Impact of discretization of the timeline for longitudinal causal inference methods

Steve Ferreira Guerra[1,2], Mireille E. Schnitzer*[1,2], Amélie Forget[1,3], and Lucie Blais[1,3]

[1]Faculté de Pharmacie, Université de Montréal, Montréal, Canada
[2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Canada
[3]Research Center, Hôpital du Sacré-Coeur de Montréal, Montréal, Canada

Abstract

In longitudinal settings, causal inference methods usually rely on a discretization of the patient timeline that may not reflect the underlying data generation process. This paper investigates the estimation of causal parameters under discretized data. It presents the implicit assumptions practitioners make but do not acknowledge when discretizing data to assess longitudinal causal parameters. We illustrate that differences in point estimates under different discretizations are due to the data coarsening resulting in both a modified definition of the parameter of interest and loss of information about time-dependent confounders. We further investigate several tools to advise analysts in selecting a timeline discretization for use with pooled Longitudinal Targeted Maximum Likelihood Estimation for the estimation of the

*Corresponding author: Mireille E. Schnitzer, Faculté de Pharmacie, Université de Montréal, Pavillon Jean-Coutu, 2940, chemin de Polytechnique. Email: mireille.schnitzer@umontreal.ca

parameters of a marginal structural model. We use a simulation study to empirically evaluate bias at different discretizations and assess the use of the cross-validated variance as a measure of data support to select a discretization under a chosen data coarsening mechanism. We then apply our approach to a study on the relative effect of alternative asthma treatments during pregnancy on pregnancy duration. The results of the simulation study illustrate how coarsening changes the target parameter of interest as well as how it may create bias due to a lack of appropriate control for time-dependent confounders. We also observe evidence that the cross-validated variance acts well as a measure of support in the data, by being minimized at finer discretizations as the sample size increases.

Keywords: electronic health data; coarsening; TMLE; semi-parametric estimation; cross-validation

1 Introduction

In health care research, causal inference has become central to the investigation of questions involving the effect of exposures in observational studies. [1] Administrative data have rapidly been embraced as a powerful tool for observational research due to their broad potential including relatively low cost, large size, longitudinal nature and long follow-up period. [2] In spite of these benefits, such databases were not intended to be used for research as the data are mainly collected for administrative purposes, which leads to inherent problems including poor data quality, absence of information on confounders, and comparability of data. [3] Furthermore, with any type of observational data that represent real-world processes, such as administrative data, the times at which exposures may change may not be controlled by the study design or through a regular schedule of follow-up visits, leading to an underlying exposure process that changes in continuous time. However, most existing methods for longitudinal causal inference assume that exposure changes only at common discrete time-points and therefore rely on a discretization of the timeline assumed to represent the true underlying data generating process. [4] Consequently, an analytical issue arises when the coarsening is left to the possibly arbitrary choice of the researcher. [5]

Through a framework that conceptualizes and defines discretization, we illustrate that arbitrary discretization can become problematic because the chosen discretization scale for the analysis can affect the definition of the target parameter and the assumptions required for estimating causal effects from longitudinal observational data. Hence, as for conditional parametric models, [6] the estimated Marginal Structural Model (MSM) parameters [7, 8] may change with different discretizations. In addition, in the presence of time-dependent confounding, if one chooses a scale that is not fine enough to capture the intricate relations between exposure, time-dependent confounders, and outcome, the estimates may be biased. On the other hand, an excessively narrow scale could lead to data sparsity at the individual timepoints, potentially resulting in finite-sample bias and inflated variance. [8–10] This leaves open the question of how to select an appropriate discretization, ideally at a fine enough scale that would capture all time-dependent confounding while balancing for inflated variance. Even though some concerns and warnings regarding arbitrary discretization have emerged, [11, 12] no criteria have yet been proposed to guide an analyst's choice of discretization. We investigate different tools which collectively inform whether there is adequate data support for a given discretization for use with pooled Longitudinal Targeted Maximum Likelihood Estimation (LTMLE). [13] Specifically, we propose that the finest possible discretization be conditionally chosen for analysis given appropriate data support, primarily informed by tables of longitudinal frequencies of exposure and censoring, by the convergence of the pooled LTMLE algorithm, and by the pooled LTMLE variance, evaluated through cross-validation. [14] We evaluate our approach through a simulation study and provide empirical evidence that the cross-validated variance supplies information about the data support. Finally, we apply our selection procedure to a real-world application of the evaluation of asthma treatment on pregnancy duration. To the best of our knowledge, this analysis is the first to account for time-dependent confounding in that setting.

It is known that traditional methods fail to produce unbiased estimation of causal effects in the presence of time-dependent confounding. [15] LTMLE [16, 17] is a doubly robust method for estimating longitudinal treatment effects. It has also been shown to be less biased and result in smaller variance than Inverse of Probability of Treatment Weighting (IPTW), [8, 9] especially in cases where data sparsity occurs, [13, 18] which makes the method appealing in this context. Pooled LTMLE [13] is a more robust ver-

sion of LTMLE for the estimation of the coefficients of an MSM that has been shown to perform better than alternative implementations of LTMLE. [13]

The paper is organized as follows. In section 2, we will present the motivating example of the effect of asthma medication on time to delivery. In section 3, we describe the general data structure and introduce the concept of a discretized dataset. In section 4, we introduce our causal parameter of interest. In section 5, we explore how discretization may affect the target parameter and the plausibility of the identifiability assumptions. In section 6, we review the pooled LTMLE algorithm to present our corresponding selection procedure. Section 7 presents the simulation study and Section 8 the real data application and results. Finally, we discuss discretization in practice, alternative approaches, and future lines of inquiry.

2 Motivating example

The investigation of discretization strategies was motivated by a study on the evaluation of the safety of asthma controller medications on pregnancy outcomes. [19] In the initial study, data on pregnant asthmatic women with deliveries between 1998 and 2008 were extracted from the linkage of the RAMQ and MED-ECHO administrative databases in the province of Québec, Canada. Information on pregnancy outcomes, exposure to asthma medication, and related confounders were assessed from prescription renewals in community pharmacies, hospitalisations, emergency room visits, and outpatient medical consultations. For additional information on the data and these administrative databases, refer to the Table 1 and the supplementary materials of the original article. [19]

It is known that uncontrolled asthma is associated with adverse effects for the fetus and that advantages of adequate control outweigh any potential risks of asthma medication. [20] Therefore asthma should continue to be controlled with medications during pregnancy. [21] However, it has been shown that about 50% of women tend to lower their controller medication during pregnancy, [22] potentially due to fear of adverse medication effects. This is specifically the case for women with mild asthma, for whom treatment may alternate between low daily doses of Inhaled Corticosteroids (ICS) or no controller medication. [21, 23] The published literature has not iden-

tified adverse effects of low ICS dosage on pregnancy outcomes. However, although asthma control is a time-dependent confounder of the longitudinal exposure to controller medication, [21] there are currently no data on the effectiveness and safety of asthma medications from studies considering the time-dependent confounded nature of asthma control. The proposed analysis is therefore aimed at fitting an MSM, estimated with pooled LTMLE, to evaluate the short-term relative effect of low ICS dose versus no ICS dose on time to delivery in women with mild asthma.

Since it is known that the relationship between asthma control and treatment happens at a finer scale than trimester, [21] interest lies in performing a time-dependent extraction of the administrative health data. While a finer scale would supposedly ensure the best possible control for time-dependent confounding, limitations exist because of the nature of the methods used to control for said time-dependent confounding. We are then faced with the dilemma of how to choose an appropriate discretization given the available data.

3 Data structure and causal parameter of interest

In the above described example, consider a longitudinal data structure where, for every individual, we observe the following data

$$\boldsymbol{O} = (\boldsymbol{L}(0),\ A(0),\ Y(1),\ \boldsymbol{L}(1),\ A(1),\ \ldots,\ Y(K),\ \boldsymbol{L}(K),\ A(K),\ Y(K+1))\ .$$

Let t index the discrete times at which time-dependent variables are observed, $t = 0, \ldots, K+1$. Let $Y(t)$ denote a time-dependent outcome, $A(t)$ a time-dependent exposure, and $\boldsymbol{L}(t)$ a time-dependent vector of covariates, for arbitrary time-point t, measured in that order. In particular, let $\boldsymbol{L}(0)$ be the baseline covariates measured at the beginning of the study, and $Y(K+1)$ be the final outcome assessed at the end of the study, and assume that all individuals are outcome-free at study entry (i.e. $Y(0) \equiv 0$). We consider a binary exposure where $A(t) = 1$ indicates that a person was exposed at time-point t. Let the overbar represent a variable's history, such that, for example, $\bar{A}(t) = (A(0), A(1), \ldots, A(t))$ denotes an individual's exposure history up until time-point t. Examples of fixed regimes (i.e. possible values

at which $\bar{A}(t)$ may be set) include "always exposed", $\bar{a}(t) = (1, 1, \ldots, 1)$, a regime where the subject is exposed to treatment at every time-point, and "never exposed", $\bar{a}(t) = (0, 0, \ldots, 0)$, where the subject is unexposed to treatment at every time-point.

