2011 JSM Proceedings, Miami Beach FL, July 30-August 4, 2011, pp. 157-171.

TECHNICAL REPORT
R-372-JSM
December 2011

External Validity and Transportability: A Formal Approach

Judea Pearl* Elias Bareinboim†

Abstract

We provide a formal definition of the notion of "transportability," or "external validity," as a license to transfer causal information from experimental studies to a different population in which only observational studies can be conducted. We introduce a formal representation called "selection diagrams" for expressing differences and commonalities between populations of interest and, using this representation, we derive procedures for deciding whether causal effects in the target population can be inferred from experimental findings in a different population. When the answer is affirmative, the procedures identify the set of experimental and observational studies that need be conducted to license the transport. We further discuss how transportability analysis can guide the transfer of knowledge in non-experimental learning to minimize re-measurement cost and improve prediction power.

Key Words: external validity, generalizability, causal inference, surrogate endpoint, meta-analysis.

1. Introduction: Threats vs. Assumptions

Science is about generalization; conclusions that are obtained in a laboratory setting are transported and applied elsewhere, in an environment that differs in many aspects from that of the laboratory.

If the target environment is arbitrary, or drastically different from the study environment nothing can be learned from the latter. However, the fact that most experiments 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 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. The standard literature on this topic, falling under rubrics such as "quasi-experiments," "meta-analysis," and "external validity," consists primarily of "threats," namely, verbal narratives of what can go wrong when we try to transport results from one study to another (e.g., [Shadish et al., 2002, chapter 3; Manski, 2007]). In contrast, we seek to establish "licensing assumptions," namely, formal conditions under which the transport of results across diverse environments is licensed from first principles.

Transportability analysis requires a formal language within which the notion of "environment" or "population" is given precise characterization, and differences among populations can be encoded and analyzed. The advent of causal diagrams [Pearl, 1995, 2009; Spirtes et al., 2000] provides such a language and renders the formalization of transportability possible. Using this language, this paper offers a precise definition for the notion of transportability and establishes formal conditions that, if held true, would permit us to transport results across studies, domains, environments, or populations.

*Cognitive Systems Laboratory, Department of Computer Science, University of California, Los Angeles, Los Angeles, CA. 90095

†Cognitive Systems Laboratory, Department of Computer Science, University of California, Los Angeles, Los Angeles, CA. 90095

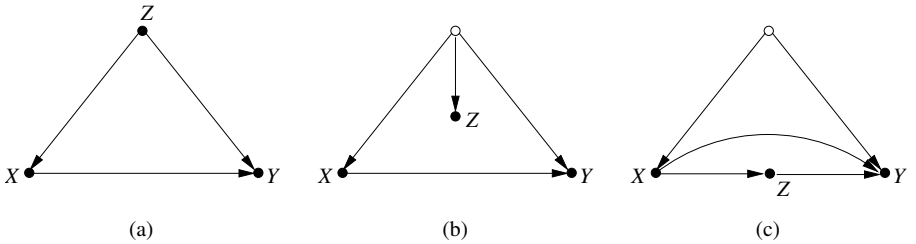

Figure 1: Causal diagrams depicting Examples 1–3. In (a) Z represents "age." In (b) Z represents "linguistic skills" while age (hollow circle) is unmeasured. In (c) Z represents a biological marker situated between the treatment (X) and a disease (Y).

2. Motivating Examples

To motivate our discussion and to demonstrate some of the subtle questions that transportability entails, we will consider three simple examples, graphically depicted in Fig. 1. The examples invoke the familiar domain of clinical trials, yet the issues raised pertain to any learning environment that can be characterized by the structure of the data-generating model. For example, a robot trained by a simulator should be able to transport causal knowledge acquired in training to challenges of a new environment, in which experiments are costly or infeasible.

Example 1 *We conduct a randomized trial in Los Angeles (LA) and estimate the causal effect of treatment X on outcome Y for every age group $Z = z$ as depicted in Fig. 1(a). We now wish to generalize the results to the population of New York City (NYC), but we find that the distribution $P(x, y, z)$ in LA is different from the one in NYC (call the latter $P^*(x, y, z)$). In particular, 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 $P^*(y|do(x))$.*[1]

If we can assume that age-specific effects $P(y|do(x), Z = z)$ are invariant across cities , the overall causal effect in NYC should be

$$P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z) \tag{1}$$

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.

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. (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 X-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 skills, which correlates with age (not measured) (see Fig. 1(b)). Given the observed disparity $P(z) \neq P^*(z)$, how are we to estimate the causal effect $P^*(y|do(x))$ in NYC from the z-specific causal effect $P(y|do(x), z)$ estimated in LA?*

[1]The $do(x)$ notation [Pearl, 1995, 2009] interprets $P(y|do(x))$ as the probability of outcomes $Y = y$ in a randomized experiment where the treatment variables X take on values $X = x$. $P(y|do(x), z)$ is logically equivalent to $P(Y_x = y|Z_x = z)$ in counterfactual notation. Likewise, the diagrams used in this paper should be interpreted as parsimonious encoding of functional relations [Pearl, 2009, p. 101], where every bi-directed arc $X \dashleftarrow\dashrightarrow Y$ stands for a set of latent variables affecting X and Y.

