Efficient causal inference for longitudinal data using targeted maximum likelihood estimation

Mireille Schnitzer[1]
Erica Moodie[1]
Robert Platt[1]
Mark van der Laan[2]

[1]Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

[2]Public Health, Division of Biostatistics
University of California at Berkeley

February 9, 2012

Contents

- The PROmotion of Breastfeeding Intervention Trial (PROBIT)
- Estimation of causal parameters for longitudinal data
- Semiparametric efficiency
- Targeted maximum likelihood estimation (TMLE)
- TMLE for longitudinal data
- Simulation results
- PROBIT results

PROmotion of Breastfeeding Intervention Trial

- Kramer et al 2001
- Cluster-randomized trial assigned breastfeeding intervention to hospitals in Belarus.
  - Subjects in study were mothers/infant pairs where the mother had chosen to breastfeed.
  - Breastfeeding intervention taught breastfeeding technique to the mother.
- Subjects in study were mothers/infant pairs where the mother had chosen to breastfeed.

PROmotion of Breastfeeding Intervention Trial

Interest lies in studying the effect of breastfeeding on gastrointestinal infections in infants.

- Previously, all evidence of protective effect of breastfeeding from observational data.
  - Infeasible/unethical to randomize breastfeeding.
  - Intervention was developed to increase duration and exclusivity of breastfeeding.
- Original purpose of study was to establish whether the intervention had effect on duration of breastfeeding and infections

We attempted to extract additional information from the dataset to obtain the causal effect of longer breastfeeding duration on infant infections...

Causal parameter of interest

Often in causal inference (potential-outcome framework), we are interested in the parameter described as

- **Marginal exposure-specific mean**
  - $\psi_a = E(Y^a)$: the population mean of the counterfactual outcome under fixed exposure a.

Same idea goes for longitudinal data.

- Consider data of the form $O = (W, L_1, A_1, L_2, ..., L_{K-1}, A_{K-1}, Y)$.
- Define a *fixed exposure pattern* as $a = (a_1, a_2, ..., a_{K-1})$.
- Then, the marginal exposure-specific mean can be defined with a time-varying exposure.

Causal parameter of interest

The PROBIT data recorded baseline information of the mother and infant at birth.

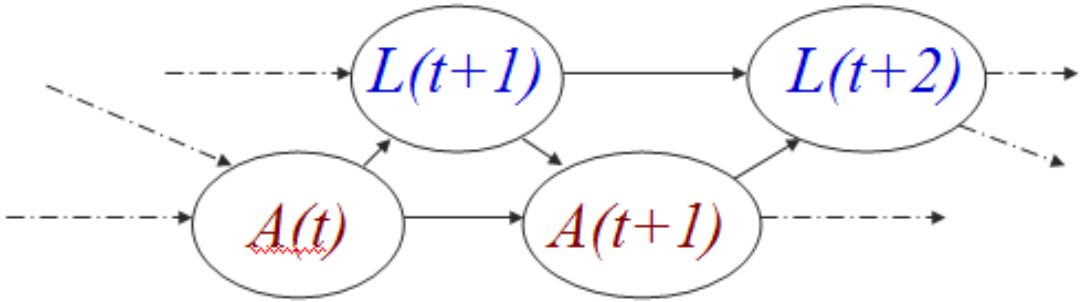
- Follow-ups at 1, 2, 3, 6, 9, and 12 months of age
- At each follow-up, breastfeeding status and infection counts over past interval were ascertained

Therefore, the corresponding parameter of interest is the marginal mean number of infection counts under breastfeeding pattern a.

- For example, $a = (1,1,1,1,1)$ ongoing breastfeeding; $a = (1,1,0,0,0)$ stopping after the second follow-up.

Longitudinal data and time-dependent confounding

Standard non-causal longitudinal modeling does not estimate causal parameters in the presence of time-dependent confounding.



