Choice of effect measure, extrapolation, and decision-making in patient care and in public health

Abstract

Decision-theoretic arguments show that only absolute effect measures, and not relative ones, suffice for utility-maximizing clinical or public health decisions (Jäntgen, 2023; Sprenger and Stegenga, 2017). This paper argues that despite the validity and importance of these results, no general conclusions about the policy-relevance of different measures follow from such arguments. This is because in a typical decision situation, the decision-relevant risks or summary effects are not directly available for the target patient or population, but must be inferred via extrapolation. Since absolute effects always depend on baseline risk, reliably extrapolating them requires controlling for all causes of the outcome that vary between the source and the target, not just for effect modifiers that mechanistically interact with the exposure. Relative effects depend on baseline risk in some circumstances but not in others, depending on the nature of the exposure (e.g. Huitfeldt et al., 2018; Sheps, 1958). An appropriately chosen relative measure may thus be extrapolatable with far less auxiliary evidence or fewer assumptions, and given the target baseline risk, provides the risk information needed for rational decisions (Jäntgen, 2023). Hence, estimating baseline risk in the target directly and extrapolating a relative effect may in many cases be the most reliable and efficient way to obtain decision-relevant evidence. Absolute measures are thus not generally superior in real-life decision contexts where extrapolation error must be dealt with. Instead, appropriate choice of effect measure to be used in inferring the decision-relevant risks depends on the properties of the exposure of interest. Lastly, the paper outlines some implications for philosophical treatments of the problem of extrapolation itself.

1 Introduction

In epidemiology, an effect of a dichotomous exposure on a dichotomous outcome is a comparison of risks between the exposed and the unexposed. Causally interpreted, this comparison is assumed to equal a comparison in counterfactual risks if, hypothetically, both exposure states were to occur at once for each subject (Hernán and Robins, 2020). These comparisons are summarized by effect measures like risk difference or risk ratio. Risk difference describes the additive influence of an exposure on an outcome, and is often called an absolute effect measure. Trials occasionally report the inverse of a risk difference, which can also be classified as an absolute measure, as inverting it again returns the risk difference. Measures like risk ratio, which describe a multiplier of risk, are called relative, or ratio measures.

Arguably, it is the absolute, risk difference scale that most adequately quantifies the importance of an effect for clinical application and public health (e.g. Rothman, 2012, chapter 4). This being the case, one might think that the default choice of effect measure in epidemiological studies should be an absolute measure. Something like this has been argued e.g. by (Broadbent, 2013, chapter 9; Fuller, 2021; Stegenga, 2015; Sprenger and Stegenga, 2017). For example, Sprenger and Stegenga write

> Unfortunately, relative measures are widely employed in clinical research, and absolute measures are underused. [...] Medical science, whether in clinical trials or in epidemiology, should always use and report absolute outcome measures (Sprenger and Stegenga, 2017, p.851).

I reconsider the case for the superiority of absolute measures, when decision-making is concerned. I take as a starting point a decision-theoretic model presented in Sprenger and Stegenga (2017), which demonstrates the superiority of absolute measures in a particular decision situation. The model describes a setting where one must choose to treat a patient either with treatment A, or treatment B, given an effect measure gleaned from a trial testing A against B, and assuming risk neutral preferences and no estimation error or uncertainty about the applicability of the effect to the target patient[1]. The model entails that absolute measures, and no others, suffice for

[1]These assumptions are not made explicit in (Sprenger and Stegenga, 2017), but the conclusion that absolute

utility-maximizing decisions. Note that the risks under different exposure assignments therefore obviously suffice too. Jäntgen (2023) generalizes this model to situations where no trial result of testing A against B is available, but results from separate trials testing A and B against other control treatments are available. Jäntgen shows that risk differences or risk ratios together with baseline risks are needed to calculate the decision-relevant risks. Taking to account all these results, the conclusion is that absolute measures are superior to relative ones in a decision context, since relative measures offer no advantages over absolute ones even in the situations considered by Jäntgen. These arguments generalize from the clinical context to public health, where A and B are population-level policies associated with particular exposure levels, e.g. A could be an intervention that prevents exposure while B corresponds to doing nothing, and the studies supplying the evidence are nonclinical studies. This is also Sprenger and Stegenga's intention (Sprenger and Stegenga, 2017, p. 851).

I consider what happens when the assumption about applicability of an effect measure to the target patient or population is relaxed. Arguably, this is the typical situation: decision-makers do not have evidence of risk under exposure, or of a summary effect, obtained in the target directly, but must infer those quantities based on evidence that is partly or wholly obtained in a different population, hence relying on extrapolation. Specifically, I examine situations where an effect is extrapolated to a target from a source population with different level of baseline risk. A risk difference is by definition dependent on baseline risk, and will vary with it. Empirical evidence suggests that this variability can be quite large (e.g. Deeks, 2002; Senn, 2011). Hence, extrapolating a risk difference across unequal baseline risks is prone to error, and correcting the error requires adjusting the effect for all causes of differences in baseline risk, i.e. all causes of the outcome that vary between the source and the target besides the exposure. This is rarely possible, and one must instead rely on unsubstantiated assumptions about the distributions of such other causes.

I then point out a fact noted by Mindel Sheps (1958; 1959) and formalized by contemporary epidemiologists (e.g. Huitfeldt et al., 2018, 2021): different effect measures have different stability

measures suffice for utility-maximizing decisions does not follow without them. If e.g. the risk difference estimate at hand does not apply to the patient, then not only the estimate, but also the magnitude of error in the estimate is needed as input to the decision, to secure utility-maximization. Or if the patient's preferences are sensitive to risk, then risk information, not a summary effect, is needed for a rational decision.

properties against variation in risk profile, depending on the nature of the exposure and the coding of the outcome in data. In particular, there are circumstances in which a relative effect will be mostly independent of baseline risk, and can thus be extrapolated with a much smaller conditioning set than an absolute effect. When this is the case, the risk information needed for a decision is most efficiently obtained by extrapolating a relative effect that is applied to an estimate of baseline risk obtained directly in the target. Unlike extrapolation of risk difference, this extrapolation does not require adjusting for all causes of the outcome, while estimating baseline risk only requires observations of the outcome. The decision-theoretic results notwithstanding, it is thus not the case that absolute measures are categorically superior or on par with relative ones as output from epidemiological studies, when the needs of decision-making are considered. This is because a summary effect observed in a study population is typically not useable as direct input into a decision, but is rather a means to the end of obtaining risk or effect estimates that apply to the target, which then support a decision. Which effect measure is best suited for this purpose depends on its transportability properties, which vary from case to case depending on the properties of the exposure being considered.

