Research Article

Thomas S. Richardson* and James M. Robins

Potential outcome and decision theoretic foundations for statistical causality

https://doi.org/10.1515/jci-2022-0012
received February 17, 2022; accepted February 10, 2023

Abstract: In a recent work published in this journal, Philip Dawid has described a graphical causal model based on decision diagrams. This article describes how single-world intervention graphs (SWIGs) relate to these diagrams. In this way, a correspondence is established between Dawid's approach and those based on potential outcomes such as Robins' finest fully randomized causally interpreted structured tree graphs. In more detail, a reformulation of Dawid's theory is given that is essentially equivalent to his proposal and isomorphic to SWIGs.

Keywords: directed acyclic graph, decision theory, finest fully randomized causally interpreted structured tree graph, potential outcome, single-world intervention graph

MSC 2020: 62A01, 62D20, 62H22

1 Introduction

In his recent article, *Decision Theoretic Foundations for Causality*, Philip Dawid elaborates on an earlier theory that he advanced previously [1]. We welcome Dawid's efforts to build a foundation for causal models that aims to develop a graphical framework, while placing an emphasis on making assumptions that are both transparent and testable. Similar concerns have also motivated much of our previous work on potential outcome models represented in terms of finest fully randomized causally interpreted structured tree graphs (FFRCISTGs) [2] and single-world intervention graphs (SWIGs) [3].

Indeed, like Dawid, we have argued that, in contrast, the assumption of independent errors that is typically adopted by users of Pearl's non-parametric structural equations (also called structural causal models) is untestable and also imposes (superexponentially) many assumptions that are unnecessary for most purposes; furthermore, the independent error assumptions allow the identification of causal quantities that cannot be identified via any randomized experiment on the observed variables [4]. Thus, this assumption contradicts the dictum "no causation without manipulation" and severs the connection between experimentation and causal inference that has been central to much of the conceptual progress during the last century. We also note that [5] cites the move to specifying causal models using potential outcomes rather than error terms as underpinning the "credibility revolution" in Econometrics.

In our view, Dawid's updated theory represents a marked advance on his earlier proposal in that it requires stronger ontological commitments, specifically, the existence of an "intent-to-treat" (ITT) variable, before a model may be called causal. ITT variables are necessary and important in order to encode the notion of ignorability and the effect of treatment on the treated.

* **Corresponding author: Thomas S. Richardson,** Department of Statistics, University of Washington, Seattle, Washington, United States, e-mail: thomasr@uw.edu
James M. Robins: Department of Epidemiology, Harvard School of Public Health, Boston, United States

In addition, as noted by Dawid, the ITT variables make it possible to connect his approach to that based on potential outcomes[1] and SWIGs. The connection between the two approaches may help to illuminate the strengths and weakness of each formalism. We also present a reformulation of Dawid's theory that is essentially equivalent to his proposal and isomorphic to SWIGs.

We thank Philip Dawid for helpful feedback on our article; in particular, for pointing out a significant omission regarding our proposed definition of distributional consistency for SWIGs. We also thank him for his patience regarding the completion of this manuscript.

2 Relating observational and experimental worlds

At a high level, every approach to causal inference relates a model describing a factual passively observed world and models describing hypothetical "interventional" worlds in which a treatment (or exposure) variable takes on a specific value.

In both the current and previous decision-theoretic conceptions advocated by Dawid, these worlds "exist" at least hypothetically, as different distributions. The relation is then created by the assertion of equalities linking different parts of these distributions. In Dawid's formalism, the set of distributions is represented using a single *kernel* object in which non-random regime indicators (also called "policy variables" by [6]) index the different distributions; there is no requirement that these distributions live on the same probability space. Dawid encodes the equalities between the observational and interventional worlds via extended conditional independence (ECI) relations, including independence from (and conditional on) regime indicators.

In the standard presentation of the potential outcome approach, random variables corresponding to the outcomes for an individual under all possible interventions[2] are assumed to exist, living on a common probability space. The consistency assumption then serves to construct the factual variables as a deterministic function of the potential outcomes. Owing to the fundamental problem of causal inference, the resulting factual distribution is consistent with many different intervention distributions. However, under additional Markov restrictions on the joint distribution of the potential outcomes, the interventional distribution becomes identified from the joint distribution of the factuals under a positivity assumption. Notwithstanding this, often in practice, data are obtained on a subset of the factual variables in which case some or even all interventional distributions become only partially identified from the available (i.e., the observed) data.

3 SWIGs

The SWIG approach is designed to provide a simple way to relate graphs representing joint distributions over the observed variables and those representing joint distributions over potential outcomes. The approach is "single world" in that each of the constraints defining the model concerns a set of potential outcomes corresponding to a single joint intervention on the target variables.[3]

Following [2,3], we will assume throughout that there is a set of variables indexed by $V = \{1, \ldots, p\}$ and that a pre-specified (possibly strict) subset $A \subseteq V$ of these variables are targets for intervention. Often, we will, with a slight abuse of notation, also refer to the corresponding sets of random variables as V and A, respectively.

---

1 Though Dawid and others distinguish between potential outcomes and counterfactuals on philosophical grounds, we do not do so here; we think that this distinction, though of interest, is a separate issue from those under discussion here.

2 This does not mean that it is assumed that all variables can be intervened on.

3 Consequently, though SWIGs define a potential outcome model, they do not impose "cross-world" assumptions such as strong ignorability $Y(0), Y(1) \perp\!\!\!\perp T$ or strong conditional ignorability, $Y(0), Y(1) \perp\!\!\!\perp T \,|\, X$.

However, for proofs and formal statements, it is sometimes necessary to distinguish between the random variables and the sets that index them. For this purpose, we introduce the following notation: we define $X_B \equiv \{X_i, i \in B \subseteq V\}$, so that the complete set of factual variables is X_V and the subset that are targets for intervention are X_A. We use $\mathfrak{X}_i$ as the state space for the variable X_i, and we will let $\mathfrak{X}_V \equiv \times_{i\in V}\mathfrak{X}_i$ and $\mathfrak{X}_A \equiv \times_{i\in A}\mathfrak{X}_i$ be the state spaces for the variables with indices in V and A, respectively. Similarly, given an assignment x_V to the variables (with indices) in V, we let x_i and x_B refer to the value assigned to X_i and to the set X_B. We also make use of the usual shorthand, using, for example, A_i to refer to X_{A_i}, A for X_A, and a_i to denote x_{a_i}.

Definition 1. Given a directed acyclic graph (DAG) $\mathcal{G}$ with vertex set V, the SWIG $\mathcal{G}(a)$ corresponding to an intervention that sets the variables in $A = \{A_1, \ldots, A_k\} \subseteq V$ to $a = (a_1, \ldots, a_k) \in \mathfrak{X}_A$ is constructed as follows:
(1) Every vertex $A_i \in A$ is split into two halves, a "random half" and a "fixed half."
(2) The random half contains A_i and inherits all of the incoming edges directed into A_i in the original graph.
(3) The fixed half inherits all of the outgoing edges directed out of A_i in the original graph and is labeled with the value a_i.
(4) Random vertices in nodes on the graph are then re-labeled according to one of the schemes mentioned subsequently.

There are three labeling schemes that may be employed in step (4):

Uniform labeling: Every random vertex Y in the SWIG $\mathcal{G}(a)$ is labeled with the full vector $Y(a_1, \ldots, a_k)$.

Temporal labeling: Given a total ordering of the vertices on the original graph, each random vertex Y is labeled $Y(a_1, \ldots a_i)$ with the values corresponding to those vertices $A_1, \ldots, A_i$ that are ordered prior to Y.

Ancestral labeling: Each random vertex Y is labeled $Y(a_{\mathrm{an}_{\mathcal{G}(a)}(Y)})$, where $a_{\mathrm{an}_{\mathcal{G}(a)}(Y)}$ corresponds to those fixed vertices a_i that are *still* ancestors of Y *after* splitting the nodes in A.

Temporal labeling may be seen as encoding the assumption that interventions in the future do not affect outcomes in the past. Thus, the potential outcome for a variable $Y(a_1, \ldots, a_k)$, in a world in which there is an intervention on $A_1, \ldots, A_k$, is a function only of those interventions $A_1, \ldots, A_i$ that took place (temporally) before Y, so that $Y(a_1, \ldots, a_k) = Y(a_1, \ldots, a_i)$. This is the natural labeling scheme to apply in the context

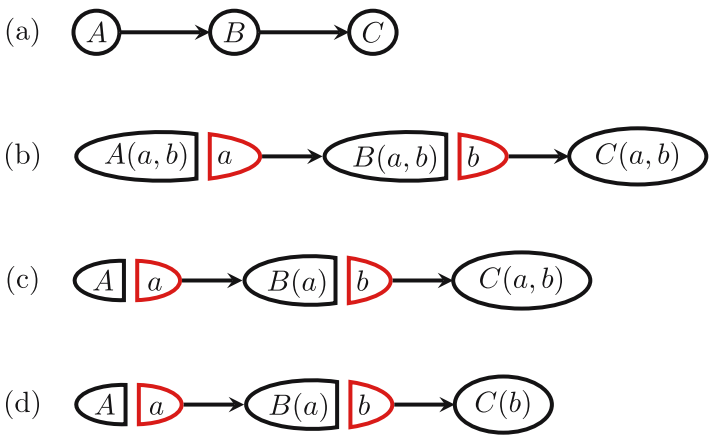

Figure 1: Illustration of SWIG labeling schemes. (a) DAG $\mathcal{G}$ representing the observed joint distribution $p(A, B, C)$; (b) SWIG $\mathcal{G}(a, b)$ with uniform labeling; (c) SWIG $\mathcal{G}(a, b)$ with temporal labeling; and (d) SWIG $\mathcal{G}(a, b)$ with ancestral labeling. (These and other figures were created using the `swings` TikZ package, available on CTAN.)

where all variables are temporally ordered and missing edges correspond (solely) to the absence of population-level direct effects.

Ancestral labeling encodes the assumption that the potential outcome $Y(a_1, \ldots, a_k)$ is solely a function of those interventions that are (still) causally antecedent to Y in the context of the other interventions that are being carried out. Thus, for example, in Figure 1(d), the vertex for C is labeled $C(b)$ and not $C(a, b)$ because after intervention on B, there is no directed path from A to C. This labeling corresponds to the interpretation of missing edges in the graph in terms of the absence of individual-level direct effects so that, for example, $C(a, b) = C(b)$ in Figure 1(d). [3, §7] also discuss more general schemes that assume a time order, but also allow some missing edges to be interpreted at the individual level and others at the population (or distribution) level; in that article ancestral labeling is termed "minimal labeling."

Uniform labeling corresponds to the absence of any assumption regarding equality of potential outcomes (as random variables) across different interventions.[4] In the potential outcome framework, this would often appear somewhat unnatural. However, in this article, we will use this labeling to show that although we may wish to adopt the additional equalities between potential outcomes that are implied by the temporal and/or causal relationships, our results do not require these equalities. In addition, SWIGs with this labeling scheme are essentially isomorphic to the augmented decision diagrams proposed in [8]. In particular, note that under the uniform labeling scheme, the set of random variables appearing in two SWIGs $\mathcal{G}(a)$ and $\mathcal{G}(a^*)$, where a, $a^* \in \mathcal{X}_A$, have no overlap; this will continue to hold when, in section 3.7, we consider SWIGs $\mathcal{G}(b)$ where we intervene on a (possibly empty) subset $B \subseteq A$.

3.1 Distributional consistency for SWIGs

In order to relate passively observed distributions to those under intervention, we introduce a consistency assumption relating sets of counterfactual distributions. For this purpose, we introduce the following notation:

$$\mathcal{P}_A \equiv \{p(V(a)) \mid a \in \mathcal{X}_A\}, \tag{1}$$

$$\mathcal{P}_A^{\subseteq} \equiv \bigcup_{D \subseteq A} \mathcal{P}_D. \tag{2}$$

Thus, $\mathcal{P}_A$ is the set of counterfactual distributions over V that arise from all possible joint interventions setting the variables in A to a value $a \in \mathcal{X}_A$. Likewise, $\mathcal{P}_A^{\subseteq}$ is the set of counterfactual distributions over V resulting from all possible joint interventions on subsets D of A; this includes the case $D = \varnothing$, corresponding to the observed distribution, so $p(V) \in \mathcal{P}_A^{\subseteq}$.

We make the following consistency assumption.[5]

Definition 2. (Distributional consistency for SWIGs) The set of distributions $\mathcal{P}_A^{\subseteq}$ will be said to obey distributional consistency if, given $B_i \in A$ and $C \subseteq A \backslash \{B_i\}$, where C may be empty, for all y, b, c:

$$p(Y(b, c) = y, B_i(b, c) = b) = p(Y(c) = y, B_i(c) = b), \tag{3}$$

where $Y = V \backslash \{B_i\}$. As a special case, if C is empty, then for all y, b:

$$p(Y(b) = y, B_i(b) = b) = p(Y = y, B_i = b). \tag{4}$$

Equalities (3) and (4) simply state that the probability of the event $\{Y = y, B_i = b\}$, where B_i is the "natural" or (in Dawid's terminology) ITT variable, remains the same whether or not there is (subsequently) an intervention that targets B_i and sets it to b.

4 Even in contexts where one assumes temporal and/or causal relationships between counterfactual random variables, uniform labeling is useful when discussing more than one SWIG, since it makes clear which variable is in which SWIG (see the potential outcome calculus in [7] as an example).

5 We thank Philip Dawid for pointing out an important omission in this definition in an earlier draft of this article.

(4) implies that $p(B_i(b) = b) = p(B_i = b)$,[6] and thus, $p(Y(b) = y \mid B_i(b) = b) = p(Y = y \mid B_i = b)$. This has the interpretation that an intervention on B_i setting it to b is "ideal" in the sense that for the remaining variables Y, the intervention does not change the distribution of Y given $B_i = b$. That $p(B_i(b) = b) = p(B_i = b)$ can be seen as following from the fact that B_i and $B_i(b)$ represent, respectively, the natural value taken by B_i in the absence of an intervention and the natural value of B_i immediately prior to an intervention.

Under a standard potential outcome model that includes equalities between random variables, (3) follows directly from the consistency assumption and recursive substitution:

$$B_i(b, c) = b \Rightarrow B_i(c) = B_i(b, c) = b \Rightarrow Y(b, c) = Y(B_i(c), c) = Y(c).$$

As with the discussion of labeling earlier, in a potential outcome theory, it is natural to assume consistency at the level of random variables. Our motivation here for formulating consistency via (3) as a relation between distributions is solely to make clear that we do not require the stronger assumption for our results. However, proceeding in this way makes the notation more cumbersome since every potential outcome variable is labeled with every intervention.

Distributional consistency may also be formulated in terms of a dynamic regime. Let g_i^* denote the dynamic regime[7] on B which "intervenes" to set the intervention target to the "natural" value that the variable B_i would take in the absence of an intervention. Let $V(g_i^*, c)$ be the set of potential outcomes that would arise under g_i^* in conjunction with an intervention setting C to c. We may then re-express (3) as:

$$p(V(g_i^*, c)) = p(V(c)). \tag{5}$$

In words, in the context of an intervention setting C to c, a dynamic regime that intervenes to set B_i to the value that it would have taken anyway has no effect on the distribution of V.[8]

Though the distributional consistency assumption involves a single variable B_i, repeated applications imply the same conclusion for a set B.