3.1 Discretization

Formally, define the timeline $\mathcal{I} = [0, \tau]$ on the real line, where τ corresponds with the maximum follow-up period of a longitudinal study at which the last outcome is assessed. A partition $\mathcal{P}$ of $\mathcal{I}$ is a finite collection of points $t_0, t_1, \ldots, t_K, t_{K+1}$ such that the union formed by the disjoint intervals $J_k = [t_k, t_{k+1}[$ is $\mathcal{I}$, where the t_k can be ordered such that $0 = t_0 < t_1 < \ldots < t_K < t_{K+1} = \tau$. It follows that each different partition $\mathcal{P}_r$, $r = r_0, r_1, \ldots, r_R$, of our timeline corresponds to a different discretized dataset O_r, where T_r is the set of time-points in $\mathcal{P}_r$ and R is the number of possible different discretizations.

In particular, let us denote $\mathcal{P}_{r_0}$ as the finest partition into which our timeline can be divided in practice, corresponding to the dataset O_{r_0} and time-points T_{r_0}. This finest discretized data structure is equal to the finest possible scale at which changes can be observed. In health administrative databases, this finest partition may consist of the set of all days during the follow-up period. Any other partition $\mathcal{P}_r$ that does not include all of these points is said to be coarser than $\mathcal{P}_{r_0}$. Examples of coarser partitions could be a set of time-points where the intervals between time-points are of 1 week. Inversely, a discretization formed by partition $\mathcal{P}_{r_1}$ is finer than another discretization $\mathcal{P}_{r_2}$ if it includes all points in $\mathcal{P}_{r_2}$ and at least one other point in $\mathcal{I}$. It follows that there are no possible refinements of $\mathcal{P}_{r_0}$. Note that a partition need not only include points that form equally spaced intervals. Figure 1 illustrates three nested discretized timelines, representing respectively a timeline where we consider the finest discretized data, a timeline where we only consider every other time-point, and a timeline where we consider every fourth time-point.

Further, let $\bar{a}_r(t)$ define an exposure regime up to time t on the discretized data O_r, with $\bar{a}_r(\tau) \equiv \bar{a}_r$ of same length as T_r. Correspondingly, $\bar{\mathcal{A}}_r$ can be defined as the set of all possible regimes of interest on T_r. From here on, with some abuse of notation, let us renumber the time-points T_r in any given

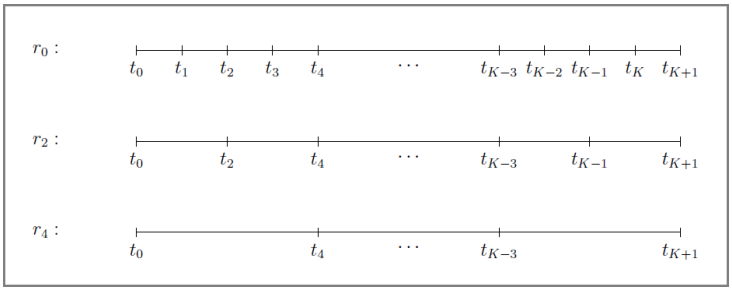

Figure 1: Illustration of the finest discretized timeline, and two coarser discretizations of the finest timeline

partition $\mathcal{P}_r$ as $T_r = \{0, 1, 2, \ldots, K_r, K_r + 1\}$.

3.2 Discretization in practice

Any observed dataset, O_r, can then be defined from the finest partitioned dataset, O_{r_0}, as a function of its partition: $O_r := \Theta(\mathcal{P}_r)(O_{r_0})$. Note that this mapping may not be unique since in practice various procedures may be employed to discretize the finest data, thus resulting in various potential discretized datasets for the analysis. For example, to create O_{r_2} corresponding to discretization r_2 in Figure 1, one might choose a mapping that completely omits information at removed time-points. Hence, for binary exposures say, $\bar{a}_{r_0}(t_2) = (0, 1, 0) \Rightarrow \bar{a}_{r_2}(t_2) = (0, 0)$. Another mapping could aim to summarize the information of the omitted time-points, for example $A_{r_2}(t_2) = I\left(A_{r_0}(t_2) = 1 \mid A_{r_0}(t_1) = 1\right)$ such that $\bar{a}_{r_0}(t_2) = (0, 1, 0) \Rightarrow \bar{a}_{r_2}(t_2) = (0, 1)$. In doing so, one may be altering the initial definitions of the variables and may be violating some causal assumptions, which we discuss in Section 5.2.

3.3 Parameter of interest

Following the Neyman-Rubin counterfactual framework, [24, 25] and for a given discretization indexed by r, let $Y^{\bar{a}_r}(t)$ be a random variable representing a subject's counterfactual outcome at time t had they followed the exposure history $(\bar{A}(t-1), t \in T_r) = \bar{a}_r(t-1)$. The hypothetical intervention represented by $\bar{a}_r(t-1)$ corresponds to exposures that are set at times T_r and sustained throughout the intervals between observed discretized time-points. In particular, discretized regimes may only contain treatment changes at

time-points in the discretization. Note that variables indexed by negative values should be taken as the null set. $E[Y^{\bar{a}_r}(t)]$ is the mean outcome had the exposure history been set to a specific fixed regime $\bar{a}_r(t-1)$ for every subject in the population of interest. In such longitudinal settings, an MSM may be used to describe the expectation of the time-dependent counterfactual outcome as a function, $m_j(\bar{a}, t, \boldsymbol{W}), j = 1, \ldots, J$, of exposure history, time, and possibly a subset of baseline characteristics, $\boldsymbol{W} \subseteq \boldsymbol{L}(0)$ [7, 8, 13]:

$$f\left(E[Y^{\bar{a}_r}(t) \mid t, \boldsymbol{W}]\right) = \eta(\boldsymbol{\beta}, \bar{a}_r, t, \boldsymbol{W}) = \sum_{j=1}^{J} \beta_j m_j(\bar{a}_r, t, \boldsymbol{W}), t \in T_r.$$

As in Petersen *et al.* (2014), [13] and corresponding to our motivating example, we may be interested in modelling the counterfactual survival probability using the variable $Y(t)$, a binary indicator of failure by time t. In this case, one could define a working MSM, for example evaluating the effect of most recent exposure on the counterfactual probability of failure by time t:

$$\eta(\boldsymbol{\beta}, \bar{a}_r, t) \equiv logit[P(Y^{\bar{a}_r}(t) = 1)] = \beta_0 + \beta_1 a_r(t-1) + \beta_2 t, \, t \in T_r \; . \qquad (1)$$

Accordingly, our target parameter, $\boldsymbol{\psi}_r$, the coefficients in the model, could be defined using the logistic log-likelihood, given the hypothetical experiment where exposure is set to some fixed regime at time-points in T_r:

$$\boldsymbol{\psi}_r = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \, E \sum_{t \in T_r} \sum_{\bar{a}_r \in \bar{\mathcal{A}}_r} \{Y^{\bar{a}_r}(t) log(expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t))) + (1 - Y^{\bar{a}_r}(t)) log(1 - expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t)))\}. \quad (2)$$

Note that the above MSM is defined on a fixed (arbitrary) discretization. By changing the discretization, we impose a different target causal parameter since different discretizations would impose summing over different sets of regimes on the corresponding counterfactual data, i.e. only those that can change at the set of time-points considered. We show this numerically in the simulation study in section 7.

