The causal inference of longitudinal exposures with marginal structural models: an overview and application of longitudinal TMLE

Mireille Schnitzer, PhD
CIHR Canada Research Chair in Causal Inference and Machine Learning in Health Sciences
Associate Professor of Biostatistics
Faculty of Pharmacy, Université de Montréal, Canada
Topics

• Longitudinal exposures: when and why do standard methods fail?
• Definition of marginal structural models
• Inverse probability weighting (IPW) and TMLE for longitudinal exposures
• Application results: the effect of breastfeeding duration on infant hospitalization in the first year

Application and related publications were joint work with Mark van der Laan, Erica Moodie, and Robert Platt.
Michael Kramer is PI of the PROmotion of Breastfeeding Intervention Trial.
Longitudinal exposures

We are interested in estimating the effects of an exposure
• Can be passive: treatment received, exposure to an environment, etc.
• It can be active: prescription by treating physician, intervention in an experimental design

When planning a study and analysis, we must consider exactly what exposure and effect we are interested in.
• How does this relate to a hypothetical intervention?

In particular, we ask: does the exposure vary over time?
Example in an RCT context

In a randomized controlled trial (RCT), two effects of interest may be:

• The effect of assignment (intent-to-treat)
• The effect of being assigned to and then following the protocol
  • Effect of sustained exposure to the protocol drug
  • Due to non-adherence, observed exposure may change over time

*We may be interested in estimating the same effect-types with observational data*

*Hernán and Robins, Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available (2016), American Journal of Epidemiology, 183(8): 758-764*
Limitations of “standard”
conditional methods

Simple example:
• Single binary *randomized* treatment, A= 0 or 1,
• Patient covariates, X, measured, related to the outcome,
• Binary end-of-study outcome, Y.

• A logistic regression will estimate a marginal odds ratio, \( \exp(\beta_1) \):
  \[
  \text{logit } E(Y|A) = \beta_0 + \beta_1 A
  \]

• A *conditional* logistic regression will estimate a conditional odds ratio \( \exp(\beta_1^*) \):
  \[
  \text{logit } E(Y|A,X) = \beta_0^* + \beta_1^* A + \beta_2 X
  \]
Why does this matter*?

Even in the RCT,
• $\exp(\beta_1) \neq \exp(\beta_1^*)$ due to non-collapsibility of the OR.
• Different sets of covariates $X$ will result in different effects being estimated.
• The conditional effect $\exp(\beta_1^*)$ has a definition that relies on the regression model specified.

In an observational study, similar issues arise.
• The conditional parameter depends on variable selection and on the parametric model used for estimation.
• Gives rise to incoherence in analysis target.

*Read a counterpoint:
Evaluating the efficacy of antiretroviral medications in patients with AIDS

- In the 1980s, short-term RCTs evaluated the efficacy of emerging antiretroviral medications for the treatment of AIDS.
- First trial on azidothymidine (AZT)\(^1\) had a maximum duration of 24 weeks.
- Outcomes included mortality and opportunistic infections.

- Patients from many such studies were followed over the longer-term in observational studies.
- A major biostatistical challenge emerged: how to evaluate time-varying exposures to changing drug regimens?

Evaluating the efficacy of antiretroviral medications in patients with AIDS

Table 1. Epidemiological biases in observational studies and analytical resolution. All of these biases occur simultaneously; depending on the actual study design and analysis, certain biases can be more relevant than others.

<table>
<thead>
<tr>
<th>Bias</th>
<th>Potential artifact</th>
<th>Resolution*a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longer survival increases the chance to begin therapy</td>
<td>Can make therapy appear better than it really is</td>
<td>Therapy use must be modelled as a time-dependent covariate or with intention to treat</td>
</tr>
<tr>
<td>Sicker individuals initiate therapy sooner than healthy individuals</td>
<td>Can make therapy appear harmful</td>
<td>Incorporate covariates that measure patients’ health status (i.e. CD4+ counts, haemoglobin, AIDS status) into the analysis</td>
</tr>
<tr>
<td>Individuals with greatest access to health care are more likely to use drug</td>
<td>Can make therapy appear better than it really is</td>
<td>Incorporate covariates measuring access to healthcare system (i.e. income, insurance) into the analysis</td>
</tr>
<tr>
<td>Individuals who use one AIDS drug are more likely to use other AIDS drugs</td>
<td>Can falsely attribute benefits from the second AIDS drugs (i.e. Pneumocystis carinii prophylaxis) to the drug of interest (i.e. zidovudine)</td>
<td>All AIDS therapies must be included (as time-dependent covariates) into the analysis</td>
</tr>
</tbody>
</table>

