Awa Diop*, Caroline Sirois, Jason R. Guertin, Mireille E. Schnitzer, James M. Brophy, Claudia Blais and Denis Talbot

History-restricted marginal structural model and latent class growth analysis of treatment trajectories for a time-dependent outcome

https://doi.org/10.1515/ijb-2023-0116
Received October 11, 2023; accepted July 11, 2024; published online August 12, 2024

Abstract: In previous work, we introduced a framework that combines latent class growth analysis (LCGA) with marginal structural models (LCGA-MSM). LCGA-MSM first summarizes the numerous time-varying treatment patterns into a few trajectory groups and then allows for a population-level causal interpretation of the group differences. However, the LCGA-MSM framework is not suitable when the outcome is time-dependent. In this study, we propose combining a nonparametric history-restricted marginal structural model (HRMSM) with LCGA. HRMSMs can be seen as an application of standard MSMs on multiple time intervals. To the best of our knowledge, we also present the first application of HRMSMs with a time-to-event outcome. It was previously noted that HRMSMs could pose interpretation problems in survival analysis when either targeting a hazard ratio or a survival curve. We propose a causal parameter that bypasses these interpretation challenges. We consider three different estimators of the parameters: inverse probability of treatment weighting (IPTW), g-computation, and a pooled longitudinal targeted maximum likelihood estimator (pooled LTMLE). We conduct simulation studies to measure the performance of the proposed LCGA-HRMSM. For all scenarios, we obtain unbiased estimates when using either g-computation or pooled LTMLE. IPTW produced estimates with slightly larger bias in some scenarios. Overall, all approaches have good coverage of the 95 % confidence interval. We applied our approach to a population of older Quebecers composed of 57,211 statin initiators and found that a greater adherence to statins was associated with a lower combined risk of cardiovascular disease or all-cause mortality.

Keywords: pooled LTMLE; g-computation; IPTW; history-restricted MSMs; survival analysis; cardiovascular disease

***Corresponding author: Awa Diop**, Département de médecine sociale et préventive, Université Laval, Centre de recherche du CHU de Québec – Université Laval, Axe santé des populations et pratiques optimales en santé, Québec, QC, Canada, E-mail: awa.diop.2@ulaval.ca. https://orcid.org/0000-0002-8646-5305
Caroline Sirois, Faculté de pharmacie, Université Laval, Centre de recherche du CHU de Québec – Université Laval, Axe santé des populations et pratiques optimales en santé, Québec, QC, Canada, E-mail: caroline.sirois@pha.ulaval.ca
Jason R. Guertin, Tissue Engineering Laboratory (LOEX), Département de médecine sociale et préventive, Université Laval, Centre de recherche du CHU de Québec – Université Laval, Axe santé des populations et pratiques optimales en santé, Québec, QC, Canada, E-mail: jason.guertin@fmed.ulaval.ca
Mireille E. Schnitzer, Faculté de pharmacie et Département de médecine sociale et préventive, ESPUM, Department of Epidemiology, Biostatistics, and Occupational Health, Université de Montréal, McGill University, Montréal, QC, Canada, E-mail: mireille.schnitzer@umontreal.ca. https://orcid.org/0000-0001-8049-9646
James M. Brophy, Hospital Center for Health Outcomes Research, McGill University, Montréal, QC, Canada, E-mail: james.brophy@mcgill.ca
Claudia Blais, Institut national de santé publique du Québec (INSPQ), Québec, QC, Canada, E-mail: claudia.blais@inspq.qc.ca
Denis Talbot, Département de médecine sociale et préventive, Université Laval, Centre de recherche du CHU de Québec – Université Laval, Axe santé des populations et pratiques optimales en santé, Québec, QC, Canada, E-mail: denis.talbot@fmed.ulaval.ca

Open Access. © 2024 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

1 Introduction

Prevention of cardiovascular diseases (CVDs) is particularly important given their high prevalence and impact on population health [1]. However, to implement primary prevention strategies in the population, a rigorous measurement of expected benefits is necessary [2]. Statins, medications that reduce cholesterol levels, are used in the primary prevention of CVD events, but their benefit for this purpose among older individuals is uncertain [3, 4]. In this study, our goal is to measure the potential of actual patterns of time-varying statin usage to prevent a first CVD or death event using medical administrative databases. A widespread approach to measuring the impact of statins on CVD events is by estimating the effect of the cumulative number of periods that subjects were exposed to the treatment [5]. Alternatively, some researchers are more interested in knowing the effect of treatment adherence versus non adherence, i.e., compliance versus non compliance of the patient to the physician's recommendations [6]. Adherence is often described using approaches such as the medication possession ratio or the proportion of days covered [7, 8]. However, these methods do not capture the complex dynamics of a time-varying treatment. Latent class growth analysis (LCGA) has been proposed as a better method to measure actual patterns of adherence [9].

In our previous work, we introduced a theoretical framework that combines LCGA and a nonparametric marginal structural model (LCGA-MSM) to measure the impact of groups of treatment trajectories (or trajectory groups) on an outcome measured at the end of the follow-up period [10]. The LCGA-MSM approach tackles some important challenges encountered when analysing longitudinal data. Indeed, the LCGA summarizes the numerous observed time-varying treatment patterns into a few trajectory groups [9–11]. The LCGA-MSM also deals adequately with time-varying variables that can have a double role as confounders and mediators when estimating the effect of these trajectory groups on the outcome. Such treatment-confounder feedback is expected in our application. For example, it is well known that diabetes is an important CVD risk factor [12, 13]. Moreover, in a meta-analysis, Sattar et al. [14] concluded that statin therapy is associated with a slightly increased risk of developing diabetes. Therefore, diabetes has a potential double role of confounder and mediator in the pathway between statins and CVDs. LCGA-MSM also gives a direct population-level causal interpretation of group effects on the outcome, an advantage of MSMs [15]. Thus, a combination of LCGA and MSM might contribute to better understanding the effects of treatment adherence and therefore help make better clinical or public health decisions. For example, LCGA-MSM may help determine whether benefits are expected among patients with imperfect adherence, and whether it is worth developing policies to improve adherence among such patients.

While the LCGA-MSM framework has multiple advantages, it has some limitations. One noteworthy limitation of the LCGA-MSM is the increasing time gap between the period during which treatment trajectories are evaluated and measurements of the outcome over the following period. Moreover, it may not currently be suitable for some complex applications in which the outcome is time-dependent. Indeed, our previous work focused on a treatment measured during a single window and an outcome measured only at the end of that window [10]. This may result in a loss of information since, in a real-life setting, events can occur during the course of the treatment trajectory. In this paper, we aim to generalize the LCGA-MSM framework that respects the time-dependent nature of the treatment and outcome. To do this, we propose combining a nonparametric history-restricted marginal structural model (HRMSM) with LCGA. HRMSMs are a generalization of standard MSMs that consist of a repeated application of standard MSMs on different time-windows [16–18].

Our work extends the current literature in several directions. First, we extend LCGA-MSMs to make possible the use of multiple measurements of the outcome through the HRMSM framework. To the best of our knowledge, we also present the first application of HRMSMs with a time-to-event outcome. It was previously noted that HRMSMs could pose important interpretation problems when targeting hazard ratios in a survival analysis context [16]. The causal parameter we propose circumvents this caveat. Finally, we introduce a pooled longitudinal targeted maximum likelihood estimation (LTMLE) estimator of the parameters of an HRMSM and an associated variance estimator based on the efficient influence curve that accounts for the correlations arising from the repeated use of the data. We consider three different estimators of the parameters of our proposed

LCGA-HRMSM: inverse probability of treatment weighting (IPTW), g-computation, and the LTMLE. In the remainder, we first present the data used in our motivating illustration. We then present a review of LCGA-MSMs, the notation and data structure, the theoretical framework of LCGA-HRMSMs and the different estimators. Simulation studies are used to evaluate and compare the different estimators. We then return to the motivating real-data analysis before ending the paper with some recommendations for the practical application of our proposed approach.

2 Data

We built a retrospective cohort for the period between April 1, 2013 and March 30, 2018 using the Quebec integrated chronic disease surveillance system (QICDSS) available at the *Institut national de santé publique du Québec (INSPQ)*. The QICDSS is updated annually and is composed of five linked databases [19]: (1) health insurance registry, (2) hospitalization database, (3) vital statistics death (4) physician claims and (5) pharmaceutical services. All five databases are merged using the unique identifier of individuals that is their health insurance number [19]. We identified individuals aged greater than 65 years old on April 1, 2013 from Quebec, Canada. Because we are interested in a primary prevention context, only individuals without any CVD history in the last five years were included in the study. The quantification of CVD was based on an algorithm approved by the public health agency of Canada. Individuals with a statin claim in the year prior to enrolment were excluded as we are interested only in statin initiators. To be included, an individual was thus required to be enrolled in the public drug insurance plan for at least one year prior to enrolment and the following years. Thus, we included in the cohort 57,211 individuals who initiated statin treatment between April 2013 and September 2017. It is noteworthy that all Quebecers are enrolled in public drug insurance at the age of 65.

3 Review of LCGA-MSMs

In this section, we briefly present the LCGA-MSM approach (see Diop et al. [10] for more details). LCGA-MSM is a two-step approach. In the first step, an LCGA is used to classify individuals into J distinct trajectory groups $z_1, \dots, z_J$ based on their treatment trajectory [11]. LCGA is a mixture modeling approach that represents the treatment trajectory within each group as a polynomial or some other (e.g., cubic spline) function of time [11, 20–22]. In a second step, a working MSM is chosen to relate the outcome to these trajectory groups. Following [23], the causal parameter of interest β is defined as a projection of the true nonparametric MSM m^* onto the working model m. From a practical point of view, LCGA first allows clustering the observed treatment trajectories in a few distinct groups. These groups allow a simplified visualization of the most commonly observed patterns of exposures over time. In a second step, the MSM allows estimating the effect of these trajectory groups as a best approximation to the true underlying causal model under the working model. The parameter of interest can be conceptualized as follows: "if we were to conduct a randomized trial with groups consisting of the possible individual treatment trajectories, and then cluster the treatment groups according to the trajectory groups found in the first step, how would the mean outcome differ between clusters?" [10].

Consider the data presented in Section 2 with follow-up times $t = 1, \dots, K$ with $K = 60$ months, where $t = 1$ corresponds to the month of statin initiation (index date). An example of an LCGA-MSM application on these data would be to use data on time points $t = 1, \dots, K'$ ($K' < K$) to classify individuals into trajectory groups and then use the remaining period ($t = K' + 1, \dots, K$) as a follow-up for the outcome. For example, months 1–12 could be used to construct trajectories and months 13–60 for the outcome follow-up. For simplicity, we temporarily assume that events cannot occur during the exposure follow-up time, noting that our previous work [10] and the extension presented in later sections accommodate the occurrence of events during the exposure follow-up. In the following, we use capital letters to represent random variables and corresponding lower case letters to represent observed or fixed values these variables take. Let $\bar{A}_t = (A_1, A_2, \dots, A_t)$ be the treatment trajectory up to time t with $\bar{A} \equiv \bar{A}_{K'}$. We denote by Y_K the observed outcome that can be binary or continuous and by $Y_K^{\bar{a}}$ the

counterfactual outcome under a specific treatment trajectory $\bar{a}$, which is measured during the outcome follow-up time (between $K'+1$ and K). Similarly, $\bar{L}_t = (L_1, L_2, \dots L_t)$ is the covariates' history up to time t with $\bar{L} \equiv \bar{L}_{K'}$. We denote by F_X the unknown distribution of all possible counterfactual variables $X = \left(\bar{L}^{\bar{a}}, Y_K^{\bar{a}} : \bar{a} \in \mathcal{A}\right)$ where $\mathcal{A}$ is the set of all possible values of $\bar{a}$. We denote by P_{F_X} the distribution of the observed data $\mathcal{O} = \left\{Y_K, \bar{A}_{K'}, \bar{L}_{K'}\right\}$, from which $i = 1, \dots, n$ independent and identically distributed observations are randomly drawn. The following nonparametric form of an MSM (the true model) is considered:

$$E_{F_X}\left(Y_K^{\bar{a}}\right) = m^*(\bar{a}). \tag{1}$$

The nonparametric identification of model (1) from the observed data can be achieved under the following causal assumptions: (1) no interference between subjects: the potential outcome of a given individual is not affected by others' exposure; (2) positivity: for each level $(\bar{a}_t, \bar{l}_t)$, $P(A_t = a_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_t = \bar{l}_t) > 0$. In other words, in each stratum defined by previous treatment and covariates, we find both exposed and unexposed individuals at time t. (3) sequential conditional exchangeability: there are no unmeasured confounders conditional on the history of treatment up to time $t-1$ and the history of the covariates up to time t [15], that is $Y_K^{\bar{a}} \perp\!\!\!\perp A_t | \bar{A}_{t-1}, \bar{L}_t$; (4) consistency: given the observed treatment trajectory, the observed outcome and the potential outcome under that given trajectory are the same, formally if $\bar{A} = \bar{a}$ then $Y_K^{\bar{a}} = Y_K$. We note that although sequential conditional exchangeability is often interpreted as a "no unmeasured confounders" assumption, this assumption is also violated under collider stratification.