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 should be

$$P^*(y|do(x)) = P(y|do(x)) \tag{2}$$

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. (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 transport formula depends on the causal context in which distributional 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 Fig. 1(b) for guidance, the Z-specific effect of X on Y in NYC is given by:

$$P^*(y|do(x), z) \;\; = \;\; \sum_{age} P(y|do(x), age)P^*(age|z) \tag{3}$$

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.

Example 3 *Examine the case where Z is a X-dependent variable, say a disease biomarker as shown in Fig. 1(c). Assume further that the disparity $P(z) \neq P^*(z)$ is discovered in each level of X 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 NYC?*

Here, Eq. (1) is wrong for two reasons. First, as in the case of age-proxy, it matters whether the disparity in $P(z)$ represents differences in susceptibility to X or differences in propensity to receiving X. In the latter case, Eq. (2) would be valid, while in the former, more information is needed. Second, the overall causal effect is no longer a simple average of the z-specific causal effects but is given by

$$P^*(y|do(x)) \;\; = \;\; \sum_{z} P^*(y|do(x), z)P^*(z|do(x)) \tag{4}$$

which reduces to (1) only in the special case where Z is unaffected by X, as is the case in Fig. 1(a). We shall see (Theorem 3 below) that the correct transport formula is

$$P^*(y|do(x)) \;\; = \;\; \sum_{z} P(y|do(x), z)P^*(z|x) \tag{5}$$

which calls for weighting the z-specific effects by $P^*(z|x)$, to be estimated in the target environment.

3. Formalizing Transportability

3.1 Selection Diagrams and Selection Variables

The examples above demonstrate that transportability is a causal, not statistical notion, requiring knowledge of the mechanisms, or processes, through which differences come about. To witness, every probability distribution $P(x, y, z)$ that is compatible with the

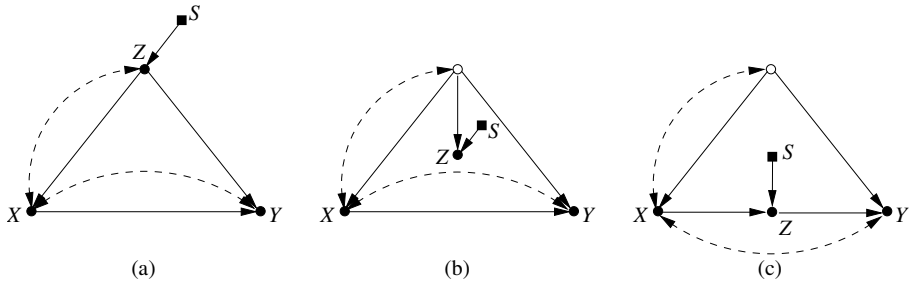

Figure 2: Selection diagrams depicting Examples 1–3. In (a) the two populations differ in age distributions. In (b) the populations differs in how Z depends on age (an unmeasured variable, represented by the hollow circle) and the age distributions are the same. In (c) the populations differ in how Z depends on X.

process of Fig. 1(b) is also compatible with that of Fig. 1(a) and, yet, the two processes dictate different transport formulas. Thus, to represent formally the differences between domains we must resort to a representation in which the causal mechanisms are explicitly encoded and in which domain differences 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 domains differ, and switching between the two domains will be represented by conditioning on different values of these S variables.

Formally, 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 Π^* instead of Π. We now attribute the difference between the two to the action of a set S of selection variables, and write[2]

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

Of equal importance is the absence of an S variable pointing to Y in Fig. 2(a), which encodes the assumption that age-specific effects are invariant across the two populations.

The variables in S represent exogenous conditions that determine the values of the variables to which they point.[3]

For example, the age disparity $P(z) \neq P^*(z)$ discussed in Example 1 will be represented by the inequality $P(z) \neq P(z|s)$ where S stands for all factors responsible for drawing subjects at age $Z = z$ to NYC rather than LA.

This graphical representation, which we will call "selection diagrams" can also represent structural differences between the two domains. For example, if the causal diagram of the study population contains an arrow between X and Y, and the one for the target population contains no such arrow, the selection diagram will be $X \rightarrow Y \leftarrow S$ where the role of variable S is to disable the arrow $X \rightarrow Y$ when $S = s^*$ (i.e., $P(y|x, s^*) = P(y|x', s^*)$ for all x') and reinstate it when $S = s$.[4] Our analysis will apply therefore to all factors by which domains may differ or that may "threaten" the transport of conclusions between domains, studies, populations, locations or environments.

[2]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.

[3]Elsewhere, we analyze S variables representing selection of units into the study pool [Bareinboim and Pearl, 2011]; there, the arrows will be pointing towards S.

[4]Pearl [1995; 2009, p. 71] 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.

For clarity, we will represent the S variables by squares, as in Fig. 2, which uses selection diagrams to encode the three examples discussed above. In particular, Fig. 2(a) and 2(b) represent, respectively, two different mechanisms responsible for the observed disparity $P(z) \neq P^*(z)$. The first (Fig. 2(a)) dictates transport formula (1) while the second (Fig. 2(b)) calls for direct, unadjusted transport (2).

