Statistical Science
2014, Vol. 29, No. 4, 459–484
DOI: 10.1214/14-STS505

Causal Etiology of the Research of James M. Robins

Thomas S. Richardson and Andrea Rotnitzky

Abstract. This issue of *Statistical Science* draws its inspiration from the work of James M. Robins. Jon Wellner, the Editor at the time, asked the two of us to edit a special issue that would highlight the research topics studied by Robins and the breadth and depth of Robins' contributions. Between the two of us, we have collaborated closely with Jamie for nearly 40 years. We agreed to edit this issue because we recognized that we were among the few in a position to relate the trajectory of his research career to date.[1]

Many readers may be unfamiliar with Robins' singular career trajectory and in particular how his early practical experience motivated many of the inferential problems with which he was subsequently involved. Robins majored in mathematics at Harvard College, but then, in the spirit of the times, left college to pursue more activist social and political goals. Several years later, Robins enrolled in Medical School at Washington University in St. Louis, graduating in 1976. His M.D. degree remains his only degree, other than his high school diploma.

After graduating, he interned in medicine at Harlem Hospital in New York. After completing the internship, Robins spent a year working as a primary care physician in a community clinic in the Roxbury neighborhood of Boston. During that year, he helped organize a vertical Service Employees International Union affiliate that included all salaried personnel, from maintenance to physicians, working at the health center. In retaliation, he was dismissed by the director of the clinic and found that he was somewhat unwelcome at the other Boston community clinics. Unable to find a job and with his unemployment insurance running out, he surprisingly was able to obtain a prestigious residency in Internal Medicine at Yale University, a testament, he says with some irony, to the enduring strength of one's Ivy League connections.

At Yale, Robins and his college friend Mark Cullen, now head of General Medicine at Stanford Medical School, founded an occupational health clinic, with the goal of working with trade unions in promoting occupational health and safety. When testifying in workers' compensation cases, Robins was regularly asked whether it was "more probable than not that a worker's death or illness was *caused* by exposure to chemicals in the workplace." Robins' lifelong interest in causal inference began with his need to provide an answer. As the relevant scientific papers consisted of epidemiologic studies and biostatistical analyses, Robins enrolled in biostatistics and epidemiology classes at Yale. He was dismayed to learn that the one question he needed to answer was the one question excluded from formal discussion in the mainstream biostatistical literature.[2] At the time, most biostatisticians insisted that evidence for causation could only be obtained through

Thomas S. Richardson is Professor and Chair, Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195, USA (e-mail: thomasr@uw.edu).
Andrea Rotnitzky is Professor, Department of Economics, Universidad Torcuato Di Tella & CONICET, Av. Figueroa Alcorta 7350, Sáenz Valiente 1010, Buenos Aires, Argentina (e-mail: arotnitzky@utdt.edu).

[1]Here, we restrict attention to Robins' contributions to the research literature. Robins has also contributed by training and mentoring leading researchers in causal inference: among others, Elizabeth Halloran, Miguel Hernán, Eric Tchetgen Tchetgen and Tyler VanderWeele worked with him as graduate students. Both of the editors of this Special Issue were greatly influenced by Robins' as a graduate student and post-doc and have been fortunate to have subsequently collaborated with him over many years.

[2]Robins and Greenland (1989a, 1989b) provided a formal definition of the probability of causation and a definitive answer to the question in the following sense. They proved that the probability of causation was not identified from epidemiologic data even in the absence of confounding, but that sharp upper and lower bounds could be obtained. Specifically, under the assumption that a workplace exposure was never beneficial, the probability $P(t)$ that a workers death occurring t years after exposure was due to that

randomized controlled trials; since, for ethical reasons, potentially harmful chemicals could not be randomly assigned, it followed that statistics could play little role in disentangling causation from spurious correlation.

1. CONFOUNDING

In his classes, Robins was struck by the gap present between the informal, yet insightful, language of epidemiologists such as Miettinen and Cook (1981) expressed in terms of "confounding, comparability, and bias," and the technical language of mathematical statistics in which these terms either did not have analogs or had other meanings. Robins' first major paper "The foundations of confounding in Epidemiology" written in 1982, though only published in 1987, was an attempt to bridge this gap. As one example, he offered a precise mathematical definition for the informal epidemiologic concept of a "confounding variable" that has apparently stood the test of time (see VanderWeele and Shpitser, 2013). As a second example, Efron and Hinkley (1978) had formally considered inference accurate to order $n^{-3/2}$ in variance conditional on exact or approximate ancillary statistics. Robins showed, surprisingly, that long before their paper, epidemiologists had been intuitively and informally referring to an estimator as "unbiased" just when it was asymptotically unbiased conditional on either exact or approximate ancillary statistics; furthermore, they intuitively required that the associated conditional Wald confidence interval be accurate to $O(n^{-3/2})$ in variance. As a third example, he solved the problem of constructing the tightest Wald-type intervals guaranteed to have conservative coverage for the average causal effect among the n study subjects participating in a completely randomized experiment with a binary response variable; he showed that this interval can be strictly narrower than the usual binomial interval even under the Neyman null hypothesis of no average causal effect. To do so, he constructed an estimator of the variance of the empirical difference in treatment means that improved on a variance estimator earlier proposed by Neyman (1923). Aronow, Green and Lee (2014) have recently generalized this result in several directions including to nonbinary responses.

exposure was sharply upper bounded by 1 and lower bounded by $\max[0, \{f_1(t) - f_0(t)\}/f_1(t)]$, where $f_1(t)$ and $f_0(t)$ are, respectively, the marginal densities in the exposed and unexposed cohorts of the random variable T encoding time to death.

2. TIME-DEPENDENT CONFOUNDING AND THE g-FORMULA

It was also in 1982 that Robins turned his attention to the subject that would become his grail: causal inference from complex longitudinal data with time-varying treatments, that eventually culminated in his revolutionary papers Robins (1986, 1987b). His interest in this topic was sparked by (i) a paper of Gilbert (1982)[3] on the healthy worker survivor effect in occupational epidemiology, wherein the author raised a number of questions Robins answered in these papers and (ii) his medical experience of trying to optimally adjust a patient's treatments in response to the evolution of the patient's clinical and laboratory data.

2.1 Overview

Robins career from this point on became a "quest" to solve this problem, and thereby provide methods that would address central epidemiological questions, for example, *is a given long-term exposure harmful or a treatment beneficial? If beneficial, what interventions, that is, treatment strategies, are optimal or near optimal?*

In the process, Robins created a "bestiary" of causal models and analytic methods.[4] There are the basic "phyla" consisting of the g-formula, marginal structural models and structural nested models. These phyla then contain "species," for example, structural nested failure time models, structural nested distribution models, structural nested (multiplicative additive and logistic) mean models and yet further "subspecies": direct-effect structural nested models and optimal-regime structural nested models.

Each subsequent model in this taxa was developed to help answer particular causal questions in specific contexts that the "older siblings" were not quite up to. Thus, for example, Robins' creation of structural nested and marginal structural models was driven by the so-called null paradox, which could lead to falsely finding a treatment effect where none existed, and was a serious nonrobustness of the estimated g-formula, his then current methodology. Similarly, his research on higher-order influence function estimators was motivated by a concern that, in the presence of confounding by continuous, high dimensional confounders, even

[3]The author, Ethel Gilbert, is the mother of Peter Gilbert who is a contributor to this special issue; see (Richardson et al., 2014).

[4]In the epidemiologic literature, this bestiary is sometimes referred to as the collection of "g-methods."

doubly robust methods might fail to adequately control for confounding bias.

This variety also reflects Robins' belief that the best analytic approach varies with the causal question to be answered, and, even more importantly, that confidence in one's substantive findings only comes when multiple, nearly orthogonal, modeling strategies lead to the same conclusion.

2.2 Causally Interpreted Structured Tree Graphs

Suppose one wishes to estimate from longitudinal data the causal effect of time-varying treatment or exposure, say cigarette smoking, on a failure time outcome such as all-cause mortality. In this setting, a time-dependent confounder is a time-varying covariate (e.g., presence of emphysema) that is a predictor of both future exposure and of failure. In 1982, the standard analytic approach was to model the conditional probability (i.e., the hazard) of failure time t as a function of past exposure history using a time-dependent Cox proportional hazards model. Robins formally showed that, even when confounding by unmeasured factors and model specification are absent, this approach may result in estimates of effect that may fail to have a causal interpretation, regardless of whether or not one also adjusts for the measured time-dependent confounders in the analysis. In fact, if previous exposure also predicts the subsequent evolution of the time-dependent confounders (e.g., since smoking is a cause of emphysema, it predicts this disease) then the standard approach can find an artifactual exposure effect even under the sharp null hypothesis of no net, direct or indirect effect of exposure on the failure time of any subject.

Prior to Robins (1986), although informal discussions of net, direct and indirect (i.e., mediated) effects of time varying exposures were to be found in the discussion sections of most epidemiologic papers, no formal mathematical definitions existed. To address this, Robins (1986) introduced a new counterfactual model, the *finest fully randomized causally interpreted structured tree graph* (FFRCISTG)[5] model that extended the point treatment counterfactual model of Neyman (1923) and Rubin (1974, 1978a)[6] to longitudinal studies with time-varying treatments, direct and indirect effects and feedback of one cause on another. Due to his lack of formal statistical training, the notation and formalisms in Robins (1986) differ from those found in the mainstream literature; as a consequence the paper can be a difficult read.[7] Richardson and Robins (2013, Appendix C) present the FFRCISTG model using a more familiar notation.[8]

We illustrate the basic ideas using a simplified example. Suppose that we obtain data from an observational or randomized study in which n patients are treated at two times. Let A_1 and A_2 denote the treatments. Let L be a measurement taken just prior to the second treatment and let Y be a final outcome, higher values of which are desirable. To simplify matters, for now we will suppose that all of the treatments and responses are binary. As a concrete example, consider a study of HIV infected subjects with (A_1, L, A_2, Y), respectively, being binary indicators of anti-retroviral treatment at time 1, high CD4 count just before time 2, anti-retroviral therapy at time 2, and survival at time 3 (where for simplicity we assume no deaths prior to assignment of A_2). There are $2^4 = 16$ possible observed data sequences for (A_1, L, A_2, Y); these may be depicted as an event tree as in Figure 1.[9] Robins (1986) referred to such event trees as "structured tree graphs."

[5] A complete list of acronyms used is given before the References.

[6] See Freedman (2006) and Sekhon (2008) for historical reviews of the counterfactual point treatment model.

[7] Robins published an informal, accessible, summary of his main results in the epidemiologic literature (Robins, 1987a). However, it was not until 1992 (and many rejections) that his work on causal inference with time-varying treatments appeared in a major statistical journal.

[8] The perhaps more familiar *Non-Parametric Structural Equation Model with Independent Errors* (NPSEM-IE) considered by Pearl may be viewed as submodel of Robins' FFRCISTG.

A *Non-Parametric Structural Equation Model* (NPSEM) assumes that all variables (V) can be intervened on. In contrast, the FFRCISTG model does not require one to assume this. However, if all variables in V can be intervened on, then the FFRCISTG specifies a set of one-step ahead counterfactuals, $V_m(\overline{v}_{m-1})$ which may equivalently be written as structural equations $V_m(\overline{v}_{m-1}) = f_m(\overline{v}_{m-1}, \varepsilon_m)$ for functions f_m and (vector-valued) random errors ε_m. Thus, leaving aside notational differences, structural equations and one-step ahead counterfactuals are equivalent. All other counterfactuals, as well as factual variables, are then obtained by recursive substitution.

However, the NPSEM-IE model of Pearl (2000) further assumes the errors ε_m are jointly independent. In contrast, though an FFRCISTG model is also an NPSEM, the errors (associated with incompatible counterfactual worlds) may be dependent—though any such dependence could not be detected in a RCT. Hence, Pearl's model is a strict submodel of an FFRCISTG model.

[9] In practice, there will almost always exist baseline covariates measured prior to A_1. In that case, the analysis in the text is to be understood as being with a given joint stratum of a set of baseline covariates sufficient to adjust for confounding due to baseline factors.

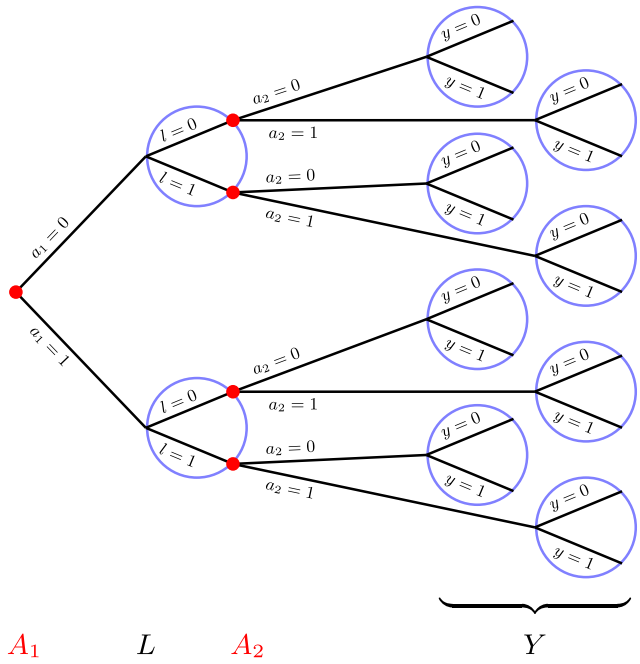

FIG. 1. *Causal tree graph depicting a simple scenario with treatments at two times A_1, A_2, a response L measured prior to A_2, and a final response Y. Blue circles indicate evolution of the process determined by Nature; red dots indicate potential treatment choices.*

We wish to assess the effect of the two treatments (a_1, a_2) on Y. In more detail, for a given subject we suppose the existence of four potential outcomes $Y(a_1, a_2)$ for $a_1, a_2 \in \{0, 1\}$ which are the outcome a patient would have if (possibly counter-to-fact) they were to receive the treatments a_1 and a_2. Then $E[Y(a_1, a_2)]$ is the mean outcome (e.g., the survival probability) if everyone in the population were to receive the specified level of the two treatments. The particular instance of this regime under which everyone is treated at both times, so $a_1 = a_2 = 1$, is depicted in Figure 4(a). We are interested in estimation of these four means since the regime (a_1, a_2) that maximizes $E[Y(a_1, a_2)]$ is the regime a new patient exchangeable with the n study subjects should follow.

There are two extreme scenarios: If in an observational study, the treatments are assigned, for example, by doctors, based on additional unmeasured predictors U of Y then $E[Y(a_1, a_2)]$ is not identified since those receiving (a_1, a_2) within the study are not representative of the population as a whole.

At the other extreme, if the data comes from a completely randomized clinical trial (RCT) in which treatment is assigned independently at each time by the flip of coin, then it is simple to see that the counterfactual $Y(a_1, a_2)$ is independent of the treatments (A_1, A_2) and that the average potential outcomes are identified since those receiving (a_1, a_2) in the study are a simple random sample of the whole population. Thus,

$$(1) \qquad Y(a_1, a_2) \perp\!\!\!\perp \{A_1, A_2\},$$

$$(2) \qquad E\big[Y(a_1, a_2)\big] = E[Y \mid A_1 = a_1, A_2 = a_2],$$

where the right-hand side of (2) is a function of the observed data distribution. In a completely randomized experiment, association is causation: the associational quantity on the right-hand side of (2) equals the causal quantity on the left-hand side.

