TECHNICAL REPORT
R-400
May 2014

# External Validity: From do-calculus to Transportability across Populations*

**Judea Pearl and Elias Bareinboim**

University of California, Los Angeles

*Abstract.* The generalizability of empirical findings to new environments, settings or populations, often called "external validity," is essential in most scientific explorations. This paper treats a particular problem of generalizability, called "transportability", defined as a license to transfer causal effects learned in experimental studies to a new population, in which only observational studies can be conducted. We introduce a formal representation called "selection diagrams" for expressing knowledge about differences and commonalities between populations of interest and, using this representation, we reduce questions of transportability to symbolic derivations in the do-calculus. This reduction yields graph-based procedures for deciding, prior to observing any data, whether causal effects in the target population can be inferred from experimental findings in the study population. When the answer is affirmative, the procedures identify what experimental and observational findings need be obtained from the two populations, and how they can be combined to ensure bias-free transport.

*Key words and phrases:* experimental design, generalizability, causal effects, external validity.

## 1. INTRODUCTION: THREATS VS. ASSUMPTIONS

Science is about generalization, and generalization requires that conclusions obtained in the laboratory be transported and applied elsewhere, in an environment that differs in many aspects from that of the laboratory.

Clearly, if the target environment is arbitrary, or drastically different from the study environment nothing can be transferred and scientific progress will come to a standstill. However, the fact that most studies are conducted with the intention of applying the results elsewhere means that we usually deem the target environment sufficiently similar to the study environment to justify the transport of experimental results or their ramifications.

Remarkably, the conditions that permit such transport have not received systematic formal treatment. In statistical practice, problems related to combining

*Computer Science Department, Los Angeles, CA, 90095-1596, USA. (e-mail: judea@cs.ucla.edu; eb@cs.ucla.edu).*

*This research was supported in parts by NIH grant #1R01 LM009961-01, NSF grant #IIS-0914211, and ONR grant #N000-14-09-1-0665.

and generalizing from diverse studies are handled by methods of meta analysis (Glass (1976); Hedges and Olkin (1985); Owen (2009)), or hierarchical models (Gelman and Hill (2007)), in which results of diverse studies are pooled together by standard statistical procedures (e.g., inverse-variance re-weighting in meta-analysis, partial pooling in hierarchical modelling) and rarely make explicit distinction between experimental and observational regimes; performance is evaluated primarily by simulation.

To supplement these methodologies, our paper provides theoretical guidance in the form of limits on what can be achieved in practice, what problems are likely to be encountered when populations differ significantly from each other, what population differences can be circumvented by clever design, and what differences constitute theoretical impediments, prohibiting generalization by any means whatsoever.

On the theoretical front, the standard literature on this topic, falling under rubrics such as "external validity" (Campbell and Stanley (1963); Manski (2007)), "heterogeneity" (Höfler et al. (2010)), "quasi-experiments" ((Shadish et al., 2002, Ch. 3); Adelman (1991)),[1] consists primarily of threats, namely, explanations of what may go wrong when we try to transport results from one study to another while ignoring their differences. Rarely do we find an analysis of "licensing assumptions," namely, formal conditions under which the transport of results across differing environments or populations is licensed from first principles.[2]

The reasons for this asymmetry are several. First, threats are safer to cite than assumptions. He who cites "threats" appears prudent, cautious and thoughtful, whereas he who seeks licensing assumptions risks suspicions of attempting to endorse those assumptions.

Second, assumptions are self destructive in their honesty. The more explicit the assumption, the more criticism it invites, for it tends to trigger a richer space of alternative scenarios in which the assumption may fail. Researchers prefer therefore to declare threats in public and make assumptions in private.

Third, whereas threats can be communicated in plain English, supported by anecdotal pointers to familiar experiences, assumptions require a formal language within which the notion "environment" (or "population") is given precise characterization, and differences among environments can be encoded and analyzed.

The advent of causal diagrams (Wright (1921); Heise (1975); Davis (1984); Verma and Pearl (1988); Spirtes et al. (1993); Pearl (1995)) together with models of interventions (Haavelmo, 1943; Strotz and Wold, 1960) and counterfactuals (Neyman, 1923; Rubin, 1974; Robins, 1986; Balke and Pearl, 1995) provides such a language and renders the formalization of transportability possible.

[1] Manski (2007) defines "external validity" as follows: "An experiment is said to have "external validity" if the distribution of outcomes realized by a treatment group is the same as the distribution of outcome that would be realized in an actual program." (Campbell and Stanley, 1963, p. 5) take a slightly broader view: ""External validity" asks the question of generalizability: to what populations, settings, treatment variables, and measurement variables can this effect be generalized?"

[2] Hernán and VanderWeele (2011) studied such conditions in the context of compound treatments, where we seek to predict the effect of one version of a treatment from experiments with a different version. Their analysis is a special case of the theory developed in this paper (Petersen, 2011). A related application is reported in Robins et al. (2008) where a treatment strategy is extrapolated between two biological similar populations under different observational regimes.

Armed with this language, this paper departs from the tradition of communicating "threats" and embarks instead on the task of formulating "licenses to transport," namely, assumptions that, if they held true, would permit us to transport results across studies.

In addition, the paper uses the inferential machinery of the do-calculus (Pearl, 1995; Koller and Friedman, 2009; Huang and Valtorta, 2006; Shpitser and Pearl, 2006) to derive algorithms for deciding whether transportability is feasible and how experimental and observational findings can be combined to yield unbiased estimates of causal effects in the target population.

The paper is organized as follows. In section 2, we review the foundations of structural equations modelling (SEM), the question of identifiability, and the do-calculus that emerges from these foundations. (This section can be skipped by readers familiar with these concepts and tools.) In section 3, we motivate the question of transportability through simple examples, and illustrate how the solution depends on the causal story behind the problem. In section 4, we formally define the notion of transportability and reduce it to a problem of symbolic transformations in do-calculus. In section 5, we provide a graphical criterion for deciding transportability and estimating transported causal effects. We conclude in section 6 with brief discussions of related problems of external validity, these include statistical transportability, and meta-analysis.

## 2. PRELIMINARIES: THE LOGICAL FOUNDATIONS OF CAUSAL INFERENCE

The tools presented in this paper were developed in the context of nonparametric Structural Equations Models (SEM), which is one among several approaches to causal inference, and goes back to (Haavelmo, 1943; Strotz and Wold, 1960). Other approaches include, for example, potential-outcomes (Rubin, 1974), Structured Tree Graphs (Robins, 1986), decision analytic (Dawid, 2002), Causal Bayesian Networks (Spirtes et al. (2000); (Pearl, 2000, Ch. 1), Bareinboim et al. (2012)), and Settable Systems (White and Chalak, 2009). We will first describe the generic features common to all such approaches, and then summarize how these features are represented in SEM.[3]

### 2.1 Causal models as inference engines

From a logical viewpoint, causal analysis relies on causal assumptions that cannot be deduced from (nonexperimental) data. Thus, every approach to causal inference must provide a systematic way of encoding, testing and combining these assumptions with data. Accordingly, we view causal modeling as an inference engine that takes three inputs and produces three outputs. The inputs are:

**I-1.** A set $A$ of qualitative causal *assumptions* which the investigator is prepared to defend on scientific grounds, and a model $M_A$ that encodes these assumptions mathematically. (In SEM, $M_A$ takes the form of a diagram or a set of unspecified functions. A typical assumption is that no direct effect

[3] We use the acronym SEM for both parametric and nonparametric representations though, historically, SEM practitioners preferred the former (Bollen and Pearl, 2013). Pearl (2011) has used the term Structural Causal Models (SCM) to eliminate this confusion. While comparisons of the various approaches lie beyond the scope of this paper, we nevertheless propose that their merits be judged by the extent to which each facilitates the functions described below.

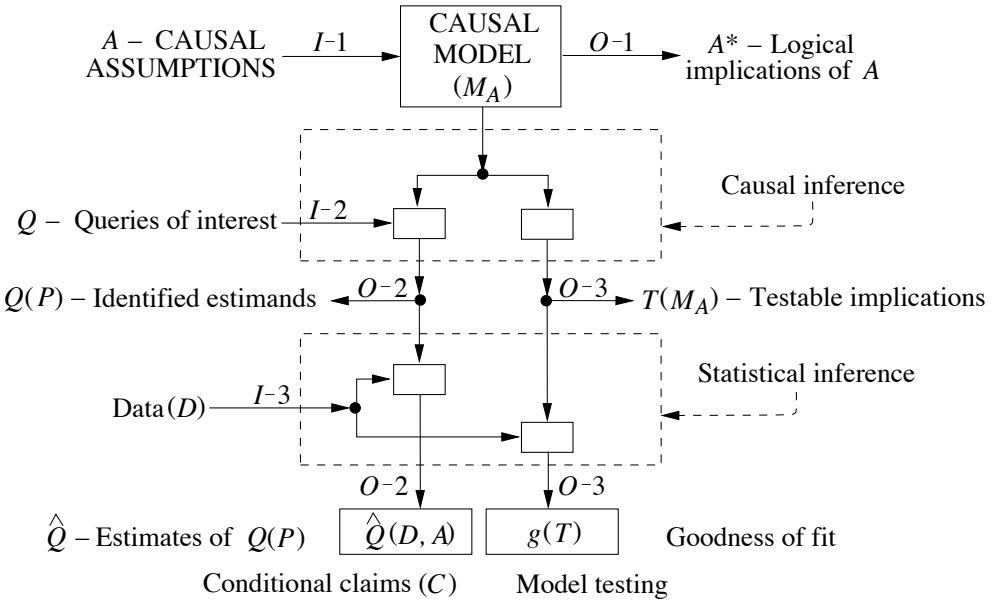


FIG 1. *Causal analysis depicted as an inference engine converting assumptions* ($A$)*, queries* ($Q$)*, and data* ($D$) *into logical implications* ($A^*$)*, conditional claims* ($C$)*, and data-fitness indices* ($g(T)$).

exists between a pair of variables (known as exclusion restriction), or that an omitted factor, represented by an error term, is independent of other such factors observed or unobserved, known as well as unknown.