In the extreme case, we could add selection nodes to all variables, which means that we have no reason to believe that the two domains share any mechanism in common, and this, of course would inhibit any exchange of conclusions between the two. Conversely, absence of a selection node pointing to a variable, say Z, represents an assumption of invariance: the local mechanism that assigns values to Z is the same in both domains.

3.2 Transportability: Definitions and Examples

Using selection diagrams as the basic representational language, and harnessing the concepts of intervention, do-calculus[5] and identifiability [Pearl, 2009, p. 77] we give the notion of transportability a formal definition.

Definition 1 *(Transportability)*
Given two domains, denoted Π *and* Π^**, characterized by probability distributions* P *and* P^**, and causal diagrams* G *and* G^**, respectively, a causal relation* R *is said to be* transportable *from* Π *to* Π^* *if* $R(\Pi)$ *is estimable from the set* I *of interventions on* Π*, and* $R(\Pi^*)$ *is identified from* P, P^*, I, G*, and* G^*.

Definition 1 provides a declarative characterization of transportability which, in theory, requires one to demonstrate the non-existence of two competing models, agreeing on $\{P, P^*, I, G, G^*\}$, and disagreeing on $R(\Pi^*)$. Such demonstrations are extremely cumbersome for reasonably sized models, and we seek therefore procedural criteria which, given the pair (G, G^*) will decide the transportability of any given relation directly from the structures of G and G^*. Such criteria will be developed in the sequel by breaking down a complex relation R into more elementary relations whose transportability can immediately be recognized. We will formalize the structure of this procedure in Theorem 1, followed by Definitions 2 and 3 below, which will identify two special cases where transportability is immediately recognizable.

Theorem 1 *Let* D *be the selection diagram characterizing* Π *and* Π^**, and* S *a set of selection variables in* D*. The relation* $R = P(y|do(x), z)$ *is transportable from* Π *to* Π^* *if and only 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:

(if part): 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 formers can be written as P^* terms and are estimable, therefore, from observations on Π^*, as required by Definition 1. All other terms, especially those involving do-operators, do not contain S; they are experimentally identifiable therefore in Π.

(only if part): If R is transportable, its transport formula T must satisfy the condition of Theorem 1, and that means that $R = T$ is a valid equality in do-calculus. Moreover, since do-calculus is complete, [Shpitser and Pearl, 2006] every valid equality can be obtained by a finite application of the three rules of the calculus. This proves the Theorem. □

[5] The three rules of do-calculus are defined in Appendix 1 and illustrated in graphical details in [Pearl, 2009, p. 87].

Definition 2 *(Direct Transportability)*
A causal relation R is said to be directly transportable *from Π to Π^*, if $R(\Pi^*) = R(\Pi)$.*

The equality $R(\Pi^*) = R(\Pi)$ means that R retains its validity without adjustment, as in Eq. (2). A graphical test for direct transportability of $P(y|do(x))$ follows immediately from do-calculus and reads: $(S \perp\!\!\!\perp Y|X)_{G_{\overline{X}}}$; i.e., X blocks all paths from S to Y once we remove all arrows pointing to X. Indeed, such condition would allow us to eliminate s from the do-expression, and write:

$$R(\Pi^*) = P(y|do(x), s) = P(y|do(x)) = R(\Pi)$$

Example 4 *Figure 4(a) represents a simple example of direct transportability. Indeed, since S merely changes the mechanism by which the value $X = x$ is selected (sometimes called "treatment assignment mechanism"), it does not change any causal effect of X [Pearl, 2009, pp. 72–73].*

Definition 3 *(Trivial Transportability)*
A causal relation R is said to be trivially transportable *from Π to Π^*, if $R(\Pi^*)$ is identifiable from (G^*, P^*).*

This criterion amounts to ordinary (nonparametric) identifiability of causal relations using graphs [Pearl, 2009, p. 77]. It permits us to estimate $R(\Pi^*)$ directly from passive observations on Π^*, un-aided by causal information from Π.

Example 5 *Let R be the causal effect $P(y|do(x))$ and let the selection diagram be $X \rightarrow Y \leftarrow S$, then R is trivially transportable, since $R(\Pi^*) = P^*(y|x)$.*

Example 6 *Let R be the causal effect $P(y|do(x))$ and let the selection diagram of Π and Π^* be $X \rightarrow Y \leftarrow S$, with X and Y confounded as in Fig. 4(b), then R is not transportable, because $P^*(y|do(x)) = P(y|do(x), s)$ cannot be decomposed into s-free or do-free expressions using do-calculus. This is the smallest graph for which the causal effect is non-transportable.*

4. 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 domains, and what the transport formula should be.

