Marginal Mean Weighting Through Stratification for Time-Varying Multivalued Treatments: A Monte Carlo Comparison with Inverse-Probability-Weighted Marginal Structural Models

Ariel Linden, DrPH
University of California, San Francisco
Department of Medicine
Division of Clinical Informatics & Digital Transformation (DoC-IT)
San Francisco, CA, USA
ariel.linden@ucsf.edu

Abstract

Marginal structural models are commonly estimated using inverse probability of treatment weighting (IPTW) to address time-varying confounding, but direct inversion of estimated treatment probabilities can produce unstable weights. Marginal mean weighting through stratification (MMWS) provides an alternative weight-construction strategy that has been extended to longitudinal multivalued treatments but has received little comparative evaluation. We compared sequential MMWS with multinomial IPTW in a Monte Carlo simulation varying sample size, confounding strength, treatment-confounder feedback, number and prevalence of treatment levels, treatment-model misspecification, and positivity. Both methods used identical fitted treatment models, isolating differences in weight construction. MMWS had lower root mean squared error than IPTW for every treatment contrast across all scenarios, with particularly large advantages under treatment-model misspecification, although it generally incurred modestly greater bias. Under positivity stress, MMWS retained lower root mean squared error but exhibited substantially greater bias and poorer confidence-interval coverage, while exclusions and cumulative-weight restarts revealed increasing support limitations as treatment complexity increased. These findings suggest that MMWS can provide a more stable and robust alternative to conventional IPTW when treatment-level overlap is adequate, but does not eliminate the positivity challenges inherent in complex longitudinal multivalued treatments.

Keywords: marginal structural models; marginal mean weighting through stratification; inverse probability of treatment weighting; multivalued treatments; time-varying confounding; Monte Carlo simulation

1 Introduction

Treatments and exposures frequently vary over time and may take more than two possible values at each treatment occasion. Examples include patients moving over successive visits among no treatment and several alternative treatment modalities [1], and cancer patients whose chemotherapy may be dynamically modified across treatment cycles through changes in dose or treatment delay [2]. Causal inference in this setting is complicated when evolving characteristics both influence subsequent treatment and are themselves affected by prior treatment. Such treatment-confounder feedback creates a dilemma for conventional regression adjustment: controlling for a time-varying confounder is necessary to address confounding of subsequent treatment, but may simultaneously block part of the causal effect of earlier treatment [3–5]. The problem becomes increasingly challenging with multivalued treatments because the number of treatment probabilities and possible treatment histories grows rapidly with the number of treatment occasions.

Marginal structural models (MSMs), typically estimated using inverse probability of treatment weighting (IPTW), were developed to address time-varying confounding of this kind [4–6]. Rather than conditioning on treatment-induced confounders in the outcome model, IPTW constructs a weighted pseudo-population in which measured predictors of treatment are balanced across treatment groups at each occasion. Longitudinal applications of this approach have appeared across several substantive settings [7,8]. Although the foundational MSM literature largely emphasized binary treatments, the framework extends naturally to multivalued treatment by estimating the probability of the observed treatment category at each occasion using a multinomial treatment model. Applications and methodological developments have extended MSMs to multivalued and other complex longitudinal treatment settings, including multiple treatment categories, joint time-varying exposures, and multiple treatment processes [1,2,9–11].

An alternative method for constructing propensity-score weights is marginal mean weighting through stratification (MMWS). MMWS replaces direct inversion of an individual's estimated propensity score with weights based on treatment proportions within propensity-score strata [12]. The method was subsequently generalized to multivalued treatments [13] and, importantly for the present study, to sequences of multivalued treatments [14]. In the longitudinal formulation, treatment probabilities are estimated sequentially conditional on the observed history, occasion-specific MMWS weights are constructed within treatment histories, and the resulting weights are multiplied across occasions. Thus, sequential MMWS and conventional sequential IPTW address the same longitudinal causal-identification problem but differ in how estimated treatment-assignment probabilities are translated into weights. Hong [12,14] argued that the stratification underlying MMWS may make the resulting estimator less sensitive to poorly estimated or extreme propensity scores and consequently more robust and efficient than direct IPTW.

These differences may be especially consequential for time-varying multivalued treatments. With K possible treatment levels observed over T occasions, as many as K^T treatment histories are possible. Increasing either dimension can produce sparse histories, limited overlap, and small estimated treatment probabilities, while treatment-model misspecification can compound across successive occasions. Despite the availability of both approaches, comparative evidence concerning the finite-sample performance of sequential MMWS and conventional multinomial IPTW in this

setting remains limited.

The purpose of this study is therefore to compare marginal mean weighting through stratification with inverse-probability-weighted marginal structural models for estimating the effects of time-varying multivalued treatments. Using Monte Carlo simulation, we examine their performance across conditions varying sample size, confounding strength, treatment-confounder feedback, number and prevalence of treatment levels, propensity-model misspecification, and positivity. We additionally develop a data-adaptive, balance-driven procedure for determining the number and boundaries of propensity-score strata required by MMWS. Performance is evaluated in terms of bias, precision, standard-error calibration, confidence-interval coverage, root mean squared error, and the behavior of the resulting weights. The sequential MMWS and IPTW procedures evaluated here are implemented in the community-contributed Stata package `mmwsl` [15].

2 Methods

2.1 The causal structure of time-varying, multivalued treatment

Consider N subjects observed at treatment occasions $t = 1, \ldots, T$. At occasion t, subject i receives one of K mutually exclusive, exhaustive treatment levels, $A_{it} \in \{0, 1, \ldots, K-1\}$, with level 0 designated the reference level. Let $\bar{A}_{it} = (A_{i1}, \ldots, A_{it})$ denote the observed treatment history through occasion t, and let L_{it} denote the time-varying covariates measured immediately prior to A_{it}, with history $\bar{L}_{it} = (L_{i1}, \ldots, L_{it})$; let U_i denote baseline covariates measured prior to occasion 1. Let Y_{it} denote an outcome ascertained at occasion t; when interest lies in a single outcome measured once, after the final wave, this is simply Y_{iT}. For a fixed treatment history $\bar{a} = (a_1, \ldots, a_T)$, with each $a_t \in \{0, \ldots, K-1\}$, and $\bar{a}_t = (a_1, \ldots, a_t)$ its truncation through occasion t, let $Y_{it}(\bar{a}_t)$ denote the potential value of the occasion-t outcome under history $\bar{a}_t$, and let $Y_{iT}(\bar{a})$ denote the corresponding potential value of a single terminal outcome measured once, after the final wave.

Figure 1 depicts the causal structure this notation is meant to capture, drawn for the $T = 4$-wave, $K = 3$-level case adopted as the baseline of the simulation described in Section 2.6. Two features of this structure are central to everything that follows. First, at every wave, L_t is a confounder of the concurrent treatment A_t: it predicts which treatment level is assigned and, through U and the accumulating history, is also associated with the outcome, so that comparing subjects who received different levels of A_t without accounting for L_t conflates the effect of treatment with pre-existing differences in L_t. Second, L_t is simultaneously a *consequence* of the prior treatment: the arrow $A_{t-1} \to L_t$ in Figure 1 means that later covariates carry forward the effects of earlier treatment decisions, so that L_t lies on the causal pathway between A_{t-1} and Y as well as being a confounder of A_t.

This dual role constitutes treatment-confounder feedback [3]. Conditioning on $\bar{L}_{iT}$ in a conventional outcome regression may block pathways through which earlier treatment affects the outcome, whereas failing to condition on L_{it} leaves subsequent treatment confounded [3–5]. Both estimators considered here address this problem through weighting rather than direct adjustment for time-varying covariates in the outcome model.