**I-2.** A set $Q$ of *queries* concerning causal or counterfactual relationships among variables of interest. In linear SEM, $Q$ concerned the magnitudes of structural coefficients but, in general, $Q$ may address causal relations directly, e.g.,

$Q_1$ : What is the effect of treatment $X$ on outcome $Y$?

$Q_2$ : Is this employer practicing gender discrimination?

In principle, each query $Q_i \in Q$ should be "well defined," that is, computable from any fully specified model $M$ compatible with $A$. (See Definition 1 for formal characterization of a model, and also Section 2.4 for the problem of identification in partially specified models.)

**I-3.** A set $D$ of experimental or non-experimental *data*, governed by a joint probability distribution presumably consistent with A.

The outputs are

**O-1.** A set $A^*$ of statements which are the logical implications of $A$, separate from the data at hand. For example, that $X$ has no effect on $Y$ if we hold $Z$ constant, or that $Z$ is an instrument relative to $\{X, Y\}$.

**O-2.** A set $C$ of data-dependent *claims* concerning the magnitudes or likelihoods of the target queries in $Q$, each contingent on $A$. $C$ may contain, for example, the estimated mean and variance of a given structural parameter, or the expected effect of a given intervention. Auxiliary to $C$, a causal model should also yield an estimand $Q_i(P)$ for each query in $Q$, or a determination that $Q_i$ is not identifiable from $P$ (Definition 2.)

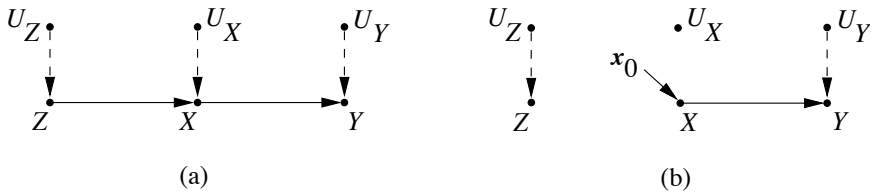


FIG 2. *The diagrams associated with (a) the structural model of equation (2.1) and (b) the modified model of equation (2.2), representing the intervention* $do(X = x_0)$.

**O-3.** A list $T$ of testable statistical implications of $A$ (which may or may not be part of O-2), and the degree $g(T_i), T_i \in T$, to which the data agrees with each of those implications. A typical implication would be a conditional independence assertion, or an equality constraint between two probabilistic expressions. Testable constraints should be read from the model $M_A$ (see Definition 3.), and used to confirm or disconfirm the model against the data.

The structure of this inferential exercise is shown schematically in Figure 1. For a comprehensive review on methodological issues, see (Pearl (2009a, 2012a)).

## 2.2 Assumptions in Nonparametric Models

A structural equation model (SEM) $M$ is defined as follows:

DEFINITION 1 (Structural Equation Model). (Pearl, 2000, p. 203)

1. *A set* $U$ *of background or exogenous variables, representing factors outside the model, which nevertheless affect relationships within the model.*
2. *A set* $V = \{V_1, ..., V_n\}$ *of endogenous variables, assumed to be observable. Each of these variables is functionally dependent on some subset* $PA_i$ *of* $U \cup V$.
3. *A set* $F$ *of functions* $\{f_1, ..., f_n\}$ *such that each* $f_i$ *determines the value of* $V_i \in V$, $v_i = f_i(pa_i, u)$.
4. *A joint probability distribution* $P(u)$ *over* $U$.

A simple SEM model is depicted in Fig. 2(a), which represents the following three functions:

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= f_X(z, u_X) \\ y &= f_Y(x, u_Y), \end{aligned} \tag{2.1}$$

where in this particular example, $U_Z$, $U_X$ and $U_Y$ are assumed to be jointly independent but otherwise arbitrarily distributed. Whenever dependence exists between any two exogenous variables, a bidirected arrow will be added to the diagram to represent this dependence (e.g., Fig. 4)[4]. Each of these functions

[4] More precisely, the absence of bidirected arrows implies marginal independences relative of the respective exogenous variables. In other words, the set of all bidirected edges constitute an i-map of $P(U)$ (Richardson, 2003).

represents a causal process (or mechanism) that determines the value of the left variable (output) from the values on the right variables (inputs), and is assumed to be invariant unless explicitly intervened on. The absence of a variable from the right-hand side of an equation encodes the assumption that nature ignores that variable in the process of determining the value of the output variable. For example, the absence of variable $Z$ from the arguments of $f_Y$ conveys the empirical claim that variations in $Z$ will leave Y unchanged, as long as variables $U_Y$ and $X$ remain constant.

It is important to distinguish between a *fully specified model* in which $P(U)$ and the collection of functions $F$ are specified and a *partially specified model*, usually in the form of a diagram. The former entails one and only one observational distribution $P(V)$; the latter entails a set of observational distributions $P(V)$ that are compatible with the graph (those that can be generated by specifying $\langle F, P(u)\rangle$).

### 2.3 Representing Interventions, Counterfactuals and Causal effects

This feature of invariance permits us to derive powerful claims about causal effects and counterfactuals, even in nonparametric models, where all functions and distributions remain unknown. This is done through a mathematical operator called $do(x)$, which simulates physical interventions by deleting certain functions from the model, replacing them with a constant $X = x$, while keeping the rest of the model unchanged (Haavelmo, 1943; Strotz and Wold, 1960; Pearl, 2012c). For example, to emulate an intervention $do(x_0)$ that sets $X$ to a constant $x_0$ in model $M$ of Figure 2(a), the equation for $x$ in equation (2.1) is replaced by $x = x_0$, and we obtain a new model, $M_{x_0}$,

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= x_0 \\ y &= f_Y(x, u_Y), \end{aligned} \tag{2.2}$$

the graphical description of which is shown in Figure 2(b).

The joint distribution associated with this modified model, denoted $P(z, y|do(x_0))$ describes the post-intervention distribution of variables $Y$ and $Z$ (also called "controlled" or "experimental" distribution), to be distinguished from the preintervention distribution, $P(x, y, z)$, associated with the original model of equation (2.1). For example, if $X$ represents a treatment variable, $Y$ a response variable, and $Z$ some covariate that affects the amount of treatment received, then the distribution $P(z, y|do(x_0))$ gives the proportion of individuals that would attain response level $Y = y$ and covariate level $Z = z$ under the hypothetical situation in which treatment $X = x_0$ is administered uniformly to the population.[5]

In general, we can formally define the postintervention distribution by the equation

$$P_M(y|do(x)) = P_{M_x}(y) \tag{2.3}$$

[5]Equivalently, $P(z, y|do(x_0))$ can be interpreted as the joint probability of $(Z = z, Y = y)$ under a randomized experiment among units receiving treatment level $X = x_0$. Readers versed in potential-outcome notations may interpret $P(y|do(x), z)$ as the probability $P(Y_x = y|Z_x = z)$, where $Y_x$ is the potential outcome under treatment $X = x$.

In words, in the framework of model $M$, the postintervention distribution of outcome $Y$ is defined as the probability that model $M_x$ assigns to each outcome level $Y = y$. From this distribution, which is readily computed from any fully specified model $M$, we are able to assess treatment efficacy by comparing aspects of this distribution at different levels of $x_0$.[6]

### 2.4 Identification, d-separation and Causal Calculus

A central question in causal analysis is the question of *identification* of causal queries (e.g., the effect of intervention $do(X = x_0)$) from a combination of data and a partially specified model, for example, when only the graph is given and neither the functions $F$ nor the distribution of $U$. In linear parametric settings, the question of identification reduces to asking whether some model parameter, $\beta$, has a unique solution in terms of the parameters of $P$ (say the population covariance matrix). In the nonparametric formulation, the notion of "has a unique solution" does not directly apply since quantities such as $Q(M) = P(y|do(x))$ have no parametric signature and are defined procedurally by simulating an intervention in a causal model $M$, as in equation (2.2). The following definition captures the requirement that $Q$ be estimable from the data:

DEFINITION 2 (Identifiability). *A causal query $Q(M)$ is identifiable, given a set of assumptions $A$, if for any two (fully specified) models, $M_1$ and $M_2$, that satisfy $A$, we have* [7]

$$P(M_1) = P(M_2) \Rightarrow Q(M_1) = Q(M_2) \tag{2.4}$$

In words, the functional details of $M_1$ and $M_2$ do not matter; what matters is that the assumptions in $A$ (e.g., those encoded in the diagram) would constrain the variability of those details in such a way that equality of $P$'s would entail equality of $Q$'s. When this happens, $Q$ depends on $P$ only, and should therefore be expressible in terms of the parameters of $P$.