Robins, however, considered an intermediate trial design in which both treatments are randomized, but the probability of receiving A_2 is dependent on both the treatment received initially (A_1) and the observed response (L); a scenario now termed a *sequential randomized trial*. Robins viewed his analysis as also applicable to observational data as follows. In an observational study, the role of an epidemiologist is to use subject matter knowledge to try to collect in L sufficient data to eliminate confounding by unmeasured factors, and thus to have the study mimic a sequential RCT. If successful, the only difference between an actual sequential randomized trial and an observational study is that in the former the randomization probabilities $\Pr(A_2 = 1 \mid L, A_1)$ are known by design while in the latter they must be estimated from the data.[10]

Robins viewed the sequential randomized trial as a collection of five trials in total: the original trial at $t = 1$, plus a set of four randomized trials at $t = 2$ nested within the original trial.[11] Let the counterfactual $L(a_1)$ be the outcome L when A_1 is set to a_1. Since the counterfactuals $Y(a_1, a_2)$ and $L(a_1)$ do not depend on the actual treatment received, they can be viewed, like a subject's genetic make-up, as a fixed (possibly unobserved) characteristic of a subject and therefore independent of the randomly assigned treatment conditional on pre-randomization covariates. That is, for each (a_1, a_2) and l:

$$(3) \quad \big\{Y(a_1, a_2), L(a_1)\big\} \perp\!\!\!\perp A_1,$$

$$(4) \qquad Y(a_1, a_2) \perp\!\!\!\perp A_2 \mid A_1 = a_1, \quad L = l.$$

These independences suffice to identify the joint density $f_{Y(a_1,a_2),L(a_1)}(y, l)$ of $(Y(a_1, a_2), L(a_1))$ from

[10] Of course, one can never be certain that the epidemiologists were successful which is the reason RCTs are generally considered the gold standard for establishing causal effects.

[11] That is, the trials starting at $t = 2$ are on study populations defined by specific (A_1, L)-histories.

the distribution of the factual variables by the "g-computation algorithm formula" (or simply *g-formula*) density

$$f^*_{a_1,a_2}(y,l) \equiv f(y \mid a_1, l, a_2) f(l \mid a_1) \tag{5}$$

provided the conditional probabilities on the right-hand side are well-defined (Robins, 1986, page 1423). Note that $f^*_{a_1,a_2}(y,l)$ is obtained from the joint density of the factuals by removing the treatment terms $f(a_2 \mid a_1, l, a_2) f(a_1)$. This is in-line with the intuition that A_1 and A_2 cease to be random since, under the regime, they are set by intervention to constants a_1 and a_2. The g-formula was later referred to as the "manipulated density" by Spirtes, Glymour and Scheines (1993) and the "truncated factorization" by Pearl (2000).

Robins (1987b) showed that under the weaker condition that replaces (3) and (4) with

$$\begin{aligned} &Y(a_1,a_2) \perp\!\!\!\perp A_1 \quad \text{and} \\ &Y(a_1,a_2) \perp\!\!\!\perp A_2 \mid A_1 = a_1, \quad L = l, \end{aligned} \tag{6}$$

the marginal density of $Y(a_1,a_2)$ is still identified by

$$f^*_{a_1,a_2}(y) = \sum_l f(y \mid a_1, l, a_2) f(l \mid a_1), \tag{7}$$

the marginal under $f^*_{a_1,a_2}(y,l)$.[12] Robins called (6) *randomization w.r.t. Y*.[13] Furthermore, he provided substantive examples of observational studies in which only the weaker assumption would be expected to hold. It is much easier to describe these studies using representations of causal systems using Directed Acyclic Graphs and Single World Intervention Graphs, neither of which existed when (Robins, 1987b) was written.

2.3 Causal DAGs and Single World Intervention Graphs (SWIGs)

Causal DAGs were first introduced in the seminal work of Spirtes, Glymour and Scheines (1993); the theory was subsequently developed and extended by Pearl (1995a, 2000) among others.

A causal DAG with random variables $V_1, \ldots, V_M$ as nodes is a graph in which (1) the lack of an arrow from node V_j to V_m can be interpreted as the absence of a direct causal effect of V_j on V_m (relative to the other variables on the graph), (2) all common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph, and (3) the Causal Markov Assumption (CMA) holds. The CMA links the causal structure represented by the Directed Acyclic Graph (DAG) to the statistical data obtained in a study. It states that the distribution of the factual variables factor according to the DAG. A distribution factors according to the DAG if nondescendants of a given variable V_j are independent of V_j conditional on pa_j, the parents of V_j. The CMA is mathematically equivalent to the statement that the density $f(v_1, \ldots, v_M)$ of the variables on the causal DAG $\mathcal{G}$ satisfies the Markov factorization

$$f(v_1, \ldots, v_M) = \prod_{j=1}^{M} f(v_j \mid \text{pa}_j). \tag{8}$$

A graphical criterion, called d-separation (Pearl, 1988), characterizes all the marginal and conditional independences that hold in every distribution obeying the Markov factorization (8).

Causal DAGs may also be used to represent the joint distribution of the observed data under the counterfactual FFRCISTG model of Robins (1986). This follows because an FFRCISTG model over the variables $\{V_1, \ldots, V_M\}$ induces a distribution that factors as (8). Figure 2(a) shows a causal Directed Acyclic Graph (DAG) corresponding to the sequentially randomized experiment described above: vertex H represents an unmeasured common cause (e.g., immune function) of CD4 count L and survival Y. Randomization of treatment implies A_1 has no parents and A_2 has only the observed variables A_1 and L as parents.

Single-World Intervention Graphs (SWIGs), introduced in (Richardson and Robins, 2013), provide a simple way to derive the counterfactual independence relations implied by an FFRCISTG model. SWIGs were designed to unify the graphical and potential outcome approaches to causality. The nodes on a SWIG are the counterfactual random variables associated with a specific hypothetical intervention on the treatment variables. The SWIG in Figure 2(b) is derived from the causal DAG in Figure 2(a) corresponding to a sequentially randomized experiment. The SWIG represents the counterfactual world in which A_1 and A_2 have been set to (a_1, a_2), respectively. Richardson and Robins (2013) show that under the (naturally associated) FFRCISTG model the distribution of the counterfactual variables on the SWIG factors according to

[12] The g-formula density for Y is a generalization of standardization of effect measures to time varying treatments. See Keiding and Clayton (2014) for a historical review of standardization.

[13] Note that the distribution of $L(a_1)$ is no longer identified under this weaker assumption.

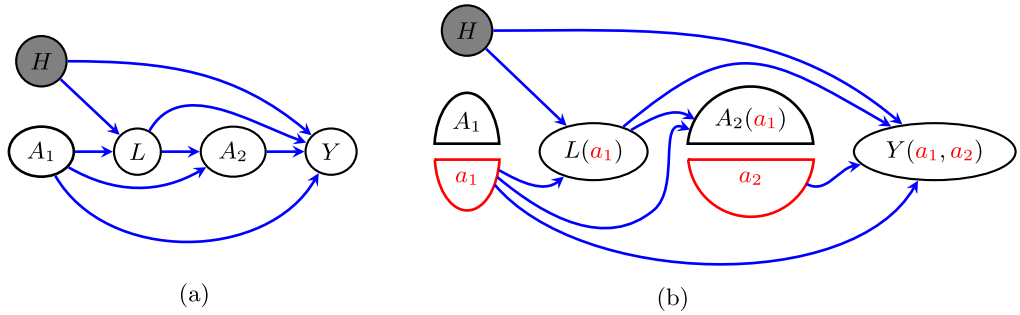

FIG. 2. (a) *A causal DAG $\mathcal{G}$ describing a sequentially randomized trial*; (b) *the SWIG $\mathcal{G}(a_1, a_2)$ resulting from intervening on A_1 and A_2.*

the graph. Applying Pearl's d-separation criterion to the SWIG we obtain the independences (3) and (4).[14]

Robins (1987b) in one of the aforementioned substantive examples described an observational study of the effect of formaldehyde exposure on the mortality of rubber workers which can represented by the causal graph in Figure 3(a). (This graph cannot represent a sequential RCT because the treatment variable A_1 and the response L have an unmeasured common cause.) Follow-up begins at time of hire; time 1 on the graph. The vertices H_1, A_1, H_2, L_2, A_2, Y are indicators of sensitivity to eye irritants, formaldehyde exposure at time 1, lung cancer, current employment, formaldehyde exposure at time 2 and survival. Data on eye-sensitivity and lung cancer were not collected. Formaldehyde is a known eye-irritant. The presence of an arrow from H_1 to A_1 but not from H_1 to A_2 reflects the fact that subjects who believe their eyes to be sensitive to formaldehyde are given the discretion to choose a job without formaldehyde exposure at time of hire but not later. The arrow from H_1 to L reflects the fact that eye sensitivity causes some subjects to leave employment. The arrows from H_2 to L_2 and Y reflects the fact that lung cancer causes both death and loss of employment. The fact that H_1 and H_2 are independent reflects the fact that eye sensitivity is unrelated to the risk of lung cancer.

From the SWIG in Figure 3(b), we can see that (6) holds so we have randomization with respect to Y but $L(a_1)$ is not independent of A_1. It follows that the g-formula $f^*_{a_1,a_2}(y)$ equals the density of $Y(a_1, a_2)$ even though (i) the distribution of $L(a_1)$ is not identified and (ii) neither of the individual terms $f(l \mid a_1)$ and $f(y \mid a_1, l, a_2)$ occurring in the g-formula has a causal interpretation.[15]

[14] More precisely, we obtain the SWIG independence $Y(a_1, a_2) \perp\!\!\!\perp A_2(a_1) \mid A_1, L(a_1)$, that implies (4) by the consistency assumption after instantiating A_1 at a_1. Note when checking d-separation on a SWIG all paths containing red "fixed" nonrandom vertices, such as a_1, are treated as always being blocked (regardless of the conditioning set).

[15] Above we have assumed the variables A_1, L, A_2, Y occurring in the g-formula are temporally ordered. Interestingly, Robins (1986, Section 11) showed identification by the g-formula can require a nontemporal ordering. In his analysis of the Healthy Worker Sur-

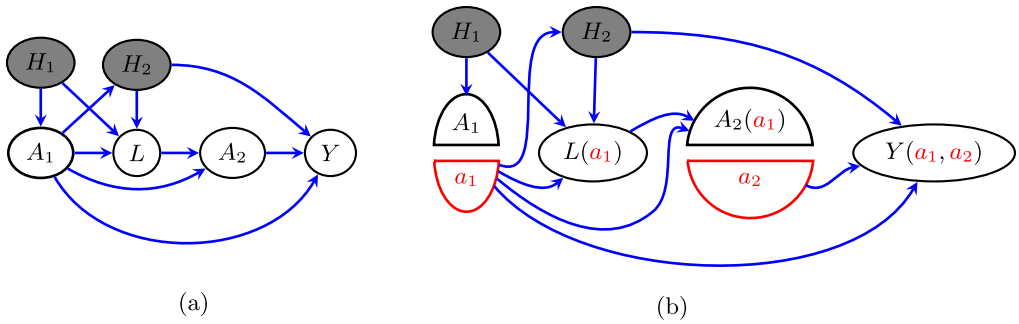

FIG. 3. *Formaldehyde study*: *H_1, indicator of sensitivity to eye irritants*; *A_1, formaldehyde exposure at time* 1; *H_2, lung cancer*; *L, current employment*; *A_2, formaldehyde exposure at time* 2; *Y, survival. H_1 and H_2 are unmeasured.* (a) *A causal DAG $\mathcal{G}$ in which initial treatment is confounded, while the second treatment is sequentially randomized*; (b) *the SWIG $\mathcal{G}(a_1, a_2)$. L is known to have no direct effect on Y except indirectly via the effect on A_2; H_1 influences A_1 but not A_2. See text for further explanation.*

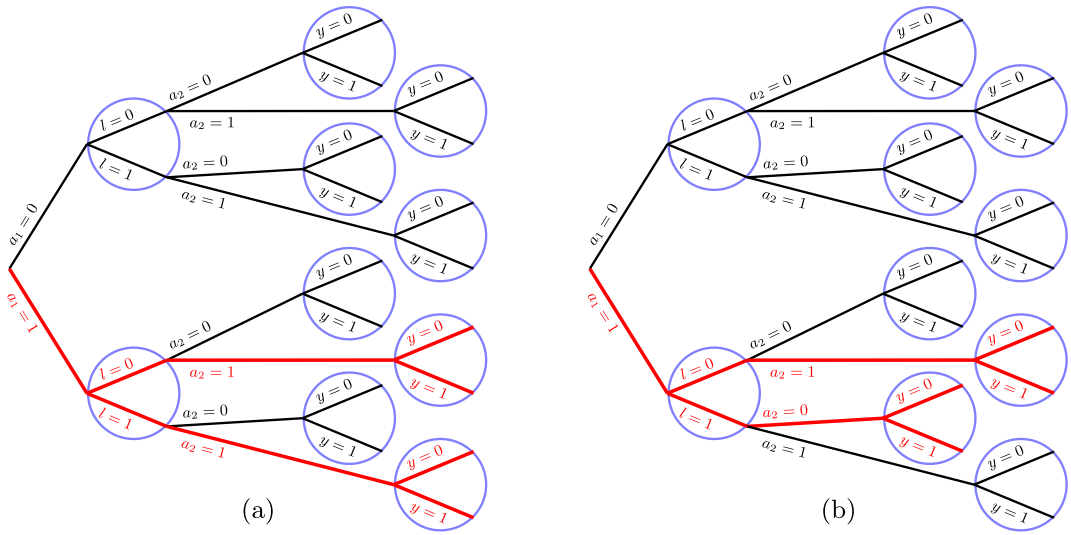

FIG. 4. *Tree graphs depicting specific treatment regimes*: (a) $a_1 = a_2 = 1$; (b) *the dynamic regime* $a_1 = 1$, $a_2 = (1 - l)$. *The red paths indicate all possible observed data sequences under these regimes.*

Subsequently, Tian and Pearl (2002a) developed a graphical algorithm for nonparametric identification that is "complete" in the sense that if the algorithm fails to derive an identifying formula, then the causal quantity is not identified (Shpitser and Pearl, 2006; Huang and Valtorta, 2006). This algorithm strictly extends the set of causal identification results obtained by Robins for static regimes.

2.4 Dynamic Regimes

The "g" in "g-formula" and elsewhere in Robins' work refers to generalized treatment regimes g. The set $\mathbb{G}$ of all such regimes includes *dynamic* regimes in which a subject's treatment at time 2 depends on the response L to the treatment at time 1. An example of a dynamic regime is the regime in which all subjects receive anti-retroviral treatment at time 1, but continue to receive treatment at time 2 only if their CD4 count at time 2 is low, indicating that they have not yet responded to anti-retroviral treatment. In our study with no baseline covariates and A_1 and A_2 binary, a dynamic regime g can be written as $g = (a_1, g_2(l))$ where the function $g_2(l)$ specifies the treatment to be given at time 2. The dynamic regime above has $(a_1 = 1, g_2(l) = 1 - l)$ and is highlighted in Figure 4. If L is binary, then $\mathbb{G}$ consists of 8 regimes comprised of the 4 earlier static regimes (a_1, a_2) and 4 *dynamic* regimes. The *g-formula* density associated with a regime $g = (a_1, g_2(l))$ is

$$f_g^*(y, l) \equiv f(l \mid a_1) f\big(y \mid A_1 = a_1, L = l, A_2 = g_2(l)\big).$$

Letting $Y(g)$ be a subject's counterfactual outcome under regime g, Robins (1987b) proves that if both of the following hold:

$$\begin{aligned} &Y(g) \perp\!\!\!\perp A_1, \\ &Y(g) \perp\!\!\!\perp A_2 \mid A_1 = a_1, \quad L = l \end{aligned} \tag{9}$$

then $f_{Y(g)}(y)$ is identified by the g-formula density for Y:

$$\begin{aligned} f_g^*(y) &= \sum_l f_g^*(y, l) \\ &= \sum_l f\big(y \mid A_1 = a_1, L = l, A_2 = g_2(l)\big) \\ &\qquad \cdot f(l \mid a_1). \end{aligned}$$

Robins (1987b) refers to (9) as the assumption that regime g *is randomized with respect to* Y. Given a causal DAG, Dynamic SWIGs (dSWIGS) can be used to check whether (9) holds. Tian (2008) gives a complete graphical algorithm for identification of the effect of dynamic regimes based on DAGs.

vivor Effect, data were available on temporally ordered variables (A_1, L_1, A_2, L_2, Y) where the L_t are indicators of survival until time year t, A_t is the indicator of exposure to a lung carcinogen and, there exists substantive background knowledge that carcinogen exposure at t cannot cause death within a year. Under these assumptions, Robins proved that equation (6) was false if one respected temporal order and chose L to be L_1, but was true if one chose $L = L_2$. Thus, $E[Y(a_1, a_2)]$ was identified by the g-formula $f^*_{a_1,a_2}(y)$ only for $L = L_2$. See (Richardson and Robins, 2013, page 54) for further details.