*a These techniques will mitigate, not eliminate, the respective biases.

Table 1 of Hoover (1995) in Drugs, describing the difficulties in the analysis of longitudinal data for evaluating long-term zidovudine in HIV patients.
Evaluating the efficacy of antiretroviral medications in patients with AIDS

- CD4+ indicates disease progression
- CD4+ was used to determine initiation/application of treatment
Evaluating the efficacy of antiretroviral medications in patients with AIDS

- CD4+ indicates disease progression
- CD4+ was used to determine initiation/application of treatment
Evaluating the efficacy of antiretroviral medications in patients with AIDS

- CD4+ indicates disease progression
- CD4+ was used to determine initiation/application of treatment
Conditional vs causal methods

Conditional methods do not estimate a coherent effect in this context.

- Context: time-dependent confounders that are also mediators of the exposure effect.
- Examples of “conditional” methods: mixed models, Cox proportional hazards models with time-dependent covariates

Developments by Dr. James Robins to define and estimate a coherent effect:

- Structural Nested Models
- Marginal Structural Models (MSM)
Marginal structural models

Marginal structural models are models for counterfactual outcomes.

Single time-point setting \((X, A, Y)\)
- \(Y(1)\) is the outcome that \textcolor{red}{\text{would have}} occurred if we set \(A=1\)
- \(Y(0)\) is the outcome that \textcolor{red}{\text{would have}} occurred if we set \(A=0\)

- Model e.g.: 
  \[
  \text{logit } E(Y(a)) = \beta_0 + \beta_1 a
  \]
- \(\exp(\beta_1)\) has the same marginal OR interpretation as in the RCT.
Marginal structural models

Marginal structural models are models for counterfactual outcomes.

Multi time-point setting, e.g. \((X_0, A_0, X_1, A_1, Y)\)

- \(Y(0,0)\) is the outcome that \textit{would have} occurred if we set \(A_0 = A_1 = 0\).
- \(Y(1,1)\) is the outcome that would have occurred if we set \(A_0 = A_1 = 1\).

- Model e.g.: \(\text{logit } E(Y(a_0, a_1)) = \beta_0 + \beta_1 a_0 + \beta_2 a_1\)
- \(\exp(\beta_1)\) and \(\exp(\beta_2)\) have the same marginal OR interpretation as in a sequentially RCT.
MSM analytical target in the AIDS example

• Idea: we imagine a hypothetical longitudinal intervention that sets A to 0 or 1 at every time point.
• In this world, there is no confounding -> don’t adjust for CD4+.
• Adjustment for CD4+ would block part of the effect pathway.
Saturated MSMs

The simplest type of MSMs are “saturated” MSMs.

- We estimate $E(Y(a_0, a_1))$ directly for a fixed $(a_0, a_1)$.
- Interpretation: The expected outcome (or risk of outcome) if we set $A_0=a_0$ and $A_1=a_1$.

- Can choose only treatment regimens corresponding with guidelines (e.g. if $A_0=1$ then necessarily $A_1=1$).

- Can also fit saturated/unsaturated MSMs under guidelines that depend on covariates. E.g. $A_t=1$ if $\{\text{CD4}_t<500\text{cells/mm}^3 \text{ or if } A_{t-1}=1\}$. 
Example: breastfeeding duration and infant hospitalizations

The PROmotion of Breastfeeding Intervention Trial (PROBIT) (Kramer et al 2001)

- Maternal hospitals and polyclinics were randomized to receive lactation management training (n=31 completed).
- Pregnant women who intended to breastfeed; only healthy full-term, singleton infants were included.

