HHS Public Access

Author manuscript

Epidemiology. Author manuscript; available in PMC 2017 November 01.

Published in final edited form as:

Epidemiology. 2016 November ; 27(6): 859–869. doi:10.1097/EDE.0000000000000547.

Diagnostics for confounding of time-varying and other joint exposures

John W. Jackson, Sc.D

[1]Harvard T.H. Chan School of Public Health, Department of Epidemiology

[2]Massachusetts General Hospital, Department of Psychiatry, Schizophrenia Clinical & Research Program, Chester M. Pierce MD Division of Global Psychiatry

Abstract

The effects of joint exposures (or exposure regimes) include those of adhering to assigned treatment vs. placebo in a randomized controlled trial, duration of exposure in a cohort study, interactions between exposures, and direct effects of exposure, among others. Unlike the setting of a single point exposure (e.g. propensity score matching), there are few tools to describe confounding for joint exposures or how well a method resolves it. Investigators need tools that describe confounding in ways that are conceptually grounded and intuitive for those who read, review, and use applied research to guide policy.

We revisit the implications of exchangeability conditions that hold in sequentially randomized trials, and the bias structure that motivates the use of g-methods such as marginal structural models. From these we develop covariate balance diagnostics for joint exposures that can (1) describe time-varying confounding (2) assess whether covariates are predicted by prior exposures given their past, the indication for g-methods (3) describe residual confounding after inverse probability weighting. For each diagnostic we present time-specific metrics that encompass a wide class of joint exposures, including regimes of multivariate time-varying exposures in censored data, with multivariate point exposures as a special case. We outline how to estimate these directly or with regression and how to average them over person-time. Using a simulated example, we show how these metrics can be presented graphically.

This conceptually grounded framework can potentially aid the transparent design, analysis, and reporting of studies that examine joint exposures. We provide easy-to-use tools to implement it.

INTRODUCTION

The effects of joint exposures (or exposure regimes) are of wide interest. These include the effect of adhering to treatment vs. placebo in a randomized controlled trial, duration of exposure in a cohort study, interactions between exposures, and direct effects of exposure, among others. Rarely are they pursued with designs that conditionally randomize each

Corresponding author: John W. Jackson, Sc.D., 677 Huntington Avenue, Boston, MA 02115, john.jackson@mail.harvard.edu.

The author has no conflicts of interest related to this work.

Software, instructions, example data and code are available on the author's academic website: http://www.hsph.harvard.edu/john-jackson/.

exposure given exposure history, so most studies suffer time-varying confounding. To control for this, investigators measure and adjust for time-varying covariates that predict exposures and the outcome.

When covariates are affected by exposure or its unmeasured cause—we call this exposure-covariate feedback—conventional adjustments may yield bias.[1–3] Other approaches known as g-methods avoid this bias.[4] These include g-estimation of structural nested models,[5] the parametric g-formula,[6] and inverse probability weighted estimation of marginal structural models.[7] This last method is by far the most popular and appears in substantive journals.[8–10]

Unfortunately, there are few tools to describe how exposure regimes are confounded. First, some studies report how time-varying covariates distribute over person-time without accounting for exposure history. Second, there are few tools to inform whether g-methods are needed to adjust for confounding. Many investigators justify their choice to use g-methods (or not) on substantive grounds alone. Even those who do use g-methods could obtain more efficient estimates if they knew which covariates they could stratify on without inducing bias. Third, implementing g-methods involves layers of modeling decisions and their impact on residual confounding is usually expressed (if at all) in opaque summary statistics. These summaries may not be intuitive for those who read, review, and use research to guide policy. This paper aims to provide a conceptually-grounded framework to describe confounding of time-varying exposures and other joint exposures, assess whether g-methods are indicated to adjust for it, and evaluate how well certain g-methods resolve it.

DIAGNOSTIC FRAMEWORK

Overview

Our framework contains three diagnostics for confounding. Each involves some form of covariate balance. Diagnostic 1 describes how exposures relate to prior covariates within levels of exposure history (for time-varying confounding). Diagnostic 2 describes how covariates relate to prior exposures given the exposures' past (for exposure-covariate feedback). Diagnostic 3 is a weighted analogue of Diagnostic 1 for residual time-varying confounding in marginal structural models (see eAppendix for the parametric g-formula). These are summarized in Table 1 for a single time-varying exposure, our focus here. In the eAppendix we extend these to multivariate time-varying exposures, of which multivariate point exposures are a special case. We begin with diagnostics for uncensored data, and then address interpretations and adaptations for right-censored data. Afterwards we outline how to estimate these time-specific metrics and their averages over person-time.

We measure covariate balance using the mean difference between any arbitrary level of exposure a' and a common referent a''. This difference contributes to bias expressions for confounding of the average exposure effect in the entire population, and also among the exposed.[11] Other measures could be used, however, including novel bias metrics that incorporate empirical data on the covariate-outcome relationship (see eAppendix). We argue that, for causal inference, these diagnostics are best applied to specific regimes of interest as they evolve. For simplicity we reserve this detail for the Discussion and eAppendix.

We now structurally define our use of the terms time-varying confounding and exposure-covariate feedback. On a causal directed acyclic graph (DAG),[12] time-varying confounding occurs when, for any exposure in a regime, conditioning on its exposure history does not block all backdoor paths linking it to the outcome.[13] Exposure-covariate feedback occurs when any measured covariate (used to block a backdoor path) is affected by exposure or an unmeasured cause of prior exposure.[3] Note that Hernán and Robins have coined this latter phenomenon treatment-confounder feedback.[14] The literature lacks standard nomenclatures for these phenomena.

SIMULATED EXAMPLE AND NOTATION

To illustrate how to report the diagnostics succinctly, we simulated data on a time-varying binary exposure $A(t)$ randomized at baseline, a vector of five binary covariates $C(t) = L(t)$, $M(t)$, $N(t)$, $O(t)$, $P(t)$ that vary over time (some affected by exposure), an unmeasured variable U, and censoring indicators $S(t)$ where 1=censored and 0=otherwise. These data were generated according to the structure shown in Figure 1a (without censoring) and also Figure 1c (with censoring). Our analyses did not require data on the outcome Y so we did not simulate it. Had we done so, it would have been affected by all prior variables. Data were generated for times $t = 0,1,2$ with $S(t)$, $C(t)$, and $A(t)$ occurring in that order. See the eAppendix '**Details of Example Simulations**' for details.

For notation, let $\bar{V}(t)$ be a historical vector of random variable V through time t i.e. $V(0)$, …, $V(t)$. Let k index the amount of time by which a variable precedes time t. Let $V_{\bar{x}}$ represent the counterfactual value that V would have realized had the sequence of $\bar{X}$ been set to regime $\bar{x}$. In our simulation an example regime $\bar{a}$ might be "always exposed" ($A(0) = 1$, $A(1) = 1$, $A(2) = 1$). Let the expression $V \coprod A \mid C$ represent statistical independence between V and A conditional on C. The indicator function $I(\cdot)$ equals one if true and zero otherwise.