Formally, this reweighting strategy is valid only if three conditions hold at every occasion t

[4,5]: sequential conditional exchangeability, $Y_{it}(\bar{a}_t) \perp\!\!\!\perp A_{it} \mid \bar{A}_{i,t-1} = \bar{a}_{t-1}, \bar{L}_{it}$, for every history $\bar{a}_t$; positivity, $0 < P(A_{it} = j \mid \bar{A}_{i,t-1} = \bar{a}_{t-1}, \bar{L}_{it})$ for every treatment level j with nonzero support in the population; and consistency, so that a subject's observed outcome equals the potential outcome under their own observed history. The first condition requires that every time-varying confounder of subsequent treatment be measured and included in L_{it} — graphically, that no arrow into A_t originates from an unmeasured node in Figure 1; the second requires that every treatment level remain genuinely attainable at every wave and history, a condition stressed directly by the "positivity stress" simulation scenario (Section 2.6); the third requires the observed outcome under the treatment history actually received to equal the corresponding potential outcome, and we additionally assume no interference between subjects and well-defined treatment levels. Sections 2.2 and 2.3 describe two different weight-construction strategies — direct inverse-probability weighting, and marginal mean weighting through stratification — that both target these same identifying conditions, and differ only in how the fitted treatment-assignment model at each wave is converted into a subject's weight.

Two related causal quantities can be defined within this framework. An MSM may target $E[Y_{iT}(\bar{a})]$, a function of the complete treatment history $\bar{a}$. The present study instead evaluates the wave-specific marginal effect of current treatment A_{it} on Y_{it}, after weighting for treatment and covariate history through occasion t. This estimand corresponds to longitudinal weighting formulations in which weights and treatment effects are estimated sequentially at each occasion [7,14]. Equation (2) in Section 2.2 states the corresponding marginal structural model.

2.2 Marginal structural models and sequential multinomial IPTW

Building on the causal structure and identification conditions established in Section 2.1, a marginal structural model (MSM) specifies a model for the marginal mean of the potential outcome as a function of treatment history,

$$E[Y_{iT}(\bar{a})] = g(\bar{a}; \boldsymbol{\beta}), \tag{1}$$

for a known link function $g(\cdot)$ and parameter vector $\boldsymbol{\beta}$ [4–6]. Because the number of distinct histories $\bar{a}$ grows as K^T, applied MSMs typically summarize history through a lower-dimensional function of $\bar{a}$ — for example, current treatment level, cumulative exposure to a given level, or a small set of history classes — rather than saturating on every possible sequence [1,7]. As established in Section 2.1, the wave-specific estimand evaluated in this paper summarizes history through current treatment level alone, giving a marginal mean of the occasion-t outcome Y_{it},

$$E[Y_{it}(a)] = \beta_0 + \sum_{j=1}^{K-1} \beta_j \mathbb{1}(a = j), \tag{2}$$

so that β_j is the marginal causal contrast between treatment level j and the reference level 0 at occasion t.

The most widely used estimator for $\boldsymbol{\beta}$ is inverse probability of treatment weighting (IPTW) [4,5]. At each occasion, a treatment-assignment model — for $K > 2$, a multinomial model — is fit

for the conditional probability of the observed treatment level given the observed history,

$$\pi_{it}(j) \;=\; P\big(A_{it} = j \mid \bar{A}_{i,t-1}, \bar{L}_{it}\big)\,, \qquad j = 0, 1, \ldots, K-1, \tag{3}$$

and, optionally, a reduced "numerator" model conditioning on a coarser history (e.g., $\bar{A}_{i,t-1}$ alone) is fit to stabilize the resulting weights. The occasion-specific contribution to subject i's weight is $\hat{\pi}^{\text{num}}_{it}(a_{it})/\hat{\pi}_{it}(a_{it})$, and the cumulative (stabilized) sequential weight through occasion t is the product of these occasion-specific contributions,

$$sw_{it} \;=\; \prod_{h=1}^{t} \frac{\hat{\pi}^{\text{num}}_{ih}(a_{ih})}{\hat{\pi}_{ih}(a_{ih})}, \tag{4}$$

with sw_{iT}, evaluated at the final wave, the special case $t = T$, and with the unstabilized weight obtained by setting all numerator probabilities equal to 1, $\hat{\pi}^{\text{num}}_{ih}(a_{ih}) \equiv 1$ for $h = 1, \ldots, t$. Fitting model (2) (or its history-level generalization), pooled across occasions, by weighted least squares with weight sw_{it} for the wave-t observation, and a subject-clustered sandwich variance estimator to account for within-subject dependence across repeated observations, yields the IPTW estimator of $\boldsymbol{\beta}$ evaluated in this paper. The finite-sample calibration of this variance estimator, whether it adequately reflects the additional uncertainty contributed by using estimated rather than known weights, is assessed empirically in Section 2.7 via the SE/SD ratio and confidence-interval coverage. This sequential-multinomial-IPTW construction generalizes directly from the binary treatment case originally developed by Robins and colleagues [3,4,6] to multivalued and multiple treatments [1,9,11], and remains the natural comparator against which alternative weight-construction strategies are evaluated. Related refinements to how the treatment-assignment weights themselves are estimated, rather than to the underlying causal framework, include covariate-balancing propensity scores [16] and jointly calibrated treatment/censoring weights [10]; these are not evaluated here, as our interest is in weight *construction strategy* (stratification versus direct inverse-probability weighting) rather than in the estimation method for the underlying treatment-assignment probabilities themselves.

2.3 Marginal mean weighting through stratification for longitudinal multivalued treatments

Marginal mean weighting through stratification (MMWS) was introduced by Hong [12], building on an approach first described by Huang et al. [17], as an alternative to direct inverse-probability weighting for a single (cross-sectional) treatment occasion. Rather than using the estimated propensity score directly in the denominator of a weight, MMWS stratifies subjects on the estimated propensity score and constructs a weight from the resulting within-stratum treatment proportions, an approach subsequently generalized to multivalued and multiple treatments [13].

For a single occasion with K treatment levels, a multinomial model provides, for each subject i, one estimated propensity score per treatment level, $\hat{p}_{ij} = \hat{P}(A_i = j \mid L_i)$ for $j = 0, \ldots, K-1$. For each level j in turn, the sample is stratified on $\hat{p}_{\cdot j}$ into a set of strata $s = 1, \ldots, S_j$ (Section 2.4 describes how S_j and the stratum boundaries are chosen). Within stratum s of the

j-th stratification, the MMWS weight assigned to a subject with observed treatment level j is

$$w_i \;=\; \frac{n_s}{n_{sj}} \times \frac{n_j}{N}, \qquad i : A_i = j,\; i \in s, \tag{5}$$

where n_s is the number of subjects (of any treatment level) in stratum s, n_{sj} is the number of subjects with observed level j in stratum s, n_j is the total number of subjects with observed level j in the full sample, and N is the total sample size. The first factor, n_s/n_{sj}, reweights level-j subjects in stratum s so that, within the stratum, they represent the stratum's full composition rather than only its level-j members; the second factor, n_j/N, rescales this within-stratum reweighting so that level j's weighted total, summed across all strata, reproduces its actual marginal (sample-wide) prevalence — the "marginal mean" property for which the method is named.