- All women initiated breastfeeding.
- Breastfeeding durations varied.
- Questions of interest: did the duration of breastfeeding impact gastrointestinal infections or hospitalizations in the first year?
Data used for exposure

2.2 Is the baby still breastfeeding? □ yes □ no
2.2.1 If no, was the baby still breastfeeding at the last study visit? □ yes □ no
2.2.2 If baby has stopped breastfeeding since last study visit, date last breastfed dd mm year
Interpretation of a saturated MSM (simplified data)

Observed data \((L_1, A_1, L_2, A_2, Y)\)
- \(A_1 = \) continued breastfeeding at month 1
- \(A_2 = \) continued breastfeeding at month 2
- \(Y = \) gastrointestinal infections (yes/no) by month 2

Contrasts of different durations of breastfeeding:
- Risk of infection under imposed durations of breastfeeding
- \(\Pr(Y(0,0)=1)\)  \(\text{No breastfeeding}\)
- \(\Pr(Y(1,0)=1)\)  \(\text{Breastfeed until 1^{st} month}\)
- \(\Pr(Y(1,1)=1)\)  \(\text{Breastfeed until at least month 2}\)
Can we fit this model using the observed data?

Some assumptions must hold, including:

• **No interference**
  The potential outcomes of one subject do not depend on the exposures of other subjects.

• **Sequential consistency**
  When we observe \((A_1=a_1, A_2=a_2)\) the observed outcome \(Y\) is equal to the counterfactual under imposed \((a_1, a_2)\).

• **Sequential conditional exchangeability**
  No unmeasured confounders of \(A_2\) and \(Y\) *(i.e. No other confounders besides \(L_1, A_1\), and \(L_2)\)*
  No unmeasured confounders of \(A_1\) and \(Y\) *(i.e. No other confounders besides \(L_1)\)*

• **Sequential positivity**
  \(1 > \Pr(A_2=1 \mid L_1, A_1=1, L_2) > 0\) and
  \(1 > \Pr(A_1=1 \mid L_1) > 0\)
Estimation using propensity scores

Likely the most popular approach to estimation is Inverse Probability Weighting (IPW).

Goal: estimate $\Pr(Y(1,1)=1)$

1. Fit two regressions to estimate
   
   $g_1 = \Pr(A_1=1 \mid L_1)$ and $g_2 = \Pr(A_2=1 \mid L_1, A_1=1, L_2)$

2. Calculate a weight for each person with $A_1=1$ and $A_2=1$
   
   $w = 1/(g_1 \times g_2)$

3. Run an intercept-only regression with weights $w$, on the subset with $A_1=1$ and $A_2=1$. 
Key strengths and weaknesses of IPW

Weaknesses
- Inefficient
- Can be unstable in finite samples due to large weights (though weight stabilization may help)
- No established properties under general machine learning for propensity scores

Strengths
- Intuitive
- Can easily check/summarize weight sizes
- Simple ad hoc weight management strategies available
- Can easily handle longitudinal censoring
- Can directly be used to fit more complex MSMs

Note: IPW and all methods rely on untestable assumption of “no unmeasured confounders”
Outcome regression modeling (g-Computation)

- Estimation of MSMs using outcome modeling is available through the g-formula. (Robins, 1986)
- One particular formulation is more straight-forward to use when there are multiple time points and time-dependent confounders. (Bang and Robins, 2005)

For estimating $\Pr(Y(1,1)=1)$, it requires two quantities:

$$Q_2 = E(Y \mid L_1, A_1=1, L_2, A_2=1) \text{ and } Q_1 = E(Q_2 \mid L_1, A_1=1)$$

Under the causal assumptions,

$$E(Q_1) = \Pr(Y(1,1)=1)$$
Key strengths and weaknesses of g-Computation

Weaknesses

• Somewhat more challenging to extend to different outcome types (longitudinal counts, survival)
• No established properties under general machine learning for Q₁ and Q₂.
• Relies on correct specification of nested expectations

Strengths

• Efficient -- maximum likelihood approach
• Quick run-time (one model per time point)
• Can easily handle longitudinal censoring
• Can be used for general MSMs
**LTMLE**

**Longitudinal TMLE (LTMLE) for saturated MSMs** (van der Laan and Gruber, 2012)

- Involves modeling all components needed for IPW and g-computation.
- Idea: improve upon the estimation of $Q_t$ using information from the propensity weights.