Diagnostics for a Time-Varying Exposure without Censoring

Diagnostic 1: time-varying confounding in the study population—The exchangeability assumptions used to estimate causal effects of exposure regimes imply certain data patterns under confounding vs. no confounding. Consider the DAG in Figure 1a where exposure is randomly assigned at baseline but not during follow-up. There is confounding because, for some exposures $A(t)$, conditioning on exposure history $\bar{A}(t-1)$ does not block all backdoor paths linking them to the outcome Y. Conditioning on exposure history in a longitudinal study of this kind would not render each exposure independent of prior covariates that predict the outcome.

When we assume exchangeability given exposure and covariate history $Y_{\bar{a}} \coprod A(t) | \bar{A}(t-1), \bar{C}(t)$, we imply that exposure and covariate histories can be used to block the backdoor paths. But in Figure 1b where each exposure is conditionally randomized given exposure history, the only backdoor paths are through prior exposures. A stronger form of exchangeability $Y_{\bar{a}} \coprod A(t) | \bar{A}(t-1)$ holds and exposure regimes are not confounded.[13] As a result, each exposure is independent of prior covariates given its own history $\bar{C}(t) \coprod A(t) | \bar{A}(t-1)$.[13,15] This testable property—statistical exogeneity—speaks to *measured* confounding.

Investigators address confounding by adjusting for covariates or their surrogates. While there are many ways to select variables for this purpose,[16,17] drawing causal inference presumes that the ones chosen can grant exchangeability $Y_{\bar{a}} \coprod A(t) | \bar{A}(t-1), \bar{C}(t)$. In this setting, every departure from statistical exogeneity is of concern.

Having selected covariates to satisfy exchangeability $Y_{\bar{a}} \coprod A(t) | \bar{A}(t-1), \bar{C}(t)$, one can describe confounding by evaluating, for each exposure measurement, the balance of prior covariates within levels of exposure history (Diagnostic 1):

$$E[C(t-k)|A(t)=a', \overline{A}(t-1)] - E[C(t-k)|A(t)=a'', \overline{A}(t-1)]$$

where $0 \leq k \leq t$. These time-specific metrics could be reported on a trellised plot,[18] as shown in Figure 2. With many measurements one can report metrics for selected times or, as we later describe, averages over person-time. Note that the diagnostic adjusts for exposure history (eFigure 1a). Otherwise it could mislead by capturing imbalance for covariates that do not confound current exposure but are affected by prior exposure (e.g. in Figure 1b there is an association between $A(1)$ and $C(1)$ but it is completely through prior exposure $A(0)$).

Diagnostic 2: exposure-covariate feedback in the study population—Having described the chosen covariate history's departure from statistical exogeneity, the next step is to assess whether covariates are predicted by any prior exposure given its past, to determine whether g-methods are needed to adjust for them. Any estimator that merely stratifies on covariates involved in feedback with exposure could block exposure effects or suffer selection bias, while g-methods avoid these problems.[4] The selection bias would arise when the association between the covariate and a prior exposure is causal (Figure 3a) or through a shared but *unmeasured* cause (Figure 3b).[3] Any metric of exposure-covariate feedback should capture both sources of association by adjusting for covariates that precede the exposure.

One could test for exposure-covariate feedback by regressing covariate measurements on prior exposures and covariates (the test is whether all exposure parameters equal zero).[19] But such models must be correctly specified. An alternative is to examine associations between covariates and prior exposures that are statistically exogenous i.e. independent of prior covariates given exposure history (after inverse probability weighting or propensity score stratification). One could evaluate the weighted balance of every covariate measurement across each prior exposure, within levels of exposure history preceding that exposure (Diagnostic 2a):

$$E\left[W_{a'(t-k)} \times I(A(t-k)=a') \times C(t)|\overline{A}(t-k-1)\right] - E\left[W_{a''(t-k)} \times I(A(t-k)=a'') \times C(t)|\overline{A}(t-k-1)\right]$$

where $1 \leq k \leq t$ and the weight is the stabilized inverse probability of the prior exposure at time $t - k$:

$$W_{a(t-k)}=\frac{P[A(t-k)=a|\overline{A}(t-k-1)]}{P[A(t-k)=a|\overline{A}(t-k-1),\overline{C}(t-k)]}$$

Estimating these weights requires correctly specified models for each exposure measurement given prior exposure and covariate histories (that block all backdoor paths linking that exposure to the outcome). This yields weighted populations where the exposure is statistically exogenous.[20] Any imbalance will reflect the association from that exposureds causal and confounded relationships with the covariate (see eFigure 1b).

The same goal can be achieved through propensity-score stratification. For binary exposures, one could diagnose exposure-covariate feedback by evaluating the balance of every covariate measurement across each prior exposure within levels of its propensity-score (Diagnostic 2b):

$$E\left[C(t)|A(t-k)=a',e_{a'(t-k)}\right]-E\left[C(t)|A(t-k)=a'',e_{a'(t-k)}\right]$$

where $1 \leq k \leq t$ and the propensity score is the probability that the prior exposure equals the non-referent value given the past:

$$e_{a'(t-k)}=P[A(t-k)=a'|\overline{A}(t-k-1),\overline{C}(t-k)]$$

For categorical exposures one must jointly condition on the estimated probability of each non-referent exposure value given the past. Provided correctly specified models for these quantities, exposures will be statistically exogenous among those with the same propensity score[21] or, more generally, the propensity score function[22] (as shown in eFigure 1c). Any imbalance will reflect the association from that exposure's causal and confounded relationships with the covariate. Propensity score values are usually coarsened into strata that retain statistical exogeneity. For a categorical exposure with n levels, predicted values for each level are coarsened into j classes, so that the final strata are locations on a j^{n-1} dimensional grid formed by crossing the classes for all but one exposure level. As shown in Figure 4, metrics from Diagnostic 2a could be reported as a trellised plot and, as with Diagnostic 1, for specific times or as averages over person-time.

Diagnostic 3: residual time-varying confounding in the weighted population— Many investigators fit marginal structural models via inverse probability weights to estimate the average causal effect of an exposure regime. Given enough covariate history to grant exchangeability $Y_{\bar{a}} \perp\!\!\!\perp A(t)|\bar{A}(t-1), \bar{C}(t)$, one could fit a marginal structural model by applying a cumulative inverse probability weight for each exposure measurement, formed from models for each exposure $A(t)$ given covariate history $\bar{C}(t)$ and exposure history $\bar{A}(t-1)$.[7,23] Hernán and Robins[20] proved that if the models for exposure are correctly specified and positivity holds, each exposure in the weighted population would be statistically exogenous. A departure from statistical exogeneity would reflect residual confounding and

can result from positivity violations, finite sample bias, or when the models or weights are misspecified (e.g. by truncating their values).

The current standard for diagnosing residual confounding in the weighted population is to examine whether, for all exposure times, the mean of the stabilized weights diverges from one.[23] The rationale is that, for each covariate stratum, weights that create equally sized exposure groups will average to one. So this practice can alert us to residual confounding, but it will not tell us which covariates to suspect, because it only assesses departures from statistical exogeneity in aggregate. And, although the mean of correctly specified weights must equal one, the mean of misspecified weights could still approach one (when the weights' mean do not equal one for some covariate strata but cancel out across strata). Because this standard can understate residual confounding, it would be better used in concert with other diagnostics.