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

Proof:

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

From Rule-1 of do-calculus [Pearl, 2009, p. 85] we have: $P(y|do(x), z, s) = P(y|do(x), z)$ whenever Z satisfies $(Y \perp\!\!\!\perp S|Z)$ in $D_{\overline{X}}$. This proves Theorem 2. □

Definition 4 *(S-admissibility)*
A set T of variables satisfying $(Y \perp\!\!\!\perp S|T)$ in $D_{\overline{X}}$ will be called S-admissible.

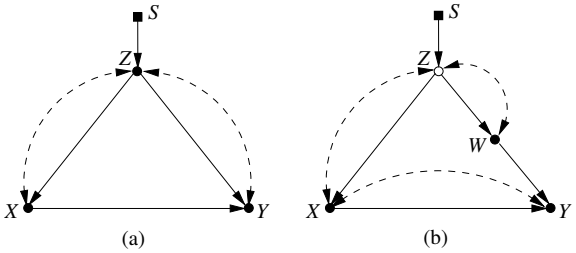

Figure 3: Selection diagrams illustrating S-admissibility. (a) has no S-admissible set while in (b), W is S-admissible.

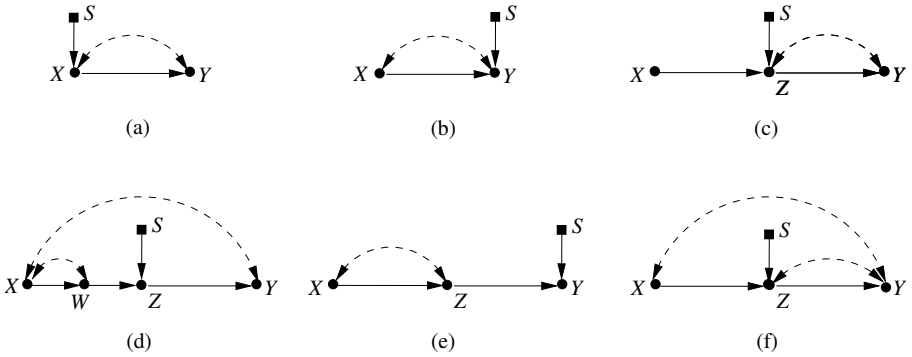

Figure 4: Selection diagrams illustrating transportability. The causal effect $P(y|do(x))$ is (trivially) transportable in (c) but not in (b) and (f). It is transportable in (a), (d), and (e) (see Corollary 2 and Example 9).

Corollary 1 *The average causal effect $P(y|do(x))$ is transportable from Π to Π^* 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. (1).*

Proof:

$$\begin{aligned} P^*(y|do(x)) &= P(y|do(x),s) \\ &= \sum_z P(y|do(x),z,s)P(z|do(x),s) \\ &= \sum_z P(y|do(x),z)P(z|s) \\ &\qquad \text{(using } S\text{-admissibility and} \\ &\qquad \text{(Rule-3 of } do\text{-calculus)} \\ &= \sum_z P(y|do(x),z)P^*(z) \end{aligned} \tag{6}$$

□

Example 7 *The causal effect is transportable in Fig. 2(a), since Z is S-admissible, and directly transportable in Fig. 2(b) and 4(a), where the empty set is S-admissible. It is also transportable in Fig. 3(b), where W is S-admissible, but not in Fig. 3(a) where no S-admissible set exists.*

Contrasting the diagrams in Figs. 2(a) and 3(a), we witness again the crucial role of causal knowledge in facilitating transportability. These two diagrams are statistically indistinguishable, yet the former is transportable, while the latter is not.

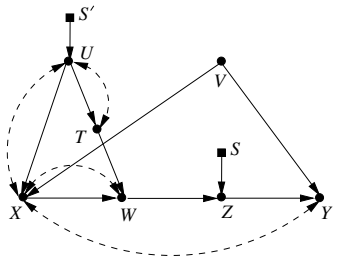

Figure 5: Selection diagram in which the causal effect is shown to be transportable in two iterations of Theorem 3.

Corollary 2 *Any S variable that is pointing directly into X as in Fig. 4(a), or that is d-connected to Y only through X can be ignored.*

Proof: 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 X-dependent Z variables, as in Fig. 2(c).

Theorem 3 *The causal effect $P(y|do(x))$ is transportable from Π to Π^* if any 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 $(X \perp\!\!\!\perp Y|W,S)_D$ and for which $P(w|do(x))$ is transportable.*

Proof:

1. *Condition (1) entails transportability.*
2. *If condition (2) holds, it implies*

$$P^*(y|do(x)) = P(y|do(x),s) \tag{7}$$
$$= \sum_z P(y|do(x),z,s)P(z|do(x),s) \tag{8}$$
$$= \sum_z P(y|do(x),z)P^*(z|do(x)) \tag{9}$$

We now note that the transportability of $P(z|do(x))$ should reduce $P^*(z|do(x))$ to a combination of do-free and star-free expressions, thus rendering $P(y|do(x))$ transportable.