The relevance of these conclusions is twofold. Firstly, the fact that the evidential demands of extrapolation vary depending on what measure is being transferred is elemental to considerations of policy-relevance, as decision-making usually relies on extrapolated effects, and collecting auxiliary evidence to support extrapolation is costly. Secondly, since those evidential demands vary from one measure to the other depending on the nature of the exposure and coding of the outcome, no measure is most transferrable by default, and all blanket statements about superior policy-relevance of a particular measure over others are therefore probably false.

As for the problem of extrapolation itself, prominent philosophical accounts overlook the specification of the effect of interest, and advise searching and controlling for all other causes of the outcome that could possibly modify the exposure's influence in some way (e.g. Bareinboim and Pearl, 2013; Cartwright and Hardie, 2012; Steel, 2007). While this guarantees correct inference when doable, it is often very difficult to do in practice. Paying attention to the choice of effect measure offers a different perspective; which other causes of the outcome need to be controlled for depends on the choice of effect measure, and the nature of the exposure. Acknowledging this shows

the problem of extrapolation in a new light.

The paper is structured as follows. Section 2 introduces common effect measures. Section 3 introduces the decision-theoretic arguments of Sprenger and Stegenga (2017) and Jäntgen (2023). Section 4 introduces the problem of extrapolation. Section 5 introduces the phenomenon of baseline risk dependence, illustrated with a stylized example. Section 6 describes the implications of extrapolation error for decision-making. Section 7 describes implications for theories of extrapolation. Conclusions are summarized in section 8.

2 Varieties of effect measure

A basic quantity of interest in epidemiology is the incidence proportion, better known as risk. Risk is the proportion of subjects who become cases, i.e. experience an outcome, in a population during a period of follow-up (Rothman, 2012). Being a proportion, risk is interpretable as a probability. For a dichotomous outcome Y (0: outcome does not occur, 1: outcome occurs), risk is defined as $R = p(Y = 1)$. The complement of risk, $1 - R$, describes the risk of avoiding an outcome, i.e. of experiencing $Y = 0$. Historically, this is called "survival risk" or "survival probability", whether or not the outcome is death. From risk, (risk-)odds is defined as the ratio of risk to the survival risk, $O = \frac{R}{1-R}$.

For a dichotomous exposure A (0: unexposed, 1: exposed), the risk of an outcome given exposure is $R_1 = p(Y = 1|A = 1)$, and risk in absence of exposure is $R_0 = p(Y = 1|A = 0)$. From these quantities, one can define summary effect measures that compare risks or odds under different exposure assignments. Absolute measures include risk difference (RD) and number needed to treat (NNT):

$$\text{Risk Difference: } RD = R_1 - R_0$$

$$\text{Number Needed to Treat: } NNT = \frac{1}{RD}$$

. RD describes the absolute change in risk associated with exposure. NNT describes the number of exposure events (typically, number of treated) associated with a single occurrence of an outcome

(typically, of remission or recovery).

Some of the most common relative effect measures include risk ratio (*RR*) and odds ratio (*OR*). The latter is usually used as an estimator of the former rather than as a summary effect in itself. I do not discuss estimation, but include *OR* in the definitions below to acknowledge its ubiquitous use in the context of studying dichotomous outcomes. Survival ratio (*SR*) is another relative measure occasionally used, and is defined below as it will become important later. These measures are defined as

$$\begin{aligned}\text{Risk Ratio: } RR &= \frac{R_1}{R_0} \\ \text{Survival Ratio: } SR &= \frac{1-R_1}{1-R_0} \\ \text{Odds Ratio: } OR &= \frac{R_1}{1-R_1} \Big/ \frac{R_0}{1-R_0} = \frac{O_1}{O_0}\end{aligned}$$

, they describe the proportional change in risk or odds associated with exposure.

To be causally interpretable, the associational measures that summarize population frequencies should be equal to the corresponding counterfactual effects (Hernán and Robins, 2020). The latter describe differences or ratios of risks under exposure assignments that are at least partly not realized. Let $Y^{a=1}$ and $Y^{a=0}$ be outcome variables hypothetically observed for each subject under exposure and absence of exposure, respectively. A causal effect for an individual exists if, and only if $Y^{a=0} \neq Y^{a=1}$, and a causal effect in population exists iff $p(Y^{a=0}=1) \neq p(Y^{a=1}=1)$. For example, a causal *RD* equals

$$[p(Y^{a=1}=1) - p(Y^{a=0}=1)]$$

, which can be interpreted as the difference in proportion of cases that would be observed if every subject was exposed, and that would be observed if every subject was unexposed. Since these exposure distributions cannot be realized at the same time, at least one of the risks is counterfactual. Ratios of counterfactual risks define causal ratio measures in a similar fashion. For example, causal *SR* is defined as

$$\frac{p(Y^{a=1}=0)}{p(Y^{a=0}=0)}$$

. Since this paper is not about estimation, I omit discussion of identifiability conditions of causal

effects, and assume that all effect measures discussed below are causally interpretable without always explicitly indicating this. The counterfactual notation is used whenever it is in order to be explicit.

Causal effects explain statistical facts about a population by answering what-if? -questions (Kuorikoski, 2012; Woodward, 2003). For example, a causal $RD = 0.05$ claims that if an intervention eliminated exposure in a population, the population proportion of cases would have been reduced by $0.05 \cdot r$, where r is the prevalence of exposure. Reality is of course messier than this, and another causative exposure may replace a prevented one. Regardless, it is the connection to reasoning about policy-interventions that makes causal effects relevant for clinical and public health decision-making.

3 The decision-theoretic argument

Consider two treatments A and B and a dichotomous outcome Y (0: no recovery; 1: recovery), and an agent that must choose between A and B for treating a patient. Values $A = 1$ and $B = 1$, denoting exposure to the treatments, are abbreviated as A and B to make the notation as consistent with (Sprenger and Stegenga, 2017) and (Jäntgen, 2023) as possible. The principle of maximising expected utility gives the following decision-rule

> (#) 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 $EU(A) > EU(B)$ (Sprenger and Stegenga, 2017, p. 845)

, where a is the cost (broadly understood) of treatment A, b the cost of B, u the utility associated with $Y = 1$, $u^{'}$ the utility of $Y = 0$, and $EU(A)$ and $EU(B)$ are expected utilities of choosing A and of choosing B, respectively.