To further investigate residual confounding, we can examine, for each exposure measurement, the weighted balance of prior covariates within levels of exposure history (Diagnostic 3):

$$E\left[W_{a'(t)} \times I(A(t)=a') \times C(t-k)|\overline{A}(t-1)\right] - E\left[W_{a''(t)} \times I(A(t)=a'') \times C(t-k)|\overline{A}(t-1)\right]$$

where $0 \leq k \leq t$ and the weight is a stabilized cumulative inverse probability weight for current exposure at time t, perhaps

$$W_{a(t)}=\prod_{k=0}^{k=t} \frac{P[A(t-k)=a|\overline{A}(t-k-1)]}{P[A(t-k)=a|\overline{A}(t-k-1),\overline{C}(t-k)]}$$

to target the average causal effect. This diagnostic appropriately adjusts for exposure history (eFigure 1d) and can summarize residual imbalance in any weighted population, even for covariates that the weights ignore. It can be estimated within levels of baseline covariates that appear in both the weight numerator and denominator (that a marginal structural model would also condition on) by using regression models that we later describe. These metrics could be plotted in the same manner as those from Diagnostic 1 (see eFigure 2). An important use of Diagnostic 3 will be to examine residual imbalance under post-hoc model and weight specifications, as we demonstrate in Figure 5.

Diagnostics for a Time-Varying Exposure with Censoring

Until now we have assumed complete follow-up for all subjects. But study-dropout and competing events can affect whether data at time t are observed. In this section we describe how to interpret and adapt the diagnostics under right censoring. In Table 1 we present formulae for diagnostics of a single time-varying exposure that accommodate censoring (see eAppendix Table 1 for multivariate time-varying exposures).

Diagnostics 1 and 3 for Censored Data—Diagnostic 1 can capture selection-bias in censored data. As an example, consider the DAG in Figure 1c where exposure $A(2)$ and

outcome Y are linked through censoring $S(2)$, a collider that has been conditioned on[3] e.g. $A(2) \leftarrow C(1) \rightarrow [S(2)] \leftarrow C(0) \rightarrow Y$. Diagnostic 1 would reflect this selection-bias as an imbalance of $C(0)$ across $A(2)$. Diagnostic 3 can also capture residual selection-bias to the degree that the cumulative inverse probability weights fail to produce statistical exogeneity in the weighted population, leaving paths like $A(2) \leftarrow C(1) \rightarrow [S(2)] \leftarrow C(0) \rightarrow Y$ possibly attenuated but nonetheless intact. Note that Diagnostics 1 and 3, as specified in Table 1, only capture selection-bias through censoring up to time t, but not after that point.

It is important to also examine statistical exogeneity for censoring. Some tools to address censoring view it as another treatment,[7] such that the underlying causal question concerns a joint intervention on exposure and censoring. Assumptions of exchangeability, positivity, and consistency apply for both, even if the implications for bias are somewhat subtle.[24] One can examine statistical exogeneity for exposures $A(t)$ first, and then censoring $S(t)$ second, conditional on uncensored exposure history, both in the actual study population and after applying joint inverse probability weights for exposure and censoring.[7] This is a special case of Diagnostics 1 and 3 for multivariate time-varying exposures (see eAppendix).

Diagnostic 2 for Censored Data—When censoring is associated with exposure and covariates, it can bias assessments of exposure-covariate feedback, even when the query concerns a randomized exposure's effect on a covariate (consider the open path $A(0) \rightarrow [S(2)] \leftarrow C(1) \rightarrow C(2)$ in Figure 1c). Whenever the exposure's association with censoring is not through measured covariates on backdoor paths, our original proposal for Diagnostic 2 will pick up the selection-bias as an imbalance. But this source of imbalance does not motivate the use of g-methods. To see this, consider Figure 1c again. If we studied the effect of a regime involving $A(0)$, $A(1)$ and $A(2)$ on Y while stratifying on $C(0)$, $C(1)$ and $C(2)$, the path $A(0) \rightarrow [S(2)] \leftarrow [C(1)] \rightarrow [C(2)] \rightarrow Y$ would be blocked. The joint effect of $A(0)$, $A(1)$ and $A(2)$ on Y would still suffer selection bias e.g. $A(0) \rightarrow [C(2)] \leftarrow U \rightarrow Y$, but this involves $A(0)$'s causal relationship with $C(2)$. Their association through censoring is irrelevant.

Under strong assumptions, we can modify Diagnostic 2 to recover associations between covariates and prior exposures that do motivate g-methods: ones that are causal or through unmeasured common causes. To do so we must measure the variables needed to block all backdoor paths linking censoring to future covariates. With these we fit models for remaining uncensored at each time $t-k$, e.g.:

$$\text{logit}\, P[S(t-k)=0|\overline{A}(t-k-1), \overline{C}(t-k-1), \overline{S}(t-k-1)=0] = \gamma_0 + \gamma_1' \overline{A}(t-k-1) + \gamma_2' \overline{C}(t-k-1)$$

were $0 \leq k \leq t$. The predicted probabilities from these models become the denominator of time-specific inverse probability weights, one for each time $t-k$. The weight numerators are the probability of remaining uncensored at time $t-k$ among those present at time $t-k-1$. Taking the product of these weights from baseline through time t gives the time-specific cumulative inverse probability weights for each uncensored observation:

$$W_{s(t)} = \prod_{k=0}^{k=t} \frac{P[S(t-k)=0|\overline{S}(t-k-1)=0]}{P[S(t-k)=0|\overline{S}(t-k-1)=0, \overline{A}(t-k-1), \overline{C}(t-k-1)]}$$

Diagnostic 2 is then estimated in the weighted population (for Diagnostic 2a the final weight is the product between this cumulative censoring weight and the time-specific exposure weight). Provided that the weights and models are specified correctly, and no strata are deterministically censored,[25] the selection-bias will be purged from the weighted data where censoring is a random process (see eFigures 3b and 3c). When these conditions are not met there will be residual bias in Diagnostic 2.

Estimating Diagnostic 2 in censored data requires care. While the final metric only uses data from the uncensored at time t, the model for the exposure is fit among the uncensored at time $t-k$. If the cumulative censoring weight is incorporated, models for remaining uncensored at each time $t-k$ are fit among those present at time $t-k-1$. (Note that while the censoring weight denominators condition on prior exposure, the numerators do not because doing so could reintroduce selection-bias via paths like $A(0) \rightarrow \boxed{S(2)} \leftarrow A(1) \rightarrow C(2)$ in Figure 1c). The cumulative censoring weight can be omitted when investigators artificially censor observations based on their exposure history before time $t-k$ to focus Diagnostic 2 on specific regimes (when that is the only censoring mechanism). For reasons elaborated elsewhere,[26,27] it seems prudent to incorporate it whenever censoring is informative of future covariates.