3. *If condition (3) holds, it implies*

$$P^*(y|do(x)) = P(y|do(x),s) \tag{10}$$
$$= \sum_w P(y|do(x),w,s)P(w|do(x),s) \tag{11}$$
$$= \sum_w P(y|w,s)P^*(w|do(x)) \tag{12}$$

(by Rule-3 of do-calculus)

$$= \sum_w P^*(y|w)P^*(w|do(x)) \tag{13}$$

Again, the transportability of $P(w|do(x))$ should reduce $P^*(w|do(x))$ to a combination of do-free and star-free expressions, thus rendering $P(y|do(x))$ transportable. This proves Theorem 3. □

Example 8 *Applying Theorem 3 to Fig. 2(c), we conclude that $R = P(y|do(x))$ is trivially transportable, for it is identifiable in Π^* through the front-door criterion [Pearl, 2009]. R is likewise (trivially) transportable in Fig. 4(c) (by the back-door criterion). R is not transportable however in Fig. 3(a), where no S-admissible set exists.*

Example 9 *Fig. 4(d) requires that we invoke both conditions of Theorem 3, iteratively, and yields transport formula (derived in Appendix 2):*

$$P^*(y|do(x)) = \sum_z P(y|do(x), z) \sum_w P(w|do(x)) P^*(z|w) \tag{14}$$

The first two factors on the right are estimable in the experimental domain, and the third in the observational domain. Surprisingly, the joint effect $P(y, w, z|do(x))$ need not be estimated in the experiment; a decomposition that results in improved estimation power.

A similar analysis applies to Fig. 4(e). The model of Fig. 4(f) however does not allow for the transportability of $P(y|do(x))$ because there is no S-admissible set in the diagram and condition 3 of Theorem 3 cannot be invoked.

Example 10 *Fig. 5 represents a more challenging selection diagram, which requires several iterations to discern transportability [Pearl and Bareinboim, 2011b], and yields:*

$$P^*(y|do(x)) = \sum_z P(y|do(x), z) \sum_w P^*(z|w) \sum_t P(w|do(x), t) P^*(t) \tag{15}$$

The main power of this formula is to guide the learning agent in deciding what measurements need be taken in each domain. It asserts, for example, that variables U and V need not be measured, that the W-specific causal effects need not be learned in the experiment and only the conditional probabilities $P^*(z|w)$ and $P^*(t)$ need be learned in the target domain.

5. Transportability Across Observational Domains

Our analysis thus far assumed that transport is needed from experimental learning because R, the relation of interest, is causal and cannot be identified solely from passive observations in the target domain. In this section we demonstrate that transporting purely observational findings can be beneficial as well, albeit for different reasons.

Assume we conduct an elaborate observational study in LA, involving dozens of variables and thousands of samples, aiming to learn some statistical relation, $R(P)$ (say a conditional distribution for prediction or classification). We now wish to estimate the same relation $R(P^*)$ in NYC. The question arises whether it is necessary to repeat the study from scratch or, in case the disparity between the two domains is localized, if we can leverage what we learned in LA, supplement it with a less elaborate study in NYC and combine the results to yield an informed estimate of $R(P^*)$.

In complex models, the savings gained by focusing on only a small subset of variables in P^* can be enormous, because any reduction in the number of measured variables translates into substantial reduction in the number of samples needed to achieve a given level of prediction accuracy. This is especially true in non-parametric models, where estimation efficiency deteriorates significantly with the number of variables involved.

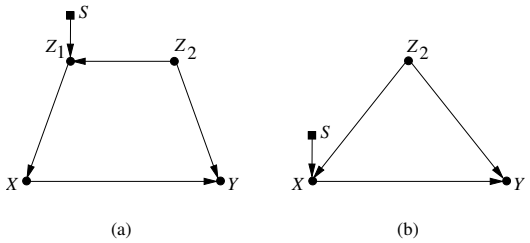

Figure 6: (a) Selection diagram in which relasuring X or Y, or, alternatively, by measuring only X and Z_2. (b) The diagram resulting from marginalizing over Z_1.

An examination of the transport formulas derived in this paper (e.g., Eqs. (14) or (15)) reveals that the methods developed for transporting causal relations are applicable to statistical relations as well, albeit with some modification. Consider Eq. (14) and its associated diagram in Fig. 4(d). If the target relation $R = P^*(y|do(x))$ was expressed, not in terms of the $do(x)$ operator, but as a conditional probability $R(P^*) = \sum_c P^*(y|x,c)P^*(c)$ where C is a sufficient set of covariates, the right hand side of (14) reveals that $P^*(z|w)$ is the only relation that need to be re-estimated at the target domain; all the other terms in that expression are estimable at the source environment, using C, X, Z, W and Y.

These considerations motivate a slightly different definition of transportability, tailored to non-experimental learning, which emphasizes narrowing the scope of observations rather than identification per se.

Definition 5 *(Observational Transportability)*
Given two domains, Π *and* Π^**, characterized by probability distributions* P *and* P^**, and causal diagrams* G *and* G^**, respectively, a statistical relation* $R(P)$ *is said to be* observationally transportable *from* Π *to* Π^* *over* V^* *if* $R(P^*)$ *is identified from* P*,* $P^*(V^*)$*,* G*, and* G^**. where* $P^*(V^*)$ *is the marginal distribution of* P^* *over a subset of variables* V^**.*

This definition requires that the relation transferred be reconstructed from data obtained in the old learning domain, plus observations conducted on a subset V^* of variables in the new domain. In the example above, $R(P)$ was shown to be observationally transportable over $V^* = \{Z, W\}$, while in the example of Fig. 5, we have $V^* = \{Z, W, T\}$ (from Eq. (15)).