Independences (3) and (4) imply that (9) is true for all $g \in \mathbb{G}$. For a drug treatment, for which, say, higher outcome values are better, the optimal regime g_{opt} maximizing $E[Y(g)]$ over $g \in \mathbb{G}$ is almost always a dynamic regime, as treatment must be discontinued when toxicity, a component of L, develops.

Robins (1989, 1986, page 1423) used the g-notation $f(y \mid g)$ as a shorthand for $f_{Y(g)}(y)$ in order to emphasize that this was the density of Y *had intervention* g *been applied to the population*. In the special case of static regimes (a_1, a_2), he wrote $f(y \mid g = (a_1, a_2))$.[16]

2.5 Statistical Limitations of the Estimated g-Formulae

Consider a sequentially randomized experiment. In this context, randomization probabilities $f(a_1)$ and $f(a_2 \mid a_1, l)$ are known by design; however, the densities $f(y \mid a_1, a_2, l)$ and $f(l \mid a_1)$ are not known and, therefore, they must be replaced by estimates $\widehat{f}(y \mid a_1, a_2, l_2)$ and $\widehat{f}(l \mid a_1)$ in the g-formula. If the sample size is moderate and l is high dimensional, these estimates must come from fitting dimension-reducing models. Model misspecification will then lead to biased estimators of the mean of $Y(a_1, a_2)$. Robins (1986) and Robins and Wasserman (1997) described a serious nonrobustness of the g-formula: the so-called "null paradox": In biomedical trials, it is frequently of interest to consider the possibility that the sharp causal null hypothesis of no effect of either A_1 or A_2 on Y holds. Under this null, the causal DAG generating the data is as in Figure 2 except without the arrows from A_1, A_2 and L into Y.[17] Then, under this null, although $f^*_{a_1,a_2}(y) = \sum_l f(y \mid a_1, l, a_2) f(l \mid a_1)$ does not depend on (a_1, a_2), nonetheless both $f(y \mid a_1, l, a_2)$ and $f(l \mid a_1)$ will, in general, depend on a_1 (as may be seen via d-connection).[18] In general, if L has discrete components, it is not possible for standard nonsaturated parametric models (e.g., logistic regression models) for both $f(y \mid a_1, a_2, l_2)$ and $f(l_2 \mid a_1)$ to be correctly specified, and thus depend on a_1 and yet for $f^*_{a_1,a_2}(y)$ not to depend on a_1.[19] As a consequence, inference based on the estimated g-formula must result in the sharp null hypothesis being falsely rejected with probability going to 1, as the trial size increases, even when it is true.

[16] Pearl (1995a) introduced an identical notation except that he substituted the word "do" for "$g =$," thus writing $f(y \mid \text{do}(a_1, a_2))$.

[17] If the $L \to Y$ edge is present, then A_1 still has an effect on Y.

[18] The dependence of $f(y \mid a_1, l, a_2)$ on a_1 does not represent causation but rather selection bias due to conditioning on the common effect L of H_1 and A_1.

[19] But see Cox and Wermuth (1999) for another approach.

2.6 Structural Nested Models[20]

To overcome the null paradox, Robins (1989) and Robins et al. (1992) introduced the semiparametric structural nested distribution model (SNDMs) for continuous outcomes Y and structural nested failure time models (SNFTMs) for time to event outcomes. See Robins (1997a, 1997b) for additional details.

Robins (1986, Section 6) defined the *g-null hypothesis* as

$$H_0 : \text{the distribution of } Y(g) \text{ is the same for all } g \in \mathbb{G}. \tag{10}$$

This hypothesis is implied by the sharp null hypothesis of no effect of A_1 or A_2 on any subject's Y. If (9) holds for all $g \in \mathbb{G}$, then the g-null hypothesis is equivalent to any one of the following assertions:

(i) $f^*_g(y)$ equals the factual density $f(y)$ for all $g \in \mathbb{G}$;
(ii) $Y \perp\!\!\!\perp A_1$ and $Y \perp\!\!\!\perp A_2 \mid L, A_1$;
(iii) $f^*_{a_1,a_2}(y)$ does not depend on (a_1, a_2) and $Y \perp\!\!\!\perp A_2 \mid L, A_1$;

see Robins (1986, Section 6). In addition, any one of these assertions exhausts all restrictions on the observed data distribution implied by the sharp null hypothesis.

Robins' goal was to construct a causal model indexed by a parameter ψ^* such that in a sequentially randomized trial (i) $\psi^* = 0$ if and only if the g-null hypothesis (10) was true and (ii) if known, one could use the randomization probabilities to both construct an unbiased estimating function for ψ^* and to construct tests of $\psi^* = 0$ that were guaranteed (asymptotically) to reject under the null at the nominal level. The SNDMs and SNFTMs accomplish this goal for continuous and failure time outcomes Y. Robins (1989) and Robins (1994) also constructed additive and multiplicative structural nested mean models (SNMMs) which satisfied the above properties except with the g-null hypothesis replaced by the *g-null mean hypothesis*:

$$H_0 : E[Y(g)] = E[Y] \quad \text{for all } g \in \mathbb{G}. \tag{11}$$

As an example, we consider an additive structural nested mean model. Define

$$\gamma(a_1, l, a_2) = E[Y(a_1, a_2) - Y(a_1, 0) \mid L = l, A_1 = a_1, A_2 = a_2]$$

[20] These models are discussed by Vansteelandt and Joffe (2014) in this issue.

and

$$\gamma(a_1) = E\big[Y(a_1, 0) - Y(0, 0) \mid A_1 = a_1\big].$$

Note $\gamma(a_1, l, a_2)$ is the effect of the last blip of treatment a_2 at time 2 among subjects with observed history (a_1, l, a_2), while $\gamma(a_1)$ is the effect of the last blip of treatment a_1 at time 1 among subjects with history a_1. An additive SNMM specifies parametric models $\gamma(a_1, l, a_2; \psi_2)$ and $\gamma(a_1; \psi_1)$ for these blip functions with $\gamma(a_1; 0) = \gamma(a_1, l, a_2; 0) = 0$. Under the independence assumptions (9), $H_2(\psi_2)d(L, A_1)\{A_2 - E[A_2 \mid L, A_1]\}$ and $H_1(\psi)\{A_1 - E[A_1]\}$ are unbiased estimating functions for the true ψ^*, where $H_2(\psi_2) = Y - \gamma(A_1, L, A_2; \psi_2)$, $H_1(\psi) = H_2(\psi_2) - \gamma(A_1; \psi_1)$, and $d(L, A_1)$ is a user-supplied function of the same dimension as ψ_2. Under the g-null mean hypothesis (11), the SNMM is guaranteed to be correctly specified with $\psi^* = 0$. Thus, these estimating functions when evaluated at $\psi^* = 0$, can be used in the construction of an asymptotically α-level test of the g-null mean hypothesis when $f(a_1)$ and $f(a_2 \mid a_1, l)$ are known (or are consistently estimated).[21] When L is a high-dimensional vector, the parametric blip models may well be misspecified when g-null mean hypothesis is false. However, because the functions $\gamma(a_1, l, a_2)$ and $\gamma(a_1)$ are nonparametrically identified under assumptions (9), one can construct consistent tests of the correctness of the blip models $\gamma(a_1, l, a_2; \psi_2)$ and $\gamma(a_1; \psi_1)$. Furthermore, one can also estimate the blip functions using cross-validation (Robins, 2004) and/or flexible machine learning methods in lieu of a prespecified parametric model (van der Laan and Rose, 2011). A recent modification of a multiplicative SNMM, the structural nested cumulative failure time model, designed for censored time to event outcomes has computational advantages compared to a SNFTM, because, in contrast to a SNFTM, parameters are estimated using an unbiased estimating function that is differentiable in the model parameters; see Picciotto et al. (2012).

Robins (2004) also introduced optimal-regime SNNMs drawing on the seminal work of Murphy (2003) on semiparametric methods for the estimation of optimal treatment strategies. Optimal-regime SNNM estimation, called A-learning in computer science, can be viewed as a semiparametric implementation of dynamic programming (Bellman, 1957).[22] Optimal-regime SNMMs differ from standard SNMMs only in that $\gamma(a_1)$ is redefined to be

$$\gamma(a_1) = E\big[Y\big(a_1, g_{2,\text{opt}}(a_1, L(a_1))\big) - Y\big(0, g_{2,\text{opt}}(0, L(0))\big) \mid A_1 = a_1\big],$$

where $g_{2,\text{opt}}(a_1, l) = \arg\max_{a_2} \gamma(a_1, l, a_2)$ is the optimal treatment at time 2 given past history (a_1, l). The overall optimal treatment strategy g_{opt} is then $(a_{1,\text{opt}}, g_{2,\text{opt}}(a_1, l))$ where $a_{1,\text{opt}} = \arg\max_{a_1} \gamma(a_1)$. More on the estimation of optimal treatment regimes can be found in Schulte et al. (2014) in this volume.

2.7 Instrumental Variables and Bounds for the Average Treatment Effect

Robins (1989, 1993) also noted that structural nested models can be used to estimate treatment effects when assumptions (9) do not hold but data are available on a time dependent instrumental variable. As an example, patients sometimes fail to fill their prescriptions and thus do not comply with their prescribed treatment. In that case, we can take $A_j = (A_j^p, A_j^d)$ for each time j, where A_j^p denotes the treatment *prescribed* and A_j^d denotes the *dose* of treatment actually received at time j. Robins defined A_j^p to be *an instrumental variable* if (9) still holds after replacing A_j by A_j^p and for all subjects $Y(a_1, a_2)$ depends on $a_j = (a_j^p, a_j^d)$ only through the actual dose a_j^d. Robins noted that unlike the case of full compliance (i.e., $A_j^p = A_j^d$ with probability 1) discussed earlier, the treatment effect functions γ are not nonparametrically identified. Consequently, identification can only be achieved by correctly specifying (sufficiently restrictive) parametric models for γ.

If we are unwilling to rely on such parametric assumptions, then the observed data distribution only implies bounds for the γ's. In particular, in the setting of a point treatment randomized trial with noncompliance and the instrument A_1^p being the assigned treatment, Robins (1989) obtained bounds on the average causal effect $E[Y(a_d = 1) - Y(a_d = 0)]$ of the received treatment A_d. To the best of our knowledge, this paper was the first to derive bounds for nonidentified causal effects defined through potential outcomes.[23] The study of such bounds has become an active area of research. Other early papers include Manski (1990) and Balke

[21] In the literature, semiparametric estimation of the parameters of a SNM based on such estimating functions is referred to as "g-estimation."

[22] Interestingly, Robins (1989, page 127 and App. 1), unaware of Bellman's work, reinvented the method of dynamic programming but remarked that, due to the difficulty of the estimation problem, it would only be of theoretical interest for finding the optimal dynamic regimes from longitudinal epidemiological data.

[23] See also Robins and Greenland (1989a, 1989b).

and Pearl (1994).[24] See Richardson et al. (2014) in this volume for a survey of recent research on bounds.

2.8 Limitations of Structural Nested Models

Robins (2000) noted that there exist causal questions for which SNMs are not altogether satisfactory. As an example, for Y binary, Robins (2000) proposed a structural nested logistic model in order to ensure estimates of the counterfactual mean of Y were between zero and one. However, he noted that knowledge of the randomization probabilities did not allow one to construct unbiased estimating function for its parameter ψ^*. More importantly, SNMs do not directly model the final object of public health interest—the distribution or mean of the outcome Y as function of the regimes g—as these distributions are generally functions not only of the parameters of the SNM but also of the conditional law of the time dependent covariates L given the past history. In addition, SNMs constitute a rather large conceptual leap from standard associational regression models familiar to most statisticians. Robins (1998, 2000) introduced a new class of causal models, marginal structural models, that overcame these particular difficulties. Robins also pointed out that MSMs have their own shortcomings, which we discuss below. Robins (2000) concluded that the best causal model to use will vary with the causal question of interest.

2.9 Dependent Censoring and Inverse Probability Weighting

Marginal Structural Models grew out of Robins' work on censoring and *inverse probability of censoring weighted* (IPCW) estimators. Robins work on dependent censoring was motivated by the familiar clinical observation that patients who did not return to the clinic and were thus censored differed from other patients on important risk factors, for example measures of cardio-pulmonary reserve. In the 1970s and 1980s, the analysis of right censored data was a major area of statistical research, driven by the introduction of the proportional hazards model (Cox, 1972; Kalbfleisch and Prentice, 1980) and by martingale methods for their analysis (Aalen, 1978; Andersen et al., 1993; Fleming and Harrington, 1991). This research, however, was focused on independent censoring. An important insight in Robins (1986) was the recognition that by reframing the problem of censoring as a causal inference problem as we will now explain, it was possible to adjust for dependent censoring with the g-formula.

Rubin (1978a) had pointed out previously that counterfactual causal inference could be viewed as a missing data problem. Robins (1986, page 1491) recognized that the converse was indeed also true: a missing data problem could be viewed as a problem in counterfactual causal inference.[25] Robins conceptualized right censoring as just another time dependent "treatment" A_t and one's inferential goal as the estimation of the outcome Y under the static regime g "never censored." Inference based on the g-formula was then licensed provided that censoring was explainable in the sense that (6) holds. This approach to dependent censoring subsumed independent censoring as the latter is a special case of the former.

Robins, however, recognized once again that inference based on the estimated g-formula could be nonrobust. To overcome this difficulty, Robins and Rotnitzky (1992) introduced IPCW tests and estimators whose properties are easiest to explain in the context of a two-armed RCT of a single treatment (A_1). The standard Intention-to-Treat (ITT) analysis for comparing the survival distributions in the two arms is a log-rank test. However, data are often collected on covariates, both pre- and post-randomization, that are predictive of the outcome as well as (possibly) of censoring. An ITT analysis that tries to adjust for dependent-censoring by IPCW uses estimates of the arm-specific hazards of censoring as functions of past covariate history. The proposed IPCW tests have the following two advantages compared to the log rank test. First, if censoring is dependent but explainable by the covariates, the log-rank test is not asymptotically valid. In contrast, IPCW tests asymptotically reject at their nominal level provided the arm-specific hazard estimators are consistent. Second, when censoring is independent, although both the IPCW tests and the log-rank test asymptotically reject at their nominal level, the IPCW tests, by making use of covariates, can be more powerful than the log-rank test even against proportional-hazards alternatives. Even under independent censoring tests based

[24] Balke and Pearl (1994) showed that Robins' bounds were not sharp in the presence of "defiers" (i.e., subjects who would never take the treatment assigned) and derived sharp bounds in that case.