**Motivation**

- Double robustness
  - Correct modeling of all $g_t$ and/or all $Q_t$ results in consistency
- Established properties when incorporating machine learning
  - Avoids parametric modeling, better consistency guarantees
- Efficient (beats IPW) when all models are correct.
LTMLE algorithm (1/2)

- Estimate $g_1 = \Pr(A_1=1|L_1)$ and $g_2 = \Pr(A_2=1|L_1, A_1=1, L_2)$
- Estimate $Q_2 = \mathbb{E}(Y|L_1, A_1=1, L_2, A_2=1)$

- Update $\hat{Q}_2$
  - Construct weight $w_2 = 1/(g_1* g_2)$
  - Run weighted logistic regression $Y \sim \text{offset}(\logit(\hat{Q}_2))$ in the subset $(A_1=1, A_2=1)$
  - Denote the intercept by $\hat{\varepsilon}_2$
  - Update $\hat{Q}_2^* = \logit^{-1}(\logit(\hat{Q}_2) + \hat{\varepsilon}_2)$
LTMLE algorithm (2/2)

• Estimate $Q_1 = \mathbb{E}(Q_2 \mid L_1, A_1=1)$ (using $\hat{Q}_2^*$ as outcome)

• *Update* $\hat{Q}_1$
  • Construct weight $w_1 = 1/(g_1)$
  • Run weighted logistic regression $\hat{Q}_2^* \sim \text{offset} \left( \logit(\hat{Q}_1) \right)$ in the subset $(A_1=1)$
  • Denote the intercept by $\hat{\varepsilon}_1$
  • Update $\hat{Q}_1^* = \logit^{-1}(\logit(\hat{Q}_1) + \hat{\varepsilon}_1)$

• Take the average of $\hat{Q}_1^*$
Key strengths and weaknesses of LTMLE

**Weaknesses**
- More challenging to extend to different outcome types (longitudinal counts, survival)
- Weights can cause instability – must be diagnosed
- Multiple moving pieces to check in practice

**Strengths**
- “Locally” efficient in the semiparametric model space
- Can easily handle longitudinal censoring
- Can be used for general MSMs (Petersen et al 2014)
- Can formally incorporate machine learning
PROBIT study

Interest lies in studying the effect of breastfeeding on infant health outcomes.

• Cluster-randomized breastfeeding encouragement program
• Previously, all evidence of protective effect of breastfeeding came from observational data.
  • Infeasible/unethical to randomize breastfeeding.
• Original purpose of study was to establish whether the intervention had an effect on the duration of breastfeeding and on infections.

Our goal: To evaluate the causal effect of breastfeeding duration on the number of infant gastrointestinal infections (Schnitzer, van der Laan, et al, 2014) and periods of hospitalization (Schnitzer, van der Laan, et al, 2018) using LTMLE.
PROBIT study

Longitudinal design: follow-up visits at 1, 2, 3, 6, 9, and 12 months.

- Baseline data: maternal demographic, educational, and smoking information, details about previous pregnancies, and infant information (sex, birth weight, gestational age, and Apgar score).
- Follow-up data: infant feeding, growth, illnesses, and hospitalization (binary).
- Within the 31 clusters, a total of **17,036 mother-infant pairs** had complete baseline data.
PROBIT data structure

- First time interval:

With 6 follow-ups, we obtain the following observed data:

\[ O = ( L_0, C_1, L_1, A_1, C_2, L_2, A_2, C_3, L_3, A_3, C_4, L_4, A_4, C_5, L_5, A_5, C_6, Y) \]

where \( Y \) is the total count of periods of hospitalization and \( L_t \) are the intermediate indicators of hospitalization.
Causal parameters of interest

We are interested in contrasting the expected counterfactual outcome under 6 different breastfeeding durations.