When a query $Q$ is given in the form of a do-expression, for example $Q = P(y|do(x), z)$, its identifiability can be decided systematically using an algebraic procedure known as the do-calculus (Pearl, 1995). It consists of three inference rules that permit us to map interventional and observational distributions whenever certain conditions hold in the causal diagram G.

The conditions that permit the application these inference rules can be read off the diagrams using a graphical criterion known as d-separation (Pearl, 1988).

DEFINITION 3 ($d$-separation).
*A set $S$ of nodes is said to block a path $p$ if either*

[6] Counterfactuals are defined similarly through the equation $Y_x(u) = Y_{M_x}(u)$ (see (Pearl, 2009b, Ch. 7)), but will not be needed for the discussions in this paper.

[7] An implication similar to (2.4) is used in the standard statistical definition of parameter identification, where it conveys the uniqueness of a parameter set $\theta$ given a distribution $P_\theta$ (Lehmann and Casella, 1998). To see the connection, one should think about the query $Q = P(y|do(x))$ as a function $Q = g(\theta)$ where $\theta$ is the pair $F \cup P(u)$ that characterizes a fully specified model $M$.

1. *p contains at least one arrow-emitting node that is in S, or*
2. *p contains at least one collision node that is outside S and has no descendant in S.*

*If S blocks* all *paths from set X to set Y, it is said to "d-separate X and Y," and then, it can be shown that variables X and Y are independent given S, written* $X \perp\!\!\!\perp Y|S$.[8]

D-separation reflects conditional independencies that hold in any distribution $P(v)$ that is compatible with the causal assumptions $A$ embedded in the diagram. To illustrate, the path $U_Z \to Z \to X \to Y$ in Figure 2(a) is blocked by $S = \{Z\}$ and by $S = \{X\}$, since each emits an arrow along that path. Consequently we can infer that the conditional independencies $U_Z \perp\!\!\!\perp Y|Z$ and $U_Z \perp\!\!\!\perp Y|X$ will be satisfied in any probability function that this model can generate, regardless of how we parametrize the arrows. Likewise, the path $U_Z \to Z \to X \leftarrow U_X$ is blocked by the null set $\{\emptyset\}$, but it is not blocked by $S = \{Y\}$ since $Y$ is a descendant of the collision node $X$. Consequently, the marginal independence $U_Z \perp\!\!\!\perp U_X$ will hold in the distribution, but $U_Z \perp\!\!\!\perp U_X|Y$ may or may not hold.[9]

### 2.5 The Rules of do-calculus

Let $X$, $Y$, $Z$, and $W$ be arbitrary disjoint sets of nodes in a causal DAG $G$. We denote by $G_{\overline{X}}$ the graph obtained by deleting from $G$ all arrows pointing to nodes in $X$. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from $G$ all arrows emerging from nodes in $X$. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$.

The following three rules are valid for every interventional distribution compatible with $G$:

**Rule 1** (Insertion/deletion of observations):

$$P(y|do(x), z, w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{2.5}$$

**Rule 2** (Action/observation exchange):

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{2.6}$$

**Rule 3** (Insertion/deletion of actions):

$$P(y|do(x), do(z), w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\overline{Z(W)}}}, \tag{2.7}$$

where $Z(W)$ is the set of $Z$-nodes that are not ancestors of any $W$-node in $G_{\overline{X}}$.

To establish identifiability of a query $Q$, one needs to repeatedly apply the rules of do-calculus to $Q$, until the final expression no longer contains a do-operator[10]; this renders it estimable from non-experimental data. The *do*-calculus

[8] See Hayduk et al. (2003), Glymour and Greenland (2008), and Pearl (2009b, p. 335) for a gentle introduction to *d*-separation.

[9] This special handling of collision nodes (or *colliders*, e.g., $Z \to X \leftarrow U_X$) reflects a general phenomenon known as *Berkson's paradox* (Berkson, 1946), whereby observations on a common consequence of two independent causes render those causes dependent. For example, the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail.

[10] Such derivations are illustrated in graphical details in (Pearl, 2009b, pp. 87).

was proven to be complete for the identifiability of causal effects in the form $Q = P(y|do(x), z)$ (Shpitser and Pearl, 2006; Huang and Valtorta, 2006), which means that if $Q$ cannot be expressed in terms of the probability of observables $P$ by repeated application of these three rules, such an expression does not exist. In other words, the query is not estimable from observational studies without making further assumptions, for example, linearity, monotonicity, additivity, absence of interactions, etc.

We shall see that, to establish transportability, the goal will be different; instead of eliminating do-operators from the query expression, we will need to separate them from a set of variables S that represent disparities between populations.

## 3. INFERENCE ACROSS POPULATIONS: MOTIVATING EXAMPLES

To motivate the treatment of Section 4, we first demonstrate some of the subtle questions that transportability entails through three simple examples, informally depicted in Fig. 3.

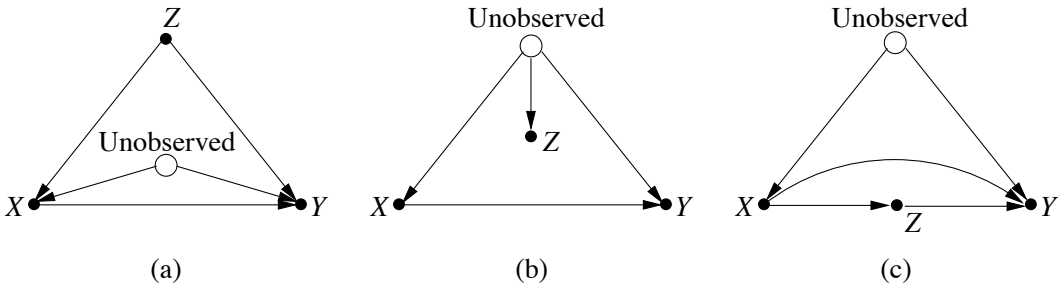


FIG 3. *Causal diagrams depicting Examples 1–3. In (a) Z represents "age." In (b) Z represents "linguistic skills" while age (in hollow circle) is unmeasured. In (c) Z represents a biological marker situated between the treatment* ($X$) *and a disease* ($Y$).

EXAMPLE 1. *Consider the graph in Fig. 3(a) that represents cause-effect relationships in the pre-treatment population in Los Angeles. We conduct a randomized trial in Los Angeles and estimate the causal effect of exposure $X$ on outcome $Y$ for every age group $Z = z$.* [11] [12] *We now wish to generalize the results to the population of New York City (NYC), but data alert us to the fact that the study distribution $P(x, y, z)$ in LA is significantly different from the one in NYC (call the latter $P^*(x, y, z)$). In particular, we notice that the average age in NYC is significantly higher than that in LA. How are we to estimate the causal effect of $X$ on $Y$ in NYC, denoted $P^*(y|do(x))$?*

Our natural inclination would be to assume that age-specific effects are invariant across cities and so, if the LA study provides us with (estimates of)

[11] Throughout the paper, each graph represents the causal structure of the population prior to the treatment, hence $X$ stands for the level of treatment taken by an individual out of free choice.

[12] The arrow from $Z$ to $X$ represents the tendency of older people to seek treatment more often than younger people, and the arrow from $Z$ to $Y$ represents the effect of age on the outcome.

age-specific causal effects $P(y|do(x), Z = z)$, the overall causal effect in NYC should be

$$P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z) \tag{3.1}$$

This *transport formula* combines experimental results obtained in LA, $P(y|do(x), z)$, with observational aspects of NYC population, $P^*(z)$, to obtain an experimental claim $P^*(y|do(x))$ about NYC.[13]

Our first task in this paper will be to explicate the assumptions that renders this extrapolation valid. We ask, for example, what must we assume about other confounding variables beside age, both latent and observed, for Eq. (3.1) to be valid, or, would the same transport formula hold if $Z$ was not age, but some proxy for age, say, language proficiency. More intricate yet, what if $Z$ stood for an exposure-dependent variable, say hyper-tension level, that stands between $X$ and $Y$?

Let us examine the proxy issue first.