[25] A viewpoint recently explored by Mohan, Pearl and Tian (2013).

on the estimated g-formula are not guaranteed to be asymptotically α-level, and hence are not robust.

To illustrate, we consider here an RCT with A_1 being the randomization indicator, L a post-randomization covariate, A_2 the indicator of censoring and Y the indicator of survival. For simplicity, we assume that any censoring occurs at time 2 and that there are no failures prior to time 2. The IPCW estimator $\widehat{\beta}$ of the ITT effect $\beta^* = E[Y \mid A = 1] - E[Y \mid A = 0]$ is defined as the solution to

$$\mathbb{P}_n\big[I(A_2 = 0)U(\beta)/\widehat{\Pr}(A_2 = 0 \mid L, A_1)\big] = 0, \tag{12}$$

where $U(\beta) = (Y - \beta A_1)(A_1 - 1/2)$, throughout $\mathbb{P}_n$ denotes the empirical mean operator and $\widehat{\Pr}(A_2 = 0 \mid L, A_1)$ is an estimator of the arm-specific conditional probability of being uncensored. When first introduced in 1992, IPCW estimators, even when taking the form of simple Horvitz–Thompson estimators, were met with both surprise and suspicion as they violated the then widely held belief that one should never adjust for a post-randomization variable affected by treatment in a RCT.

2.10 Marginal Structural Models

Robins (1993, Remark A1.3, pages 257–258) noted that, for any treatment regime g, if randomization w.r.t. Y, that is, (9), holds, $\Pr\{Y(g) > y\}$ can be estimated by IPCW if one defines a person's censoring time as the first time he/she fails to take the treatment specified by the regime. In this setting, he referred to IPCW as *inverse probability of treatment weighted* (IPTW). In actual longitudinal data in which either (i) treatment A_k is measured at many times k or (ii) the A_k are discrete with many levels or continuous, one often finds that few study subjects follow any particular regime. In response, Robins (1998, 2000) introduced MSMs. These models address the aforementioned difficulty by borrowing information across regimes. Additionally, MSMs represent another response to the g-null paradox complementary to Structural Nested Models.

To illustrate, suppose that in our example of Section 2, A_1 and A_2 now have many levels. An instance of an MSM for the counterfactual means $E[Y(a_1, a_2)]$ is a model that specifies that

$$\Phi^{-1}\big\{E\big[Y(a_1, a_2)\big]\big\} = \beta_0^* + \gamma(a_1, a_2; \beta_1^*),$$

where Φ^{-1} is a given link function such as the logit, log, or identity link and $\gamma(a_1, a_2; \beta_1)$ is a known function satisfying $\gamma(a_1, a_2; 0) = 0$. In this model, $\beta_1 = 0$ encodes the *static-regime mean null hypothesis* that

$$H_0 : E\big[Y(a_1, a_2)\big] \text{ is the same for all } (a_1, a_2). \tag{13}$$

Robins (1998) proposed IPTW estimators $(\widehat{\beta}_0, \widehat{\beta}_1)$ of (β_0^*, β_1^*). When the treatment probabilities are known, these estimators are defined as the solution to

$$\mathbb{P}_n\big[W v(A_1, A_2)\big(Y - \Phi\{\beta_0 + \gamma(A_1, A_2; \beta_1)\}\big)\big] = 0 \tag{14}$$

for a user supplied vector function $v(A_1, A_2)$ of the dimension of (β_0^*, β_1^*) where

$$W = 1/\big\{f(A_1)f(A_2 \mid A_1, L)\big\}.$$

Informally, the product $f(A_1)f(A_2 \mid A_1, L)$ is the "probability that a subject had the treatment history he did indeed have."[26] When the treatment probabilities are unknown, they are replaced by estimators.

Intuitively, the reason why the estimating function of (14) has mean zero at (β_0^*, β_1^*) is as follows: Suppose the data had been generated from a sequentially randomized trial represented by DAG in Figure 2. We may create a pseudo-population by making $1/\{f(A_1)f(A_2 \mid A_1, L)\}$ copies of each study subject. It can be shown that in the resulting pseudo-population $A_2 \perp\!\!\!\perp \{L, A_1\}$, and thus is represented by the DAG in Figure 2, except with both arrows into A_2 removed. In the pseudo-population, treatment is completely randomized (i.e., there is no confounding by either measured or unmeasured variables), and hence causation is association. Further, the mean of $Y(a_1, a_2)$ takes the same value in the pseudo-population as in the actual population. Thus if, for example, $\gamma(a_1, a_2; \beta_1) = \beta_{1,1}a_1 + \beta_{1,2}a_2$ and Φ^{-1} is the identity link, we can estimate (β_0^*, β_1^*) by OLS in the pseudo-population. However, OLS in the pseudo-population is precisely weighted least squares in the actual study population with weights $1/\{f(A_1)f(A_2 \mid A_1, L)\}$.[27]

Robins (2000, Section 4.3) also noted that the weights W can be replaced by the so-called stabilized weights $SW = \{f(A_1)f(A_2 \mid A_1)\}/\{f(A_1)f(A_2 \mid$

[26] IPTW estimators and IPCW estimators are essentially equivalent. For instance, in the censoring example of Section 2.9, on the event $A_2 = 0$ of being uncensored, the IPCW denominator $\widehat{pr}(A_2 = 0 \mid L, A_1)$ equals $f(A_2 \mid A_1, L)$, the IPTW denominator.

[27] More formally, recall that under (6), $E[Y(a_1, a_2)] = \Phi\{\beta_0^* + \gamma(a_1, a_2; \beta_1^*)\}$ is equal to the g-formula $\int y f^*_{a_1,a_2}(y)\,dy$. Now, given the joint density of the data $f(A_1, L, A_2, Y)$, define

$$\widetilde{f}(A_1, L, A_2, Y) = f(Y \mid A_1, L, A_2)\widetilde{f}_2(A_2)f(L \mid A_1)\widetilde{f}_1(A_1),$$

where $\widetilde{f}_1(A_1)\widetilde{f}_2(A_2)$ are user-supplied densities chosen so that $\widetilde{f}$ is absolutely continuous with respect to f. Since the g-formula depends on the joint density of the data only through $f(Y \mid A_1, L, A_2)$ and $f(L \mid A_1)$, then it is identical under $\widetilde{f}$ and under f. Furthermore, for each a_1, a_2 the g-formula under $\widetilde{f}$ is just equal to $\widetilde{E}[Y \mid A_1 = a_1, A_2 = a_2]$ since, under $\widetilde{f}$, A_2 is independent of

$A_1, L)\}$, and described settings where, for efficiency reasons, using SW is preferable to using W.

MSMs are not restricted to models for the dependence of the mean of $Y(a_1, a_2)$ on (a_1, a_2). Indeed, one can consider MSMs for the dependence of any functional of the law of $Y(a_1, a_2)$ on (a_1, a_2), such as a quantile or the hazard function if Y is a time-to-event variable. If the study is fully randomized, that is, (1) holds, then an MSM model for a given functional of the law of $Y(a_1, a_2)$ is tantamount to an associational model for the same functional of the law of Y conditional on $A_1 = a_1$ and $A_2 = a_2$. Thus, under (1), the MSM model can be estimated using standard methods for estimating the corresponding associational model. If the study is only sequentially randomized, that is, (6) holds but (1) does not, then the model can still be estimated by the same standard methods but weighting each subject by W or SW.

Robins (2000) discussed disadvantages of MSMs compared to SNMs. Here, we summarize some of the main drawbacks. Suppose (9) holds for all $g \in \mathbb{G}$. If the g-null hypothesis (10) is false but the static regime null hypothesis that the law of $Y(a_1, a_2)$ is the same for all (a_1, a_2) is true, then by (iii) of Section 2.6, $f(y \mid A_1 = a_1, A_2 = a_2, L = l)$ will depend on a_2 for some stratum (a_1, l) thus implying a causal effect of A_2 in that stratum; estimation of an SNM model would, but estimation of an MSM model would not, detect this effect. A second drawback is that estimation of MSM models, suffers from marked instability and finite-sample bias in the presence of weights W that are highly variable and skewed. This is not generally an issue in SNM estimation. A third limitation of MSMs is that when (6) fails but an instrumental variable is available, one can still consistently estimate the parameters of a SNM but not of an MSM.[28]

An advantage of MSMs over SNMs that was not discussed in Section 2.8 is the following. MSMs can be constructed that are indexed by easily interpretable parameters that quantify the overall effects of a subset of all possible dynamic regimes (Hernán et al., 2006; van der Laan and Petersen, 2007; Orellana, Rotnitzky and Robins, 2010a, 2010b). As an example consider a longitudinal study of HIV infected patients with baseline CD4 counts exceeding 600 in which we wish to determine the optimal CD4 count at which to begin anti-retroviral treatment. Let g_x denote the dynamic regime that specifies treatment is to be initiated the first time a subject's CD4 count falls below x, $x \in \{1, 2, \ldots, 600\}$. Let $Y(g_x)$ be the associated counterfactual response and suppose few study subjects follow any given regime. If we assume $E[Y(g_x)]$ varies smoothly with x, we can specify and fit (by IPTW) a dynamic regime MSM model $E[Y(g_x)] = \beta_0^* + \beta_1^{*T} h(x)$ where, say, $h(x)$ is a vector of appropriate spline functions.

3. DIRECT EFFECTS

Robins' analysis of sequential regimes leads immediately to the consideration of direct effects. Thus, perhaps not surprisingly, all three of the distinct direct effect concepts that are now an integral part of the causal literature are all to be found in his early papers. Intuitively, all the notions of direct effect consider whether "the outcome (Y) would have been different had cause (A_1) been different, but the level of (A_2) remained unchanged." The notions differ regarding the precise meaning of A_2 "remained unchanged."

3.1 Controlled Direct Effects

In a setting in which there are temporally ordered treatments A_1 and A_2, it is natural to wonder whether the first treatment has any effect on the final outcome were everyone to receive the second treatment. Formally, we wish to compare the potential outcomes $Y(a_1 = 1, a_2 = 1)$ and $Y(a_1 = 0, a_2 = 1)$. Robins (1986, Section 8) considered such contrasts, that are now referred to as *controlled direct effects*. More generally, the *average controlled direct effect of* A_1 *on* Y *when* A_2 *is set to* a_2 is defined to be

$$\text{CDE}(a_2) \equiv E\big[Y(a_1 = 1, a_2) - Y(a_1 = 0, a_2)\big], \tag{15}$$

where $Y(a_1 = 1, a_2) - Y(a_1 = 0, a_2)$ is the individual level direct effect. Thus, if A_2 takes k-levels then there are k such contrasts.

Under the causal graph shown in Figure 5(a), in contrast to Figures 2 and 3, the effect of A_2 on Y is unconfounded, by either measured or unmeasured variables, association is causation and thus, under the associated FFRCISTG model:

$$\text{CDE}(a_2) = E[Y \mid A_1 = 1, A_2 = a_2] - E[Y \mid A_1 = 0, A_2 = a_2].$$

$\{L, A_1\}$. Consequently, for any $q(A_1, A_2)$

$$\begin{aligned} 0 &= \widetilde{E}[q(A_1, A_2)(Y - \Phi\{\beta_0^* + \gamma(A_1, A_2; \beta_1^*)\})] \\ &= E[q(A_1, A_2)\{\widetilde{f}(A_1)\widetilde{f}(A_2)/\{f(A_1)f(A_2 \mid A_1, L)\}\} \\ &\qquad \cdot (Y - \Phi\{\beta_0^* + \gamma(A_1, A_2; \beta_1^*)\})], \end{aligned}$$

where the second equality follows from the Radon–Nikodym theorem. The result then follows by taking $q(A_1, A_2) = v(A_1, A_2)/\{\widetilde{f}(A_1)\widetilde{f}(A_2)\}$.

[28] Note that, as observed earlier, in this case identification is achieved through parametric assumptions made by the SNM.

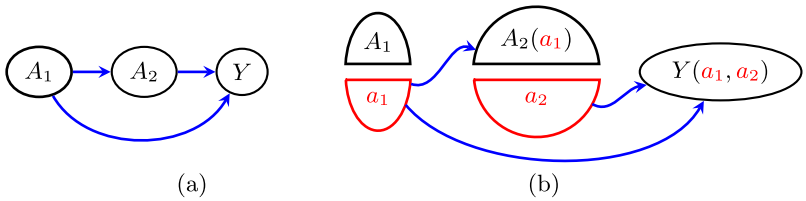

FIG. 5. (a) *A causal DAG $\mathcal{G}$ with no (measured or unmeasured) confounding of A_2 on Y*; (b) *the SWIG $\mathcal{G}(a_1, a_2)$ resulting from intervening on A_1 and A_2.*

The CDE can be identified even in the presence of time-dependent confounding. For example, in the context of the FFRCISTG associated with either of the causal DAGs shown in Figures 2 and 3, the CDE(a_2) will be identified via the difference in the expectations of Y under the g-formula densities $f^*_{a_1=1,a_2}(y)$ and $f^*_{a_1=0,a_2}(y)$.[29]

The CDE requires that the potential outcomes $Y(a_1, a_2)$ be well-defined for all values of a_1 and a_2. This is because the CDE treats both A_2 and A_1 as causes, and interprets "A_2 remained unchanged" to mean "had there been an intervention on A_2 fixing it to a_2."

This clearly requires that the analyst be able to describe a well-defined intervention on the mediating variable A_2.

There are many contexts in which there is no clear well-defined intervention on A_2 and thus it is not meaningful to refer to $Y(a_1, a_2)$. The CDE is not applicable in such contexts.

3.2 Principal Stratum Direct Effects (PSDE)

Robins (1986) considered causal contrasts in the situation described in Section 2.9 in which death from a disease of interest, for example, a heart attack, may be censored by death from other diseases. To describe these contrasts, we suppose A_1 is a treatment of interest, $Y = 1$ is the indicator of death from the disease of interest (in a short interval subsequent to a given fixed time t) and $A_2 = 0$ is the "at risk indicator" denoting the absence of death either from other diseases or the disease of interest prior to time t.

Earlier Kalbfleisch and Prentice (1980) had argued that if $A_2 = 1$, so that the subject does not survive to time t, then the question of whether the subject would have died of heart disease subsequent to t had death before t been prevented is meaningless. In the language of counterfactuals, they were saying (i) that if $A_1 = a_1$ and $A_2 \equiv A_2(a_1) = 1$, the counterfactual $Y(a_1, a_2 = 0)$ is not well-defined and (ii) the counterfactual $Y(a_1, a_2 = 1)$ is never well-defined.

Robins (1986, Section 12.2) observed that if one accepts this then the only direct effect contrast that is well-defined is $Y(a_1 = 1, a_2 = 0) - Y(a_1 = 0, a_2 = 0)$ and that is well-defined only for those subjects who would survive to t regardless of whether they received $a_1 = 0$ or $a_1 = 1$. In other words, even though $Y(a_1, a_2)$ may not be well-defined for all subjects and all a_1, a_2, the contrast:

$$E\big[Y(a_1 = 0, a_2) - Y(a_1 = 1, a_2) \mid A_2(a_1 = 1) = A_2(a_1 = 0) = a_2\big] \tag{16}$$

is still well-defined when $a_2 = 0$. As noted by Robins, this could provide a solution to the problem of defining the causal effect of the treatment A_1 on the outcome Y in the context of censoring by death due to other diseases.