Note that, despite the purely statistical nature of the task, a causal selection diagram is still needed to identify the mechanisms by which the two domains differ. The probabilities P and P^*, being descriptive, cannot convey information about the locality of the mechanism that accounts for their differences. In Fig. 5, for example, changes in S' will propagate to the entire probability $P(t, u, x, w, y)$ and could not be distinguished from changes in an S-node that points, say, at W or at Y. Moreover P and P^* can be deceptively identical, and hide profound differences in mechanisms. A typical example is the structural differences between two statistically-indistinguishable models.

While selection diagrams are still an essential tool for formulating differences among domains, the mechanics of solving observational transportability problems is somewhat different. Since the transported relations are not cast in do-expressions, the do-calculus is no longer helpful, and we must rely mainly on the conditional independencies encoded in the selection diagram.

An example of this mechanics will be illustrated through the selection diagram of Fig. 6. Assume that, after learning $P(x, y, z_1, z_2)$ in the source environment, one is interested in classifying X from observations on Y in the target environment. If hand labeling data in the new domain is a costly enterprise, we may ask whether we can learn the desired classifier $P^*(x|y)$ without taking any measurement of X or Y in the new domain. Formally, this

amounts to asking whether $P(x|y)$ is transportable over $V^* = \{Z_1, Z_2\}$. The answer is of course positive, since, given the selection diagram of Fig. 6, the conditional probability $P(z_1|z_2)$ is the only factor that changes in the Markovian factorization of P. Therefore, we can simply re-learn $P^*(z_1|z_2)$ and compute our target relation $P^*(x|y)$ from the factorization $P^*(x, y, z_1, z_2) = P(y|z_2, x)P(x|z_1)P(z_2)P^*(z_1|z_2)$, with all but the last factor transportable from the source environment.

We see that the conditional independencies embedded in the diagram have the capacity to narrow the scope V^* of variables that need be measured, so as to minimize measurement cost and sample variability. For example, if Z_1 is multi-dimensional or more costly to measure than X, $R = P(x|y)$ can be transported over $V^* = \{Z_2, X\}$. This can be seen by ignoring (or marginalizing over) Z_1, which yields the diagram of Fig. 6(b).

Remarkably, the transport of certain relations across disparate domains can sometimes be accomplished with no re-measurements whatsoever, thus exhibiting "direct transportability" (Definition 2). The selection diagram of Fig. 6(a), for example, permits the transport of the relation

$$R' = \sum_{z_1} P(y|x, z_1)P(z_1) \tag{16}$$

over the null set $V^* = \{\}$. This becomes evident from the fact that Z_2 can replace Z_1 in R' [Pearl and Paz, 2010] and, using the independencies $(S \perp\!\!\!\perp Y|X, Z_2)$ and $(S \perp\!\!\!\perp Z_2)$ shown in the diagram, the transported relation becomes s-free:

$$R'(P^*) = \sum_{z_2} P(y|x, z_2)P(z_2) \tag{17}$$

While a systematic analysis of observational transportability is beyond the scope of this paper, Definition 5 offers a formal characterization of this ubiquitous class of information transfer problems, and identifies the basic elements needed for their solution.

Conclusions

Given judgmental assessments 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 experiments conducted elsewhere. When such inference is possible, the criteria provided by Theorems 1-3 yield transport formulae, namely, principled ways of recalibrating the learned relations so as to account for differences in the populations. These formulae enable the learner to select the essential measurements in both the experimental and observational studies, and thus minimize measurement costs and sample variability.

Extending these results to observational studies, we showed that there is also benefit in transporting statistical findings from one population to another in that it enables learners to avoid repeated measurements that are not absolutely necessary for reconstructing the relation transferred. Procedures for deciding whether such reconstruction is feasible when certain re-measurements are forbidden or too costly were shown capable of substantial savings in sample size or increase in estimation power.

A second extension of transportability analysis led to a causally principled definition of "surrogate endpoint," namely, a variable Z such that knowing the effect of treatment on Z allows predictions of the effect of X on the more clinically relevant outcome Y [Joffe and Green, 2009]. [Pearl and Bareinboim, 2011a] have argued that a surrogate should serve not merely as a good predictor of outcomes, but also as a $robust$ predictor of effects in the face of changing external conditions. Therefore, any formal definition of

surrogacy must make change in conditions an integral part of the definition.[6] Accordingly, Z is defined as a surrogate of Y if observation of Z in Π^* enables the effect of X on Y to be transported from Π to Π^* without re-measurement of Y and regardless of the mechanism responsible for variations in Z. The procedure developed from this transportability-based definition allows for the identification of valid surrogates in a complex set of causal relations.