ESTIMATING TIME-SPECIFIC AND SUMMARY DIAGNOSTICS

Each diagnostic can be estimated using a special data structure. With a slight shift in notation, they can all be expressed as weighted measures of covariate balance $Q_{x(\tau)}$ contrasting covariate means measured at time τ_C, across exposure values a' versus a'' measured at time τ, within strata $X(\tau)$ defined by exposure history $H(\tau)$ (i.e. $\bar{A}(\tau-1)$) or the propensity score $e(\tau)$. The distance between exposure and covariate times can be represented as k (which equals $|\tau-\tau_C|$). To estimate $Q_{x(\tau)}$, one could first structure the data so that each row is a person-time observation classified by the exposure value, its measurement time and history; the covariate name, its value and measurement time; and finally a weight that equals one for unweighted diagnostics. Data scientists would call this format “tidy” because the key observation of interest is the pairing of exposure-covariate measurements.[28] For a given covariate, one would take weighted means within joint levels of the exposure value, timing of the exposure and covariate measurements, and strata $X(\tau)$. The results could be used to directly estimate any of the metrics.

The metrics can be averaged over the person-time distribution up to a desired level. For example, when regimes of interest incorporate values from categorical exposures one could begin by standardizing over the non-referent values within levels of history, time, and distance. Going on to standardize over exposure history yields the average balance for each time and distance pair τ, k. Standardizing further over time yields the average balance for each distance k (an analogous plot appears in eFigure 4). One could finally standardize over distance, across all k or among partitions of k (an example plot appears in Figure 6). These summary metrics and their interpretations, which are formally presented in Table 2, will be most useful in longitudinal studies with many measurement times.

Alternatively, one could use models to estimate $Q_{x(\tau)}$ and summary metrics from this same person-time dataset. Take Diagnostic 1 or 3 with a binary exposure, for example. For each covariate name, one could subset to where covariates precede exposure ($\tau_c \leq \tau$) and regress the covariate value $C(\tau_c)$ on the exposure value $A(\tau)$, exposure history $H(\tau)$, exposure time τ, and covariate distance k e.g. a linear model for $E[C(\tau_c)|A(\tau), H(\tau), \tau, k]$, possibly weighted by the inverse probability of exposure and/or censoring. If this model were saturated, parameter combinations involving $A(\tau)$ would yield the same estimates as the formulae in Table 1. This model would be hard to fit, so one might specify a simpler model with just main effects and first-order interactions with exposure $A(\tau)$:

$$E[C(\tau_c)|A(\tau), H(\tau), \tau, k] = \beta_0 + \beta_1 A(\tau) + \beta_2 H(\tau) + \beta_3 \tau + \beta_4 k + A(\tau) \times (\beta_5 H(\tau) + \beta_6 \tau + \beta_7 k)$$

where $\tau_c \leq \tau$. In this model, the balance metric is allowed to vary by exposure history, time, and distance. If $\beta_5 = 0$ then the diagnostic parameter is constant over exposure history $H(\tau)$; if also $\beta_6 = 0$ it is constant over time τ; and if also $\beta_7 = 0$, it is constant over distance k. One could assume these constraints and omit the corresponding parameters. But the summary estimates would be biased if the assumptions made were false. In contrast, pooling estimates by standardization does not require that the diagnostic parameter be constant over any partition of person-time. Nor does it require a correctly specified model for covariate measurements. Although pooling estimates by regression might handle values from continuous exposures and yield summaries for sparse data, these may rely on extrapolations that are hard to justify. The regression approach may be favored when investigators need to condition metrics on baseline covariates. These arguments also apply to models for Diagnostic 2 i.e. models for $E[C(\tau_c)|A(\tau), H(\tau), \tau, k]$ where $\tau < \tau_c$.

DISCUSSION

We have outlined a framework to assess confounding for one or more point exposures or time-varying exposures by examining: (1) departures from statistical exogeneity in the study population (2) whether covariates associate with prior exposures, causally or through unmeasured causes (3) departures from statistical exogeneity in a weighted or stratified population (see eAppendix). Diagnostics 1 and 2 can be used to understand confounding regardless of how investigators choose to control for it. Diagnostic 3 applies after particular methods are chosen. They can be used for regimes involving one or more binary, categorical, or continuous exposures. Importantly, the metrics map to exchangeability conditions that define confounding for joint exposures (Diagnostics 1 and 3), and to the bias that motivates g-methods (Diagnostic 2). We have outlined ways to estimate them directly or with regression models and summarize them over person-time. The results can be reported as trellised plots that portray a wealth of substantive information.

To implement this framework, one can articulate precise contrasts between two or more specific exposure regimes and artificially censor observations once their exposure history becomes incompatible with them (see eAppendix). Diagnostics need only be estimated at times when there is exposure variation given exposure history. When no one follows a regime of interest, though, this represents a fundamental positivity violation. Problems will

also arise under left-censoring or when the start of exposure regimes is unclear. In these situations, one might pursue better data for the causal contrast at hand. Future work could explore alternative implementations for sparse data that leverage functions of exposure history through modeling.

Selecting covariates for causal inference should be deeply guided by subject matter knowledge. Several strategies have been discussed elsewhere[16,17] and they all caution against choosing instruments. Such variables do not predict the outcome, can bias point estimates, and inflate standard errors.[29,30] In longitudinal settings an instrument could be a fixed or relative phenomenon, depending on the outcome. For time-to-event and repeated measures outcomes an early covariate might predict them initially but less so as follow-up accrues. Investigators ultimately choose some set of covariate history for confounding control, and the diagnostics may help illuminate implicit assumptions (see, for example, eFigure 6).

The diagnostics should be applied after choosing covariates to control for confounding. Each imbalance for Diagnostics 1 and 3 is to be addressed unless one's assumptions involve a series of bias cancellations. Any imbalance for Diagnostic 2 would call for using g-methods to adjust for confounding. Some g-methods can segregate covariates into separate groups, one to stratify the analysis, and the other to weight or standardize by; the choice may affect bias and precision. For marginal structural models, weights that adjust for too many covariates may explode. For the parametric g-formula, correct models for each covariate are required. One might choose to weight or standardize only with covariates that show imbalances for Diagnostic 2 and stratify on the rest. Different strategies might be pursued for structural nested models which are less sensitive to positivity violations and do not require models for covariates. These issues deserve further study as conditioning on covariates that precede exposure can induce selection-bias in some cases[31] and this may compound or cancel over time.

Using covariate balance to diagnose confounding has its limits. It does not measure bias at the covariate level or in aggregate. The magnitude and direction of bias depends on the covariate-outcome association and Diagnostics 1 and 3 can be modified to incorporate empirical data on this relationship (see eAppendix). Our proposal does not account for covariate-covariate correlations, though. And the joint distribution of measured covariates can still suffer imbalance even though the marginal means are balanced.[32] This issue can be ameliorated by adapting stricter definitions of balance.[33] But potential bias from unmeasured confounding, unmeasured selection-bias, and also measurement error may still go undetected. Nevertheless, the proposed diagnostics could support design-based analyses that avoid outcome data until the final stage.[34]

We have outlined a framework for diagnosing confounding of exposure regimes. In an era of "Big Data" and significant advances in causal modeling, epidemiologists must remain intimate with data, using diagnostics grounded in the causal theory we rely on. This intimacy can inform our statistical analyses, and can help communicate our methods and their performance to stakeholders. We developed diagnostics and plots that we hope enable

these aspirations and provide easy-to-use R functions (Austria, Vienna) to implement them. SAS macros (Cary, NC) are forthcoming.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author would like to thank Susan Gruber, another anonymous reviewer, and the editor for challenging comments that improved the paper. The author would also like to acknowledge James Robins, Tyler VanderWeele, Xabier Garcia-De-Albeniz, and Jennifer Weuve for very helpful conversations and comments on earlier versions of this paper. Deborah Blacker also provided helpful feedback on the software manual associated with this paper.