Rubin (1998) and Frangakis and Rubin (1999, 2002) later used this same contrast to solve precisely the same problem of "censoring by death."[30]

In the terminology of Frangakis and Rubin (2002) for a subject with $A_2(a_1 = 1) = A_2(a_1 = 0) = a_2$, the *individual principal stratum direct effect* is defined to be:[31]

$$Y(a_1 = 1, a_2) - Y(a_1 = 0, a_2)$$

(here, A_1 is assumed to be binary). The *average PSDE in principal stratum a_2* is then defined to be

$$\begin{aligned}\mathrm{PSDE}(a_2) &\equiv E\big[Y(a_1 = 1, a_2) - Y(a_1 = 0, a_2) \mid A_2(a_1 = 1) = A_2(a_1 = 0) = a_2\big] \\ &= E\big[Y(a_1 = 1) - Y(a_1 = 0) \mid A_2(a_1 = 1) = A_2(a_1 = 0) = a_2\big],\end{aligned} \tag{17}$$

[29] See (7).

[30] The analysis of Rubin (2004) was also based on this contrast, with A_2 no longer a failure time indicator so that the contrast (16) could be considered as well-defined for any value of a_2 for which the conditioning event had positive probability.

[31] For subjects for whom $A_2(a_1 = 1) \neq A_2(a_1 = 0)$, no principal stratum direct effect (PSDE) is defined.

where the second equality here follows, since $Y(a_1, A_2(a_1)) = Y(a_1)$.[32] In contrast to the CDE, the PSDE has the advantage that it may be defined, via (17), without reference to potential outcomes involving intervention on a_2. Whereas the CDE views A_2 as a treatment, the PSDE treats A_2 as a response. Equivalently, this contrast interprets "had A_2 remained unchanged" to mean "we restrict attention to those people whose value of A_2 would still have been a_2, even under an intervention that set A_1 to a different value."

Although the PSDE is an interesting parameter in many settings (Gilbert, Bosch and Hudgens, 2003), it has drawbacks beyond the obvious (but perhaps less important) ones that neither the parameter itself nor the subgroup conditioned on are nonparametrically identified. In fact, having just defined the PSDE parameter, Robins (1986) criticized it for its lack of transitivity when there is a non-null direct effect of A_1 and A_1 has more than two levels; that is, for a given a_2, the PSDEs comparing $a_1 = 0$ with $a_1 = 1$ and $a_1 = 1$ with $a_1 = 2$ may both be positive but the PSDE comparing $a_1 = 0$ with $a_1 = 2$ may be negative. Robins, Rotnitzky and Vansteelandt (2007) noted that the PSDE is undefined when A_1 has an effect on every subject's A_2, a situation that can easily occur if A_2 is continuous. In that event, a natural strategy would be to, say, dichotomize A_2. However, Robins, Rotnitzky and Vansteelandt (2007) showed that the PSDE in principal stratum a_2^* of the dichotomized variable may fail to retain any meaningful substantive interpretation.

3.3 Pure Direct Effects (PDE)[33]

Once it has been established that a treatment A_1 has a causal effect on a response Y, it is natural to ask what "fraction" of a the total effect may be attributed to a given causal pathway. As an example, consider a RCT in nonhypertensive smokers of the effect of an anti-smoking intervention (A_1) on the outcome myocardial infarction (MI) at 2 years (Y). For simplicity, assume everyone in the intervention arm and no one in the placebo arm quit cigarettes, that all subjects were tested for new-onset hypertension A_2 at the end of the first year, and no subject suffered an MI in the first year. Hence, A_1, A_2 and Y occur in that order. Suppose the trial showed smoking cessation had a beneficial effect on both hypertension and MI. It is natural to consider the query: "What fraction of the total effect of smoking cessation A_1 on MI Y is through a pathway that does not involve hypertension A_2?"

Robins and Greenland (1992) formalized this question via the following counterfactual contrast, which they termed the "pure direct effect":

$$Y\{a_1 = 1, A_2(a_1 = 0)\} - Y\{a_1 = 0, A_2(a_1 = 0)\}.$$

The second term here is simply $Y(a_1 = 0)$.[34] The contrast is thus the difference between two quantities: first, the outcome Y that would result if we set a_1 to 1, while "holding fixed" a_2 at the value $A_2(a_1 = 0)$ that it would have taken had a_1 been 0; second, the outcome Y that would result from simply setting a_1 to 0 [and thus having A_2 again take the value $A_2(a_1 = 0)$]. Thus, the Pure Direct Effect interprets had "A_2 remained unchanged" to mean "had (somehow) A_2 taken the value that it would have taken had we fixed A_1 to 0." The contrast thus represents the effect of A_1 on Y had the effect of A_1 on hypertension A_2 been blocked. As for the CDE, to be well-defined, potential outcomes $Y(a_1, a_2)$ must be well-defined. As a summary measure of the direct effect of (a binary variable) A_1 on Y, the PDE has the advantage (relative to the CDE and PSDE) that it is a single number.

The average pure direct effect is defined as[35]

$$\begin{aligned}\text{PDE} = {} & E\big[Y\{a_1 = 1, A_2(a_1 = 0)\}\big] \\ & - E\big[Y\big(a_1 = 0, A_2(a_1 = 0)\big)\big].\end{aligned}$$

Thus, the ratio of the PDE to the total effect $E[Y\{a_1 = 1\}] - E[Y\{a_1 = 0\}]$ is the fraction of the total that is through a pathway that does not involve hypertension (A_2).

Unlike the PSDE, the PDE is an average over the full population. However, unlike the CDE, the PDE is not nonparametrically identified under the FFRCISTG model associated with the simple DAG shown in Figure 5(a). Robins and Richardson (2011, App. C) computed bounds for the PDE under the FFRCISTG associated with this DAG.

Pearl (2001) obtains identification of the PDE under the DAG in Figure 5(a) by imposing stronger counterfactual independence assumptions, via a Nonparametric Structural Equation Model with Independent Errors

[32] This follows from consistency.

[33] Pearl (2001) adopted the definition given by Robins and Greenland (1992) but changed nomenclature. He refers to the pure direct effect as a "natural" direct effect.

[34] This follows by consistency.

[35] Robins and Greenland (1992) also defined the total indirect effect (TIE) of A_1 on Y through A_2 to be

$$E\big[Y\{a_1 = 1, A_2(a_1 = 1)\}\big] - E\big[Y\{a_1 = 1, A_2(a_1 = 0)\}\big].$$

It follows that the total effect $E[Y\{a_1 = 1\}] - E[Y\{a_1 = 0\}]$ can then be decomposed as the sum of the PDE and the TIE.

(NPSEM-IE).[36] Under these assumptions, Pearl (2001) obtains the following identifying formula:

$$\sum_{a_2}\{E[Y \mid A_1 = 1, A_2 = a_2] - E[Y \mid A_1 = 0, A_2 = a_2]\} \cdot P(A_2 = a_2 \mid A_1 = 0), \tag{20}$$

which he calls the "Mediation Formula."

Robins and Richardson (2011) noted that the additional assumptions made by the NPSEM-IE are not testable, even in principle, via a randomized experiment. Consequently, this formula represents a departure from the principle, originating with Neyman (1923), that causation be reducible to experimental interventions, often expressed in the slogan "no causation without manipulation."[37] Robins and Richardson (2011) achieve a rapprochement between these opposing positions by showing that the formula (20) is equal to the g-formula associated with an intervention on two treatment variables not appearing on the graph (but having deterministic relations with A_1) under the assumption that one of the variables has no direct effect on A_2 and the other has no direct effect on Y. Hence, under this assumption and in the absence of confounding, the effect of this intervention on Y is point identified by (20).[38]

Although there was a literature on direct effects in linear structural equation models (see, e.g., Blalock, 1971) that preceded Robins (1986) and Robins and Greenland (1992), the distinction between the CDE and PDE did not arise since in linear models these notions are equivalent.[39]

3.4 The Direct Effect Null

Robins (1986, Section 8) considered the null hypothesis that $Y(a_1, a_2)$ does not depend on a_1 for all a_2, which we term the *sharp null-hypothesis of no direct effect of* A_1 *on* Y (*relative to* A_2) or more simply as the "sharp direct effect null."

In the context of our running example with data (A_1, L, A_2, Y), under (6) the sharp direct effect null implies the following constraint on the observed data distribution:

$$f^*_{a_1,a_2}(y) \quad \text{is not a function of } a_1 \text{ for all } a_2. \tag{21}$$

Robins (1986, Sections 8 and 9) noted that this constraint (21) is *not* a conditional independence. This is in contrast to the g-null hypothesis which we have seen is equivalent to the independencies in (ii) of Section 2.6 [when equation (9) holds for all $g \in \mathbb{G}$].[40] He concluded that, in contrast to the g-null hypothesis, the constraint (21), and thus the sharp direct effect null, cannot be tested using case control data with unknown case and control sampling fractions.[41] This constraint (21) was later independently discovered by Verma and Pearl (1990) and for this reason is called the "Verma constraint" in the Computer Science literature.

Robins (1999b) noted that, though (21) is not a conditional independence in the observed data distribution, it does correspond to a conditional independence,

[36] In more detail, the FFRCISTG associated with Figures 5(a) and (b) assumes for all a_1, a_2,

$$Y(a_1, a_2), A_2(a_1) \perp\!\!\!\perp A_1, \quad Y(a_1, a_2) \perp\!\!\!\perp A_2(a_1) \mid A_1, \tag{18}$$

which may be read directly from the SWIG shown in Figure 5(b); recall that red nodes are always blocked when applying d-separation. In contrast, Pearl's NPSEM-IE also implies the independence

$$Y(a_1, a_2) \perp\!\!\!\perp A_2(a_1^*) \mid A_1, \tag{19}$$

when $a_1 \neq a_1^*$. Independence (19), which is needed in order for the PDE to be identified, is a "cross-world" independence since $Y(a_1, a_2)$ and $A_2(a_1^*)$ could never (even in principle) both be observed in any randomized experiment.

[37] A point freely acknowledged by Pearl (2012) who argues that causation should be viewed as more primitive than intervention.

[38] This point identification is not a "free lunch": Robins and Richardson (2011) show that it is these additional assumptions that have reduced the FFRCISTG bounds for the PDE to a point. This is a consequence of the fact that these assumptions induce a model *for the original variables* $\{A_1, A_2(a_1), Y(a_1, a_2)\}$ that is a strict submodel of the original FFRCISTG model.

Hence to justify applying the mediation formula by this route one must first be able to specify in detail the additional treatment variables and the associated intervention so as to make the relevant potential outcomes well-defined. In addition, one must be able to argue on substantive grounds for the plausibility of the required no direct effect assumptions and deterministic relations.

It should also be noted that even under Pearl's NPSEM-IE model the PDE is not identified in causal graphs, such as those in Figures 2 and 3 that contain a variable (whether observed or unobserved) that is present both on a directed pathway from A_1 to A_2 and on a pathway from A_1 to Y.

[39] Note that in a linear structural equation model the PSDE is not defined unless A_1 has no effect on A_2.

[40] Results in Pearl (1995b) imply that under the sharp direct effect null the FFRCISTGs associated with the DAGs shown in Figures 2 and 3 also imply inequality restrictions similar to Bell's inequality in Quantum Mechanics. See Gill (2014) for discussion of statistical issues arising from experimental tests of Bell's inequality.

[41] To our knowledge, it is the first such causal null hypothesis considered in Epidemiology for which this is the case.

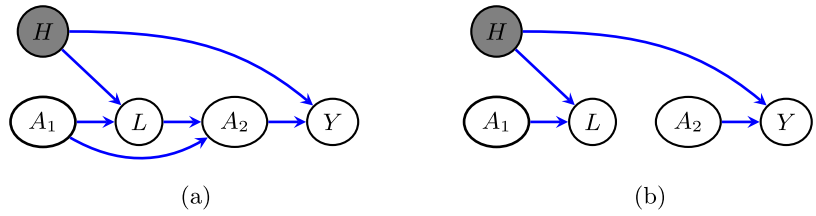

FIG. 6. (a) *A DAG representing the sequentially randomized experiment shown in Figure 2 but where there is no direct effect of* A_1 *on* Y *relative to* A_2; (b) *a DAG representing the pseudo-population obtained by re-weighting the distribution with weights proportional to* $1/f(A_2 \mid L, A_1)$.

but in a weighted distribution with weights proportional to $1/f(A_2 \mid A_1, L)$.[42] This can be understood from the informal discussion following equation (14) in the previous section: there it was noted that given the FFRCISTG corresponding to the DAG in Figure 2, reweighting by $1/f(A_2 \mid A_1, L)$ corresponds to removing both edges into A_2. Hence, if the edges $A_1 \to Y$ and $L \to Y$ are not present, so that the sharp direct effect null holds, as in Figure 6(a), then the reweighted population is described by the DAG in Figure 6(b). It then follows from the d-separation relations on this DAG that $Y \perp\!\!\!\perp A_1 \mid A_2$ in the reweighted distribution.

This fact can also be seen as follows. If, in our running example from Section 2.2, A_1, A_2, Y are all binary, the sharp direct effect null implies that $\beta_1^* = \beta_3^* = 0$ in the saturated MSM with

$$\Phi^{-1}\{E[Y(a_1, a_2)]\} = \beta_0^* + \beta_1^* a_1 + \beta_2^* a_2 + \beta_3^* a_1 a_2.$$

Since β_1^* and β_3^* are the associational parameters of the weighted distribution, their being zero implies the conditional independence $Y \perp\!\!\!\perp A_1 \mid A_2$ under this weighted distribution.

In more complex longitudinal settings, with the number of treatment times k exceeding 2, all the parameters multiplying terms containing a particular treatment variable in a MSM may be zero, yet there may still be evidence in the data that the sharp direct effect null for that variable is false. This is directly analogous to the limitation of MSMs relative to SNMs with regard to the sharp null hypothesis (10) of no effect of any treatment that we noted at the end of Section 2.10. To overcome this problem, Robins (1999b) introduced direct effect structural nested models. In these models, which involve treatment at k time points, if all parameters multiplying a given a_j take the value 0, then we can conclude that the distribution of the observables do not refute the natural extension of (21) to k times. The latter is implied by the sharp direct effect null that a_j has no direct effect on Y holding $a_{j+1}, \ldots, a_k$ fixed.

[42] This observation motivated the development of graphical "nested" Markov models that encode constraints such as (21) in addition to ordinary conditional independence relations; see the discussion of "Causal Discovery" in Section 7 below.

4. THE FOUNDATIONS OF STATISTICS AND BAYESIAN INFERENCE

Robins and Ritov (1997) and Robins and Wasserman (2000) recognized that the lack of robustness of estimators based on the g-formula in a sequential randomized trial with known randomization probabilities had implications for the foundations of statistics and for Bayesian inference. To make their argument transparent, we will assume in our running example (from Section 2.2) that the density of L is known and that $A_1 = 1$ with probability 1 (hence we drop A_1 from the notation). We will further assume the observed data are n i.i.d. copies of a random vector (L, A_2, Y) with A_2 and Y binary and L a $d \times 1$ continuous vector with support on the unit cube $(0, 1)^d$. We consider a model for the law of (L, A_2, Y) that assumes that the density $f^*(l)$ of L is known, that the treatment probability $\pi^*(l) \equiv \Pr(A_2 = 1 \mid L = l)$ lies in the interval $(c, 1-c)$ for some known $c > 0$ and that $b^*(l, a_2) \equiv E[Y \mid L = l, A_2 = a_2]$ is continuous in l. Under this model, the likelihood function is

$$\mathcal{L}(b, \pi) = \mathcal{L}_1(b)\mathcal{L}_2(\pi), \tag{22}$$

where

$$\mathcal{L}_1(b) = \prod_{i=1}^{n} f^*(L_i) b(L_i, A_{2,i})^Y \cdot \{1 - b(L_i, A_{2,i})\}^{1-Y}, \tag{23}$$

$$\mathcal{L}_2(\pi) = \prod_{i=1}^{n} \pi_2(L_i)^{A_{2,i}} \{1 - \pi_2(L_i)\}^{1-A_{2,i}}, \tag{24}$$

and $(b, \pi) \in \mathcal{B} \times \mathbf{\Pi}$. Here $\mathcal{B}$ is the set of continuous functions from $(0, 1)^d \times \{0, 1\}$ to $(0, 1)$ and $\mathbf{\Pi}$ is the set of functions from $(0, 1)^d$ to $(c, 1 - c)$.