The parameter of interest is defined on the resulting counterfactual data $\boldsymbol{O}_r^{\bar{a}_r} = (Y^{\bar{a}_r}(t), \boldsymbol{L}^{\bar{a}_r}(t); t \in T_r)$, where $\boldsymbol{L}^{\bar{a}_r}(t)$ is the counterfactual covariate value at time t had past exposures been set to $\bar{a}_r(t-1)$. For any given r, $\boldsymbol{O}_r^{\bar{a}_r}$ consist of n independent, identically distributed observations from a true

underlying distribution $Q_0(\boldsymbol{O}_r^{\bar{a}_r})$, which can be decomposed according to the time-dependent distribution of the data as:

$$Q_0(\boldsymbol{O}_r^{\bar{a}_r}) = \underbrace{\prod_{t \in T_r \setminus \{0\}} P_0(Y^{\bar{a}_r}(t) \mid \bar{\boldsymbol{L}}^{\bar{a}_r}(t), \bar{Y}^{\bar{a}_r}(t-1))}_{Q_{0Y}(\boldsymbol{O}_r^{\bar{a}_r})} \underbrace{\prod_{t \in T_r \setminus \{K_r+1\}} P_0(\boldsymbol{L}^{\bar{a}_r}(t) \mid \bar{Y}^{\bar{a}_r}(t), \bar{\boldsymbol{L}}^{\bar{a}_r}(t-1))}_{Q_{0\boldsymbol{L}}(\boldsymbol{O}_r^{\bar{a}_r})}$$

Here we suppose that $Q_0(\boldsymbol{O}_r^{\bar{a}_r})$ is a member of a statistical model space $\mathcal{M}$ that can be decomposed into $\mathcal{Q}_Y$ and $\mathcal{Q}_{\boldsymbol{L}}$, the set of all possible values of $Q_{0Y}(\boldsymbol{O}_r^{\bar{a}_r})$ and $Q_{0\boldsymbol{L}}(\boldsymbol{O}_r^{\bar{a}_r})$, respectively. For every coarsened dataset, our causal parameter of interest is thus defined as a mapping $\boldsymbol{\psi}_r \equiv \boldsymbol{\Psi}(Q_0(\boldsymbol{O}_r^{\bar{a}_r})) : \mathcal{M} \to \mathbb{R}^p$ for some function $\boldsymbol{\Psi}$ that takes as argument a member from the statistical model space $\mathcal{M}$ into the parameter space $\mathbb{R}^p$. As a result, we see that the true value of the parameter being targeted, $\boldsymbol{\psi}_r$, directly depends on the data discretization. Consequently, interpretation of the target parameter relies on the chosen discretization.

4 Causal assumptions

To obtain consistent estimates of causal effects from longitudinal observational data one must assume some identifiability conditions. We review the relevant assumptions in order to examine how they are affected by discretization.

4.1 Causal assumptions for longitudinal data

The time-ordering assumption states that $\boldsymbol{L}(t-1)$ precedes $A(t-1)$, which precedes $Y(t)$, for any time-point on an arbitrary discretization indexed as $t = 0, 1, \ldots, K+1$. Another necessary assumption is the so called no unmeasured confounders assumption, [26] formally, the (weak) sequential randomization assumption (SRA) [27]:

$$Y^{\bar{a}}(t) \coprod A(t-1) \mid \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1), \bar{A}(t-2), t = 1, ..., K+1.$$

This assumption can be thought of as in a sequential randomized trial, where at each follow-up time t, exposure is randomly assigned conditional on the observed history. Here, it is assumed that the measured $\bar{\boldsymbol{L}}(t-1)$ is a sufficient

set of confounders such that the SRA holds. We further assume positivity [17, 26] which requires that:

$$P(A(t) = a(t) \mid \bar{\boldsymbol{L}}(t), \bar{Y}(t), \bar{A}(t-1) = \bar{a}(t-1)) > 0, \forall a(t), t = 0, \ldots, K,$$

for every combination of the values of the confounders $\bar{\boldsymbol{L}}(t)$ and exposure history for which $P(\bar{\boldsymbol{L}}(t) = \bar{\boldsymbol{l}}(t), \bar{A}(t-1) = \bar{a}(t-1)) > 0$. Even if the positivity assumption holds in theory, estimated near practical positivity violations may occur when the support in the data for a given regime is not sufficient. Hence, for estimation purposes, the positivity assumption must also hold in practice since, under near practical positivity violations, standard causal methods have been known to produce biased estimates and exhibit substantial variability. [28]

Furthermore, the no interference assumption [29] indicates that an individual's counterfactual is not affected by another individual's exposure. Finally, consistency implies that $Y^{\bar{a}}(t) = Y(t)$ and $\boldsymbol{L}^{\bar{a}}(t) = \boldsymbol{L}(t)$ when $\bar{A}(t-1) = \bar{a}(t-1)$. [25, 30] This assumption also implies that the levels of exposure have to correspond to well-defined interventions. [31] Below we will examine some of the implications of discretization on the plausibility of these causal assumptions.

4.2 Causal assumptions for discretized data

Preservation of time-ordering is strongly dependent on the discretization process. A case in which coarsened data could fail to preserve time-ordering is when, instead of removal of information, the next observed time-point would constitute of a summary measure of the unobserved time-points, similarly to the example in section 3.2. Let us take as illustration a case where the finest discretization constitutes five time-points, $T_{r_0} = 0, 1, 2, 3, 4$, and suppose that $\bar{a}(4) = (0, 1, 0, 0, 0)$ and $\bar{l}(4) = (0, 0, 1, 0, 0)$. By coarsening the data such as to only observe every fourth time-point (i.e. $T_{r_4} = 0, 4$), one could consider that $A_{r_4}(4) = 1$ if exposure occurred in any of the time-points $t = 1, 2, 3, 4$, similarly for $L_{r_4}(4)$. In this specific example, we would thus obtain $A_{r_4}(4) = 1$ and $L_{r_4}(4) = 1$. However, in the finest discretization, exposure happened prior to the covariate event, although this is not represented in the coarsened data. Thus correct time-ordering at the finest

discretization does not imply preserved time-ordering in the coarsened data. One commonly adopted approach to maintain time-ordering in this scenario is to use lagged values of confounders. For example, for the discretized data in the above example, confounder information at time-point $t = 4$, $L_{r_4}(4)$, would only consist of the information contained in $L_{r_4}(0)$. It is thus clear that time-ordering is preserved since $L_{r_4}(0) = L_{r_0}(0)$ occurred prior to the information summarized in $A_{r_4}(4)$. Note that this may, however, lead to problems regarding the SRA assumption since exposure is in truth informed based on confounder information just prior to exposure but this information is now omitted.

In general, excessive discretization may fail to capture sufficient relations between the variables to remove time-dependent confounding bias. Thus by imposing a discretization, we are potentially wrongfully assuming that the SRA still holds. Suppose we assume that the SRA holds at the finest discretization r_0:

$$Y^{\bar{a}_{r_0}}(t) \coprod A(t-1) \mid \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1), \bar{A}(t-2), t \in T_{r_0}.$$

In addition to the example of lagging covariate values given above, a second example in which the SRA may not hold is illustrated in the Directed Acyclic Graph (DAG) in Figure 2. This DAG depicts the causal relations between exposure $A(t)$, covariate $L(t)$, and outcome $Y(t+1)$ at three time-points. One can see that when discretization is carried out by removing information about $A(1)$, $L(1)$, and $Y(1)$, an unmeasured confounder $U(1)$ is created for the relationship between $A(2)$ and $Y(3)$. By failing to adjust for $U(1)$, confounding bias may be introduced when estimating the effect of $A(2)$ on $Y(3)$. A similar loss of information may arise even if the variables at $t = 2$ are redefined as summaries of the information at times $t = 1$ and $t = 2$.

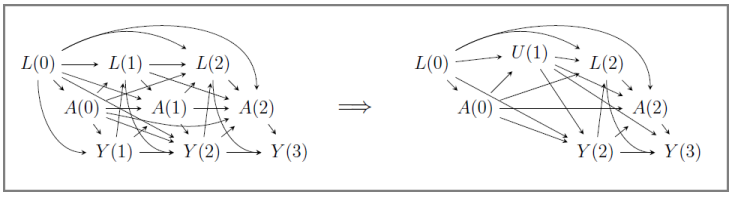

Figure 2: Figure 2. DAG illustrating the result of discretization on time-dependent confounding

In some very specific settings, there may exist some coarser discretization, $r^{\dagger}$, where the SRA still holds. For example, Figure 3 gives a case where the SRA would hold when calculating the effect of most recent exposure, $A(2)$, on subsequent outcome, $Y(3)$, even in the presence of $U(1)$. Note, however, that whether a specific coarsening approach leads to a discretization in which the SRA is preserved is untestable.