The parameter of interest is defined as a projection of the true nonparametric MSM m^* onto a parametric working model that is a function of the trajectory groups. Let z_i be the trajectory group assigned to subject i as defined previously and $z^*(\bar{a}) = (z_1^*(\bar{a}), \dots, z_J^*(\bar{a}))$ dummy variables indicating whether the individual trajectory $\bar{a}$ is clustered into trajectory group $1, \dots, J$. For example, when the outcome Y_K is binary, the true model could be projected onto the following logistic LCGA-MSM:

$$m(\bar{a}|\beta) \equiv E\left(Y_K^{\bar{a}}\right) = \text{expit}(\beta_0 + \beta_1 z_1^*(\bar{a}) + \beta_2 z_2^*(\bar{a}) + \cdots + \beta_{J-1} z_{J-1}^*(\bar{a})). \tag{2}$$

To define our parameter of interest β we also need to choose a loss function and a projection weight function $\lambda(\bar{a})$. For example, we consider the negative log-likelihood of a logistic regression. Denote by $\mathcal{M}^*$ the infinite-dimensional space of all possible functions m^* relating the counterfactual expectation with the treatment trajectories and $m \in \mathcal{M}^*$ the working model. Considering our example working model and loss-function, the parameter is then defined as:

$$\beta(F_X | m, \lambda) = \underset{\beta \in \mathbf{R}^k}{\arg\min}\, E_{F_X}\left\{ -\sum_{\bar{a} \in \bar{\mathcal{A}}} \left[\log\left(m(\bar{a}|\beta)^{Y_K^{\bar{a}}} (1 - m(\bar{a}|\beta))^{(1 - Y_K^{\bar{a}})} \right) \right] \lambda(\bar{a}) \right\}.$$

In our previous work, we proposed an IPTW estimator of this parameter. The expression of the IPTW is given by:

$$W(\bar{A}) = \frac{\lambda(\bar{A})}{g(\bar{A}|X)}, \tag{3}$$

where we define, $g(\bar{A}|X) = \prod_{t=1}^{K'} P(A_t = 1 | \bar{A}_{t-1}, \bar{L}_t)$ and $\lambda(\bar{A})$ is a projection weight function. If stabilized weights are of interest, we set $\lambda(\bar{A}) = \prod_{t=1}^{K'} P(\bar{A}_t | \bar{A}_{t-1})$. For unstabilized weights $\lambda(\bar{A}) = 1$. Thus, the IPTW estimating function is given by:

$$D_{h_\lambda}(O|\beta) = \frac{h_\lambda(\bar{A}, L_1)\epsilon_{\bar{a}}(\beta)}{\prod_{t=1}^{K'} P(A_t = 1 | \bar{A}_{t-1}, \bar{L}_t)}, \tag{4}$$

with $h_\lambda(\bar{A}) \equiv \lambda(\bar{A}) \frac{\partial}{\partial \beta} m(\bar{a}|\beta) \times m(\bar{a}|\beta)$, $\lambda(\bar{A}) = \prod_{t=0}^{K'} P(\bar{A}_t | \bar{A}_{t-1})$ and $\epsilon_{\bar{a}}(\beta) = Y_K^{\bar{a}} - m(\bar{a}|\beta)$.

In practice, parameters of the working model can be directly estimated using a weighted logistic regression of Y_K on z with observations weighted according to the weights defined in Equation (3). The logistic regression model can be fitted using a weighted *generalized estimating equations* (GEE) estimator, with, for example, the

function *geeglm* from the **R** package *geepack*. A simple practical solution for inference is to use the sandwich variance estimator of the GEE routine, which treats the weights as known and results in conservative statistical inferences [24]. Alternative solutions for inferences include (a) bootstrapping the estimation of the weights and the GEE regression, but not the LCGA, or (b) stacking the estimating equations of the weights and the GEE regression [see e.g., 25], both solutions providing consistent estimation of the variance. Note that the distribution of the trajectory groups is only a function of the joint distribution of the treatments $\bar{A}$ and does not depend on the parameter of interest β; thus z is an ancillary statistic [10, 26, 27]. Therefore, the data-driven estimation of the trajectory groups can be ignored when estimating the parameters of our LCGA-MSM. However, it should be noted that inferences are conditional on the selected LCGA. As argued by multiple authors, clustering that does not involve using the outcome does not invalidate inferences [28, 29]. Methods that assess the sensitivity of the results to the uncertainty of subjects' classification to trajectory groups have been developed and could be considered within the LCGA-MSM framework [30].

In Figure 1, we present an example in which adherence changes considerably during the follow-up period of the outcome. In an LCGA-MSM analysis, the outcome at any point of the follow-up would be modeled as a function of the trajectory determined using the single treatment window. Opposingly, the LCGA-HRMSM framework we introduce next considers multiple treatment windows and allows the subjects' trajectory groups to change from one window to another. Consequently, the outcome at any time point can be modeled as a function of the treatment trajectory group in the most recent treatment window.

In the example in Figure 2, if we were to apply an LCGA-MSM, we would consider that the outcome is observed at the end of the follow-up at $K = 6$ which might not be the best use of the data. We also see in Figure 2 that events might be observed during the exposure measurement period. That is, the treatment trajectory does not necessarily precede the outcome; instead they are two concomitant phenomena. HRMSMs can thus be useful to make the most of the available information.

4 History restricted MSM and LCGA

HRMSMs can be seen as a repeated application of a standard MSM to different subsets of the original data across time points. More precisely, estimating the parameters of an HRMSM first requires splitting the data into multiple windows of fixed size. For each window, the counterfactual outcome at the end of the window is modeled as a function of the treatments measured during that window. This process is performed simultaneously for all windows, thus resulting in repeated measurements of both the trajectory groups and the outcome. We emphasize that our focus is on HRMSMs for which there is a single baseline for each outcome time point. HRMSMs with multiple baselines for each outcome time-point are susceptible to theoretical challenges [31]. In the following,

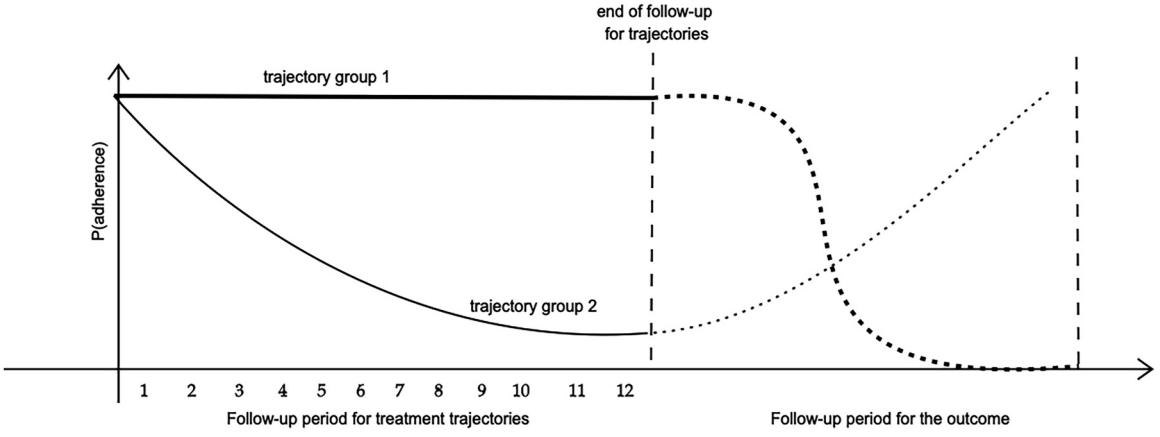

Figure 1: Illustration of LCGA-MSM after the follow-up period of treatment trajectories.

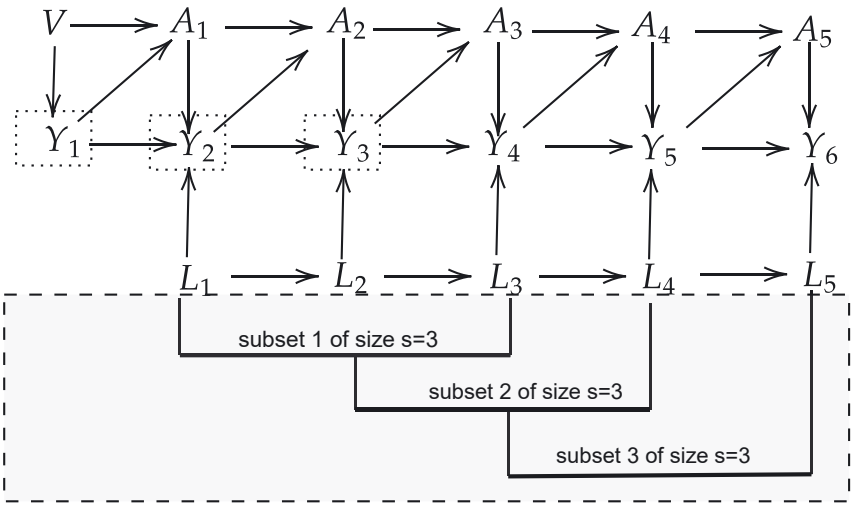

Figure 2: Simplified causal graph for a follow-up of length $K = 6$ and for subsets of length $s = 3$. Associations between the time-varying covariates (L_1, L_2, L_3, L_4 and L_5) and the time-varying treatment (A_1, A_2, A_3, A_4 and A_5) were omitted to avoid overloading the figure. Similarly, unmeasured common causes between L and Y nodes or arrows from Y nodes to L nodes are allowed but omitted. Note that V represents baseline characteristics.

we present the notation in the t-specific framework used for HRMSMs in [16], the data structure, the LCGA in the t-specific framework and, the definition and identification of the causal parameter of interest before giving a definition of HRMSMs.

4.1 Notation in the t-specific framework

In the conventional counterfactual framework, the data are considered from baseline to the end of the follow-up as one set. In HRMSMs, a new form of counterfactual framework is defined: the t-specific (for time-specific) counterfactual framework [16]. In this t-specific framework, multiple time intervals of fixed history size s are chosen from the original data according to the context of the study. We define $\mathcal{T}_s$, the set indexing all time intervals. When the total follow-up is K, we can form $K - s + 1$ time intervals of length s. For example, if $K = 60$ and $s = 6$ months as in our application, we can form 55 overlapping time intervals: [1, 6], [2, 7], [3, 8] … [53, 58], [54, 59], [55, 60]; therefore, $\mathcal{T}_s = \{1, 2, \dots, 55\}$.

We denote by $\bar{A}_d \equiv \bar{A}_{d,d+s-1} = \{A_d, \dots, A_{d+s-1}\}$ the treatment trajectories in the dth time interval $I_d = [d, d + s - 1]$. Similarly, $\bar{L}_{d,d+s-1}$ denote the covariates' history in the dth interval of follow-up, i.e., the history up to the time point $d + s - 1$. We denote by $Y^{\bar{a}_d}_{d+s}$ the counterfactual outcome measured at the end of the time interval I_d and corresponding to a treatment trajectory $\bar{a}_d$ between time points d and $d + s - 1$. In the following, our outcome is the occurrence of a first event in line with our application regarding primary prevention but our approach can easily be extended to a setting with repeated events. In our application, the outcome $Y^{\bar{a}_d}_{d+s} = 1$ if an event would have counterfactually occurred under the treatment regime $\bar{a}_d$ and 0 otherwise. Similarly, $\bar{L}^{\bar{a}_d} \equiv \bar{L}^{\bar{a}_d}_{d,d+s-1}$ denotes a counterfactual covariate process in the dth interval. We denote by F_X the unknown distribution of all possible counterfactual variables $X = \left(\bar{L}^{\bar{a}_d}, Y^{\bar{a}_d}_{d+s} : \bar{a}_d \in \mathcal{A}_d\right)$ where $\mathcal{A}_d$ is the set of all possible values of $\bar{a}_d$ for the dth interval. The exposure at each time t in the dth time interval is considered as bivariate $A_{d,t} = (A_{d,t}(1), A_{d,t}(2))$. The first exposure $A_{d,t}(1)$ represents the statin intake. Thus $A_{d,t}(1) = 1$ indicates that the subject did take their prescriptions at time t in the dth time interval and $A_{d,t}(1) = 0$ otherwise. The second exposure, $A_{d,t}(2)$ is a censoring variable and takes the value 1 if there is right censoring (e.g., end of follow-up or loss to follow-up). When $A_{d,t}(2) = 1$, all subsequent variables of the subject are treated as missing. If $A_{d,t}(2) = 0$, the subject is uncensored in the dth time interval. As will be seen shortly, $A_{d,t}(2)$ will be used to account for

individuals with incomplete exposure follow-up time, either because of loss to follow-up or because an outcome occurred during the exposure follow-up time. This will allow avoiding the potential selection bias that may arise because trajectory groups can only be determined for people with complete exposure follow-up.

