Effect heterogeneity and variable selection for standardizing causal effects to a target population

Anders Huitfeldt (*)[1,2], Sonja A. Swanson[3,4], Mats J. Stensrud[4,5] and Etsuji Suzuki[4,6]

[1]Norwegian Institute of Public Health
[2]PharmacoEpidemiology and Drug Safety Research Group, Department of Pharmacy, and PharmaTox Strategic Initiative, Faculty of Mathematics and Natural Sciences, University of Oslo
[3]Department of Epidemiology, Erasmus MC
[4]Department of Epidemiology, Harvard T.H. Chan School of Public Health
[5]Department of Biostatistics, University of Oslo
[6]Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University

January 1, 2020

Abstract

The participants in randomized trials and other studies used for causal inference are often not representative of the populations seen by clinical decision-makers. To account for differences between populations, researchers may consider standardizing results to a target population. We discuss several different types of homogeneity conditions that are relevant for standardization: Homogeneity of effect measures, homogeneity of counterfactual outcome state transition parameters, and homogeneity of counterfactual distributions. Each of these conditions can be used to show that a particular standardization procedure will result in an unbiased estimate of the effect in the target population, given assumptions about the relevant scientific context. We compare and contrast the homogeneity conditions, in particular their implications for selection of covariates for standardization and their implications for how to compute the standardized causal effect in the target population. While some

of the recently developed counterfactual approaches to generalizability rely upon homogeneity conditions that avoid many of the problems associated with traditional approaches, they often require adjustment for a large (and possibly unfeasible) set of covariates.

1. BACKGROUND

The participants in randomized trials and other studies used for causal inference are often not representative of the populations seen by clinical decision-makers [1]. Several statistical methods have been proposed to standardize estimates of a causal effect to the distribution of baseline covariates in a clinically relevant target population, in order to account for differences between populations. However, less attention has been given to how an investigator should reason about which covariates need to be standardized over. The choice of variables is important not only for the standardization procedure, but also for determining which personal characteristics of the participants must be considered when reasoning qualitatively about how representative a study is, relative to the intended target population. In this paper, we discuss different ways to select such covariates, and show that this problem is closely related to how one chooses to operationalize effect homogeneity between populations.

For all examples, we will consider the effect of a binary treatment A (for example, a pharmaceutical; 1 = treated, 0 = not treated) on a binary outcome Y (for example, a side effect; 1 = occurred, 0 = did not occur). Counterfactuals will be denoted using superscripts. We will let V denote a set of baseline covariates which are potential effect modifiers (for example: gender, nationality, etc). In some examples, we will consider an individual binary potential effect modifier (an element of V), which we will call W. We will consider two separate populations: The study population ($P = s$), in which we have valid evidence for the causal effect of the treatment; and the target population ($P = t$) in which we either have no data on the exposure or outcomes variables, or only have observational data and are unable to rule out confounding. We will consider membership in the study population to be defined at baseline, and focus on issues that arise due to non-random selection into the study. We note that while it is certainly possible that there is selection out of the study post-baseline, this is better considered as a form of selection bias [2, 3] related to internal as opposed to external validity. We will focus primarily on binary outcomes, but note that most methods and concepts discussed in this paper (except counterfactual outcome state transition (COST) parameters) extend readily to continuous and time-to-event outcomes. We consider several measures of causal effect including the risk difference (RD), the risk ratio (RR), the survival ratio (SR) (which can be understood as the RR where the coding of the outcome variable is reversed), and the odds ratio (OR). These effect measures may be defined in a specific population or subgroup, which we denote using subscript as needed. For instance, RD_t is the RD in population t.

Epidemiologists and clinical scientists have traditionally defined effect homogeneity in terms of a specific effect measure. For example, one may consider effect homogeneity as the absence of effect modification on the RR, RD or OR scale. These definitions of effect homogeneity are associated with several established conceptual and practical shortcomings, including lack of biological interpretation, baseline risk dependence, zero-bounds,

prediction outside the range of valid probabilities, non-collapsibility and asymmetry [4]. There have also been several recent methodological developments in defining effect homogeneity based on counterfactual distributions rather than specific measures of effect [5]. These approaches consider the outcome under the active treatment separately from the outcome under the control condition. VanderWeele described this as "effect modification in distribution" [6], to contrast with the traditional approach, which was termed "effect modification in measure".

Both these types of effect homogeneity may occur between two subgroups which are both in the study population (for example: between men in the study population and women in the study population), or between one subgroup which is in the study population and another subgroup which is in the target population (for example: between men in the study population and men in the target population). Homogeneity between two groups that are both in the study population is often invoked in meta-analysis and model specification. Homogeneity between one group that is in the study population and another that is in the target population is necessary in settings where the goal is to extrapolate the findings to settings outside of the observed data ("generalizability" or "transportability"). Conceptually, these types of homogeneity are closely related, and differ primarily in that the former is testable from the observed data, whereas the latter is not.

The paper is organized as follows. First, we consider approaches to generalizability that are based on conditional homogeneity of standard effect measures (such as RR and RD) between the study population and the target population. We then describe the recently introduced COST parameters [4], and show how this framework can be used to overcome some of the shortcomings of traditional effect measures. Finally, we review approaches to generalizability based on conditional homogeneity of individual counterfactual distributions, with a particular emphasis on methods based on inverse probability weighted estimators, and methods based on causal diagrams. As we introduce each approach, we repeatedly refer to two tables throughout the text: Table 1 shows an overview of different ways an investigator can operationalize effect homogeneity; Table 2 shows five different approaches to standardization which rely on different homogeneity conditions. Proofs of the standardization formulas in Table 2 are shown in appendix 1.