Sprenger and Stegenga (2017, p. 846) prove that $EU(A) > EU(B)$ iff

$$[p(Y = 1|A) - p(Y = 1|B)] > \frac{a-b}{u-u^{'}} \tag{1}$$

. Since $[p(Y = 1|A) - p(Y = 1|B)]$ is a difference in risks under two different exposures, treatments

A and B, 1 can be written as

$$RD > \frac{a-b}{u-u'} \tag{1}$$

, where RD stands for the relevant risk difference.

This means that given a trial testing A against B on subjects relevantly similar to the target patient, the trial RD together with costs and utilities would suffice for a rational decision by allowing one to check whether equation 1 is satisfied. Since inverting NNT returns RD, NNT also suffices (Sprenger and Stegenga, 2017, p. 846). To intuitively grasp what this means, ignore for the moment the assumption that $a > b$ and consider a situation where the costs of A and B are equal. In that situation, one should choose A over B as long as A confers any benefit at all over B in probability of recovery. Equation 1 confirms the intuition. When $a = b$, the costs-to-utilities ratio $\frac{a-b}{u-u'}$ goes to zero, and equation 1 becomes $RD > 0$: one should choose A over B iff the RD testing A against B is above zero. Relative measures do not suffice for rational decision-making, because they underdetermine RD. A $RR = 2$, for example, is compatible with RDs ranging from infinitesimally tiny to $RD = 0.5$. Knowing nothing but RR, considerable uncertainty remains about whether equation 1 is satisfied. Sprenger & Stegenga give a somewhat more complicated demonstration of this problem (Sprenger and Stegenga, 2017, p. 847), but the fundamental point is the same: a relative measure alone can neither determine risks, nor the difference in risks that appears in equation 1, leaving room for error in the would-be decision.

Jäntgen (2023) considers situations where A and B are tested in separate trials against other control treatments, and no RD comparing A against B is available. Jäntgen shows that RDs established in separate trials do not suffice for rational decisions, because the RDs calculated against other (not-A nor B) control treatments do not determine an RD between A and B. She then shows that RDs or RRs from the separate trials together with control group risks do suffice, by allowing one to construct the decision-relevant RD. For example, from knowing an RD between A and comparator A' together with the trial-specific baseline risk $p(Y = 1|A')$, one obtains $p(Y = 1|A)$ as the sum of the RD and $p(Y = 1|A')$. $p(Y = 1|B)$ is obtained similarly from an RD between B and B' together with $p(Y = 1|B')$. From $p(Y = 1|A)$ and $p(Y = 1|B)$ one gets an RD between A and B. If $p(Y = 1|A') = p(Y = 1|B')$ holds in both trial populations and is equal between the populations,

the *RD*s from the trials alone provide the *RD* between *A* and *B* (Jäntgen, 2023, p. 1190). The basic point remains the same: *RD* has overall superior policy-relevance since there is no case where a relative measure instead of *RD* would be needed to obtain the decision-relevant information, but there are conditions in which *RD*s alone, but not relative measures alone, are enough to obtain that information.

Sprenger and Stegenga's and Jäntgen's examples consider settings where the outcome of interest is beneficial, e.g. recovery from disease. But the same reasoning applies also to choices between policies to prevent an exposure that causes a harmful outcome: one should choose preventive policy *A* over policy *B* iff $EU(A) > EU(B)$, which requires comparing the risk of outcome under the exposure status associated with policy *A* to that associated with *B*. When the outcome is harmful, we may still apply the exact same model by interpreting *RD* in equation 1 as difference in survival probabilities, or by simply recoding the outcome. So intepreted, the same condition holds for choosing a prevention strategy: to rationally choose policy *A* over *B*, the difference in the survival probabilities conditional on exposures associated with *A* and *B* must satisfy equation 1.

4 Decision-making and extrapolation

Sprenger and Stegenga (2017) and Jäntgen (2023) present compelling decision-theoretic arguments in favor of absolute effect measures. But note that the arguments assume that the absolute measure at hand, whether reported from a single study or constructed from results of separate ones, does apply to the target patient or population. For if it does not, then there is room for error just like with relative measures, and the conclusion about the superiority of absolute measures does not follow. In reality, a summary effect reported in an epidemiological study usually does not apply to a particular target patient or population as is. Rather, the evidence that clinicians, patients, or policymakers work with is empirical estimates typically obtained in a different population than the one that is the target of the decision. Based on this evidence, one must then infer the target risks, or the *RD*, that suffice for a decision in the way described by (Sprenger and Stegenga, 2017). This complicates things in two ways.

Firstly, there is uncertainty due to estimation error. I will not comment on this problem further

than to point out that *NNT* has undesirable properties in this respect: as the underlying *RD* approaches zero, *NNT* approaches positive or negative infinity, resulting in practically meaningless confidence intervals whenever the confidence interval of the *RD* incorporates zero or is bounded close to it (McAlister, 2008). Secondly, one faces the problem of extrapolation: even if an *RD* estimate equals the causal *RD* in the source population, how to know if the causal claim so established applies to a different target? If the *RD* extrapolated from the source does not apply to the target, decisions based on it may turn out to be suboptimal.

There is a substantial cross-disciplinary body of literature on extrapolation, and no comprehensive review is attempted here. Instead, I give a broad outline of two general approaches that more specific proposals can be seen as exemplifying, mostly based on sources from the philosophy of science literature on the topic.

One approach, that I call the ”causal approach”, focuses on other causes of the outcome than the exposure that may modify the exposure's effect (e.g. Bareinboim and Pearl, 2013; Cartwright and Hardie, 2012). In particular, some such causes are co-factors that interact conjunctively with the exposure, such that certain configurations of the co-factors and the exposure suffice to bring about the outcome, but no proper part of such configuration will suffice (Baumgartner and Falk, 2018; Cartwright and Hardie, 2012; Deaton and Cartwright, 2018; Mackie, 1974; Rothman, 2012, chapter 3.). Distributions of such co-factors may vary from one population to the next, manifesting as variation in the exposure's effect. There may also be sufficient configurations of causes that do not include the exposure at all and distributions of these may vary, causing differences in baseline risk. In the face of such threats to extrapolation, the idea of the causal approach is to find a conditioning set of other causes of the outcome that renders the exposure-outcome relation independent of other influences. A detailed proposal of this sort is described in (Bareinboim and Pearl, 2013).