4.2 Data structure of HRMSMs

In the following, we present the organization of the data in the context of HRMSMs. For simplicity, we consider a data-generating mechanism adapted from Schnitzer et al. [32]. The follow-up time is of length $K = 6$ and the time-dependent outcome is observed at each time $t = 1, 2, \ldots, 6$ (see Figure 2). However, only outcomes Y_4, Y_5 and Y_6, measured at the end of each subset of length $s = 3$, are considered as dependent variables; indeed a period of three measures has to be spared to construct treatment trajectories [33].

In a typical longitudinal causal inference problem, the observed data are usually represented as: $\mathcal{O} = \{Y_1, L_1, A_1, Y_2, L_2, A_2, Y_3, L_3, A_3, Y_4, L_4, A_4, Y_5, L_5, A_5, Y_6\}$ where L_1, L_2, L_3, L_4, L_5 represent the time-varying covariates, A_1, A_2, A_3, A_4, A_5 the bivariate treatment indicators and $Y_1, Y_2, Y_3, Y_4, Y_5, Y_6$ the time-dependent outcome. This representation of the data entails temporal ordering. The baseline outcome is $Y_1 = 0$ for all individuals. Put differently, all individuals were at risk of experiencing the event at the beginning of the follow-up.

In an HRMSM, the data are rearranged in an augmented dataset that includes a separate row for each interval to which a subject contributes. Continuing the previous example, we get three subsets from the original data: $O^1 = \{Y_1, L_1, A_1, Y_2, L_2, A_2, Y_3, L_3, A_3, Y_4\}$, $O^2 = \{Y_2, L_2, A_2, Y_3, L_3, A_3, Y_4, L_4, A_4, Y_5\}$ and $O^3 = \{Y_3, L_3, A_3, Y_4, L_4, A_4, Y_5, L_5, A_5, Y_6\}$. In a general form, the observed data are represented as $K - s + 1$ data structures O^d: $O^d = \{\bar{A}_{d,d+s-1}, \bar{L}_{d,d+s-1}, \bar{Y}_{d,d+s}\}$ with distribution $P^d_{F_X}, d = 1, \ldots, K - s + 1$. In the t-specific counterfactual framework, these subsets are considered simultaneously [16]. Denote by L^*_t, A^*_t, the covariates, the exposure and the outcome at the tth time point of a given interval, after rearranging the data. The following data structure represents the augmented version of the data: $\mathcal{O} = \left\{Y^*_1 = (Y_1, Y_2, Y_3), L^*_1 = (L_1, L_2, L_3), A^*_1 = (A_1, A_2, A_3), Y^*_2 = (Y_2, Y_3, Y_4), L^*_2 = (L_2, L_3, L_4), A^*_2 = (A_2, A_3, A_4), Y^*_3 = (Y_3, Y_4, Y_5), L^*_3 = (L_3, L_4, L_5), A^*_3 = (A_3, A_4, A_5), Y^*_4 = (Y_4, Y_5, Y_6)\right\}$. Figure 3 summarizes this process. Note that because we require a complete exposure follow-up, individuals with events occurring during the exposure follow-up are considered as censored, that is, if $Y^*_t = 1$ then $A^*_t(2) = 1$. Censoring could also occur in absence of an event during the exposure follow-up if there is a loss to follow-up. Also note that in the following, we only present quantities using the "original" wide format data notation, not the transformed augmented data notation.

4.3 LCGA in the t-specific framework

After rearranging the data in an augmented dataset, the observed treatment trajectories measured in all time intervals are simultaneously summarized into J distinct trajectory groups $z_1, \ldots, z_J$ using LCGA [11]. In other words, an LCGA is applied to the entire pooled data $O^d, d = 1, \ldots, K - s + 1$, simultaneously. This procedure leads to a repeated version of the LCGA. Note that individuals can get assigned to different trajectory groups in different subsets. In the setting where individuals can be lost to follow-up, the LCGA is applied to the available data and the missing data mechanism is considered to be missing at random (MAR) [34]. Note that medico-administrative databases cover the vast majority of the population of interest and there are thus very few losses to follow-up.

4.4 Definition of nonparametric HRMSMs

Given that we are interested in a time-to-event outcome, it may seem natural to define our nonparametric HRMSM as a model for the hazard ratio (HR). However, HRs lead to two problems of interpretation as discussed by Neugebauer et al. [16]. The first problem is that HRs have a well known built-in selection bias [35]. In our context, one possible scenario where selection bias can occur with HRs, is when there exists an unmeasured risk factor U of cardiovascular diseases (Y) as illustrated through the directed acyclic graph (DAG) in Figure 4.

ID	Y_1	L_1	$A_1(1)$	$A_1(2)$	Y_2	L_2	$A_2(1)$	$A_2(2)$	Y_3	L_3	$A_3(1)$	$A_3(2)$	Y_4	L_4	$A_4(1)$	$A_4(2)$	Y_5	L_5	$A_5(1)$	$A_5(2)$	Y_6
1	0	0	1	0	0	1	0	0	0	0	1	0	1	.	.	1	1	.	.	1	1
…	…	..	..	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
n	0	1	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	1	1	0	1

Original data split into three subsets

ID	Y_1	L_1	$A_1(1)$	$A_1(2)$	Y_2	L_2	$A_2(1)$	$A_2(2)$	Y_3	L_3	$A_3(1)$	$A_3(2)$	Y_4
1	0	0	1	0	0	1	0	0	0	0	1	0	1
…	…	..	..	…	…	…	…	…	…	…	…	…	…
n	0	1	0	0	0	0	1	0	0	1	1	0	0

ID	Y_2	L_2	$A_2(1)$	$A_2(2)$	Y_3	L_3	$A_3(1)$	$A_3(2)$	Y_4	L_4	$A_4(1)$	$A_4(2)$	Y_5
1	0	1	0	0	0	0	1	0	1	.	.	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…
n	0	0	1	0	0	1	1	0	0	1	0	0	0

ID	Y_3	L_3	$A_3(1)$	$A_3(2)$	Y_4	L_4	$A_4(1)$	$A_4(2)$	Y_5	L_5	$A_5(1)$	$A_5(2)$	Y_6
1	0	0	1	0	1	.	.	1	1	.	.	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…
n	0	1	1	0	0	1	0	0	0	1	1	0	1

Renaming columns and adding a column for the time interval

ID	d	Y_1^*	L_1^*	$A_1^*(1)$	$A_1^*(2)$	Y_2^*	L_2^*	$A_2^*(1)$	$A_2^*(2)$	Y_3^*	L_3^*	$A_3^*(1)$	$A_3^*(2)$	Y_4^*
1	1	0	0	1	0	0	1	0	0	0	0	1	0	1
…	…	…	..	..	…	…	…	…	…	…	…	…	…	…
n	1	0	1	0	0	0	0	1	0	0	1	1	0	0

ID	d	Y_1^*	L_1^*	$A_1^*(1)$	$A_1^*(2)$	Y_2^*	L_2^*	$A_2^*(1)$	$A_2^*(2)$	Y_3^*	L_3^*	$A_3^*(1)$	$A_3^*(2)$	Y_4^*
1	2	0	1	0	0	0	0	1	0	1	.	.	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
n	2	0	0	1	0	0	1	1	0	0	1	0	0	0

ID	d	Y_1^*	L_1^*	$A_1^*(1)$	$A_1^*(2)$	Y_2^*	L_2^*	$A_2^*(1)$	$A_2^*(2)$	Y_3^*	L_3^*	$A_3^*(1)$	$A_3^*(2)$	Y_4^*
1	3	0	0	1	0	1	.	.	1	1	.	.	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
n	3	0	1	1	0	0	1	0	0	0	1	1	0	1

Augmented dataset

ID	d	Y_1^*	L_1^*	$A_1^*(1)$	$A_1^*(2)$	Y_2^*	L_2^*	$A_2^*(1)$	$A_2^*(2)$	Y_3^*	L_3^*	$A_3^*(1)$	$A_3^*(2)$	Y_4^*
1	1	0	0	1	0	0	1	0	0	0	0	1	0	1
1	2	0	1	0	0	0	0	1	0	1	.	.	1	1
1	3	0	0	1	0	1	.	.	1	1	.	.	1	1
…	…	…	..	..	…	…	…	…	…	…	…	…	…	…
n	1	0	1	0	0	0	0	1	0	0	1	1	0	0
n	2	0	0	1	0	0	1	1	0	0	1	0	0	0
n	3	0	1	1	0	0	1	0	0	0	1	1	0	1

Figure 3: Example of data structure of HRMSMs for a follow-up of length $K = 5$ and with binary variables.

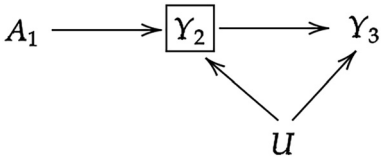

Figure 4: Illustration of the built-in bias problem of HRs with two time points.

Estimation of the HR at the second period of follow-up is conditional on being event-free at the first period of follow-up ($Y_2 = 0$). This conditioning is expected to result in an underestimation of the benefits of a preventive treatment, like statins, at the second period of follow-up. This selection bias is induced by what Hernán [35] called differential selection of less susceptible individuals to the risk factor U or differential depletion of susceptible individuals to the risk factor U. Indeed, individuals most at risk are going to experience the event quicker in the control group than in the treatment group, since the treatment helps to compensate for the increased risk of

these individuals. Thus, at the second period, the treatment group differs from the control group with respect to the risk factor U: the treatment group comprises a greater proportion of individuals at high risk than the control group.

Because of the limitations of HRs as an effect measure, the survival curve or its complement, the cumulative risk, have been recommended [35]. In a standard MSM, survival curves under different trajectories can be constructed using HRs, allowing the comparison of survival curves between treatment trajectories. This leads us to the second problem. Indeed in HRMSMs, it is impossible to construct a survival curve using HRs [16]. In fact, because the trajectory group may change from one window to another, the HR at a given time point t does not correspond to a single trajectory group. To bypass these problems, we propose to use a model for the absolute risk. More precisely, we propose the following form for the nonparametric HRMSM with a time-dependent outcome for the dth interval $I_d = [d, d+s-1]$:

$$E_{F_X}\left[Y_{d+s}^{\bar{a}_d} | Y_d = 0\right] = m^*\left(Y_d = 0\right). \tag{5}$$

For the first interval, the absolute risk corresponds to the number of individuals that experienced the event in the interval I_1 divided by the number of individuals at risk at the beginning of the interval I_1 under the counterfactual regime of treatment $\bar{A}_{1,s}(1) = \bar{a}_{1,s}(1)$ with a counterfactual intervention $\bar{A}_{1,s}(2) = 0$ that prevents censoring or events from happening during the exposure follow-up period for all individuals. The nonparametric HRMSM is thus defined as $P(Y_{s+1}^{\bar{a}_{1,s}(1),\bar{a}_{1,s}(2)=0} = 1 | Y_1 = 0) \equiv P\left(Y_{s+1}^{\bar{a}_{1,s}} = 1\right)$ as $\bar{a}_{1,s} \equiv (\bar{a}_{1,s}(1), \bar{a}_{1,s}(2))$. In the second interval $I_2 = [2, s+1]$, treatment at time point $t = 1$ (a_1) is left random (it is not part of the intervention defined by $\bar{a}_{2,s+1}$). Therefore, we have to condition on $Y_2 = 0$ to ensure that we still have a population at risk at the beginning of the interval I_2. The nonparametric HRMSM is $P(Y_{s+2}^{\bar{a}_{2,s+1}(1),\bar{a}_{2,s+1}(2)=0} = 1 | Y_2 = 0) \equiv P\left(Y_{s+2}^{\bar{a}_{2,s+1}} = 1 | Y_2 = 0\right)$. The risk in the interval I_2 is measured as the number of individuals that experienced the event divided by the number of individuals still at risk at the beginning of interval I_2 under the treatment regime $\bar{A}_{2,s+1}(1) = \bar{a}_{2,s+1}(1)$ and where $\bar{A}_{2,s+1}(2) = 0$ prevents the occurrence of censoring or events during the follow-up period I_2. For the dth interval $I_d = [d, d+s-1]$, the nonparametric HRMSM is $P(Y_{s+d}^{\bar{a}_{d,d+s-1}(1),\bar{a}_{d,d+s-1}(2)=0} = 1 | Y_d = 0) \equiv P\left(Y_{s+d}^{\bar{a}_{d,d+s-1}} = 1 | Y_d = 0\right)$.