Hong [14] (Chapter 8) extends this construction to a longitudinal treatment process by applying it separately within each treatment occasion and each distinct observed treatment history up to that occasion. Let g index the set of subjects who share an identical observed treatment history $\bar{A}_{i,t-1}$ through occasion $t-1$ (a *treatment-history subgroup*). Within wave t and history subgroup g, a multinomial model for A_t is fit using only the covariates and subjects belonging to that subgroup, yielding per-level propensity scores $\hat{p}_{ij}^{(t,g)}$; these are stratified as above, and an occasion-specific weight is computed using the same marginal-mean formula, now indexed by (t, g):

$$w_{it}^{(t)} \;=\; \frac{n_{tg,s}}{n_{tg,s,j}} \times \frac{n_{tg,j}}{n_{tg}}, \tag{6}$$

the direct K-category generalization of Hong's per-occasion binary weight. The cumulative sequential MMWS weight through occasion t is then the product of these occasion-specific weights,

$$W_{it} \;=\; \prod_{h=1}^{t} w_{ih}^{(h)}, \tag{7}$$

with W_{iT}, evaluated at the final wave, the special case $t = T$, and mirroring the multiplicative structure of the sequential IPTW weight in equation (4), but replacing each occasion's inverse-probability contribution with a stratification-based marginal-mean contribution. Fitting model (2), pooled across occasions, by weighted least squares with weight W_{it} for the wave-t observation, again using a subject-clustered sandwich variance estimator, yields the MMWS estimator of $\boldsymbol{\beta}$ evaluated against sequential IPTW in Section 2.6. Because stratifying on the propensity score, rather than inverting it directly, smooths the influence of any single subject's estimated propensity score on their weight, Hong argued that this construction may be less sensitive to extreme or poorly estimated propensity scores than direct IPTW while relying on the same identifying assumptions given in Section 2.1 [12,14].

We implemented equations (6)–(7) in the Stata command `mmwsl` [15], which proceeds sequentially by wave and observed treatment history. Within each wave-history subgroup, the command fits the multinomial treatment model, obtains treatment-specific propensity scores, constructs the occasion-specific MMWS weights (Section 2.4 describes how the underlying strata are formed), and accumulates them through the current occasion. For direct comparison, `mmwsl` also constructs

stabilized sequential IPTW weights from the same fitted treatment models, ensuring that the two estimators differ only in how the estimated treatment probabilities are converted into weights.

2.4 Data-adaptive propensity-score stratification

The MMWS estimator requires specification of both the number of propensity-score strata and their boundaries. The original cross-sectional MMWS proposal and its longitudinal extension do not provide a general rule for making these choices. Although five equal-frequency strata have traditionally been used based on results for binary propensity-score subclassification [18,19], a fixed stratum count is poorly suited to the longitudinal multivalued setting, in which sample size and treatment prevalence may vary substantially across treatment occasions, treatment histories, and treatment levels.

We therefore developed a data-adaptive, balance-driven procedure that determines the number and boundaries of strata separately for each treatment-specific propensity score, using an iterative search rather than a single fixed quantile count. The procedure begins by partitioning the propensity score into an automatically determined number of equal-frequency quantile bins: when the starting count is not specified by the user, it is set to

$$S_0 \;=\; \max\!\left(5,\; \left\lfloor \frac{N \cdot \hat{\pi}_{\min}}{100} \right\rfloor\right), \quad \text{capped at a user-set ceiling } S_{\max}, \tag{8}$$

where N is the analytic sample size and $\hat{\pi}_{\min}$ is the observed sample share of the rarest treatment level, so that the initial partition targets approximately 100 members of the rarest level per starting bin (subject to a floor of five bins). The target of 100 is a pragmatic starting value rather than a formally optimal one: it is intended only to keep the number of starting bins reasonable, and thin enough that the subsequent balance-driven splitting described below has room to operate, across the range of sample sizes and rarest-level prevalences considered here; because S_0 is merely a starting point for iterative refinement, and is bounded above by $S_{\max}$, the final number of strata is intended to be less sensitive to this particular choice. Each candidate stratum is then refined independently: within a stratum, the maximum pairwise standardized mean difference (SMD) of the propensity score is computed over every pair of treatment levels present in that stratum, and the stratum is retained if this maximum SMD is at or below a pre-specified balance threshold, by default 0.25, following the convention proposed by Rubin [20]. A stratum failing this criterion is split into two new sub-strata (again by equal-frequency quantiles of the propensity score restricted to that stratum's own members) and each is independently re-evaluated. A stratum lacking even one member of some treatment level present elsewhere in the data is excluded outright: no amount of further splitting can manufacture a missing comparator, so this is treated as a stratum-level positivity failure rather than a balance failure, and the observations it contains are left without a defined MMWS weight. Finally, because further splitting a small stratum can produce cells too thin to yield a stable SMD, a stratum whose thinnest present treatment level falls below the per-group reliability floor

$$n_{\min} \;=\; \left\lceil \frac{5.4}{\mathrm{SMD}_{\mathrm{crit}}^2} \right\rceil \tag{9}$$

is retained rather than split further. This floor follows from the standard large-sample approximation for the variance of an estimated standardized mean difference between two groups of size n each, $\mathrm{Var}(\widehat{\mathrm{SMD}}) \approx 2/n$: requiring the half-width of a two-sided 90% confidence interval for the SMD, $z_{0.95}\sqrt{2/n}$ with $z_{0.95} = 1.645$, to be no larger than the balance criterion $\mathrm{SMD}_{\mathrm{crit}}$ itself, and solving for n, gives $n_{\min} = 2\, z^2_{0.95}/\mathrm{SMD}^2_{\mathrm{crit}} \approx 5.4/\mathrm{SMD}^2_{\mathrm{crit}}$ (87 observations at the default $\mathrm{SMD}_{\mathrm{crit}} = 0.25$); below this size, an observed SMD close to the criterion is not estimated precisely enough for the balance check itself to be trustworthy, which is why it is flagged as *floor-forced*, to distinguish it from a stratum that cleanly satisfied the balance criterion. This split/accept/exclude/floor-force refinement continues, subject to a ceiling on the total number of strata formed, until every region of the propensity score has been resolved, thereby supplying a fully automatic, reproducible, and locally adaptive answer to the stratum-count question left open by Hong [12, 13, 14], one that responds separately to each wave's, each history subgroup's, and each treatment level's own sample size and degree of overlap. This general approach — combining propensity-score stratification with weighting, and choosing the stratification data-adaptively — extends a line of work on combining these two propensity-score strategies for a single binary treatment occasion [21–24] to the longitudinal, multivalued setting considered here.

This procedure is implemented in the community-contributed Stata command `pstrata` [24] and is invoked automatically by `mmwsl` whenever the number of strata is not fixed by the user, so that, unless otherwise noted, every stratification performed in the simulation study reported below (Section 2.6) used this balance-driven procedure rather than an arbitrary fixed stratum count.

2.5 Illustrative application

To illustrate the longitudinal multivalued-treatment setting, we extend the three-level congestive heart failure disease-management example of Linden et al. [25] from a single treatment occasion to repeated treatment decisions. The original alternatives were usual care, nurse-based telephone management, and remote telemonitoring (RTM). In this hypothetical extension, $A_t \in \{0, 1, 2\}$ denotes these three management levels at each of four quarterly follow-up visits, with management intensity allowed to change as a patient's clinical status evolves.

At each visit, L_t represents current CHF status, such as symptom burden, recent weight change, and unplanned utilization, measured before the management decision. Worsening status may prompt escalation from usual care to nurse calls or RTM, whereas improvement may permit de-escalation. Thus, a patient stabilized under RTM might follow the history RTM $\rightarrow$ RTM $\rightarrow$ nurse calls $\rightarrow$ usual care, whereas a patient whose condition deteriorates might follow usual care $\rightarrow$ nurse calls $\rightarrow$ RTM $\rightarrow$ RTM. Because prior management may itself affect subsequent clinical status, L_t is both a predictor of current treatment and a consequence of earlier treatment, producing the treatment-confounder feedback represented in Figure 1.

In this setting, the same multinomial treatment model would estimate the probabilities of usual care, nurse calls, and RTM at each occasion conditional on treatment and covariate history. Sequential IPTW would construct weights directly from these fitted probabilities, whereas MMWS would use the same probabilities to stratify patients and construct weights from treatment proportions within the resulting strata. The two approaches therefore address the same longitudinal

causal problem while differing only in how estimated treatment probabilities are translated into weights.