I call the other broad approach the ”mechanism approach”. Something like it is proposed for example in (Steel, 2007; Stegenga, 2015; Tonelli and Williamson, 2020; Wilde and Parkkinen, 2019; Williamson, 2019). This approach is similar to the causal approach – one should control for causally meaningful differences between source and target – but comes with a heuristic for identifying the relevant differences: one should trace the exposure's mechanism of action in the source and in the target, and compare them for similarity. Differences in the mediating mechanism

itself may result in differences in the effect, as may differences in factors capable of interfering with the mechanism's parts. It is these differences that one should search and control for when extrapolating a causal effect. The mechanistic heuristic makes the extrapolation problem more tractable by constraining the search for factors to control for, at the cost of risking a mistake when the relevant factors bear no obvious relation to the exposure's mechanism of action. Note that the suggested heuristic for extrapolation is a flipside of a heuristic for explaining effect heterogenity: when observing effect heterogeneity, one should look for differences in the mediating mechanisms, or differences in factors that interfere with the mechanisms.

I say more about these approaches in section 7. The following section describes a type of effect heterogeneity that is particularly challenging for both approaches.

5 Entanglement of effect with baseline risk

RD is dependent on baseline risk as a matter of definition. To see this intuitively, note that risk is bound between zero and one. Hence, only a small increase in risk for the exposed is possible for high baseline risk, and conversely for preventive exposures and low risks. *RD*s for the same exposure-outcome pair will thus vary across populations or individuals that differ in risk profile, even when there is no heterogeneity due to causal interaction between the exposure and covariates, confounding, or differences in mechanism of action. To illustrate, I use a hypothetical example introduced by Huitfeldt (Huitfeldt, 2019) and redescribed and analyzed by Cinelli and Pearl (2021) and Colnet et al. (2023). My description follows Colnet et al. (2023) with minor modifications.

A team of hypothetical epidemiologists is trying to answer the following question: What effect does playing a single round of Russian roulette have on total mortality? We assume that the epidemiologists are unaware of how Russian roulette works, and rely on running a cohort study – perhaps an RCT – to estimate the effect.

The example is of course unrealistic, but not dissimilar in structure to realistic scenarios. Say that a new, highly potent opioid is introduced to the illegal drug market and pushed to addicts who lack the means to measure doses accurately: each dose comes with a risk of deadly overdose. An obvious thing that public health authorities would want to know is the effect of the drug on mortality

in the addict population. For a historical example of a structurally identical model as presented below, but applied to real world data, see Sheps (1959). The point of imagining Russian roulette as the exposure is that the mechanism of Russian roulette allows for an unambiguous description of the risk-generating process. The elements of the example are thus a population where some are exposed to playing Russian roulette (perhaps through random assignment) and others are not, together with the following stipulated conditions:

- Dichotomous outcome Y (0: survival; 1: death).
- Identical baseline risks: Every subject has the same risk of dying from other causes than exposure during follow-up.
- Exposure's mechanism of action: one of six chambers of a revolver is loaded with a bullet, giving a $\frac{1}{6}$ chance of death.
- Monotonicity: the exposure can only cause death, no subject who would die unexposed would be saved by exposure.
- Homogeneity: There is no causal interaction with covariates, exposure affects every subject the same way.
- No confounding.

Given the above assumptions, the process that generates risk of death can be described with

$$p(Y^a = 1) = b + a \cdot \frac{1}{6}(1 - b) \tag{2}$$

. Equation 2 is slightly adapted from (Colnet et al., 2023), and should be read as follows. $p(Y^a = 1)$ is the marginal risk of death, averaged over both (counterfactual) exposure assignments as indicated by the superscript a. b is the baseline risk, i.e. the probability of dying from other causes than Russian roulette. Exposure status is denoted by the dichotomous exposure indicator a (0: not exposed; 1: exposed). $\frac{1}{6}$ is the probability of dying from exposure, and assumed to apply similarly

to everyone (homogeneity). This risk can however only be realized given exposure, $a = 1$, and given that the baseline risk is not realized, hence the multiplication by $(1-b)$.

Equation 2 also describes risks in the exposed and the unexposed, when a is assigned to equal either 1 or 0. Risk in the unexposed is simply $R_0 = p(Y^{a=0} = 1) = b$: an unexposed subject cannot die from exposure, so their risk of death is the baseline risk b. Risk in the exposed requires a bit of explanation. Like every other subject, each exposed subject may die of other causes than Russian roulette during follow-up, i.e. the baseline risk b applies. The exposed can also die as a consequence of playing Russian roulette, with probability $\frac{1}{6}$. A precondition for this risk being realized is that the baseline risk does not realize. Hence, the additional risk that applies to the exposed over and above the baseline risk is $\frac{1}{6}(1-b)$, and the risk in the exposed becomes $R_1 = p(Y^{a=1} = 1) = b + \frac{1}{6}(1-b)$. Risks in exposed and unexposed give a risk difference of

$$RD = R_1 - R_0 = \underbrace{b + \frac{1}{6}(1-b)}_{R_1} - \underbrace{b}_{R_0} = \frac{1}{6}(1-b)$$

which, as expected, is exactly the risk increase beyond baseline risk inflicted by exposure.

Colnet et al. (2023) use the generative process described by equation 2 to show how the risk difference is a function of baseline risk, evident in the presence of b in the RD above. Say that the epidemiologists run a perfectly designed and executed cohort study in a population with baseline risk of $b = 0.01$, thus obtaining $RD = 0.165$. Even assuming homogeneity and absence of confounding, this effect would not be transferrable to a target population with a different level of baseline risk. For a higher baseline risk of 0.1, for example, this RD predicts that risk in the exposed is 0.265, or 26.5%, but this is an overestimate. For the 26.5% risk to apply in the population with higher baseline risk, the exposure would there have to kill a larger proportion of subjects who would have otherwise survived, since the proportion of subjects at risk of experiencing the exposure's effect is smaller in the higher baseline population. That is, the exposure would have to be more lethal in that population compared to the source population. But this makes no sense – the mechanism of the exposure is assumed to be the same for every individual, everywhere. With baseline risk of 0.1, the true risk in the exposed is in fact 0.25, giving an $RD = 0.15$. On the RD scale, the effect of Russian roulette thus appears heterogeneous after all, even though a key assumption of the generative

process is that the exposure affects every subject the same way. This apparent heterogeneity is due to entanglement of the *RD* with baseline risk. Of note, entanglement is also present on the *RR* scale, so if we imagine that the Roulette study was one of the primary studies fed to Jäntgen's (2023) procedure of constructing target *RD*s from separate trial population *RD*s or *RR*s, the constructed target *RD* would be incorrect either way, unless all baseline risks are equal.