Assumptions in the t-specific counterfactual framework are similar to those described for the usual counterfactual framework. The main difference is that each assumption is made relative to each interval or t-specific set of data (see Appendix Section 10.1). Under these assumptions, the nonparametric HRMSM defined in Equation (5) is identified from the observed data as:

$$\mathbb{E}_{F_X}\left(Y_{d+s}^{\bar{a}_d} | Y_d = 0\right) = \int_{l_d} \dots \int_{l_{d+s-1}} \mathbb{E}(Y_{d+s} | Y_d = 0, \bar{A}_{d,d+s-1}(1) = \bar{a}_{d,d+s-1}(1), \bar{A}_{d,d+s-1}(2) = 0, \bar{L}_{d,d+s-1})$$
$$f(l_{d+s-1} | \bar{l}_{d,d+s-2}, \bar{a}_{d,d+s-2}) \mathrm{d}\mu(L_{d+s-2}), \dots f(l_d) \mathrm{d}\mu(L_d), \tag{6}$$

where μs are dominating measures. HRMSMs are a class of MSMs with multiple advantages like a gain of statistical power. In addition, HRMSM allows investigating effect modification according to time-dependent baseline covariates since multiple baselines are considered simultaneously [16].

5 Definition and estimation of the parameters of the LCGA-HRMSM

5.1 Definition

Our causal parameter of interest β is defined as the vector that minimizes the statistical distance between the true causal marginal risk and the marginal risk defined through a working model, using a negative log-likelihood loss function. The statistical estimation problem is defined as the projection of the true model (the nonparametric HRMSM) $m^*\left[Y_d = 0\right] \equiv \mu_d^{\bar{a}_d}$ onto the working model $m(\bar{a}_d | \beta)$ with projection weights $\lambda(\bar{a}_d)$ under all values of $\bar{a}_d$. Because the nonparametric HRMSM is a model for the absolute risk, a natural choice for the working model is a log-linear model, such that the exponential of its coefficients can be interpreted as relative risks. For such a

log-linear model, two common choices to define the loss function are the log-binomial or Poisson negative log-likelihoods. One potential challenge with the log-binomial likelihood is the issue where the common maximum likelihood algorithms failing to converge because predicted probabilities must remain between 0 and 1, which is not ensured by the log-link [36]. The Poisson log-likelihood does not share this limitation, but may yield predicted probabilities greater than 1 [36, 37]. Both loss functions are presented in the following. A specific example of such a log-linear working model is:

$$\log(m(\bar{a}_d|\beta)) = \beta_{0,d} + \beta_1 z_1^*(\bar{a}_d) + \beta_2 z_2^*(\bar{a}_d) + \cdots + \beta_{J-1} z_{J-1}^*(\bar{a}_d). \tag{7}$$

Denoting $z^*(\bar{a}_d) = \left[1, z_1^*(\bar{a}_d), z_2^*(\bar{a}_d), \ldots, z_{J-1}^*(\bar{a}_d)\right]$, the working model can thus be rewritten:

$$\log(m(\bar{a}_d|\beta)) = z^*(\bar{a}_d)\beta. \tag{8}$$

Formally, the causal parameter of interest is defined as:

$$\beta_{m,\lambda} = \arg\min_{\beta} \; E\left[\mathcal{L}_\lambda(\mu_d^{\bar{a}_d}, m(\bar{a}_d|\beta))\right],$$

where $\mathcal{L}_\lambda$ is a loss function that depends on projection weights $\lambda(\bar{a}_d)$. We consider next the full data estimating equations that arise from considering either the log-binomial or the Poisson negative log-likelihood as loss-functions. Intuitively, the estimates of β are chosen such that the working model is as close as possible to the true model. The target parameter of interest pools parameters beta over all time intervals, putting an equal weight on each.

5.1.1 Log-binomial negative log-likelihood loss

The log-binomial likelihood and log-likelihood are respectively given by:

$$L(\beta|\bar{a}) \propto \prod_{d\in\mathcal{T}_s}\prod_{i=1}^{n} m(\bar{a}_d|\beta)^{y_{i,d+s}^{\bar{a}_d}}(1 - m(\bar{a}_d|\beta))^{1-y_{i,d+s}^{\bar{a}_d}},$$

$$\log(L(\beta|\bar{a})) \propto \sum_{d\in\mathcal{T}_s}\sum_{i=1}^{n} y_{i,d+s}^{\bar{a}_d} \log(m(\bar{a}_d|\beta) + \left(1 - y_{i,d+s}^{\bar{a}_d}\right)\log(1 - m(\bar{a}_d|\beta)).$$

The parameter of interest is defined as:

$$\beta = \arg\min_{\beta\in\mathbf{R}^k} E_{F_X}\left\{-\sum_{d\in\mathcal{T}_s}\sum_{\bar{a}_d\in\bar{\mathcal{A}}}\left[Y_{d+s}^{\bar{a}_d}\log(m(\bar{a}_d|\beta) + \left(1 - Y_{d+s}^{\bar{a}_d}\right)\log(1 - m(\bar{a}_d|\beta))\right]\lambda(\bar{a}_d)\right\}. \tag{9}$$

The score $D(X|\beta)$ is given by a first order derivative of the loss function defined in Equation (9). Therefore the score is given by:

$$D(X|\beta) = E_{F_X}\left\{-\sum_{d\in\mathcal{T}_s}\sum_{\bar{a}_d\in\bar{\mathcal{A}}_d}\frac{\partial}{\partial\beta}m(\bar{a}_d|\beta)\left[\frac{Y_{d+s}^{\bar{a}_d} - m(\bar{a}_d|\beta)}{m(\bar{a}_d|\beta)(1 - m(\bar{a}_d|\beta))}\right]\lambda(\bar{a}_d)\right\}. \tag{10}$$

Inserting $m(\bar{a}_d|\beta) = e^{z^*(\bar{a}_d)\beta}$ in Equations (9) and (10), we get:

$$\beta = \arg\min_{\beta\in\mathbf{R}^k} E_{F_X}\left\{-\sum_{d\in\mathcal{T}_s}\sum_{\bar{a}_d\in\bar{\mathcal{A}}}\left[Y_{d+s}^{\bar{a}_d}\log(e^{z^*(\bar{a}_d)\beta}) + \left(1 - Y_{d+s}^{\bar{a}_d}\right)\log(1 - e^{z^*(\bar{a}_d)\beta})\right]\lambda(\bar{a}_d)\right\}, \tag{11}$$

$$D(X|\beta) = E_{F_X}\left\{-\sum_{d\in\mathcal{T}_s}\sum_{\bar{a}_d\in\bar{\mathcal{A}}_d} z^*(\bar{a}_d)\left[\frac{Y_{d+s}^{\bar{a}_d} - e^{z^*(\bar{a}_d)\beta}}{e^{z^*(\bar{a}_d)\beta}(1-e^{z^*(\bar{a}_d)\beta})}\right]\lambda(\bar{a}_d)\right\}. \tag{12}$$

5.1.2 Poisson negative log-likelihood loss

The Poisson likelihood and log-likelihood are:

$$L(\beta|\bar{a}) = \prod_{d\in\mathcal{T}_s}\prod_{i=1}^{n}\frac{e^{y_i^{\bar{a}}\log(m(\bar{a}_d|\beta))}e^{-m(\bar{a}_d|\beta)}}{y_{i,d+s}^{\bar{a}_d}!}$$

$$\log(L(\beta|\bar{a})) = \sum_{d\in\mathcal{T}_s}\sum_{i=1}^{n} y_{i,d+s}^{\bar{a}_d}\log(m(\bar{a}_d|\beta)) - m(\bar{a}_d|\beta) - \log\left(y_{i,d+s}^{\bar{a}_d}!\right).$$

The parameter of interest is defined as:

$$\beta = \underset{\beta\in\mathbf{R}^k}{\arg\min}\, E_{F_X}\left\{-\sum_{d\in\mathcal{T}_s}\sum_{\bar{a}_d\in\bar{\mathcal{A}}}\left[Y_{d+s}^{\bar{a}_d}\log(m(\bar{a}_d|\beta)) - m(\bar{a}_d|\beta)\right]\lambda(\bar{a}_d)\right\}. \tag{13}$$

The score equation $D(X|\beta)$ is given by a first order derivative of the loss function defined in Equation (13). Thus, we have:

$$D(X|\beta) = E_{F_X}\left\{-\sum_{d\in\mathcal{T}_s}\sum_{\bar{a}_d\in\bar{\mathcal{A}}_d}\frac{\partial}{\partial\beta}\left[Y_{d+s}^{\bar{a}_d}\log(m(\bar{a}_d|\beta)) - m(\bar{a}_d|\beta)\right]\lambda(\bar{a}_d)\right\}. \tag{14}$$

Inserting $m(\bar{a}_d|\beta) = e^{z^*(\bar{a}_d)\beta}$ in Equation (13) and $\frac{\partial}{\partial\beta}m(\bar{a}_d|\beta) = z^*e^{z^*(\bar{a}_d)\beta}$ in Equation (14), we get the expression of the loss function and the score equation:

$$\beta = \underset{\beta\in\mathbf{R}^k}{\arg\min}\, E_{F_X}\left\{-\sum_{d\in\mathcal{T}_s}\sum_{\bar{a}_d\in\bar{\mathcal{A}}}\left[Y_{d+s}^{\bar{a}_d}z^*(\bar{a}_d)\beta - e^{z^*(\bar{a}_d)\beta}\right]\lambda(\bar{a}_d)\right\} \tag{15}$$

$\Leftrightarrow$

$$D(X|\beta) = E_{F_X}\left\{-\sum_{d\in\mathcal{T}_s}\sum_{\bar{a}_d\in\bar{\mathcal{A}}_d} z^*(\bar{a}_d)\left[Y_{d+s}^{\bar{a}_d} - e^{z^*(\bar{a}_d)\beta}\right]\lambda(\bar{a}_d)\right\}. \tag{16}$$

5.2 Estimation

In what follows, we present three different estimators of the parameters of the LCGA-HRMSM: IPTW, g-computation and pooled LTMLE. Inferences for each estimator take into account that the same individual can contribute multiple observations in the augmented dataset.

5.2.1 IPTW estimator

Based on the work of Neugebauer and van der Laan [23] and Neugebauer et al. [16], we first propose an IPTW estimator of the parameters of our LCGA-HRMSM. Under the sequential conditional exchangeanbility assumption,

the expression of the IPTW estimator for the dth interval is equivalent to the expression of the IPTW estimator in the LCGA-MSM and is given by:

$$D_{h_{\lambda_d}}(\mathcal{O}|\beta) = \frac{h_{\lambda_d}(\bar{a}_d)\epsilon_{\bar{a}_d}(\beta)}{\prod_{j=d}^{d+s-1} g_d(A_{d,j}(1))g_d(A_{d,j}(2))},$$

where for the dth interval, $g_d(A_{d,j}(1))g_d(A_{d,j}(2)) = P(A_{d,j}(1)|\bar{A}_{d,j-1}(1), A_{d,j}(2) = 0, \bar{L}_{d,j}, Y_d = 0)P(A_{d,j}(2) = 0|\bar{A}_{d,j-1}(1), \bar{A}_{d,j-1}(2) = 0, \bar{L}_{d,j}, Y_d = 0)$, $h_{\lambda_d}(\bar{a}_d) \equiv \lambda_d(\bar{a}_d)\frac{\partial}{\partial\beta}m(\bar{a}_d|\beta) \times m(\bar{a}_d|\beta)$ and $\epsilon_{\bar{a}_d}(\beta) = Y_{d+s}^{\bar{a}_d} - m(\bar{a}_d|\beta)$. The IPTW estimator pooled over time intervals is given by: $D_{h_\lambda}(\mathcal{O}|\beta) = \sum_{d\in\mathcal{T}_s}\frac{h_{\lambda_d}(\bar{a}_d)\epsilon_{\bar{a}_d}(\beta)}{\prod_{j=d}^{d+s-1} g_d(A_{d,j}(1))g_d(A_{d,j}(2))}$. In the case of a log-binomial model, $m(\bar{a}_d|\beta) = [m(\bar{a}_d|\beta)(1 - m(\bar{a}_d|\beta))]^{-1}$ and in the case of a Poisson model $\mathrm{Var}^{-1}\, m(\bar{a}_d|\beta) = [m(\bar{a}_d|\beta)]^{-1}$. The IPTW estimating equations are: $\sum_{d\in\mathcal{T}_s}\sum_{i=1}^{n} D_{h_{\lambda_d}}(\mathcal{O}_i|g_{d,n}, \lambda_{d,n}, \beta) = 0$. Under regularity conditions, the IPTW estimator of β is consistent and asymptotically linear if $\lambda_{d,n}$ and $g_{d,n}$ are consistent estimators of λ_d and g_d [38]. In other words, the causal risk can be estimated consistently using the IPTW estimator if the model for the treatment and the model for the censoring are correctly specified and we have a consistent estimator of the projection weight λ_d.