Lemma 3. *If $\mathcal{P}_A^{\subseteq}$ obeys distributional consistency, B and C are disjoint subsets of A, where C may be empty, then for all y, b, c:*

$$p(Y(b, c) = y, B(b, c) = b) = p(Y(c) = y, B(c) = b), \tag{6}$$

where $Y = V \backslash B$.

Proof. We prove this by induction on the size of B. The base case follows by definition of distributional consistency. Let B_i be a variable in B, and let $B_{-i} = B \backslash \{B_i\}$.

---

6 In a standard potential outcome model, it would follow by definition of B_i, as the "natural value" of treatment, that $B_i(b) = B_i$. Indeed, in the standard potential outcome approach, there will not be a need to write $B_i(b)$. Readers familiar with the *do*-operator [9] should be aware that whereas in that theory, intervention on a variable precludes observing the natural value, in the potential outcome theory, at least conceptually, we are supposing that the natural value could be observed and then, an instant later, we could intervene upon it without the natural value having any downstream causal effects.
Also, note that the SWIG local Markov property (Definition 7) will imply the stronger condition that $p(B_i(b^*) = b) = p(B_i = b)$ for all $b, b^* \in \mathcal{X}_B$ (see Lemma 8).

7 A dynamic regime is an intervention in which the value to which the variable is set is a function of the values taken by earlier variables. With g_i^*, the earlier variable is the natural value that the variable would take on (see footnote 6).

8 Note that for (5) to be equivalent to (3), we require that when $B_i(b, c) = b$, then $V(g_i^*, c) = V(b, c)$. This will hold if we assume: (i) that the values taken by variables that occur prior to the intervention on B_i are unaffected by this intervention, and (ii) for variables that arise after the intervention, it makes no difference whether the value b is imposed due to the dynamic regime g_i^* and natural value $B_i(g^*, c) = b$, or due to a regime that uniformly imposes b. Both assumptions will hold if one views earlier variables, including the natural value $B_i(b, c)$, as having values that are determined prior to the decision to intervene on B_i (see §4.2.1 for more discussion of this point).

$$\begin{aligned} p(Y(b,c)=y, B(b,c)=b) &= p(Y(b_i,b_{-i},c)=y, B_i(b_i,b_{-i},c)=b_i, B_{-i}(b_i,b_{-i},c)=b_{-i}) \\ &= p(Y(b_{-i},c)=y, B_i(b_{-i},c)=b_i, B_{-i}(b_{-i},c)=b_{-i}) \\ &= p(Y(c)=y, B_i(c)=b_i, B_{-i}(c)=b_{-i}) \\ &= p(Y(c)=y, B(c)=b). \end{aligned}$$

Here, the second equality applies distributional consistency, taking "C" to be $B_{-i} \cup C$; the third applies the induction hypothesis, taking "Y" to be $Y \cup \{B_i\}$ and "B" to be B_{-i}. □

The next lemma relates equality of conditional distributions with and without an intervention on B.

Lemma 4. *Suppose $\mathcal{P}_A^{\subseteq}$ obeys distributional consistency. Let B and C be disjoint subsets of A, where C may be empty, and let Y and W be disjoint subsets of $V \setminus B$. It then follows that:*

$$p(Y(b,c)=y \mid B(b,c)=b, W(b,c)=w) = p(Y(c)=y \mid B(c)=b, W(c)=w). \tag{7}$$

Proof. This follows by applying Lemma 3 to $p(Y(b,c), B(b,c), W(b,c))$, and $p(B(b,c), W(b,c))$. □

In addition, we have the following:

Lemma 5. *Suppose $\mathcal{P}_A^{\subseteq}$ obeys distributional consistency, and let B and C be disjoint subsets of A, where C may be empty. If $B \subseteq W \subseteq V$ and $p(W(b,c))$ is not a function of b, then it follows from distributional consistency that $p(W(b,c)) = p(W(c))$.*

Proof.

$$\begin{aligned} p(X_W(b,c)=w) &= p(X_{W\setminus B}(b,c)=w_{W\setminus B}, X_B(b,c)=w_B) \\ &= p(X_{W\setminus B}(w_B,c)=w_{W\setminus B}, X_B(w_B,c)=w_B) \\ &= p(X_{W\setminus B}(c)=w_{W\setminus B}, X_B(c)=w_B) \\ &= p(X_W(c)=w). \end{aligned}$$

Here, we use that $p(W(b,c))$ is not a function of b in the second equality and distributional consistency via Lemma 3 in the third. □

Note that distributional consistency (3) does not imply the analogous result for conditional distributions. In particular, it is possible to have $B_i \in Y$, $p(Y(b) \mid M(b))$ not be a function of b and yet $p(Y(b) \mid M(b)) \neq p(Y \mid M)$. This is because even if $p(Y(b) \mid M(b))$ is not a function of b, both $p(Y(b), M(b))$ and $p(M(b))$ may still be functions of b, in which case there is no way to apply (3) to relate them to distributions in which B is not intervened on.

However, when the conditioning set contains B, we have the following:

Lemma 6. *Suppose $\mathcal{P}_A^{\subseteq}$ obeys distributional consistency, with B and C disjoint subsets of A, where C may be empty. Further, let Y and W be disjoint sets with $B \subseteq W$. If $p(Y(b,c) \mid W(b,c))$ is not a function of b, then*

$$p(Y(b,c) \mid W(b,c)) = p(Y(c) \mid W(c)). \tag{8}$$

Proof.

$$\begin{aligned} p(X_Y(b,c)=y \mid X_W(b,c)=w) &= p(X_Y(b,c)=y \mid X_{W\setminus B}(b,c)=w_{W\setminus B}, X_B(b,c)=w_B) \\ &= p(X_Y(w_B,c)=y \mid X_{W\setminus B}(w_B,c)=w_{W\setminus B}, X_B(w_B,c)=w_B) \\ &= p(X_Y(c)=y \mid X_{W\setminus B}(c)=w_{W\setminus B}, X_B(c)=w_B). \end{aligned}$$

Here, the second equality uses the fact that $p(X_Y(b,c)=y \mid X_W(b,c)=w)$ is not a function of b, while the third follows from distributional consistency via Lemma 4. □

3.2 Local Markov property defining the SWIG model

Although we derive a SWIG graphically from the original DAG by node splitting, we will define the model by associating a local Markov property with the SWIG and the potential outcome distribution. The resulting model corresponds to the FFRCISTG model of [2] (see [3, Appendix C]). We will then derive the Markov property for the original DAG and the observed distribution from these by applying distributional consistency.

Given a DAG $\mathcal{G}$ with vertices $V = \{1, \ldots, p\}$, we will use $\mathrm{pa}_{\mathcal{G}}(i)$ to indicate the (index) set of variables that are the parents of W_i in the original DAG $\mathcal{G}$, and let $\mathrm{pre}_{\prec}(i)$ indicate $\{1, \ldots, i-1\}$, the predecessors of i under a total ordering $\prec$ that is consistent with the edges in $\mathcal{G}$. We will drop the subscript when the DAG or ordering is clear from context.

The SWIG local Markov property is defined on the set of distributions $\mathcal{P}_A \equiv \{p(V(a)) | a \in \mathcal{X}_A\}$, where $A \subseteq V$ is the maximal set of variables that may be intervened on see ([10, §1.2.4] and [11]).

Definition 7. A set of potential outcome distributions $\mathcal{P}_A$ obeys *the SWIG ordered local Markov property for DAG $\mathcal{G}$ under $\prec$* if for all $i \in V$, $a \in \mathcal{X}_A$, and $w \in \mathcal{X}_{\mathrm{pre}_{\prec}(i)}$,

$$p(X_i(a) \,|\, X_{\mathrm{pre}_{\prec}(i)}(a) = w) \tag{9}$$

is a function only of $a_{\mathrm{pa}_{\mathcal{G}}(i) \cap A}$ and $w_{\mathrm{pa}_{\mathcal{G}}(i) \setminus A}$.

In words, (9) states that after intervening on A, the distribution of $X_i(a)$ given its predecessors depends solely on the values taken by intervention targets in A that are parents of i, and by any other (random) variables that are parents of i but that are not intervened on, and hence are not in A.[9]

Though the function of the local property is to define and characterize the potential outcome model, intuition may be gained by observing that the local property follows from d-separation applied to the SWIG $\mathcal{G}(a)$.[10] Specifically, the condition (9) corresponds to two sets of d-separations.

***d*-separation from fixed nodes:** That $p(X_i(a) \,|\, X_{\mathrm{pre}(i)}(a))$ does not depend on $a_{A \setminus \mathrm{pa}_{\mathcal{G}}(i)}$ is encoded in the SWIG $\mathcal{G}(a)$ by the d-separation of $X_i(a)$ from fixed nodes a_j that correspond to vertices A_j that are not the parents of X_i in $\mathcal{G}$ given the parents of $X_i(a)$ in $\mathcal{G}(a)$, both random and fixed (see [7,10,12]). Specifically, we have:

$$X_i(a) \perp\!\!\!\perp_d a_{A \setminus \mathrm{pa}(i)} | a_{A \cap \mathrm{pa}(i)}, X_{\mathrm{pa}(i) \setminus A}(a), \tag{10}$$

where here we used $\perp\!\!\!\perp_d$ to indicate d-separation[11] in the SWIG $\mathcal{G}(a)$ and use lowercase letters, e.g., $a_{A \setminus \mathrm{pa}(i)}$, to refer to fixed nodes. We may further decompose the set of fixed nodes $a_{A \setminus \mathrm{pa}(i)}$:

$$X_i(a) \perp\!\!\!\perp_d \overbrace{a_{A \setminus \mathrm{pre}(i)}}^{\text{time order}}, \overbrace{a_{(A \cap \mathrm{pre}(i)) \setminus \mathrm{pa}(i)}}^{\text{causal Markov prop.}} \mid \underbrace{a_{A \cap \mathrm{pa}(i)}}_{\text{fixed parents}}, \underbrace{X_{\mathrm{pa}(i) \setminus A}(a).}_{\text{random parents}} \tag{11}$$

The set of fixed nodes in $a_{A \setminus \mathrm{pre}(i)}$ correspond to interventions on variables that occur after X_i and thus do not change $p(X_i(a) \,|\, X_{\mathrm{pre}(i)}(a))$. Likewise, the effects of the fixed nodes in $a_{(A \cap \mathrm{pre}(i)) \setminus \mathrm{pa}(i)}$ are screened off by the random and fixed nodes that are parents of $X_i(a)$.

---

9 In [3, §8, Def.44], a weaker Markov property was stated that did not require that (9) is not a function of $a_{A \setminus \mathrm{pa}_{\mathcal{G}}(i)}$. This weaker condition does not imply the Markov property for the observed distribution (unless $A = V$). Consequently, Propositions 45 and 46 and Theorem 65(c) in [3] are incorrect. Correct reformulations are given in Theorems 10, 11, and 12.

10 This is as to be expected since d-separation encodes the global property that is implied by the local property.

11 Note that in other articles [3,7,10] d-connection for SWIGs is defined such that fixed nodes may never occur as non-endpoint vertices on d-connecting paths. In those articles, we never formally condition on fixed nodes. Here, in (10) for the purpose of formulating the local property, we formally include the fixed parents of $X_i(a)$ in the set that is (graphically) conditioned on. This is solely in order to make the development similar to the decision diagram approach we consider subsequently.

Table 1: Defining properties for the SWIG $\mathcal{G}(\mathbf{x})$ in Figure 2(b), expressed via factorization

Local Markov property for $\mathcal{G}(x_1, x_2)$ via factorization terms
$p(H(x_0, x_1))$
$p(X_0(x_0, x_1) \mid H(x_0, x_1))$
$p(Z(x_0, x_1) \mid H(x_0, x_1), X_0(x_0, x_1))$
$p(X_1(x_0, x_1) \mid H(x_0, x_1), X_0(x_0, x_1), Z(x_0, x_1))$
$p(Y(x_0, x_1) \mid H(x_0, x_1), X_0(x_0, x_1), Z(x_0, x_1), X_1(x_0, x_1))$

Arguments in $p(V_i(\mathbf{x}) \mid V_{\mathrm{pre}(i)}(\mathbf{x}))$ on which this term does not depend are colored red. Note that the arguments in $V_{\mathrm{pre}(i)}(\mathbf{x})$ on which the term depends, correspond to the parents of $V_i(\mathbf{x})$ in $\mathcal{G}(\mathbf{x})$; these are written in black. For example, for the term corresponding to $V_i = Y$, the arguments are x_1 and $Z(\mathbf{x})$, and these are the parents of $Y(\mathbf{x})$ in $\mathcal{G}(\mathbf{x})$.

***d*-separation from random nodes:** That $p(X_i(a) \mid X_{\mathrm{pre}(i)}(a) = w)$ does not depend on $w_{\mathrm{pre}(i)\setminus(\mathrm{pa}_{\mathcal{G}}(i)\setminus A)}$ is encoded in $\mathcal{G}(a)$ by the d-separation of $X_i(a)$ from $X_{\mathrm{pre}(i)\setminus(\mathrm{pa}(i)\setminus A)}(a)$ conditioning on the parents of $X_i(a)$ in $\mathcal{G}(a)$, both random and fixed:

$$X_i(a) \perp\!\!\!\perp_d X_{\mathrm{pre}(i)\setminus(\mathrm{pa}(i)\setminus A)}(a) \mid a_{A\cap\mathrm{pa}(i)}, X_{\mathrm{pa}(i)\setminus A}(a). \tag{12}$$

The random vertices $X_{\mathrm{pre}(i)\setminus(\mathrm{pa}(i)\setminus A)}(a)$ may be further decomposed:

$$X_i(a) \perp\!\!\!\perp_d \overbrace{X_{\mathrm{pre}(i)\setminus\mathrm{pa}(i)}(a)}^{\text{assoc. Markov prop.}}, \overbrace{X_{\mathrm{pa}(i)\cap A}(a)}^{\text{ignorability}} \mid \underbrace{a_{A\cap\mathrm{pa}(i)}}_{\text{fixed parents}}, \underbrace{X_{\mathrm{pa}(i)\setminus A}(a)}_{\text{random parents}}. \tag{13}$$

The d-separation of $X_i(a)$ from nodes representing the natural value of variables that are in A and parents of X_i in $\mathcal{G}$ corresponds to ignorability. On the other hand, the d-separation of $X_i(a)$ from variables that are predecessors, but not parents, of X_i in $\mathcal{G}$ can be regarded as an associational Markov property.