Instead of *RD* (or *RR*), the scale on which the homogeneous effect of Russian roulette becomes evident is the survival ratio scale (Colnet et al., 2023; Huitfeldt, 2019). This can quickly be verified by recalling the definition of survival ratio

$$SR = \frac{1-R_1}{1-R_0}$$

, which for the example makes

$$SR = \frac{1-\left(b+\frac{1}{6}(1-b)\right)}{1-b} = \frac{5}{6}$$

: the exact level of baseline risk b is irrelevant to *SR*, and *SR* is therefore transferrable across populations that differ in baseline risk. The $SR = \frac{5}{6}$ is also immediately interpretable in light of the mechanism of Russian roulette, while the *RD* is not.

Generally speaking, there is no scale that by default provides transportable results, and in particular, no reason to believe that the popular *RR* scale is always appropriate (see Broadbent, 2013, chapter 9.). But as Mindel Sheps (1958; 1959) pointed out half a century ago, there are certain situation-specific patterns. For monotonically risk-increasing exposures, the *SR* scale is least affected by baseline risk. The reason is that in that situation, *SR* tracks the probability of avoiding an outcome among those who are at risk of experiencing it due to the exposure, i.e. among those for whom the baseline risk is not realized (more on this in section 7). For preventive exposures the *RR* scale tends to be appropriate for similar reasons. Indicative of these patterns is that *SR* cannot predict invalid risks for a risk-increasing exposure but *RR* can, and the other way around for risk-decreasing exposures. Recoding the outcome reverses these relations. *RD* is worse than either ratio measure in that it can predict risks outside $[0,1]$ no matter the nature of the exposure.

This alone says little about transportability except that patently absurd extrapolation errors are sometimes avoidable by choosing the right measure. Sheps's deeper insight concerns the connection between effect measures and population parameters of epidemiological interest: how effect measures computable from observed data track such parameters depends on the nature of the exposure and the coding of the outcome. In the example, the probability of surviving exposure among those at risk of dying from it is identified by *SR*, making *SR* transportable between populations that are equal in this parameter, regardless of baseline risk. Ideas akin to Sheps's are formalized and analyzed in (Huitfeldt et al., 2018, 2019, 2021), who describe relations between effect measures and particular population parameters defined in a counterfactual causal model. I briefly return to these ideas in section 7.

The conclusions from the present section are the following. Effect measures vary in transportability in situation-specific ways. *RD*s will always, and ratio measures sometimes, be unstable across varying baseline risks, and hence variably risky to extrapolate.

6 Implications for decision-making

Entanglement of *RD* with baseline risk may compromise the use of study-population *RD* estimates as direct input into decision-making. However, the error in extrapolating an *RD* across baseline risks is of course not always so large that it would lead to suboptimal decisions, and it is possible to gauge how tolerant rational decision-making is to such error. The problem is rather that it is difficult to assess when the extrapolation error is within those tolerable bounds. Note also that even though the absolute size of the error might appear small when the baseline risk difference is small, its impact can still be significant especially in public health settings where costly, population-wide interventions are at stake.

More formally, note that the extrapolation error ε is just the difference between the extrapolated *RD*, denoted RD_{source}, and the true *RD* in the target, RD_{target},

$$\varepsilon = RD_{source} - RD_{target}$$

and RD_{target} equals the difference between the extrapolated RD_{source} and the extrapolation error

$$RD_{target} = RD_{source} - \varepsilon$$

. Now recall from Sprenger and Stegenga (2017) that $EU(A) > EU(B)$ iff equation 1 is satisfied for the relevant *RD*. Thus, using the extrapolated RD_{source} in decision-making leads to mistakenly choosing treatment or policy *A* over *B* iff equation 1 holds for RD_{source} but does not hold for the directly unknown RD_{target} between *A* and *B*. To avoid such a mistake, equation 1 must hold also after correcting for ε. This bounds the benign size of the extrapolation error:

$$\overbrace{RD_{source} - \varepsilon}^{RD_{target}} > \frac{a-b}{u-u'}$$
$$\varepsilon < RD_{source} - \frac{a-b}{u-u'}$$

.

In words, the extrapolation error must be smaller than the difference between the extrapolated RD_{source} and the costs-to-utilities ratio. To check this result against intuition, consider the extreme case where the costs are equal: $a = b$. In this situation, one should choose one treatment or policy over the other as long as the relevant *RD* is above zero in the target population, meaning that the extrapolation error is smaller than the extrapolated RD_{source} by any amount. When $a = b$, $\frac{a-b}{u-u'}$ becomes zero, and the benign size of ε is bounded by

$$\varepsilon < RD_{source}$$

, as expected.

The smaller the difference between RD_{source} and the costs-to-utilities ratio, the smaller the maximum extrapolation error that rational decision-making tolerates. The problem is that it is difficult to know the exact value ε; if one knew it, one would already know the correct RD_{target}. Prominent approaches to extrapolation provide limited relief, as explained in section 7 below. Rough

guesses concerning the size of the error are possible if one knows baseline risks in both populations, though. If an exposure increases risk and the baseline risk in target is higher (respectively, lower) than in the source, RD_{target} is probably smaller (larger) than RD_{source}. The opposite applies for risk-decreasing exposures. Much uncertainty will remain regardless, even in unrealistic circumstances where other effect modifiers are equally distributed.

These problems do not always necessitate running a new study in the target population to obtain RD_{target}. It may be possible to extrapolate an effect on a different scale than the *RD* scale, and combine that with baseline risk information in the target to calculate RD_{target}, give or take estimation error. The running example provides illustration. In the example, a source population baseline risk was assumed to be 0.01, giving $RD_{source} = 0.165$, while a hypothetical target population has baseline risk of 0.1 and $RD_{target} = 0.15$, and direct extrapolation of RD_{source} is defeated by the baseline risk difference. Since *SR* is in this case insensitive to baseline risk, *SR* can be extrapolated between the populations with no adjustment. Given baseline risk of 0.1 in the target, the baseline survival probability to which the modification by exposure applies is $1 - 0.1 = 0.9$. Applying the extrapolated *SR*, the survival probability in the exposed in the target is $0.9 \cdot \frac{5}{6} = 0.75$, risk in exposed is $1 - 0.75 = 0.25$, and the correct $RD_{target} = 0.25 - 0.1 = 0.15$.

Of course, in a real-world application even the most transportable measure would probably require some adjustment for heterogeneity, which can have other sources than baseline risk dependence. The purpose of the example is merely to demonstrate that it is not obvious that a study population *RD* is optimal evidence for decision-making, all things considered. Depending on the nature of the exposure, adjusting a RD_{source} approriately may be much more difficult and costly than adjusting a relative measure and obtaining an estimate of baseline risk in the target. To adjust RD_{source}, one needs to control for all dissimilarly distributed causes of the outcome that are not screened off the outcome by the exposure of interest; a lot of causal and distributional information about the target is needed. If one can identify a relative effect that is at least approximately free of baseline risk dependence, only some of that information is needed, together with observations of the outcome to estimate baseline risk.