EXAMPLE 2. *Let the variable $Z$ in Example 1 stand for subjects language proficiency, and let us assume that $Z$ does not affect exposure ($X$) or outcome ($Y$), yet it correlates with both, being a proxy for age which is not measured in either study (see Fig. 3(b)). Given the observed disparity $P(z) \neq P^*(z)$, how are we to estimate the causal effect $P^*(y|do(x))$ for the target population of NYC from the $z$-specific causal effect $P(y|do(x), z)$ estimated at the study population of LA?*

The inequality $P(z) \neq P^*(z)$ in this example may reflect either age difference or differences in the way that $Z$ correlates with age. If the two cities enjoy identical age distributions and NYC residents acquire linguistic skills at a younger age, then, since $Z$ has no effect whatsoever on $X$ and $Y$, the inequality $P(z) \neq P^*(z)$ can be ignored and, intuitively, the proper transport formula would be

$$P^*(y|do(x)) = P(y|do(x)) \tag{3.2}$$

If, on the other hand, the conditional probabilities $P(z|age)$ and $P^*(z|age)$ are the same in both cities, and the inequality $P(z) \neq P^*(z)$ reflects genuine age differences, Eq. (3.2) is no longer valid, since the age difference may be a critical factor in determining how people react to $X$. We see, therefore, that the choice of the proper transport formula depends on the causal context in which population differences are embedded.

This example also demonstrates why the invariance of $Z$-specific causal effects should not be taken for granted. While justified in Example 1, with $Z$ = age, it fails in Example 2, in which $Z$ was equated with "language skills." Indeed, using

[13] At first glance, Eq. (3.1) may be regarded as a routine application of "standardization" or "recalibration" – a statistical extrapolation method that can be traced back to a century-old tradition in demography and political arithmetic (Westergaard, 1916; Yule, 1934; Lane and Nelder, 1982). On a second thought it raises the deeper question of why we consider age-specific effects to be invariant across populations. See discussion following Example 2.

Fig. 3(b) for guidance, the $Z$-specific effect of $X$ on $Y$ in NYC is given by:

$$\begin{aligned} P^*(y|do(x),z) &= \sum_{age} P^*(y|do(x),z,age)P^*(age|do(x),z) \\ &= \sum_{age} P^*(y|do(x),age)P^*(age|z) \\ &= \sum_{age} P(y|do(x),age)P^*(age|z) \end{aligned}$$

Thus, if the two populations differ in the relation between age and skill, i.e.,

$$P(age|z) \neq P^*(age|z)$$

the skill-specific causal effect would differ as well.

The intuition is clear. A NYC person at skill level $Z = z$ is likely to be in a totally different age group from his skill-equals in Los Angeles and, since it is age, not skill that shapes the way individuals respond to treatment, it is only reasonable that Los Angeles residents would respond differently to treatment than their NYC counterparts at the very same skill level.

The essential difference between Examples 1 and 2 is that age is normally taken to be an exogenous variable (not assigned by other factors in the model) while skills may be indicative of earlier factors (age, education, ethnicity) capable of modifying the causal effect. Therefore, conditional on skill, the effect may be different in the two populations.

EXAMPLE 3. *Examine the case where $Z$ is a $X$-dependent variable, say a disease bio-marker, standing on the causal pathways between $X$ and $Y$ as shown in Fig. 3(c). Assume further that the disparity $P(z|x) \neq P^*(z|x)$ is discovered and that, again, both the average and the $z$-specific causal effect $P(y|do(x),z)$ are estimated in the LA experiment, for all levels of $X$ and $Z$. Can we, based on information given, estimate the average (or $z$-specific) causal effect in the target population of NYC?*

Here, Eq. (3.1) is wrong because the overall causal effect (in both LA and NYC) is no longer a simple average of the $z$-specific causal effects. The correct weighing rule is

$$P^*(y|do(x)) = \sum_z P^*(y|do(x),z)P^*(z|do(x)) \tag{3.3}$$

which reduces to (3.1) only in the special case where $Z$ is unaffected by $X$. Eq. (3.2) is also wrong because we can no longer argue, as we did in Ex. 2, that $Z$ does not affect $Y$, hence it can be ignored. Here, $Z$ lies on the causal pathway between $X$ and $Y$ so, clearly, it affects their relationship. What then is the correct transport formula for this scenario?

To cast this example in a more realistic setting, let us assume that we wish to use $Z$ as a "surrogate endpoint" to predict the efficacy of treatment $X$ on outcome $Y$, where $Y$ is too difficult and/or expensive to measure routinely (Prentice, 1989; Ellenberg and Hamilton, 1989). Thus, instead of considering experimental and observational studies conducted at two different locations, we consider two such studies taking place at the same location, but at different times. In the first

study, we measure $P(y, z|do(x))$ and discover that $Z$ is a good surrogate, namely, knowing the effect of treatment on $Z$ allows prediction of the effect of treatment on the more clinically relevant outcome ($Y$) (Joffe and Green, 2009). Once $Z$ is proclaimed a "surrogate endpoint", it invites efforts to find direct means of controlling $Z$. For example, if cholesterol level is found to be a predictor of heart diseases in a long-run trial, drug manufacturers would rush to offer cholesterol-reducing substances for public consumption. As a result, both the prior $P(z)$ and the treatment-dependent probability $P(z|do(x))$ would undergo a change, resulting in $P^*(z)$ and $P^*(z|do(x))$, respectively.

We now wish to re-assess the effect of the drug $P^*(y|do(x))$ in the new population and do it in the cheapest possible way, namely, by conducting an observational study to estimate $P^*(z, x)$, acknowledging that confounding exists between $X$ and $Y$ and that the drug affects $Y$ both directly and through $Z$, as shown in Fig. 3(c).

Using a graphical representation to encode the assumptions articulated thus far, and further assuming that the disparity observed stems only from a difference in people's susceptibility to $X$ (and not due to a change in some unobservable confounder), we will prove in Section 5 that the correct transport formula should be

$$P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z|x), \tag{3.4}$$

which is different from both (3.1) and (3.2). It calls instead for the $z$-specific effects to be re-weighted by the conditional probability $P^*(z|x)$, estimated in the target population. [14]

To see how the transportability problem fits into the general scheme of causal analysis discussed in Section 2.1 (Fig. 1), we note that, in our case, the data comes from two sources, experimental (from the study) and non-experimental (from the target), assumptions are encoded in the form of selection diagrams, and the query stands for the causal effect (e.g., $P^*(y|do(x))$). Although this paper does not discuss the goodness-of-fit problem, standard methods are available for testing the compatibility of the selection diagram with the data available.

## 4. FORMALIZING TRANSPORTABILITY

### 4.1 Selection diagrams and selection variables

The pattern that emerges from the examples discussed in Section 3 indicates that transportability is a causal, not statistical notion. In other words, the conditions that license transport as well as the formulas through which results are transported depend on the causal relations between the variables in the domain, not merely on their statistics. For instance, it was important in Example 3 to ascertain that the change in $P(z|x)$ was due to the change in the way $Z$ is affected by $X$, but not due to a change in confounding conditions between the two. This cannot be determined solely by comparing $P(z|x)$ and $P^*(z|x)$. If $X$ and $Z$ are confounded (e.g., Fig. 6(e)), it is quite possible for the inequality

[14] Quite often the possibility of running a second randomized experiment to estimate $P^*(z|do(x))$ is also available to investigators, though at a higher cost. In such cases, a transport formula would be derivable under more relaxed assumptions, for example, allowing for $X$ and $Z$ to be confounded.

$P(z|x) \neq P^*(z|x)$ to hold, reflecting differences in confounding, while the way that $Z$ is affected by $X$ (i.e., $P(z|do(x))$) is the same in the two populations – a different transport formula will then emerge for this case.

Consequently, licensing transportability requires knowledge of the mechanisms, or processes, through which population differences come about; different localization of these mechanisms yield different transport formulae. This can be seen most vividly in Example 2 (Fig. 3(b)) where we reasoned that no re-weighing is necessary if the disparity $P(z) \neq P^*(z)$ originates with the way language proficiency depends on age, while the age distribution itself remains the same. Yet, because age is not measured, this condition cannot be detected in the probability distribution $P$, and cannot be distinguished from an alternative condition,

$$P(age) \neq P^*(age) \quad \text{and} \quad P(z|age) = P^*(z|age),$$

one that may require re–weighting according to to Eq. (3.1). In other words, every probability distribution $P(x, y, z)$ that is compatible with the process of Fig. 3(b) is also compatible with that of Fig. 3(a) and, yet, the two processes dictate different transport formulas.

Based on these observations, it is clear that if we are to represent formally the differences between populations (similarly, between experimental settings or environments), we must resort to a representation in which the causal mechanisms are explicitly encoded and in which differences in populations are represented as local modifications of those mechanisms.

To this end, we will use causal diagrams augmented with a set, $S$, of "selection variables," where each member of $S$ corresponds to a mechanism by which the two populations differ, and switching between the two populations will be represented by conditioning on different values of these $S$ variables.[15]

Intuitively, if $P(v|do(x))$ stands for the distribution of a set $V$ of variables in the experimental study (with X randomized) then we designate by $P^*(v|do(x))$ the distribution of $V$ if we were to conduct the study on population $\Pi^*$ instead of $\Pi$. We now attribute the difference between the two to the action of a set $S$ of selection variables, and write[16] [17]

$$P^*(v|do(x)) = P(v|do(x), s^*).$$

The selection variables in $S$ may represent all factors by which populations may differ or that may "threaten" the transport of conclusions between populations. For example, in Fig. 4(a) the age disparity $P(z) \neq P^*(z)$ discussed in Example 1 will be represented by the inequality

$$P(z) \neq P(z|s)$$

Text

[15] Disparities among populations or sub-populations can also arise from differences in design; for example, if two samples are drawn by different criteria from a given population. The problem of generalizing between two such sub-populations is usually called *sampling selection bias* (Heckman, 1979; Hernán et al., 2004; Cole and Stuart, 2010; Pearl, 2013; Bareinboim et al., 2014). In this paper, we deal only with nature-induced, not man-made disparities.