Table 1: Definitions of conditional effect homogeneity between study population and target population

Homogeneity condition	Definition
Effect Homogeneity in Measure	
On the risk difference scale	$\mathrm{RD}_{s,v} = \mathrm{RD}_{t,v}$
On the risk ratio scale	$\mathrm{RR}_{s,v} = \mathrm{RR}_{t,v}$
On the survival ratio scale	$\mathrm{SR}_{s,v} = \mathrm{SR}_{t,v}$
On the odds ratio scale	$\mathrm{OR}_{s,v} = \mathrm{OR}_{t,v}$
Homogeneity of COST Parameters	
For introducing treatment	$Y^{a=1} \perp\!\!\!\perp P \vert Y^{a=0}, V$
For removing treatment	$Y^{a=0} \perp\!\!\!\perp P \vert Y^{a=1}, V$
Effect Homogeneity in Distribution	
S-ignorability	$Y^a \perp\!\!\!\perp P \vert V \ \ (\forall a)$
S-admissibility	$Y^a \perp\!\!\!\perp P^a \vert V^a \ \ (\forall a)$

Table 2: Five Approaches to Effect Transportation

	Interpretation of result	Validity Conditions
$$\mathrm{RR}_t = \sum_v \mathrm{RR}_{s,v} \times \Pr(V=v \mid Y^{a=0}=1, P=t)$$	Effect measure in target population. In this specific example, the weights ($\Pr(V=v \mid Y^{a=0}=1, P=t)$) are specific to the risk ratio; similar weights exist for other collapsible effect measures but not for non-collapsible effect measures.	Conditional effect homogeneity in measure. (i.e. V must comprise a sufficient set of effect measure modifiers)
$$\Pr(Y^{a=1}=1 \mid P=t) = \sum_v \left[\Pr(Y^{a=0}=1 \mid P=t, V=v) \times RR_{s,v} \times \Pr(V=v \mid P=t)\right]$$	Average outcome under treatment in target population	Conditional effect homogeneity in measure
$$\Pr(Y^{a}=1 \mid P=t) = \sum_v \left[\Pr(Y^{a}=1 \mid V=v, P=s) \times \Pr(V=v \mid P=t)\right]$$	Average outcome under treatment (or under no treatment) in target population	Conditional effect homogeneity in distribution (i.e. V must comprise a set of variables sufficient to block all paths between P and Y)
$$\Pr(Y^{a}=1) = \sum_v \left[\frac{\Pr(Y=1 \mid A=a, V=v, P=s) \times \Pr(A=a, V=v, P=s)}{\Pr(A=a \mid P=s, V=v) \times \Pr(P=s \mid V=v)}\right]$$	Average outcome under treatment (or under no treatment) in population from which the (non-random) sample was taken	Conditional effect homogeneity in distribution
$$\Pr(Y^{a}=1 \mid P \neq s) = \sum_v \left[\frac{\Pr(Y=1 \mid A=a, V=v, P=s) \times \Pr(A=a, V=v, P=s)}{\Pr(A=a \mid P=s, V=v) \times \frac{\Pr(P=s \mid V=v)}{\Pr(P \neq s \mid V=v)} \times \Pr(P \neq s)}\right]$$	Average outcome under treatment (or under no treatment) in those who were eligible to be selected for the study, but weren't. Note that we do not in general assume that P is binary. In the special situation where P is binary (such that all members of a well-defined source population belong either to the study population or the target population), approach 5 estimates the effect in the same subgroup as approach 3	Conditional effect homogeneity in distribution

Note that in approaches 1 through 3, the identifying expressions for the effect in the target population are written in terms of counterfactual variables in order to focus on the part of the analysis that is made necessary in order to account for heterogeneity between populations. In practice, it will be necessary to find an identifying expression that is written terms of observable quantities, which will require either marginal or conditional exchangeability ($(Y^a \perp\!\!\!\perp A \mid P=s$ or $Y^a \perp\!\!\!\perp A \mid P=s, V$ for $\forall a)$ in the study population. In approaches 4 and 5, the identifying expressions are written in terms of observable quantities; these expressions rely upon conditional exchangeability for their derivation. For simplicity of notation, we are here assuming that the same set of variables is sufficient to account both for confounding in the study population and for differences between the populations.

Note also that the expression in approach 4 can be rewritten on the individual level as $\mathbb{E}\left[\frac{Y \times I(A=a, P=s)}{\Pr(A=a \mid V=v_i, P=s) \times \Pr(P=s \mid V=v_i)}\right]$ in order to illustrate that it can be computed from the data by taking a weighted average in a pseudopopulation where all individuals have been weighted by the inverse of their probability of exposure, and their probability of selection, given their covariates V

2. EFFECT HOMOGENEITY IN MEASURE

Effect homogeneity in measure occurs whenever the effect in one population (or subgroup) is equal to the effect in another population (or subgroup) in terms of a particular effect measure, such as the RD or the RR. For example, if the RD in the study population (i.e. RD_s) is equal to RD in the target population (i.e. RD_t) we say that there is effect homogeneity on the RD scale.

Many commonly used methods in epidemiology rely on assumptions that are equivalent to conditional effect homogeneity in measure. For example, the Mantel-Haenszel estimator only has a clear population-level interpretation if the conditional OR is equal between all strata of the covariates [7] (i.e if there is effect homogeneity in measure between groups in the study population). Epidemiologists also often rely on effect homogeneity in measure when they omit interaction terms from regression models. For example, suppose we fit the the following logistic regression model in the study population:

$$\text{logit } \Pr(Y = 1|A, W, P = s) \ = \beta_0 + \beta_1 A + \beta_2 W.$$

In this model, by omitting a product term $\beta_3 \times A \times W$, we encode the assumption that the OR of A on Y in the group $W = 1$ is equal to the OR in $W = 0$, or in other words, that there is effect homogeneity on the OR scale between two subgroups of the study population: $\text{OR}_{s,w=1} = \text{OR}_{s,w=0}$.

If we are willing to assume homogeneity of an effect measure between two groups in the study population in order to justify the absence of a product term, we may be tempted to ask if we could use a similar homogeneity assumption between one group that is in the study population, and another group that is outside of the study population (e.g. $\text{OR}_{s,v} = \text{OR}_{t,v}$ for all values of v) in order to justify extrapolation of an effect to the target population. In this paper, if such a homogeneity condition holds on any scale, we say that there is conditional effect homogeneity on that scale, and that V is a sufficient set of effect measure modifiers for the transportation from the study population to the target population.