Fitting Procedure: LCGA-HRMSM with IPTW

In practice, implementation of the IPTW estimator follows these steps:

Step 1: Under the sequential conditional exchangeability assumption, estimate the treatment probability for the dth subset:

$$g_d(A_{d,j}(1))g_d(A_{d,j}(2)) = \prod_{j=d}^{d+s-1} P(A_{d,j}(1)|\bar{A}_{d,j-1}(1), A_{d,j}(2) = 0, \bar{L}_{d,j}, Y_d = 0) \times P\big(A_{d,j}(2) = 0| \times \bar{A}_{d,j-1}(1), \bar{A}_{d,j-1}(2) = 0, \bar{L}_{d,j}, Y_d = 0\big),$$

where $\bar{A}_{d,j-1}(1) \equiv \emptyset$ if $j - 1 < d$. This can be done, for example, using a logistic regression to model exposure at each time point as a function of previous exposure and covariates within the dth subset. Similarly, a logistic regression can be used to model the censoring variable.

Step 2: Construct the unstabilized or stabilized inverse of the treatment mechanism by defining λ_d as in Section 3. If stabilized weights are of interest, set $\lambda_d = \prod_{j=d}^{d+s-1} P(A_{d,j}(1)|\bar{A}_{d,j-1}(1), A_{d,j}(2) = 0, Y_d = 0)P(A_{d,j}(2) = 0|\bar{A}_{d,j-1}(1), \bar{A}_{d,j-1}(2) = 0, Y_d = 0)$. For unstabilized weights, $\lambda_d = 1$.

Step 3: Summarize treatment trajectories into J distinct groups, then assign each individual in one of these groups based on their highest "posterior" probability.

Step 4: Specify a weighted generalized estimating equation model using the IPTW as weights. For example, a pooled weighted Poisson regression model of the outcome with the trajectory groups G (which were obtained from the LCGA) and time d as regressors could be fitted using the augmented dataset. For inferences, a robust variance estimator that accounts for the fact that each individual can contribute multiple observations in the augmented dataset needs to be used. Such a robust estimator is produced by GEE routines. For example, the function *geeglm* from the **R** package *geepack* can be used to fit the robust Poisson model.

5.2.2 G-computation estimator

Another approach to estimate the parameters of an LCGA-HRMSM is g-computation. The IPTW estimator sequentially creates pseudo-populations where $A_{d,t}$ is independent of confounders $\bar{L}_{d,t}$ at each time t of each time interval d. Opposingly, g-computation does not remove the association between the treatment and the confounders. Instead, the counterfactual expectation is estimated under each possible (fixed) treatment regime in the observational data [24]. For example, we would estimate the conditional expectation of the outcome had everyone in the population been perfectly adherent to statins, that is under $\bar{A}_d = 1, \ldots, 1$ for all d. Various algorithms for implementing g-computation have been proposed (e.g., noniterative conditional expectation, iterated conditional expectations) [39, 40]. Here we consider an algorithm based on iterated conditional expectations that

has the advantage of not requiring to model the covariates' distribution. This algorithm for implementing the g-computation estimator of the parameters of LCGA-HRMSMs requires fitting a sequence of outcome models to estimate each counterfactual expectation for all time intervals. Then, each deterministic treatment regime is assigned to a trajectory group using the same LCGA that was fitted on the observed data. To estimate the parameters of LCGA-HRMSMs, the counterfactual expectations are regressed on the trajectory groups and the index of time intervals according to the structural working model specification. In the current application, we are using a weighted log-binomial and a Poisson model with weights λ_n. The latter model gives the parameter estimates of the LCGA-HRMSM.

Fitting Procedure: LCGA-HRMSM with G-computation

- a. Estimation of counterfactual expectations

Under each of the 2^s possible treatment regimes $\bar{a}_d, d = 1 \dots K - s + 1$, the g-computation can be implemented as follows:

Step 0: Set the initial value of the conditional outcome $Q^{\bar{a}_d}_{d,d+s} = Y_{d+s}$

Moving backward, for $j = d + s - 1, \dots, d$:

Step 1: With the observed data fit a logistic regression:

$$E(Q^{\bar{a}_d}_{d,j+1} | \bar{A}_{d,j}(1), \bar{A}_{d,j}(2) = 0, \bar{L}_{d,j}, Y_d = 0) = \text{expit}(\gamma_0 + \gamma_1 \bar{A}_{d,j} + \gamma_2 \bar{L}_{d,j});$$

Step 2: Compute the predicted value under the treatment regime up to time j:

$$\hat{Q}^{\bar{a}_d}_{d,j} = \hat{E}(Q^{\bar{a}_d}_{d,j+1} | \bar{A}_{d,j}(1) = \bar{a}_{d,j}(1), \bar{A}_{d,j}(2) = 0, \bar{L}_{d,j}, Y_d = 0);$$

At the end of the algorithm, compute $\hat{Q}^{\bar{a}_d}_{d,1} = \frac{1}{n_d} \sum_{i=1}^{n_d} Q^{\bar{a}_d}_{1,n_d}(L_{1,j})$. The quantity $\hat{Q}^{\bar{a}_d}_{d,1}$ is an estimate for the counterfactual outcome under treatment regime $\bar{a}_d$, $\mathbb{E}\left[Y^{\bar{a}_d}_{d+s} | Y_d = 0\right]$. Stack the 2^s estimates in a vector Q. Note that this fitting procedure makes it easy to integrate parallel computation and fit simultaneously all $K - s + 1$ time intervals.

- b. Prediction of trajectory groups

Compute the probability of observing each possible treatment regime $\bar{a}_d(1)$ under each of the trajectory groups produced by the LCGA model, that is $P(\bar{a}_d(1) | z = j)$. Then, assign each possible treatment regime to the trajectory group that yields the largest conditional ("posterior") probability.

- c. Estimation of the parameters of a LCGA-HRMSM and inferences

To estimate the parameters of the LCGA-HRMSM, the vector of the estimated counterfactual means Q is regressed on the predicted trajectory groups using a log-binomial regression or a Poisson regression as a working model with weight λ_n.

The standard errors are estimated using block bootstrapping. For each replication of the bootstrap, n individuals and their repeated measures across all subsets are selected with replacement. G-computation is then performed on each subset before performing a pooled regression of the counterfactual means onto the trajectory groups. Standard errors are estimated by taking the standard deviation of the vector of the HRMSM parameter estimates over the B bootstrap replications. Note that the recourse to bootstrap for estimating standard errors makes the g-computation estimator very computationally intensive.

5.2.3 Pooled LTMLE estimator

The IPTW requires fitting a series of models for the treatment, whereas g-computation requires fitting a series of outcome models. Both estimators are consistent only if all the models involved are correctly specified. Doubly robust approaches, such as pooled LTMLE, combine the strength of both methods to yield consistent estimates if either all the models for the treatment or all the models for the outcome are correct. Moreover, the pooled LTMLE is a locally efficient estimator; in other words, the LTMLE achieves the semiparametric asymptotic bounds which guarantees a minimum asymptotic variance among all regular asymptotically linear (RAL) estimators making the same assumptions on the model space when all models are correctly parametrically specified [41, 42]. In practice, the pooled LTMLE is a flexible algorithm that can easily be combined with machine learning techniques.

The pooled LTMLE defines a plug-in estimator for a target parameter denoted ψ. In our context, we denote the LTMLE $\psi^{\bar{a}_d}$ for the dth time interval. The target parameter is expressed as an average over time periods

of the time-specific vector of conditional means. The fitting procedure is similar to the fitting procedure for g-computation. As in g-computation, the procedure first entails estimating the 2^s counterfactual expectations under each possible treatment regime $\bar{a}_d$ for each of the $K - s + 1$ subsets. Each possible treatment regime is then assigned to a trajectory group. In fact, the g-computation estimate serves as an initial estimate in the LTMLE procedure; LTMLE sequentially updates each initial estimate using information from the treatment models that are used in constructing the IPTW estimator. Finally, the estimated counterfactual means are regressed on the predicted trajectory groups and the time interval index using either a log-binomial or a Poisson model, with weight λ_n. It is noteworthy that the target parameter ψ represents the updated counterfactual means upon which we aim to regress the trajectory groups, in order to estimate our causal parameter of interest β.

LCGA-HRMSM with Pooled LTMLE

- a. Estimation of updated counterfactual means

LTMLE estimates are obtained following two major steps: (i) initial estimation of the counterfactual means using a plug-in estimator (i.e., g-computation), (ii) and fluctuation of the initial plug-in estimator with an error such that it solves the equations of the efficient influence curve (EIC) (updating phase) [43, 44].

For each time $j = d + s - 1, \ldots, d$:

Step 1: Estimate the product $\prod_{t=d}^{j} g_d(A_{d,t}(1))g_d(A_{d,t}(2)) = \prod_{t=d}^{j} g_d(A_t(1)|\bar{A}_{d,t-1}(1), A_{d,t}(2) = 0, \bar{L}_{d,t}, Y_d = 0)g_d(\bar{A}_{d,t-1}(2) = 0|\bar{L}_{d,t}, Y_d = 0)$.

Step 2: For all the 2^s treatment regimes, set the initial value of the conditional outcome $Q^{\bar{a}_d}_{d,j+1} = Y_{s+d}$.

Step 3: With the observed data fit a logistic regression for the conditional outcome:

$$E(Q^{\bar{a}_d}_{d,j+1}|Y_d = 1, \bar{A}_{d,j}(1), \bar{A}_{d,j}(2) = 0, \bar{L}_{d,j})) = \text{expit}(\gamma_0 + \gamma_1 \bar{A}_{d,j} + \gamma_2 \bar{L}_{d,j});$$

Step 4: For all 2^s treatment regimes, compute the predicted value:

$$\hat{Q}^{\bar{a}_d}_{d,j} = \hat{E}(Q^{\bar{a}_d}_{d,j+1}|\bar{A}_{d,j}(1) = \bar{a}_{d,j}, \bar{A}_{d,j}(2) = 0, \bar{L}_{d,j} = \bar{l}_{d,j}, Y_d = 0);$$

Step 5: For all 2^s treatment regimes compute the clever covariates:

$$H_{d,j} = \frac{I(\bar{A}_{d,j}(1) = \bar{a}_{d,j}(1), I(\bar{A}_{d,j}(2) = 0))}{\prod_{t=d}^{j} g_d(A_{d,t}(1))g_d(A_{d,t}(2))} \lambda(\bar{a}_{d,t}) \frac{\partial}{\partial \beta} m(\bar{a}_d|\beta) \operatorname{Var}^{-1} m(\bar{a}_d|\beta).$$

The term $\frac{\partial}{\partial \beta} m(\bar{a}_d|\beta) \operatorname{Var}^{-1} m(\bar{a}_d|\beta)$ corresponds to the vector $(1, z_1, z_2 \ldots, z_{J-1})$ for the derivative taken with respect to $\beta_0, \beta_1 \ldots$ and β_{J-1}. Therefore, the dimension of the weights $\mathbf{H}_d$ for all j in the pooled TMLE is $n \times 2^s \times \dim(\beta)$.

Step 6: Estimate the vector of fluctuation errors ϵ by fitting a pooled logistic regression over all 2^s treatment regimes on $\mathbf{H}_d$ without an intercept and with $\hat{Q}^{\bar{a}_d}_{d,j+1}$ as an offset.

$$\text{logit}\left(Q^{\bar{a}^*_d}_{d,j}\right) = \text{logit}\left(\hat{Q}^{\bar{a}_d}_{d,j+1}\right) + \epsilon_d \mathbf{H}_d.$$

The final step of the algorithm consists of updating the target parameter with:

$$Q^{\bar{a}^*_d}_{d,j} = \text{expit}(\text{logit}\left(\hat{Q}^{\bar{a}_d}_{d,j+1}\right) + \epsilon_d \mathbf{H}_d).$$

Steps - **b. Prediction of trajectory groups** and - **c. Estimation of the parameters β of a LCGA-HRMSM**, are performed in the same way as in the g-computation algorithm presented in Section 5.2.2.