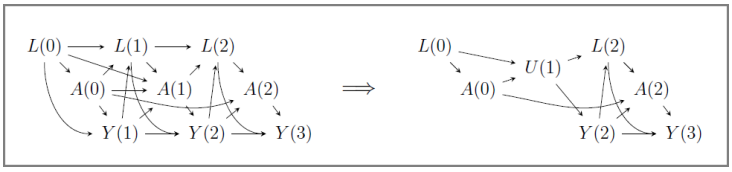

Figure 3: DAG illustrating the result of discretization on time-dependent confounding

The positivity assumption must also hold for every discretization r, at each time-point $t \in T_r$:

$$P(A(t) = a(t) \mid \bar{\boldsymbol{L}}(t), \bar{Y}(t), \bar{A}(t-1)) > 0, \forall a(t), t \in T_r.$$

Because the set of discretized regimes may be a subset of regimes at the finest scale, if positivity holds at the finest scale, then it is plausible that it would hold for any coarser discretization conditional on the true covariate history. Furthermore, in longitudinal contexts with many time-points, causal methods relying on inverse weight calculations may be more at risk of so called near practical positivity violations because they involve the product of the above probabilities over all time-points. [32] When the discretization is overly fine and the number of time-points is large these probabilities may vanish. Coarsening is enticing because coarser discretizations of the timeline may be used as a quick fix since the product would be taken over fewer time-points. Hence, as the discretization gets coarser, the practical positivity assumption may be relaxed by only requiring that the estimated probabilities be larger than zero at fewer observed discretized times.

For consistency, note that the counterfactual intervention is well-defined relative to the chosen discretization, where exposure at selected time-points would be intervened on and sustained between time-points, with changes only allowed at discretized time-points. It would then be important to assess whether it is likely that the resulting counterfactuals corresponding to

this hypothetical experiment are equal to the observed outcomes, for subjects whose summarized observed exposure corresponds to the intervention under the experiment. Whether or not this is the case would rely on how exposure between time-points is summarized, and how well observed between-time-point exposures correspond to the sustained exposure assignment of the hypothetical intervention. In section 1 of the Supplementary Materials, we provide an example where consistency does not hold.

Finally, it is logical to assume that if no interference holds at T_{r_0} it holds for all T_r. If one subject's exposure does not affect another subject's potential outcome under the finest discretization, then it will also not affect the potential outcome under a coarsened discretization.

5 Methods

In this section, we first recapitulate the steps of the pooled LTMLE algorithm to estimate the parameters of an MSM [13] in order to describe the cross-validated variance that can inform the selection of the discretization of the timeline.

5.1 Pooled LTMLE

In order to explain the algorithm, let us define the conditional expected value of the outcome at time t given past covariate history and a fixed exposure history $\bar{a}_r(t-1)$ as $\bar{Q}_t^{\bar{a}_r}(t) = E(Y^{\bar{a}_r}(t) \mid \bar{A}(t-1) = \bar{a}_r(t-1), \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1))$. Recursively, for $j = t, ..., 1$, define $\bar{Q}_t^{\bar{a}_r}(j) = E(\bar{Q}_t^{\bar{a}_r}(j+1) \mid \bar{A}(j-1) = \bar{a}_r(j-1), \bar{\boldsymbol{L}}(j-1), \bar{Y}(j-1))$ where $\bar{Q}_t^{\bar{a}_r}(t+1) := Y^{\bar{a}_r}(t)$. $\bar{Q}_t^{\bar{a}_r}(j)$ is sometimes conveniently referred to as the "outcome regression". Under the described causal assumptions, our causal parameter of interest can be identified with respect to these iterated conditional expectations [13, 17] as

$$\begin{aligned}\boldsymbol{\psi}_r \;=\; \text{argmax}_{\boldsymbol{\beta}}\, E \sum_{t=1}^{K_r+1} \sum_{\bar{a}_r \in \bar{\mathcal{A}}_r} \{\bar{Q}_t^{\bar{a}_r}(1) log(expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t, \boldsymbol{W}))) \;+ \\ (1 - \bar{Q}_t^{\bar{a}_r}(1)) log(1 - expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t, \boldsymbol{W})))\}.\end{aligned}$$

We can construct a simple plug-in estimator that fits a regression of the estimate of $\bar{Q}_t^{\bar{a}_r}(1)$ on $\bar{a}_r, t,$ and $\boldsymbol{W}$ according to our MSM.

Likewise, let us define for every $j = t, \ldots, 1$, $\bar{g}^{\bar{a}_r}(j) = \prod_{k=0}^{j} P(A(k) = a(k) \mid \bar{\boldsymbol{L}}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}_r(k-1))$ which can be described as the "propensity scores" and represent the probabilities of receiving exposure $\bar{A}(j) = \bar{a}_r(j)$ given covariate history $\bar{\boldsymbol{L}}(j)$ and a fixed exposure history $\bar{A}(j-1) = \bar{a}_r(j-1)$.

For each time t, implementation of the pooled LTMLE, as developed in Petersen *et al.* (2014), [13] involves the calculation of an initial estimate $\bar{Q}^{\bar{a}_r}_{t,n}(j)$ and the update of this initial estimate to $\bar{Q}^{\bar{a}_r,*}_{t,n}(j)$ along a path that uses information from $\bar{g}^{\bar{a}_r}_n(j)$. Here the subscript n indicates estimated quantities. A submodel that defines the update path is described in section 2 of the Supplementary Materials, along with the specific logistic regression used to fit this submodel. This update is recursively carried out for $j = t, \ldots, 1$ to obtain $\bar{Q}^{\bar{a}_r,*}_{t,n}(1)$. This is then repeated for each t, after which we can compute our targeted substitution estimator $\boldsymbol{\psi}_{r,n}$ using a pooled logistic regression over all $\bar{Q}^{\bar{a}_r,*}_{t,n}(1), t = 1, \ldots, K_r + 1$. The specific algorithm, as given in Petersen *et al.* (2014), [13] is reproduced in detail for our setting in the Supplementary Materials.

The variance of $\boldsymbol{\psi}_{r,n}$ can be approximated using a sandwich estimator based on the efficient influence curve (EIC). [13, 17] Note that the EIC is a function of the inverse of the exposure probabilities which can take large values in practice. As a consequence, under data sparsity, the above algorithm may lead to unstable inference. In particular, since the generalized score of the submodel used for the updating step spans the EIC, [33] failure to solve the score equation will then result in non-convergence of the algorithm. To alleviate this problem, it is known that using the inverse of the exposure probabilities as weights in the logistic regression model used for the update improves the performance of the estimator in the face of data sparsity problems. [34, 35] If this approach also fails, implementation of pooled LTMLE in the *ltmleMSM* function of the *ltmle* R package version 1.1-0 [36, 37] proceeds without updating $\bar{Q}^{\bar{a}_r}_{t,n}(j)$, i.e. $\bar{Q}^{\bar{a}_r,*}_{t,n}(j) \coloneqq \bar{Q}^{\bar{a}_r}_{t,n}(j)$. We view failure to converge as a serious warning for potential lack of support in the data. Thus, we consider this convergence issue as a criterion for the selection of candidate discretizations for analysis.

5.2 Discretization selection

Ideally, the selected discretization, r^*, would allow to adjust sufficiently for time-dependent confounding while avoiding inflated variance caused by excessive refining. Hence, to more efficiently discriminate between discretizations leading to excessively high-variance estimates, our selection approach incorporates the cross-validated variance estimate of the target parameter, $\widehat{\text{Var}}(\boldsymbol{\psi_{r,n}})$. Note that this term is asymptotically small, and thus coherent as a measure of adequate data support. Below, we outline steps to identify potential discretizations as close as possible to the finest discretization, with sufficient data support.

Step 1: Create candidate discretizations

Beforehand, the candidate discretizations that we want to compare must be chosen, ideally including one at a sufficiently fine scale. Other choices could be motivated by convention or a priori knowledge of scales that may sufficiently allow to control for time-dependent confounding, although, whether the SRA is satisfied for any particular discretization remains an untestable assumption.

One should then create a table of prevalent and incident exposures at each time-point. Results from this table can be used as a tool to investigate which discretizations offer sufficient support for every exposure regime of interest, and thus determine if a given discretization should even be considered as a candidate. The table may also include numbers of uncensored subjects at each time point. An example of such a table and discussion of its usefulness in assessing data support can be found in the application results in section 8.1.

Step 2: Cross-validation [14]

For each candidate discretization r, separate each discretized dataset into V folds of size $\frac{n}{V}$. Let each of these folds be indicated by $v = 1, \ldots, V$. Let the observations in a specific fold v constitute the validation sample and let the training sample comprise of the

remaining $V-1$ folds. Let $P^0_{0,v}$ and $P^1_{0,v}$ represent the true probability distributions of the training and validation samples, respectively, and let $P^0_{n,v}$ and $P^1_{n,v}$ represent estimations of $P^0_{0,v}$ and $P^1_{0,v}$, respectively.

For each fold $v = 1, \ldots, V$, repeat steps 3 and 4:

Step 3: Pooled LTMLE

Fit the pooled LTMLE algorithm using the training sample, i.e. estimators $P^0_{n,v}$ of $P^0_{0,v}$ are built in the training sample.

Discard any candidate discretizations for which the pooled LTMLE algorithm failed to converge.

