Discussion Assignment #3

Sabrina Boyce, Shelley Facente, and Steph Holm

11/6/2019

Question 1: Specify the question of interest. How does this question differ from that addressed in the original analysis of the randomized trial? What was the authors' motivation for doing a secondary analysis?

The question of interest is: *What causal effect does breastfeeding duration have on the number of infections a newborn is expected to experience in their first year?* This differs from the original analysis in the trial because they were originally evaluating the effect of longer breastfeeding duration (via the PROBIT intervention) on gastrointestinal tract infection, using an intent-to-treat analysis (breastfeeding was not randomized). While their results indicated a significant reduction in infection incidence for infants whose mothers were in the intervention group, since breastfeeding was not randomized, causality could not reasonably be determined. Since breastfeeding can't ethnically be randomized, a secondary analysis using statistical techniques to assess causality was required.

Question 2: Specify the longitudinal causal model $\mathcal{M}^{\mathcal{F}}$ for individual level data (with endogenous nodes and time ordering corresponding to the observed data as specified in Section 3). What do you think about the temporal ordering assumptions? Any concerns? *Bonus:* How might you modify this causal model to more accurately describe knowledge about the true data generating process?

We know from the article that the endogenous nodes are:

$W =$ baseline covariates

$C_t, t = 1...K =$ Whether the subject was censored before the tth timepoint (including the final timepoint, K, which would indicate censoring at the time the outcome was measured.)

$L_t, t = 1...K-1 =$ Whether the infant had any gastrointestinal infections between timepoints t - 1 and t. Any subject already censored during or prior to any given timepoint received an $L(t)$ value of zero.

$A_t, t = 1...K-1 =$ Breastfeeding status at timepoint t (*yes/no*).

$Y =$ the outcome: the total number of infections accrued up until and including visit K. Any subject censored during or prior to time K received a Y value of zero.

Therefore, the longitudinal structural causal model $\mathcal{M}^{\mathcal{F}}$ is:

$$
\begin{aligned}
W &= f_W(U_W) \\
C(t) &= f_C(\bar{C}(t-1), W, U_{\bar{C}}(t)), t = 1, 2, ...6 \\
L(t) &= f_L(\bar{L}(t-1), C(t), W, U_{\bar{L}}(t)), t = 1, 2, ...5 \\
A(t) &= f_A(\bar{A}(t-1), \bar{L}(t), \bar{C}(t), W, U_{\bar{A}}(t)), t = 1, 2, ...5 \\
Y &= f_Y(\bar{A}(5), \bar{L}(5), \bar{C}(6), W, U_Y(6)))
\end{aligned}
$$

One concern about the time-ordering is that at each timepoint t, $L(t)$ measures whether an infection occurred between time $t-1$ and time t, and $A(t)$ measures whether the mother is still breastfeeding at time t or stopped at some point between time $t-1$ and time t. Therefore there is uncertainty about the time ordering of infection and breastfeeding cessation, which must be considered relative to inference.

Question 3: Specify the counterfactual outcomes of interest. Hint: First define the interventions of interest. Then define the counterfactual outcome under these interventions (the counterfactual outcome is a random variable). How are these counterfactual outcomes generated using an intervention on the causal model?

Intervention of interest: A_t=contintued breastfeeding from t-1 to t (binary: 0, 1) for k=1,2,3,4,5 The intervention regime $\bar{A}(t)$ includes all possible combinations of 0 and 1 at every time point 1-5. There is a restriciton on these possible intervention regimes, though, limiting the combinations of 0/1 across time points to only those that assume that once you have a 0, you will have 0 for all subsequent time points. This is a restriction requirement that is given by real world constraints.

C_t=an indicator variable if infant is censored at time t (K=1-6). We intervene on censoring such that $\bar{C} = 0$, meaning that no censoring occurs.

Our counterfactual outcome is $Y_{\bar{A},\bar{C}}$, which is the number of infant infections observed under breastfeeding regime $\bar{A}(t)$ and no censoring.

These outcomes are generated by setting the value of A_t to $\bar{A}(t) = \bar{a}(t)$ for each time point and setting $\bar{C}(t) = 0$ and observe the number of infant infections at t=6.

Question 4: Specify the target causal parameter.

$$\psi_{\bar{a}} = E(Y_{\bar{a}})$$

The target causal parameter is the marginal mean of the counterfactual outcome.