3.3 Example

The d-separations given by (11) and (13) can obviously be stated as a single graphical condition for each random vertex $V_i(a)$ in $\mathcal{G}(a)$. In Tables 1 and 2, we give the SWIG local Markov property corresponding to the SWIG $\mathcal{G}(\mathbf{x}) \equiv \mathcal{G}(x_1, x_2)$, shown in Figure 2(b),[12] under the ordering (H, X_0, Z, X_1, Y): Table 1 in terms of factorization; Table 2 via d-separation. Note that for each V_i, the number of arguments on which $p(V_i(\mathbf{x}) \mid V_{\mathrm{pre}(i)}(\mathbf{x}))$ depends corresponds exactly to the number of parents (random and fixed) of the corresponding random variable in $\mathcal{G}(\mathbf{x})$ in Figure 2(b): zero for $H(\mathbf{x})$ and $X_0(\mathbf{x})$ and two for $Z(\mathbf{x})$, $X_1(\mathbf{x})$, and $Y(\mathbf{x})$. This is also the number of terms listed to the right of the conditioning bar in Table 2. Here, as elsewhere in this article, we use the uniform labeling because we wish to emphasize that our results do not require any equalities between random variables.

3.4 Consequences of the local Markov property

Under distributional consistency, it follows from the SWIG local Markov property that whether or not interventions in the future occur has no effect on the distribution of prior variables.

12 [8, Figure 15] gives the corresponding SWIG under ancestral labeling. In Dawid's discussion of this example [8, Figure 15], two conditions are stated as supporting g-computation. The first of these is correct, but the second should be $Y(x_0, x_1) \perp\!\!\!\perp X_0$, not $Z(x_0) \perp\!\!\!\perp X_0$.

Table 2: d-separation relations corresponding to the SWIG local Markov property in the SWIG $\mathcal{G}(x_1, x_2)$ in Figure 2(b)

Local Markov property for $\mathcal{G}(x_1, x_2)$ via d-separation				
$H(x_0, x_1)$	$\perp\!\!\!\perp_d$	x_0, x_1		
$X_0(x_0, x_1)$	$\perp\!\!\!\perp_d$	$H(x_0, x_1), x_0, x_1$		
$Z(x_0, x_1)$	$\perp\!\!\!\perp_d$	$X_0(x_0, x_1), x_1$	|	$H(x_0, x_1), x_0$
$X_1(x_0, x_1)$	$\perp\!\!\!\perp_d$	$X_0(x_0, x_1), x_0, x_1$	|	$H(x_0, x_1), Z(x_0, x_1)$
$Y(x_0, x_1)$	$\perp\!\!\!\perp_d$	$X_0(x_0, x_1), X_1(x_0, x_1), H(x_0, x_1), x_0$	|	$Z(x_0, x_1), x_1$

Here, x_1 and x_2 refer to the fixed nodes, and $\perp\!\!\!\perp_d$ indicates d-separation in the SWIG (see also footnote 11 regarding the formal inclusion of fixed nodes on the RHS of the conditioning bar).

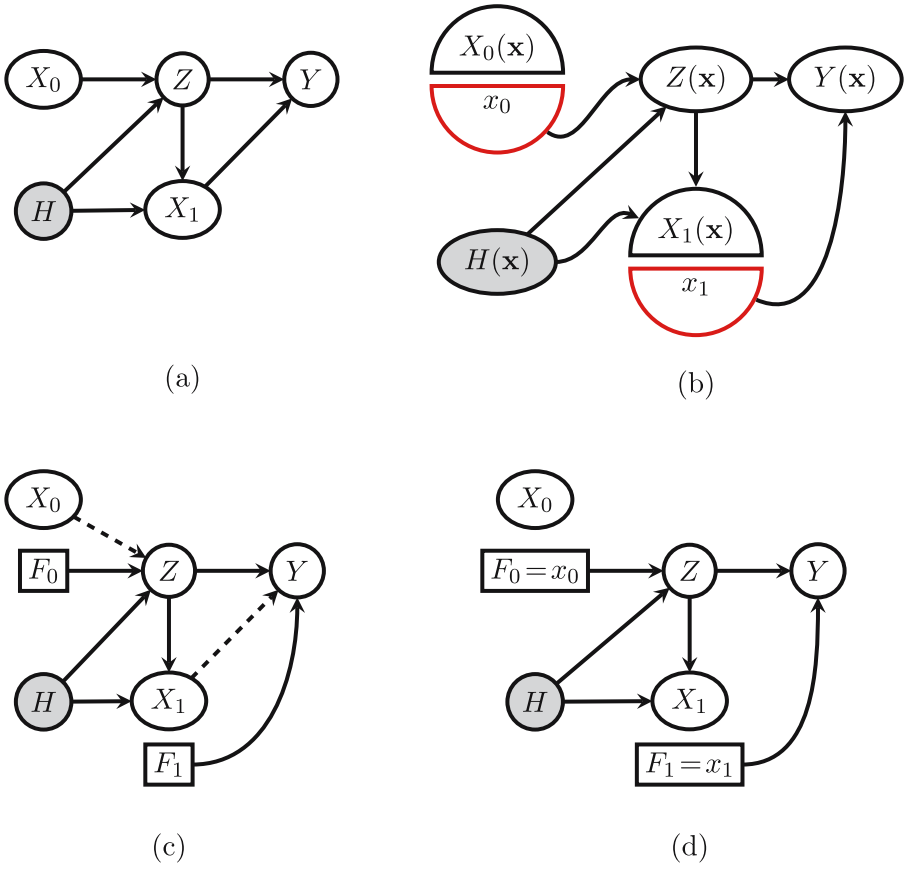

Figure 2: (a) The DAG $\mathcal{G}$ originally considered in Ex. 11.3.3, Fig. 11.12 in [15, p. 353]; here, H is unobserved; (b) the SWIG $\mathcal{G}(\mathbf{x})$ under uniform labeling, $\mathbf{x} \equiv (x_1, x_2)$; Figure 15 in [8] shows the SWIG with ancestral labeling; (c) the reformulated augmented graph $\mathcal{G}^*$ (this corresponds to Dawid's ITT DAG $\mathcal{G}^*$ shown in Figure 13 of [8, p. 62] after marginalizing X_0, X_1 and then removing * from the ITT variables); and (d) the resulting graph under the regime $F_0 = x_0$, $F_1 = x_1$; this is the graph that encodes the reformulated Markov property; see Definition 18.

Lemma 8. *If $\mathcal{P}_A^{\subseteq}$ obeys distributional consistency and $\mathcal{P}_A$ obeys the SWIG ordered local Markov property for DAG $\mathcal{G}$ under $\prec$, then for all $k \in V$ and $a \in \mathcal{X}_A$:*

$$p(X_1(a), \ldots, X_k(a)) = p(X_1(a_{\mathrm{pre}(k)\cap A}), \ldots, X_k(a_{\mathrm{pre}(k)\cap A})). \tag{14}$$

Proof. First observe that since

$$p(X_1(a), \ldots, X_k(a)) = \prod_{i=1}^{k} p(X_i(a) \,|\, X_{\mathrm{pre}(i)}(a)),$$

and the local Markov property implies that $p(X_i(a) \mid X_{\mathrm{pre}(i)}(a))$ does not depend on $a_{A \setminus \mathrm{pre}(i)}$, it follows that $p(X_1(a), \ldots, X_k(a))$ does not depend on $a_{A \setminus \mathrm{pre}(k)}$.

We now prove the claim by reverse induction on the ordering of the vertices in V.

For the base case, suppose k is the maximal vertex in V. If $k \notin A$, then (14) holds trivially since $A = \mathrm{pre}(k) \cap A$. If $k \in A$, then since $k \notin \mathrm{pre}(k)$, $p(X_1(a), \ldots, X_k(a))$ does not depend on a_k, and thus, by Lemma 5, $p(X_1(a), \ldots, X_k(a)) = p(X_1(a_{A \setminus \{k\}}), \ldots, X_k(a_{A \setminus \{k\}}))$.

Our inductive hypothesis is that (14) holds for $k = j + 1$, so that

$$p(X_1(a), \ldots, X_{j+1}(a)) = p(X_1(a_{\mathrm{pre}(j+1) \cap A}), \ldots, X_{j+1}(a_{\mathrm{pre}(j+1) \cap A})).$$

Summing both sides over x_{j+1}, we obtain:

$$p(X_1(a), \ldots, X_j(a)) = p(X_1(a_{\mathrm{pre}(j+1) \cap A}), \ldots, X_j(a_{\mathrm{pre}(j+1) \cap A})). \tag{15}$$

If $j \notin A$, then (15) establishes the claim since $\mathrm{pre}(j + 1) \cap A = \mathrm{pre}(j) \cap A$. If $j \in A$, then note that we have already established earlier that the left-hand side (LHS) of (15) is not a function of a_j. Consequently, the right-hand side (RHS) is also not a function of a_j. It then follows from Lemma 5 that

$$p(X_1(a_{\mathrm{pre}(j+1) \cap A}), \ldots, X_j(a_{\mathrm{pre}(j+1) \cap A})) = p(X_1(a_{\mathrm{pre}(j) \cap A}), \ldots, X_j(a_{\mathrm{pre}(j) \cap A})).$$

This completes the proof. □

The next lemma gives a simple characterization of the consequences of the SWIG local Markov property in conjunction with distributional consistency.

Lemma 9. *If $\mathcal{P}^{\subseteq}_A$ obeys distributional consistency and $\mathcal{P}_A$ obeys the SWIG ordered local Markov property for DAG $\mathcal{G}$ under $\prec$, then:*

$$p(X_i(a) \mid X_{\mathrm{pre}(i)}(a)) \tag{16}$$

$$= p(X_i(a_{\mathrm{pre}(i) \cap A}) \mid X_{\mathrm{pre}(i)}(a_{\mathrm{pre}(i) \cap A})) \tag{17}$$

$$= p(X_i(a_{\mathrm{pa}(i) \cap A}) \mid X_{\mathrm{pre}(i)}(a_{\mathrm{pa}(i) \cap A})) \tag{18}$$

$$= p(X_i(a_{\mathrm{pa}(i) \cap A}) \mid X_{\mathrm{pa}(i)}(a_{\mathrm{pa}(i) \cap A})) \tag{19}$$

$$= p(X_i(a_{\mathrm{pa}(i) \cap A}) \mid X_{\mathrm{pa}(i) \setminus A}(a_{\mathrm{pa}(i) \cap A})). \tag{20}$$

Since the SWIG local Markov property (9) states that (16) is not a function of $a_{A \setminus \mathrm{pa}_{\mathcal{G}}(i)}$, the equality of (16) and (18) may appear to follow immediately. However, as noted in the discussion prior to Lemma 6, the fact that a counterfactual conditional distribution $p(Y(a_j) \mid W(a_j))$ does not depend on the specific value, a_j, of an intervention on A_j does *not* imply that $p(Y(a_j) \mid W(a_j)) = p(Y \mid W)$.

Proof. Here, (17) follows since by Lemma 8

$$p(X_i(a), X_{\mathrm{pre}(i)}(a)) = p(X_i(a_{\mathrm{pre}(i) \cap A}), X_{\mathrm{pre}(i)}(a_{\mathrm{pre}(i) \cap A})).$$

(18) follows from Definition 7 and Lemma 6. Finally, (19) and (20) follow from the SWIG local Markov property via (8) since $p(X_i(a) \mid X_{\mathrm{pre}(i)}(a) = x_{\mathrm{pre}(i)})$ does not depend on $x_{\mathrm{pre}(i) \setminus (\mathrm{pa}(i) \setminus A)} = (x_{\mathrm{pre}(i) \setminus \mathrm{pa}(i)}, x_{\mathrm{pa}(i) \cap A})$. □

3.5 Markov property for the observed distribution

We now show that distributional consistency together with the SWIG local Markov property implies the usual local Markov property [13] for the observed distribution.

Theorem 10. *If $\mathcal{P}^{\subseteq}_A$ obeys distributional consistency and $\mathcal{P}_A$ obeys the SWIG ordered local Markov property for $\mathcal{G}$ and $\prec$, then $p(V)$ obeys the usual DAG ordered local Markov property w.r.t. $\mathcal{G}$ and $\prec$.*

Proof. Let $v^* \in \mathcal{X}_{\mathrm{pre}(i)}$.

$$\begin{aligned} p(X_i = v \mid X_{\mathrm{pre}(i)} = v^*) &= p(X_i(v^*_{\mathrm{pre}(i)\cap A}) = v \mid X_{\mathrm{pre}(i)}(v^*_{\mathrm{pre}(i)\cap A}) = v^*) \\ &= p(X_i(v^*_{\mathrm{pa}(i)\cap A}) = v \mid X_{\mathrm{pa}(i)\setminus A}(v^*_{\mathrm{pa}(i)\cap A}) = v^*_{\mathrm{pa}(i)\setminus A}). \end{aligned} \tag{21}$$

Here, the first equality follows from distributional consistency via Lemma 4. The second follows directly from the equality of (17) and (20) in Lemma 9. Since the last line is not a function of $v^*_{\mathrm{pre}(i)\setminus \mathrm{pa}(i)}$, the ordered local Markov property for the DAG holds. □

3.5.1 Discussion of relation to Dawid

Dawid takes the reverse approach to ours: he proposes additional extended Markovian conditions that, when added to the usual Markov property for the observable law, will imply the Markov property for his extended graph. However, as we describe in detail below, our approach appears to be simpler in that, given distributional consistency, it requires only one property per variable, giving $|V|$ constraints in total; in contrast, Dawid requires one property for every observed variable in V, together with two additional properties for each intervention target in A for a total of $(|V| + 2|A|)$.

In addition, our approach captures context-specific independences, corresponding to "dashed" edges in Dawid's diagrams; furthermore, these are not captured directly in Dawid's A+B formulation. We show that by restating the SWIG local property in Dawid's notation, we are able to provide a characterization of the (extended) Markov properties for the augmented graph and the original graph that also requires only one constraint per variable, plus distributional consistency.

It is the case that Dawid incorporates distributional consistency into his defining independences, whereas we state it as a separate property that precedes the definition of the model. However, as we have shown earlier, distributional consistency may be seen as a tautologous property, the truth of which is implicit in the notion of an ideal intervention: distributional consistency states that if B would naturally take the value b, then an ideal intervention that would set B to b has no effect on the distribution of (all) the other variables. For this reason, we believe it is natural to distinguish consistency from the other properties being used to define the model.

However, in the spirit of Dawid's approach, in Appendix A.1, we show that if $\mathcal{P}^{\subseteq}_A$ obeys distributional consistency, then $\mathcal{P}_A$ will obey the SWIG local Markov property corresponding to $\mathcal{G}$ if: (i) $p(V)$ is positive and obeys the (ordinary) local Markov property for the graph $\mathcal{G}$; and (ii) $\mathcal{P}_A$ obeys the SWIG local Markov property corresponding to $\overline{\mathcal{G}}$, a complete supergraph of $\mathcal{G}$. This formulation requires $2|V|$ restrictions.

3.6 Identification of the potential outcome distribution $p(V(a))$ from $p(V)$

We show that it follows from the SWIG local Markov property that $p(V(a))$ is identified given the distribution over the observables provided that the relevant conditional distributions are identified from the distribution of the observables.