5.2.3.1 Variance estimation

Two possible ways to estimate the variance of the pooled LTMLE estimator are to use block bootstrapping as in g-computation or to use the empirical influence curve of the estimator of β. However, in the case of an HRMSMs, using influence curves is not straightforward. Indeed, when using LTMLE to estimate the parameters of a usual MSM (i.e., not an HRMSM), the variance of the estimated target parameter $\hat{\psi}\left(Q^{\bar{a}_t}\right) \equiv \hat{\psi}$ is given by:

$$\mathrm{Var}(\hat{\psi}) \approx \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} IF_{i,\beta}\right) = \frac{1}{n}\mathrm{Var}\left(IF_{i,\beta}\right).$$

The EIC is defined as:

$$IF_{\beta} = \sum_{d\in\mathcal{T}_s} C(Q^d)^{-1}\sum_{t,\bar{a}_d} H(\bar{a}_d,t)(\bar{Q}^{\bar{a}_d,t} - m(\bar{a}_d|\beta)) + \sum_{d\in\mathcal{T}_s} C(Q^d)^{-1}\sum_{t}\sum_{k\in I_d}\sum_{\bar{a}_d} H(\bar{a}_d,t)\frac{I(\bar{A}_d=\bar{a}_d)}{g_d}\left(\bar{Q}^{\bar{a}_d,k+1} - \bar{Q}^{\bar{a}_d,k}\right).$$

where $C(Q^d) = E_{Q^d(L_{d,1})}\sum_{t,\bar{a}_d} H(\bar{a}_d,t)\frac{\partial}{\partial\beta}m(\bar{a}_d|\beta)$ and $H(\bar{a}_d,t)$ designates the clever covariate function of treatment regime $\bar{a}_d$ at time t. The target parameter $\psi^{\bar{a}_d}$ obtained with the pooled LTMLE solves the EIC equation $P_n IF_{\beta}\left(\bar{Q}_n^{\bar{a}_d^*}, g_n, \psi^{\bar{a}_d}\left(Q_n^{\bar{a}_d^*}, Q_{L_{d,1},n}^{\bar{a}_d^*}\right)\right) = 0$ [43].

This result holds when the observations are independent and identically distributed. In the case of an HRMSM, a given individual can contribute data to multiple observations (subsets), thus inducing correlation between observations. When estimating the variance using influence functions, we have to account for this correlation. In our setting, because only subjects who are event-free at the start of a given subset are included in that subset, the number of individuals included decreases ($n_1 \geq n_2 \geq \ldots \geq n_{K-s+1}$). Based on Schnitzer et al. [45], we can show that the variance of the estimated target parameter $\hat{\psi}\left(\sum_{d\in\mathcal{T}_s} Q_{d,1}^{\bar{a}_d}\right) \equiv \hat{\psi}$ using the influence functions is given by:

$$\mathrm{Var}(\hat{\psi}) \approx \mathrm{Var}\left(\frac{1}{n}\sum_{d\in\mathcal{T}_s}\sum_{i=1}^{n_d} IF_{id,\hat{\beta}}\right)$$

$$\begin{aligned}\mathrm{Var}(\hat{\psi}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{d\in\mathcal{T}_s}\sum_{i=1}^{n_d} IF_{id}\right)\\ &= (1/n^2)\left[\sum_{d\in\mathcal{T}_s} n_d\,\mathrm{Var}(IF_{id}) + 2\sum_{d\in\mathcal{T}_s}\sum_{d'>d} n_{d'}\,cov\left(IF_{id}, IF_{id'}\right)\right]\\ &= (1/n^2)\left[\sum_{d\in\mathcal{T}_s} n_d\sigma_d^2 + 2\sum_{d\in\mathcal{T}_s}\sum_{d'>d} n_{d'}\rho_{d,d'}\right]\end{aligned}$$

with $\sigma_d^2 = \mathrm{Var}(IF_{id})$ and $\rho_{d,d'} = cov\left(IF_{id}, IF_{id'}\right)$. This derivation assumes that all individuals are independent and identically distributed with variance σ_d^2 in each of the $d = 1,\ldots,K-s+1$ subsets and $\rho_{d,d'}$ designates the correlation between the influence functions from subsets d and d' among observations arising from the same individual. In other words, this derivation allows for correlations between observations from the same individual across subsets, but not for correlations between observations from different individuals, whether within a given subset or between subsets. Note that $\mathrm{Var}(IF_{id}) = E\left(IF_{id}^2\right)$ since $E(IF_{id}) = 0$ by construction. Finally, it is worth noting that the variance estimator based on the EIC is only consistent if both the treatment and outcome models are correctly specified [46].

As a concrete example, with two subsets we get $\sigma^2 = 1/n^2\left[n_1\sigma_1^2 + n_2\sigma_2^2 + 2n_2\rho_{1,2}\right]$ with $\sigma_1^2 = \mathrm{Var}(IF_{i1})$, $\sigma_2^2 = \mathrm{Var}(IF_{i2})$ and $\rho_{1,2} = cov(IF_{i1}, IF_{i_2})$.

6 Simulation study

6.1 Description

We now present a simulation study that aims to investigate the relative performance of four estimators of the parameters of an LCGA-HRMSM: unstabilized IPTW, g-computation, pooled LTMLE and pooled LTMLE combined with SuperLearner denoted pooled LTMLE + SL. As algorithms for the SuperLearner, we used a

generalized additive model (GAM) and generalized linear model (GLM). We also present the crude unadjusted model. The data-generating mechanism is adapted from [32]. The data generated are of the form $(V, Y_1, L_1, A_1, Y_2, L_2, A_2, Y_3, L_3, A_3, Y_4, L_4, A_4, Y_5, L_5, A_5, Y_6)$ as presented in Section 4. Confounders $V, L_t, t = 1, 2, 3, 4, 5$ are binary variables and $V \subseteq L_1$ is a set of baseline variables. The outcome variable $Y_t, t = 1, 2, 3, 4, 5, 6$ indicates the occurrence ($Y_t = 1$) or absence ($Y_t = 0$) of an event at time t. The exposure $A_t, t = 1, 2, 3, 4, 5$ is binary. In this simulation, the outcome is measured at times $t = 1, 2, \ldots, 6$ but only outcomes measured at times $t = 4, 5$ and 6 are considered as dependent variables ($Y_1 = Y_2 = Y_3 = 0$).

We considered scenarios with 1, 2 and 3 windows of size $s = 3$. We summarized the observed treatment trajectories using 3 trajectory groups. We generated 1,000 datasets of size 5,000 for the observed data. To determine the true values of the parameters, we simulated a population based on the t-specific counterfactual framework by generating 1,000 datasets of size 5,000 for each of the 2^3 treatment regimes. In the classical counterfactual framework, counterfactual data would be generated using the same equations as those used to generate the observed data except that treatment regimes would be fixed instead of random. In the t-specific counterfactual framework, the same principle is applied but only treatment regimes between time point $t - s + 1$ and t are deterministic; treatment regimes between time points 1 and $t - s$ are left random.

For g-computation and pooled LTMLE/pooled LTMLE + SL (Q models), we estimated the mean counterfactual outcome using a logistic regression (see Appendix for the model specification). For g-computation, the standards errors were estimated with 50 replications of block bootstrapping (a number of replications between 50 and 200 is recommended for the estimation of standard errors; see [47] for reference) and for pooled LTMLE/pooled LTMLE + SL, the standards errors were estimated using the influence functions. We considered as working models the log-binomial and Poisson models. For each estimation method, we evaluated the performance by measuring the bias, the standard errors of the estimates (SEE) and the coverage probability of the 95 % confidence intervals.

6.2 Data-generating mechanism

We present here the data-generating mechanism adapted from [32].

$$
\begin{aligned}
&Y_1 = 0\\
&V \sim \mathcal{N}(0,1)/4 + 1\\
&A_1 \sim \mathit{Bernoulli}(\text{expit}(-0.5 + 2.5V))\\
&L_1 \sim \mathit{Bernoulli}(\text{expit}(1 + V + 0.5A_1))\\
&Y_2 = 0\\
&A_2 \sim \mathit{Bernoulli}(\text{expit}(-0.5 + V + 1.2L1))\\
&L_2 \sim \mathit{Bernoulli}(\text{expit}(1 + L_1 + 0.5A_2))\\
&Y_3 = 0\\
&A_3 \sim \mathit{Bernoulli}(\text{expit}(-0.5 + V + 1.2L_2))\\
&L_3 \sim \mathit{Bernoulli}(\text{expit}(1 + L_2 + 0.5A_3))\\
&Y_4 \sim \begin{cases} 1 & \text{if } Y_3 = 1\\ \mathit{Bernoulli}(1 - \text{expit}(1 + V - 0.7L_3 - 0.5A_3)) & \text{if } Y_3 = 0 \end{cases}\\
&A_4 \sim \mathit{Bernoulli}(\text{expit}(-0.5 + V + 1.2L_3)).\\
&L_4 \sim \mathit{Bernoulli}(\text{expit}(1 + L_3 + 0.5A_4))\\
&Y_5 \sim \begin{cases} 1 & \text{if } Y_4 = 1\\ \mathit{Bernoulli}(1 - \text{expit}(0.8V - 0.7L_4 - 0.5A_4)) & \text{if } Y_4 = 0 \end{cases}\\
&A_5 \sim \mathit{Bernoulli}(\text{expit}(-0.5 + V + 1.2L_4)).\\
&L_5 \sim \mathit{Bernoulli}(\text{expit}(1 + L_4 + 0.5A_4))
\end{aligned}
$$

$$Y_6 \sim \begin{cases} 1 & \text{if } Y_5 = 1 \\ Bernoulli(1 - \text{expit}(0.5V - 0.7L_5 - 0.5A_5)) & \text{if } Y_5 = 0 \end{cases}$$

6.3 Specification of the models for the outcome and the treatment

For 1 time interval

$$g_1: A1 \sim V$$
$$g_2: A2 \sim A1 + L1 + V$$
$$g_3: A3 \sim A2 + A1 + L1 + L2 + V$$

$$Q_1: Q2 \sim A1 + L1 + V$$
$$Q_2: Q3 \sim A2 + A1 + L2 + L1 + V$$
$$Q_3: Y \sim A3 + A2 + A1 + L3 + L2 + L1 + V$$

For 2 time intervals

$$g_1: A1 \sim V$$
$$g_2: A2 \sim A1 + L1 + V$$
$$g_3: A3 \sim A2 + A1 + L1 + L2 + V$$
$$g_4: A4 \sim A3 + A2 + A1 + L1 + L2 + L3 + V | Y4 = 0$$

$$Q_1: Q2 \sim A1 + L1 + V1 + V2 + V3$$
$$Q_2: Q3 \sim A2 + A1 + L2 + L1 + V1 + V2 + V3$$
$$Q_3: Y \sim A3 + A2 + A1 + L3 + L2 + L1 + V1 + V2 + V3$$

For 3 time intervals

$$g_1: A1 \sim V$$
$$g_2: A2 \sim A1 + L1 + V$$
$$g_3: A3 \sim A2 + A1 + L1 + L2 + V$$
$$g_4: A4 \sim A3 + A2 + A1 + L1 + L2 + L3 + V | Y4 = 0$$
$$g_5: A5 \sim A4 + A3 + A2 + A1 + L1 + L2 + L3 + L4 + V | Y5 = 0$$

$$Q_1: Q2 \sim A1 + L1 + V1 + V2 + V3 + V4 + V5$$
$$Q_2: Q3 \sim A2 + A1 + L2 + L1 + V1 + V2 + V3 + V4 + V5$$
$$Q_3: Y \sim A3 + A2 + A1 + L3 + L2 + L1 + V1 + V2 + V3 + V4 + V5$$

6.4 Results

Results of the scenarios with 1, 2 and 3 time intervals are shown respectively in Tables 1–3. In the scenario with a single time interval (Table 1), all estimation methods yielded unbiased estimates, except for the crude model. Moreover, results obtained when using a log-binomial or a Poisson loss function were almost identical. In the other scenarios, g-computation and pooled LTMLE/pooled LTMLE + SL also yielded estimates with little to no

Table 1: Simulation study results with 1 time interval.

	Log-binomial model			Poisson model		
	Bias	**SEE**	**C95 %**	**Bias**	**SEE**	**C95 %**
Group 1						
IPTW	0.00	0.11	94 %	0.00	0.11	94 %
G-computation + bootstrap	−0.01	0.10	90 %	−0.01	0.10	90 %
Pooled LTMLE	0.00	0.10	98 %	0.00	0.10	98 %
Pooled LTMLE + SL	0.00	0.10	98 %	0.00	0.10	98 %
Crude model	−0.04	0.14	80 %	−0.04	0.14	80 %
Group 2						
IPTW	0.00	0.08	94 %	0.00	0.08	94 %
G-computation + bootstrap	−0.01	0.06	92 %	−0.01	0.06	92 %
Pooled LTMLE	−0.01	0.06	100 %	−0.01	0.06	100 %
Pooled LTMLE + SL	0.00	0.06	100 %	0.00	0.06	100 %
Crude model	−0.14	0.05	9 %	−0.14	0.05	9 %

Bias, *SEE* and C95 % indicate respectively the bias, the standard errors of the estimates and the coverage probability of the 95 % confidence interval. We considered three trajectory groups; the group with the lowest adherence is the reference group (group 3).