We assume the goal is inference about $\mu(b)$ where $\mu(b) = \int b(l,1) f^*(l)\, dl$. Under randomization, that is (3) and (4), $\mu(b^*)$ is the counterfactual mean of Y when treatment is given at both times.

When π^* is unknown, Robins and Ritov (1997) showed that no estimator of $\mu(b^*)$ exists that is uniformly consistent over all $\mathcal{B} \times \mathbf{\Pi}$. They also showed that even if π^* is known, any estimator that does not use knowledge of π^* cannot be uniformly consistent over $\mathcal{B} \times \{\pi^*\}$ for all π^*. However, there do exist estimators that depend on π^* that are uniformly $\sqrt{n}$-consistent for $\mu(b^*)$ over $\mathcal{B} \times \{\pi^*\}$ for all π^*. The Horvitz–Thompson estimator $\mathbb{P}_n\{A_2 Y/\pi^*(L)\}$ is a simple example.

Robins and Ritov (1997) concluded that, in this example, any method of estimation that obeys the likelihood principle such as maximum likelihood or Bayesian estimation with independent priors on b and π, must fail to be uniformly consistent. This is because any procedure that obeys the likelihood principle must result in the same inference for $\mu(b^*)$ regardless of π^*, even when π^* becomes known. Robins and Wasserman (2000) noted that this example illustrates that the likelihood principle and frequentist performance can be in severe conflict in that any procedure with good frequentist properties must violate the likelihood principle.[43] Ritov et al. (2014) in this volume extends this discussion in many directions.

5. SEMIPARAMETRIC EFFICIENCY AND DOUBLE ROBUSTNESS IN MISSING DATA AND CAUSAL INFERENCE MODELS

Robins and Rotnitzky (1992) recognized that the inferential problem of estimation of the mean $E[Y(g)]$ (when identified by the g-formula) of a response Y under a regime g is a special case of the *general problem* of estimating the parameters of an arbitrary semiparametric model in the presence of data that had been coarsened at random (Heitjan and Rubin, 1991).[44]

This viewpoint led them to recognize that the IPCW and IPTW estimators described earlier were not fully efficient. To obtain efficient estimators, Robins and Rotnitzky (1992) and Robins, Rotnitzky and Zhao (1994) used the theory of semiparametric efficiency bounds (Bickel et al., 1993; van der Vaart, 1991) to derive representations for the efficient score, the efficient influence function, the semiparametric variance bound, and the influence function of any asymptotically linear estimator in this *general* problem. The books by Tsiatis (2006) and by van der Laan and Robins (2003) provide thorough treatments. The generality of these results allowed Robins and his principal collaborators Mark van der Laan and Andrea Rotnitzky to solve many open problems in the analysis of semiparametric models. For example, they used the efficient score representation theorem to derive locally efficient semiparametric estimators in many models of importance in biostatistics. Some examples include conditional mean models with missing regressors and/or responses (Robins, Rotnitzky and Zhao, 1994; Rotnitzky and Robins, 1995), bivariate survival (Quale, van der Laan and Robins, 2006) and multivariate survival models with explainable dependent censoring (van der Laan, Hubbard and Robins, 2002).[45]

[43] In response Robins (2004, Section 5.2) offered a Bayes–frequentist compromise that combines honest subjective Bayesian decision making under uncertainty with good frequentist behavior even when, as above, the model is so large and the likelihood function so complex that standard (uncompromised) Bayes procedures have poor frequentist performance. The key to the compromise is that the Bayesian decision maker is only allowed to observe a specified vector function of X [depending on the known $\pi^*(X)$] but not X itself.

[44] Given complete data X, an always observed coarsening variable R, and a known coarsening function $x_{(r)} = c(r,x)$, *coarsening at random* (CAR) is said to hold if $\Pr(R = r \mid X)$ depends only on $X_{(r)}$, the observed data part of X. Robins and Rotnitzky (1992), Gill, van der Laan and Robins (1997) and Cator (2004) showed that in certain models assuming CAR places no restrictions on the distribution of the observed data. For such models, we can pretend CAR holds when our goal is estimation of functionals of the observed data distribution. This trick often helps to derive efficient estimators of the functional. In this section, we assume that the distribution of the observables is compatible with CAR, and further, that in the estimation problems that we consider, CAR may be assumed to hold without loss of generality.

In fact, this is the case in the context of our running causal inference example from Section 2.2. Specifically, let $X = \{Y(a_1,a_2), L(a_1); a_j \in \{0,1\}, j = 1,2\}$, $R = (A_1, A_2)$, and $X_{(a_1,a_2)} = \{Y(a_1,a_2), L(a_1)\}$. Consider a model M_X for X that specifies (i) $\{Y(1,a_2), L(1); a_2 \in \{0,1\}\} \perp\!\!\!\perp \{Y(0,a_2), L(0); a_2 \in \{0,1\}\}$ and (ii) $Y(a_1,1) \perp\!\!\!\perp Y(a_1,0) \mid L(a_1)$ for $a_1 \in \{0,1\}$. Results in Gill and Robins (2001, Section 6) and Robins (2000, Sections 2.1 and 4.2) show that (a) model M_X places no further restrictions on the distribution of the observed data $(A_1, A_2, L, Y) = (A_1, A_2, L(A_1), Y(A_1, A_2))$, (b) given model M_X, the additional independences $X \perp\!\!\!\perp A_1$ and $X \perp\!\!\!\perp A_2 \mid A_1, L$ together also place no further restrictions on the distribution of the observed data (A_1, A_2, L, Y) and are equivalent to assuming CAR. Further, the independences in (b) imply (9) so that $f_{Y(g)}(y)$ is identified by the g-formula $f^*_g(y)$.

[45] More recently, in the context of a RCT, Tsiatis et al. (2008) and Moore and van der Laan (2009), following the strategy of Robins and Rotnitzky (1992), studied variants of the locally efficient tests

In coarsened at random data models, whether missing data or causal inference models, locally efficient semiparametric estimators are also doubly robust (Scharfstein, Rotnitzky and Robins, 1999, pages 1141–1144) and (Robins and Rotnitzky, 2001). See the book (van der Laan and Robins, 2003) for details and for many examples of doubly robust estimators. Doubly robust estimators had been discovered earlier in special cases. In fact, Firth and Bennett (1998) note that the so-called model-assisted regression estimator of a finite population mean of Cassel, Särndal and Wretman (1976) is design consistent which is tantamount to being doubly robust. See Robins and Rotnitzky (2001) for other precursors.

In the context of our running example, from Section 2.2, suppose (6) holds. An estimator $\widehat{\mu}_{\mathrm{dr}}$ of $\mu = E[Y(a_1, a_2)] = f^*_{a_1,a_2}(1)$ for, say $a_1 = a_2 = 1$, is said to be *doubly robust* (DR) if it is consistent when either (i) a model for $\pi(L) \equiv \Pr(A_2 = 1 \mid A_1 = 1, L)$ or (ii) a model for $b(L) \equiv E[Y \mid A_1 = 1, L, A_2 = 1]$ is correct. When L is high dimensional and, as in an observational study, $\pi(\cdot)$ is unknown, double robustness is a desirable property because model misspecification is generally unavoidable, even when we use flexible, high dimensional, semiparametric models in (i) and (ii). In fact, DR estimators have advantages even when, as is usually the case, the models in (i) and (ii) are both incorrect. This happens because the bias of the DR estimator $\widehat{\mu}_{\mathrm{dr}}$ is of second order, and thus generally less than the bias of a non-DR estimator (such as a standard IPTW estimator). By second order, we mean that the bias of $\widehat{\mu}_{\mathrm{dr}}$ depends on the product of the error made in the estimation of $\Pr(A_2 = 1 \mid A_1 = 1, L)$ times the error made in the estimation of $E[Y \mid A_1 = 1, L, A_2 = 1]$.

Scharfstein, Rotnitzky and Robins (1999) noted that the locally efficient estimator of Robins, Rotnitzky and Zhao (1994)

$$\widetilde{\mu}_{\mathrm{dr}} = \{\mathbb{P}_n[A_1]\}^{-1} \cdot \mathbb{P}_n\left[A_1\left\{\frac{A_2}{\widehat{\pi}(L)}Y - \left\{\frac{A_2}{\widehat{\pi}(L)} - 1\right\}\widehat{b}(L)\right\}\right]$$

is doubly robust where $\widehat{\pi}(L)$ and $\widehat{b}(L)$ are estimators of $\pi(L)$ and $b(L)$. Unfortunately, in finite samples this estimator may fail to lie in the parameter space for μ, that is, the interval $[0, 1]$ if Y is binary. In response, Scharfstein, Rotnitzky and Robins (1999) proposed a plug-in DR estimator, the doubly robust regression estimator

$$\widehat{\mu}_{\mathrm{dr,reg}} = \{\mathbb{P}_n[A_1]\}^{-1}\mathbb{P}_n\{A_1\widehat{b}(L)\},$$

where now $\widehat{b}(L) = \mathrm{expit}\{m(L; \widehat{\eta}) + \widehat{\theta}/\widehat{\pi}(L)\}$ and $(\widehat{\eta}, \widehat{\theta})$ are obtained by fitting by maximum likelihood the logistic regression model $\Pr(Y = 1 \mid A_1 = 1, L, A_2 = 1) = \mathrm{expit}\{m(L; \eta) + \theta/\widehat{\pi}(L)\}$ to subjects with $A_1 = 1$, $A_2 = 1$. Here, $m(L; \eta)$ is a user-specified function of L and of the Euclidean parameter η.

Robins (1999a) and Bang and Robins (2005) obtained plug-in DR regression estimators in longitudinal missing data and causal inference models by reexpressing the g-formula as a sequence of iterated conditional expectations.

van der Laan and Rubin (2006) proposed a clever general method for obtaining plug-in DR estimators called targeted maximum likelihood. In our setting, the method yields an estimator $\widehat{\mu}_{\mathrm{dr,TMLE}}$ that differs from $\widehat{\mu}_{\mathrm{dr,reg}}$ only in that $\widehat{b}(L)$ is now given by $\mathrm{expit}\{\widehat{m}(L) + \widehat{\theta}_{\mathrm{greedy}}/\widehat{\pi}(L)\}$ where $\widehat{\theta}_{\mathrm{greedy}}$ is again obtained by maximum likelihood but with a fixed offset $\widehat{m}(L)$. This offset is an estimator of $\Pr(Y = 1 \mid A_1 = 1, L, A_2 = 1)$ that might be obtained using flexible machine learning methods. Similar comments apply to models considered by Bang and Robins (2005). Since 2006 there has been an explosion of research that has produced doubly robust estimators with much improved large sample efficiency and finite sample performance; Rotnitzky and Vansteelandt (2014) give a review.

We note that CAR models are not the only models that admit doubly robust estimators. For example, Scharfstein, Rotnitzky and Robins (1999) exhibited doubly robust estimators in models with nonignorable missingness. Robins and Rotnitzky (2001) derived sufficient conditions, satisfied by many non-CAR models, that imply the existence of doubly robust estimators. Recently, doubly robust estimators have been obtained in a wide variety of models. See Dudik et al. (2014) in this volume for an interesting example.

6. HIGHER ORDER INFLUENCE FUNCTIONS

It may happen that the second-order bias of a doubly-robust estimator $\widehat{\mu}_{\mathrm{dr}}$ decreases slower to 0 with n than $n^{-1/2}$, and thus the bias exceeds the standard error of the estimator. In that case, confidence intervals for μ based on $\widehat{\mu}_{\mathrm{dr}}$ fail to cover at their nominal rate even in large samples. Furthermore, in such a case, in terms of mean squared error, $\widehat{\mu}_{\mathrm{dr}}$ does not optimally trade off bias and variance. In an attempt to address these

and estimators of Scharfstein, Rotnitzky and Robins (1999) to increase efficiency and power by utilizing data on covariates.

problems, Robins et al. (2008) developed a theory of point and interval estimation based on higher order influence functions and use this theory to construct estimators of μ that improve on $\widehat{\mu}_{\mathrm{dr}}$. Higher order influence functions are higher order U-statistics. The theory of Robins et al. (2008) extends to higher order the first order semiparametric inference theory of Bickel et al. (1993) and van der Vaart (1991). In this issue, van der Vaart (2014) gives a masterful review of this theory. Here, we present an interesting result found in Robins et al. (2008) that can be understood in isolation from the general theory and conclude with an open estimation problem.

Robins et al. (2008) consider the question of whether, for estimation of a conditional variance, random regressors provide for faster rates of convergence than do fixed regressors, and, if so, how? They consider a setting in which n i.i.d. copies of (Y, X) are observed with X a d-dimensional random vector, with bounded density $f(\cdot)$ absolutely continuous w.r.t. the uniform measure on the unit cube $(0, 1)^d$. The regression function $b(\cdot) = E[Y \mid X = \cdot]$ is assumed to lie in a given Hölder ball with Hölder exponent $\beta < 1$.[46] The goal is to estimate $E[\mathrm{Var}\{Y \mid X\}]$ under the homoscedastic semiparametric model $\mathrm{Var}[Y \mid X] = \sigma^2$. Under this model, the authors construct a simple estimator $\widehat{\sigma}^2$ that converges at rate $n^{-(4\beta/d)/(1+4\beta/d)}$, when $\beta/d < 1/4$.

Wang et al. (2008) and Cai, Levine and Wang (2009) earlier proved that if $X_i, i = 1, \ldots, n$, are nonrandom but equally spaced in $(0, 1)^d$, the minimax rate of convergence for the estimation of σ^2 is $n^{-2\beta/d}$ (when $\beta/d < 1/4$) which is slower than $n^{-(4\beta/d)/(1+4\beta/d)}$. Thus, randomness in X allows for improved convergence rates even though no smoothness assumptions are made regarding $f(\cdot)$.

To explain how this happens, we describe the estimator of Robins et al. (2008). The unit cube in $\mathbb{R}^d$ is divided into $k = k(n) = n^\gamma$, $\gamma > 1$ identical subcubes each with edge length $k^{-1/d}$. A simple probability calculation shows that the number of subcubes containing at least two observations is $O_p(n^2/k)$. One may estimate σ^2 in each such subcube by $(Y_i - Y_j)^2/2$.[47] An estimator $\widehat{\sigma}^2$ of σ^2 may then be constructed by simply averaging the subcube-specific estimates $(Y_i - Y_j)^2/2$ over all the sub-cubes with at least two observations. The rate of convergence of the estimator is maximized at $n^{-(4\beta/d)/(1+4\beta/d)}$ by taking $k = n^{2/(1+4\beta/d)}$.[48]

Robins et al. (2008) conclude that the random design estimator has better bias control, and hence converges faster than the optimal equal-spaced fixed X estimator, because the random design estimator exploits the $O_p(n^2/n^{2/(1+4\beta/d)})$ random fluctuations for which the X's corresponding to two different observations are only a distance of $O(\{n^{2/(1+4\beta/d)}\}^{-1/d})$ apart.