[16] Alternatively, one can represent the two populations' distributions by $P(v|do(x), s)$, and $P(v|do(x), s^*)$, respectively. The results, however, will be the same, since only the location of $S$ enters the analysis.

[17] Pearl (1993; 2009b, p. 71), Spirtes et al. (1993), and Dawid (2002), for example, use conditioning on auxiliary variables to switch between experimental and observational studies. Dawid (2002) further uses such variables to represent changes in parameters of probability distributions.

where $S$ stands for all factors responsible for drawing subjects at age $Z = z$ to NYC rather than LA.

Of equal importance is the absence of an $S$ variable pointing to $Y$ in Fig. 4(a), which encodes the assumption that age-specific effects are invariant across the two populations.

This graphical representation, which we will call "selection diagrams" is defined as follows:[18]

DEFINITION 4 (Selection Diagram). *Let $\langle M, M^* \rangle$ be a pair of structural causal models (Definition 1) relative to domains $\langle \Pi, \Pi^* \rangle$, sharing a causal diagram $G$. $\langle M, M^* \rangle$ is said to induce a selection diagram $D$ if $D$ is constructed as follows:*

1. *Every edge in $G$ is also an edge in $D$;*
2. *$D$ contains an extra edge $S_i \rightarrow V_i$ whenever there might exist a discrepancy $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$ between $M$ and $M^*$.*

In summary, the $S$-variables locate the *mechanisms* where structural discrepancies between the two populations are suspected to take place. Alternatively, the absence of a selection node pointing to a variable represents the assumption that the mechanism responsible for assigning value to that variable is the same in the two populations. In the extreme case, we could add selection nodes to all variables, which means that we have no reason to believe that the populations share any mechanism in common, and this, of course would inhibit any exchange of information among the populations. The invariance assumptions between populations, as we will see, will open the door for the transport of some experimental findings.

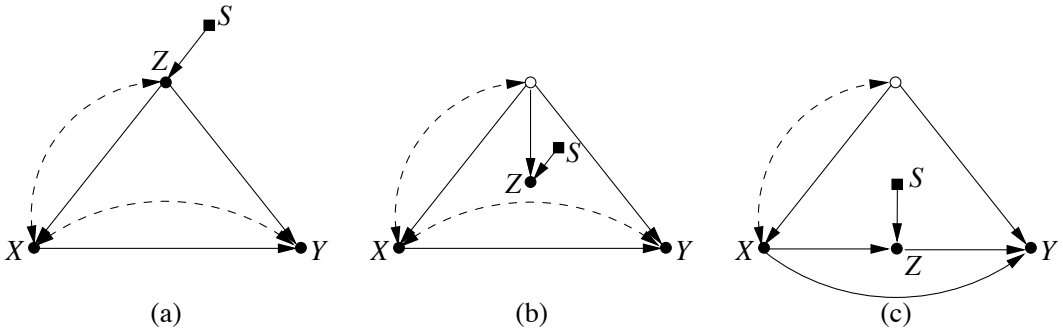


FIG 4. *Selection diagrams depicting specific versions of Examples 1–3. In (a) the two populations differ in age distributions. In (b) the populations differs in how $Z$ depends on age (an unmeasured variable, represented by the hollow circle) and the age distributions are the same. In (c) the populations differ in how $Z$ depends on $X$. In all diagrams, dashed arcs (e.g., $X \dashleftarrow\dashrightarrow Y$) represent the presence of latent variables affecting both $X$ and $Y$.*

For clarity, we will represent the $S$ variables by squares, as in Fig. 4, which uses selection diagrams to encode the three examples discussed in Section 3. (Besides

[18]The assumption that there are no structural changes between domains can be relaxed starting with $D = G^*$ and adding $S$-nodes following the same procedure as in Def. 4, while enforcing acyclicity. In extreme cases in which the two domains differ in causal directionality (Spirtes et al., 2000, pp. 298–99), acyclicity cannot be maintained. This complication as well as one created when $G$ is a edge-super set of $G^*$ require a more elaborated graphical representation and lie beyond the scope of this paper.

the $S$ variables, these graphs also include additional latent variables, represented by bidirected edges, which makes the examples more realistic.) In particular, Fig. 4(a) and 4(b) represent, respectively, two different mechanisms responsible for the observed disparity $P(z) \neq P^*(z)$. The first (Fig. 4(a)) dictates transport formula (3.1), while the second (Fig. 4(b)) calls for direct, unadjusted transport (3.2). This difference stems from the location of the $S$ variables in the two diagrams. In Fig. 4(a), the $S$ variable represents unspecified factors that cause age differences between the two populations, while in Fig. 4(b), $S$ represents factors that cause differences in reading skills ($Z$) while the age distribution itself (unobserved) remains the same.

In this paper, we will address the issue of transportability assuming that scientific knowledge about invariance of certain mechanisms is available and encoded in the selection diagram through the S nodes. Such knowledge is, admittedly, more demanding than that which shapes the structure of each causal diagram in isolation. It is, however, a prerequisite for any attempt to justify transfer of findings across populations, which makes selection diagrams a mathematical object worthy of analysis

### 4.2 Transportability: Definitions and Examples

Using selection diagrams as the basic representational language, and harnessing the concepts of intervention, *do*-calculus, and identifiability (Section 2), we can now give the notion of transportability a formal definition.

DEFINITION 5 (Transportability). *Let $D$ be a selection diagram relative to domains $\langle \Pi, \Pi^* \rangle$. Let $\langle P, I \rangle$ be the pair of observational and interventional distributions of $\Pi$, and $P^*$ be the observational distribution of $\Pi^*$. The causal relation $R(\Pi^*) = P^*(y|do(x), z)$ is said to be transportable from $\Pi$ to $\Pi^*$ in $D$ if $R(\Pi^*)$ is uniquely computable from $P, P^*, I$ in any model that induces $D$.*

Two interesting connections between identifiability and transportability are worth noting. First, note that all identifiable causal relations in $D$ are also transportable, because they can be computed directly from $P^*$ and require no experimental information from $\Pi$. Second, note that given causal diagram $G$, one can produce a selection diagram $D$ such that identifiability in $G$ is equivalent to transportability in $D$. First set $D = G$, and then add selection nodes pointing to all variables in $D$, which represents that the target domain does not share any mechanism with its counterpart – this is equivalent to the problem of identifiability because the only way to achieve transportability is to identify $R$ from scratch in the target population.

While the problems of identifiability and transportability are related, proofs of non-transportability are more involved than those of non-identifiability for they require one to demonstrate the non-existence of two competing models compatible with $D$, agreeing on $\{P, P^*, I\}$, and disagreeing on $R(\Pi^*)$.

Definition 5 is declarative, and does not offer an effective method of demonstrating transportability even in simple models. Theorem 1 offers such a method using a sequence of derivations in do-calculus.

THEOREM 1. *Let $D$ be the selection diagram characterizing two populations, $\Pi$ and $\Pi^*$, and $S$ a set of selection variables in $D$. The relation $R = P^*(y|do(x), z)$*

*is transportable from* $\Pi$ *to* $\Pi^*$ *if the expression* $P(y|do(x), z, s)$ *is reducible, using the rules of do-calculus, to an expression in which* $S$ *appears only as a conditioning variable in do-free terms.*

PROOF. Every relation satisfying the condition of Theorem 1 can be written as an algebraic combination of two kinds of terms, those that involve $S$ and those that do not. The former can be written as $P^*$-terms and are estimable, therefore, from observations on $\Pi^*$, as required by Definition 5. All other terms, especially those involving *do*-operators, do not contain $S$; they are experimentally identifiable therefore in $\Pi$. □

This criterion was proven to be both sufficient and necessary for causal effects, namely $R = P^*(y|do(x))$ (Bareinboim and Pearl, 2012). Theorem 1, though procedural, does not specify the sequence of rules leading to the needed reduction when such a sequence exists. (Bareinboim and Pearl, 2013b) derived a complete procedural solution for this, based on graphical method developed in (Tian and Pearl, 2002; Shpitser and Pearl, 2006). Despite its completeness, however, the procedural solution is not trivial, and we take here an alternative route to establish a simple and transparent procedure for confirming transportability, guided by two recognizable subgoals.

DEFINITION 6. *(Trivial Transportability)*
*A causal relation* $R$ *is said to be* trivially transportable *from* $\Pi$ *to* $\Pi^*$*, if* $R(\Pi^*)$ *is identifiable from* $(G^*, P^*)$.

This criterion amounts to an ordinary test of identifiability of causal relations using graphs, as given by Definition 2. It permits us to estimate $R(\Pi^*)$ directly from observational studies on $\Pi^*$, un-aided by causal information from $\Pi$.