The author was funded by the Alonzo Smythe Yerby Fellowship at the Harvard T.H. Chan School of Public Health and also a NIMH Training Grant in Training in Psychiatric Genetics and Translational Research (T32 MH017119).

References

1. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect. Mathematical Modelling. 1986; 7:1393–1512.
2. Robins JM. Addendum to “A new approach to causal inference in mortality studies with sustained exposure periods - Application to the control of the healthy worker survivor effect. Computers and Mathematics with Applications. 1987; 14:923–945.
3. Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004; 15(5):615–25. [PubMed: 15308962]
4. Robins, JM., Hernán, MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice, G.Davidian, M.Verbeke, G., Molenberghs, G., editors. Longitudinal Data Analysis. New York, NY: Chapman and Hall/CRC Press; 2009. p. 553-599.
5. Vansteelandt S, Joffe M. Structural nested models and g-estimation: the partially realized promise. Statistical Science. 2014; 29(4):707–731.
6. Taubman SL, Robins JM, Mittleman MA, Hernan MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. Int J Epidemiol. 2009; 38(6):1599–611. [PubMed: 19389875]
7. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11(5):550–560. [PubMed: 10955408]
8. Yang S, Eaton CB, Lu J, Lapane KL. Application of marginal structural models in pharmacoepidemiologic studies: a systematic review. Pharmacoepidemiol Drug Saf. 2014; 23(6): 560–71. [PubMed: 24458364]
9. Safford MM. Comparative effectiveness research and outcomes of diabetes treatment. JAMA. 2014; 311(22):2275–6. [PubMed: 24915257]
10. Jackson JW, Gagne JJ. SMART designs in observational studies of opioid therapy duration. J Gen Intern Med. 2014; 29(3):429–31. [PubMed: 24449033]
11. VanderWeele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology. 2011; 22(1):42–52. [PubMed: 21052008]
12. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999; 10(1):37–48. [PubMed: 9888278]
13. Robins JM. Association, causation, and marginal structural models. Synthese. 1999; 121:151–179.
14. Hernán, MA., Robins, JM. Longitudinal Causal Inference. In: Wright, JD., editor. International Encyclopedia of the Social & Behavioral Sciences. 2. Oxford: Elsevier; 2015. p. 340-344.

15. Robins, JM. Marginal structural models versus structural nested models as tools for causal inference. In: Holloran, M., Berry, D., editors. Statistical Models in Epidemiology: The Environment and Clinical Trials. New York: Springer-Verlag; 1999. p. 95-134.
16. Sauer BC, Brookhart MA, Roy J, VanderWeele TJ. A review of covariate selection for non-experimental comparative effectiveness research. Pharmacoepidemiol Drug Saf. 2013; 22(11): 1139–45. [PubMed: 24006330]
17. VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat. 2013; 41(1):196–220. [PubMed: 25544784]
18. Becker RA, Cleveland WS, Shyu M. The visual design and control of trellis display. Journal of Computational and Graphical Statistics. 1996; 5(2):123–155.
19. Robins JM, Greenland S, Hu F-C. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. Journal of the American Statistical Association. 1999; 94:687–700.
20. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association. 2001; 96(454):440–448.
21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1):41–55.
22. Imai K, Van Dyk DA. Causal Inference With General Treatment Regimes: Generalizing the Propensity Score. Journal of the American Statistical Association. 2004; 99(467):854–866.
23. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology. 2008; 168(6):656–664. [PubMed: 18682488]
24. Howe CJ, Cole SR, Lau B, Napravnik S, Eron JJ Jr. Selection bias due to loss to follow up in cohort studies. Epidemiology. 2015
25. Howe CJ, Cole SR, Chmiel JS, Muñoz A. Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. American Journal of Epidemiology. 2011; 173(5):569–577. [PubMed: 21289029]
26. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. Stat Methods Med Res. 2012; 21(3):243–56. [PubMed: 21389091]
27. Westreich D. Berkson's bias, selection bias, and missing data. Epidemiology. 2012; 23(1):159–64. [PubMed: 22081062]
28. Wickham H. Tidy Data. Journal of Statistical Software. 2014; 59(10):1–23. [PubMed: 26917999]
29. Pearl, J. On a class of bias-amplifying variables that endanger effect estimates. In: Grunwald, P., Sprites, P., editors. Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010); Corvallis, OR: Association for Uncertainty in Artificial Intelligence; 2010.
30. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol. 2011; 174(11):1213–22. [PubMed: 22025356]
31. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology. 2003; 14(3):300–6. [PubMed: 12859030]
32. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Association, Series A. 2008; 171(Part 2): 481–502.
33. Imbens, GW., Rubin, DB. Causal Inference for Statistics, Social, and Biomedical Sciences. New York, NY: Cambridge University Press; 2015. Chapter 14. Assessing Overlap in Covariate Distributions.
34. Rubin DB. For objective causal inference, design trumps analysis. The Annals of Applied Statistics. 2008; 2(3):808–840.
35. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. Basic Clin Pharmacol Toxicol. 2006; 98(3):237–42. [PubMed: 16611197]
36. VanderWeele TJ. Unmeasured confounding and hazard scales: sensitivity analysis for total, direct, and indirect effects. Eur J Epidemiol. 2013; 28(2):113–7. [PubMed: 23371044]

37. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. Epidemiology. 2009; 20(1):6–13. [PubMed: 19234396]

38. Jackson JW, VanderWeele TJ, Viswanathan A, Blacker D, Schneeweiss S. The explanatory role of stroke as a mediator of the mortality risk difference between older adults who initiate first- versus second-generation antipsychotic drugs. American Journal of Epidemiology. 2014; 180(8):547–852.

39. Jackson JW, VanderWeele TJ, Blacker D, Schneeweiss S. Mediators of first versus second-generation antipsychotic-related mortality in older adults. Epidemiology. 2015; 26(5):700–709. [PubMed: 26035686]

40. Achy-Brou AC, Frangakis CE, Griswold M. Estimating treatment effects of longitudinal designs using regression models on propensity scores. Biometrics. 2010; 66(3):824–833. [PubMed: 19817741]

41. Shinohara RT, Narayan AK, Hong K, Kim HS, Coresh J, Streiff MB, Frangakis CE. Estmiating parsimonious models of longitudinal causal effects using regressions on propensity scores. Statistics in Medicine. 2013; 30(22):3829–3837.