An Open Problem[49]

Consider again the above setting with random X. Suppose that β/d remains less than $1/4$ but now $\beta > 1$. Does there still exist an estimator of σ^2 that converges at $n^{-(4\beta/d)/(1+4\beta/d)}$? Analogy with other nonparametric estimation problems would suggest the answer is "yes," but the question remains unsolved.[50]

7. OTHER WORK

The available space precludes a complete treatment of all of the topics that Robins has worked on. We provide a brief description of selected additional topics and a guide to the literature.

Analyzing Observational Studies as Nested Randomized Trials

Hernán et al. (2008) and Hernán, Robins and García Rodríguez (2005) conceptualize and analyze observational studies of a time varying treatment as a nested sequence of individual RCTs trials run by nature. Their analysis is closely related to g-estimation of SNM (discussed in Section 2.6). The critical difference is that in these papers Robins and Hernán do not specify a SNM to coherently link the trial-specific ef-

[46] A function $b(\cdot)$ lies in the Hölder ball $H(\beta, C)$ with Hölder exponent $\beta > 0$ and radius $C > 0$, if and only if $b(\cdot)$ is bounded in supremum norm by C and all partial derivatives of $b(x)$ up to order $\lfloor\beta\rfloor$ exist, and all partial derivatives of order $\lfloor\beta\rfloor$ are Lipschitz with exponent $(\beta - \lfloor\beta\rfloor)$ and constant C.

[47] If a subcube contains more than two observations, two are selected randomly, without replacement.

[48] Observe that $E[(Y_i - Y_j)^2/2 \mid X_i, X_j] = \sigma^2 + \{b(X_i) - b(X_j)\}^2/2$, $|b(X_i) - b(X_j)| = O(\|X_i - X_j\|^\beta)$ as $\beta < 1$, and $\|X_i - X_j\| = d^{1/2}O(k^{-1/d})$ when X_i and X_j are in the same subcube. It follows that the estimator has variance of order k/n^2 and bias of order $O(k^{-2\beta/d})$. Variance and the squared bias are equated by solving $k/n^2 = k^{-4\beta/d}$ which gives $k = n^{2/(1+4\beta/d)}$.

[49] Robins has been trying to find an answer to this question without success for a number of years. He suggested that it is now time for some crowd-sourcing.

[50] The estimator given above does not attain this rate when $\beta > 1$ because it fails to exploit the fact that $b(\cdot)$ is differentiable. In the interest of simplicity, we have posed this as a problem in variance estimation. However, Robins et al. (2008) show that the estimation of the variance is mathematically isomorphic to the estimation of θ in the semi-parametric regression model $E[Y \mid A, X] = \theta A + h(X)$, where A is a binary treatment. In the absence of confounding, θ encodes the causal effect of the treatment.

fect estimates. This has benefits in that it makes the analysis easier and also more familiar to users without training in SNMs. The downside is that, in principle, this lack of coherence can result in different analysts recommending, as optimal, contradictory interventions (Robins, Hernán and Rotnitzky 2007).

Adjustment for "Reverse Causation"

Consider an epidemiological study of a time-dependent treatment (say cigarette smoking) on time to a disease of interest, say clinical lung cancer. In this setting, uncontrolled confounding by undetected preclinical lung cancer (often referred to as "reverse causation") is a serious problem. Robins (2008) develops analytic methods that may still provide an unconfounded effect estimate, provided that (i) all subjects with preclinical disease severe enough to affect treatment (i.e., smoking behavior) at a given time t will have their disease clinically diagnosed within the next x, say 2 years and (ii) based on subject matter knowledge an upper bound, for example, 3 years, on x is known.

Causal Discovery

Spirtes, Glymour and Scheines (1993) and Pearl and Verma (1991) proposed statistical methods that allowed one to draw causal conclusions from associational data. These methods assume an underlying causal DAG (or equivalently an FFRCISTG). If the DAG is incomplete, then such a model imposes conditional independence relations on the associated joint distribution (via d-separation). Spirtes, Glymour and Scheines (1993) and Pearl and Verma (1991) made the additional assumption that *all* conditional independence relations that hold in the distribution of the observables are implied by the underlying causal graph, an assumption termed "stability" by Pearl and Verma (1991), and "faithfulness" by Spirtes, Glymour and Scheines (1993). Under this assumption, the underlying DAG may be identified up to a ("Markov") equivalence class. Spirtes, Glymour and Scheines (1993) proposed two algorithms that recover such a class, entitled "PC" and "FCI." While the former presupposes that there are no unobserved common causes, the latter explicitly allows for this possibility.

Robins and Wasserman (1999) and Robins et al. (2003) pointed out that although these procedures were consistent they were not uniformly consistent. More recent papers (Kalisch and Bühlmann, 2007; Colombo et al., 2012) recover uniform consistency for these algorithms by imposing additional assumptions. Spirtes and Zhang (2014) in this volume extend this work by developing a variant of the PC Algorithm which is uniformly consistent under weaker assumptions.

Shpitser et al. (2012, 2014), building on Tian and Pearl (2002b) and Robins (1999b) develop a theory of *nested Markov models* that relate the structure of a causal DAG to conditional independence relations that arise after re-weighting; see Section 3.4. This theory, in combination with the theory of graphical Markov models based on Acyclic Directed Mixed Graphs (Richardson and Spirtes, 2002; Richardson, 2003; Wermuth, 2011; Evans and Richardson, 2014; Sadeghi and Lauritzen, 2014), will facilitate the construction of more powerful[51] causal discovery algorithms that could (potentially) reveal much more information regarding the structure of a DAG containing hidden variables than algorithms (such as FCI) that solely use conditional independence.

Extrapolation and Transportability of Treatment Effects

Quality longitudinal data is often only available in high resource settings. An important question is when and how can such data be used to inform the choice of treatment strategy in low resource settings. To help answer this question, Robins, Orellana and Rotnitzky (2008) studied the extrapolation of optimal dynamic treatment strategies between two HIV infected patient populations. The authors considered the treatment strategies g_x, of the same form as those defined in Section 2.10, namely, "start anti-retroviral therapy the first time at which the measured CD4 count falls below x." Given a utility measure Y, their goal is to find the regime $g_{x_{\text{opt}}}$ that maximizes $E[Y(g_x)]$ in the second low-resource population when good longitudinal data are available only in the first high-resource population. Due to differences in resources, the frequency of CD4 testing in the first population is much greater than in the second and, furthermore, for logistical and/or financial reasons, the testing frequencies cannot be altered. In this setting, the authors derived conditions under which data from the first population is sufficient to identify $g_{x_{\text{opt}}}$ and construct IPTW estimators of $g_{x_{\text{opt}}}$ under those conditions. A key finding is that owing to the differential rates of testing, a necessary condition for identification is that CD4 testing has no direct causal effect on Y not through anti-retroviral therapy. In this issue, Pearl and Bareinboim (2014) study the related question of transportability between populations using graphical tools.

[51] But still not uniformly consistent!

Interference, Interactions and Quantum Mechanics

Within a counterfactual causal model, Cox (1958) defined there to be *interference between treatments* if the response of some subject depends not only on their treatment but on that of others as well. On the other hand, VanderWeele and Robins (2009) defined two binary treatments (a_1, a_2) to be *causally interacting* to cause a binary response Y if for some unit $Y(1,1) \neq Y(1,0) = Y(0,1)$; VanderWeele (2010a) defined the interaction to be *epistatic* if $Y(1,1) \neq Y(1,0) = Y(0,1) = Y(0,0)$. VanderWeele with his collaborators has developed a very general theory of empirical tests for causal interaction of different types (VanderWeele and Robins, 2009; VanderWeele, 2010a, 2010b; VanderWeele and Richardson, 2012).

Robins, VanderWeele and Gill (2012) showed, perhaps surprisingly, that this theory could be used to give a simple but novel proof of an important result in quantum mechanics known as Bell's theorem. The proof was based on two insights: The first was that the consequent of Bell's theorem could, by using the Neyman causal model, be recast as the statement that there is interference between a certain pair of treatments. The second was to recognize that empirical tests for causal interaction can be reinterpreted as tests for certain forms of interference between treatments, including the form needed to prove Bell's theorem. VanderWeele et al. (2012) used this latter insight to show that existing empirical tests for causal interactions could be used to test for interference and spillover effects in vaccine trials and in many other settings in which interference and spillover effects may be present. The papers Ogburn and VanderWeele (2014) and VanderWeele, Tchetgen Tchetgen and Halloran (2014) in this issue contain further results on interference and spillover effects.

Multiple Imputation

Wang and Robins (1998) and Robins and Wang (2000) studied the statistical properties of the multiple imputation approach to missing data (Rubin, 1987). They derived a variance estimator that is consistent for the asymptotic variance of a multiple imputation estimator even under misspecification and incompatibility of the imputation and the (complete data) analysis model. They also characterized the large sample bias of the variance estimator proposed by Rubin (1978b).

Posterior Predictive Checks

Robins, van der Vaart and Ventura (2000) studied the asymptotic null distributions of the posterior predictive p-value of Rubin (1984) and Guttman (1967) and of the conditional predictive and partial posterior predictive p-values of Bayarri and Berger (2000). They found the latter two p-values to have an asymptotic uniform distribution; in contrast they found that the posterior predictive p-value could be very conservative, thereby diminishing its power to detect a misspecified model. In response, Robins et al. derived an adjusted version of the posterior predictive p-value that was asymptotically uniform.

Sensitivity Analysis

Understanding that epidemiologists will almost never succeed in collecting data on all covariates needed to fully prevent confounding by unmeasured factors and/or nonignorable missing data, Robins with collaborators Daniel Scharfstein and Andrea Rotnitzky developed methods for conducting sensitivity analyses. See, for example, Scharfstein, Rotnitzky and Robins (1999), Robins, Rotnitzky and Scharfstein (2000) and Robins (2002, pages 319–321). In this issue, Richardson et al. (2014) describe methods for sensitivity analysis and present several applied examples.

Public Health Impact

Finally, we have not discussed the large impact of the methods that Robins introduced on the substantive analysis of longitudinal data in epidemiology and other fields. Many researchers have been involved in transforming Robins' work on time-varying treatments into increasingly reliable, robust analytic tools and in applying these tools to help answer questions of public health importance.

LIST OF ACRONYMS USED

CAR:	Section 5	coarsened at random.
CD4:	Section 2.2	(medical) cell line depleted by HIV.
CDE:	Section 3.1	controlled direct effect.
CMA:	Section 2.3	causal Markov assumption.
DAG:	Section 2.3	directed acyclic graph.
DR:	Section 5	doubly robust.
dSWIG:	Section 2.4	dynamic single-world intervention graph.
FFRCISTG:	Section 2.2	finest fully randomized causally interpreted structured tree graph.
HIV:	Section 2.2	(medical) human immunodeficiency virus.
IPCW:	Section 2.9	inverse probability of censoring weighted.
IPTW:	Section 2.10	inverse probability of treatment weighted.
ITT:	Section 2.9	intention to treat.
MI:	Section 3.3	(medical) myocardial infarction.
MSM:	Section 2.10	marginal structural model.

NPSEM:	Section 2.2	nonparametric structural equation model.
NPSEM-IE:	Section 2.2	nonparametric structural equation model with independent errors.
PDE:	Section 3.3	pure direct effects.
PSDE:	Section 3.2	principal stratum direct effects.
RCT:	Section 2.2	randomized clinical trial.
SNM:	Section 2.6	structural nested model.
SNDM:	Section 2.6	structural nested distribution model.
SNFTM:	Section 2.6	structural nested failure time model.
SNMM:	Section 2.6	structural nested mean model.
SWIG:	Section 2.3	single-world intervention graph.
TIE:	Section 3.3	total indirect effect.

ACKNOWLEDGMENTS

This work was supported by the US National Institutes of Health Grant R01 AI032475.

REFERENCES

AALEN, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726. MR0491547

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. MR1198884

ARONOW, P. M., GREEN, D. P. and LEE, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *Ann. Statist.* **42** 850–871. MR3210989

BALKE, A. and PEARL, J. (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of the* 12*th Conference on Artificial Intelligence* 1 230–237. MIT Press, Menlo Park, CA.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

BAYARRI, M. J. and BERGER, J. O. (2000). p values for composite null models. *J. Amer. Statist. Assoc.* **95** 1127–1142, 1157–1170. MR1804239

BELLMAN, R. (1957). *Dynamic Programming*. Princeton Univ. Press, Princeton, NJ. MR0090477

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

BLALOCK, H. M., ed. (1971). *Causal Models in the Social Sciences*. Aldine Publishing, Chicago, IL.

CAI, T. T., LEVINE, M. and WANG, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivariate Anal.* **100** 126–136. MR2460482

CASSEL, C. M., SÄRNDAL, C. E. and WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63** 615–620. MR0445666

CATOR, E. A. (2004). On the testability of the car assumption. *Ann. Statist.* **32** 1957–1980. MR2102499

COLOMBO, D., MAATHUIS, M. H., KALISCH, M. and RICHARDSON, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.* **40** 294–321. MR3014308

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York. MR0095561

COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **34** 187–220. MR0341758

COX, D. R. and WERMUTH, N. (1999). Likelihood factorizations for mixed discrete and continuous variables. *Scand. J. Stat.* **26** 209–220. MR1707595

DUDIK, M., ERHAN, D., LANGFORD, J. and LI, L. (2014). Doubly robust policy evaluation and learning. *Statist. Sci.* **29** 485–511.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65** 457–487. MR0521817

EVANS, R. J. and RICHARDSON, T. S. (2014). Markovian acyclic directed mixed graphs for discrete data. *Ann. Statist.* **42** 1452–1482. MR3262457

FIRTH, D. and BENNETT, K. E. (1998). Robust models in probability sampling. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 3–21. MR1625672

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York. MR1100924

FRANGAKIS, C. E. and RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86** 365–379. MR1705410

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

FREEDMAN, D. A. (2006). Statistical models for causation: What inferential leverage do they provide? *Eval. Rev.* **30** 691–713.

GILBERT, E. S. (1982). Some confounding factors in the study of mortality and occupational exposures. *American J. Epidemiology* **116** 177–188.

GILBERT, P. B., BOSCH, R. J. and HUDGENS, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59** 531–541. MR2004258

GILL, R. D. (2014). Statistics, causality and Bell's theorem. *Statist. Sci.* **29** 512–528.

GILL, R. D. and ROBINS, J. M. (2001). Causal inference for complex longitudinal data: The continuous case. *Ann. Statist.* **29** 1785–1811. MR1891746

GILL, R. D., VAN DER LAAN, M. J. and ROBINS, J. M. (1997). Coarsening at random: Characterizations, conjectures, counterexamples. In *Survival Analysis. Proceedings of the First. Seattle Symposium in Biostatistics. Lecture Notes in Statistics* **123** 255–294. Springer, New York.

GUTTMAN, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **29** 83–100. MR0216699

HEITJAN, D. F. and RUBIN, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253. MR1135174

HERNÁN, M. A., ROBINS, J. M. and GARCÍA RODRÍGUEZ, L. A. (2005). Discussion on "Statistical issues in the women's health initiative." *Biometrics* **61** 922–930. MR2216183

HERNÁN, M. A., LANOY, E., COSTAGLIOLA, D. and ROBINS, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology* **98** 237–242.