2.6 Simulation strategy and design

Following the design approach of Linden et al. [25], who compared several modeling strategies for multivalued treatments in a point-treatment setting, we conducted a Monte Carlo simulation in which features of the data-generating process (DGP) expected to influence the relative performance of sequential MMWS and sequential multinomial IPTW were varied one at a time from a common baseline. The baseline DGP generated $N = 2{,}000$ subjects followed over $T = 4$ waves. At the first wave, a single continuous confounder was drawn $X_{i1} \sim \mathrm{N}(0, 1)$; at subsequent waves it evolved as $X_{it} = 0.4\, X_{i,t-1} + \delta\, [\mathbb{1}(A_{i,t-1} = 1) - \mathbb{1}(A_{i,t-1} = 2)] + \mathrm{N}(0, 0.9^2)$, where δ is the treatment–confounder feedback coefficient (baseline $\delta = 0$, i.e., no treatment–confounder feedback at baseline). At each wave, treatment was generated from a three-level ($K = 3$; levels 0, A, B) multinomial (softmax) model linear in X_{it},

$$P(A_{it} = j \mid X_{it}) \;=\; \frac{\exp(\eta_{ijt})}{1 + \sum_{j'=1}^{K-1} \exp(\eta_{ij't})}, \qquad \eta_{ijt} = \alpha_j + \gamma_j \cdot c \cdot X_{it}, \tag{10}$$

where c is a confounding-strength scalar (baseline $c = 0.8$), γ_j is a level-specific sign/multiplier applied to c, and α_j is a level-specific intercept controlling relative treatment-level prevalence. The wave-specific outcome followed $Y_{it} = 10 + 0.5\, X_{it} + q\, X_{it}^2 + \sum_{j=1}^{K-1} \tau_j \mathbb{1}(A_{it} = j) + \mathrm{N}(0, 2^2)$, with quadratic-confounding coefficient q (baseline $q = 2.2$) and true occasion-specific treatment effects $\tau_A = 3$, $\tau_B = 6$ (baseline). The simulation implements a restricted, wave-specific case developed in Section 2.1, in which the direct effect on Y_{it} operates only through the current treatment A_{it}, while earlier treatment affects Y_{it} indirectly, through its influence on the evolving confounder X_{it}. Each of 14 scenarios altered exactly one baseline element, summarized in Table 1: overall sample size ($N = 500$; $N = 5{,}000$); confounding strength (weak, $c = 0.4$; strong, $c = 1.2$); outcome functional form (linear-only, $q = 0$); treatment–confounder feedback (moderate, $\delta = 0.3$; strong, $\delta = 0.6$); number of treatment levels ($K = 4$; $K = 5$, with additional levels and true effects added in the same pattern); treatment-level imbalance (one non-reference level's intercept α_j shifted sharply negative, making it rare); propensity-model misspecification (a true quadratic term in X_{it} added to the treatment-assignment log-odds at a mild and a severe dose, while the treatment model actually fit by `mmwsl` remained linear in X_{it} in every scenario, reflecting an investigator who has correctly measured but not correctly functionally specified the confounder); and positivity stress (confounding strength increased further, to $c = 2.2$, so that the estimated propensity for one treatment level approaches zero in the tails of X_{it}'s distribution). Each scenario was replicated $R = 1{,}000$ times.

Within each replicated dataset, five estimators of the occasion-specific contrasts τ_j were compared: (1) a *naive* estimator regressing Y on treatment-level indicators alone, ignoring X; (2) a *linear-adjusted* estimator additionally conditioning on X; (3) a *correctly specified outcome-adjusted* estimator additionally conditioning on X^2 (matches the functional form of the baseline outcome model, and available only because X and its true functional form are known by design); (4) the

MMWS estimator, weighting by the cumulative sequential MMWS weight from `mmwsl` (equation 7), with strata formed automatically by `pstrata`; and (5) the *IPTW* estimator, weighting by the corresponding cumulative sequential stabilized IPTW weight (equation 4) computed by `mmwsl` from the identical wave/subgroup-specific multinomial models used to construct the MMWS weight. All five estimators were fit as a subject-clustered linear regression of Y on treatment-level indicators (with appropriate weights for estimators 4 and 5), matching the working model in equation (2). The naive and adjusted (linear and correctly specified) comparators anchor the weighted estimators against the two conventional, non-propensity-score alternatives — no adjustment, and direct outcome-model adjustment for the observed and, in the correctly specified case, correctly transformed confounder.

2.7 Performance measures

Following established conventions for reporting Monte Carlo simulation studies in the biostatistical and epidemiological literature [26,27], and consistent with the layout adopted by Linden et al. [25], estimator performance for each treatment-level contrast τ_j, within each scenario and estimator, was summarized across the $R = 1{,}000$ replications using: (1) *percentage relative bias*, $100 \times (\bar{\hat{\tau}}_j - \tau_j)/\tau_j$, where $\bar{\hat{\tau}}_j$ is the mean of the estimated contrast across replications; (2) the *empirical standard deviation* of $\hat{\tau}_j$ across replications; (3) the *mean model-based (cluster-robust) standard error* of $\hat{\tau}_j$ across replications; (4) the *SE/SD ratio*, the mean model-based standard error divided by the empirical standard deviation, with values near 1 indicating well-calibrated inference, values below 1 indicating anticonservative (understated) standard errors, and values above 1 indicating conservative (overstated) standard errors; (5) *95% confidence interval coverage*, the proportion of replications in which the nominal 95% Wald interval, $\hat{\tau}_j \pm 1.96 \times \widehat{\mathrm{SE}}(\hat{\tau}_j)$, contained the true value τ_j; and (6) *root mean squared error* (RMSE), $\sqrt{R^{-1}\sum_{r=1}^{R}(\hat{\tau}_{jr} - \tau_j)^2}$.

For the MMWS and IPTW estimators, we additionally monitored the consequences of wave-level exclusions for cumulative weight construction. When a wave-specific weight was undefined because the relevant propensity-score stratum did not contain all treatment levels represented in the analytic sample, cumulative weighting resumed at the next occasion for which a valid weight could be computed; we refer to this resumption as a *restart*. Four diagnostics were recorded for every replication: (a) the number of wave-level observations excluded from weighting; (b) the number of restarts of the cumulative weight, separately for MMWS and IPTW; (c) the number of subjects experiencing at least one restart; and (d) the mean and maximum number of restarts per affected subject. These diagnostics distinguish wave-level positivity failures from their consequences for subsequent cumulative weight construction.

In total, the simulation encompassed 14 scenarios, each replicated 1,000 times, yielding 14,000 simulated panel datasets; all analyses were conducted in Stata using `mmwsl` and `pstrata` as described in Sections 2.3 and 2.4.

3 Results

3.1 Overall estimator performance

Table 2 reports performance of all five estimators in the baseline scenario (N = 2,000, T = 4, K = 3). The naive estimator had relative bias of 36.3% and 8.3% for contrasts A and B, respectively, and the linear-adjusted estimator had corresponding biases of 26.4% and 13.2%. Coverage was 0% for both contrasts with both estimators. The correctly specified outcome-adjusted estimator was essentially unbiased, with RMSE of 0.057 and coverage of 95.8% and 95.2% for contrasts A and B, respectively.

IPTW had relative bias of 0.3% and −0.1% for contrasts A and B, compared with −1.6% and −2.0% for MMWS. MMWS had lower RMSE than IPTW for both contrasts (0.154 vs. 0.349 for contrast A and 0.189 vs. 0.434 for contrast B), corresponding to IPTW-to-MMWS RMSE ratios of 2.27 and 2.30. Coverage was 95.5% and 91.2% for IPTW and 96.4% and 89.9% for MMWS.