- The L_t are time-dependent covariates that are affected by prior treatment.
- Referred to as *time-dependent confounders*.
- Crude estimates, controlling for baseline/time-dep confounders give biased results (Fewell et al, 2004)

Estimation methods

- Typical (causal) approaches may involve
  - Inverse probability of treatment weighting (IPTW) – weighting outcomes
  - G-computation – fitting the likelihood

Estimation methods: IPTW

Example of IPTW without stabilization:

$$\hat{\psi}_a^{IPTW} = \frac{1}{n}\sum_{i=1}^{n} I(\bar{A}_i = a) Y_i w_i$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{I(\bar{A}_i = a) Y_i}{p(A = a|X)}$$

Concerns with IPTW:

- These weights can sometimes get very big!
- Not an efficient method

Estimation methods: G-computation

G-computation (Gill & Robins, 2001) plug-in estimator: likelihood computation, holding exposure pattern constant.

- point treatment (1 time point)

$$\hat{\Psi}_a^{GCOMP}(p_n) = \sum_{i=1}^{n} E(Y \mid A(1) = a_1, W_i) p(W_i)$$

- 2 treatment times

$$\hat{\Psi}_a^{GCOMP}(p_n) = \sum_{i=1}^{n} \sum_{l_1 \in \mathscr{L}_1} E(Y \mid A(2) = a_2, L(1) = l_1, A(1) = a_1, W_i) \times p(L(1) = l_1 | A(1) = a_1, W_i) p(W_i)$$

- G-computation is a plug-in estimator because it computes an estimate of the parameter using a function of the underlying data density, $\hat{p}$

Better estimation in causal inference

Taking a step back, what are our goals in causal inference?

- We want to estimate a well-defined, interpretable parameter.
  - Efficient, unbiased estimation of this parameter
  - Small sample performance (boundedness, speed of convergence)
- We may not actually care about estimation of many nuisance parameters (e.g. density components in the G-computation)

Better estimation in causal inference

Plug-in estimation: we are interested in estimating a parameter that can be expressed as a smooth function Ψ of a density p_0.

- True value is $\psi = \Psi(p_0)$
- May only require a portion of the underlying density, $p_0 = Q_0 \cdot g_0$ so that $\psi = \Psi(Q_0)$.
- Q_0 must be estimated with some fitting method, resulting in estimate $\hat{\psi} = \Psi(Q_n)$.

Plugging in a correctly specified parametric density estimate will produce efficient, unbiased inference. But if your parametric modeling assumptions are wrong, you're in trouble.

Semiparametric inference for plug-in estimators

- Semiparametric theory: any asymptotically linear estimator has a unique influence function (IF).
- If $\Psi(Q_n)$ estimates $\Psi(Q_0) = \psi$ then $D(p_0)$ is the IF if

$$n^{1/2}\{\Psi(Q_n) - \Psi(Q_0)\} = n^{-1/2} \sum_{i=1}^{n} D_i(p_0) + o_p(1).$$

- By application of CLT,

$$n^{1/2}\{\Psi(Q_n) - \Psi(Q_0)\} \xrightarrow{\mathscr{D}} N\left(0, E(DD^T)\right)$$

Semiparametric inference for plug-in estimators

- The variance of the estimator is the variance of its IF.
  - So we want to get an estimator associated with an IF with minimal variance.
  - This is called the *efficient influence function* (EIF)
- Semiparametric theory has provided tools to find the EIF for many parameters (e.g. Semiparametric Theory and Missing Data by A A Tsiatis).
- Ideally, wed like to obtain efficient estimators with this associated EIF.

Semi-parametric inference for plug-in estimators

Influence functions can be characterized as components of a Hilbert space of mean zero functions with finite variance.