Step 4: Evaluation of $\widehat{\mathrm{Var}}(\boldsymbol{\psi_{r,n}}(P^0_{n,v}))$ among remaining discretizations

Define $\bar{Q}^{*1}_{n,v}(P^0_{n,v})$ as the pooled LTMLE updated values evaluated on the validation sample using the estimations $P^0_{n,v}$ built with the training sample in Step 3. The subscript 1 indicates evaluation on the validation sample. Specifically, using the remaining sample v, calculate estimates of $\bar{Q}^{\bar{a}_r,*}_{t,n}(1), t = 1, \ldots, K_r + 1$ from the estimations $P^0_{n,v}$. With the validation sample v, evaluate $\widehat{\mathrm{Var}}(\boldsymbol{\psi_{r,n}}(P^0_{n,v}))$. The cross-validated variance is the average of $\widehat{\mathrm{Var}}(\boldsymbol{\psi_{r,n}}(P^0_{n,v}))$ across samples ν.

We propose that the analyst select the finest discretization that has a cross-validated variance "comparable" to the lowest cross-validated variance, in the set of remaining candidate discretizations for which the algorithm has converged. While not a strict decision rule, this procedure provides an analyst with three tools to diagnose estimation under various discretizations. Importantly, when making their selection, the analyst should be blind to the estimates and confidence intervals to avoid p-value hacking.

The resulting estimated parameter $\boldsymbol{\psi}_{r^*,n}$ is obtained by using pooled LTMLE on the selected discretized dataset $\boldsymbol{O}_{r^*}$, for the selected discretization r^*.

We empirically evaluate the usage of the cross-validated variance for the selection of a discretization through a simulation study in the next section. Sample code of this procedure for a two time-point example is provided in the following GitHub repository: `https://github.com/steveferreiraguerra/PLTMLE`.

6 Simulation study

The first aim of this simulation study is to assess the impact of discretization on the estimation of a causal quantity of interest using pooled LTMLE. It further aims to evaluate the performance of the cross-validated variance criterion on selection of a discretization.

The simulated data consist of n *i.i.d.* observations with structure $\boldsymbol{O} = (L(0), A(0), Y(1), L(1), A(1), Y(2), \ldots, L(11), A(11), Y(12))$. These data are aimed to mimic the data in the motivating example consisting of a study where individuals remain exposed after first exposure. Hence, defining all variables as binary, let $A(t) = 1$ indicate whether a person was exposed by time t, $L(t) = 1$ indicate whether a person had covariate L at time t, with $L(0)$ representing a sole baseline covariate, and $Y(t) = 1$ indicate the occurrence of the outcome by time t. The exposure and outcome processes are monotone (i.e. if $Y(t-1) = 1$ then $Y(t) = 1$, and similarly for $A(t)$) and all subjects were outcome-free at study entry (i.e. $Y(0) \equiv 0$). The data is generated at the finest scale, at fixed equally-spaced intervals, conditional on the past three time-points, $t-1, t-2, t-3$. The full data generating process is described in greater detail in the Supplementary Materials.

Candidate discretizations were created by sequentially removing a single time-point from the finest generated data. The candidate discretized datasets consist of $\mathcal{O} = \{\boldsymbol{O}_{r_0}, \boldsymbol{O}_{r_1}, \boldsymbol{O}_{r_2}, \boldsymbol{O}_{r_3}, \boldsymbol{O}_{r_4}, \boldsymbol{O}_{r_5}, \boldsymbol{O}_{r_6}, \boldsymbol{O}_{r_7}, \boldsymbol{O}_{r_8}, \boldsymbol{O}_{r_9}, \boldsymbol{O}_{r_{10}}\}$. T_r indicates which time-points are included in the discretization, such that $T_{r_0} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, $T_{r_1} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12\}$, $T_{r_2} =$

$\{0,1,2,4,5,6,7,8,9,10,12\}$, $T_{r_3} = \{0,1,2,4,5,6,8,9,10,12\}$, $T_{r_4} = \{0,1,2,4,5,6,8,10,12\}$, $T_{r_5} = \{0,2,4,5,6,8,10,12\}$, $T_{r_6} = \{0,2,4,6,8,10,12\}$, $T_{r_7} = \{0,2,4,6,8,12\}$, $T_{r_8} = \{0,4,6,8,12\}$, $T_{r_9} = \{0,4,8,12\}$, $T_{r_{10}} = \{0,4,12\}$. We adopted a simplistic discretization method that consists of omitting the observed information of unselected time-points, as if it had never been observed. For example, $\boldsymbol{O}_{r_9} = (L(0), A(0), Y(4), L(4), A(4), Y(8), L(8), A(8), Y(12))$, where $(Y(t), L(t), A(t)) \in \boldsymbol{O}_{r_9}$ are identical to $(Y(t), L(t), A(t)) \in \boldsymbol{O}_{r_0}$ for $t \in \{0,4,8,12\}$.

The target parameter is defined as in equation (2). Here the goal is therefore to analyze the effect of most-recent exposure on the probability of event by time t across different discretizations. This choice is motivated by the fact that the definition of most-recent exposure is consistent across discretizations and can consistently be interpreted as the effect of exposure at the previous observed time-point on current outcome. The true value of the parameter of interest for the finest discretization was assessed numerically and is equal to -0.442 up to the third decimal. True values for every other discretization were also attained numerically and are presented in Table 1. In the simulation, the target parameter is estimated through pooled LTMLE. Pooled LTMLE code for the parameters of an MSM is currently available in R software [38] as the *ltmleMSM* function of the *ltmle* [36, 37] package version 1.1-0, developed in Petersen *et al.* (2014). [13] We constructed pooled LTMLE code specific to this estimation problem, in which we also implemented a similar numerical approach to solve for $\boldsymbol{\epsilon}$ in the updating step if the approach using a regression model did not converge . This was done in order to reduce the computational complexity of the more general function *ltmleMSM*, to adequately capture failures in convergence, and to more efficiently perform the cross-validation steps. The point estimates were found to be nearly identical between both codes with sample datasets.

500 sets of simulated data were analyzed. In Table 1, for each discretization, the mean estimates and Monte Carlo (MC) variance are reported for the corresponding pooled LTMLE estimate. As mentioned previously, due to the nature of the generated data, which may present severe practical positivity violations at certain levels of discretization, the pooled LTMLE algorithm may not converge. In such situations, we report an estimated value NA. We therefore report mean estimates and MC variance from non-missing estimates only. It is also often recommended to respond to such practical positivity

violations using ad-hoc methods such as truncation, [28] a common approach to reducing the variance, but which may introduce bias due to misspecification of the treatment model. We therefore apply truncation at the 5th and 95th percentiles of the cumulative weights at each time-point. Additionally in Table 1, we present the percentage of times each discretization had the smallest cross-validated variance. In order to assess how the support in the data affects the discretization choice, data with sample sizes $n = 500$, 1000, 2500 were analyzed.

6.1 Simulation results

The results for each discretization and for the selection procedure are presented in Table 1, which is separated in three sections: one for each sample size. Each line corresponds with the results for a candidate discretization. The true values of the target parameter of interest for every discretization are displayed in the second column. It can be observed that the true value changes according to the discretization of the data, with the true value at the finest discretization (-0.442) being far from the true value at the coarsest discretization (-0.651). Note that convergence of true values to the finest parameter is non-monotonic, meaning that there is no guarantee that we approach the true value at the finest discretization as we refine a given coarser discretization.

The next two columns contain the mean pooled LTMLE estimates and MC variance for every discretization and for the selected discretization. We can see from these results that at coarser discretizations the mean estimate was biased for the corresponding true value of the parameter of interest and that, at finer discretizations, the mean estimates were roughly unbiased for the true value. For example, with a sample size of $n = 500$, the mean estimate at the coarser discretization, r_{10}, was equal to -0.810 and the true value equalled -0.651. This indicates that coarser discretizations may not have fully controlled for time-dependent confounding resulting in biased estimates. It is of interest to note that this bias is non-monotonic, and that it is thus possible that we attain local minima which are less biased than at certain finer discretizations. Note also that the Monte Carlo variance decreased with finer discretization. This corresponds with the logic that as the number of time-points increases our estimation becomes more precise since we have more pooled data. However, these measures were only computed

on available estimates since, for finer discretizations and small sample sizes, most of the estimates were returned as NA, with proportions summarized in the last column. This may underestimate the true variability of the estimator under finer discretizations. For example, at $n = 500$, discretizations finer than r_5 produced an estimate for fewer than 26% of simulated sets. Missing estimates were due to large weights and sparse observations that occurred frequently at finer discretizations, even after employing truncation. This was particularly problematic at smaller sample sizes.