3.2 Performance across data-generating conditions

Table 3 reports relative bias and RMSE for IPTW and MMWS across the 14 simulation scenarios. MMWS had lower RMSE than IPTW for every treatment contrast in every scenario, with IPTW-to-MMWS RMSE ratios ranging from 1.10 to 14.73.

For sample size, the RMSE ratios ranged from 1.57 to 1.64 at N = 500 and from 2.18 to 2.74 at N = 5,000. MMWS relative bias ranged from −3.6% to −2.9% at N = 500 and from −1.2% to −0.4% at N = 5,000; IPTW relative bias remained within 0.8% of the true effect at both sample sizes. Under weak confounding, RMSE ratios were 1.39 and 1.52; under strong confounding, they were 1.90 and 1.76. MMWS relative bias under strong confounding was −4.2% and −3.7% for contrasts A and B.

With moderate treatment-confounder feedback, IPTW relative bias was 1.6% and −0.5% and MMWS relative bias was −1.5% and −2.6% for contrasts A and B, respectively. Under strong feedback, corresponding biases were 5.1% and −0.9% for IPTW and 1.8% and −3.1% for MMWS. RMSE ratios ranged from 1.23 to 2.23 across the two feedback scenarios.

With four treatment levels, RMSE ratios ranged from 1.45 to 2.31 across the three contrasts; with five levels, they ranged from 1.40 to 1.82 across the four contrasts. In the treatment-imbalance scenario, the RMSE ratios were 1.37 for contrast A and 1.96 for the rare-level contrast B. For contrast B, relative bias was −2.0% for IPTW and 3.3% for MMWS.

The largest IPTW-to-MMWS RMSE ratios occurred under propensity-model misspecification. Under severe misspecification, relative bias for contrast A was 62.1% with IPTW and 5.3% with MMWS, with RMSEs of 3.230 and 0.219, respectively (RMSE ratio 14.73). For contrast B, relative bias was −10.2% with IPTW and −4.3% with MMWS, and the RMSE ratio was 3.76. Under mild misspecification, relative bias was 6.6% versus −0.2% for contrast A and −3.2% versus −2.4% for contrast B, for IPTW and MMWS respectively; corresponding RMSE ratios were 4.89 and 1.74.

Under positivity stress, MMWS had greater relative bias than IPTW for both contrasts: −12.3% versus −0.2% for contrast A and −8.4% versus −2.8% for contrast B. MMWS nevertheless had lower RMSE for both contrasts, with IPTW-to-MMWS RMSE ratios of 1.28 and 1.52.

3.3 Standard-error calibration and coverage

Table 4 reports the SE/SD ratio and empirical coverage of nominal 95% confidence intervals. For IPTW, the SE/SD ratio was below 1 for every scenario and treatment contrast, ranging from 0.323 to 0.992. Coverage was between 90% and 96% for most scenarios but was lower for the rare-level contrast under treatment imbalance (75.7%), both contrasts under severe propensity-model misspecification (35.8% and 36.0%), and contrast B under mild misspecification (78.8%).

For MMWS, SE/SD ratios ranged from 0.664 to 1.269 and were at or above 0.90 except for the two contrasts under positivity stress. Coverage was below 90% for contrast B under strong confounding (79.5%), moderate feedback (84.0%), strong feedback (81.7%), severe propensity-model misspecification (55.1%), positivity stress (61.6%), and mild propensity-model misspecification (84.1%); coverage for contrast A was 85.2% under severe misspecification, 74.4% under positivity stress, and 89.7% under treatment imbalance.

Under positivity stress, SE/SD ratios were 0.617 and 0.516 for IPTW and 0.664 and 0.732 for MMWS for contrasts A and B, respectively. Coverage was 90.5% and 88.2% for IPTW and 74.4% and 61.6% for MMWS.

3.4 Weight behavior, positivity, and restarts

Table 5 reports wave-level exclusions from weighting. The mean exclusion rate was 3.0% in the baseline scenario. Exclusion rates were 22.1% at $N = 500$ and 0.4% at $N = 5{,}000$, and 0.9% and 7.0% under weak and strong confounding, respectively. With moderate and strong treatment-confounder feedback, exclusion rates were 4.0% and 5.8%.

The exclusion rate increased to 18.5% with four treatment levels and 32.7% with five treatment levels. It was 13.2% when one treatment level was rare and 22.4% under positivity stress. Rates under severe and mild propensity-model misspecification were 3.1% and 3.0%, respectively.

Under positivity stress, the mean maximum stabilized IPTW weight was 570.1 and the mean Kish effective sample size was 298.2 from a nominal sample of $N = 2{,}000$.

Table 6 reports subject-level restart diagnostics for the six scenarios instrumented to record them: baseline, $N = 500$, $K = 4$, $K = 5$, treatment imbalance, and positivity stress. For MMWS, the mean number of cumulative-weight restarts per replication was 17.0 at baseline, 99.5 at $N = 500$, 146.6 at $K = 4$, 638.8 at $K = 5$, 427.8 under treatment imbalance, and 941.0 under positivity stress. The proportion of subjects experiencing at least one restart followed the same pattern: 0.9% at baseline, 19.5% at $N = 500$, 7.3% at $K = 4$, 31.8% at $K = 5$, 19.7% under treatment imbalance, and 40.5% under positivity stress.

For IPTW, restarts were rare: the cumulative weight restarted in only one of the six scenarios (treatment imbalance, a mean of 0.07 restarts per replication, affecting a mean of 0.07 subjects, well under 0.1% of the sample) and did not restart at all at baseline, $N = 500$, $K = 4$, $K = 5$, or under positivity stress.

Among affected MMWS subjects, the mean number of restarts per subject ranged from 1.00 at baseline and $K = 4$ to 1.15 under positivity stress. The mean replication-level maximum ranged from 1.00 at baseline to 2.14 under positivity stress and was 1.90 under treatment imbalance. Where IPTW restarted at all (treatment imbalance), the mean and maximum were both 1.00.

4 Discussion

This study provides direct finite-sample evidence comparing sequential MMWS with conventional IPTW for time-varying, multivalued treatments. Across most conditions examined, MMWS exhibited somewhat greater bias but substantially lower variability and RMSE. Its advantage was particularly pronounced under treatment-model misspecification. Positivity stress represented the clearest boundary condition: MMWS retained lower RMSE but exhibited substantially greater bias and poorer coverage than IPTW. Increasing the number of treatment levels also produced substantial losses of support, reflected in wave-level exclusions and MMWS restarts.

These findings are consistent with the different ways in which the two estimators convert propensity scores into weights. IPTW directly inverts each subject's estimated probability of the treatment received, so errors or extreme estimated probabilities enter the weight directly and may compound across occasions. MMWS instead constructs weights from treatment proportions within propensity-score strata, thereby smoothing the influence of individual estimated propensities. This mechanism, originally proposed by Hong [12,14], provides a plausible explanation for MMWS's lower variability and robustness to treatment-model misspecification. At the same time, smoothing cannot compensate for inadequate treatment-level support and may introduce bias as overlap deteriorates. This limitation becomes increasingly important with multivalued longitudinal treatments as the number of possible treatment histories grows as K^T. The increasing exclusions and restarts observed with additional treatment levels and under positivity stress illustrate this underlying support problem. IPTW restarts were essentially absent because direct inverse-probability weighting does not additionally require that each propensity-score stratum contain every treatment level.

The inferential results also differed between the estimators. IPTW model-based standard errors underestimated empirical sampling variability throughout the simulations, although confidence-interval coverage remained reasonably close to nominal in many conditions. A plausible explanation is that occasional highly influential estimated weights disproportionately increase the across-replication empirical standard deviation while having less influence on the typical replication-specific sandwich standard error; coverage need not deteriorate proportionately when such replications are uncommon. MMWS standard errors were generally better calibrated, although coverage deteriorated when estimator bias became substantial, particularly under positivity stress and severe treatment-model misspecification.