Theorem 11. *Suppose that $\mathcal{P}^{\subseteq}_A$ obeys distributional consistency and $\mathcal{P}_A$ obeys the SWIG ordered local Markov property for $\mathcal{G}$ and $\prec$. Let $a \in \mathcal{X}_A$ be an assignment to the intervention targets in A, and let $v \in \mathcal{X}_V$. Then, for all i:*

$$p(X_i(a) = v_i \mid X_{\mathrm{pre}(i)}(a) = v_{\mathrm{pre}(i)}) = p(X_i = v_i \mid X_{\mathrm{pa}(i)\setminus A} = v_{\mathrm{pa}(i)\setminus A}, X_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}). \tag{22}$$

Consequently, $p(V(a))$ is identified from $p(V)$ and obeys d-separation in the SWIG $\mathcal{G}(a)$, whenever the conditional distributions on the RHS of (22) *are identified by $p(V)$.*

The equality (22) here corresponds to the property referred to as "modularity" in [3]; this is also an instance of the extended g-formula of [2,14].

Proof.

$$\begin{aligned} & p(X_i(a) = v_i \mid X_{\mathrm{pre}(i)}(a) = v_{\mathrm{pre}(i)}) \\ &= p(X_i(a_{\mathrm{pa}(i)\cap A}) = v_i \mid X_{\mathrm{pa}(i)\cap A}(a_{\mathrm{pa}(i)\cap A}) = v_{\mathrm{pa}(i)\cap A}, X_{\mathrm{pa}(i)\setminus A}(a_{\mathrm{pa}(i)\cap A}) = v_{\mathrm{pa}(i)\setminus A}) \\ &= p(X_i(a_{\mathrm{pa}(i)\cap A}) = v_i \mid X_{\mathrm{pa}(i)\cap A}(a_{\mathrm{pa}(i)\cap A}) = a_{\mathrm{pa}(i)\cap A}, X_{\mathrm{pa}(i)\setminus A}(a_{\mathrm{pa}(i)\cap A}) = v_{\mathrm{pa}(i)\setminus A}) \\ &= p(X_i = v_i \mid X_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}, X_{\mathrm{pa}(i)\setminus A} = v_{\mathrm{pa}(i)\setminus A}). \end{aligned} \tag{23}$$

Here, the first equality follows from the equality of (16) and (19); the second follows from the equality of (19) and (20); the third follows from distributional consistency via (7). □

3.7 Distributions resulting from fewer interventions

Finally, we show that if $p(V(a))$ obeys the SWIG local Markov property for $\mathcal{G}$ and distributional consistency, then if we intervene on $B \subset A$, the resulting distribution $p(V(b))$ will obey the SWIG local Markov property for $\mathcal{G}$ with respect to this reduced set of intervention targets. The two previous theorems can be seen as the special case in which $B = \varnothing$.

Theorem 12. *Suppose that $\mathcal{P}^{\subseteq}_A$ obeys distributional consistency and $\mathcal{P}_A$ obeys the SWIG ordered local Markov property for $\mathcal{G}$ and $\prec$. Let b be an assignment to the intervention targets in $B \subseteq A$, and let $v \in \mathcal{X}_V$. Then for all i:*

$$p(X_i(b) = v_i \mid X_{\mathrm{pre}(i)}(b) = v_{\mathrm{pre}(i)}) = p(X_i = v_i \mid X_{\mathrm{pa}(i)\cap B} = b_{\mathrm{pa}(i)\cap B}, X_{\mathrm{pa}(i)\setminus B} = v_{\mathrm{pa}(i)\setminus B}). \tag{24}$$

Consequently, every $p(V(b)) \in \mathcal{P}_B$ obeys the Markov property for the SWIG $\mathcal{G}(b)$ and is identified whenever the conditional distributions on the RHS of (24) *are identified by $p(V)$.*

Proof.

$$\begin{aligned} & p(X_i(b) = v_i \mid X_{\mathrm{pre}(i)}(b) = v_{\mathrm{pre}(i)}) \\ &= p(X_i(b_{\mathrm{pre}(i)\cap B}) = v_i \mid X_{\mathrm{pre}(i)}(b_{\mathrm{pre}(i)\cap B}) = v_{\mathrm{pre}(i)}) \\ &= p(X_i(b_{\mathrm{pre}(i)\cap B}, v_{\mathrm{pre}(i)\cap(A\setminus B)}) = v_i \mid X_{\mathrm{pre}(i)}(b_{\mathrm{pre}(i)\cap B}, v_{\mathrm{pre}(i)\cap(A\setminus B)}) = v_{\mathrm{pre}(i)}) \\ &= p(X_i = v_i \mid X_{\mathrm{pa}(i)\setminus A} = v_{\mathrm{pa}(i)\setminus A}, X_{\mathrm{pa}(i)\cap B} = b_{\mathrm{pa}(i)\cap B}, X_{\mathrm{pa}(i)\cap(A\setminus B)} = v_{\mathrm{pa}(i)\cap(A\setminus B)}) \\ &= p(X_i = v_i \mid X_{\mathrm{pa}(i)\setminus B} = v_{\mathrm{pa}(i)\setminus B}, X_{\mathrm{pa}(i)\cap B} = b_{\mathrm{pa}(i)\cap B}). \end{aligned}$$

Here, the first equality is by Lemma 5; the second is distributional consistency via Lemma 4; the third follows from Theorem 11 applied to $\mathcal{G}(a)$; the fourth is a simplification. □

4 Critique of Dawid's proposal

We have the following four main issues, which we describe in detail as follows:

(1) The inclusion of ITT variables within Dawid's theory appears necessary in order to distinguish causal relationships from happenstance agreement between observational and ("fat hand") intervention distributions. However, including all three of T (the "actual" treatment), T^* (the ITT variable), and F_T (the regime indicator) introduces deterministically related variables and thereby obscures the content of Dawid's defining conditional independences A and B.

(2) Related to the previous point, d-separation is no longer a complete criterion for determining conditional independence on a graph in which there are definitional deterministic relationships between the variables.[13]

13 This is also an issue for the twin network approach developed in [15].

(3) Dawid's ITT augmented diagrams incorporate context-specific independence (via dashed edges) but his results do not establish that the resulting distribution obeys all of the implied context-specific independences; these are not implied by his defining conditional independences $A + B$; these independences will not hold without additional information concerning the relation of T to T^* and F_T that is not captured in $A + B$.
(4) Dawid makes use of what he terms "fictitious" independence relations, but he argues that these are assumptions that can be made without loss of generality. This is not the case in general, though, as we show, in the context of his arguments, the resulting logical "gap" can be filled.

We show that all of these issues may be avoided by re-formulating his theory in two simple ways:
(I) Marginalizing out the post-intervention treatment variable T while keeping the ITT variable T^*.[14]
(II) Formulating the defining extended independence relations in terms of distributional consistency and the augmented ITT diagram (after marginalizing T) and intervening on all the variables in A; the local Markov property for the original variables is then implied.

The resulting theory is formally isomorphic to the SWIG theory described earlier; the augmented ITT graph can be viewed as containing the union of the nodes and edges in the original DAG $\mathcal{G}$ and the SWIG $\mathcal{G}(a)$, with the fixed nodes in the SWIG corresponding to the (non-idle) regime indicators in the augmented DAG.

4.1 The simplest setting

Consider the setting in which there is a single exposure T and an outcome Y; suppose that T takes a finite set of states $\mathfrak{T}$. Dawid's augmented causal graph with the intention-to-treat variable T^* is shown in Figure 3(d). Here, T^* represents the natural value of treatment which an individual is "selected to receive" [8, p. 52] in the absence of an intervention that would override this. This is distinct from T the "treatment actually applied" [8, p. 54, Def. 1]; F_T is a regime indicator taking values in $\mathfrak{T} \cup \{\varnothing\}$. Under Dawid's proposal the graph in Figure 3(d) represents the kernel $p(T^*, T, Y \mid F_T)$; $F_T = \varnothing$ indicates the observational regime in which case $T = T^*$ (see Figure 3(e)) where we have used a colored edge, $T^* \to T$, to indicate the deterministic relationship between T and T^*. Similarly, $F_T = t \in \mathfrak{T}$ indicates the interventional regime in which case $T = t$ (see Figure 3(f)). Note that $T \perp\!\!\!\perp T^* \mid F_T \neq \varnothing$, which is represented by the dashed edge from T^* to T in Figure 3(d) and by the absence of the edge between T^* and T in Figure 3(e).

For comparison, Figure 3(a) and (b), respectively, show the representations of the observed distribution $p(T^*, Y)$ and the joint distribution $p(T^*, Y(t))$; as suggested by the graphical structures, there is a close correspondence between these approaches when ITT variables are included in the decision theory graph. In what follows, we will show that in fact, the two theories can be shown to be isomorphic up to labeling of variables (Table 3).

Although Dawid includes ITT variables in the development here, they were absent in [16] and ultimately his goal is to remove the ITT variables, leaving the DAG shown in Figure 3(c) containing only the original variables and the treatment indicators (see bottom of [8, p. 65]). Dawid states that the augmented graphs without ITT variables are sufficient for reasoning about point interventions.

Given this, one may ask why it is necessary to introduce the ITT variables into the theory in the first place. One issue that arises is that without the ITT variables, the decision theoretic approach lacks the language to describe concepts such as the effect of treatment on the treated. In addition, the approach lacks the concepts necessary to distinguish different scenarios where there is equality between distributions in the observed and interventional worlds: those scenarios where the equality reflects agreement between an observational study and a randomized experiment due to the absence of confounding, versus those where the equality is purely "contingent" or spurious.

14 Dawid instead proposes to marginalizes out the ITT variables.

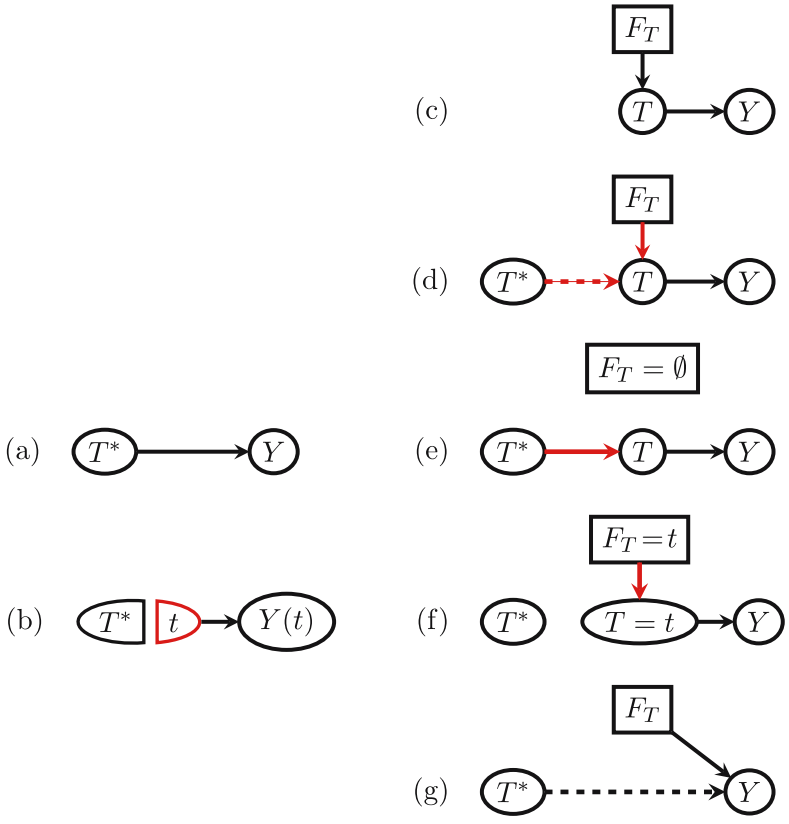

Figure 3: The simplest case of a single treatment T and outcome Y in the absence of confounding. (a) DAG $\mathcal{G}$ representing the observed joint distribution $p(T^*, Y)$; (b) SWIG $\mathcal{G}(t)$ corresponding to $\mathcal{G}$ representing $p(T^*, Y(t))$; (c) Dawid's augmented DAG representing the set of kernels $p(Y, T|F_T)$, where F_T is a regime indicator; (d) Dawid's augmented DAG with ITT variables, representing the kernels $p(Y, T^*, T|F_T)$, where F_T is a regime indicator; the dashed edge indicates that the edge between T^* and T is absent in the interventional regime, while the red edges indicate deterministic relationships; (e) the ITT augmented graph representing the observational regime $p(T^*, T, Y|F_T = \emptyset) = p(T^*, T, Y)$ under which $T^* = T$; (f) the ITT augmented graph for $p(T^*, T, Y|F_T = t) = p(T^*, t, Y|F_T = t)$, an intervention setting T to t, so $F_T = t \neq \emptyset$; and (g) the latent projection of the graph in (d) after marginalizing T. Note that in (a), (b) we use T^* (rather than T) for the natural value of treatment in order to highlight the correspondence to the ITT variables in Dawid's proposal. The graph in (g) corresponds to (a) and (b), under the correspondence $t \Leftrightarrow F_T = t$, $Y(t) \Leftrightarrow Y \mid F_T = t$.

Table 3: Correspondence between the potential outcome/SWIG approach and the decision theoretic approach

	Potential outcome	**Decision theoretic**
Graph for observed data	$\mathcal{G}$	ITT DAG, $F_T = \emptyset$
Graph representing intervention on T	$\mathcal{G}(t)$	ITT DAG, $F_T = t$
Observed distribution	$p(T^*, Y)$	$p(T^*, Y|F_T = \emptyset)$
Distribution resulting from setting $T = t$ directly after observing T^*	$p(T^*, Y(t))$	$p(T^*, Y|F_T = t)$

Here, in the potential outcome approach we use T^* (rather than T) to denote the natural value of treatment so as to make the correspondence more self-evident

To illustrate this, consider the following story. Suppose that a manufacturer of dietary supplements carries out an observational study. They find that those who regularly consume the supplement ($T = 1$) have lower levels of "bad" cholesterol (Y) than the people who do not ($T = 0$). Buoyed by these results, the manufacturer hires a company to perform a randomized trial. The results of the previous study are given to the company; it is made clear that the manufacturer would like these results confirmed and that repeat

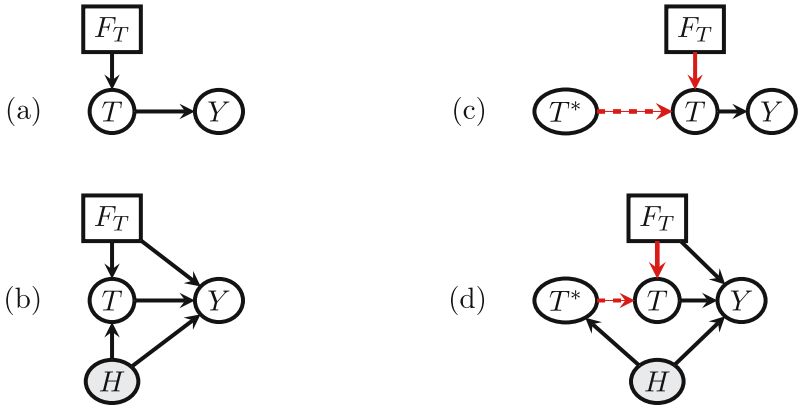

Figure 4: Illustration of the necessity of ITT (aka "natural value of treatment") variable T^* in Dawid's proposal. (a) An augmented DAG (without ITT nodes) corresponding to an observational study without confounding and a perfect intervention on T. (b) An augmented DAG (without ITT nodes) representing an observational study with confounding (H) and a mis-targeted ("fat-hand") intervention affecting both T and Y. If the mis-targeted intervention matches the effect of confounding, then there will be equality of the observational and interventional distributions $p(Y|T=t, F_T=\emptyset) = p(Y|T=t, F_T=t)$ so that the extended independence $Y \perp\!\!\!\perp F_T|T$ will hold, and hence the causal diagram shown in (a) cannot be refuted. The inclusion of T^* resolves this. (c) The DAG with ITT variables corresponding to the study without confounding, this implies $Y \perp\!\!\!\perp F_T, T^*|T$, which is not implied by the ITT augmented DAG (d) when confounding is present.

business depends on the firm achieving this. In order to comply with this, the testing company carry out a non-blinded study and also modify the software in the cholesterol-measuring system to ensure that the results agree with those in the observational study (see Figure 4(b)), here H represents unobserved confounding and the edge $F_T \to Y$ indicates the compromised measurement process.[15] Since the experimental and observational distributions agree, it will hold that $Y \perp\!\!\!\perp F_T \,|\, T$, as implied by the decision-theoretic graph in Figure 4(a).