- One way to estimate with the EIF is to directly solve it as an estimating equation
- i.e. if you find some

$$p_n \text{ s.t. } \frac{1}{n}\sum_{i=1}^{n} D_i^*(p_n) = 0, \text{ then } \hat{\psi} = \Psi(p_n).$$

- Such an estimator will produce semiparametric efficient estimators that are also *doubly robust*
  - i.e. when both a model for the exposure and a model for the other variables must be fit, you will only need to correctly specify one of them to obtain asymptotically unbiased inference.

TMLE

So why isn't this *good enough*?

- Treating the EIF as an estimating equation can lead to several problems:
  - Unstable inference under data sparsity,
  - Possibility of multiple solutions,
  - Unbounded.
- The doubly robust estimators in Kang & Schafer (2007) are examples of efficient estimators
  - perform poorly under mild misspecification of density

TMLE

TMLE (van der Laan and Rubin, 2006) is a framework that uses plug-in estimators while solving the EIF.

- Plug-in estimation:
  - single solution
  - respects global bounds
  - can be formulated as a loss-based estimator

TMLE Procedure

Suppose, as in the G-computation case, that the density can be decomposed into $p_0 = Q_0 \cdot g_0$. The TMLE for $\Psi(Q_0)$ then proceeds as follows:

- Estimate density components Q_n and g_n
- Create a parametric fluctuation $Q_n(\varepsilon)$ with the constraints
  - $Q_n(0) = Q_n$
  - the score of $Q_n(\varepsilon)$ at $\varepsilon = 0$ spans the EIF, i.e.

$$\frac{d \ln Q_{n,j}(\varepsilon)}{d\varepsilon} \propto D_j^*(p_n)$$

TMLE Procedure

- Maximize the likelihood w.r.t. ε

$$\hat{\varepsilon}^1 = \text{argmax}_\varepsilon \ln Q_n(\varepsilon)$$

  (i.e. solve the score equations).
- Plug in the ε estimate to obtain the updated density

$$Q_n^1 = Q_n(\hat{\varepsilon}^1).$$

- Fluctuate this updated density, and iterate until convergence ($\varepsilon^k = 0$).
- Let Q_n^∞ be the limit density. The final estimator is given as $\Psi(Q_n^\infty)$.

TMLE Procedure

- This iterative maximization is the process that obtains a density estimate that solves the EIF.
- Since Q_n^∞ is a solution to the EIF, $\Psi(Q_n^\infty)$ is semiparametric efficient.

TMLE Procedure

- Intuitively, the MLE of the density optimizes the fit of the density model.
- The TMLE adjusts the density estimate to efficiently estimate the parameter of interest.
  - We don't really care about the underlying density.
- TMLE can be rephrased as a loss-based estimator by simply substituting the requirement of a likelihood specification (in the fluctuation step) with the choice of a loss function.

TMLE for 2 time points

- Consider a 2 time point observations structure

$$O = (W, A(1), L(1), A(2), Y).$$

  - A is treatment at time point 1 and 2.
  - $L(1)$ is a binary covariate.
  - Y is an outcome that can be assigned an exponential family member working model.

- We wish to estimate the parameter

$$\psi = \Psi(p_0) = E(Y_{A(2)=1,A(1)=1})$$

TMLE for 2 time points

- As before, the G-computation formula for the marginal exposure-specified mean is

$$\hat{\Psi}_a^{GCOMP}(p_n) = \sum_{i=1}^{n} \sum_{l_1 \in \mathscr{L}_1} E(Y \mid A(2) = a_2, L(1) = l_1, A(1) = a_1, W_i) \times p(L(1) = l_1 | A(1) = a_1, W_i) p(W_i)$$

- Therefore, this plug-in estimator requires conditional density estimates for 3 density components: $\bar{Q}_Y, \bar{Q}_{L_1}$, and $\bar{Q}_W$.

TMLE for 2 time points

- The first step is estimating the required underlying density components: $\bar{Q}_Y, \bar{Q}_{L_1}$, and $\bar{Q}_W = \frac{1}{n}$.
- A TMLE procedure can then be constructed to update the 3 conditional densities required in the G-computation estimator (van der Laan, 2010).
- This can be done by using a generalized linear loss function for the GLM outcome with a linear fluctuation.
- The intermediate density is updated using a logistic loss function with a linear fluctuation.
- Since it is already nonparametric efficient, the estimate of p_W requires no update.