The overall idea behind this approach is to identify a set of measured covariates V such that, within levels of the covariates, the magnitude of the effect (when measured on that particular scale) is equal between the populations. To illustrate, it is possible that the RR for adverse effects of Codeine differs between Norway and Japan because the two countries have different distributions of variants of CYP2D6 [8], a gene associated with drug metabolism, but that on average, the RR associated with the use of the drug is equal between Norwegians and Japanese who have the same variant of the gene. If that is the case, then we have effect measure homogeneity conditional on CYP2D6 variant, and CYP2D6 is a sufficient set of effect modifiers on the RR scale. Of note, a sufficient set of effect measure modifiers may not exist among the measured covariates.

An example of the utility of such homogeneity conditions occurs when an investigator attempts to account for heterogeneity between populations by standardizing effect esti-

mates to the distribution of covariates V in the target population. The first two formulas in table 2 can be used to standardize estimates of a causal effect to a target population, if one has measured a sufficient set of effect modifiers. Approach one, which is a weighted average of the effect measure, is valid for collapsible effect measures [9], whereas approach two, which is a weighted average of the predicted stratum-specific average outcome under treatment, is valid for any effect measure.

A large literature exists on statistical tests for detecting and quantifying any effect heterogeneity in measure between groups in the observed data (for example, between groups in the study population, or between multiple study populations). Examples of this include Cochran's Q test [10] and the I^2 statistic [11]. While homogeneity of an effect measure is to some extent an empirical question [12, 13, 14], convincing arguments for stability of the effect measure outside of the observed data will often require additional, explicit assumptions about the data generating mechanism. Unfortunately, few examples of data generating mechanisms which result in stability of an effect measure exist in the published literature, and in many settings, finding convincing mechanisms may not be feasible. However, in the next subsection, we discuss the COST parameters framework to demonstrate that at least in some settings, such mechanisms can be found.

2.1 COST parameters

COST parameters are a new class of effect parameters that were proposed in order to formalize a counterfactual causal model that may result in effect homogeneity in terms of standard observable measures of effect. The COST parameters for introducing treatment are defined as follows:

$$\mathrm{G} = \Pr(Y^{a=1} = 1 \mid Y^{a=0} = 1)$$

$$\mathrm{H} = \Pr(Y^{a=1} = 0 \mid Y^{a=0} = 0)$$

The COST parameters can be understood as the proportion of cases and non-cases that would not have had the opposite outcome if their exposure status had been altered. In other words, these are the probabilities that the outcome does not "switch" in response to treatment (see Figure 1). In Huitfeldt et al [4], it was shown that if certain cofactors that determine treatment effect have equal prevalence between two groups, and if the interaction between these cofactors and treatment A operates according to certain simple biological principles, then the COST parameters for introducing treatment are equal between populations, which can mathematically be written as $Y^{a=1} \perp\!\!\!\perp P | Y^{a=0}$. If the cofactors instead interact with treatment according to a different biological mechanism, this would instead result in homogeneity of COST parameters for removing treatment ($Y^{a=0} \perp\!\!\!\perp P | Y^{a=1}$).

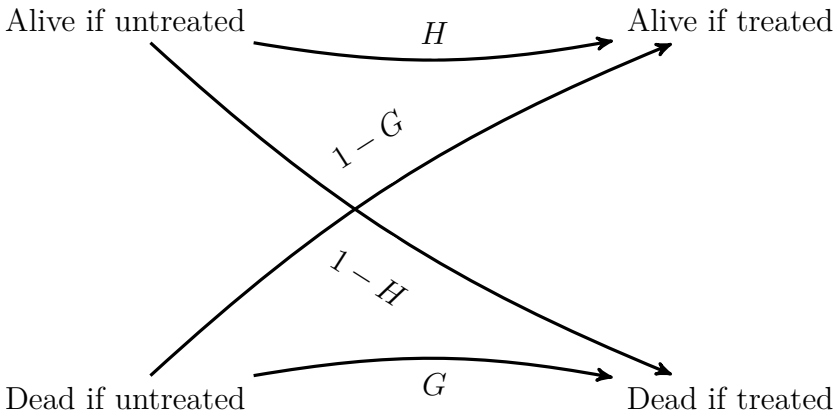

Figure 1: Counterfactual outcome state transition parameters associated with introduction of treatment. 1-G is the probability that someone who would otherwise die survives if treated. 1-H is the probability that someone who would otherwise survive dies if treated.

Thus, by using the condition $Y^{a=1} \perp\!\!\!\perp P | Y^{a=0}$ to operationalize effect homogeneity, we reframe homogeneity of the "magnitude of effect" as a matter of equal distribution of the cofactors that determine whether individuals respond to treatment. If the prevalences of those cofactors differ between populations, such effect equality may hold within levels of covariates V, in which case, there is conditional equality of the effect of introducing (or removing) treatment. Conditioning on V is then seen as an attempt to account for those variables that are predictors of the prevalence of the potentially unmeasured background cofactors which determine whether an individual "switches outcome" in response to treatment.

Much of the intuition behind traditional approaches to choice of effect modifiers translates readily to the COST parameter framework, but with the added advantage that the line of reasoning is specific to the relevant effect measure. To illustrate, suppose we are interested in generalizing findings about the adverse effects of Codeine. The goal is to account for all cofactors that determine whether a patient will respond to treatment, either by conditioning on those cofactors directly, or by finding observable markers for their prevalence. For example, if the CYP2D6 variant partly determines whether patients respond to Codeine, we could either condition on the gene, or condition on ethnicity as a marker for its prevalence (that is, include either the genetic variant or ethnicity in the set V). Further, pre-study drug use may be an observable marker for the prevalence of susceptibility, due to depletion of susceptibles [15]. We will therefore measure and control for ethnicity and pre-study drug use, as well as any other covariates that are relevant according to similar criteria.