To be clear, the critique here is *not* that someone who was unaware of the presence of confounding and the devious activities of the company running the trial would infer the wrong causal effect. Rather, it is that without the ITT variables, the decision-theoretic approach lacks the conceptual apparatus necessary to distinguish the situations in Figure 4(a) and (b).[16] In contrast, if the ITT variables T^* are included, then no such difficulty arises: the corresponding augmented DAG, shown in Figure 4(c), now additionally requires that $Y \perp\!\!\!\perp T^* \,|\, F_T = t$, which will fail to hold if there is unobserved confounding between T^* and Y. Note that this latter condition is essentially equivalent to the ignorability condition $Y(t) \perp\!\!\!\perp T^*$ in the potential outcome framework; we return to this point below.

4.2 Dawid's defining ECI relations

Under Dawid's formalism, the augmented graph with ITT variables, shown in Figure 3(d), defines a causal model via the following ECI relations:

$$A: \quad T^* \perp\!\!\!\perp F_T, \tag{25}$$

$$B: \quad Y \perp\!\!\!\perp T^*, F_T \,|\, T, \tag{26}$$

(see [8, Eq. (62), (63)]).

15 Within the potential outcome framework, this would correspond to a failure of consistency, since among people with $T^* = t$, it need not hold that their observed outcome Y is the same as the outcome they would have had, had they been in an experiment and assigned to t, namely $Y(t)$.

16 Here, we are assuming that there is no information available regarding the nature or identity of the possible confounding variables H.

4.2.1 Dawid's independence *A*

The first independence (25) states that whether or not there is an intervention on T has no effect on the (distribution of the) ITT value T^*. Indeed, Dawid states:

> Now, T^* is determined prior to any (actual or hypothetical) treatment application, and behaves as a covariate [...] this distribution is then the same in all regimes [8, Section 8, p.54].

Similarly, in the potential outcome framework, it is assumed that intervention on a treatment variable does not affect variables whose values are realized prior to that intervention, including the natural value of that treatment variable, T^*, so that $T^*(t) = T^*$.

However, Dawid's reference to T^* being a covariate that is *determined prior* to an actual or *hypothetical* treatment application is perhaps surprising: if the value taken by T^* is determined prior to the decision regarding the regime F_T, then this would appear to imply that, in fact, the random variables in the distributions $p(T^* \mid F_T = \varnothing)$ and $p(T^* \mid F_T = t)$ must live on a common probability space. But in this case, it is hard to see why the random variables in the distributions $p(T^*, T, Y \mid F_T = \varnothing)$ and $p(T^*, T, Y \mid F_T = t)$ should not also live on a common probability space! The primary obstacle to so doing appears to be the use of Y and T to indicate what are distinct random variables (corresponding to different regimes) that are defined on the same space. This problem can obviously be overcome by simply using (T^*, T, Y) and $(T^*, T(t), Y(t))$ to refer to the random variables under the idle and intervention regimes, respectively; following Definition 1 in [8], this would imply that $T = T^*$ (under the idle regime) and $T(t) = t$ (under an intervention).

An analyst who adopted this notation is not obligated to impose any additional equalities relating these random variables – such as those implied by consistency – should they not wish to do so. As we did earlier in Section 3, one might choose instead to follow Dawid by merely imposing distributional consistency (see also further discussion below). However, from the perspective of the potential outcome framework, this leads to an unnecessary multiplicity of random variables and more cumbersome notation. For example, in the simple case of a binary treatment, this approach requires three random variables $\{Y, Y(0), Y(1)\}$ corresponding to the response, rather than just two $\{Y(0), Y(1)\}$ with consistency at the level of random variables.[17] It is unclear what is gained by assuming consistency at the level of distributions rather than individuals.

4.2.2 Dawid's independence *B*

The fact that T is a deterministic function of T^* and F_T means that the number of non-trivial conditional independence statements in (26) is not self-evident. A casual reader might imagine that in (26) the pair (F_T, T^*) might take $(|\mathfrak{T}| + 1)|\mathfrak{T}|$ different values for each value of the conditioning variable T. However, given $T = t$, there are only $|\mathfrak{T}| + 1$ possible values for (F_T, T^*):

$$T = t \Rightarrow (F_T, T^*) \in \{(\varnothing, t)\} \cup \{(t, s) : s \in \mathfrak{T}\},$$

since either we are in the idle regime, $F_T = \varnothing$ and $T = T^*$, or we are in the interventional regime, in which case $F_T = t$ and T^* may take any value. Thus, given $T = t$, (26) corresponds to a set of $|\mathfrak{T}|$ equalities:

$$p(Y \mid T^* = t, F_T = t, T = t) = p(Y \mid T^* = t, F_T = \varnothing, T = t), \tag{27}$$

$$p(Y \mid T^* = t, F_T = t, T = t) = p(Y \mid T^* = s, F_T = t, T = t), \quad \text{for } s \neq t. \tag{28}$$

4.2.3 Distributional consistency in *B*

Equation (27) corresponds to distributional consistency, which [8, eq. (14)] defines as:

17 In other words, with Y given by a deterministic function, so $Y = (1 - T)Y(0) + TY(1)$.

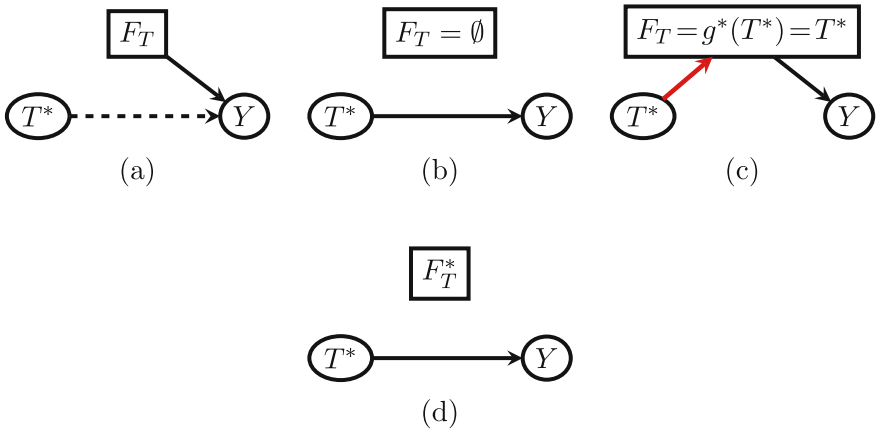

Figure 5: Encoding distributional consistency via a special dynamic regime in the setting of a reformulated decision diagram (having marginalized the intervention target "T"). (a) Reformulated augmented graph $\mathcal{G}^*$ representing the observed joint distribution $p(T^*, Y \,|\, F_T)$; (b) augmented graph $\mathcal{G}^*$ corresponding to $p(T^*, Y \,|\, F_T = \emptyset)$; (c) augmented graph $\mathcal{G}^*$ corresponding to $p(T^*, Y \,|\, F_T = g^*)$; corresponding to the dynamic "regime" that "intervenes" on the target setting it to the natural value T^*; (d) a graph illustrating distributional consistency (31); here, F_T^* is a special regime indicator taking only the values $\emptyset$ and g^*. The graph (d) encodes the distributional consistency assumption: the distribution over Y and T^* resulting from the "intervention" g^* is identical to having no intervention.

$$p(Y \,|\, F_T = \emptyset, T = t) = p(Y \,|\, T^* = t, F_T = t). \tag{29}$$

Dawid notes that this implies:

$$Y \perp\!\!\!\perp F_T \,|\, T^*, T \tag{30}$$

(see [8, Lemma 1]). However, this formulation also somewhat obscures the actual number of constraints: if $T^* \neq T = t$, then $F_T = t$ so that the statement becomes trivial, while if $T^* = T = t$, then F_T only takes two possible values $\emptyset$ and t. Given this, it becomes clear that (30) may be reformulated by defining a dynamic regime g^* that "intervenes" to set T to be T^*. By defining a special regime indicator, denoted F_T^*, that takes only two values $\emptyset$ or g^*, we can re-express (30) as:

$$Y, T^* \perp\!\!\!\perp F_T^*. \tag{31}$$

Note that in so doing, we do not need to refer to T[18] (see Figure 5(d) for a graphical depiction).

In terms of potential outcomes, the independence (31) may be expressed as:

$$p(Y, T^* = t) = p(Y, T^* = t \,|\, F_T^* = \emptyset) = p(Y, T^* = t \,|\, F_T^* = g^*) = p(Y(t), T^*(t) = t), \tag{32}$$

which corresponds to distributional consistency (see (4)).

4.2.4 Ignorability in B

Equation (28) expresses the property of ignorability, which Dawid [8, eq. (20)] expresses as:

$$Y \perp\!\!\!\perp T^* \,|\, F_T, T. \tag{33}$$

However, as Dawid himself notes, given $T = t$, then either $F_T = \emptyset$ in which case $T^* = t$ (and independence holds trivially), or $F_T = t$, so that this constraint is identical to:

$$Y \perp\!\!\!\perp T^* \,|\, F_T = t \quad \text{for } t \in \mathfrak{T} \tag{34}$$

(see Figure 3(f) and (g)). Equivalently in terms of potential outcomes,

18 Along similar lines, in his equation (14) [8] notes that the LHS of (29) is equivalent to $p(Y \,|\, F_T = \emptyset, T^* = t)$.

Figure 6: (a) Reformulated augmented graph $\mathcal{G}^*$ representing the observed joint distribution $p(T^*, Y|F_T)$; (b) graph illustrating that, if desired, the "applied treatment" variable T may be added to $\mathcal{G}^*$ since it is a deterministic function of T^* and F_T. Note that although it may seem counterintuitive that T is not a parent of Y in this graph, this is formally correct.

$$Y(t) \perp\!\!\!\perp T^* \quad \text{for } t \in \mathfrak{T} \tag{35}$$

(see Figure 3(b)). Again, we note that T is not required for the purpose of expressing this condition.

4.3 Simplification

From a graphical perspective, it is perhaps natural to wish to express the invariance of the distribution of Y given T across observational and interventional distributions by examining whether a regime indicator F_T is d-separated from Y given T. However, as we have seen, it is necessary to include what Dawid calls the ITT variable (aka the natural value of treatment) T^* in order to rule out cases of spurious invariance. Furthermore, T^* plays a central role in certain notions, such as the effect of treatment on the treated, that are widely used in many studies that apply the potential outcome framework.

As shown previously, there is no need to condition on T when describing the defining independences, and in fact doing so arguably obscures the nature of the specific assumption being made. This suggests that T should be marginalized from the ITT augmented graph, rather than T^* as Dawid proposes. Note that, if we distinguish the cases $F_T = \varnothing$ and $F_T = t$, the resulting graphs (modulo labeling) are isomorphic to those used in the SWIG framework (compare Figure 3(a) to (e), and (b) to (f)).

We carry out this reformulation in full generality in the next section.

5 Reformulation of decision graphs

Our proposed reformulation of decision graphs follows a strategy similar to that used for SWIGs. In contrast, Dawid aims to give ECI relations that, together with the usual independence relations over the observed variables, will yield the Markov property for the augmented decision graph with ITT variables. As an alternative, we begin by defining a Markov property associated with the augmented decision graph, and then, using distributional consistency we derive the usual observed conditional independences.[19]

It should be noted that Dawid's independences do not actually imply the full Markov property for the ITT graph because, as noted by the presence of a dashed edge, there are context-specific independences implied by the graph.[20] However, these are not implied by the independence relations A and B. (To see this, note that the conditions A and B would also hold for a decision DAG with the same structure, but in which T was not a deterministic function of T^* and F_T, in which case the context-specific independence relations would not hold.)

19 However, see Appendix A.1 for an alternative reformulation that, similar to Dawid's approach, starts from the assumption that the observed distribution $p(V)$ obeys the usual Markov property for $\mathcal{G}$.

20 Note that in Figure 12 in [8], there is an edge $Z \to X_1$ that is not dashed, but which it appears should be dashed: note that the relationship of Z and H to X_1 are symmetric and the edge $H \to X_1$ is dashed.

Note also that these extra ECI relations are not restricted solely to those involving T. Consider, for example, the front door graph shown in Figure 7(a). Since Dawid's augmented decision diagram, shown in Figure 7(c), includes a dashed edge from T^* to T, indicating that this edge should be removed conditional on $F_T = t$, the diagram implies that Y will be d-separated from F_T given M and $F_T \neq \varnothing$. However, even though it is encoded in the augmented graph, the corresponding ECI:

$$Y \perp\!\!\!\perp F_T \,|\, M, F_T \neq \varnothing \tag{36}$$

does not follow from the independences $A + B$.

In the potential outcome framework, the constraint (36) corresponds to:

$$p(Y(t) \,|\, M(t)) = p(Y(t^*) \,|\, M(t^*)). \tag{37}$$

This constraint is naturally encoded by the d-separation of $Y(t)$ from the fixed variable t given $M(t)$ on the SWIG $\mathcal{G}(t)$ shown in Figure 7(b) (see [7,10,12]).

As these examples suggest, in order to capture the full Markov structure of the augmented decision diagram, including those constraints corresponding to dashed edges, it is natural to use the constraints implied by the decision diagram when no regime indicators are idle, which we express in shorthand as $F_A \neq \varnothing$; graphically, this corresponds to removing (temporarily) all of the dashed edges. We show below that the independences encoded then imply, via distributional consistency, the Markov property for the observed data that is encoded in the original graph.