- E.g. $E( Y (1, 1, 1, 1, 1) )$ versus $E( Y (1, 1, 0, 0, 0) )$
  contrasts outcomes under duration past 9 months vs up to 2 months (stopping between 2-3 months).
Complex data structure: additional challenges

Longitudinal count data
• Conditioning on updated covariates $L_t$ means that only part of the outcome is random.
• Standard algorithms can’t be used directly.

Clustered observations by study center
• Can use cluster bootstrap to estimate variance
• Developed a closed-form asymptotic approximation for LTMLE

Longitudinal censoring

Details in book chapter
Results for hospitalizations

Table: Estimated differences in the expected counterfactual number of hospitalizations under different breastfeeding durations.

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>S.E.</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3-6 months vs 1-2 months</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-Comp (sequential)</td>
<td>-0.12</td>
<td>0.03</td>
<td>(-0.17,-0.06)</td>
</tr>
<tr>
<td>IPW</td>
<td>-0.05</td>
<td>0.03</td>
<td>(-0.10,0.01)</td>
</tr>
<tr>
<td>parametric TMLE</td>
<td>-0.06</td>
<td>0.01</td>
<td>(-0.08,-0.04)</td>
</tr>
<tr>
<td>TMLE with SL</td>
<td>-0.06</td>
<td>0.01</td>
<td>(-0.08,-0.04)</td>
</tr>
<tr>
<td><strong>9+ months vs 3-6 months</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-Comp (sequential)</td>
<td>-0.00</td>
<td>0.01</td>
<td>(-0.03,0.02)</td>
</tr>
<tr>
<td>IPW</td>
<td>0.05</td>
<td>0.02</td>
<td>(0.01,0.08)</td>
</tr>
<tr>
<td>parametric TMLE</td>
<td>-0.00</td>
<td>0.01</td>
<td>(-0.02,0.02)</td>
</tr>
<tr>
<td>TMLE with SL</td>
<td>-0.00</td>
<td>0.01</td>
<td>(-0.02,0.02)</td>
</tr>
</tbody>
</table>
Results for hospitalizations

Expected number of periods with hospitalizations with 95% confidence region as estimated by TMLE with Super Learner

Marginal expected number of hospitalizations

Breastfeeding termination
Results for number of gastrointestinal infections
Simulation results overview

- TMLE can reduce or remove bias due to the misspecification of models in the g-Computation.

- Adjusting for cluster (dummy variable) can protect against unmeasured confounding related to the cluster.

- Essential to implement TMLE with machine learning when the true model forms are unknown.

- TMLE rivals g-Computation, beats IPW, in terms of standard error size.

- Implementation of modified algorithm to incorporate the count outcome had a large impact (12% vs 0% bias for g-Computation).
Conclusions

• Causal inference methods are warranted in particular in settings where exposures vary over time.
• Choosing the counterfactual parameters of interest is an important part of the analysis.
• All methods rely on a set of assumptions that can be discussed and refuted. The extent of the resulting bias may also be discussed or approximated.
• LTMLE is desirable because of its efficiency and ability to integrate machine learning methods.
  • May need non-standard implementations with some data structures.
References (1/2)


Funding

• I currently hold a Canadian Institutes of Health Research (CIHR) Canada Research Chair in Causal Inference and Machine Learning in Health Sciences.

• My course on causal inference was developed under my current position in the Faculté de pharmacie at Université de Montréal.

• My PhD work (2009-2012) was supported by the Natural Sciences and Engineering Research Council of Canada (CGS-D) and by the Fonds de recherche du Québec – Nature et technologies (Bourse de stage international).

• The Promotion of Breastfeeding Intervention Trial, PI Michael S. Kramer, was originally funded by the National Health Research and Development Program (Health Canada), the Thrasher Research Fund, the United Nations Children’s Fund (UNICEF), and the European Regional Office of the World Health Organization (WHO).