COST parameters are generally not identified from the data without strong monotonicity assumptions. If the treatment has a positive monotonic effect (meaning that the treatment does not prevent anyone from having the outcome), and if the COST pa-

rameters G and H are equal between the study population and the target population conditional on covariates V, there will be homogeneity on the SR scale for exposures which increase the incidence of the outcome (SR becomes equivalent to H, whereas G is trivially 1). Analogous discussion applies when considering a situation in which the treatment of interest is negatively monotonic (protective), in which case RR becomes equivalent to G, and H is trivially 1.

If treatment has monotonic effects, the COST parameters can therefore be used as a "bridge" between the biological knowledge on the one side, and homogeneity of observable measures of effects on the other side, thereby allowing the investigator to standardize effect measure from a study to a target population using either approach 1 or approach 2 from Table 2. The necessary weights are discussed elsewhere [9]. The bias which is associated with the use of COST parameters in the presence of non-monotonicity is small either if non-monotonicity is negligible, or if the baseline risks are similar between the target population and the study population. If non-monotonicity is not negligible and baseline risks differ between the study population and the target population, the bias associated with COST parameters may be substantial; in such settings, the COST parameter approach should not be used.

The COST parameter approach often results in a recommendation to consider effect homogeneity in terms of the RR scale for exposures that reduce incidence, and in terms of the SR scale for exposures that increase the incidence, while keeping the coding of the exposure such that the "natural state" of exposure has value 0 and the intervention has value 1. Variations of this suggestion have arisen independently a number of times in the previous literature [16, 17, 18]. This approach is also consistent with the Cochrane Handbook [19], which states that "When the study aims to reduce the incidence of an adverse outcome there is empirical evidence that risk ratios of the adverse outcome are more consistent than risk ratios of the non-event" (the handbook does not take a position on what effect measure to use when the study attempts to estimate the increase in incidence of an adverse outcome). When the disease is rare, this approach is closely approximated by the earlier suggestion to consider "relative benefits and absolute harms" of interventions [20].

Finally, we note an important limitation of COST parameters, which is that they have so far only been defined for binary outcomes. Extensions to continuous and time-to-event outcomes have not yet been established.

3. EFFECT HOMOGENEITY IN DISTRIBUTION

An alternative approach is to operationalize effect homogeneity in terms of the individual counterfactual distributions under treatment and no treatment. Effect homogeneity in distribution between the study population and the target population holds whenever the following two conditions hold simultaneously: (1) If everyone in both populations were untreated, you would observe the same distribution of outcomes in the two populations

($Y^{a=0} \perp\!\!\!\perp P$) and (2) if everyone in both populations were treated, you would observe the same distribution of outcomes in the two populations ($Y^{a=1} \perp\!\!\!\perp P$). This condition was referred to as "S-ignorability" by Bareinboim and Pearl, and as "exchangeability between populations" by Lesko et al [21]. VanderWeele [6] showed that effect homogeneity in distribution implies homogeneity of all standard effect measures; effect homogeneity in distribution is therefore a stronger assumption than effect homogeneity in measure on standard scales.

In order to illustrate the difference between effect homogeneity in distribution and effect homogeneity in measure, we again consider the logistic regression model discussed in the previous section:

$$\text{logit } \Pr(Y = 1|A, W, P = s) \ = \beta_0 + \beta_1 A + \beta_2 W$$

This model is restricted to the study population, and omits the product term $\beta_3 \times A \times W$. We showed that this model is justified under effect homogeneity in measure on the OR scale between groups in the study population. We note that this model could also be justified under effect homogeneity in distribution between the same two groups. However, this modeling approach has an immediate implication: If effect homogeneity in distribution holds and the effect of A on Y is unconfounded conditional on W and $P = s$, then β_2 must be equal to zero (see Appendix 2). This makes the model subject to an empirical test: if e.g. the Wald test rejects $\beta_2 = 0$ the model is misspecified. While we do not recommend this as a test of the homogeneity assumption, we believe this example illustrates that effect homogeneity in distribution is a very strong concept, and that investigators often have to rely on a weaker form of effect homogeneity.

As with effect homogeneity in measure, effect homogeneity in distribution may hold within levels of a set of covariates V. If effect homogeneity in distribution between the study population and the target population holds conditional on V, one can use a third standardization formula (approach 3 in Table 2), based on separately standardizing the conditional risk under treatment and the conditional risk under no treatment from the study population, to the distribution of V in the target population.

Although methods based on assuming conditional homogeneity of the distribution of a counterfactual variable are mathematically elegant and avoid most of the limitations of defining homogeneity with respect to effect measures, they require strong assumptions which go well beyond the conditions that epidemiologists and clinical scientists have traditionally considered necessary for generalizability. Specifically, whereas approaches that are based on conditional effect homogeneity in measure aim only to control for those covariates that are associated with the magnitude of the effect, methods that rely on conditional effect homogeneity in distribution are valid only if they account for every cause of the outcome that differs between the study population and the target population. In other words, approaches based on conditional effect homogeneity in distribution may lead to biased transportability estimates in the presence of unmeasured causes of the outcome

whose distributions differ between the study population and the target population. An example of the implications of such bias was shown recently in the closely related context of agent-based models used for extrapolation[22].

Effect homogeneity in distribution may occasionally be a reasonable assumption if the imbalance in covariates arises due to a fully understood non-random sampling mechanism, for example, if the investigators enroll participants from an enumerated source population, with selection probability determined by measured baseline covariates. However, outside of such stylized examples, it is more challenging to see good justifications for this type of homogeneity assumption.