Another advantage of this approach is that we will only require the ITT variables T^*; the "applied treatment," which Dawid [8] denotes "T," will not be required.[21]

Specifically, consider a set of variables $V_1, \ldots, V_p$. Let $A \subset \{1, \ldots, p\}$ be the (index) set of the targets of intervention. If $i \in A$, then let V_i be the corresponding ITT variable (which Dawid denotes by X_i^*). Thus, the set $V_1, \ldots, V_p$ consists of ITT variables as well as variables that are not in A and hence not targets of intervention.[22] Thus, under the regime where every intervention target has been intervened upon, so that $F_i \neq \varnothing$ for all $i \in A$, the variables in $V_1, \ldots, V_p$ correspond to the random variables in the SWIG $\mathcal{G}(a)$.

For every intervention target $i \in A$, let g_i^* denote the dynamic regime that "intervenes" to set the intervention target to its natural value V_i. Let F_i^* be a regime indicator taking the states $\varnothing$ or g_i^*.

Definition 13. (Distributional consistency for decision diagrams) The kernel $p(V \mid F_A)$ is said to obey distributional consistency if, given $B_i \in A$ and $C \subseteq A \backslash \{B_i\}$, where C may be empty,

$$V \perp\!\!\!\perp F_i^* \,|\, F_C, F_C \neq \varnothing, \tag{38}$$

where we use the shorthand $F_C \neq \varnothing$ to indicate that for all $j \in C$, $F_j \neq \varnothing$.

Note that, taking $Y = V \backslash \{B_i\}$, (38) is equivalent to the following equality, which corresponds exactly to (3):

$$\begin{aligned} p(Y = y, B_i = b \,|\, F_i = b, F_C = c) &= p(Y = y, B_i = b \,|\, F_i^* = g_i^*, F_C = c, F_C \neq \varnothing) && (39) \\ &= p(Y = y, B_i = b \,|\, F_i^* = \varnothing, F_C = c, F_C \neq \varnothing) \\ &= p(Y = y, B_i = b \,|\, F_i = \varnothing, F_C = c) && (40) \\ &= p(Y = y, B_i = b \,|\, F_C = c). \end{aligned}$$

Here, the second equality follows from (38), while the first and third equalities are via the definition of F^* and $\varnothing$.

As observed by Dawid, in place of Definition 13, we could instead have defined distributional consistency, without reference to the dynamic regime g_i^*, by simply equating (39) and (40). We have chosen to

21 Since T is a deterministic function of T^* and F_T, it is possible to add back in these variables if we wish (see Figure 6).

22 In this formulation, we will not use the post-intervention target variables, which Dawid denotes, X_i.

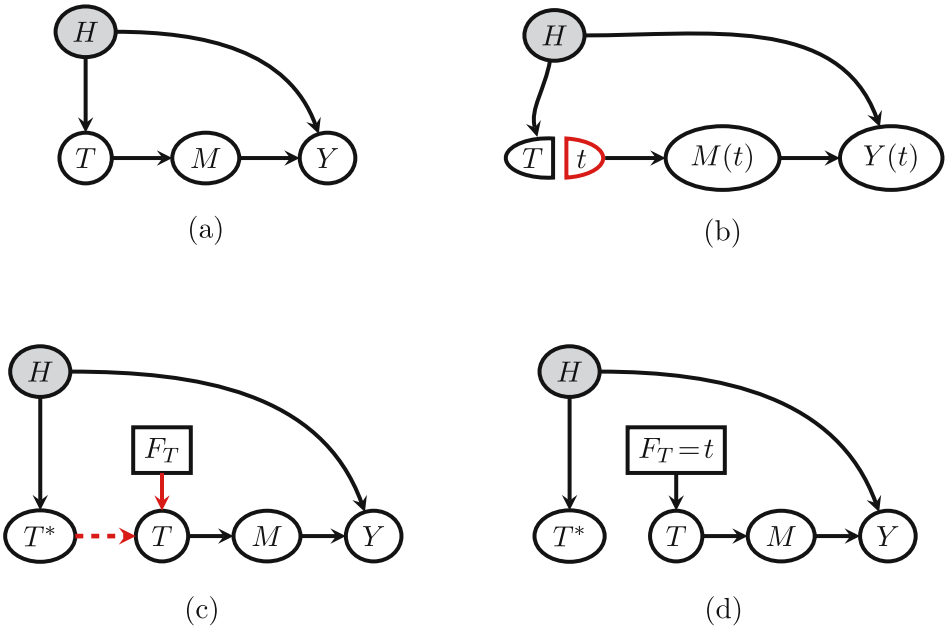

Figure 7: (a) Front-door graph $\mathcal{G}$; (b) the SWIG $\mathcal{G}(t)$ (with ancestral labeling); (c) the augmented decision diagram $\mathcal{G}^*$; and (d) the augmented decision diagram given $F_T = t$ in which the dashed edge from T^* to T is removed. Note that in (d) F_T is d-separated from Y given M. However, the corresponding extended independence, $Y \perp\!\!\!\perp F_T | M, F_T \neq \emptyset$, is not implied by Dawid's conditions A+B.

make use of g_i^* in order to emphasize what we see as the tautological nature of distributional consistency, while also formulating condition (38) as a conditional independence.

The following four Lemmas are reformulations of Lemmas 3–6 in the decision diagram framework. Though the proofs are largely translations of those lemmas, we include them here for completeness.

Lemma 14. *If $p(V \mid F_A)$ obeys distributional consistency, B and C are disjoint subsets of A, where C may be empty, then for all y, b, and c:*

$$V \perp\!\!\!\perp F_B^* \mid F_C, F_C \neq \emptyset, \tag{41}$$

where F_B^* is the set $\{F_i^*, i \in B\}$.

Proof. This follows by induction on the size of B. □

Lemma 15. *Let B and C be disjoint subsets of A, where C may be empty, and let Y and W be disjoint subsets of $V \setminus B$, then distributional consistency implies:*

$$Y \perp\!\!\!\perp F_B^* \mid B, W, F_C, F_C \neq \emptyset. \tag{42}$$

Proof. This follows by applying (extended) graphoid axioms to (41). □

Lemma 16. *Let B and C be disjoint subsets of A, where C may be empty. If $B \subseteq W$, then under distributional consistency:*

$$W \perp\!\!\!\perp F_B \mid F_C, F_{B \cup C} \neq \emptyset \quad \Rightarrow \quad W \perp\!\!\!\perp F_B \mid F_C, F_C \neq \emptyset. \tag{43}$$

Proof. Let $b \in \mathfrak{X}_B$, $c \in \mathfrak{X}_C$, and $w \in \mathfrak{X}_W$. Given the LHS of (43), it is sufficient to prove $p(X_W \mid F_B = b, F_C = c) = p(X_W \mid F_B = \emptyset, F_C = c)$. Now:

$$\begin{aligned} p(X_W = w \mid F_B = b, F_C = c) &= p(X_{W\setminus B} = w_{W\setminus B}, X_B = w_B \mid F_B = b, F_C = c) \\ &= p(X_{W\setminus B} = w_{W\setminus B}, X_B = w_B \mid F_B = w_B, F_C = c) \\ &= p(X_{W\setminus B} = w_{W\setminus B}, X_B = w_B \mid F_B^* = g_B^*, F_C = c) \\ &= p(X_{W\setminus B} = w_{W\setminus B}, X_B = w_B \mid F_B^* = \varnothing, F_C = c) \\ &= p(X_{W\setminus B} = w_{W\setminus B}, X_B = w_B \mid F_B = \varnothing, F_C = c). \end{aligned}$$

The second equality uses the premise of (43), the third is by definition of g_B^*, and the fourth is distributional consistency via Lemma 14. □

Lemma 17. *Let B and C be disjoint subsets of A, where C may be empty. Let Y and W be disjoint sets with $B \subseteq W$, then under distributional consistency*:

$$Y \perp\!\!\!\perp F_B \mid W, F_C, F_{B\cup C} \neq \varnothing \quad \Rightarrow \quad Y \perp\!\!\!\perp F_B \mid W, F_C, F_C \neq \varnothing. \tag{44}$$

Proof. Similar to the proof of Lemma 16, given the premise in (44), it suffices to show that $p(X_Y \mid X_W, F_B = b, F_C = c) = p(X_Y \mid X_W, F_B = \varnothing, F_C = c)$.

$$\begin{aligned} p(X_Y \mid X_W = w, F_B = b, F_C = c) &= p(X_Y \mid X_{W\setminus B} = w_{W\setminus B}, X_B = w_B, F_B = b, F_C = c) \\ &= p(X_Y \mid X_{W\setminus B} = w_{W\setminus B}, X_B = w_B, F_B = w_B, F_C = c) \\ &= p(X_Y \mid X_{W\setminus B} = w_{W\setminus B}, X_B = w_B, F_B^* = g_B^*, F_C = c) \\ &= p(X_Y \mid X_{W\setminus B} = w_{W\setminus B}, X_B = w_B, F_B = \varnothing, F_C = c). \end{aligned}$$

As in the previous proof, the second equality uses the premise of (44), the third is by definition of g_B^*, and the fourth is distributional consistency via Lemma 14. □

5.1 Reformulated augmented decision diagrams

Let $\mathcal{G}$ be a DAG with a topologically ordered vertex set $V = \{1, \ldots, p\}$ representing an observed distribution $p(W_V)$.[23] Let $A \subseteq V$ be the subset of vertices for which interventions are well defined, let $F = \{F_i, i \in A\}$ be the corresponding set of regime indicators. Let $\mathcal{G}^*$ be the extended DAG with vertex set $V \cup F$, representing the kernels $p(W_V \mid F_A)$. As before, we use $\mathrm{pa}(i)$ to indicate the (index) set of the variables that are the parents of W_i in the original DAG $\mathcal{G}$, and let $\mathrm{pre}(i)$ denote $\{1, \ldots, i-1\}$, the predecessors of i under a total ordering $\prec$ consistent with $\mathcal{G}$.[24]

Definition 18. The kernel $p(W_V \mid F_A)$ will be said to obey the augmented DAG local Markov property for the DAG $\mathcal{G}^*$ if for all $i \in V$:

$$W_i \perp\!\!\!\perp F_{A\setminus \mathrm{pa}(i)}, W_{\mathrm{pre}(i)\setminus(\mathrm{pa}(i)\setminus A)} \mid W_{\mathrm{pa}(i)\setminus A}, F_{A\cap \mathrm{pa}(i)}, F_A \neq \varnothing, \tag{45}$$

where $F_{A\cap \mathrm{pa}(i)} \neq \varnothing$ is a shorthand for $F_j \neq \varnothing$ for all $j \in A \cap \mathrm{pa}(i)$.

---

23 In this section of this article, when we wish to distinguish random variables from index sets, we use W_i, rather than X_i. This is because in Dawid's development, X is reserved to denote intervention targets. However, we depart from Dawid's notation in that we will not use an asterisk to indicate ITT variables, this is because we will only include the ITT variables associated with intervention targets on the reformulated diagram.

24 Note that the vertex set for $\mathcal{G}^*$ corresponds to the sets of variables that in Dawid's notation would be written $(V_i, X_1^*, \ldots, V_k, X_k^*, V_{k+1})$; in other words, it consists of domain variables and ITT variables associated with intervention targets. We will have no need to include what Dawid calls "the intervention targets," which he denotes X_i, in our reformulated decision diagram, though they may be added (see Figure 6).

This formulation captures the Markov property necessary for the augmented diagram including the context-specific independences that arise from interventions (that are not captured directly in Dawid's $A + B$ formulation).

Note that this property follows from d-separation applied to the graph in which we intervene on every vertex in A. We will show that under distributional consistency, this property implies factorization of the observed distribution with respect to the original graph.

However, it is useful first to further decompose the sets on the RHS of the independence. Specifically, we divide the regime indicators that are not the parents of i into those that occur after i and those that are prior to i:

$$F_{A \setminus \mathrm{pa}(i)} = (F_{A \setminus \mathrm{pre}(i)}, F_{(A \cap \mathrm{pre}(i)) \setminus \mathrm{pa}(i)}).$$

Similarly, we divide the set of random variables that are prior to i and either in A or not parents of i into those that are not parents and those that are the parents that are in A:

$$W_{\mathrm{pre}(i) \setminus (\mathrm{pa}(i) \setminus A)} = (W_{\mathrm{pre}(i) \setminus \mathrm{pa}(i)}, W_{\mathrm{pa}(i) \cap A}).$$

Thus, independence (45) becomes:

$$W_i \perp\!\!\!\perp \overbrace{F_{A \setminus \mathrm{pre}(i)}}^{\text{time order}}, \overbrace{F_{(A \cap \mathrm{pre}(i)) \setminus \mathrm{pa}(i)}}^{\text{causal Markov prop.}}, \overbrace{W_{\mathrm{pre}(i) \setminus \mathrm{pa}(i)}}^{\text{assoc. Markov prop.}}, \overbrace{W_{\mathrm{pa}(i) \cap A}}^{\text{ignorability}} \mid \underbrace{F_{A \cap \mathrm{pa}(i)}}_{\text{fixed parents}}, \underbrace{W_{\mathrm{pa}(i) \setminus A}}_{\text{random parents}}, \underbrace{F_A \neq \varnothing}_{\text{intervene on all of } A}. \tag{46}$$

Consequently, independence (46) captures the following:

- Later interventions have no effect on earlier distributions (time order).
- Given intervention on all earlier targets, the specific value of an intervention does not affect the distribution of a variable given its non-intervened parents unless the intervened on variable is itself a parent (causal Markov property).
- Independence from earlier random variables given non-intervened parents (associational Markov property).
- An intervention on a parent of a variable renders that variable independent of the natural value of the intervention target conditional on its other non-intervened parents (ignorability).

5.2 Example

In Table 4, we show the reformulated decision diagram Markov property corresponding to the augmented DAG $\mathcal{G}^*$, as shown in Figure 2(c). Note that the local property here corresponds naturally to the graph $\mathcal{G}^*$ under the regime $F_0 = x_0$, $F_1 = x_1$ displayed in Figure 2(d). In particular, note that for each random vertex, the size of the conditioning set in the defining independence (ignoring the term $F_{01} \neq \varnothing$) is equal to the number of parents that the vertex has in Figure 2(d).

Table 4: Defining properties for reformulated decision diagram corresponding to Figure 2, under the ordering (H, X_0, Z, X_1, Y)

Reformulated decision diagram local property				
H	$\perp\!\!\!\perp$	F_0, F_1	$\mid$	$F_{01} \neq \varnothing$
X_0	$\perp\!\!\!\perp$	H, F_0, F_1	$\mid$	$F_{01} \neq \varnothing$
Z	$\perp\!\!\!\perp$	X_0, F_1	$\mid$	$H, F_0, F_{01} \neq \varnothing$
X_1	$\perp\!\!\!\perp$	X_0, F_0, F_1	$\mid$	$H, Z, F_{01} \neq \varnothing$
Y	$\perp\!\!\!\perp$	H, X_0, X_1, F_0	$\mid$	$Z, F_1, F_{01} \neq \varnothing$

Here, as elsewhere, $F_{01} \neq \varnothing$ is a shorthand for $(F_0 \neq \varnothing \ \& \ F_1 \neq \varnothing)$.