Finally, a natural extension of these results provide a principled model-based method of conducting meta-analysis Here, the target of analysis is not to minimize the scope V* at the target population but, rather, to share samples from diverse studies so as to minimize sampling variation [Pearl, 2011b].

Consider the selection diagrams depicted in Fig. 7, each representing a study conducted

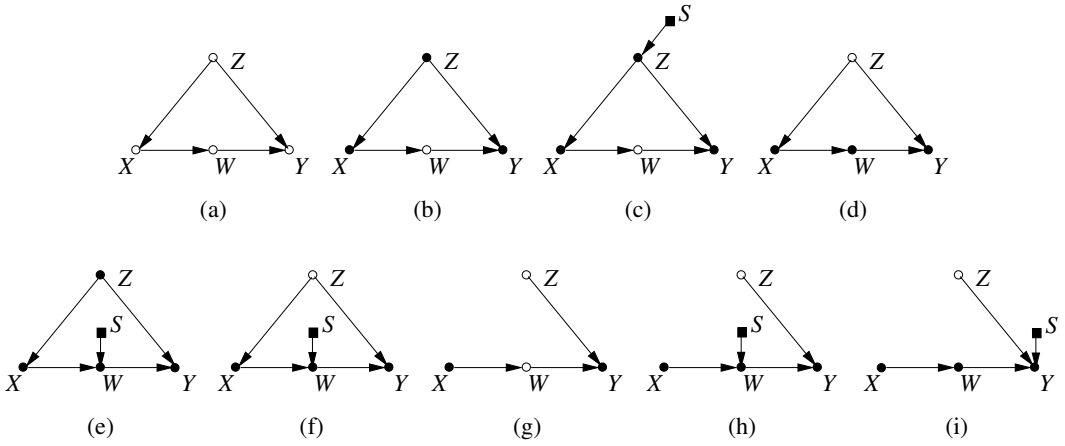

Figure 7: Diagrams representing 8 studies ((b)–(i)) conducted under different conditions on different populations, aiming to estimate the causal effect of X on Y in the target population, shown in 7(a).

on a different population and under a different set of conditions. Here, solid circles represent variables that were measured in the respective study and hollow circles variables that remained unmeasured. An arrow $\blacksquare\!\!\rightarrow$ represents an external influence affecting a mechanism by which the study population is assumed to differ from the target population Π^*, shown in Fig. 7(a). For example, Fig. 1(c) represents an observational study on population Π_c in which variables X, Z and Y were measured, W was not measured and the prior probability $P_c(Z)$ differs from that of the target population $P^*(Z)$. Diagrams (b)–(f) represent observational studies while (g)–(j) stand for experimental studies with X randomized (hence the missing arrows into X).

Despite differences in populations, measurements and conditions, each of the studies may provide information that bears on the target relation $R(\Pi^*)$ which, in this example, we take to be the causal effect of X on Y, $P^*(y|do(x))$ or; given the structure of Fig. 7(a),

$$R(\Pi^*) = P^*(y|do(x)) = \sum_z P^*(y|x,z)P^*(z).$$

While $R(\Pi^*)$ can be estimated directly from some of the studies (e.g., (g)) and indirectly from others (e.g., (b) and (d)), it cannot be estimated separately from any study for which the population differs substantially from Π^* (e.g., (c), (e), (f)). The estimates of R provided

[6]Traditional definitions of surrogacy [Prentice, 1989] as well as those based on "principal strata" [Frangakis and Rubin, 2002] lack this feature and are, therefore, problematic [Pearl, 2011a].

by the former studies may differ from each other due to sampling variations and measurement errors, and can be aggregated in the standard tradition of meta-analysis. The latter studies, however, should not be averaged with the former, since they do not provide unbiased estimates of R. They are not totally useless, though, for they can provide information that renders the former estimates more precise. For example, although we cannot identify R from study 7(c), since $P_c(z)$ differs from $P^*(z)$, we can nevertheless use the estimates of $P_c(x|z)$, $P_c(y|z,x)$ that 7(c) provides to improve the accuracy of $P^*(x|z)$ and $P^*(y|z,x)$[7] which may be needed for estimating R by indirect methods. For example, $P^*(y|z,x)$ is needed in study 7(b) if we use the estimator $R = \sum_z P^*(y|x,z)P^*(z)$, while $P^*(x|z)$ is needed if we use the inverse probability estimator $R = \sum_z P^*(x,y,z)/P^*(x|z)$.

Similarly, consider the randomized studies depicted in 7(h) and 7(i). None is sufficient for identifying R in isolation, yet taken together, they permit us to borrow $P_i(w|do(x))$ from 7(i) and $P_h(y|w,do(x))$ from 7(h) (likewise, from 7(d)) and synthesize a bias-free estimator:

$$\begin{aligned} R &= \sum_w P^*(y|w,do(x))P^*(w|do(x)) \\ &= \sum_w P_h(y|w,do(x)P_i(w|do(x)) \end{aligned}$$

The challenge of synthetic meta-analysis is to take a collection of studies, annotated with their respective selection diagrams (as in Fig. 7), and construct an estimator of a specified relation $R(\Pi^*)$ that makes maximum use of the samples available, by exploiting the commonalities among the populations studied and the target population Π^*.