We note that approaches based upon effect homogeneity in distribution do not make use of possible information contained in the relationship between what happens if the pharmaceutical is taken, and what happens if the pharmaceutical is not taken. To illustrate, consider a situation where we have conducted a randomized controlled trial on the effect of homeopathy vs no treatment on the incidence of cardiovascular disease, and concluded that the effect in the study population is null. Suppose we are interested in predicting the effect in a different target population, but we believe there may be unmeasured causes of cardiovascular disease that differ between the study population and the target population. In such situations, if we operationalize effect homogeneity using a notion of effect homogeneity in distribution, we are likely forced to conclude that we are unable to make predictions for the target population. In contrast, investigators using an approach based on effect homogeneity in measure could potentially be able to clarify plausible conditions under which the null findings can be extrapolated to the target population.

3.1 Weighted estimators for generalizability and transportability

One particular implementation of generalization based on effect homogeneity in distribution originated with work by Stuart and Cole [23, 24]. These methods extend inverse probability based estimators [25], which play a key role in previous work on causal modelling [26, 27] to the setting of external validity. The validity conditions of these methods are equal to those of standardization based methods discussed above.

Users of these methods often distinguish between "transportability" (where the analytic goal is to extrapolate the findings to a target population that does not include those in the study population, i.e. a target population that looks like those who were eligible for, but were not sampled in the study), and "generalizability" (where the target population also includes those in the study population). The methods used for each objective differs in that inverse probability of selection weights [21] (approach 4 in table 2) are used for generalizability, whereas inverse odds of selection weights[28] (approach 5 in table 2) are used for transportability.

Stated slightly differently, the choice between weights is determined by whether the target population is similar to the entire source population from which the study partici-

pants were selected, or similar to the subset of the source population that was not selected for the study. We believe the first type of target population is more common; in such settings, inverse probability of selection weights should be used. Inverse odds weights may be appropriate if the study participants are sampled for a pilot study to determine whether the intervention will be implemented in those who were eligible to be selected, but weren't.

Inverse probability weighted methods have been applied to generalize the results of trials on the effect of HIV medication [29] and treatments for substance use disorder [30]. Lesko et al provided a full description of how these methods can be used in practice [21]. Buchanan et al [31] provided results about the statistical properties of inverse probability weighted estimators for external validity. Dahabreh et al [32] discussed estimators based on augmented inverse probability weights, which are doubly roubst. Nguyen et al [33] showed how to conduct sensitivity analyses on deviations from conditional effect homogeneity. Breskin et al [34] provided results on bounds, i.e. intervals that show how wrong the point estimates can be in either direction if the assumption of conditional effect homogeneity in distribution is not fully met, in the presence and/or absence of confounding. If one suspects both confounding and lack of effect homogeneity in distribution, these bounds can be used to reason about target validity [35], that is, how much bias there may be in the estimates for the target population as a result of deviations both from internal and external validity.

We note that for collapsible effect measures, re-weighting methods based upon conditional effect homogeneity in measure may be feasible, but to the best of our knowledge, the theory for such methods has not yet been fully developed.

3.2 Causal diagrams for transportability

One example of a class of data generating mechanisms that guarantees effect homogeneity in distribution (and therefore also effect homogeneity in measure for all standard effect measures) was provided by Bareinboim and Pearl [36, 37, 38, 39], based on causal diagrams [40]. These diagrams are, to our knowledge, the first published formal framework for reasoning about which variables to adjust for when using approaches based on effect homogeneity in distribution. In particular, they use a generalization of effect homogeneity in distribution that allows the covariates that are adjusted for, and membership in the populations that the counterfactual distributions are equal between, to be downstream consequences of treatment.

A selection diagram is constructed as follows: First, the investigator must provide a causal directed acyclic graph (DAG) that is valid both for the study population and for the target population. For this to be possible, the variables must be in the same temporal order between the two populations. If that requirement is met, a DAG which is valid for both populations can be constructed by including every node and edge from the causal DAG in each population. After a shared causal DAG has been constructed,

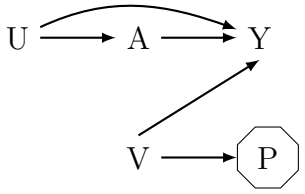

Figure 2: In this causal diagram, findings from the study population may be transported to the target population if we have measured sufficient covariates V to block all paths between the selection node P, and the outcome Y. We have chosen to represent the selection node P with an octagon.

one must also add (1) selection variable nodes (P) associated with all variables whose assignment mechanism differs between the study population and the target population, and all variables that depend upon background causes whose distribution differs between the study population and the target population (2) all paths between P and Y that one is not able to rule out based on the temporal structure or expert knowledge. Generally, such paths will exist whenever there are causes of the outcome that differ between the populations. Note that when the goal is to account for "man-made" differences between the study population and the target population, i.e. those differences which arise due to how the sample was selected, P is a single binary "sink node" representing membership in the study population, which has the same interpretation as the P variable that we have considered so far in this paper. An example of a causal diagram used for transportability is shown in Figure 2.

Once a selection diagram has been constructed, one can check for transportability of the results by determining whether Y is d-separated from P, given some set of measured covariates V, in a manipulated graph where all arrows going into A have been deleted. If such d-separation holds in the manipulated selection diagram, there will exist a transport formula which identifies the causal effect in the target population based on a combination of observed quantities in the study population and observed quantities in the target population. If V consists only of baseline covariates, then the transport formula is equal to the standardization formula discussed in section 3.1.

Note that the independence relation that is queried by this d-separation approach can be written algebraically as

$$Y^a \perp\!\!\!\perp P^a | V^a \ \ \forall a$$

which Bareinboim and Pearl referred to as "S-admissibility". When P and V are pre-treatment variables, $P^a = P$ and $V^a = V$ so the independence relation can be simplified as

$$Y^a \perp\!\!\!\perp P | V \ \ \forall a$$

(or "S-ignorability"). This simplified version is identical to the previously discussed operationalization of conditional effect homogeneity in distribution, which illustrates the

equivalence between the graphical approach and approaches based on standardization or inverse probability weights when V and P are pre-treatment.