5.3 Consequences of the local Markov property

Lemma 19. *If the kernel $p(W_V \mid F_A)$ obeys distribution consistency and the augmented DAG local Markov property w.r.t. $\mathcal{G}^*$, then*:

$$p(W_i \mid W_{\mathrm{pre}(i)}, F_A = a) \tag{47}$$

$$= p(W_i \mid W_{\mathrm{pre}(i)}, F_{\mathrm{pre}(i)\cap A} = a_{\mathrm{pre}(i)\cap A}) \tag{48}$$

$$= p(W_i \mid W_{\mathrm{pre}(i)}, F_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}) \tag{49}$$

$$= p(W_i \mid W_{\mathrm{pa}(i)}, F_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}) \tag{50}$$

$$= p(W_i \mid W_{\mathrm{pa}(i)\setminus A}, F_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}). \tag{51}$$

Proof. Here, (48) follows from Lemma 16 since by Definition 18,

$$W_{\mathrm{pre}(i)\cup\{i\}} \perp\!\!\!\perp F_{A\setminus \mathrm{pre}(i)} \mid F_{A\cap \mathrm{pre}(i)}, F_A \neq \varnothing.$$

Similarly, (49) follows from Lemma 17 since by the local Markov property:

$$W_i \perp\!\!\!\perp F_{(A\cap \mathrm{pre}(i))\setminus \mathrm{pa}(i)} \mid F_{A\cap \mathrm{pa}(i)}, F_{A\cap \mathrm{pre}(i)} \neq \varnothing.$$

Finally, (50) and (51) again follow from the local Markov property since

$$W_i \perp\!\!\!\perp W_{\mathrm{pre}(i)\setminus \mathrm{pa}(i)}, W_{\mathrm{pa}(i)\cap A} \mid W_{\mathrm{pa}(i)\setminus A}, F_A = a,$$

hence $p(W_i \mid W_{\mathrm{pre}(i)} = w, F_A = a)$ does not depend on $w_{\mathrm{pre}(i)\setminus(\mathrm{pa}(i)\setminus A)} = (w_{\mathrm{pre}(i)\setminus \mathrm{pa}(i)}, w_{\mathrm{pa}(i)\cap A})$. □

5.4 Markov property for the observed distribution

The following result shows that the reformulated local Markov property implies, via distributional consistency, the ordinary local Markov property for the observed distribution. This result corresponds to Theorem 10.

Theorem 20. *If the kernel $p(W_V \mid F_A)$ obeys distribution consistency and the augmented DAG local Markov property w.r.t. $\mathcal{G}^*$, then $p(W_V)$ obeys the usual local Markov property w.r.t. $\mathcal{G}$.*

Proof. Let $w^* \in \mathcal{X}_{\mathrm{pre}(i)}$.

$$\begin{aligned} p(W_i = w \mid W_{\mathrm{pre}(i)} = w^*) &= p(W_i = w \mid W_{\mathrm{pre}(i)} = w^*, F_{\mathrm{pre}(i)\cap A} = w^*_{\mathrm{pre}(i)\cap A}) \\ &= p(W_i = w \mid W_{\mathrm{pa}(i)\setminus A} = w^*_{\mathrm{pa}(i)\setminus A}, F_{\mathrm{pa}(i)\cap A} = w^*_{\mathrm{pa}(i)\cap A}). \end{aligned} \tag{52}$$

Here, the first equality follows by distributional consistency. The second follows directly from the equality of (48) and (51) in Lemma 19. Since the last line only depends on $w^*_{\mathrm{pa}(i)}$, the ordered local Markov property for the DAG holds. □

5.5 Identifiability

The next result shows that the reformulated local Markov property implies that the kernel $p(V \mid F_A)$ will be identified from the distribution of the observables provided that the relevant conditional distributions are identified (from the distribution of the observables). This result corresponds to Theorem 11.

Theorem 21. *Suppose the kernel $p(W_V \mid F_A)$ obeys distribution consistency and the augmented DAG local Markov property w.r.t. $\mathcal{G}^*$. Let a be an assignment to the intervention targets in A, and let v be an assignment to W_V. Then, for every i*:

$$p(W_i = v_i \mid F_A = a, W_{\mathrm{pre}(i)} = v_{\mathrm{pre}(i)}) = p(W_i = v_i \mid W_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}, W_{\mathrm{pa}(i)\setminus A} = v_{\mathrm{pa}(i)\setminus A}). \tag{53}$$

Consequently, $p(W_V \mid F_A = a)$ is identified given $p(W_V)$ and obeys the Markov property for the DAG formed from $\mathcal{G}^*$ by removing all outgoing edges from vertices in A.

As before, we note that equality (53) corresponds to the property referred to as "modularity" in the SWIG formulation, which is also an instance of the extended g-formula of [2,14].

Proof. Let $a \in \mathfrak{X}_A$, $v \in \mathfrak{X}_V$. Now:

$$\begin{aligned} & p(W_i = v_i \mid W_{\mathrm{pre}(i)} = v_{\mathrm{pre}(i)}, F_A = a) \\ &= p(W_i = v_i \mid W_{\mathrm{pa}(i)\cap A} = v_{\mathrm{pa}(i)\cap A}, W_{\mathrm{pa}(i)\setminus A} = v_{\mathrm{pa}(i)\setminus A}, F_{A\cap \mathrm{pa}(i)} = a_{\mathrm{pa}(i)\cap A}) \\ &= p(W_i = v_i \mid W_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}, W_{\mathrm{pa}(i)\setminus A} = v_{\mathrm{pa}(i)\setminus A}, F_{A\cap \mathrm{pa}(i)} = a_{\mathrm{pa}(i)\cap A}) \\ &= p(W_i = v_i \mid W_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}, W_{\mathrm{pa}(i)\setminus A} = v_{\mathrm{pa}(i)\setminus A}). \end{aligned} \tag{54}$$

Here, the first equality follows from the equality of (47) and (50); the second follows from the equality of (50) and (51); the third by distributional consistency. □

5.6 Distributions resulting from fewer interventions

As in the SWIG case, a similar argument applies if we consider interventions on a subset $B \subseteq A$. This result corresponds to Theorem 12.

Theorem 22. *Suppose the kernel $p(W_V \mid F_A)$ obeys distribution consistency and the augmented DAG local Markov property w.r.t. $\mathcal{G}^*$. Let b be an assignment to the intervention targets in $B \subseteq A$, and let w^* be an assignment to W_V. Then, for every i:*

$$p(W_i = w_i^* \mid F_B = b, W_{\mathrm{pre}(i)} = w^*_{\mathrm{pre}(i)}) = p(W_i = w_i^* \mid W_{\mathrm{pa}(i)\cap B} = b_{\mathrm{pa}(i)\cap B}, W_{\mathrm{pa}(i)\setminus B} = w^*_{\mathrm{pa}(i)\setminus B}). \tag{55}$$

Consequently, $p(W_V \mid F_B = b)$ is identified given $p(W_V)$ and obeys the Markov property for the augmented DAG $\mathcal{G}^{**}$ formed from $\mathcal{G}^*$ by removing all outgoing edges from vertices in B and removing the regime indicators $F_{A\setminus B}$.

Proof.

$$\begin{aligned} & p(W_i = w_i \mid W_{\mathrm{pre}(i)} = w_{\mathrm{pre}(i)}, F_B = b) \\ &= p(W_i = w_i \mid W_{\mathrm{pre}(i)} = w_{\mathrm{pre}(i)}, F_{\mathrm{pre}(i)\cap B} = b_{\mathrm{pre}(i)\cap B}) \\ &= p(W_i = w_i \mid W_{\mathrm{pre}(i)} = w_{\mathrm{pre}(i)}, F_{\mathrm{pre}(i)\cap B} = b_{\mathrm{pre}(i)\cap B}, F_{\mathrm{pre}(i)\cap (A\setminus B)} = w_{\mathrm{pre}(i)\cap (A\setminus B)}) \\ &= p(W_i = w_i \mid W_{\mathrm{pa}(i)\setminus A} = w_{\mathrm{pa}(i)\setminus A}, W_{\mathrm{pa}(i)\cap B} = b_{\mathrm{pa}(i)\cap B}, W_{\mathrm{pa}(i)\cap (A\setminus B)} = w_{\mathrm{pa}(i)\cap (A\setminus B)}) \\ &= p(W_i = w_i \mid W_{\mathrm{pa}(i)\setminus B} = w_{\mathrm{pa}(i)\setminus B}, W_{\mathrm{pa}(i)\cap B} = b_{\mathrm{pa}(i)\cap B}). \end{aligned}$$

Here, the first equality is by Lemma 16; the second is distributional consistency; the third follows from Theorem 21 applied to $\mathcal{G}^*$; the fourth is a simplification. □

6 The role of "fictitious" independence in Dawid's development

Dawid in [8] uses what he terms a "fictitious" independence in his proofs that the distribution of the kernels that condition on the regime indicators F_i obey the Markov property for the augmented DAG with ITT variables. Specifically, in his proof of Lemma 4, though not the statement, he makes the formal assumption that

$$F_1 \perp\!\!\!\perp F_0, \tag{56}$$

and similarly in the proof of Theorem 1, he assumes that all the regime indicators are mutually independent [8, p. 76, eqn. (82)]

$$F_1 \perp\!\!\!\perp F_2 \perp\!\!\!\perp \cdots \perp\!\!\!\perp F_{k-1} \perp\!\!\!\perp F_k. \tag{57}$$

Such an independence assumption does not fit into the ECI framework used by Dawid to describe the Markov property for augmented graphs. This is because, as stated by Dawid, an ECI statement $A \perp\!\!\!\perp B \mid C$ must satisfy: "(a) no non-stochastic variable occurs in A, and (b) all non-stochastic variables are included in $B \cup C$" [8, fn. 3]; these conditions allow independences to be viewed as well defined restrictions on $p(A \mid B, C)$ since all of the non-stochastic variables appear on the right of the conditioning bar. However, an independence of the form $F_i \perp\!\!\!\perp F_j$ violates both of these conditions.

Perhaps for this reason, Dawid argues that although his proofs make use of the assumptions (56) and (57), there is no loss of generality:

> So long as all our assumptions and conclusions are in the form described in footnote 3 [i.e., satisfy (a) and (b)], any proof that uses this extended understanding only internally will remain valid [...] [8, p. 63]
>
> [...] because the premisses and conclusions of the argument relate only to distributions conditioned on the regime indicators, the extra assumption of variation independence is itself inessential, and can be regarded as just another "trick." [8, p. 63, fn.24]

We will show via an example that Dawid's inference here is not valid: in general, the conclusion will not hold for a kernel without additional assumptions regarding the set of states taken by the non-stochastic variables. However, notwithstanding this, as we also show below, Dawid's conclusions are still correct owing to the special structure that is present in the possible states taken by regime indicators.

6.1 Invalid implication

To illustrate the issue with the proof, we re-write Dawid's equations so as to make the argument transparent. Dawid makes the following claim:

Claim 23. Consider a kernel $q(x, y \mid a, b)$, with stochastic variables X and Y and non-stochastic variables A and B. If the following ECI restrictions hold:

$$Y \perp\!\!\!\perp A \mid B, X, \tag{58}$$

$$Y \perp\!\!\!\perp B \mid A, X, \tag{59}$$

then it follows that:

$$Y \perp\!\!\!\perp A, B \mid X. \tag{60}$$

To relate this to Dawid's proof of Lemma 4 in [8], $A = F_0$, $B = F_1$, $X = X_0$, and $Y = \{H, Z, X_1^*\}$. Thus (56), (58), (59), and (60) correspond to Dawid's equations (41), (49), (50), and (51), respectively.

In the proof of this claim, Dawid makes use of the "fictitious" independence $A \perp\!\!\!\perp B$, but as noted above, he argues that this "internal" assumption may be made without loss of generality. To see that this implication does not hold without additional conditions on the state spaces for A and B, suppose that the non-stochastic pair $(A, B) \in \mathfrak{S} \equiv \{-2, -1\}^2 \cup \{1, 2\}^2$, so that (A, B) take one of the following eight states:

$$(-2, -2), (-2, -1), (-1, -2), (-1, -1), (1, 1), (1, 2), (2, 1), (2, 2).$$

Note that, by construction, the non-stochastic variables are not variation independent; they always share the same sign. Now, let $P^-(Y, X)$ and $P^+(Y, X)$ be any pair of distributions over (X, Y) such that $P^+(Y \mid X) \neq P^-(Y \mid X)$ and define the kernel $p(Y, X \mid a, b)$ for $(a, b) \in \mathfrak{S}$ as follows:

$$p(Y, X \mid a, b) = \begin{cases} p^-(Y, X) & \text{if both } a, b < 0, \\ p^+(Y, X) & \text{if both } a, b > 0, \\ \text{undefined} & \text{otherwise}. \end{cases} \tag{61}$$

By construction, if $B = b < 0$, then for all a such that $(a, b) \in \mathfrak{S}$, i.e., $a \in \{-2, -1\}$, it holds that

$$p(Y \mid A = a, B = b, X) = p^{-}(Y \mid X), \tag{62}$$

so (58) holds when $B = b < 0$. The argument when $B = b > 0$ is symmetric, since in this case for all a such that $(a, b) \in \mathfrak{S}$, we have $p(Y \mid A = a, B = b, X) = p^{+}(Y \mid X)$. Hence, (58) holds for all $B \in \{-2, -1, 1, 2\}$. A symmetric argument replacing A with B shows that (59) also holds.

However, the conclusion (60) fails since by construction:

$$p(Y \mid A = -2, B = -2, X) = p^{-}(Y \mid X) \neq p^{+}(Y \mid X) = p(Y \mid A = 2, B = 2, X). \tag{63}$$

The implication in the claim corresponds to an ECI instance of the Intersection Axiom (CI5) introduced by Dawid in his classic articles [17,18]. As he notes in several places in [8], this implication is well known not to hold in general.[25] Our aforementioned counterexample simply serves to show that even though there is no distribution over the non-stochastic variables, the implication will not hold if the non-stochastic variables are not variation independent.

6.2 Validity of the conclusion for regime indicators

That the implications used in Dawid's proofs of Lemma 4 and Theorem 1 do not hold – without conditions on the joint space for the regime indicators – may at first seem to call into question Dawid's conclusions. However, at least in causal theories making use of DAG representations and involving multiple treatments, the decisions as to whether to intervene, and if so, which value to enforce are unconstrained. Consequently, variation independence will hold, and hence, the conclusion will be valid.

However, there are situations in which interventions may be constrained. For example, suppose that there are two strategies for a medical condition; each treatment involves two separate stages (A_1 and A_2). At time $t = 1$, the doctor must decide between strategies "1," "2." It is easy to imagine situations in which, if treatment was commenced at time 1, the treatment at time 2 involves "completing" the treatment that was started at time 1, for example, removing surgical stitches from the specific operation performed at time 1. In this case, the treatment options available at time 2 are constrained by the decision at time 1.

Reflecting this, there have been causal decision theories proposed in which variables do not live in a product space (see [22]). Likewise, in the potential outcome framework, the formulation of causally interpreted structured tree graphs given by [2] also allows for this possibility.

However, even in this case, Dawid's implication will still hold, provided that the following condition obtains.