This goes to show that the decision-theoretic conclusion about the superiority of absolute effects does not generally apply to actual decision-making where extrapolation error is an issue; a decision-

maker may well be better off relying on an appropriately chosen relative effect together with a baseline risk estimate in the target, rather than trying to adjust an *RD*. One should certainly not read the decision-theoretic arguments as implicating that epidemiological studies should preferentially report absolute effects rather than relative ones. When it comes to reporting, the advice of textbook epidemiology seems apt: whenever study design allows estimating risks, the primary result to report should be risks (Rothman, 2012, p. 84). Risk information allows the user of the evidence to calculate any summary effect deemed appropriate for the purpose of calculating exposed risk in their target population. This conclusion is similar to Jäntgen's, who recommends always reporting baseline risks alongside *RD*s or *RR*s, but for different reasons.

Recall that Jäntgen considers a situation where the decision-relevant risks or *RD* must be derived from results of separate trials testing A and B against other control treatments. The baseline risks are risks given control treatments A' and B', only observed in the trial populations where A and B were tested, respectively. Jäntgen suggests calculating $p(Y = 1|A)$ from an *RD* or *RR* and baseline risk $p(Y = 1|A')$ in the population where A was tested, and $p(Y = 1|B)$ the same way in the population where B was tested. From $p(Y = 1|A)$ and $p(Y = 1|B)$, one gets an *RD* between A and B that is applied to the target. In the unlikely case where the baseline risks are equal within and between the trial populations, the *RD* between A and B can be calculated from the trial *RD*s alone. (Jäntgen, 2023, p. 1190).

To do the above in practice would require not one but many extrapolations of risks and summary effects across populations. For things to go wrong, it is enough that either baseline risk $p(Y = 1|A')$ or $p(Y = 1|B')$ differs between the target and the trial population where it was observed. This is true no matter what summary effect is used to calculate $p(Y = 1|A)$ and $p(Y = 1|B)$ from the trial baseline risks: if either baseline risk varies, then at least one of the calculated $p(Y = 1|A)$ or $p(Y = 1|B)$ will not apply to the target. It is usually dubious to assume that a risk estimated in a trial applies to a different target population as is, let alone that many risks from separate trials do. In practice, then, one would first try to estimate $p(Y = 1|A')$ and $p(Y = 1|B')$ in the target, or if that is not feasible, adjust the trial risks for prognostic factors that predict their level in the target. Then one would extrapolate summary effects from the trial populations to calculate $p(Y = 1|A)$ and $p(Y = 1|B)$ that one hopes to apply to the target. Any effect measure, absolute or relative, will do,

but as the preceding discussion suggests, *RD* may well be a riskier choice than some relative effect.

None of this of course means that Jäntgen's modeling results are invalid. It is just that decision-makers are rarely in a situation where they can assume that baseline risks *and* absolute effects observed in trial populations constrained by strict exclusion criteria apply to the intended target without adjustment. Since one normally cannot assume this, one would normally not apply the reasoning described in (Jäntgen, 2023) in practice. The modeling results therefore do not demonstrate superior policy-relevance of absolute measures. The advice to always report risks, however, is certainly reasonable, if not for exactly the reasons that Jäntgen gives.

The difference in motivation behind Jäntgen's proposed procedure and what I would recommend is hopefully now clearer. Jäntgen recommends using the trial-specific baseline risks alongside the trial *RD*s or *RR*s to obtain $p(Y = 1|A)$ and $p(Y = 1|B)$ that one applies to the target. I suggest using the risks obtained in the trials to calculate a summary effect that, in light of what is known of the nature of the exposure or treatment, is least sensitive to inevitable differences in baseline risks, and thus transportable with fewest adjustments. Whichever measure is believed to be least sensitive should then be used to calculate $p(Y = 1|A)$ and $p(Y = 1|B)$ in the target from estimates of the target baseline risks. As study reporting goes, the crucial information decision-makers need is risks, not summary effects, absolute or otherwise, as the latter can easily be calculated from risks when needed.

As a final point from this section, note that *RD* is a far from ideal basis for a decision even if estimation or extrapolation error is not an issue: Relying on *RD* assumes that expected utility scales linearly with risk. E.g. for a decision taken on $RD = 0.05$, a risk difference between 0.05 and 0.1 must have the same impact on expected utility as the difference between 0.35 and 0.4, or 0.5 and 0.55, and so on. Whenever possible, instead of *RD* or any other summary effect, one should use the risks under different exposure alternatives to calculate expected utilities of the choice-alternatives separately, as then the decision can accommodate any utility function (cf. Huitfeldt et al., 2021, Appendix 1). Given an estimate of the baseline risk in the target, one can calculate risk under the exposure or treatment of interest based on an extrapolated effect measure, which can be either absolute or relative. When the baseline risk in the source population differs notably from that in the target, it may be safer to use either *RR* or *SR* than *RD* for this purpose.

7 Implications for theories of extrapolation

For the mechanism approch to extrapolation, effect heterogeneity implies differences in the mediating mechanisms, or differences in factors capable of interfering with components of the mechanism. To extrapolate an effect, one should control for such differences. But heterogeneity that is due to the effect's dependence on baseline risk cannot be explained by such mechanistic differences, and investing effort into finding them would be futile. The mechanistic heuristic for solving the problem of extrapolation is thus unlikely to work in such cases. In the running example, the hypothetical epidemiologists would be unable to appropriately adjust RD_{source} even if they learned everything about the mechanism of action in both populations; the mechanisms are identical by stipulation. Given the simplistic structure of the example, they would of course be able to calculate any effect measure in the target based on learning the mechanism and the target baseline risk, but this would not involve extrapolation.

The causal approach correctly suggests that heterogeneity due to baseline differences indeed has a causal explanation: it is explained by differences in factors that cause differences in baseline risk. Thus, an *RD* estimate could be extrapolated if it was adjusted for a set of other causes of the outcome that suffices for isolating the effect of the exposure in the target. But such an adjustment can be hard to apply: the choice of the conditioning set would often have to rest on many unsubstantiated causal assumptions, and even if those were correct, data availability would pose another problem.