Thus, while the graphical approach and the inverse probability weighted approach will result in very similar analyses if P and V are pre-treatment (the analyses will be non-parametrically equivalent but may differ in practice as they may be associated with different modelling assumptions on the joint distribution of variables), the graphical model allows a potentially useful generalization to settings where it is necessary to adjust for post-baseline covariates. In practice, we are not aware of any published examples where a convincing argument was made that a causal effect is transportable only by measuring and adjusting for covariates that were causally affected by treatment.

Other authors have constructed causal diagrams for generalizability in different ways. In particular, Dahabreh et al [41] use Single World Intervention Graphs (SWIGs) to examine conditions under which causal parameters can be generalized from a randomized trial to all trial eligible individuals.

4. CONCLUSIONS

Causal effects may differ between populations, and investigators will often have to standardize their estimates over a set of effect modifiers in order to make the results applicable to clinically relevant populations. Before it is possible to begin reasoning about which covariates must be standardized over, it is necessary to provide a definition of effect homogeneity. Several different approaches have been proposed.

If effect homogeneity is to be operationalized in terms of stability of a measure of effect, the analytic objective is to account for all those covariates that are associated with the magnitude of the effect on the chosen scale. COST parameters have been developed to formalize conditions that result in homogeneity of observable effect measures. This approach requires that the investigators have accounted for all variables that predict treatment response, that only baseline covariates are necessary for this purpose, and that the effect of treatment is monotonic. When these conditions are met, using COST parameters allows investigators to retain much of the underlying intuition behind traditional approaches to effect modification. Future work may be necessary to develop new classes of causal models that result in homogeneity of other effect measures, including effect measures relevant to time-to-event data.

If instead effect homogeneity is to be operationalized in terms of conditional homogeneity of the distributions of counterfactual variables (such as in methods based on inverse probability weights and causal diagrams), the analytic objective shifts to accounting for all covariates that are associated with the counterfactual outcome and whose distribution differs between populations. This will generally require a much larger set of covariates. Controlling for all the necessary covariates will sometimes be feasible in situations where the goal is to recover the effect estimates for the full source population in the presence of a fully understood non-random selection mechanism, but may be less realistic in other

settings. If the required conditions are met, methods based on effect homogeneity in distribution have considerable advantages, as they do not rely on parametric assumptions or monotonicity conditions.

All approaches have considerable limitations, and the choice between them will generally depend on expert beliefs about which assumptions are most likely to be approximately true in the specific scientific context.

REFERENCES

[1] Noel S. Weiss. Generalizing from the results of randomized studies of treatment: Can non-randomized studies be of help? European Journal of Epidemiology, 34(8):715–718, August 2019.

[2] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A Structural Approach to Selection Bias. Epidemiology (Cambridge, Mass.), 15(5):615–25, 2004.

[3] Etsuji Suzuki, Toshihide Tsuda, Toshiharu Mitsuhashi, Mohammad Ali Mansournia, and Eiji Yamamoto. Errors in causal inference: an organizational schema for systematic error and random error. Annals of Epidemiology, 26(11):788–793.e1, 2016.

[4] Anders Huitfeldt, Andrew Goldstein, and Sonja A. Swanson. The Choice of Effect Measure for Binary Outcomes: Introducing Counterfactual Outcome State Transition Parameters. Epidemiologic Methods, 7(1), 2018.

[5] Issa J. Dahabreh and Miguel A. Hernán. Extending inferences from a randomized trial to a target population. European Journal of Epidemiology, 34(8):719–722, August 2019.

[6] Tyler J VanderWeele. Confounding and Effect Modification: Distribution and Measure. Epidemiologic Methods, 1(1):55–82, 2012.

[7] Sander Greenland. Interpretation and estimation of summary ratios under heterogeneity. Statistics in Medicine, 1(3):217–227, 1982.

[8] Stephen Bernard, Kathleen A Neville, Anne T Nguyen, and David A Flockhart. Interethnic Differences in Genetic Polymorphisms of CYP2d6 in the U.S. Population: Clinical Implications. The oncologist, 11(2):126–35, 2006.

[9] Anders Huitfeldt, Mats Julius Stensrud, and Etsuji Suzuki. On the Collapsibility of Measures of Effect in the Counterfactual Causal Framework. Emerging Themes in Epidemiology, 16(1), 2019.

[10] W. G. Cochran. The Comparison of Percentages in Matched Samples. Biometrika, 37(3/4):256–266, 1950.

[11] Julian P T Higgins, Simon G Thompson, Jonathan J Deeks, and Douglas G Altman. Measuring inconsistency in meta-analyses. BMJ : British Medical Journal, 327(7414):557–560, September 2003.

[12] Motoki Iwasaki, Seiichiro Yamamoto, Tetsuya Otani, Manami Inoue, Tomoyuki Hanaoka, Tomotaka Sobue, Shoichiro Tsugane, and Japan Public Health Center-based Prospective Study (JPHC Study) Group. Generalizability of Relative Risk Estimates from a Well-defined Population to a General Population. European Journal of Epidemiology, 21(4):253–262, April 2006.

[13] Charles Poole, Ian Shrier, and Tyler J VanderWeele. Is the Risk Difference Really a More Heterogeneous Measure? Epidemiology, 26(5):714–8, 2015.

[14] Donna Spiegelman and Tyler J. VanderWeele. Evaluating Public Health Interventions: 6. Modeling Ratios or Differences? Let the Data Tell Us. American Journal of Public Health, 107(7):1087–1091, 2017.

[15] Guide on Methodological Standards in Pharmacoepidemiology (Revision 7)., 2018.

[16] Mindel Cherniack Sheps. Shall We Count the Living of the Dead? The New England Journal of Medicine, 259(25):1210–4, 1958.

[17] Jonathan J Deeks. Issues in the Selection of a Summary Statistic for Meta-Analysis of Clinical Trials with Binary Outcomes. Statistics in medicine, 21(11):1575–600, 2002.

[18] Rose Baker and Dan Jackson. A new measure of treatment effect for random-effects meta-analysis of comparative binary outcome data. ArXiv:1806.03471, 2018.

[19] Julian PT Higgins and Sally Green, editors. Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). Chapter 9: Analysing data and undertaking meta-analyses, 2011.