EXAMPLE 4. *Let* $R$ *be the causal effect* $P^*(y|do(x))$ *and let the selection diagram of* $\Pi$ *and* $\Pi^*$ *be given by* $X \to Y \leftarrow S$*, then* $R$ *is trivially transportable, since* $R(\Pi^*) = P^*(y|x)$.

Another special case of transportability occurs when a causal relation has identical form in both domains – no recalibration is needed.

DEFINITION 7. *(Direct Transportability)*
*A causal relation* $R$ *is said to be* directly transportable *from* $\Pi$ *to* $\Pi^*$*, if* $R(\Pi^*) = R(\Pi)$.

A graphical test for direct transportability of $R = P^*(y|do(x), z)$ follows from do-calculus and reads: $(S \perp\!\!\!\perp Y|X, Z)_{G_{\overline{X}}}$; in words, $X$ blocks all paths from $S$ to $Y$ once we remove all arrows pointing to $X$ and condition on $Z$. As a concrete example, this test is satisfied in Fig. 4(a), and therefore, the $z$-specific effects is the same in both populations; it is directly transportable.

**Remark.**
The notion of "external validity" as defined by Manski (2007) (footnote 1) corresponds to Direct Transportability, for it requires that $R$ retains its validity without adjustment, as in Eq. (3.2). Such conditions preclude the use of information from $\Pi^*$ to recalibrate $R$.

EXAMPLE 5. *Let $R$ be the causal effect of $X$ on $Y$, and let $D$ have a single $S$ node pointing to $X$, then $R$ is directly transportable, because causal effects are independent of the selection mechanism (see Pearl, 2009b, pp. 72–73).*

EXAMPLE 6. *Let $R$ be the $z$-specific causal effect of $X$ on $Y$ $P^*(y|do(x), z)$ where $Z$ is a set of variables, and $P$ and $P^*$ differ only in the conditional probabilities $P(z|pa(Z))$ and $P^*(z|pa(Z))$ such that $(Z \perp\!\!\!\perp Y|pa(Z))$, as shown in Fig. 4(b). Under these conditions, $R$ is not directly transportable. However, the $pa(Z)$-specific causal effects $P^*(y|do(x), pa(Z))$ are directly transportable, and so is $P^*(y|do(x))$. Note that, due to the confounding arcs, none of these quantities is identifiable.*

## 5. TRANSPORTABILITY OF CAUSAL EFFECTS - A GRAPHICAL CRITERION

We now state and prove two theorems that permit us to decide algorithmically, given a selection diagram, whether a relation is transportable between two populations, and what the transport formula should be.

THEOREM 2. *Let $D$ be the selection diagram characterizing two populations, $\Pi$ and $\Pi^*$, and $S$ the set of selection variables in $D$. The strata-specific causal effect $P^*(y|do(x), z)$ is transportable from $\Pi$ to $\Pi^*$ if $Z$ d-separates $Y$ from $S$ in the $X$-manipulated version of $D$, that is, $Z$ satisfies $(Y \perp\!\!\!\perp S|Z, X)_{D_{\overline{X}}}$.*

PROOF.

$$P^*(y|do(x), z) = P(y|do(x), z, s^*)$$

From Rule-1 of *do*-calculus we have: $P(y|do(x), z, s^*) = P(y|do(x), z)$ whenever $Z$ satisfies $(Y \perp\!\!\!\perp S|Z, X)$ in $D_{\overline{X}}$. This proves Theorem 2. □

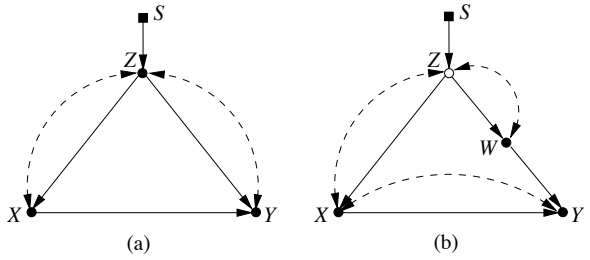


FIG 5. *Selection diagrams illustrating S-admissibility. (a) has no S-admissible set while in (b), $W$ is S-admissible.*

DEFINITION 8. *(S-admissibility)*
*A set $T$ of variables satisfying $(Y \perp\!\!\!\perp S|T, X)$ in $D_{\overline{X}}$ will be called S-admissible (with respect to the causal effect of $X$ on $Y$).*

COROLLARY 1. *The average causal effect $P^*(y|do(x))$ is transportable from $\Pi$ to $\Pi^*$ if there exists a set $Z$ of observed pre-treatment covariates that is S-admissible. Moreover, the transport formula is given by the weighting of Eq. (3.1).*

EXAMPLE 7. *The causal effect is transportable in Fig. 4(a), since $Z$ is $S$-admissible, and in Fig. 4(b), where the empty set is $S$-admissible. It is also transportable by the same criterion in Fig. 5(b), where $W$ is $S$-admissible, but not in Fig. 5(a) where no $S$-admissible set exists.*

COROLLARY 2. *Any $S$ variable that is pointing directly into $X$ as in Fig. 6(a), or that is d-separated from $Y$ in $D_{\overline{X}}$ can be ignored.*

This follows from the fact that the empty set is $S$-admissible relative to any such $S$ variable. Conceptually, the corollary reflects the understanding that differences in propensity to receive treatment do not hinder the transportability of treatment effects; the randomization used in the experimental study washes away such differences.

We now generalize Theorem 2 to cases involving treatment-dependent $Z$ variables, as in Fig. 4(c).

THEOREM 3. *The average causal effect $P^*(y|do(x))$ is transportable from $\Pi$ to $\Pi^*$ if either one of the following conditions holds*

1. *$P^*(y|do(x))$ is trivially transportable;*
2. *There exists a set of covariates, $Z$ (possibly affected by $X$) such that $Z$ is $S$-admissible and for which $P^*(z|do(x))$ is transportable;*
3. *There exists a set of covariates, $W$ that satisfy $(X \perp\!\!\!\perp Y|W)_{D_{\overline{X(W)}}}$ and for which $P^*(w|do(x))$ is transportable.*

PROOF. 1. Condition (1) entails transportability.

2. If condition (2) holds, it implies

$$P^*(y|do(x)) = P(y|do(x), s) \tag{5.1}$$

$$= \sum_z P(y|do(x), z, s)P(z|do(x), s) \tag{5.2}$$

$$= \sum_z P(y|do(x), z)P^*(z|do(x)) \tag{5.3}$$

We now note that the transportability of $P(z|do(x))$ should reduce $P^*(z|do(x))$ to a star-free expression and would render $P^*(y|do(x))$ transportable.

3. If condition (3) holds, it implies

$$P^*(y|do(x)) = P(y|do(x), s) \tag{5.4}$$

$$= \sum_w P(y|do(x), w, s)P(w|do(x), s) \tag{5.5}$$

$$= \sum_w P(y|w, s)P^*(w|do(x)) \tag{5.6}$$

(by Rule-3 of *do*-calculus)

$$= \sum_w P^*(y|w)P^*(w|do(x)) \tag{5.7}$$

We similarly note that the transportability of $P^*(w|do(x))$ should reduce $P(w|do(x), s)$ to a star-free expression and would render $P^*(y|do(x))$ transportable. This proves Theorem 3.

□

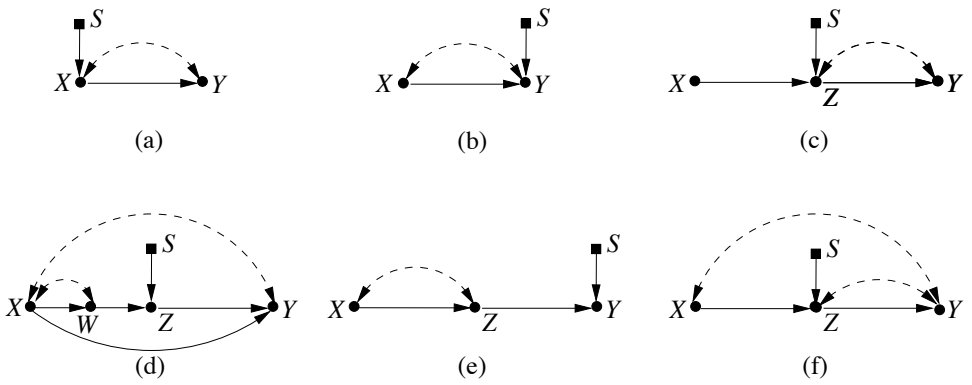


FIG 6. *Selection diagrams illustrating transportability. The causal effect $P(y|do(x))$ is (trivially) transportable in (c) but not in (b) and (f). It is transportable in (a), (d), and (e) (see Corollary 2).*

EXAMPLE 8. *To illustrate the application of Theorem 3, let us apply it to Fig. 4(c), which corresponds to the surrogate endpoint problem discussed in Section 3 (Example 3). Our goal is to estimate $P^*(y|do(x))$ – the effect of $X$ on $Y$ in the new population created by changes in how $Z$ responds to $X$. The structure of the problem permits us to satisfy condition 2 of the theorem, since $Z$ is $S$-admissible and $P^*(z|do(x))$ is trivially transportable. The former can be seen from $(S \perp\!\!\!\perp Y|X,Z)_{G_{\overline{X}}}$, hence $P^*(y|do(x),z) = P(y|do(x),z)$); the latter can be seen from the fact that $X$ and $Z$ and unconfounded, hence $P^*(z|do(x)) = P^*(z|x)$. Putting the two together, we get*

$$P^*(y|do(x)) = \sum_z P(y|do(x),z)P^*(z|x), \tag{5.8}$$

*which proves Eq. (3.4).*

**Remark.**
The test entailed by Theorem 3 is recursive, since the transportability of one causal effect depends on that of another. However, given that the diagram is finite and acyclic, the sets $Z$ and $W$ needed in conditions 2 and 3 of Theorem 3 would become closer and closer to $X$, and the iterative process will terminate after a finite number of steps. This occurs because the causal effects $P^*(z|do(x))$ (likewise, $P^*(w|do(x))$) is trivially transportable and equals $P(z)$ for any $Z$ node that is not a descendant of $X$. Thus, the need for reiteration applies only to those members of $Z$ that lie on the causal pathways from $X$ to $Y$. Note further that the analyst need not terminate the procedure upon satisfying the condition in Theorem 3. If one wishes to reduce the number of experiments, it can continue until no further reduction is feasible.