Finally, neither approach offers much guidance for identifying a scale on which an effect would be transferrable with fewest adjustments, like the *SR* scale in the example. This highlights a point urged by epidemiologic methodologists but ignored by philosophical commentators: different effect measures have different transportability properties that depend in complex ways on the objective nature of the (disease) phenomenon of interest and the subjective choice of its coding as an outcome variable in data (e.g. Deeks, 2002; Doi et al., 2022; Huitfeldt et al., 2018, 2019; Panagiotou and Trikalinos, 2015; Senn, 2011; Sheps, 1958; Webster-Clark and Keil, 2023). While philosophers, too, have touched upon this topic, it is only to debunk the idea that *RR* measures are transportable by default, not to systematically study the conditions that transportability of effects depends on (e.g. Broadbent, 2013; Fuller, 2021).

This obviously does not mean that the two approaches are false in that they could never work. To the contrary, any viable solution to the extrapolation problem must involve studying causally relevant differences somehow. The issue is rather that these approaches ignore the fact that other causes of the outcome only become relevant to the extrapolation problem relative to the chosen effect measure. If that choice is ignored, an extensive search for differences in the causes of the outcome, whether constrained by the mechanistic heuristic or not, risks either focusing on the wrong things, or being too demanding in practice. It may be that a more tractable problem where auxiliary causal-mechanical evidence is of use is that of identifying a measure that is least sensitive to context-specific causal detail (Huitfeldt et al., 2018). This proposal deserves attention from philosophers of science.

Briefly, for the running example this idea works as follows. Consider a distribution of counterfactual outcomes given exposure, $Y^{a=1}$, and absence of exposure, $Y^{a=0}$, for each subject. Monotonicity is assumed: Russian roulette can only cause death, not prevent death for anyone, so the individual effect $(Y^{a=0}=1, Y^{a=1}=0)$ is impossible and $p(Y^{a=0}=1, Y^{a=1}=0)=0$. The exposure's effect can hence only be realized in subjects who would survive if unexposed, $Y^{a=0}=0$, and the risk of death among those at risk of experiencing the effect is then $p(Y^{a=1}=1|Y^{a=0}=0)$. This is one of the components of the risk of death in the exposed, $p(Y^{a=1}=1)$. The other component is the risk of death from exposure among those destined to die anyway $p(Y^{a=1}=1|Y^{a=0}=1)$. Finally,

$$p(Y^{a=1}=1) = p(Y^{a=1}=1|Y^{a=0}=0)p(Y^{a=0}=0) + p(Y^{a=1}=1|Y^{a=0}=1)p(Y^{a=0}=1)$$

.

As the baseline risk $p(Y^{a=0}=1)$ increases, $p(Y^{a=0}=0)$ decreases and $p(Y^{a=1}=1|Y^{a=0}=0)$ contributes less to $p(Y^{a=1}=1)$, shrinking the causal risk difference $[p(Y^{a=1}=1)-p(Y^{a=0}=1)]$. This is just the entaglement that the example aims to illustrate. Hence, for an *RD* to be equal between two populations with different baseline risks, a larger proportion of subjects at risk of experiencing the exposure's effect would have to die in the population with higher baseline risk. This does not make sense given the assumed mechanism of Russian roulette: Each subject's response to exposure is determined only by the exposure's

mechanism of action, which is the same for all, and no other factors affect susceptibility.

The assumed mechanism of Russian roulette entails two stable counterfactual probabilities: the risk of dying from exposure among those at risk of dying from it, $p(Y^{a=1} = 1|Y^{a=0} = 0) = \frac{1}{6}$, and that of surviving the exposure among the same subjects $p(Y^{a=1} = 0|Y^{a=0} = 0) = \frac{5}{6}$. Huitfeldt et al. (2018, pp. 4-5, appendix B) show that given a monotonically risk-increasing exposure, the latter probability is identified in data by *SR*, and the former is not identified by any common effect measure. Huitfeldt et al.'s proof follows from more general principles that entail analogous results for other effect measures, but a short proof about *SR* only can be given as follows.

Given monotonicity, i.e. no subject is saved from death by exposure, there cannot be subjects who would die in absence of exposure among those who would survive the exposure, giving $p(Y^{a=0} = 0|Y^{a=1} = 0) = 1$. Then,

$$p(Y^{a=1} = 0|Y^{a=0} = 0) = \frac{p(Y^{a=0} = 0|Y^{a=1} = 0)p(Y^{a=1} = 0)}{p(Y^{a=0} = 0)} = \frac{p(Y^{a=1} = 0)}{p(Y^{a=0} = 0)} = SR$$

: $p(Y^{a=1} = 0|Y^{a=0} = 0)$ equals the causal survival ratio, estimated by $\frac{1-R_1}{1-R_0}$. Hence, *SR* is stable across populations as long as $p(Y^{a=1} = 0|Y^{a=0} = 0)$ is, which it will be unless there are other factors that determine susceptibility than the mechanism of action itself, and those vary between populations. If there were such factors, the *SR* could still be extrapolated controlling for those factors, without controlling for any other causes of the outcome. Had the exposure been one that decreases risk, or the coding of the outcome been the opposite, *RR* would be transportable, and *SR* not. In neither case would *RD* be transportable without adjusting for numerous causes of baseline risk, unless baseline risks are equal.

The approach described in Huitfeldt et al. (2018) links stability of effect measures to parameters that describe probabilities that subjects' counterfactual outcomes stay unchanged when their exposure status changes. The result about *SR* illustrated above follows from these considerations. This suggests an approach to extrapolation that focuses on the co-factors of the exposure that determine susceptibility to the exposure's effect, rather than all other causes of the outcome or a subset comprising the mechanism of action and its interfering factors. The extrapolation problem is still demanding: If one's qualitative understanding of the mechanism of action is limited, one

may misidentify the factors that determine susceptibility. But if the relevant co-factors can be determined, one may then justify assumptions about stabilities of particular counterfactual population parameters based on the estimated or assumed distributions of the co-factors. These justify beliefs about conditional transportability of different effect measures, where the conditioning set includes just those co-factors whose distribution differs between populations.

This proposal is contentious, and I make no claims of it being generally more applicable than any alternative; no philosophical argument alone can establish such claims. Its philosophical significance is in showing how the difficulty of extrapolating a causal claim about an exposure-outcome pair can vary significantly depending on the exact specification of the effect that is being transported. Philosophers of science have hitherto ignored this fact.