Our analysis is based on the assumption that the investigator is in possession of sufficient knowledge to determine, at least qualitatively, where two populations may differ. In practice, such knowledge may only be partially available and, as is the case in every mathematical exercise, the benefit of the analysis lies primarily in understanding what knowledge is needed for the task to succeed and how sensitive conclusions are to knowledge that we do not possess.

References

[Bareinboim and Pearl, 2011] Bareinboim, E. and Pearl, J. (2011). Controlling selection bias in causal inference. Technical Report R-381, <http://ftp.cs.ucla.edu/pub/stat_ser/r381.pdf>, Department of Computer Science, University of California, Los Angeles.

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

[Frangakis and Rubin, 2002] Frangakis, C. and Rubin, D. (2002). Principal stratification in causal inference. *Biometrics*, 1(58):21–29.

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

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

[7] The absence of boxed arrows into X and Y in Fig. 7(c) implies the equalities

$$P_c(x|z) = P^*(x|z) \text{ and } P_c(y|z,x) = P^*(y|z,x).$$

[Pearl, 1995] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–710.

[Pearl, 2009] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition.

[Pearl, 2011a] Pearl, J. (2011a). Principal stratification - a goal or a tool? *The International Journal of Biostatistics*, 7.

[Pearl, 2011b] Pearl, J. (2011b). Some thoughts concerning transfer learning, with applications to meta-analysis and data-sharing estimation. Technical Report R-387, <http://ftp.cs.ucla.edu/pub/stat_ser/r387.pdf>, Department of Computer Science, University of California, Los Angeles.

[Pearl and Bareinboim, 2011a] Pearl, J. and Bareinboim, E. (2011a). Transportability across studies: A formal approach. Technical Report R-372, <http://ftp.cs.ucla.edu/pub/stat_ser/r372.pdf>, Department of Computer Science, University of California, Los Angeles.

[Pearl and Bareinboim, 2011b] Pearl, J. and Bareinboim, E. (2011b). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pages 247–254. AAAI Press, Menlo Park, CA.

[Pearl and Paz, 2010] Pearl, J. and Paz, A. (2010). Confounding equivalence in causal equivalence. In Grünwald, P. and Spirtes, P., editors, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 433–441. AUAI, Corvallis, OR.

[Prentice, 1989] Prentice, R. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8:431–440.

[Shadish et al., 2002] Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston, second edition.

[Shpitser and Pearl, 2006] Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In Dechter, R. and Richardson, T., editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444. AUAI Press, Corvallis, OR.

[Spirtes et al., 2000] Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition.

Appendix 1

The do-calculus [Pearl, 1995] consists of three rules that permit us to transform expressions involving do-operators into other expressions of this type, whenever certain conditions hold in the causal diagram G. (See footnote 1 for semantics.)

We consider a DAG G in which each child-parent family represents a deterministic function $x_i = f_i(pa_i, \epsilon_i), i = 1, \ldots, n$, where pa_i are the parents of variables X_i in G; and $\epsilon_i, i = 1, \ldots, n$ are arbitrarily distributed random disturbances, representing background factors that the investigator chooses not to include in the analysis.

Let X, Y, and Z be arbitrary disjoint sets of nodes in a causal DAG G. An expression of the type $E = P(y|do(x), z)$ is said to be compatible with G if the interventional distribution described by E can be generated by parameterizing the graph with a set of functions f_i and a set of distributions of $\epsilon_i, i = 1, \ldots, n$

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):

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

Rule 2 (Action/observation exchange):

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

Rule 3 (Insertion/deletion of actions):

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

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

The *do*-calculus was proven to be complete [Shpitser and Pearl, 2006], in the sense that if an equality cannot be established by repeated application of these three rules, it is not valid.

Appendix 2

Derivation of the transport formula for the causal effect in the model of Fig. 4(d) (Eq. (14)),

$$\begin{aligned}
&P^*(y|do(x)) = P(y|do(x), s) \\
&= \textstyle\sum_z P(y|do(x), s, z)P(z|do(x), s) \\
&= \textstyle\sum_z P(y|do(x), z)P(z|do(x), s) \\
&\big(\text{2nd cond. of thm. 2, } S\text{-admissibility of } Z \text{ for } CE(X, Y)\big) \\
&= \textstyle\sum_z P(y|do(x), z) \sum_w P(z|do(x), w, s)P(w|do(x), s) \\
&= \textstyle\sum_z P(y|do(x), z) \sum_w P(z|w, s)P(w|do(x), s) \\
&\big(\text{3rd cond. of thm. 2, } (X \perp\!\!\!\perp Z|S, W)\big) \\
&= \textstyle\sum_z P(y|do(x), z) \sum_w P(z|w, s)P(w|do(x)) \\
&\big(\text{2nd cond. of thm. 2, } S\text{-admissibility of } \{\} \text{ for } CE(X, W)\big) \\
&= \textstyle\sum_z P(y|do(x), z) \sum_w P^*(z|w)P(w|do(x)) \qquad (18)
\end{aligned}$$