The comparison evaluated here concerns two approaches to constructing weights within the marginal structural model framework rather than competing causal-identification strategies. Sequential MMWS and IPTW rely on the same assumptions of sequential exchangeability, positivity, and consistency and, in this study, were constructed from identical fitted treatment models; they differ in how the resulting treatment probabilities are converted into weights. The present findings extend Hong's work on MMWS [12–14] by providing direct finite-sample evidence in a longitudinal multivalued setting and complement other approaches to improving longitudinal weights, including covariate-balancing propensity scores [16] and calibrated treatment and censoring weights [10]. Other methods for treatment-confounder feedback, most notably g-estimation of structural nested models [6], provide alternatives to marginal structural models but use fundamentally different modeling and estimation strategies. They were therefore outside the scope of a comparison designed

specifically to isolate differences in weight construction within a common marginal structural model framework.

A strength of this study is the controlled head-to-head design: MMWS and IPTW were constructed from identical fitted treatment models in every replication, isolating differences attributable to weight construction rather than treatment-model estimation. The simulations also examined several features particularly relevant to multivalued longitudinal treatment, including the number and prevalence of treatment levels, treatment-confounder feedback, treatment-model misspecification, and positivity. The findings should nevertheless be interpreted in light of the simulation design. We considered four treatment occasions and a relatively parsimonious data-generating process, and the scenarios varied individual features from a common baseline rather than attempting to represent every combination encountered in practice. The wave-specific estimand studied here also represents one of several possible ways of summarizing longitudinal treatment effects. These design choices allowed the comparison to isolate the behavior of the two weighting strategies; evaluating more complex treatment and covariate structures and alternative longitudinal estimands provides opportunities for future research.

5 Conclusion

Sequential MMWS provided lower RMSE than conventional IPTW across all simulated conditions and was particularly robust to treatment-model misspecification. These advantages diminished as treatment-level support deteriorated, emphasizing that MMWS does not eliminate the positivity challenges associated with complex multivalued treatment histories. When overlap is adequate, MMWS provides a useful alternative approach to constructing weights for marginal structural models, with exclusions and cumulative-weight restarts serving as important diagnostics of treatment-level support.

References

[1] Beth Ann Griffin, Rajeev Ramchand, Daniel Almirall, Mary E. Slaughter, Lane F. Burgette, and Daniel F. McCaffrey. Estimating the causal effects of cumulative treatment episodes for adolescents using marginal structural models and inverse probability of treatment weighting. *Drug and Alcohol Dependence*, 136:69–78, 2014. doi: 10.1016/j.drugalcdep.2013.12.017.

[2] Carlo Lancia, Cristian Spitoni, Jakob Anninga, Jeremy Whelan, Matthew R. Sydes, Gordana Jovic, and Marta Fiocco. Marginal structural models with dose-delay joint-exposure for assessing variations to chemotherapy intensity. *Statistical Methods in Medical Research*, 28(9):2787–2801, 2019. doi: 10.1177/0962280218780619.

[3] James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

[4] James M. Robins, Miguel A. Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[5] Miguel A. Hernán, Babette Brumback, and James M. Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454):440–448, 2001.

[6] James M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In M. Elizabeth Halloran and Donald Berry, editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 95–134. Springer, New York, 1999.

[7] Guanglei Hong and Stephen W. Raudenbush. Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, 33(3):333–362, 2008. doi: 10.3102/1076998607307355.

[8] Ariel Linden and John L. Adams. Evaluating health management programmes over time: Application of propensity score-based weighting to longitudinal data. *Journal of Evaluation in Clinical Practice*, 16:180–185, 2010.

[9] David Suarez, Josep Maria Haro, Diego Novick, and Susana Ochoa. Marginal structural models might overcome confounding when analyzing multiple treatment effects in observational studies. *Journal of Clinical Epidemiology*, 61(6):525–530, 2008. doi: 10.1016/j.jclinepi.2007.11.007.

[10] Sean Yiu and Li Su. Joint calibrated estimation of inverse probability of treatment and censoring weights for marginal structural models. *Biometrics*, 78(1):115–127, 2022. doi: 10.1111/biom.13411.

[11] Liangyuan Hu, Jiayi Ji, Himanshu Joshi, Erick R. Scott, and Fan Li. Estimating the causal effects of multiple intermittent treatments with application to COVID-19. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 72(5):1162–1186, 2023. doi: 10.1093/jrsssc/qlad076.

[12] Guanglei Hong. Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, 35(5):499–531, 2010. doi: 10.3102/1076998609359785.

[13] Guanglei Hong. Marginal mean weighting through stratification: A generalized method for evaluating multivalued and multiple treatments with nonexperimental data. *Psychological Methods*, 17(1):44–60, 2012. doi: 10.1037/a0024918.

[14] Guanglei Hong. *Causality in a Social World: Moderation, Mediation and Spill-over*. John Wiley & Sons, 2015.

[15] Ariel Linden. MMWSL: Stata module for implementing longitudinal marginal mean weighting through stratification for time-varying binary or nominal treatments, 2026. Statistical Software Components s459891, Boston College Department of Economics.

[16] Kosuke Imai and Marc Ratkovic. Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association*, 110(511):1013–1023, 2015. doi: 10.1080/01621459.2014.956872.

[17] I-Chan Huang, Constantine Frangakis, Francesca Dominici, Gregory B. Diette, and Albert W. Wu. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research*, 40(1):253–278, 2005. doi: 10.1111/j.1475-6773.2005.00350.x.

[18] William G. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24(2):295–313, 1968.

[19] Paul R. Rosenbaum and Donald B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[20] Donald B. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3–4):169–188, 2001. doi: 10.1023/A:1020363010465.

[21] Ariel Linden and John L. Adams. Improving participant selection in disease management programs: Insights gained from propensity score stratification. *Journal of Evaluation in Clinical Practice*, 14(6):914–918, 2008. doi: 10.1111/j.1365-2753.2008.01021.x.

[22] Ariel Linden. Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *Journal of Evaluation in Clinical Practice*, 20(6):1065–1071, 2014. doi: 10.1111/jep.12226.

[23] Ariel Linden. Improving causal inference with a doubly robust estimator that combines propensity score stratification and weighting. *Journal of Evaluation in Clinical Practice*, 2017. doi: 10.1111/jep.12714.

[24] Ariel Linden. A comparison of approaches for stratifying on the propensity score to reduce bias. *Journal of Evaluation in Clinical Practice*, 2017. doi: 10.1111/jep.12701.

[25] Ariel Linden, S. Derya Uysal, Andrew Ryan, and John L. Adams. Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine*, 35(4):534–552, 2016. doi: 10.1002/sim.6768.

[26] Andrea Burton, Douglas G. Altman, Patrick Royston, and Roger L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006. doi: 10.1002/sim.2673.

[27] Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. doi: 10.1002/sim.8086.

Table 1. Simulation design and inputs.