Regarding selection based on minimization of cross-validated variance, for which the results are shown in the column % select CV-Var, we first notice that, as the sample size increased, finer discretizations tended to be selected more often. Indeed, the most selected discretizations were r_8, r_6, and r_1 for sample sizes $n = 500, 1000,$ and 2500, respectively. This illustrates that this criterion behaves as expected and that finer discretizations are preferred as data support increases due to efficiency gained with the pooling of data over more time points.

7 Data application

This section revisits the motivating example of Section 2 on the comparison of the effect of low ICS dose versus no ICS on pregnancy duration. To reflect the underlying nature of the data and to capture all possible changes in covariates, a day-by-day follow-up of asthma medication exposure, asthma related covariates and pregnancy related covariates was longitudinally extracted. The final cohort consists of pregnancies with no gaps in the woman's insurance plan coverage from 1 year before and throughout pregnancy, of women who were less than 45 years of age, had a singleton delivery, and women contributing to a maximum of two pregnancies. The presence of asthma was established based on at least one diagnosis of asthma combined with at least one filled prescription for an asthma medication during the pregnancy or one year prior to pregnancy. Women taking theophylline, cromoglycate, nedocromil, ketotifen, or LABA without an ICS were excluded.

Finally, a subsetted cohort of women with mild asthma in the year prior to pregnancy with no use of ICS during the first trimester was created. Mild asthma was defined using a validated severity indicator developed in our re-

Table 1: Mean estimate, MC variance, percentage minimal cross-validated variance and percentage NA for every discretization for various sample sizes

	Discretization	True value†	Mean Est.*	MC variance*	% min CV-Var	% NA
n = 500	r_{10}	-0.651	-0.811	0.160	9.6	2.4
	r_9	-0.601	-0.666	0.091	32.4	1.2
	r_8	-0.518	-0.527	0.055	39.2	9.2
	r_7	-0.519	-0.516	0.050	7.6	32.0
	r_6	-0.512	-0.500	0.036	8.8	37.6
	r_5	-0.471	-0.422	0.027	0.8	68.8
	r_4	-0.495	-0.462	0.036	0	76.0
	r_3	-0.472	-0.480	0.032	0.4	79.2
	r_2	-0.438	-0.456	0.031	0.4	78.0
	r_1	-0.440	-0.414	0.025	0	84.8
	r_0	-0.442	-0.449	0.026	0	85.2
n = 1000	r_{10}	-0.651	-0.787	0.071	0.4	0.0
	r_9	-0.601	-0.648	0.039	2.8	0.4
	r_8	-0.518	-0.525	0.027	27.2	0.0
	r_7	-0.519	-0.513	0.023	4.4	7.2
	r_6	-0.512	-0.502	0.017	42.4	8.0
	r_5	-0.471	-0.433	0.014	11.2	30.0
	r_4	-0.495	-0.480	0.019	2	42.4
	r_3	-0.472	-0.475	0.016	1.6	40.8
	r_2	-0.438	-0.441	0.014	3.6	41.6
	r_1	-0.440	-0.434	0.011	2.4	48.0
	r_0	-0.442	-0.442	0.010	2	48.8
n = 2500	r_{10}	-0.651	-0.764	0.036	0	0.0
	r_9	-0.601	-0.641	0.019	0	0.0
	r_8	-0.518	-0.518	0.013	0.8	0.0
	r_7	-0.519	-0.513	0.010	0	0.0
	r_6	-0.512	-0.502	0.010	3.6	0.0
	r_5	-0.471	-0.440	0.010	28	0.8
	r_4	-0.495	-0.490	0.010	0	7.2
	r_3	-0.472	-0.481	0.010	0	7.6
	r_2	-0.438	-0.453	< 0.005	11.6	6.4
	r_1	-0.440	-0.437	< 0.005	42.8	6.0
	r_0	-0.442	-0.445	< 0.005	13.2	6.8

† indicates true value for every discretization
* computed using non missing values only

search group. [39] The start of follow-up was established at 20 weeks since no deliveries occurred before week 20 by definition. In this application, candidate discretized datasets to be analyzed consisted of 3-week intervals ($\boldsymbol{O}_{r_3}$), 4-week intervals ($\boldsymbol{O}_{r_4}$), 5-week intervals ($\boldsymbol{O}_{r_5}$), and 6-week intervals ($\boldsymbol{O}_{r_6}$) from start of follow-up. The finest discretized dataset consisting of daily data was not considered as a candidate discretization due to the resulting data being far too voluminous and sparse. Discretized data were created from the finest daily data and defined as $\boldsymbol{O}_r = (Y(t), \boldsymbol{L}(t), \boldsymbol{A}(t))$, $t \in T_r$, where $\boldsymbol{L}(t)$ is a vector of confounder variables measured during $[t-1, t[$, $\boldsymbol{A}(t) = (A_1(t), A_2(t))$ represents a multivariate "treatment" measured at t composed of an ongoing exposure indicator $A_1(t)$ and a censoring indicator $A_2(t)$, where $A_1(t) = 1$ indicates exposure to low daily doses of ICS and $A_1(t) = 0$ indicates no exposure to ICS. Identically, $A_2(t) = 1$ indicates that a subject has been censored by time t, $A_2(t) = 0$ otherwise. Specifically, a subject could be censored if their asthma treatment differed from one of the above defined regimes for mild asthma. For example, a subject could be censored if they begin receiving a higher ICS daily dose or the concomitant usage of LABA with ICS, which both indicate an increase in asthma severity. Finally, $Y(t) = 1$ represents a delivery occurring during $[t-1, t[$. Figure 4 displays the time-ordering of the observed data according to different discretizations.

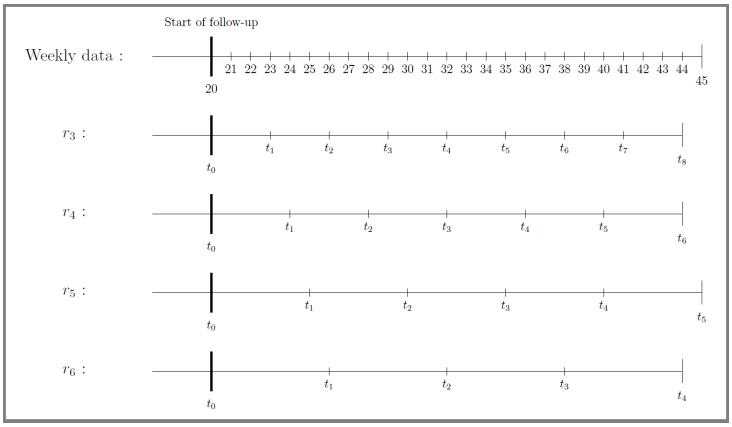

Figure 4: Candidate discretized timelines in the data application

Potential confounders were divided into four categories: characteristics of the mother, chronic maternal diseases, pathologies related to pregnancy, and

maternal asthma control related variables. A complete list of confounders may be found in Table 2. In this application, baseline characteristics were evaluated from the year prior to pregnancy to the start of follow-up at 20 weeks of gestation and are equal regardless of discretization. They are presented in Table 2.

The parameter of interest was defined according to equation (2). In this study, since we defined ICS exposure as monotone (once a woman receives a medication, she is considered to be exposed for the remainder of the pregnancy, unless censored), regimes of interest were defined as initiation of low ICS doses at any time during pregnancy. For instance, initiating at the third time-period is $\bar{a}_r = (0, 0, 1, \ldots, 1)$.

Finally, the usual influence curve based sandwich estimator for the variance of pooled LTMLE is said to be anti-conservative when practical positivity violations occur. [13] Hence, in order to obtain a valid estimate of the variance, we used the original *ltmleMSM* function from the *ltmle* R package which proposes an alternative robust variance estimate. [36, 37] Efficiency in the estimation of the pooled LTMLE requires consistent estimation of both the models for the probabilities of exposure and outcome processes. Therefore, it may be preferred to estimate these quantities using machine learning techniques or Super Learner, [40] an ensemble-learning approach. However, due to computational reasons, we opted not to use such data-adaptive methods for this application. Therefore, simple logistic regressions conditional on all past covariates were used to fit the models in steps 1 and 2 of the pooled LTMLE algorithm. As in common practice, in order to avoid overly large weights, truncation at a level of 5% was applied.

7.1 Application Results

The final cohort of women with mild asthma in the year prior to pregnancy and no ICS use during the first trimester comprised of 2878 pregnancies. The pregnancy duration in this cohort had a mean of 38.5 weeks and a range of 20 to 42 weeks. The baseline characteristics of the pregnancy cohort are given in Table 2 by exposure status at start of follow-up. For gestational hypertension / PECL / ECL, a missing value at baseline is given in the table since these characteristics could only be measured after the 20^{th} week of pregnancy.

