1–18. http://dx.doi.org/10.1080/00913367.2024.2306401.

Marcatto, F., Rolison, J. J., & Ferrante, D. (2013). Communicating clinical trial outcomes: effects of presentation method on physicians' evaluations of new treatments. *Judgment and Decision Making*, *8*(1), 29–33. http://dx.doi.org/10.1017/s1930297500004472.

McDowell, M., Gigerenzer, G., Wegwarth, O., & Rebitschek, F. G. (2019). Effect of tabular and icon fact box formats on comprehension of benefits and harms of prostate cancer screening: a randomized trial. *Medical Decision Making*, *39*(1), 41–56. http://dx.doi.org/10.1177/0272989X18818166.

McDowell, M., Rebitschek, F. G., Gigerenzer, G., & Wegwarth, O. (2016). A simple tool for communicating the benefits and harms of health interventions: a guide for creating a fact box. *MDM Policy & Practice*, *1*(1), Article 2381468316665365. http://dx.doi.org/10.1177/2381468316665365.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning.. *Psychological Review*, *121*(3), 277–301. http://dx.doi.org/10.1037/a0035944.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal*, *340*(mar23 1), c869. http://dx.doi.org/10.1136/bmj.c869.

Murad, M. H., Wang, Z., Xiao, M., Chu, H., & Lin, L. (2024). Variability of relative treatment effect among populations with low, moderate and high control group event rates: a meta-epidemiological study. *BMC Medical Research Methodology*, *24*(1), 263.

Natter, H. M., & Berry, D. C. (2005). Effects of presenting the baseline risk when communicating absolute and relative risk reductions. *Psychology, Health & Medicine*, *10*(4), 326–334. http://dx.doi.org/10.1080/13548500500093407.

Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain.. *Psychological Review*, *112*(4), 979–999. http://dx.doi.org/10.1037/0033-295X.112.4.979.

Nelson, J. D., Rosenauer, C., Crupi, V., Tentori, K., & Meder, B. (2022). The likelihood difference heuristic and binary test selection given situation-specific utilities. *Decision*, *9*(3), 285–319. http://dx.doi.org/10.1037/dec0000160.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence.. *Psychological Review*, *111*(2), 455–485. http://dx.doi.org/10.1037/0033-295X.111.2.455.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, URL: https://ci.nii.ac.jp/ncid/BA04815189.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J., & Mackenzie, D. (2018). The book of why. Hachette.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., et al. (2020). Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New England Journal of Medicine*, *383*(27), 2603–2615. http://dx.doi.org/10.1056/NEJMoa2034577.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, *2*(7), 369–375. http://dx.doi.org/10.1038/s42256-020-0197-y.

Rothwell, P. M. (1995). Can overall results of clinical trials be applied to all patients? *The Lancet*, *345*(8965), 1616–1619. http://dx.doi.org/10.1016/s0140-6736(95)90120-5.

Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, *9*(8), Article 220638. http://dx.doi.org/10.48550/arxiv.2205.11402.

Schmid, C. H., Lau, J., McIntosh, M. W., & Cappelleri, J. C. (1998). An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Statistics in Medicine*, *17*(17), 1923–1942. http://dx.doi.org/10.1002/(sici)1097-0258(19980915)17:17<1923::aid-sim874>3.0.co;2-6.

Schölkopf, B. (2022). Causality for machine learning. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), *vol. 36*, *Probabilistic and causal inference: the works of judea pearl* (pp. 765–804). Association for Computing Machinery, http://dx.doi.org/10.1145/3501714.3501755.

Schwartz, L. M., & Woloshin, S. (2013). The drug facts box: Improving the communication of prescription drug information. *Proceedings of the National Academy of Sciences*, *110*(3), 14069–14074. http://dx.doi.org/10.1073/pnas.1214646110.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*(11), 966–972.

Sheridan, S. L., Pignone, M. P., & Lewis, C. L. (2003). A randomized comparison of patients' understanding of number needed to treat and other common risk reduction formats. *Journal of General Internal Medicine*, *18*(11), 884–892. http://dx.doi.org/10.1046/j.1525-1497.2003.21102.x.