Parameter	Value / Description
Baseline regression/outcome model	
Outcome	$Y_{it} = 10 + 0.5X_{it} + qX_{it}^2 + \sum_{j=1}^{K-1} \tau_j \mathbb{1}(A_{it} = j) + \varepsilon_{it}$, $\varepsilon_{it} \sim \mathrm{N}(0, 2^2)$
Confounder process	$X_{i1} \sim \mathrm{N}(0,1)$; $X_{it} = 0.4X_{i,t-1} + \delta[\mathbb{1}(A_{i,t-1}{=}1) - \mathbb{1}(A_{i,t-1}{=}2)] + \mathrm{N}(0, 0.9^2)$, $t > 1$
Treatment model	multinomial softmax, $\eta_{ijt} = \alpha_j + \gamma_j c X_{it}$ (equation 10); linear in X_{it} at baseline
True effects (τ)	$\tau_A = 3$, $\tau_B = 6$ at baseline ($K = 3$); additional levels added in the same increasing pattern for $K = 4, 5$
Panel design	
Sample size (N)	2,000 (baseline); 500 and 5,000 examined (scenarios 02–03)
Waves (T)	4 (fixed across all scenarios)
Treatment levels (K)	3 (baseline); 4 and 5 examined (scenarios 09–10)
Scenarios (one factor varied from baseline per scenario)	
01 baseline	every factor at its baseline value
02 $N = 500$	smaller sample size
03 $N = 5{,}000$	larger sample size
04 confounding, weak	$c = 0.4$
05 confounding, strong	$c = 1.2$
06 linear outcome only	$q = 0$ (no quadratic confounding of the outcome)
07 feedback, moderate	$\delta = 0.3$
08 feedback, strong	$\delta = 0.6$
09 $K = 4$ levels	one additional non-reference treatment level added
10 $K = 5$ levels	two additional non-reference treatment levels added
11 imbalanced levels	one non-reference level's softmax intercept shifted sharply negative (rare level)
12 PS misspecification, severe	true softmax log-odds include a quadratic term in X_{it} (coefficient 0.2) omitted from the fitted (linear) treatment model
13 positivity stress	$c = 2.2$; propensity for one level approaches 0 in the tails of X_{it}
14 PS misspecification, mild	as scenario 12, quadratic coefficient 0.05 (dose-response companion to scenario 12)
Estimators compared	
Naive	Y regressed on treatment-level indicators alone
Linear-adjusted	naive model plus X

Continued on next page

Table 1 continued

Parameter	**Value / Description**
Correctly specified	naive model plus X and X^2
MMWS	weighted by cumulative sequential MMWS weight (`mmwsl`; equation 7), strata chosen automatically by `pstrata`
IPTW	weighted by cumulative sequential stabilized IPTW weight (`mmwsl`, `iptw` option; equation 4), same underlying wave/subgroup propensity models as MMWS
Stratification (`pstrata`), when not fixed by the user	
Starting number of strata	auto-derived, equation 8 (floor of 5)
Balance criterion	maximum pairwise SMD $\leq$ 0.25 across treatment levels present in a stratum
Per-group reliability floor	equation 9 (87 observations at the default SMD criterion)
Stratum-level exclusion	any treatment level entirely absent from a stratum (positivity failure)
Monte Carlo settings	
Replications per scenario	1,000
Variance estimation	subject-clustered (sandwich) robust variance for all five estimators
Significance level	$\alpha = 0.05$ (two-sided Wald test), for coverage evaluation

Note: DGP = data-generating process; IPTW = inverse probability of treatment weighting; MMWS(L) = (longitudinal) marginal mean weighting through stratification; PS = propensity score; SMD = standardized mean difference. All 14 scenarios share the baseline DGP except for the single factor listed for that scenario; scenario 01 (baseline) uses every baseline value.

Table 2. Estimator performance at the baseline scenario (scenario 01; $N = 2{,}000$, $T = 4$, $K = 3$, $R = 1{,}000$ replications), by treatment contrast.

Estimator	Mean estimate	Bias (%)	RMSE	SE/SD	Coverage (%)
Contrast A (true effect = 3)					
Naive	4.0875	36.25	1.093	0.978	0.0
Linear-adjusted	3.7924	26.41	0.798	1.013	0.0
Correctly specified	3.0000	-0.00	0.057	1.007	95.8
MMWS	2.9519	-1.60	0.154	1.008	96.4
IPTW	3.0094	0.31	0.349	0.676	95.5
Contrast B (true effect = 6)					
Naive	6.4960	8.27	0.504	1.038	0.0
Linear-adjusted	6.7907	13.18	0.796	1.021	0.0
Correctly specified	6.0016	0.03	0.057	0.997	95.2
MMWS	5.8804	-1.99	0.189	1.029	89.9
IPTW	5.9941	-0.10	0.434	0.596	91.2

Note: RMSE = root mean squared error; SE/SD = mean model-based standard error divided by the empirical standard deviation of the point estimate across replications; Coverage = empirical coverage of the nominal 95% Wald confidence interval. Contrast A: treatment level A vs. reference (true effect = 3). Contrast B: treatment level B vs. reference (true effect = 6).

Table 3. Relative bias and root mean squared error (RMSE) of the IPTW and MMWS estimators, all 14 simulation scenarios.

			IPTW		MMWS		
Scenario	Contrast	True	Bias (%)	RMSE	Bias (%)	RMSE	Ratio[a]
01 baseline	A	3	0.31	0.349	-1.60	0.154	2.27
	B	6	-0.10	0.434	-1.99	0.189	2.30
02 $N = 500$	A	3	0.78	0.487	-3.63	0.310	1.57
	B	6	-0.46	0.541	-2.85	0.329	1.64
03 $N = 5{,}000$	A	3	0.14	0.242	-0.39	0.088	2.74
	B	6	-0.10	0.248	-1.17	0.114	2.18
04 confounding, weak	A	3	-0.42	0.132	0.13	0.095	1.39
	B	6	-0.10	0.141	-0.40	0.093	1.52
05 confounding, strong	A	3	0.13	0.502	-4.23	0.264	1.90
	B	6	-1.00	0.566	-3.66	0.322	1.76
06 linear outcome only	A	3	0.24	0.110	1.17	0.100	1.10
	B	6	-0.06	0.146	-0.59	0.100	1.46
07 feedback, moderate	A	3	1.55	0.317	-1.50	0.142	2.23
	B	6	-0.52	0.348	-2.58	0.210	1.66
08 feedback, strong	A	3	5.14	0.301	1.75	0.170	1.78
	B	6	-0.93	0.293	-3.05	0.239	1.23
09 $K = 4$ levels	A	3	0.50	0.378	-1.83	0.164	2.31
	B	6	-0.08	0.417	-1.92	0.193	2.16
	C	9	-0.12	0.241	-0.51	0.166	1.45
10 $K = 5$ levels	A	3	1.71	0.323	-1.32	0.177	1.82
	B	6	-0.16	0.354	-1.83	0.203	1.74
	C	9	0.04	0.277	-0.56	0.173	1.61
	D	12	-0.30	0.271	-0.79	0.194	1.40
11 imbalanced levels	A	3	0.35	0.166	-1.79	0.121	1.37
	B	6	-1.96	0.592	3.28	0.302	1.96
12 PS misspec., severe	A	3	62.05	3.230	5.27	0.219	14.73
	B	6	-10.20	1.080	-4.26	0.287	3.76
13 positivity stress	A	3	-0.18	0.806	-12.33	0.629	1.28
	B	6	-2.81	1.038	-8.36	0.682	1.52
14 PS misspec., mild	A	3	6.64	0.683	-0.17	0.140	4.89
	B	6	-3.19	0.352	-2.39	0.202	1.74

Note: RMSE = root mean squared error. [a]RMSE ratio = IPTW RMSE / MMWS RMSE; values above 1 indicate MMWS achieved the smaller RMSE. Scenario labels correspond to Table 1.

Table 4. Standard-error calibration (SE/SD ratio) and empirical coverage of nominal 95% confidence intervals for the IPTW and MMWS estimators, all 14 simulation scenarios.