TMLE for 2 time points

- All fitting procedures should be done optimally for CV fit, while avoiding modeling assumptions.
- The success of this procedure also rests on correctly fitting the exposure models:

$$g_1(a_1) = p(A(1) = a_1 | W_i), \text{ and}$$
$$g_2(a_2, a_1) = p(A(2) = a_2 | L(1), A(0) = a_1, W_i).$$

  - However, if you get $g_n = (g_1, g_2)$ or $Q_n = (\bar{Q}_Y, \bar{Q}_{L_1})$ you are guaranteed unbiasedness.

TMLE for 2 time points

- This procedure converges after the first round of updates.
- The final step of the TMLE procedure is to take the updated fits and plug them into the G-computation estimator.

Some simulation results with this estimator

- We generated a continuous outcome conditional on two binary outcomes, a binary intermediate variable and baseline variables in the time-dependent form:

$$O = (W, A(1), L(1), A(2), Y).$$

- Estimation challenges were produced by:
  - Misspecifying the exposure/outcome/both by omitting a confounder
  - Producing positivity violations in the data generation.
- All scenarios were tested with IPTW, G-Comp, TMLE, and the efficient EE method proposed by Bang & Robins (2005).
  - For positivity violations, also scaled the outcome to [0,1] and used a logistic loss function instead of a squared-error loss function.

Simulation results: Omitted confounder

	Correct Specification				Misspecified Exposure			
	% Bias	S.E.	rMSE	Cover	% Bias	S.E.	rMSE	Cover
$n = 200$								
TMLE	-13	27	27	94	-22	25	25	95
G-COMP	-10	23	23	95	-10	23	23	95
IPTW	-2	25	25	94	440	23	30	84
BR	-12	33	28	95	-11	28	27	95
$n = 1,000$								
TMLE	-12	12	12	93	-11	11	11	93
G-COMP	-14	10	10	94	-14	10	10	94
IPTW	-14	11	11	93	430	10	21	53
BR	-13	12	12	94	-11	12	12	94
$n = 10,000$								
TMLE	-3	4	4	93	-3	4	4	92
G-COMP	-2	3	3	94	-2	3	3	94
IPTW	-3	3	4	94	443	3	20	0
BR	-3	4	4	92	-4	4	4	93

Simulation results: Omitted confounder

	Misspecified Outcome				Total Misspecification			
	% Bias	S.E.	rMSE	Cover	% Bias	S.E.	rMSE	Cover
$n = 200$								
TMLE	-10	27	27	93	436	25	31	86
G-COMP	-437	23	30	84	-437	23	30	84
IPTW	-2	25	25	94	440	23	30	84
BR	-20	28	27	94	437	26	32	87
$n = 1,000$								
TMLE	-12	12	12	93	437	11	22	59
G-COMP	426	10	21	54	426	10	21	54
IPTW	-14	11	11	93	430	10	21	53
BR	-12	12	12	93	438	11	22	58
$n = 10,000$								
TMLE	-3	4	4	93	451	4	20	0
G-COMP	440	3	19	0	440	3	19	0
IPTW	-3	3	4	94	443	3	20	0
BR	-3	4	4	93	448	4	20	0

Simulation results: Near positivity violations

	Mild Data Sparsity				Severe Data Sparsity			
	% Bias	S.E.	rMSE	Cover	% Bias	S.E.	rMSE	Cover
$n = 200$								
TMLE	233	3,110	441	93	-	-	-	-§
TMLE_{log}	104	92	95	93	38	85	98	92
G-COMP	13	39	39	94	9	31	31	95
IPTW	113	49	57	92	91	41	52	92
BR	1,128	2,119	913	96	-	-	-	-§
BR_{log}	196	115	114	95	85	100	112	95
$n = 1,000$								
TMLE	20	76	68	93	857	20,131	8,708	93
TMLE_{log}	-13	41	43	93	-38	50	56	92
G-COMP	3	18	18	94	-4	14	14	95
IPTW	11	26	33	92	63	25	37	91
BR	-79	100	103	92	-4,068	868	4,182	93
BR_{log}	-1	47	49	92	-58	57	63	93