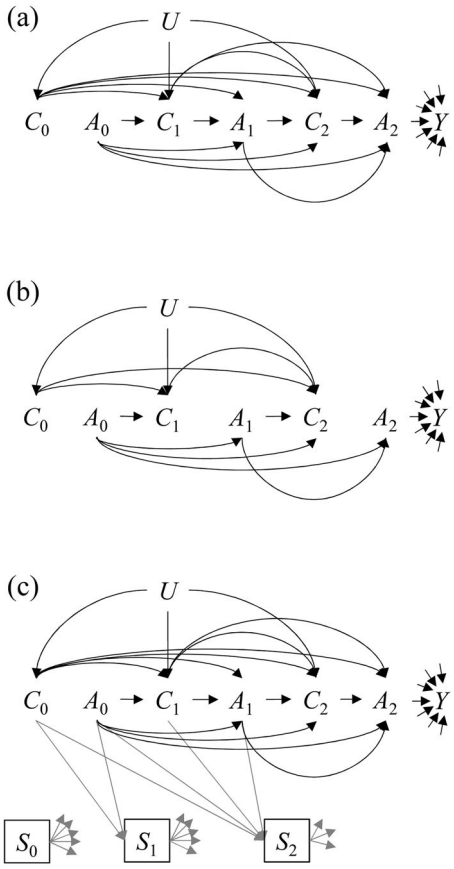

Figure 1.

Causal Directed Acyclic Graphs describing unmeasured covariate U, measured covariates $C(t)$, exposures $A(t)$, and outcome Y for a hypothetical trial where exposure is randomized (a) at baseline only (b) at each time within levels of past exposure (c) at baseline only, with continuing data collection among the uncensored at time t i.e. $S(t) = 0$.

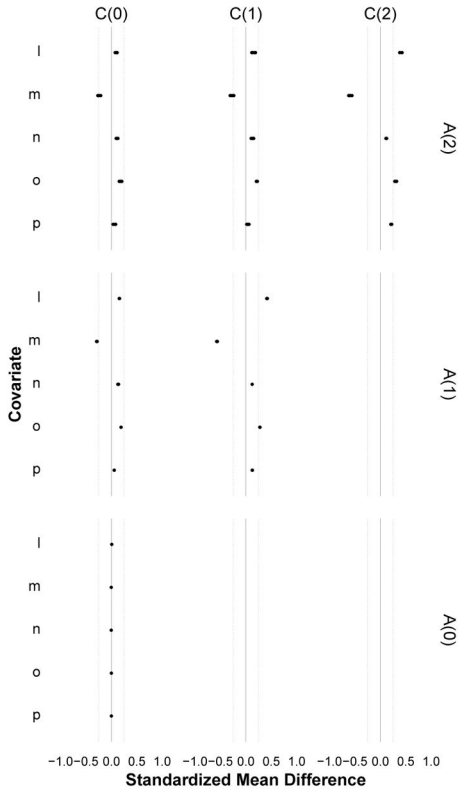

Figure 2.

A trellised covariate balance plot for Diagnostic 1 (time varying confounding), indexed by measurement times for exposure (rows) and covariates (columns). Each subplot evaluates the balance of covariates C (i.e. L,M,N,O,P) at time $t - k$ across levels of exposure $A = 1$ versus $A = 0$ at time t, for each pattern of exposure history through time $t - 1$ (dots). When every dot in a horizontal plane aligns at zero, that exposure measurement is not confounded by that covariate. When this holds for all exposure measurements (across all covariates), the exposure is statistically exogenous and there is no measured time-varying confounding. In the simulated data example shown here, imbalance was largest for the most recent covariates and decayed with increasing distance between exposure and covariate times; there was no confounding at $t = 0$ but the pattern of confounding was the same for $t = 1$ and $t = 2$; the least balanced covariates were L (higher among the exposed) and M (higher among the unexposed), and N was the least imbalanced. Note that here, all observed exposure regimes were examined.

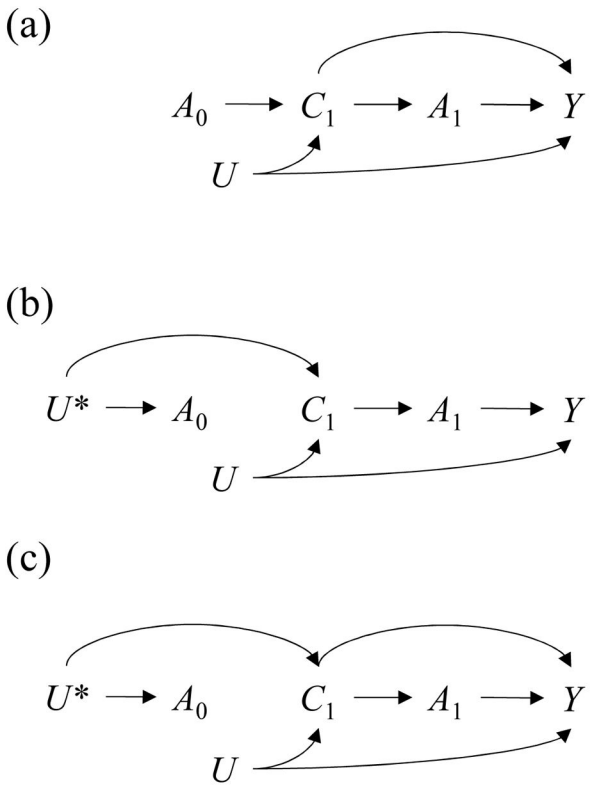

Figure 3.

Causal Directed Acyclic Graphs describing a hypothetical study with exposures $A(t)$, measured covariates $C(t)$, and unmeasured variables U^* and U, and also outcome Y. Each scenario represents a different structure of bias due to exposure-covariate feedback. In scenario (a) stratifying on $C(1)$ will block the effect of $A(0)$ and Y and open the non-causal path $A(0) \rightarrow \boxed{C(1)} \leftarrow U \rightarrow Y$. In scenarios (b) and (c) stratifying on $C(1)$ opens the non-causal pathway $A(0) \leftarrow U^* \rightarrow \boxed{C(1)} \leftarrow U \rightarrow Y$. In all scenarios, adjusting for $C(1)$ by stratification leads to bias. Adjusting by a g-method e.g. removing the arrow from $C(1)$ to $A(1)$ does not induce bias. Note that in (c) there is unmeasured confounding by U^* from the path $A(0) \leftarrow U^* \rightarrow C(1) \rightarrow Y$ and g-methods will not remove it. Because Diagnostic 2 would pick up this unmeasured confounding as an imbalance e.g. $A(0) \leftarrow U^* \rightarrow C(1)$, it is most interpretable when investigators are diagnosing covariate history sufficient enough to support exchangeability assumptions (i.e. no unmeasured confounding).