EXAMPLE 9. *Fig. 6(d) requires that we invoke both conditions of Theorem 3, iteratively. To satisfy condition 2 we note that $Z$ is $S$-admissible, and we need to prove the transportability of $P^*(z|do(x))$. To do that, we invoke condition 3 and note that $W$ d-separates $X$ from $Z$ in $D$. There remains to confirm the transportability of $P^*(w|do(x))$, but this is guaranteed by the fact that the empty set*

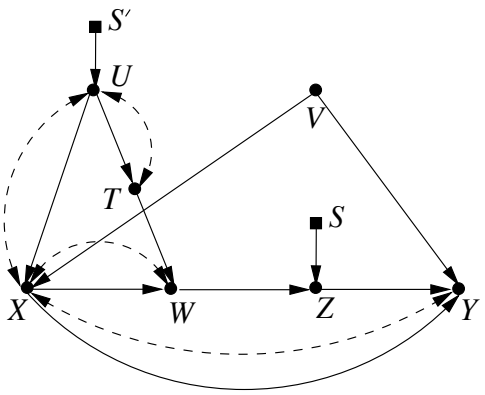


FIG 7. *Selection diagram in which the causal effect is shown to be transportable in multiple iterations of Theorem 3 (see Appendix 1).*

*is $S$-admissible relative to $W$, since $(W \perp\!\!\!\perp S)$. Hence, by Theorem 2 (replacing $Y$ with $W$) $P^*(w|do(x))$ is transportable, which bestows transportability on $P^*(y|do(x))$. Thus, the final transport formula (derived formally in Appendix 1) is:*

$$P^*(y|do(x)) = \sum_z P(y|do(x), z) \sum_w P(w|do(x)) P^*(z|w) \tag{5.9}$$

*The first two factors of the expression are estimable in the experimental study, and the third through observational studies on the target population. Note that the joint effect $P(y, w, z|do(x))$ need not be estimated in the experiment; a decomposition that results in decrease of measurement cost and sampling variability.*

A similar analysis proves the transportability of the causal effect in Fig. 6(e) (see Pearl and Bareinboim (2011)). The model of Fig. 6(f) however does not allow for the transportability of $P^*(y|do(x))$ as witnessed by the absence of $S$-admissible set in the diagram, and the inapplicability of condition 3 of Theorem 3.

EXAMPLE 10. *To illustrate the power of Theorem 3 in discerning transportability and deriving transport formulae, Fig. 7 represents a more intricate selection diagram, which requires several iteration to discern transportability. The transport formula for this diagram is given by (derived formally in Appendix 1):*

$$P^*(y|do(x)) = \sum_z P(y|do(x), z) \sum_w P^*(z|w) \sum_t P(w|do(x), t) P^*(t) \tag{5.10}$$

The main power of this formula is to guide investigators in deciding what measurements need be taken in both the experimental study and the target population. It asserts, for example, that variables $U$ and $V$ need not be measured. It likewise asserts that the $W$-specific causal effects need not be estimated in the experimental study and only the conditional probabilities $P^*(z|w)$ and $P^*(t)$ need be estimated in the target population. The derivation of this formulae is given in Appendix 1.

Despite its power, Theorem 3 in not complete, namely, it is not guaranteed to approve all transportable relations or to disapprove all non-transportable ones.

An example of the former is contrived in Bareinboim and Pearl (2012), where an alternative, necessary and sufficient condition is established in both graphical and algorithmic form. Theorem 3 provides, nevertheless, a simple and powerful method of establishing transportability in practice.

## 6. CONCLUSIONS

Given judgements of how target populations may differ from those under study, the paper offers a formal representational language for making these assessments precise and for deciding whether causal relations in the target population can be inferred from those obtained in an experimental study. When such inference is possible, the criteria provided by Theorems 2 and 3 yield transport formulae, namely, principled ways of calibrating the transported relations so as to properly account for differences in the populations. These transport formulae enable the investigator to select the essential measurements in both the experimental and observational studies, and thus minimize measurement costs and sample variability.

The inferences licensed by Theorem 2 and 3 represent worst case analysis, since we have assumed, in the tradition of nonparametric modeling, that every variable may potentially be an effect-modifier (or moderator.) If one is willing to assume that certain relationships are non interactive, or monotonic as is the case in additive models, then additional transport licenses may be issued, beyond those sanctioned by Theorems 2 and 3.

While the results of this paper concern the transfer of causal information from experimental to observational studies, the method can also benefit in transporting statistical findings from one observational study to another (Pearl and Bareinboim (2011)). The rationale for such transfer is two fold. First, information from the first study may enable researchers to avoid repeated measurement of certain variables in the target population. Second, by pooling data from both populations, we increase the precision in which their commonalities are estimated and, indirectly, also increase the precision by which the target relationship is transported. Substantial reduction in sampling variability can be thus achieved through this decomposition (Pearl (2012b)).

Clearly, the same data-sharing philosophy can be used to guide Meta-Analysis (Glass, 1976; Hedges and Olkin, 1985; Rosenthal, 1995; Owen, 2009), where one attempts to combine results from many experimental and observational studies, each conducted on a different population and under a different set of conditions, so as to construct an aggregate measure of effect size that is "better," in some formal sense, than any one study in isolation. While traditional approaches aims to average out differences between studies, our theory exploits the commonalities among the populations studied and the target population. By pooling together commonalities and discarding areas of disparity we gain maximum use of the available samples (Bareinboim and Pearl (2013c)).

To be of immediate use, our method relies on the assumption that the analyst is in possession of sufficient background knowledge to determine, at least qualitatively, where two populations may differ from one another. This knowledge is not vastly different from that required in any principled approach to causation in observational studies, since judgement about possible effects of omitted factors is crucial in any such analysis. Whereas such knowledge may only be partially

available, the analysis presented in this paper is nevertheless essential for understanding what knowledge is needed for the task to succeed and how sensitive conclusions are to knowledge that we do not possess.

Real-life situations will be marred, of course, with additional complications that were not addressed directly in this paper; for example, measurement errors, selection bias, finite sample variability, uncertainty about the graph structure, and the possible existence of unmeasured confounders between any two nodes in the diagram. Such issues are not unique to transportability; they plague any problem in causal analysis, regardless of whether they are represented formally or ignored by avoiding formalism. The methods offered in this paper are representative of what theory permits us to do in ideal situations, and the graphical representation presented in this paper makes the assumptions explicit and transparent. Transparency is essential for reaching tentative consensus among researchers and for facilitating discussions to distinguish that which is deemed plausible and important from that which is negligible or implausible.

Finally, it is important to mention two recent extensions of the results reported in this article. (Bareinboim and Pearl (2013a)) have addressed the problem of transportability in cases where only a limited set of experiments can be conducted at the source environment. Subsequently, the results were generalized to the problem of meta-transportability, that is, pooling experimental results from multiple and disparate sources to synthesize a consistent estimate of a causal relation at yet another environment, potentially different from each of the former (Bareinboim and Pearl (2013c)). It is shown that such synthesis may be feasible from multiple sources even in cases where it is not feasible from any one source in isolation.

## ACKNOWLEDGMENT

This paper benefited from discussions with Onyebuchi Arah, Stuart Baker, Sander Greenland, Michael Hoefler, Marshall Joffe, William Shadish, Ian Shrier, and Dylan Small. We are grateful to two anonymous referees for thorough reviews of this manuscript and for suggesting a simplification in the transport formula of Example 10.

## REFERENCES

Adelman, L. (1991). Experiments, quasi-experiments, and case studies: A review of empirical methods for evaluating decision support systems. *Systems, Man and Cybernetics, IEEE Transactions on* **21** 293 –301.

Balke, A. and Pearl, J. (1995). Counterfactuals and policy analysis in structural models. In *Uncertainty in Artificial Intelligence, Proceedings of the Eleventh Conference* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 11–18.

Bareinboim, E., Brito, C. and Pearl, J. (2012). Local characterizations of causal Bayesian networks. *Lecture Notes in Artificial Intelligence* **7205** 1–17.