Simple analysis of the PROBIT

- 2 time-points considered (subjects with missing data removed):
  - Want to know whether duration of breastfeeding up until 6 months affects expected number of infections between 3-6 months.
  - Treatment parameters of interest are
    - bf at 3 months compared to stopping before 3 months
    - bf at 6 months compared to stopping at 3 months
  - Many baseline suspected confounders collected
  - Strong intermediary variable is presence of infection between 0 and 3 months.

Breastfeeding effect estimates for each model

Model	Estimate	S.E.	95% C.I.
Effect of breastfeeding for 6 months vs. 3 months			
TMLE	-13	4	(-20,-5)
G-comp	-13	4	(-20,-6)
IPTW	-12	3	(-20,-6)
BR	-13	4	(-20,-6)
Effect of breastfeeding for 6 months vs. $<$3 months			
TMLE	-20	3	(-28,-14)
G-comp	-21	4	(-30,-14)
IPTW	-19	4	(-28,-13)
BR	-20	4	(-30,-14)

All values given as $\times 10^3$ the original value.

Simple analysis of the PROBIT

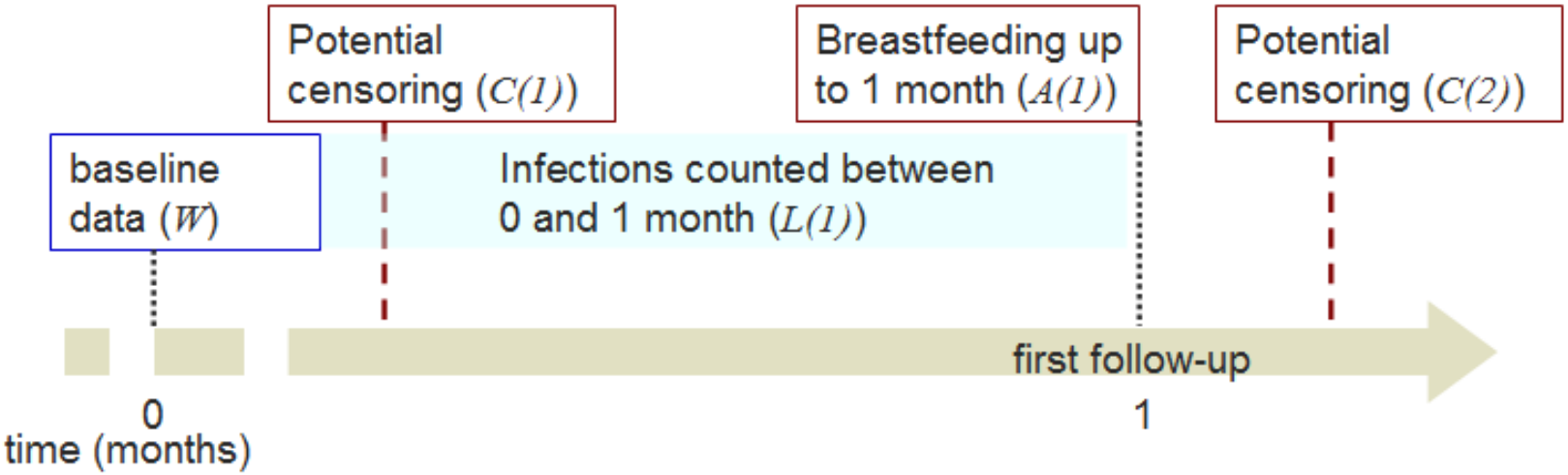
- A reduction of 0.013 in expected mean infection count corresponds with a 42% reduction (compared to the expected number with between 3-6 months of breastfeeding).
- A reduction of 0.020 in expected mean infection count corresponds with a 51% reduction (compared to the expected number with $<$3 months breastfeeding).
- This last value corresponds with a NNT of 50.