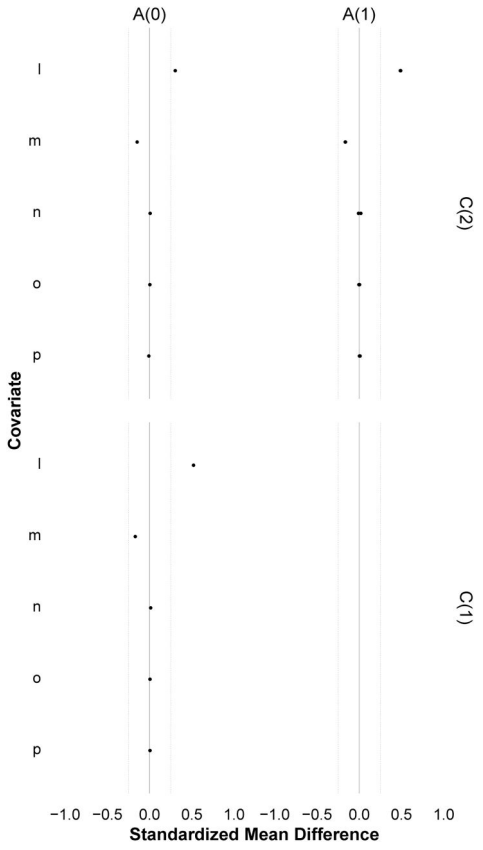

Figure 4.

A trellised covariate balance plot for Diagnostic 2a (exposure-covariate feedback), indexed by measurement times for covariates (rows) and exposure (columns). Each subplot evaluates the balance of covariates C (i.e. L,M,N,O,P) at time t across levels of exposure $A = 1$ versus $A = 0$ at time $t - k$, for each pattern of exposure history through time $t - k - 1$ (dots). For Diagnostic 2b, the plot would pertain to a specific propensity score stratum and unlike Diagnostic 2a, the dots would not represent a specific exposure history. When all the dots in a horizontal plane align at zero, the covariate measurement is independent of every prior exposure and does not contribute to exposure-covariate feedback. When this holds for every covariate measurement, there is no exposure-covariate feedback. In the simulated data example, covariate measurements for L and M were associated with previous exposures: for L the associations were strong and positive; for M they were weaker and negative. In this simulated data, any causal analysis of exposure regimes would want to consider adjusting for L and M using g-methods. Note that here, all observed exposure regimes were examined.

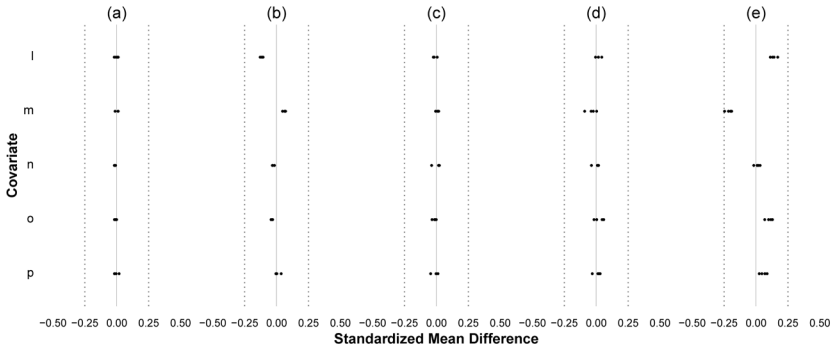

Figure 5.

Demonstration of Diagnostic 3 (residual imbalance in the weighted population) under various sources of residual confounding. Here, we present the balance for covariates and exposures measured at time $t = 2$. Scenario (a) correct specification of the exposure model. Scenario (b) mis-specification of the exposure model for $A(t)$ by only including recent covariates $C(t)$ and omitting covariate-covariate interactions. Scenarios (c) through (e) have correct specification of the exposure model but with random positivity violations built into the data-generating model for $A(t)$ such that $P[A(t)|\bar{A}(t-1), \bar{C}(t)] = 1/10{,}000{,}000$ when $L(t) = 0$ and $O(t) = 1$. In Scenario (c) the weights are not truncated. In (d) they are truncated to the 1st and 99th percentiles. In (e) they are truncated to the 10th and 90th percentiles. Scenarios (b) through (d) all contained residual confounding which appeared as imbalances for Diagnostic 3.

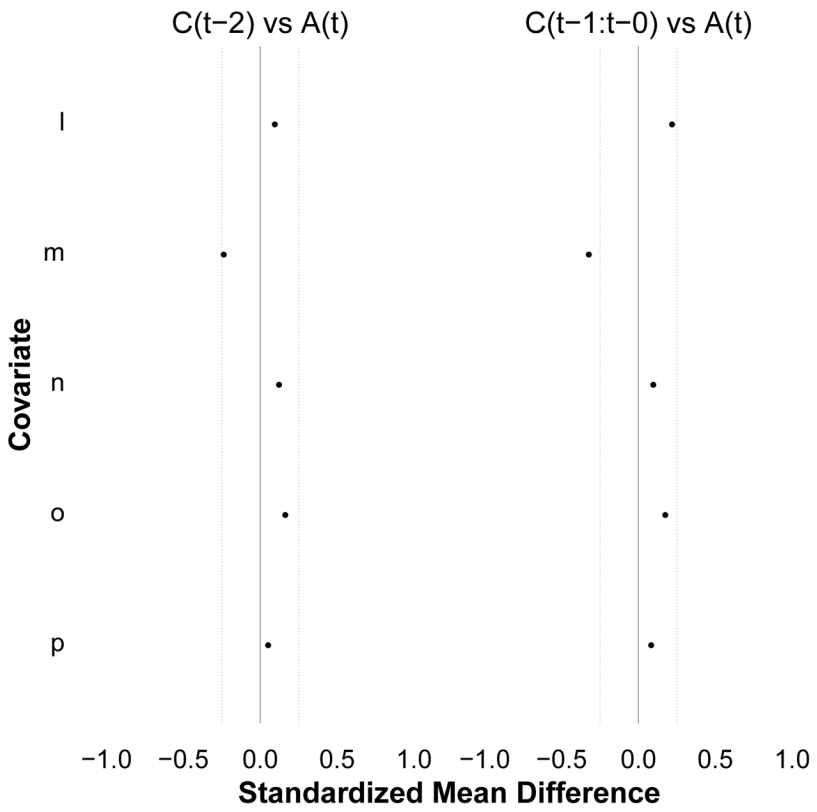

Figure 6.

A trellised plot for Diagnostic 1 among the uncensored, after averaging over exposure history, time, and segments of distance. The panels are indexed by segments of distance between exposure and covariate measurement times (columns). The right panel reports the summary metrics that assess the average balance (adjusting for exposure history) between exposures and proximal covariates at one unit of distance or less: $A(1)$ vs. $C(0)$, $A(2)$ vs. $C(1)$, $A(0)$ vs $C(0)$, $A(1)$ vs $C(1)$, $A(2)$ vs. $C(2)$. The left panel reports the analogous balance metric at two units of distance: $A(2)$ vs. $C(0)$. The average balance across these pools of person-time appear similar, and reflect the same patterns described in Figure 2. Note that here, all observed exposure regimes among the uncensored were examined. Note also that, because censoring is present, this diagnostic should be repeated to assess statistical exogeneity for censoring at time t within levels of uncensored exposure history through time $t - 1$ (e.g. eFigure 5).