8 Discussion and conclusions

I have argued that the decision-theoretic results of Sprenger and Stegenga (2017) and Jäntgen (2023) do not entail that absolute effect measures are generally superior to relative ones in policy-relevance. There are circumstances where relative measures have preferable properties from a decision-making perspective, even when an estimate of an absolute measure is known. This is because an effect found in a particular study population is rarely usable as direct input into a decision, due to inevitable differences in risk profile between populations and individuals. Rather, study-population effect estimates, together with auxiliary evidence from the target, are used to infer target risks or summary effects, and the latter are the actual input to a decision about the target. It is thus the extrapolatability of effect measures that mostly matters for decision-makers. By definition, an *RD* (or *NNT*) for a given exposure-outcome pair cannot be stable against variation in baseline risk even in unrealistic conditions where there is no confounding, no heterogeneity due to causal interaction with covariates, no differences in exposure's mechanism of action, and no estimation error. To rely on an *RD*, a decision-maker would possibly have to adjust the estimate for numerous causes of baseline risk, many of which may be unknown at the time of the decision.

While *RD*s must vary to some degree across baseline risks, it is of course partly an empirical question how often this heterogeneity is of degree that jeopardizes policy-relevance. There are

studies that suggest that *RD*s for a given exposure-outcome relation do indeed vary notably, often more so than ratio measures (e.g. Deeks, 2002; Senn, 2011). But empirical evidence cannot conclusively tell how serious a threat this is to policy-relevance, since the magnitude of error tolerated by rational decision-making depends on the costs and utilities at stake (section 6). It also seems that empirical evidence about transportability of different measures will always remain contestable (see Poole et al., 2015).

General statements about the policy-relevance of particular measures are thus somewhat dubious. Rather than making such claims, I want to highlight the following: Given the definitions of common effect measures, it is inevitable that there are scenarios that present tradeoffs between them. When differences in baseline risks prohibit extrapolating *RD* without controlling for numerous other causes of the outcome, there may be a relative effect that can be extrapolated relying on a smaller conditioning set. Extrapolating such a relative measure, when one exists, is thus less risky, and when combined with baseline risk information in the target, provides the basis for a rational decision. In that case, a decision-maker is better off relying on the relative measure and estimating the target baseline risk directly, rather than attempting to adjust the *RD*. In a different scenario, maybe no clear tradeoff to exploit can be identified, and one may try to adjust the *RD*.

If any general recommendation can be made about what to report, it would be to always report all estimated risks, as this allows the user of the evidence to calculate any summary effects deemed appropriate for their use case. This conclusion is similar to that of Jäntgen (2023), who recommends always reporting baseline risks, but for quite different reasons, as explained in section 6.

The preceding discussion has implications also for accounts of extrapolation (section 7). The philosophy of science literature on the topic has so far paid little attention to the choice of effect measure. But that choice has consequences for extrapolation: Different effect measures have different transportability properties depending on the process that generates the outcome, and the coding of the outcome in data. It may be that the causal-mechanical evidence that purportedly solves the problem of extrapolation is sometimes best employed in choosing which effect measure to use in the first place.

References

Bareinboim, E. and J. Pearl (2013). A general algorithm for deciding transportability of experimental results. *Journal of causal Inference 1*(1), 107–134.

Baumgartner, M. and C. Falk (2018). Boolean difference-making: a modern regularity theory of causation. *The British Journal for the Philosophy of Science*.

Broadbent, A. (2013). *Philosophy of Epidemiology*. London: Palgrave Macmillan.

Cartwright, N. and J. Hardie (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Cinelli, C. and J. Pearl (2021). Generalizing experimental results by leveraging knowledge of mechanisms. *European Journal of Epidemiology 36*, 149–164.

Colnet, B., J. Josse, G. Varoquaux, and E. Scornet (2023). Risk ratio, odds ratio, risk difference... which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*.

Deaton, A. and N. Cartwright (2018). Understanding and misunderstanding randomized controlled trials. *Social science & medicine 210*, 2–21.

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.

Doi, S. A., L. Furuya-Kanamori, C. Xu, L. Lin, T. Chivese, and L. Thalib (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.

Fuller, J. (2021). The myth and fallacy of simple extrapolation in medicine. *Synthese 198*, 2919–2939.

Hernán, M. A. and J. M. Robins (2020). *Causal inference: What if*. CRC Boca Raton, FL.

Huitfeldt, A. (2019). Effect heterogeneity and external validity in medicine. https://www.lesswrong.com/posts/wwbrvumMWhDfeo652. Accessed: 19-02-2024.

Huitfeldt, A., M. P. Fox, E. J. Murray, A. Hróbjartsson, and R. M. Daniel (2021). Shall we count the living or the dead? *arXiv preprint arXiv:2106.06316*.

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

Huitfeldt, A., M. J. Stensrud, and E. Suzuki (2019). On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging themes in epidemiology 16*, 1–5.

Huitfeldt, A., S. A. Swanson, M. J. Stensrud, and E. Suzuki (2019). Effect heterogeneity and variable selection for standardizing causal effects to a target population. *European journal of epidemiology 34*, 1119–1129.

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

Kuorikoski, J. (2012). Contrastive statistical explanation and causal heterogeneity. *European Journal for Philosophy of Science 2*, 435–452.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford: Clarendon Press.

McAlister, F. A. (2008). The "number needed to treat" turns 20—and continues to be used and misused. *Canadian medical association journal 179*(6), 549–553.

Panagiotou, O. A. and T. A. Trikalinos (2015). On effect measures, heterogeneity, and the laws of nature. *Epidemiology 26*(5), 710.

Poole, C., I. Shrier, and T. J. VanderWeele (2015). Is the risk difference really a more heterogeneous measure? *Epidemiology 26*(5), 714–718.

Rothman, K. J. (2012). *Epidemiology: an introduction*. Oxford university press.

Senn, S. (2011). U is for unease: Reasons for mistrusting overlap measures for reporting clinical trials. *Statistics in Biopharmaceutical Research 3*(2), 302–309.

Sheps, M. C. (1958). Shall we count the living or the dead? *New England Journal of Medicine 259*(25), 1210–1214.

Sheps, M. C. (1959). An examination of some methods of comparing several rates or proportions. *Biometrics 15*(1), 87–97.

Sprenger, J. and J. Stegenga (2017). Three arguments for absolute outcome measures. *Philosophy of Science 84*(5), 840–852.

Steel, D. (2007). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.

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.

Tonelli, M. R. and J. Williamson (2020). Mechanisms in clinical practice: use and justification. *Medicine, Health Care and Philosophy 23*, 115–124.

Webster-Clark, M. and A. P. Keil (2023). How choice of effect measure influences minimally sufficient adjustment sets for external validity. *American Journal of Epidemiology 192*(7), 1148–1154.

Wilde, M. and V.-P. Parkkinen (2019). Extrapolation and the Russo–Williamson thesis. *Synthese 196*(8), 3251–3262.

Williamson, J. (2019). Establishing causal claims in medicine. *International Studies in the Philosophy of Science 32*(1), 33–61.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford university press.