Issues with this formulation

- While this factorization of the density is intuitive there are several difficulties with implementation for $K > 2$:
  - Requirement of binary decomposition of any intermediate variables;
    - For continuous variables, should it be selected by user or selected data-adaptively?
  - Computational intensity of G-comp;
    - May be r variables at each $L(t)$ node, giving $(r * K)!$ operations
  - Tough to model components of influence function ($K > 2$).
    - Some require G-computation themselves.

Full PROBIT data structure

The PROBIT data structure (first time interval):



Hence, with 6 follow-ups, we obtain the time-ordering

$$O = (\quad W, C(1), L(1), \quad A(1), C(2), L(2),$$
$$A(2), C(3), L(3), \quad A(3), C(4), L(4),$$
$$A(4), C(5), L(5), \quad A(5), C(6), Y \quad)$$

Full PROBIT data structure

For this extended data structure, we also incorporate censoring into the definition of the parameter and the estimation technique.

- parameter, $\psi_{a,C=0} = E(Y^{a,C=0})$ is the marginal exposure-specific mean if all subjects were uncensored.

With $K > 2$ data points, we also take advantage of an easier TMLE technique...

New formulation

Inspired by Bang & Robins (2005), write the target parameter as a product of conditional outcomes (van der Laan & Gruber, 2011).

- This uses the same concept as the property of iterated expectations:

$$E(X) = EE(X|Y).$$

- we can define sequentially:

$$\begin{aligned}
\bar{Q}_K^a &= E(Y^{a,C=0}|\bar{L}_{K-1}^{a,C=0}) \\
\bar{Q}_{K-1}^a &= E(\bar{Q}_K^a|\bar{L}_{K-2}^{a,C=0}) \\
&\vdots \\
\bar{Q}_1^a &= E(\bar{Q}_2^a|W)
\end{aligned}$$

New formulation

- With this new notation, the parameter can be written as

$$\begin{aligned}\psi_{a,C=0} &= E(E(...E(E(Y^{a,C=0}|\bar{L}_K^{a,C=0})|\bar{L}_{K-1}^{a,C=0})...|W)),\\ &= E(\bar{Q}_1^a).\end{aligned}$$

New formulation

An alternative likelihood-based formula relying on this factorization is straight-forward. Algorithm:

- Fit $\bar{Q}_K^a$ by regressing Y on the full history,
- Fit $\bar{Q}_{K-1}^a$ by regressing $\bar{Q}_K^a$ on the history of $L(K-1)$,
- ...
- Fit $\bar{Q}_1^a$ by regressing $\hat{Q}_2^a$ on W.
- The mean estimate is given as the mean of $\bar{Q}_1^a$.

New formulation

- Correspondingly, a new TMLE can be constructed with the EIF presented in terms of these sequential conditional densities.
- For each density model, the outcome is always the prediction of the last $\bar{Q}_t^a$.
  - To update each density, each $\bar{Q}_t^a$ can be scaled to [0,1] and assigned a logistic loss function and a linear fluctuation.

New formulation – benefits

- No restriction on the data structure, just need the ability to assign a loss function to the sequential conditional density outcomes.
- Only one model has to be fit per time point + 1 logistic regression to minimize the loss function.
- Easy to code, no additional complications for $K > 2$ time points.
- Has the same benefits of usual TMLE (doubly-robust, efficient, etc).