Definition 24. Let $\mathfrak{F}_A \subsetneq \times_{i \in A} \mathfrak{X}_i \cup \{\varnothing\}$ indicate the (constrained) state-space for the set of regime indicators F_A.

$$\begin{aligned} &\text{For all } f \in \mathfrak{F}_A, \quad \text{s.t. } f \neq \varnothing \\ &\quad \Rightarrow \text{there exists } i \in A \text{ s.t. } f_i \neq \varnothing \text{ and } (f_{-i}, \varnothing) \in \mathfrak{F}_A, \end{aligned} \tag{64}$$

where f_{-i} indicates the values assigned to $A \backslash \{i\}$ by f.

In words, this states that for any possible setting of the regime indicators, in which they are not all "idle," there exists some intervention target A_i that is intervened upon under f, which could have not been intervened upon, such that the resulting vector $(f_{-i}, \varnothing)$ is still a valid value for F_A.

This condition may still hold in settings in which, if a later target is intervened upon, the regime under which an earlier target is set to "idle" is not well defined. For example, an intervention on A_2 setting $F_2 = 1$

25 For recent related work giving general conditions under which this implication holds for ordinary (not extended) conditional independence, see [19], [20, Ch.4], [21].

may only be well defined if $F_1 = 1$, but not $F_1 = \varnothing$. In the aforementioned treatment completion example this would be the case if, in the absence of an intervention on A_1, some patients would receive treatment 2 at time 1, so that the subsequent intervention $F_2 = 1$ would not be well defined. If the same holds for $F_2 = 2$, then $F_1 = \varnothing \Rightarrow F_2 = \varnothing$.[26] The condition (64) will always hold provided that treatment decisions follow a time order, and that, regardless of the decisions that have occurred previously, it is always possible to decide to replace the "last" intervention with the idle regime.

It is easy to see that under the condition (64) for any $f \in \mathfrak{F}_A$, there will exist a sequence $(f = f^0, f^1, \ldots, f^q = \varnothing)$ such that for $j = 1, \ldots, q$, $f^j \in \mathfrak{F}_A$, and f^j contains one more idle regime indicator than f^{j-1}. It then follows under this condition that:

$$\text{if for all } i \in A, \quad W \perp\!\!\!\perp F_i \,|\, F_{A\setminus\{i\}} \quad \text{then } W \perp\!\!\!\perp F_A, \tag{65}$$

where here the conditional independence statements implicitly quantify over all the assignments to F_A that are in $\mathfrak{F}_A$, and hence valid.

Acknowledgments: We thank Ilya Shpitser and Philip Dawid for helpful comments and discussions.

Funding information: The authors completed work on this article while visiting the American Institute for Mathematics and the Simons Institute, Berkeley. The authors were supported by ONR Grant N000141912446; Robins was also supported by NIH Grant R01 AI032475.

Conflict of interest: The authors state no conflicts of interest.

References

[1] Dawid AP. Influence diagrams for causal modelling and inference. Int Stat Rev. 2002;70:161–89.
[2] Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. Math Model. 1986;7:1393–512.
[3] Richardson TS, Robins JM. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Center for Statistics and the Social Sciences Technical Report. 2013. University of Washington, Seattle, Washington, USA. https://www.csss.washington.edu/Papers/wp128.pdf.
[4] Robins JM, Richardson TS. Alternative graphical causal models and the identification of direct effects. Causality and psychopathology: finding the determinants of disorders and their cures. United Kingdom: Oxford University Press; 2010.
[5] Imbens GW. Causality in econometrics: choice vs chance. Econometrica. 2022;90(6):2541–66.
[6] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. New York: Springer Verlag; 1993.
[7] Malinsky D, Shpitser I, Richardson TS. A potential outcomes calculus for identifying conditional path-specific effects. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research. Naha, Okinawa, Japan: MLResearch Press; 2019.
[8] Dawid AP. Decision-theoretic foundations for statistical causality. J Causal Inference. 2021;9:39–77.
[9] Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669–709.
[10] Shpitser I, Richardson TS, Robins JM. Multivariate counterfactual systems and causal graphical models; 2021. arXiv:2008.06017.
[11] Ghassami A, Shpitser I, Richardson TS, Robins JM. Causal models with restricted interventions; 2023. In preparation.
[12] Robins JM. Personal Communication; 2018.
[13] Lauritzen SL, Dawid AP, Larsen B, Leimer HG. Independence properties of directed Markov fields. Networks. 1990;20:491–505.
[14] Robins JM, Hernán MA, Siebert U. Effects of multiple interventions. In: Ezzati M, Murray CJL, Lopez AD, Rodgers A, editors. Comparative quantification of health risks : global and regional burden of disease attributable to selected major risk factors. vol. 2. Geneva: World Health Organization; 2004. p. 2191–230.
[15] Pearl J. Causality. 2nd ed. Cambridge, UK: Cambridge University Press; 2009.

26 Here, we are implicitly supposing that only static regimes are under consideration so that F_2 can only take the values 1, 2, or $\varnothing$.

[16] Dawid AP. Causal inference without counterfactuals. J Amer Stat Assoc. 2000;95:407–48.
[17] Dawid AP. Conditional independence in statistical theory. J R Stat Soc Ser B (Methodological). 1979;41(1):1–31.
[18] Dawid AP. Conditional independence for statistical operations. Ann Statist. 1980;8:598–617.
[19] Gill R. The intersection axiom of conditional probability; 2019. https://www.slideshare.net/gill1109/the-intersection-axiom-of-conditional-probability.
[20] Sullivant S. Algebraic statistics. Providence, Rhode Island, USA: American Mathematical Society; 2018.
[21] Peters J. On the intersection property of conditional independence and its application to causal discovery. J Causal Inference. 2015;3(1):97–108.
[22] Thwaites P, Smith JQ, Riccomagno E. Causal analysis with chain event graphs. Artif Intelligence. 2010;174(12):889–909.
[23] Robins JM, Richardson TS, Shpitser I. An interventionist approach to mediation analysis. 2020. https://arxiv.org/abs/2008.06019.

Appendix

A.1 Conditions implying the SWIG local Markov property for $\mathcal{G}$ given that $p(V)$ factors with respect to $\mathcal{G}$

Here, we show that if $\mathcal{P}_A$ obeys the SWIG local Markov property corresponding to a complete graph $\overline{\mathcal{G}}$, and further, the observed distribution $p(V)$ is positive and obeys the local Markov property for a subgraph $\mathcal{G}$ of $\overline{\mathcal{G}}$, then it follows from distributional consistency that $\mathcal{P}_A$ also obeys the SWIG local Markov property corresponding to $\mathcal{G}$.

Theorem 25. *Suppose $\mathcal{P}_A^{\subseteq}$ obeys distributional consistency and $\mathcal{P}_A$ obeys the SWIG ordered local Markov property for $\overline{\mathcal{G}}$, a complete DAG.*[27] *If $p(V)$ is positive and obeys the Markov property for a subgraph $\mathcal{G}$ of $\overline{\mathcal{G}}$, then $\mathcal{P}_A$ obeys the SWIG ordered local Markov property for $\mathcal{G}$.*

Proof. Let $v \in \mathcal{X}_{\mathrm{pre}(i)}$, $v_i^* \in \mathcal{X}_i$, and $a \in \mathcal{X}_A$.

$$p(X_i(a) = v_i^* \mid X_{\mathrm{pre}(i)}(a) = v_{\mathrm{pre}(i)}) \tag{A1}$$

$$= p(X_i = v_i^* \mid X_{\mathrm{pre}(i)\setminus A} = v_{\mathrm{pre}(i)\setminus A}, X_{\mathrm{pre}(i)\cap A} = a_{\mathrm{pre}(i)\cap A}) \tag{A2}$$

$$= p(X_i = v_i^* \mid X_{\mathrm{pa}(i)\setminus A} = v_{\mathrm{pa}(i)\setminus A}, X_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}). \tag{A3}$$

Here, the first equality follows from (22) that holds under the local Markov property for $\overline{\mathcal{G}}$. The second equality is due to the local Markov property for $p(V)$. Consequently, we see that (A1) does not depend on $v_{\mathrm{pre}(i)\setminus(\mathrm{pa}(i)\setminus A)}$ nor on $a_{A\setminus \mathrm{pa}(i)}$ as required by the SWIG local Markov property. Note that positivity is used here in order to ensure that (A2) and (A3) are equal for all assignments to the variables in the conditioning events. □

This result is similar in spirit to Dawid's construction in that it provides conditions that, in conjunction with the observed distribution $p(V)$ obeying the Markov property for $\mathcal{G}$, are sufficient to imply that $\mathcal{P}_A$ obeys the SWIG local Markov property for $\mathcal{G}$. The SWIG ordered local Markov property on $\mathcal{P}_A$ for a complete graph $\overline{\mathcal{G}}$ corresponds to the FFRCISTG of [2], in the case where A represents the finest, i.e., largest, set of treatment variables for which well defined counterfactuals exist, and there are no (population- or individual-level) exclusion restrictions.

Note that for a complete graph $\overline{\mathcal{G}}$, for every variable i, $\mathrm{pre}(i) = \mathrm{pa}(i)$. Consequently, the SWIG local Markov property (11) and (13) reduce to requiring that for every i:

$$X_i(a) \perp\!\!\!\perp_d \overbrace{X_{\mathrm{pre}(i)\cap A}(a)}^{\text{ignorability}}, \quad \overbrace{x_{A\setminus \mathrm{pre}(i)}}^{\text{time order}} \mid \underbrace{x_{A\cap \mathrm{pre}(i)}}_{\text{fixed predecessors}}, \quad \underbrace{X_{\mathrm{pre}(i)\setminus A}(a).}_{\text{random predecessors}} \tag{A4}$$

Thus, we see that the SWIG local Markov property for the complete graph $\overline{\mathcal{G}}$ solely imposes ignorability and that interventions in the future do not change (the distribution of) variables in the past.

The single-graph approach given by Definition 7 and the two-graph construction of Theorem 25 each have their own strengths and weaknesses:

- In the single-graph approach, the model places restrictions on $\mathcal{P}_A$; distributional consistency for $\mathcal{P}_A^{\subseteq}$ then implies the relevant SWIG Markov properties for all the other distributions in $\mathcal{P}_A^{\subseteq}$, including the factual distribution $p(V)$. This approach is more concise insofar as it requires fewer conditions. The approach does not require $p(V)$ to be positive.

27 A DAG is complete if there is an edge between every pair of variables. Note that in this case, there is only one topological ordering.

- In the two-graph construction, the graph $\mathcal{G}$ specifies conditional independence restrictions on the observed distribution $p(V)$ via an ordinary Markov property, while the SWIG Markov property for the complete supergraph $\overline{\mathcal{G}}$ imposes ignorability and a total time order on $\mathcal{P}_A$. Under positivity for $p(V)$, distributional consistency for $\mathcal{P}_A^{\subseteq}$ then implies the relevant SWIG Markov properties for every distribution in $\mathcal{P}_A^{\subseteq}$. Though it requires more conditions, this approach has the advantage that it clearly demarcates a set of additional conditions that, when added to the assumption that $p(V)$ obeys the Markov property for $\mathcal{G}$, suffice to construct the full model on $\mathcal{P}_A^{\subseteq}$.

The fact that the single-graph approach does not require positivity can be seen as an advantage since it does not restrict the set of observed distributions. As a consequence, the graph in the single-graph approach may include edges that indicate effects arising from interventions on A that set variables to configurations that have probability zero under the observed distribution. Even in the absence of confounding, such effects may only be detectable via randomized experiments (see [23] for further discussion).

A.2 Derivation of part of the augmented DAG local Markov property for $\mathcal{G}$ from $p(V)$

Similar to our development in Section A.1, and also to Dawid's construction, we provide conditions on the kernel $p(W_V \mid F_A)$ that, in conjunction with a positive observed distribution $p(V)$ that obeys the local Markov property for a subgraph $\mathcal{G}$, suffice to ensure $p(W_V \mid F_A)$ obeys the local property for the corresponding augmented graph $\mathcal{G}^*$. These conditions are formulated in terms of a decision diagram $\overline{\mathcal{G}}^*$ corresponding to a complete DAG $\overline{\mathcal{G}}$ that contains $\mathcal{G}$ as a subgraph.

Theorem 26. *Suppose that $p(W_V \mid F_A)$ obeys distribution consistency and the augmented DAG local Markov property with respect to $\overline{\mathcal{G}}^*$ where $\overline{\mathcal{G}}$ is a complete DAG. If $p(V)$ is positive and obeys the (ordinary) Markov property for a subgraph $\mathcal{G}$ of $\overline{\mathcal{G}}$ then $p(W_V \mid F_A)$ also obeys the augmented DAG local Markov property for $\mathcal{G}^*$.*

For a complete graph $\overline{\mathcal{G}}$, for every variable i, $\mathrm{pre}(i) = \mathrm{pa}(i)$, and thus, the local Markov property for $\overline{\mathcal{G}}^*$ requires that for every i,

$$W_i \perp\!\!\!\perp_d \overbrace{W_{\mathrm{pre}(i)\cap A}}^{\text{ignorability}}, \quad \overbrace{F_{A\setminus \mathrm{pre}(i)}}^{\text{time order}} \mid \underbrace{F_{A\cap \mathrm{pre}(i)}}_{\text{fixed predecessors}}, \quad \underbrace{W_{\mathrm{pre}(i)\setminus A}}_{\text{random predecessors}}, \quad \underbrace{F_A \neq \varnothing}_{\text{intervene on all of } A}. \tag{A5}$$

Thus, similar to (A4), this imposes ignorability and that interventions in the future do not change (the distribution of) variables in the past.

Proof. Let $v \in \mathcal{X}_{\mathrm{pre}(i)}$, $v^* \in \mathcal{X}_i$, and $a \in \mathcal{X}_A$.

$$p(W_i = v_i^* \mid F_A = a, W_{\mathrm{pre}(i)} = v_{\mathrm{pre}(i)}) \tag{A6}$$

$$= p(W_i = v_i^* \mid W_{\mathrm{pre}(i)\cap A} = a_{\mathrm{pre}(i)\cap A}, W_{\mathrm{pre}(i)\setminus A} = v_{\mathrm{pre}(i)\setminus A}) \tag{A7}$$

$$= p(W_i = v_i^* \mid W_{\mathrm{pa}(i)\cap A} = a_{\mathrm{pa}(i)\cap A}, W_{\mathrm{pa}(i)\setminus A} = v_{\mathrm{pa}(i)\setminus A}). \tag{A8}$$

Here, the first equality follows from (53) which holds under the augmented local Markov property for $\overline{\mathcal{G}}^*$. The second equality is due to $p(V)$ obeying the local Markov property for $\mathcal{G}$. Consequently, we see that (A6) does not depend on $v_{\mathrm{pre}(i)\setminus(\mathrm{pa}(i)\setminus A)}$ nor on $a_{A\setminus \mathrm{pa}(i)}$, as required by the augmented graph local Markov property. Note that positivity ensures that (A7) and (A8) are equal for all assignments to the variables in the conditioning events. □