Table 2: Simulation study results with 2 time intervals.

	Log-binomial model			Poisson model		
	Bias	**SEE**	**C95 %**	**Bias**	**SEE**	**C95 %**
Group 1						
IPTW	−0.07	0.11	100 %	−0.02	0.10	92 %
G-computation + bootstrap	0.00	0.10	95 %	0.00	0.10	95 %
Pooled LTMLE	0.00	0.10	100 %	0.00	0.10	100 %
Pooled LTMLE + SL	0.00	0.11	99 %	0.00	0.11	99 %
Pooled LTMLE + bootstrap	0.00	0.11	97 %	0.00	0.11	96 %
Crude model	−0.06	0.12	77 %	−0.06	0.12	77 %
Group 2						
IPTW	0.02	0.14	100 %	−0.02	0.07	94 %
G-computation + bootstrap	−0.01	0.05	96 %	−0.01	0.05	96 %
Pooled LTMLE	−0.01	0.05	99 %	−0.01	0.05	99 %
Pooled LTMLE + SL	−0.01	0.05	100 %	−0.01	0.05	100 %
Pooled LTMLE + bootstrap	−0.01	0.05	96 %	−0.01	0.05	96 %
Crude model	−0.12	0.04	12 %	−0.12	0.04	12 %

Bias, *SEE* and C95 % indicate respectively the bias, the standard errors of the estimates and the coverage probability of the 95 % confidence interval. In the case of two time intervals, convergence problems were noted with the log-binomial model. We considered three trajectory groups; the group with the lowest adherence is the reference group (group 3).

bias for all coefficients. The IPTW estimates were overall slightly more biased, particularly in the Scenario with 2 time intervals when employing a log-binomial loss function. This may be due to the convergence issues that were encountered in this specific case. This bias decreased in simulations with a larger sample size (see Appendix 1.4) Of note, the bias was similar for both pooled LTMLE implementations, regardless if SuperLearner was used or not.

Table 3: Simulation study results with 3 time intervals.

	Log-binomial model			Poisson model		
	Bias	**SEE**	**C95 %**	**Bias**	**SEE**	**C95 %**
Group 1						
IPTW	0.00	0.11	88 %	−0.02	0.09	90 %
G-computation + bootstrap	−0.01	0.09	97 %	−0.01	0.09	97 %
Pooled LTMLE	−0.01	0.09	100 %	−0.01	0.09	100 %
Pooled LTMLE + SL	−0.01	0.09	100 %	−0.01	0.09	99 %
Pooled LTMLE + bootstrap	−0.01	0.09	99 %	−0.01	0.09	99 %
Crude model	−0.05	0.11	75 %	−0.05	0.11	75 %
Group 2						
IPTW	0.00	0.10	96 %	−0.03	0.07	90 %
G-computation + bootstrap	−0.01	0.05	98 %	−0.01	0.05	98 %
Pooled LTMLE	−0.01	0.05	100 %	−0.01	0.05	100 %
Pooled LTMLE + SL	−0.01	0.05	100 %	−0.01	0.05	100 %
Pooled LTMLE + bootstrap	0.00	0.05	99 %	−0.01	0.05	99 %
Crude model	−0.13	0.04	4 %	−0.13	0.04	4 %

Bias, *SEE* and C95 % indicate respectively the bias, the standard errors of the estimates and the coverage probability of the 95 % confidence interval. We considered three trajectory groups; the group with the lowest adherence is the reference group (group 3).

In all scenarios, the SEE of the different estimators were similar, except for IPTW that had overall greater SEE. Overall, the estimators had a lower SEE in the scenarios with more time intervals (see Table 3). In all scenarios, pooled LTMLE/pooled LTMLE + SL yielded comparable confidence interval coverages, between 94 % and 100 %. IPTW had coverage probabilities between 90 % and 100 % in all scenarios and g-computation yielded coverage probabilities between 89 % and 92 % in the scenario with one time interval (Table 1), and between 95 % and 98 % with the scenarios with two and three time intervals (Tables 2 and 3). The Monte Carlo standard error for the bias estimated as $SEE/\sqrt{R}$ was between 0.001 and 0.004; the Monte Carlo error for the coverage probability estimated as $\sqrt{coverage \times (1 - coverage)/R}$ is between 0 and 0.009 with $R = 1,000$ replications [48].

7 Application

We applied our LCGA-HRMSM approach to estimate the effect of statin usage trajectories for primary prevention of CVD or all-cause mortality among older adults using the data described in Section 2. For a better understanding of how the data are organized see Figure 5.

The first time interval was composed by 57,211 statin initiators and the last interval by 2,315 individuals who were still alive and CVD-free. A pooled LCGA was fitted with all the individuals at risk at the beginning of each time interval. Treatment trajectories were summarized into 3 trajectory groups and with a log-linear function of time (see Figure 6). We have shown in a previous work with the same dataset that the choice of 3 groups and a linear form was suitable according to the Bayesian Information Criterion (BIC) and the Entropy Information Criterion (see Diop et al. [10]). Groups were ranked from highest to lowest adherence. The group with the lowest adherence was considered as the reference group. To estimate the relative risks, we considered log-binomial and Poisson working models. We also considered all three adjustment methods: IPW (product of IPTW and inverse probability of censoring weights (IPCW)), g-computation and pooled LTMLE. For the bootstrap 100 replications were considered to estimate the standard errors and construct the 95 % confidence intervals.

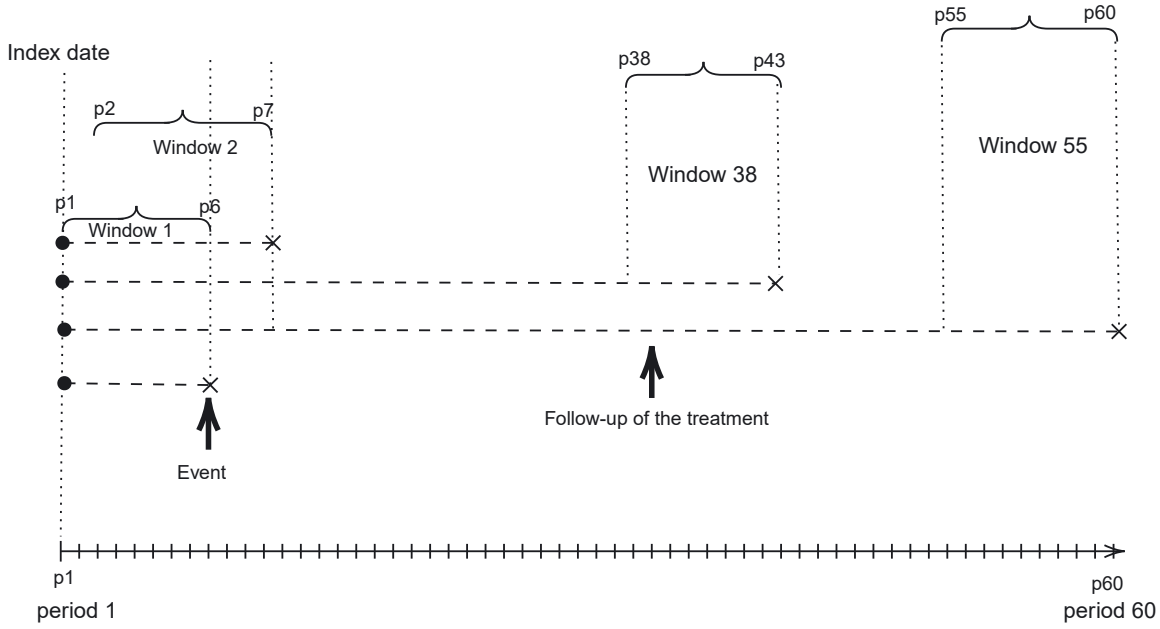

Figure 5: Illustration of the data structure for a follow-up of 60 months (from April 2013 to March 2018) and *p* indicates the follow-up period. Each participant is followed from statin initiation (index date) until occurrence of an event.

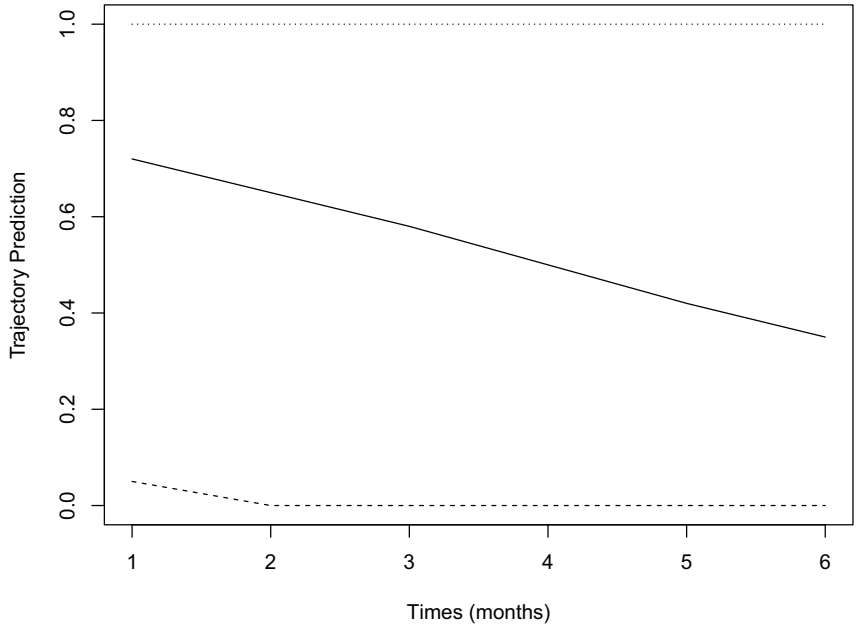

Figure 6: Predicted trajectory groups in the analysis of statin usage for primary prevention among older adults in Quebec, Canada.

7.1 Results for the marginal models

The mean of adherence across all windows was 0.99 in the first group, 0.54 in the second group and 0.02 in the third group. Overall, the data seemed to indicate a decreasing risk of CVD/all-cause mortality events among participants with a better statin adherence (see Table 4). The estimated relative risk is lower than 1 for the group with the highest adherence compared to the reference group with the lowest adherence for all three

Table 4: Application of LCGA-HRMSM to estimate statin usage trajectories for primary prevention of CVD or all-cause mortality among older adults using IPW, g-computation and pooled LTMLE. Both log-binomial and Poisson were considered as working models with group 1 being the group with the lowest adherence probability followed by group 2.

	Log-binomial				Poisson			
	RR	**SE**	**Lower 0.95**	**Upper 0.95**	**RR**	**SE**	**Lower 0.95**	**Upper 0.95**
				Crude model				
Group 1	0.87	0.10	0.71	1.06	0.87	0.10	0.71	1.19
Group 2	1.57	0.13	1.20	2.06	1.56	0.14	1.06	2.04
				IPW				
Group 1	0.77	0.19	0.53	1.10	0.76	0.19	0.53	1.10
Group 2	1.07	0.23	0.69	1.67	1.04	0.23	0.67	1.64
				G-computation + bootstrap				
Group 1	0.65	0.07	0.57	0.75	0.66	0.08	0.58	0.77
Group 2	0.81	0.27	0.48	1.39	0.83	0.27	0.49	1.40
				Pooled LTMLE				
Group 1	0.75	0.13	0.59	0.97	0.75	0.13	0.59	0.97
Group 2	0.89	0.33	0.47	1.71	0.89	0.33	0.47	1.71

RR, relative risk; SE, standard errors; lower 0.95 and upper 0.95 the lower and the upper bounds of 95 % confidence intervals.

estimation methods: IPW, g-computation and pooled LTMLE (IPW – log-binomial: RR = 0.77; Poisson: RR = 0.76; g-computation, log-binomial: RR = 0.65 and Poisson: RR = 0.66; pooled LTMLE, log-binomial and Poisson RR = 0.75). However, the 95 % confidence interval is wider when estimating the relative risk with IPW and includes 1, 95 % CI: [0.53, 1.10] for both log-binomial and Poisson models, suggesting that the effect of a better adherence to statins might also be null or detrimental up to a relative risk of 1.10. For g-formula and pooled LTMLE the 95 % confidence intervals were 95 % CI: [0.57, 0.75] and 95 % CI: [0.59, 0.97] both suggesting beneficial effects of a high statin adherence over a low adherence.