Simulations in the form of the PROBIT data

- We generated data with the same structure as the PROBIT data (6 tp, censoring) and compared different methods of estimation.
- 3 different data scenarios were considered: correct specification of propensity (g_n models), unmeasured confounder, and near positivity violations.
- Data was generated incrementally (as before) so the sequential density components $\bar{Q}_t^a$ were *always* misspecified.
- We calculated standard errors by bootstrap (when available, inference was essentially the same using the influence curve).

Simulation results: exposure model correct

$n = 1000$; 500 data sets were generated.

Method	% bias	SE (BS)	MSE	% Cover (BS)
		no unmeasured confounders		
G-Comp (likelihood)	-0.1	0.07	0.003	94.4
G-Comp (sequential)	2.4	0.12	0.016	96.0
parametric TMLE	-0.3	0.06	0.003	94.2
IPTW	-0.4	0.10	0.010	92.2
Efficient EE	-0.2	0.06	0.003	94.0

Simulation results: unmeasured confounder and near positivity violations

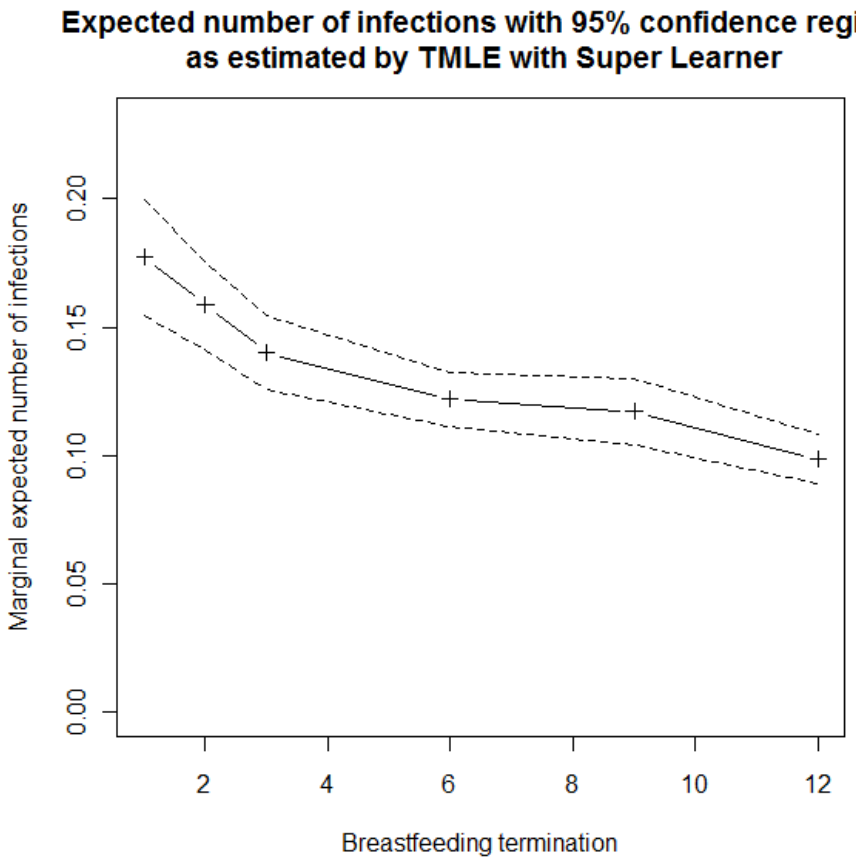
Method	% bias	SE (BS)	MSE	% Cover (BS)
		unmeasured confounder		
G-Comp (likelihood)	1.1	0.07	0.004	94.4
G-Comp (sequential)	4.4	0.04	0.023	81.4
parametric TMLE	0.8	0.06	0.004	93.0
IPTW	0.6	0.10	0.010	94.2
Efficient EE	-0.9	0.06	0.004	93.4
		positivity violations		
G-Comp (likelihood)	0.0	0.06	0.003	96.2
G-Comp (sequential)	5.8	0.13	0.027	76.4
parametric TMLE	-0.8	0.08	0.005	95.4
IPTW	-0.7	0.13	0.019	90.6
Efficient EE	2.1	0.29	0.057	92.2