[20] Paul P Glasziou and Les M Irwig. An Evidence Based Approach to Individualising Treatment. BMJ, 311(7016), 1995.

[21] Catherine R. Lesko, Ashley L. Buchanan, Daniel Westreich, Jessie K. Edwards, Michael G. Hudgens, and Stephen R. Cole. Generalizing Study Results: A Potential Outcomes Perspective. Epidemiology, 28:553–561, 2017.

[22] Eleanor J. Murray, James M. Robins, George R. Seage, Kenneth A. Freedberg, and Miguel A. Hernán. A Comparison of Agent-Based Models and the Parametric G-Formula for Causal Inference. American Journal of Epidemiology, 186(2):131–142, July 2017.

[23] Elizabeth A. Stuart, Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society. Series A, (Statistics in Society), 174(2):369–386, April 2001.

[24] Stephen R Cole and Elizabeth A Stuart. Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. American journal of epidemiology, 172(1):107–15, 2010.

[25] D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement From a Finite Universe. Journal of the American Statistical Association, 47(260):663–685, 1952.

[26] James M. Robins. Association, Causation, and Marginal Structural Models. Synthese, 121(1-2):151–179, 1999.

[27] J. M. Robins, M. A. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass.), 11(5):550–560, September 2000.

[28] Daniel Westreich, Jessie K. Edwards, Catherine R. Lesko, Elizabeth Stuart, and Stephen R. Cole. Transportability of Trial Results Using Inverse Odds of Sampling Weights. American Journal of Epidemiology, 186(8):1010–1014, 2017.

[29] Haidong Lu, Stephen R Cole, H Irene Hall, Enrique F Schisterman, Tiffany L Breger, Jessie K Edwards, and Daniel Westreich. Generalizing the per-protocol treatment effect: The case of ACTG A5095. Clinical Trials, 16(1):52–62, October 2018.

[30] Ryoko Susukida, Rosa M. Crum, Cyrus Ebnesajjad, Elizabeth A. Stuart, and Ramin Mojtabai. Generalizability of findings from randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. Addiction, 112(7):1210–1219, 2017.

[31] Ashley L. Buchanan, Michael G. Hudgens, Stephen R. Cole, Katie R. Mollan, Paul E. Sax, Eric S. Daar, Adaora A. Adimora, Joseph J. Eron, and Michael J. Mugavero. Generalizing evidence from randomized trials using inverse probability of sampling weights. Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(4):1193–1209, 2018.

[32] Issa J. Dahabreh, Sarah E. Robertson, Eric J. Tchetgen Tchetgen, Elizabeth A. Stuart, and Miguel A. Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics, 0(ja), 2018.

[33] Trang Quynh Nguyen, Cyrus Ebnesajjad, Stephen R. Cole, and Elizabeth A. Stuart. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. The Annals of Applied Statistics, 11(1):225–247, March 2017.

[34] Alexander Breskin, Daniel Westreich, Stephen R. Cole, and Jessie K. Edwards. Using Bounds to Compare the Strength of Exchangeability Assumptions for Internal and External Validity. American Journal of Epidemiology, Forthcoming, 2019.

[35] Daniel Westreich, Jessie K. Edwards, Catherine R. Lesko, Stephen R. Cole, and Elizabeth A. Stuart. Target Validity and the Hierarchy of Study Designs. American Journal of Epidemiology, 188(2):438–443, February 2019.

[36] Judea Pearl and Elias Bareinboim. Transportability of Causal and Statistical Relations: A Formal Approach. In Twenty-Fifth AAAI Conference on Artificial Intelligence, August 2011.

[37] Elias Bareinboim and Judea Pearl. A General Algorithm for Deciding Transportability of Experimental Results. Journal of Causal Inference, 1(1):107–134, 2013.

[38] Judea Pearl and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. Statistical Science, 29(4):579–595, 2014.

[39] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, 113(27):7345–7352, July 2016.

[40] Judea Pearl. Causality: Models, Reasoning and Inference. Cambridge University Press, Cambridge, 2 edition, 2009.

[41] Issa J. Dahabreh, James M. Robins, Sebastien J.-P. A. Haneuse, and Miguel A. Hernán. Generalizing causal inferences from randomized trials: counterfactual and graphical identification. arXiv:1906.10792 [stat], June 2019. arXiv: 1906.10792.

FUNDING

The authors received no specific funding for this work. Dr. Stensrud is supported by the Research Council of Norway, grant NFR239956/F20 - Analyzing clinical health registries: Improved software and mathematics of identifiability. Dr. Swanson is supported by NWO/ZonMw Veni grant (91617066). Dr. Suzuki is supported by Japan Society for the Promotion of Science (KAKENHI grant numbers JP17K17898, JP15K08776, and JP18K10104) and The Okayama Medical Foundation. Dr. Huitfeldt was supported by the Effective Altruism Hotel Blackpool during revision of the manuscript.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. Issa Dahabreh and two anonymous reviewers for suggestions that greatly improved the manuscript. Any remaining errors are our own.

A. APPENDIX 1

Proofs of identifying expressions from table 2. We note that these proofs are not new to this paper, and are included here only for completeness:

A.1 Approach 1

$$
\begin{aligned}
&\sum_v \Big[\mathrm{RR}_{s,v} \times \Pr(V=v|Y^{a=0}=1, P=t)\Big] \\
&= \sum_v \Big[\mathrm{RR}_{t,v} \times \Pr(V=v|Y^{a=0}=1, P=t)\Big] (\because \mathrm{RR}_{s,v} = \mathrm{RR}_{t,v}) \\
&= \sum_v \left[\frac{\Pr(Y^{a=1}=1|V=v, P=t) \times \Pr(V=v|Y^{a=0}=1, P=t)}{\Pr(Y^{a=0}=1|V=v, P=t)}\right] \\
&= \sum_v \left[\frac{\Pr(Y^{a=1}=1|V=v, P=t) \times \Pr(Y^{a=0}=1|V=v, P=t) \times \Pr(V=v|P=t)}{\Pr(Y^{a=0}=1|V=v, P=t) \times \Pr(Y^{a=0}=1|P=t)}\right] \\
&= \sum_v \left[\frac{\Pr(Y^{a=1}=1|V=v, P=t) \times \Pr(V=v|P=t)}{\Pr(Y^{a=0}=1|P=t)}\right] \\
&= \frac{\sum_v \left[\Pr(Y^{a=1}=1|V=v, P=t) \times \Pr(V=v|P=t)\right]}{\Pr(Y^{a=0}=1|P=t)} \\
&= \frac{\Pr(Y^{a=1}=1|P=t)}{\Pr(Y^{a=0}=1|P=t)} \\
&= \mathrm{RR}_t
\end{aligned}
\tag{1}
$$

A.2 Approach 2

$$
\begin{aligned}
&\sum_v \Big[\Pr(Y^{a=0}=1|V=v,P=t)\times \mathrm{RR}_{s,v}\times \Pr(V=v|P=t)\Big]\\
&=\sum_v \Big[\Pr(Y^{a=0}=1|V=v,P=t)\times \mathrm{RR}_{t,v}\times \Pr(V=v|P=t)\Big]\ (\because \mathrm{RR}_{s,v}=\mathrm{RR}_{t,v})\\
&=\sum_v \left[\Pr(Y^{a=0}=1|V=v,P=t)\times \frac{\Pr(Y^{a=1}=1|V=v,P=t)}{\Pr(Y^{a=0}=1|V=v,P=t)}\times \Pr(V=v|P=t)\right]\\
&=\sum_v \Big[\Pr(Y^{a=1}=1|V=v,P=t)\times \Pr(V=v|P=t)\Big]\\
&=\Pr(Y^{a=1}=1|P=t)
\end{aligned}
\tag{2}
$$

A.3 Approach 3

$$
\begin{aligned}
&\sum_v [\Pr(Y^a=1|V=v,P=s)\times \Pr(V=v|P=t)]\\
&=\sum_v [\Pr(Y^a=1|V=v,P=t)\times \Pr(V=v|P=t)]\ (\because Y^a \perp\!\!\!\perp P|V\ \forall a)\\
&=\Pr(Y^a=1|P=t)
\end{aligned}
\tag{3}
$$

A.4 Approach 4

We are here assuming that Y is a binary variable, the proof generalizes readily to settings with continuous or time-to-event outcomes. In order to simplify the logic, we will further assume that the same set of baseline covariates V is sufficient to control both for confounding for A, and for differences between populations. In other words, we will assume conditional exchangeability in the study population ($Y^a \perp\!\!\!\perp A|V=v,P=s\ \ \forall a$) and conditional effect homogeneity in distribution ($Y^a \perp\!\!\!\perp P|V=v\ \ \forall a$). Before we begin, it is useful to note that $\frac{\Pr(A=a,V=v,P=s)}{\Pr(A=a|P=s,V=v)\times\Pr(P=s|V=v)}=\Pr(V=v)$. This follows from sequential application of the definition of conditional probability.

$$
\begin{aligned}
&\sum_v \left[\frac{\Pr(Y=1|A=a,V=v,P=s) \times \Pr(A=a,V=v,P=s)}{\Pr(A=a|P=s,V=v) \times \Pr(P=s|V=v)} \right] \\
&= \sum_v \left[\Pr(Y=1|A=a,V=v,P=s) \times \Pr(V=v) \right] \\
&= \sum_v \left[\Pr(Y^a=1|A=a,V=v,P=s) \times \Pr(V=v) \right] (\because \text{Consistency}) \\
&= \sum_v \left[\Pr(Y^a=1|V=v,P=s) \times \Pr(V=v) \right] (\because Y^a \perp\!\!\!\perp A|V,P=s \ \ \forall a) \\
&= \sum_v \left[\Pr(Y^a=1|V=v) \times \Pr(V=v) \right] (\because Y^a \perp\!\!\!\perp P|V \ \ \forall a) \\
&= \Pr(Y^a=1)
\end{aligned}
\tag{4}
$$

A.5 Approach 5

The proof of approach 5 is closely related to that for approach 4. Westreich et al [28] provide a full proof in the appendix.

B. APPENDIX 2

Here, we prove that if there is effect homogeneity in distribution between the groups $W=1$ and $W=0$, then the parameter β_2 must be equal to zero in the regression model

$$\text{logit } \Pr(Y=1|A,W,P=s) = \beta_0 + \beta_1 A + \beta_2 W \tag{5}$$

Note here that we are discussing a regression model fit within the study population, and where the homogeneity assumption is between groups of baseline covariate W. In contrast to the rest of the paper, we are therefore using the homogeneity assumption $Y^a \perp\!\!\!\perp W|P=s \ \ \forall a$ rather than $Y^a \perp\!\!\!\perp P|V=v \ \ \forall a$.

Additionally, we will make the following assumptions:

$$Y^a \perp\!\!\!\perp A|W,P=s \ \ \forall a (\text{Conditional exchangeability in study population})$$
$$Y^a = Y \text{ if } A=a (\text{Consistency})$$

By consistency and exchangeability, the model can be rewritten as a structural model:

$$\text{logit } \Pr(Y^a=1|W,P=s) = \beta_0 + \beta_1 a + \beta_2 W \tag{6}$$

If $W = 0$, we have:

$$\text{logit } \Pr(Y^a = 1|W = 0, P = s) = \beta_0 + \beta_1 a \tag{7}$$

If $W = 1$, we have:

$$\text{logit } \Pr(Y^a = 1|W = 1, P = s) = \beta_0 + \beta_1 a + \beta_2 \tag{8}$$

By the assumption of effect homogeneity in distribution, we can set these equal:

$$\beta_0 + \beta_1 a = \beta_0 + \beta_1 a + \beta_2$$

Solving this for β_2 we get $\beta_2 = 0$.