Table 1

Time-specific balance metrics for confounding of a time-varying exposure $A(t)$ by time-varying covariates $C(t)$ in the presence of censoring $S(t)$

Diagnostic 1: Time-varying confounding i.e. $C(t-k)$ across levels of $A(t)$ (for all $t \in \{0, \ldots, t\}$ and chosen $k \in \{0, \ldots, t\}$ where $0 \leq k \leq t$)	
Balance metric[a]	Definitions
$E[C(t-k)\|A(t)=a', \bar{A}(t-1), \bar{S}(t)=0] - E[C(t-k)\|A(t)=a'', \bar{A}(t-1), \bar{S}(t)=0]$	N/A
Diagnostic 2: Exposure-covariate feedback i.e. $C(t)$ across levels of $A(t-k)$ (for all $t \in \{0, \ldots, t\}$ and all $k \in \{1, \ldots, t\}$ where $1 \leq k \leq t$)	
Balance metric[a]	
a. by inverse probability weighting	Weight $W_{as(t-k)} = W_{a(t-k)} \times W_{s(t)}$
$E[W_{a's(t-k)} \times I(A(t-k)=a') \times C(t)\|\bar{A}(t-k-1), \bar{S}(t)=0] - E[W_{a''s(t-k)} \times C(t)\|\bar{A}(t-k-1), \bar{S}(t)=0]$	$W_{a(t-k)}=\frac{P[A(t-k)=a\|\overline{A}(t-k-1),\overline{S}(t-k)=0]}{P[A(t-k)=a\|\overline{A}(t-k-1),\overline{C}(t-k),\overline{S}(t-k)=0]} W_{s(t)}=\prod_{k=0}^{k=t}\frac{P[S(t-k)=0\|\overline{S}(t-k-1)=0]}{P[S(t-k)=0\|\overline{S}(t-k-1)=0,\overline{A}(t-k-1),\overline{C}(t-k-1)]}$
b. by propensity score stratification[b]	Propensity score $e_{a(t-k)}$ and weight $W_{s(t)}$
$E[W_{s(t-k)} \times I(A(t-k)=a' \times C(t)\|e_{a'(t-k)}, \bar{S}(t)=0] - E[W_{s(t-k)} \times I(A(t-k)=a'' \times C(t)\|e_a(t-k), \bar{S}(t)=0]$	$e_{a(t-k)} = P[A(t-k)=a'\|\bar{A}(t-k-1), \bar{S}(t-k)=0$ $W_{s(t)}=\prod_{k=0}^{k=t}\frac{P[S(t-k)=0\|\overline{S}(t-k-1)=0]}{P[S(t-k)=0\|\overline{S}(t-k-1)=0,\overline{A}(t-k-1),\overline{C}(t-k-1)]}$
Diagnostic 3: Residual time-varying confounding i.e. $C(t-k)$ across levels of $A(t)$ (for all $t \in \{0, \ldots, t\}$ and chosen $k \in \{0, \ldots, t\}$ where $0 \leq k \leq t$)	
Balance metric[a]	Weight $W_{a(t)}$
$E[W_{a'(t)} \times I(A(t)=a') \times C(t-k)\|\bar{A}(t-1), \bar{S}(t)=0] - E[W_{a''(t)} \times I(A(t)=a'') \times C(t-k)\|\bar{A}(t-1), \bar{S}(t)=0]$	$W_{a(t)}=\prod_{k=0}^{k=t}\frac{P[A(t-k)=a\|\overline{A}(t-k-1),\overline{S}(t-k)=0]}{P[A(t-k)=a\|\overline{A}(t-k-1),\overline{C}(t-k),\overline{S}(t-k)=0]}$

[a] These balance metrics are on the mean difference scale. They can be reported on the standardized mean difference scale by dividing by the (unweighted) pooled standard deviation conditional on exposure history (for Diagnostics 1, 2a and 3) or propensity-score strata (for Diagnostic 2b). Doing so places metrics for binary and continuous covariates on the same scale.

[b] For categorical exposures one needs to jointly condition on the predicted probabilities for each non-referent exposure level

Table 2

Summary balance metrics averaged over person-time by standardization

Time-specific metrics for Diagnostics 1, 2a, and 3 have the general form[a,b]:

$$Q_{H(\tau),\tau,k} = E[W(\tau) \times I(A(\tau) = a') \times C(\tau_c) \mid H(\tau), \tau, k] - E[W(\tau) \times I(A(\tau) = a'') \times C(\tau_c) \mid H(\tau), \tau, k]$$

Where k is the distance between the covariate time τ_c and the exposure time τ i.e. $k = |\tau - \tau_c|$. $H(\tau)$ is the exposure history at time τ i.e. $A(\tau - 1)$; $W(\tau)$ is equal to one or an inverse probability weight, and follows the exposure time τ except for Diagnostic 2 under censoring where it follows covariate time τ_c. Each diagnostic $Q_{H(\tau),\tau,k}$ is estimated from weighted person-time[c]:

$$N_{H(\tau),\tau,k} = \sum W(\tau) \times I(A(\tau) = a') \mid H(\tau), \tau, k + \sum W(\tau) \times I(A(\tau) = a'') \mid H(\tau), \tau, k$$

Summary metric	Example Interpretation for Diagnostic 1
Average over exposure history $H(\tau)$[c]	
$\overline{Q}_{\tau,k} = \sum_{H(\tau)} (Q_{H(\tau),\tau,k} \times N_{H(\tau),\tau,k}) / \sum_{H(\tau)} N_{H(\tau),\tau,k}$	The average balance for exposure A assessed at time 2 and covariate C assessed at time 1 is…
Average over exposure history $H(\tau)$ and time τ	
$\overline{Q}_{k} = \sum_{H(\tau),\tau} (Q_{H(\tau),\tau,k} \times N_{H(\tau),\tau,k}) / \sum_{H(\tau),\tau} N_{H(\tau),\tau,k}$	The average balance for exposure A at any given time τ and covariate C assessed at time $\tau - 1$ is…
Average over exposure history $H(\tau)$, time τ, and distance k[d]	
$\overline{Q} = \sum_{H(\tau),\tau,k} (Q_{H(\tau),\tau,k} \times N_{H(\tau),\tau,k}) / \sum_{H(\tau),\tau,k} N_{H(\tau),\tau,k}$	The average balance for exposure A assessed at any given time and covariate C assessed at any prior time is…

[a] For Diagnostic 2b replace exposure history $H(\tau)$ with propensity score strata $e(\tau)$. The standardization over person-time defined by $e(\tau)$ gives the average balance for an exposure assessed at a specific time τ and a covariate assessed at a specific time τ_C

[b] We omitted censoring here to simplify the expressions but this can be incorporated as in Table 1 or eAppendix Table 1

[c] The time-specific metrics are implicitly indexed by their non-referent value a'. When there are many such values, one can first average metrics over the non-referent person-time distribution within levels of exposure history $H(\tau)$, time τ, distance k. The person-time contributing to this metric is the sum over all exposure levels. This metric can then be averaged over exposure history, time, and distance.

[d] One can first define disjoint ranges of distance p, and then average over distance k for each range p. An example interpretation for Diagnostic 1 would be "the average balance for exposure A assessed at time τ and covariate C assessed anywhere between time $\tau - 1$ and $\tau - 10$."