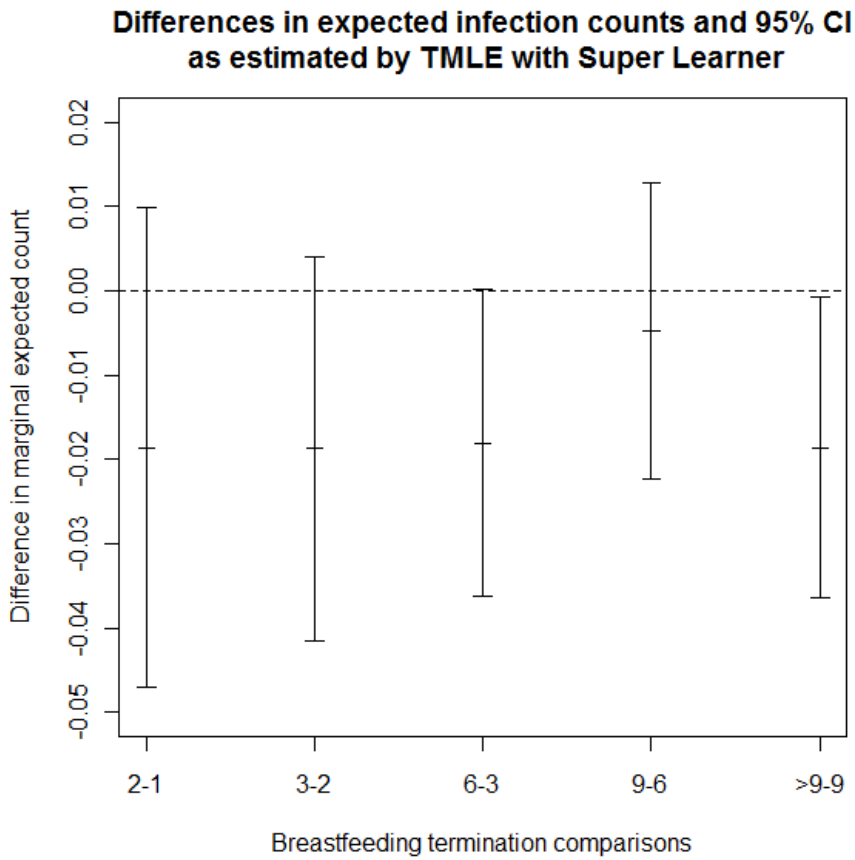
Full analysis of the PROBIT

We used each of our estimators to estimate the causal parameter:

- **Parameters of interest for the full PROBIT analysis**
  - $\psi_{a,C=0}$: the marginal population mean number of infection counts over 12 months, under each possible exposure history, a, and no censoring.
- Using this, we can also compare differences between the monotone exposure patterns.
- The TMLE method was implemented with a non-parametric method that optimizes the CV fit by fitting a library of models.

Full analysis of the PROBIT: TMLE results





Full analysis of the PROBIT: TMLE results

- Additional comparisons can be made.
  - For instance, comparing breastfeeding termination before one month to between 3-6 months, exposure effect

$$-0.055 \quad (95\% CI : -0.08, -0.03),$$

  - Comparing bf less than one month vs. over 9 months, exposure effect

$$-0.079 \quad (95\% CI : -0.1, -0.05).$$

  - Above corresponds to a NNT of 13 to avoid 1 gastrointestinal infection in the first year of life.

Conclusions

- Time-dependent confounding in longitudinal data requires causal modeling.
- Choose an interpretable (causal) parameter to estimate.
- Rather than optimizing the model fit for many nuisance parameters, optimize the fit of the target parameter.
- Efficient methods have the added benefit of double robustness.
- TMLE has the additional benefit of stability, as a plug-in estimator.

Thank you

Research supported by grants from NSERC, FQRNT and CRM.