Smith, G. L., & McFadden, G. (2002). Smallpox: anything to declare? *Nature Reviews Immunology*, *2*(7), 521–527. http://dx.doi.org/10.1038/nri845.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd Ed.). MIT Press.

Sprenger, J. (2018). Foundations of a probabilistic theory of causal strength. *Philosophical Review*, *127*(3), 371–398. http://dx.doi.org/10.1215/00318108-6718797.

Sprenger, J., & Stegenga, J. (2017). Three arguments for absolute outcome measures. *Philosophy of Science*, *84*(5), 840–852. http://dx.doi.org/10.1086/693930.

Stegenga, J. (2015). Measuring effectiveness. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *54*, 62–71. http://dx.doi.org/10.1016/j.shpsc.2015.06.003.

Stephan, S., Placi, S., & Waldmann, M. R. (2021). Evaluating general versus singular causal prevention. In W. T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *vol. 43*, *Proceedings of the annual meeting of the cognitive science society* (pp. 1402–1408). Cognitive Science Society.

Stovitz, S. D., & Shrier, I. (2019). Causal inference for clinicians. *BMJ Evidence-Based Medicine*, *24*(3), 109–112. http://dx.doi.org/10.1136/bmjebm-2018-111069.

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*(1), 107–119. http://dx.doi.org/10.1016/j.cognition.2005.09.006.

Tentori, K., Passerini, A., Timberlake, B., & Pighin, S. (2021). The misunderstanding of vaccine efficacy. *Social Science & Medicine*, *289*, Article 114273. http://dx.doi.org/10.1016/j.socscimed.2021.114273.

Trevena, L. J., Davey, H. M., Barratt, A., Butow, P., & Caldwell, P. (2006). A systematic review on communicating with patients about evidence. *Journal of Evaluation in Clinical Practice*, *12*(1), 13–23. http://dx.doi.org/10.1111/j.1365-2753.2005.00596.x.

Voysey, M., Clemens, S. A. C., Madhi, S. A., Weckx, L. Y., Folegatti, P. M., Aley, P. K., Angus, B., Baillie, V. L., Barnabas, S. L., Bhorat, Q. E., et al. (2021). Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *The Lancet*, *397*(10269), 99–111. http://dx.doi.org/10.1016/s0140-6736(20)32661-1.

Waldmann, M. R. (2017). The oxford handbook of causal reasoning. *Oxford handbooks online*. Oxford University Press, http://dx.doi.org/10.1093/oxfordhb/9780199399550.001.0001.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition.. *Journal of Experimental Psychology: General*, *121*(2), 222–236. http://dx.doi.org/10.1037//0096-3445.121.2.222.

Weinberg, G. A., & Szilagyi, P. G. (2010). Vaccine epidemiology: efficacy, effectiveness, and the translational research roadmap. *The Journal of Infectious Diseases*, *201*(11), 1607–1610. http://dx.doi.org/10.1086/652404.

Woloshin, S., Yang, Y., & Fischhoff, B. (2023). Communicating health information with visual displays. *Nature Medicine*, *29*(5), 1085–1091. http://dx.doi.org/10.1038/s41591-023-02328-1.

Wu, C. M., Meder, B., Filimon, F., & Nelson, J. D. (2017). Asking better questions: How presentation formats influence information search. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *8*(8), 1274–1297. http://dx.doi.org/10.1037/xlm0000374.

Xiao, M., Chen, Y., Cole, S. R., MacLehose, R. F., Richardson, D. B., & Chu, H. (2022). Controversy and debate: Questionable utility of the relative risk in clinical research: Paper 2: Is the odds ratio "portable" in meta-analysis? Time to consider bivariate generalized linear mixed model. *Journal of Clinical Epidemiology*, *142*, 280–287.

Yuille, A. L., & Lu, H. (2007). The noisy-logical distribution and its application to causal inference. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *20*, *Advances in neural information processing systems* (pp. 1673–1680). Curran Associates, Inc.,