The sample was primarily composed of women who had their asthma controlled in the year prior to pregnancy (80.40%). Women with no ICS usage at start of follow-up tended to have their asthma less well controlled in the year prior to pregnancy. The women were also mostly aged between 18 and 34 years at delivery (85.79%), mostly living in urban areas (83.22%), with roughly half receiving social assistance (54.76%). With respect to mother characteristics, chronic maternal diseases, and pathologies related to pregnancy, no great disparities existed between exposure groups. On the other hand, asthma control related variables were dissimilar for women taking low ICS doses versus no ICS doses. Indeed, a higher proportion of women with low ICS doses had at least one hospitalization or emergency room visit for asthma. Identically, these women had also a higher proportion of oral and nasal corticosteroids, and had a higher number of doses per week of short-acting beta2-agonists. Markers for poor asthma control were much higher in the low ICS group due to differential indication since in the latent period between the end of first trimester and cohort entry their asthma may have been poorly controlled, which motivated the clinical decision to change asthma treatment to low ICS doses. This table also demonstrates potential practical positivity problems in the data as evidenced by the few counts in many cells.

Table 3 shows exposed individuals, individuals who changed treatment from no ICS to low doses of ICS, and new censored individuals at each time-point $t \in T_r$ for every candidate discretization. For example, for the discretized dataset $\boldsymbol{O}_{r_3}$, there were 275 exposed women at t_0, the start of follow-up. At t_1 there were now 366 women exposed, 100 of whom changed from no ICS at t_0 to low doses of ICS at t_1. In total, 36 women from either exposure group were censored at t_1. Missing values indicate end of exposure measurement. For example, for the discretized data $\boldsymbol{O}_{r_6}$, all values were missing after time-point t_3, since only four time-points were used. Generally, the results from this table show that as the discretizations get coarser, the data offers more support for every exposure regime of interest.

In Table 4, Pooled LTMLE estimates of the parameter of interest and corresponding standard errors and 95 % confidence intervals (CI) are reported for every discretization. The corresponding values of the cross-validated variance are also presented for each candidate discretization. The reported values in the second column correspond to estimates of the β_1 parameter in equation

Table 2: Women's baseline characteristics per exposure status - *n(%)*

	Exposure		
	No ICS	Low ICS	Neither (Censored)
	$n =$ 2426	$n =$ 275	$n =$ 177
Characteristics of the mother			
Age at beginning of pregnancy			
< 18	30 (1.24)	8 (2.91)	0 (0.00)
18-34	2095 (86.36)	236 (85.82)	138 (77.97)
> 34	301 (12.41)	31 (11.27)	39 (22.03)
Social assistance benificiary in the year prior to pregnancy	1322 (54.49)	147 (53.45)	107 (60.45)
Rural location of residence at delivery	413 (17.02)	44 (16.00)	26 (14.69)
Chronic maternal diseases			
Chronic hypertension	67 (2.76)	7 (2.55)	3 (1.69)
Diabetes mellitus	77 (3.17)	3 (1.09)	7 (3.95)
Uterine disorders	44 (1.81)	4 (1.45)	1 (0.56)
Other chronic diseases	17 (0.70)	3 (1.09)	1 (0.56)
Pathologies related to pregnancy			
Gestational diabetes	31 (1.28)	9 (3.27)	2 (1.13)
Gestational hypertension / PECL / ECL	-	-	-
Placental complications	20 (0.82)	0 (0.00)	2 (1.13)
Other pregnancy-related pathologies	475 (19.58)	52 (18.91)	29 (16.38)
Maternal asthma control related variables			
Asthma control in the year prior to pregnancy	1991 (82.07)	218 (79.27)	105 (59.32)
SABA (doses/week)			
0	1159 (47.77)	44 (16.00)	20 (11.30)
> 0-3	1267 (52.23)	231 (84.00)	157 (88.70)
> 3	0 (0.00)	0 (0.00)	0 (0.00)
LTRA	30 (1.24)	1 (0.36)	9 (5.08)
OCS	179 (7.38)	55 (20.00)	41 (23.16)
NCS	266 (10.96)	53(19.27)	43 (24.29)
Hospitalization for asthma	20 (0.82)	9 (3.27)	13 (7.34)
Emergency room visit for asthma	207 (8.53)	77 (28.00)	32 (18.08)

1, interpreted as the effect of most recent exposure on delivery. Consequently, $exp(\beta_1)$ corresponds to the odds of delivery for women on low ICS dose versus no ICS. Hence, regardless of discretization, the point estimates can be interpreted as a reduced odds of delivery after switching to low ICS dose at any given time t, which is consistent with clinical hypotheses. For all discretizations, the results were statistically non significant, except for the estimate obtained from the discretized data $\boldsymbol{O}_{r_4}$.

The cross-validated variance was minimized at the discretization r_6. However, the selected discretization for analysis was r_5, since it is the finest discretization with cross-validated variance similar to the minimal cross-validated variance. Hence, we obtain an OR estimate of 0.843 with CI = [0.646,1.10]. In contrast with the simulation results, the data application point estimates did not display clear convergence to a value as the dis-

Table 3: Number of exposures, treatment changes, and censorings at every time-point for each candidate discretization

	Disc.	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
	r_6	275	465	624	619	-	-	-	-	-
Exposure	r_5	275	422	576	684	271	-	-	-	-
(Prevalent)	r_4	275	395	517	627	692	270	-	-	-
	r_3	275	366	460	538	624	682	616	77	-
	r_6	275	200	177	117	-	-	-	-	-
Treatment change	r_5	275	158	167	128	33	-	-	-	-
(Incident)	r_4	275	130	131	116	98	21	-	-	-
	r_3	275	100	98	87	90	75	56	4	-
	r_6	177	45	47	19	-	-	-	-	-
Censoring	r_5	177	43	39	26	6	-	-	-	-
(Incident)	r_4	177	36	33	25	18	4	-	-	-
	r_3	177	31	18	31	17	13	7	0	-

cretization became finer. The fact that the candidate discretizations were not necessarily nested may have contributed to this. The standard error estimates decreased progressively for discretizations r_6, r_5, and r_4. This can be explained since additional data translates into a gain in precision in the pooled model. Yet, at the finest discretization r_3, the standard error estimate increased. This suggests that this discretization led to large weights and standard errors. This is also displayed by the larger value of the cross-validated variance, which was much greater than for other discretizations, providing evidence that the cross-validated variance diagnoses lack of support in the data.

8 Discussion

Only a few methods, and extensions of these, handle a continuous underlying data generating distribution. [4,41–47] While some of these are limited in the

Table 4: Pooled LTMLE estimates with corresponding standard error, 95% CI, and cross-validated variance for every candidate discretization

Discretization	Estimate	Standard error	95% CI	CV-var
r_6	-0.232	0.164	[-0.554, 0.090]	1.412
r_5	-0.171	0.136	[-0.437, 0.095]	1.635
r_4	-0.310	0.133	[-0.570, -0.050]	4.587
r_3	-0.258	0.272	[-0.792, 0.275]	17.228

data structure and estimation problem to which they can be applied, others are not specific to the estimation of MSMs. Apart from these methods, most causal inference methods have relied on arbitrary discretization of the patient timeline in longitudinal studies. We note that arbitrary discretization is analogous to the choice of follow-up time-points in any observational study where treatment may change between follow-ups. Hence, our results may also be relevant when designing an observational study which measures exposure at discrete time-points, although this would not apply to settings where patients with chronic conditions are expected to maintain treatment or only change treatment at pre-specified follow-up visits at regularly spaced intervals. However, this is not the case in our motivating example in mild asthma. We have shown that such arbitrary timeline coarsening may result in bias and affects the value and identification of the underlying parameter of interest one is estimating. Furthermore, we evaluated the usage of the cross-validated variance of the pooled LTMLE to inform selection of the timeline. This procedure is readily adaptable to any MSM and LTMLE specification.

Since certain causal inference methods have been known to be sensitive to practical positivity violations, which may occur much more frequently in longitudinal contexts, one appeal of coarser discretization is that it may limit such violations by decreasing the number of observed time-points . Although pooled LTMLE has shown practical advantages over IPTW methods, notably in contexts with small support for certain regimes of interest, such methods may still be vulnerable to severe practical positivity violations. Such violations occurred in our simulation study when the number of

time-points was large and the sample size small and in the data application, particularly when the number of time-points increased, given the relatively high-dimensional confounder space. To this purpose, our selection approach included the pooled LTMLE cross-validated variance in order to better identify discretizations in which such violations would occur and hence inflate the estimated variance. The variance estimator used is based on the influence curve, which has been found to be substantially anti-conservative in settings where data sparsity due to rare outcomes or practical positivity violations occurs. Implementation of the selection procedure using the robust variance estimate provided in [36,37] should be investigated as it could produce better inference.