Bareinboim, E. and Pearl, J. (2012). Transportability of causal effects: Completeness results. In *Proceedings of the Twenty-Sixth National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 698–704.

Bareinboim, E. and Pearl, J. (2013a). Causal transportability with limited experiments. In *Proceedings of the Twenty-Seventh National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 95–101.

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

Bareinboim, E. and Pearl, J. (2013c). Meta-transportability of causal effects: A formal approach. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*. JMLR W&CP 31, 135–143.

Bareinboim, E., Tian, J. and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of The Twenty-Eighth Conference on Artificial Intelligence* (C. E. Brodley and P. Stone, eds.). AAAI Press, Menlo Park, CA, in press.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* **2** 47–53.

Bollen, K. A. and Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of Causal Analysis for Social Research* (S. L. Morgan, ed.). Springer, New York, Chapter 15.

Campbell, D. and Stanley, J. (1963). *Experimental and Quasi-Experimental Designs for Research*. Wadsworth Publishing, Chicago.

Cole, S. and Stuart, E. (2010). Generalizing evidence from randomized clinical trials to target populations. *American Journal of Epidemiology* **172** 107–115.

Davis, J. A. (1984). Extending Rosenberg's technique for standardizing percentage tables. *Social Forces* **62** pp. 679–708.

Dawid, A. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review* **70** 161–189.

Ellenberg, S. and Hamilton, J. (1989). Surrogate endpoints in clinical trials: Cancer. *Statistics in Medicine* 405–413.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*, vol. Analytical methods for social research. Cambridge University Press, New York.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher* **5** pp. 3–8.

Glymour, M. and Greenland, S. (2008). Causal diagrams. In *Modern Epidemiology* (K. Rothman, S. Greenland and T. Lash, eds.), 3rd ed. Lippincott Williams & Wilkins, Philadelphia, PA, 183–209.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

Hayduk, L., Cummings, G., Stratkotter, R., Nimmo, M., Grygoryev, K., Dosman, D., Gillespie, M. and Pazderka-robinson, H. (2003). Pearl's d-separation: One more step into causal thinking. *Structural Equation Modeling* **10** 289–311.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press.

Heise, D. (1975). *Causal Analysis*. John Wiley and Sons, New York.

Hernán, M., Hernández-Díaz, S. and Robins, J. (2004). A structural approach to selection bias. *Epidemiology* **15** 615–625.

Hernán, M. and VanderWeele, T. (2011). Compound treatments and transportability of causal inference. *Epidemiology* **22** 368–377.

Höfler, M., Gloster, A. and Hoyer, J. (2010). Causal effects in psychotherapy: Counterfactuals counteract overgeneralization. *Psychotherapy Research* DOI: 10.1080/10503307.2010.501041.

Huang, Y. and Valtorta, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 217–224.

Joffe, M. and Green, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65** 530–538.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Lane, P. and Nelder, J. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics* **38** 613–621.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation (Springer Texts in Statistics)*. 2nd ed. Springer, New York.

Manski, C. (2007). *Identification for Prediction and Decision*. Harvard University Press, Cambridge, Massachusetts.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. English translation of excerpts (1990) by D. Dabrowska and T. Speed,

in *Statistical Science*, 5:463–472.
Owen, A. B. (2009). Karl pearsons meta-analysis revisited. *Annals of Statistics* **37** 3867–3892.
Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.
Pearl, J. (1993). Graphical models, causality, and intervention. *Statistical Science* **8** 266–273.
Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.
Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.
Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96–146.
Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.
Pearl, J. (2011). The structural theory of causation. In *Causality in the Sciences* (J. W. P. McKay Illari, F. Russo, ed.). Clarendon Press, Oxford, Chapter 33.
Pearl, J. (2012a). The causal foundations of structural equation modeling. In *Handbook of Structural Equation Modeling* (R. H. Hoyle, ed.). Guilford Press, New York.
Pearl, J. (2012b). Some thoughts concerning transfer learning, with applications to meta-analysis and data-sharing estimation. Tech. Rep. R-387, Cognitive Systems Laboratory, Department of Computer Science, UCLA.
Pearl, J. (2012c). Trygve Haavelmo and the emergence of causal calculus. Tech. Rep. R-391, Cognitive Systems Lab, Department of Computer Science, UCLA; To appear: Econometric Theory, special issue on Haavelmo Centennial.
Pearl, J. (2013). Lindear models: A useful "microscope" for causal analysis. *Journal of Causal Inference* **1** 155–170.
Pearl, J. and Bareinboim, E. (2011). Transportability across studies: A formal approach. Tech. Rep. R-372, Cognitive Systems Laboratory, Department of Computer Science, UCLA.
Petersen, M. (2011). Compound treatments, transportability, and the structural causal model: The power and simplicity of causal graphs. *Epidemiology* **22** 378–381.
Prentice, R. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8** 431–440.
Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics* **30** 145–157.
Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling* **7** 1393–1512.
Robins, J., Orellana, L. and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine* **27** 4678–4721.
Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin* **118** 183–192.
Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701.
Shadish, W., Cook, T. and Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. 2nd ed. Houghton-Mifflin, Boston.
Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 437–444.
Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.
Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search*. 2nd ed. MIT Press, Cambridge, MA.
Strotz, R. and Wold, H. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* **28** 417–427.
Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, 567–573.
Verma, T. and Pearl, J. (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*. Mountain View, CA. Also in R. Shachter, T.S. Levitt, and L.N. Kanal (Eds.), *Uncertainty in AI 4*, Elesevier Science Publishers, 69–76, 1990.
Westergaard, H. (1916). Scope and method of statistics. *Publications of the American Statistical Association* **15** 229–276.

WHITE, H. and CHALAK, K. (2009). Settable systems: An extension of Pearl's causal model with optimization, equilibrium and learning. *Journal of the Machine Learning Research* **10** 1759–1799.

WRIGHT, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20** 557–585.

YULE, G. (1934). On some points relating to vital statistics, more especially statistics of occupational mortality. *Journal of the Royal Statistical Society* **97** 1–84.

## APPENDIX 1

Derivation of the transport formula for the causal effect in the model of Fig. 6(d), (Eq. (5.9)),

$$
\begin{aligned}
P^*(y|do(x)) =& P(y|do(x),s) \\
=& \sum_z P(y|do(x),s,z)P(z|do(x),s) \\
=& \sum_z P(y|do(x),z)P(z|do(x),s) \\
& \qquad \text{(2nd condition of thm. 3, } S\text{-admissibility of } Z \text{ of } CE(X,Y)) \\
=& \sum_z P(y|do(x),z) \sum_w P(z|do(x),w,s)P(w|do(x),s) \\
=& \sum_z P(y|do(x),z) \sum_w P(z|w,s)P(w|do(x),s) \\
& \qquad \text{(3rd condition of thm. 3, } (X \perp\!\!\!\perp Z|W,S)_{D_{\overline{X(W)}}}) \\
=& \sum_z P(y|do(x),z) \sum_w P(z|w,s)P(w|do(x)) \\
& \qquad \text{(2nd condition of thm. 3, } S\text{-admissibility of the empty set } \{\} \text{ of } CE(X,W)) \\
=& \sum_z P(y|do(x),z) \sum_w P^*(z|w)P(w|do(x))
\end{aligned}
\tag{6.1}
$$

Derivation of the transport formula for the causal effect in the model of Fig. 7, (Eq. (5.10)).

$$
\begin{aligned}
P^*(y|do(x)) =& P(y|do(x),s,s') = \sum_z P(y|do(x),s,s',z)P(z|do(x),s,s') \\
=& \sum_z P(y|do(x),z)P(z|do(x),s,s') \\
& \qquad \text{(2nd condition of thm. 3, } S\text{-admissibility of } Z \text{ of } CE(X,Z)) \\
=& \sum_z P(y|do(x),z) \sum_w P(z|do(x),s,s',w)P(w|do(x),s,s') \\
=& \sum_z P(y|do(x),z) \sum_w P(z|s,s',w)P(w|do(x),s,s') \\
& \qquad \text{(3rd condition of thm. 3, } (X \perp\!\!\!\perp Z|W,S,S')_{D_{\overline{X(W)}}}) \\
=& \sum_z P(y|do(x),z) \sum_w P(z|s,s',w) \sum_t P(w|do(x),s,s',t)P(t|do(x),s,s') \\
=& \sum_z P(y|do(x),z) \sum_w P(z|s,s',w) \sum_t P(w|do(x),t)P(t|do(x),s,s') \\
& \qquad \text{(2nd condition of thm. 3, } S\text{-admissibility of } T \text{ on } CE(X,W)) \\
=& \sum_z P(y|do(x),z) \sum_w P(z|s,s',w) \sum_t P(w|do(x),t)P(t|s,s') \\
& \qquad \text{(1st condition of thm. 3 / 3rd rule of } do\text{-calculus, } (X \perp\!\!\!\perp T|S,S')_D) \\
=& \sum_z P(y|do(x),z) \sum_w P^*(z|w) \sum_t P(w|do(x),t)P^*(t)
\end{aligned}
\tag{6.2}
$$