HERNÁN, M. A., ALONSO, A., LOGAN, R., GRODSTEIN, F., MICHELS, K. B., STAMPFER, M. J., WILLETT, W. C., MANSON, J. E. and ROBINS, J. M. (2008). Observational studies

analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* **19** 766.

HUANG, Y. and VALTORTA, M. (2006). Pearl's calculus of interventions is complete. In *Proceedings of the* 22*nd Conference on Uncertainty in Artificial Intelligence* (*UAI*-06). AUAI Press, Corvallis, OR.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York. MR0570114

KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

KEIDING, N. and CLAYTON, D. (2014). Standardization and control for confounding in observational studies: A historical perspective. *Statist. Sci.* **29** 529–580.

MANSKI, C. (1990). Non-parametric bounds on treatment effects. *American Economic Review* **80** 351–374.

MIETTINEN, O. S. and COOK, E. F. (1981). Confounding: Essence and detection. *American J. Epidemiology* **114** 593–603.

MOHAN, K., PEARL, J. and TIAN, J. (2013). Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems* 26 1277–1285.

MOORE, K. L. and VAN DER LAAN, M. J. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Stat. Med.* **28** 39–64. MR2655550

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 331–366. MR1983752

NEYMAN, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **X** 1–51. In Polish. English translation by D. Dabrowska and T. Speed in *Statist. Sci.* **5** (1990) 463–472.

OGBURN, E. L. and VANDERWEELE, T. J. (2014). Causal diagrams for interference. *Statist. Sci.* **29** 559–578.

ORELLANA, L., ROTNITZKY, A. and ROBINS, J. M. (2010a). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes. Part I: Main content. *Int. J. Biostat.* **6** Art. 8, 49. MR2602551

ORELLANA, L., ROTNITZKY, A. and ROBINS, J. M. (2010b). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes. Part II: Proofs of results. *Int. J. Biostat.* **6** Art. 9, 19. MR2602552

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA. MR0965765

PEARL, J. (1995a). Causal diagrams for empirical research. *Biometrika* **82** 669–710. With discussion and a rejoinder by the author. MR1380809

PEARL, J. (1995b). On the testability of causal models with latent and instrumental variables. In *Proceedings of the* 11*th Annual Conference on Uncertainty in Artificial Intelligence* 435–443. Morgan Kaufmann, San Francisco, CA. MR1615027

PEARL, J. (2000). *Causality. Models*, *Reasoning*, *and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the* 17*th Annual Conference on Uncertainty in Artificial Intelligence* 411–442. Morgan Kaufmann, San Francisco, CA.

PEARL, J. (2012). Eight myths about causality and structural equation models. Technical Report R-393, Computer Science Dept., UCLA.

PEARL, J. and BAREINBOIM, E. (2014). External validity: From do-calculus to transportability across populations. *Statist. Sci.* **29** 579–595.

PEARL, J. and VERMA, T. S. (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning* (*Cambridge*, *MA*, 1991). *Morgan Kaufmann Ser. Represent. Reason.* 441–452. Morgan Kaufmann, San Mateo, CA. MR1142173

PICCIOTTO, S., HERNÁN, M. A., PAGE, J. H., YOUNG, J. G. and ROBINS, J. M. (2012). Structural nested cumulative failure time models to estimate the effects of interventions. *J. Amer. Statist. Assoc.* **107** 886–900. MR3010878

QUALE, C. M., VAN DER LAAN, M. J. and ROBINS, J. R. (2006). Locally efficient estimation with bivariate right-censored data. *J. Amer. Statist. Assoc.* **101** 1076–1084. MR2324147

RICHARDSON, T. S. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* **30** 145–157. MR1963898

RICHARDSON, T. S. and ROBINS, J. M. (2013). Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report 128, Center for Statistics and the Social Sciences, Univ. Washington, Seattle, WA.

RICHARDSON, T. S. and SPIRTES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166

RICHARDSON, A., HUDGENS, M. G., GILBERT, P. B. and FINE, J. P. (2014). Nonparametric bounds and sensitivity analysis of treatment effects. *Statist. Sci.* **29** 596–618.

RITOV, Y., BICKEL, P. J., GAMST, A. C. and KLEIJN, B. J. K. (2014). The Bayesian analysis of complex, high-dimensional models: Can it be CODA? *Statist. Sci.* **29** 619–639.

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. Mathematical models in medicine: Diseases and epidemics. Part 2. *Math. Modelling* **7** 1393–1512. MR0877758

ROBINS, J. M. (1987a). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J. Chronic. Dis.* **40 Suppl 2** 139S–161S.

ROBINS, J. M. (1987b). Addendum to: "A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect." *Comput. Math. Appl.* **14** 923–945. MR0922792

ROBINS, J. M. (1989). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology*: *A Focus on AIDS* (L. Sechrest, H. Freeman and A. Mulley, eds.). U.S. Public Health Service, Washington, DC.

ROBINS, J. M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79** 321–334. MR1185134

ROBINS, J. M. (1993). Analytic methods for estimating HIV-treatment and cofactor effects. In *Methodological Issues in AIDS Behavioral Research* 213–288. Springer, New York.

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185

ROBINS, J. M. (1997a). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to*

Causality (*Los Angeles*, *CA*, 1994). *Lecture Notes in Statist*. **120** 69–117. Springer, New York. MR1601279

ROBINS, J. M. (1997b). Structural nested failure time models. In *Survival Analysis* (P. K. Andersen and N. Keiding, Section eds.), *Encyclopedia of Biostatistics* (P. Armitage and T. Colton, eds.) 4372–4389. Wiley, New York.

ROBINS, J. M. (1998). Marginal structural models. In 1997 *ASA Proceedings of the Section on Bayesian Statistical Science* 1–10. Amer. Statist. Assoc., Alexandria, VA.

ROBINS, J. M. (1999a). Robust estimation in sequentially ignorable missing data and causal inference models. In *ASA Proceedings of the Section on Bayesian Statistical Science* 6–10. Americ. Statist. Assoc., Alexandria, VA.

ROBINS, J. M. (1999b). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In *Computation*, *Causation*, *and Discovery* (C. Glymour and G. Cooper, eds.) 349–405. AAAI Press, Menlo Park, CA. MR1696459

ROBINS, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology*, *the Environment*, *and Clinical Trials* (*Minneapolis*, *MN*, 1997). *IMA Vol. Math. Appl.* **116** 95–133. Springer, New York. MR1731682

ROBINS, J. M. (2002). Comment on "Covariance adjustment in randomized experiments and observational studies" by P. R. Rosenbaum. *Statist. Sci.* **17** 309–321.

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. *Lecture Notes in Statist.* **179** 189–326. Springer, New York. MR2129402

ROBINS, J. M. (2008). Causal models for estimating the effects of weight gain on mortality. *Int. J. Obes.* (*Lond.*) **32 Suppl 3** S15–S41.

ROBINS, J. M. and GREENLAND, S. (1989a) Estimability and estimation of excess and etiologic fractions. *Stat. Med.* **8** 845–859.

ROBINS, J. M. and GREENLAND, S. (1989b). The probability of causation under a stochastic model for individual risk. *Biometrics* **45** 1125–1138. MR1040629

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

ROBINS, J. M., HERNÁN, M. A. and ROTNITZKY, A. (2007). Invited commentary: Effect modification by time-varying covariates. *American J. Epidemiology* **166** 994–1002.

ROBINS, J. M. and MORGENSTERN, H. (1987). The foundations of confounding in epidemiology. *Comput. Math. Appl.* **14** 869–916. MR0922790

ROBINS, J. M., ORELLANA, L. and ROTNITZKY, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Med.* **27** 4678–4721. MR2528576

ROBINS, J. M. and RICHARDSON, T. S. (2011). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology*: *Finding the Determinants of Disorders and Their Cures* (P. Shrout, K. Keyes and K. Ornstein, eds.) 1–52. Oxford Univ. Press, Oxford.

ROBINS, J. M. and RITOV, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Stat. Med.* **16** 285–319.

ROBINS, J. M. and ROTNITZKY, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology* (N. P. Jewell, K. Dietz and V. T. Farewell, eds.) 297–331. Birkhäuser, Boston, MA.

ROBINS, J. M. and ROTNITZKY, A. (2001). Comment on "Inference for semiparametric models: Some questions and an answer," by P. Bickel. *Statist. Sinica* **11** 920–936.

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology*, *the Environment*, *and Clinical Trials* (*Minneapolis*, *MN*, 1997). *IMA Vol. Math. Appl.* **116** 1–94. Springer, New York. MR1731681

ROBINS, J. M., ROTNITZKY, A. and VANSTEELANDT, S. (2007). Discussion of "Principal stratification designs to estimate input data missing due to death" by C. E. Frangakis, D. B. Rubin, M.-W. An and E. MacKenzie. MR2395697. *Biometrics* **63** 650–653. MR2395698

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROBINS, J. M., VANDERWEELE, T. J. and GILL, R. D. (2012). A proof of Bell's inequality in quantum mechanics using causal interactions. Available at arXiv:1207.4913.

ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of p values in composite null models. *J. Amer. Statist. Assoc.* **95** 1143–1167, 1171–1172. MR1804240

ROBINS, J. M. and WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124. MR1766832

ROBINS, J. M. and WASSERMAN, L. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the* 13*th Conference on Uncertainty in Artificial Intelligence* 309–420. Morgan Kaufmann, San Francisco, CA.

ROBINS, J. M. and WASSERMAN, L. (1999). On the impossibility of inferring causation from association without background knowledge. In *Computation*, *Causation*, *and Discovery* (C. Glymour and G. Cooper, eds.) 305–321. MIT Press, Cambridge, MA.

ROBINS, J. M. and WASSERMAN, L. (2000). Conditioning, likelihood, and coherence: A review of some foundational concepts. *J. Amer. Statist. Assoc.* **95** 1340–1346. MR1825290

ROBINS, J. M., BLEVINS, D., RITTER, G. and WULFSOHN, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3** 319–336.

ROBINS, J. M., SCHEINES, R., SPIRTES, P. and WASSERMAN, L. (2003). Uniform consistency in causal inference. *Biometrika* **90** 491–515. MR2006831

ROBINS, J. M., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics*: *Essays in Honor of David A. Freedman*. *Inst. Math. Stat. Collect.* **2** (D. Nolan and T. Speed, eds.) 335–421. IMS, Beachwood, OH. MR2459958

ROTNITZKY, A. and ROBINS, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82** 805–820. MR1380816

ROTNITZKY, A. and VANSTEELANDT, S. (2014). Double-robust methods. In *Handbook of Missing Data Methodology* (G. Fitzmaurice, M. Kenward, G. Molenberghs, A. Tsiatis and G. Verbeke, eds.). Chapman & Hall/CRC Press, Boca Raton, FL.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educational Psychology* **66** 688–701.

RUBIN, D. B. (1978a). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1978b). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse (C/R: P29–34). In *ASA Proceedings of the Section on Survey Research Methods* 20–28. Americ. Statist. Assoc., Alexandria, VA.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

RUBIN, D. B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Stat. Med.* **17** 371–385; discussion 387–389.

RUBIN, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scand. J. Stat.* **31** 161–170. MR2066246

SADEGHI, K. and LAURITZEN, S. (2014). Markov properties for mixed graphs. *Bernoulli* **20** 676–696. MR3178514

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478

SCHULTE, P. J., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2014). *Q*- and *A*-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.* **29** 640–661.

SEKHON, J. S. (2008). The Neyman–Rubin model of causal inference and estimation via matching methods. In *The Oxford Handbook of Political Methodology* (J. M. Box-Steffensmeier, H. E. Brady and D. Collier, eds.) 271–299. Oxford Handbooks Online, Oxford.

SHPITSER, I. and PEARL, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the* 21*st National Conference on Artificial Intelligence* 1219–1226. AAAI Press, Menlo Park, CA.

SHPITSER, I., RICHARDSON, T. S., ROBINS, J. M. and EVANS, R. J. (2012). Parameter and structure learning in nested Markov models. In *Causal Structure Learning Workshop of the* 28*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-12).

SHPITSER, I., EVANS, R. J., RICHARDSON, T. S. and ROBINS, J. M. (2014). Introduction to nested Markov models. *Behaviormetrika* **41** 3–39.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search*. *Lecture Notes in Statistics* **81**. Springer, New York. MR1227558

SPIRTES, P. and ZHANG, J. (2014). A uniformly consistent estimator of causal effects under the k-triangle-faithfulness assumption. *Statist. Sci.* **29** 662–678.

TIAN, J. (2008). Identifying dynamic sequential plans. In *Proceedings of the* 24*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-08) 554–561. AUAI Press, Corvallis, OR.

TIAN, J. and PEARL, J. (2002a). A general identification condition for causal effects. In *Proceedings of the* 18*th National Conference on Artificial Intelligence* 567–573. AUAI Press, Menlo Park, CA.

TIAN, J. and PEARL, J. (2002b). On the testable implications of causal models with hidden variables. In *Proceedings of the* 18*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-02) 519–527. Morgan Kaufmann, San Francisco, CA.

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Stat. Med.* **27** 4658–4677. MR2528575

VANDERWEELE, T. J. (2010a). Epistatic interactions. *Stat. Appl. Genet. Mol. Biol.* **9** Art. 1, 24. MR2594940

VANDERWEELE, T. J. (2010b). Sufficient cause interactions for categorical and ordinal exposures with three levels. *Biometrika* **97** 647–659. MR2672489

VANDERWEELE, T. J. and RICHARDSON, T. S. (2012). General theory for interactions in sufficient cause models with dichotomous exposures. *Ann. Statist.* **40** 2128–2161. MR3059079

VANDERWEELE, T. J. and ROBINS, J. M. (2009). Minimal sufficient causation and directed acyclic graphs. *Ann. Statist.* **37** 1437–1465. MR2509079

VANDERWEELE, T. J. and SHPITSER, I. (2013). On the definition of a confounder. *Ann. Statist.* **41** 196–220. MR3059415

VANDERWEELE, T. J., TCHETGEN TCHETGEN, E. J. and HALLORAN, M. E. (2014). Interference and sensitivity analysis. *Statist. Sci.* **29** 687–706.

VANDERWEELE, T. J., VANDENBROUCKE, J. P., TCHETGEN TCHETGEN, E. J. and ROBINS, J. M. (2012). A mapping between interactions and interference: Implications for vaccine trials. *Epidemiology* **23** 285–292.

VANSTEELANDT, S. and JOFFE, M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statist. Sci.* **29** 707–731.

VAN DER LAAN, M. J., HUBBARD, A. E. and ROBINS, J. M. (2002). Locally efficient estimation of a multivariate survival function in longitudinal studies. *J. Amer. Statist. Assoc.* **97** 494–507. MR1941466

VAN DER LAAN, M. J. and PETERSEN, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *Int. J. Biostat.* **3** Art. 3, 54. MR2306841

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York. MR1958123

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning. Causal Inference for Observational and Experimental Data*. Springer, New York. MR2867111

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11, 40. MR2306500

VAN DER VAART, A. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204. MR1091845

VAN DER VAART, A. (2014). Higher order tangent spaces and influence functions. *Statist. Sci.* **29** 679–686.

VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the* 6*th Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*-90) 220–227. Elsevier, New York.

WANG, N. and ROBINS, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935–948. MR1666715

WANG, L., BROWN, L. D., CAI, T. T. and LEVINE, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *Ann. Statist.* **36** 646–664. MR2396810

WERMUTH, N. (2011). Probability distributions with summary graph structure. *Bernoulli* **17** 845–879. MR2817608