In the simulation study, the employed discretization method may not have been the most representative of common approaches which preserve confounder and exposure information rather than discarding it. However, we also showed how summarizing information between time-points may violate the time-ordering assumption. We demonstrated that the true value of our parameter of interest may change across discretizations. This indicates that the interpretation of said parameter should be made with respect to the chosen discretization. Nonetheless, theoretical results about the large-sample properties of the resulting LTMLE on the selected discretization remain to be investigated. It is possible that the variance of the LTMLE – estimated using the efficient influence function on the selected dataset – is underestimated due to the selection of the timeline. One solution may be to use sample splitting to select a discretization and obtain estimates on separate data splits. [48]

In the data application, of the options provided, our approach indicated that we should select 5 week intervals. The results could not conclude that, for women with mild asthma, a protective effect of low ICS treatment versus no ICS treatment on delivery time exists, although point estimates were indicative of such. However, given the absence of adjustment for important confounders such as smoking, body mass index, etc., these estimates remain potentially biased. While we evaluated most recent exposure, these methods may be employed for other exposure measures such as cumulative exposure or the effects of exposure at multiple time-points. Most importantly, the different results across discretizations illustrate how the underlying data discretization may be used to alter the final conclusion, and the need for trans-

parent practice.

In summary, this paper serves as an early investigation of the causal inference problem of data discretization and proposes several investigative methods. Given the widespread usage of arbitrary discretization, more investigation is needed to evaluate related problems. Data-adaptive approaches to data extraction from administrative data may provide statistical advantages and an unambiguous decision making procedure.

9 Acknowledgements

The authors would like to thank Joshua Schwab for the insightful comments and help in the use of pooled LTMLE. MES and SFG would also like to thank Miguel Hernán for his helpful comments on the manuscript. The authors would also like to thank the reviewers for their knowledgeable comments throughout the revision process. MES is supported by a CIHR New Investigator Salary Award and an NSERC Discovery Grant.

References

[1] Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. *Annual Review of Public Health* 2013; 34: 61-75.

[2] Mazzali C, Duca P. Use of administrative data in healthcare research. *Internal and Emergency Medicine* 2015; 10(4): 517–524.

[3] Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects – advantages and disadvantages. *Nature Reviews Rheumatology* 2007; 3(12): 725–732.

[4] Zhang M, Joffe MM, Small DS. Causal inference for continuous-time processes when covariates are observed only at discrete times. *Annals of Statistics* 2011; 39(1): 131–173.

[5] Neugebauer R, Silverberg MJ, Laan v. dMJ. Observational study and individualized antiretroviral therapy initiation rules for reducing cancer incidence in HIV-infected patients. *U.C. Berkeley Division of Biostatistics Working Paper Series* 2010(Working Paper 272).

[6] Grøn R, Gerds TA, Andersen PK. Misspecified poisson regression models for large-scale registry data: inference for 'large n and small p'. *Statistics in Medicine* 2016; 35(7): 1117–1129.

[7] Robins JM. Marginal structural models. In: ; 1997.

[8] Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in Epidemiology. *Epidemiology* 2000; 11(5): 550-560.

[9] Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; 11(5): 561-570.

[10] Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of epidemiology and community health* 2006; 60(7): 578–586.

[11] Sofrygin O, Zhu Z, Schmittdiel JA, et al. Targeted Learning with Daily EHR Data. In: ; 2017: arXiv e-prints: 1705.09874.

[12] Kreif N, Sofrygin O, Schmittdiel J, et al. Evaluation of adaptive treatment strategies in an observational study where time-varying covariates are not monitored systematically. In: ; 2018: arXiv e-prints: 1806.11153.

[13] Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, Laan v. dM. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of causal inference* 2014; 2(2): 147–185.

[14] Van Der Laan MJ, Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. 2003.

[15] Robins JM. 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* 1986; 7(9-12): 1393–1512.

[16] Laan v. dMJ, Gruber S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics* 2012; 8(1).

[17] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; 61: 962-972.

[18] Schnitzer ME, Laan v. dMJ, Moodie EE, Platt RW. Effect of breastfeeding on gastrointestinal infection in infants: a targeted maximum likelihood approach for clustered longitudinal data. *The annals of applied statistics* 2014; 8(2): 703.

[19] Cossette B, Forget A, Beauchesne MF, et al. Impact of maternal use of asthma-controller therapy on perinatal outcomes. *Thorax* 2013: thoraxjnl–2012.

[20] Murphy V, Clifton V, Gibson P. Asthma exacerbations during pregnancy: incidence and association with adverse pregnancy outcomes. *Thorax* 2006; 61(2): 169–176.

[21] GINA . Global Strategy for Asthma Management and Prevention. 2016. URL : http://ginasthma.org/2016-gina-report-global-strategy-for-asthma-management-and-prevention/ (Accessed 20 Sept. 2016).

[22] Blais L, Firoozi F, Kettani FZ, et al. Relationship between changes in inhaled corticosteroid use and markers of uncontrolled asthma during pregnancy. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 2012; 32(3): 202–209.

[23] Busse WW. NAEPP expert panel report: managing asthma during pregnancy: recommendations for pharmacologic treatment - 2004 update. *Journal of Allergy and Clinical Immunology* 2005; 115(1): 34–46.

[24] Splawa-Neyman J, Dabrowska D, Speed T, others . On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* 1990; 5(4): 465–472.

[25] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies.. *Journal of educational Psychology* 1974; 66(5): 688.

[26] Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983; 70(1): 41-55.

[27] Robins JM. Causal inference from complex longitudinal data. In: Springer. 1997 (pp. 69–117).

[28] Petersen ML, Porter KE, Gruber S, Wang Y, Laan v. dMJ. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research* 2012: 0962280210386207.

[29] Cox DR. Planning of experiments.. *New York: John Wiley & Sons* 1958: 308 p.

[30] VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009; 20(6): 880–883.

[31] Rubin DB. Comment on: "Statistics and causal inference" by P. Holland. *Journal of the American Statistical Association* 1983; 81: 961–962.

[32] Laan v. dMJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer Series in StatisticsSpringer . 2011.

[33] Laan v. dMJ, Rubin D. Targeted maximum likelihood learning. *The International Journal of Biostatistics* 2006; 2(1): Article 11.

[34] Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* 2007; 22(4): 523–539.

[35] Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science* 2007; 22(4): 544–559.

[36] Schwab J, Lendle S, Petersen M, Laan v. dM. *LTMLE: longitudinal targeted maximum likelihood estimation*. 2013. R package.

[37] Lendle S, Schwab J, Petersen M, Laan v. dM. ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data. *Journal of Statistical Software* 2017; 81(1): 1–21.

[38] R Development Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2011. ISBN 3-900051-07-0.

[39] Firoozi F, Lemière C, Beauchesne MF, Forget A, Blais L. Development and validation of database indexes of asthma severity and control. *Thorax* 2007; 62(7): 581–587.

[40] Laan V. dMJ, Polley EC, Hubbard AE. Super learner. *Statistical applications in genetics and molecular biology* 2007; 6(1).

[41] Saarela O, Liu ZA. A flexible parametric approach for estimating continuous-time inverse probability of treatment and censoring weights. *Statistics in medicine* 2016; 35(23): 4238–4251.

[42] Lok JJ. Statistical modeling of causal effects in continuous time. *The Annals of Statistics* 2008: 1464–1507.

[43] Røysland K. A martingale approach to continuous-time marginal structural models. *Bernoulli* 2011; 17(3): 895–915.

[44] Yang S, Pieper K, Cools F. Semiparametric estimation of structural failure time model in continuous-time processes. In: ; 2018.

[45] Yang S, Tsiatis AA, Blazing M. Modeling survival distribution as a function of time to treatment discontinuation: A dynamic treatment regime approach. *Biometrics* 2018; 74(3): 900–909.

[46] Hu L, Hogan JW. Causal comparative effectiveness analysis of dynamic continuous-time treatment initiation rules with sparsely measured outcomes and death. *Biometrics* 2019.

[47] Lok JJ. Mimicking counterfactual outcomes to estimate causal effects. *Annals of statistics* 2017; 45(2): 461.

[48] Laan v. dMJ, Luedtke AR, Díaz I. Discussion of Identification, Estimation and Approximation of Risk under Interventions that Depend on the Natural Value of Treatment Using Observational Data, by Jessica Young, Miguel Hernán, and James Robins. *Epidemiologic Methods* 2014; 3(1): 21–31.