For the group with the second highest adherence to statins, 95 % confidence intervals obtained with IPW, g-formula and pooled LTMLE included 1. More precisely, the relative risk when comparing the second group to the reference group was close to 1 when using the IPW (log-binomial: RR = 1.07, 95 % CI: [0.69, 1.67]; Poisson: RR = 1.04, 95 % CI: [0.67, 1.64]). The estimated relative risk was below 1 when using either g-computation or pooled LTMLE (RR = 0.81, 95 % CI: [0.48, 1.39]; RR = 0.89, 95 % CI: [0.47, 1.71]).

8 Discussion

In this study, we proposed an extension of the LCGA-MSM framework to a time-dependent outcome using a non-parametric HRMSM. LCGA is used to summarize treatment trajectories into few trajectory groups; then HRMSM is used to relate the trajectory groups to the time-dependent outcome. HRMSMs are seen as a generalization of a standard MSM and allow modeling the risk of an event at each time interval according to the recent trajectory. As far as we know, we present the first application of HRMSMs with a time-to-event outcome. It was previously noted that HRMSMs could pose interpretation problems in survival analysis when either targeting a hazard ratio or a survival curve [16]. To bypass these interpretation challenges, we proposed as causal parameter the absolute risk. We considered a weighted log-binomial and Poisson working models to estimate the absolute risk with weight λ_n. To estimate the parameters of an LCGA-HRMSM, we used unstabilized IPTW, g-computation and pooled LTMLE without and with SuperLearner. We also proposed an approach to estimate the variance based on influence functions when using the pooled LTMLE.

We conducted a simulation study to assess the empirical performance of the proposed LCGA-HRMSM estimators. Overall, results are similar when considering a log-binomial or a Poisson working model. For all scenarios, we obtained unbiased estimates when using either g-computation or pooled LTMLE/pooled LTMLE + SL. This result was particularly expected with the pooled LTMLE as it is a doubly robust method. Unstabilized IPTW had a less good performance when considering two time intervals. Indeed, IPTW is well known in the literature for inducing more variability in the estimation of parameters. All approaches had good coverage of the 95 % confidence intervals. Estimates of the variance after correction for the pooled LTMLE are similar to the estimates of the variance obtained trough block-bootstrapping. This result validates our proposed correction for the dependence between influence functions. We also applied our approach to a population of 57,211 statin initiators to investigate the benefits of using statin for primary prevention of CVD or death events among older Quebecers. We found that the estimates obtained with the g-computation and pooled LTMLE lead to the same conclusions that high adherence to statins might be beneficial for primary prevention and reduce the risk of CVD or death among older adults. However, the IPW gave a slightly different, more inconclusive, result. For an average statin adherence, the three estimation methods produce inconclusive results, since the 95 % confidence intervals show that the data are compatible with both clinically meaningful beneficial and detrimental associations.

Along with the strength of LCGA-HRMSMs, there are some limitations. In practice, a user can encounter challenges regarding the choice of the hyperparameters s which is the size of a time interval or J the number of trajectory groups. With the wrong choice of s, estimation problems might arise. For example, this can happen when in a time interval all individuals are exposed or unexposed. Verification can be done for all time intervals, to verify if the chosen s allows a good distribution of the data. As for the choice of the number of groups J, it can be chosen based on different criterias as the Bayesian information criterion, through cross-validation or bootstrap [11, 49, 50]. As argued in Nagin [11] and Diop et al. [10], trajectory groups are not meant to represent the true data-generating distribution but can be seen as points of support. Another limitation is that the LCGA step assumes that missingness in the treatment trajectory is MAR conditional on the groups. If this assumption does not hold, the LCGA may fail to identify the true trajectory groups, if trajectory groups truly exist. However, because we take the perspective that the trajectory groups are a convenient approximation to the truth, and define our estimand conditional on the groups that are formed, we do not see this limitation as fundamental. Nonetheless, extensions that relax the assumptions on the missingness mechanism for the LCGA step may be an interesting direction for future work. For example, Vermunt et al. [34] discuss ways of accommodating not missing at random data.

We believe that LCGA-HRMSM is an interesting and useful approach when one needs to investigate the effect of different adherence profiles on a time-dependent outcome. In this paper, our focus was on primary intervention. Thus, we did not take into account recurrent events. However, our approach LCGA-HRMSM is easily extendable to the case where we encounter multiple events which makes it interesting for more complex settings. Moreover, important gains in computation time can be made when using g-computation or pooled LTMLE/pooled LTMLE + SL as the fitting procedure can be applied simultaneously on all time intervals using, for example, parallel computation.

Acknowledgments: The authors acknowledge the usage of Laval University computing resources.
Research ethics: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no competing interests.
Research funding: This work was supported by the Canadian Institutes of Health Research – CIHR, Grant Number: .A Diop was also supported by the Centre de Recherche en Santé Durable – VITAM and a scolarship for the end of her doctoral studies by Laval University. D Talbot, C Sirois and J R Guertin are a Fonds de recherche du Québec – Santé (FRQS) Chercheur-Boursier. Mireille E. Schnitzer holds the Canada Research Chair in Causal Inference and Machine Learning in Health Science.

Data availability: The data used in this study are available from the Institut national de santé publique du Québec. The R code to perform these methods is implemented in the R package trajmsm, which is available on CRAN: https://cran.r-project.org/web/packages/trajmsm/index.html.

References

1. World Health Organization. Cardiovascular diseases (CVDs); 2019. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.
2. Pigeot I, De Henauw S, Foraita R, Jahn I, Ahrens W. Primary prevention from the epidemiology perspective: three examples from the practice. BMC Med Res Methodol 2010;10:1—11.
3. Habib AR, Katz MH, Redberg RF. Statins for primary cardiovascular disease prevention: time to curb our enthusiasm. JAMA Intern Med 2022;182:1021—4.
4. Mortensen MB, Falk E. Primary prevention with statins in the elderly. J Am Coll Cardiol 2018;71:85—94.
5. de Keyser CE. Pharmacogenetic epidemiology of statins in an ageing population [Ph.D. thesis]. Erasmus Universiteit Rotterdam; 2015.
6. Brown MT, Bussell JK. Medication adherence: who cares? In: Mayo clinic proceedings. Elsevier; 2011, vol 86:304—14 pp.
7. Ho PM, Bryson CL, Rumsfeld JS. Medication adherence: its importance in cardiovascular outcomes. Circulation 2009;119:3028—35.
8. Nau DP. Proportion of days covered (PDC) as a preferred method of measuring medication adherence. Springfield, VA: Pharmacy Quality Alliance; 2012.
9. Franklin JM, Shrank WH, Pakes J, Sanfélix-Gimeno G, Matlin OS, Brennan TA, et al. Group-based trajectory models: a new approach to classifying and predicting long-term medication adherence. Med Care 2013;51:789—96.
10. Diop A, Sirois C, Guertin JR, Schnitzer ME, Candas B, Cossette B, et al. Marginal structural models with latent class growth analysis of treatment trajectories: statins for primary prevention among older adults. Stat Methods Med Res 2023;32:2207—25.
11. Nagin DS. Group-based modeling of development. Cambridge: Harvard University Press; 2005.
12. Johnstone MT, Veves A. Diabetes and cardiovascular disease. Totowa: Springer Science & Business Media; 2005.
13. Kannel WB, McGee DL. Diabetes and cardiovascular risk factors: the Framingham study. Circulation 1979;59:8—13.
14. Sattar N, Preiss D, Murray HM, Welsh P, Buckley BM, Craen AJMD, et al. Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials. Lancet 2010;375:735—42.
15. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Statistical models in epidemiology, the environment, and clinical trials. New York: Springer; 2000.
16. Neugebauer R, van der Laan MJ, Joffe MM, Tager IB. Causal inference in longitudinal studies with history-restricted marginal structural models. Electron J Stat 2007;1:119—54.
17. Petersen ML, Deeks SG, Martin JN, Van Der Laan MJ. History-adjusted marginal structural models for estimating time-varying effect modification. Am J Epidemiol 2007;166:985—93.
18. van der Laan MJ, Petersen ML, Joffe MM. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. Int J Biostat 2005;1:4. https://doi.org/10.2202/1557-4679.1003.
19. Blais C, Jean S, Sirois C, Rochette L, Plante C, Larocque I, et al. Quebec integrated chronic disease surveillance system (QICDSS), an innovative approach. Chronic Dis Inj Can 2014;34:226—35.
20. Francis B, Elliott A, Weldon M. Smoothing group-based trajectory models through B-splines. J Dev Life-Course Criminol 2016;2:113—33.
21. Leisch F. FlexMix: a general framework for finite mixture models and latent glass regression in R. J Stat Software 2004;11:1—18.
22. Muthen B. Latent variable mixture modeling. In: New developments and techniques in structural equation modeling. New York: Taylor & Francis; 2001.
23. Neugebauer R, van der Laan M. Nonparametric causal effects based on marginal structural models. J Stat Plann Inference 2007;137:419—34.
24. Hernan MA, Robins JM. Causal inference: what if. Boca Raton: Chapman and Hall/CRC; 2020.
25. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med 2004;23:2937—60.
26. Basu D. The family of ancillary statistics. Sankya A 1959;21:247—56.
27. Leo Lehmann E, Scholz FW. Ancillarity. Lect Notes Monogr Ser 1992;17:32—51.
28. Reid S, Tibshirani R. Sparse regression and marginal testing using cluster prototypes. Biostatistics 2016;17:364—76.
29. Reid S, Taylor J, Tibshirani R. A general framework for estimation and inference from clusters of features. J Am Stat Assoc 2018;113:280—93.
30. Loh WW, Kim J-S. Evaluating sensitivity to classification uncertainty in latent subgroup effect analyses. BMC Med Res Methodol 2022;22:1—18.

31. Robins JM, Hernan MA, Rotnitzky A. Invited commentary: effect modification by time-varying covariates. Am J Epidemiol 2007;166:994—1002.
32. Schnitzer ME, Moodie EEM, Platt RW. Targeted maximum likelihood estimation for marginal time-dependent treatment effects under density misspecification. Biostatistics 2013;14:1—14.
33. Curran PJ, Obeidat K, Losardo D. Twelve frequently asked questions about growth curve modeling. J Cognit Dev 2010;11:121—36.
34. Vermunt JK, Tran B, Magidson J. Latent class models in longitudinal research. In: Handbook of longitudinal research: design, measurement, and analysis. Burlington: Elsevier; 2008.
35. Hernán MA. The hazards of hazard ratios. Epidemiology 2010;21:13—15.
36. Williamson T, Eliasziw M, Fick GH. Log-binomial models: exploring failed convergence. Emerg Themes Epidemiol 2013;10:1—10.
37. Chen W, Qian L, Shi J, Franklin M. Comparing performance between log-binomial and robust Poisson regression models for estimating risk ratios under model misspecification. BMC Med Res Methodol 2018;18:1—12.
38. Neugebauer R, van der Laan MJ. Locally efficient estimation of nonparametric causal effects on mean outcomes in longitudinal studies. In: U.C. Berkeley division of biostatistics working paper series; 2003, vol 132:1—25 pp.
39. Chatton A, Le Borgne F, Leyrat C, Foucher Y. G-computation and doubly robust standardisation for continuous-time data: a comparison with inverse probability weighting. Stat Methods Med Res 2022;31:706—18.
40. Wen L, Young JG, Robins JM, Hernán MA. Parametric g-formula implementations for causal survival analyses. Biometrics 2021;77:740—53.
41. Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA. Efficient and adaptive estimation for semiparametric models. New York: Springer; 1993, vol 4.
42. Tsiatis A. Semiparametric theory and missing data. New York: Springer Science & Business Media; 2007.
43. Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan M. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. J Causal Inference 2014;2:147—85.
44. Rosenblum M, van der Laan MJ. Targeted maximum likelihood estimation of the parameter of a marginal structural model. Int J Biostat 2010;6:19. https://doi.org/10.2202/1557-4679.1238.
45. Schnitzer ME, van der Laan MJ, Moodie EEM, Platt RW. Effect of breastfeeding on gastrointestinal infection in infants: a targeted maximum likelihood approach for clustered longitudinal data. Ann Appl Stat 2014;8:703—25.
46. van der Laan MJ, Gruber S. Targeted minimum loss based estimation of an intervention specific mean outcome. In: U.C. Berkeley division of biostatistics working paper series; 2011, vol 290:1—38 pp.
47. Tibshirani RJ, Efron B. An introduction to the bootstrap. Monogr Stat Appl Probab 1993;57:1—436.
48. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med 2019;38:2074—102.
49. Mésidor M, Sirois C, Simard M, Talbot D. A bootstrap approach for evaluating uncertainty in the number of groups identified by latent class growth models. Am J Epidemiol 2023;192:1896—903.
50. Nielsen JD, Rosenthal JS, Sun Y, Day DM, Bevc I, Duchesne T. Group-based criminal trajectory analysis using cross-validation criteria. Commun Stat Theor Methods 2014;43:4337—56.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/ijb-2023-0116).