		IPTW		MMWS	
Scenario	Contrast	SE/SD[a]	Coverage (%)	SE/SD[a]	Coverage (%)
01 baseline	A	0.676	95.5	1.008	96.4
	B	0.596	91.2	1.029	89.9
02 $N = 500$	A	0.748	94.5	0.999	94.3
	B	0.720	93.5	1.054	93.3
03 $N = 5{,}000$	A	0.695	94.8	1.119	98.1
	B	0.742	92.1	1.158	92.8
04 confounding, weak	A	0.976	94.2	1.172	98.8
	B	0.992	95.5	1.269	98.2
05 confounding, strong	A	0.697	94.8	0.897	92.4
	B	0.669	91.2	0.900	79.5
06 linear outcome only	A	0.980	95.3	1.011	93.9
	B	0.743	94.1	0.998	92.5
07 feedback, moderate	A	0.713	96.7	1.107	96.5
	B	0.707	92.3	1.071	84.0
08 feedback, strong	A	0.911	92.0	1.020	94.5
	B	0.883	93.3	1.090	81.7
09 $K = 4$ levels	A	0.634	94.6	1.076	96.3
	B	0.669	91.5	1.110	91.4
	C	0.890	95.5	1.055	96.4
10 $K = 5$ levels	A	0.793	95.6	1.092	96.6
	B	0.772	92.8	1.136	94.0
	C	0.850	95.7	1.153	97.4
	D	0.893	94.0	1.140	94.7
11 imbalanced levels	A	0.902	94.7	0.923	89.7
	B	0.666	75.7	1.087	92.2
12 PS misspec., severe	A	0.323	35.8	1.070	85.2
	B	0.366	36.0	1.104	55.1
13 positivity stress	A	0.617	90.5	0.664	74.4
	B	0.516	88.2	0.732	61.6
14 PS misspec., mild	A	0.451	94.2	1.069	97.0

	B	0.781	78.8	1.051	84.1

Note: SE/SD = mean model-based (sandwich) standard error divided by the empirical standard deviation of the point estimate across replications; a value below 1 indicates the model-based SE understates sampling variability (anticonservative), a value above 1 indicates it overstates it (conservative). Coverage = empirical coverage of the nominal 95% Wald confidence interval. [a]Ratio of mean SE to empirical SD, not its reciprocal. Scenario labels correspond to Table 1.

Table 5. Wave-level exclusions from weighting and, for the positivity-stress scenario, extreme-weight diagnostics, all 14 simulation scenarios.

Scenario	N	Records[a]	Excl.[b]	Excl. (%)	Max wt.[c]	Kish ESS[c]
01 baseline	2,000	8,000	236.6	3.0	—	—
02 $N = 500$	500	2,000	443.0	22.1	—	—
03 $N = 5{,}000$	5,000	20,000	79.9	0.4	—	—
04 confounding, weak	2,000	8,000	68.9	0.9	—	—
05 confounding, strong	2,000	8,000	563.7	7.0	—	—
06 linear outcome only	2,000	8,000	235.4	2.9	—	—
07 feedback, moderate	2,000	8,000	318.9	4.0	—	—
08 feedback, strong	2,000	8,000	463.0	5.8	—	—
09 $K = 4$ levels	2,000	8,000	1480.4	18.5	—	—
10 $K = 5$ levels	2,000	8,000	2615.5	32.7	—	—
11 imbalanced levels	2,000	8,000	1056.7	13.2	—	—
12 PS misspec., severe	2,000	8,000	245.3	3.1	—	—
13 positivity stress	2,000	8,000	1793.4	22.4	570.1	298.2
14 PS misspec., mild	2,000	8,000	237.0	3.0	—	—

Note: $T = 4$ in every scenario. [a]Records $= N \times T$, the total number of wave-level observations analyzed per replication. [b]Mean, across 1,000 replications, of the number of wave-level observations excluded from weighting because the relevant propensity-score stratum did not contain all treatment levels represented in the analytic sample (Section 2.4); this is a property of the stratification step specifically, computed from propensity models shared with IPTW, but its consequence is not: an excluded observation's MMWS weight is undefined, while its IPTW weight, computed directly from the same fitted propensities, is generally still defined (Section 3.4). [c]Mean, across 1,000 replications, of the largest stabilized IPTW weight and the Kish effective sample size, out of N subjects; recorded only for the positivity-stress scenario (13). Scenario labels correspond to Table 1.

Table 6. Subject-level cumulative-weight restart diagnostics, the six instrumented scenarios.

		MMWS			IPTW		
Scenario	N	Restarts[a]	Subj. ≥ 1[b] (%)	Mean/Max[c]	Restarts[a]	Subj. ≥ 1[b] (%)	Mean/Max[c]
01 baseline	2,000	17.0	17.0 (0.9)	1.00/1.00	0.00	0.00 (0.0)	—/—
02 $N = 500$	500	99.5	97.7 (19.5)	1.02/1.26	0.00	0.00 (0.0)	—/—
09 $K = 4$ levels	2,000	146.6	146.4 (7.3)	1.00/1.02	0.00	0.00 (0.0)	—/—
10 $K = 5$ levels	2,000	638.8	636.0 (31.8)	1.00/1.16	0.00	0.00 (0.0)	—/—
11 imbalanced	2,000	427.8	394.4 (19.7)	1.08/1.90	0.07	0.07 (0.0)	1.00/1.00
13 positivity stress	2,000	941.0	809.7 (40.5)	1.15/2.14	0.00	0.00 (0.0)	—/—

Note: A restart occurs when cumulative weight construction resumes at the next occasion for which a valid wave-specific weight can be computed, following an occasion at which it was undefined (Section 2.7). Only these six of the 14 scenarios were instrumented to capture subject-level restart diagnostics. [a]Mean, across 1,000 replications, of the number of wave-level restarts of the cumulative weight. [b]Mean, across 1,000 replications, of the number of subjects experiencing at least one restart; the percentage in parentheses is this count divided by N. [c]Mean, across replications, of the per-replication mean and maximum number of restarts among subjects experiencing at least one restart; recorded as — when no subject in any replication restarted. Scenario labels correspond to Table 1.

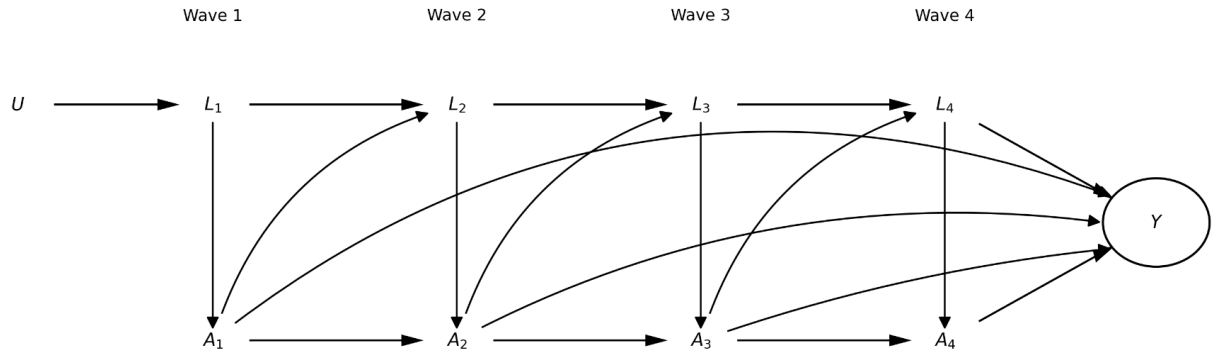

Figure 1: Directed acyclic graph for a four-wave model with three treatment levels, illustrating time-varying treatment, time-varying confounding, and an outcome measured after the final wave. L_t: time-varying covariates, measured before A_t; $A_t \in \{0, 1, 2\}$: treatment at occasion t; Y: outcome, measured after wave 4; U: baseline covariates, measured before wave 1. Each L_t is affected by the prior treatment A_{t-1} and, in turn, confounds the concurrent treatment A_t; each A_t may also affect the outcome Y directly, independent of later treatment. Not every arrow from U is drawn (it is a common cause of every subsequent L_t, A_t, and Y); the graph is acyclic, consistent with the causal structure assumed by both the marginal structural model of Section 2.2 and the MMWS estimator of Section 2.3.