Question 5: What are the observed data? What is the assumed link between the observed data and the structural causal model $\mathcal{M}^{\mathcal{F}}$? Factorize the observed data distribution $P_0(O)$ according to the time-ordering.

The observed data are from the PROBIT study of 17,036 mother/infant pairs from 31 hospitals in Belarus who were provided data at 6 time points. We asssume that the observed data were generated by sampling n i.i.d. copies from the data generating process that is compatible with our structural equation model.

n i.i.d. copies of $O = (W, C_1, L_1, A_1....L_{K-1}, A_{K-1}, C_K, Y)$distributed over$P_0$

Factorizing the distribution in the following way:

$$P(O=o) = Q_Y(Y|\bar{C}_K = 0, \bar{A}_{K-1} = \bar{a}(K-1), \bar{L}_{K-1}\bar{l}(K-1), W = w)\times$$

$$\prod_{t=1}^{K-1} Q_{L_t}(L(t) = l(t)|\bar{C}_t = 0, \bar{A}(t-1) = \bar{a}(t-1), \bar{L}(t-1) = \bar{l}(t-1), W = w) \times Q_w(W = w) \quad \times$$

$$\prod_{t=1}^{K-1} g_{A_t}(A(t) = a(t)|\bar{A}(t-1) = \bar{a}(t-1), \bar{L}(t) = \bar{l}(t), \bar{C}_t = 0, W = w) \quad \times$$

$$\prod_{t=1}^{K} g_{C_t}(C(t) = 0|\bar{A}(t-1) = \bar{a}(t-1), \bar{L}(t-1) = \bar{l}(t-1), \bar{C}(t-1) = 0, W = w))$$

Question 6: What are the needed identifiability assumptions? Do they seem reasonable here? Any particular concerns?

The needed identifiability assumptions are sequential randomization and positivity, as well as the assumption of uninformative censoring (censoring is not associated with exposure and outcome)

Sequential randomization for this data structure:

$Y_{\bar{a}} \perp A_1 | L_1, C_1 = 0, W$
$Y_{\bar{a}} \perp A_2 | L_2, A_1 = a_1, L_1, \bar{C}_2 = 0, W \; \ldots.$
$Y_{\bar{a}} \perp A_5 | L_5, A_4 = a_4, L_4, A_3 = a_3, L_3, A_2 = a_2, L_2, A_1 = a_1, L_1, \bar{C}_5 = 0, W$

Positivity

$P(A(k) = a(k) | \bar{A}_{K-1} = \bar{a}_{K-1}, \bar{L}_K, \bar{C}_K = 0, W) > 0$ - a.e., for t=1,2..5
$P(C(K) = 0 | C(K-1) = 0, \bar{A}_K - 1) > 0$

These assumptions do seem reasonable here, if we have limited our possible intervention regimes to account for women likely not restarting breastfeeding after stopping. We may have practical positivity because there are a lot of time points, but this is likely okay given the large sample.

Question 7: Under the assumption of sequential randomization and positivity, write out the (counterfactual or "post intervention") distribution $Q^{\bar{a}}$ a of the counterfactual non-intervention variables (W; $\bar{L}_5^{\bar{a}}$, $Y^{\bar{a}}$) as a function of the observed data distribution P_0.

The "Post Intervention" distribution $Q^{\bar{a}}$ of the counterfactual non-intervention variables is:

$$P_O^{\bar{a}} = Q^{\bar{a}}(Y^{\bar{a}}, \bar{L}_K^{\bar{a}}, W) = Q_Y(Y | \bar{C}_K = 0, \bar{A}_{K-1} = \bar{a}_{K-1}, \bar{L}_{K-1}, W) \;\; \times$$
$$\prod_{t=1}^{K-1} Q_{L_t}(L_t | \bar{C}_t = 0, \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1}, W) Q_W(W)$$

Question 8: Specify the statistical estimand using the traditional G-computation formula.

$$E(Y_{\bar{a},\bar{C}=0}) = \sum_{\bar{w}} \sum_{\bar{l}} E_0(Y | \bar{A}(5) = \bar{a}(5), \bar{L}(5) = \bar{l}(5), \bar{C}(6) = 0, W) \;\; \times$$
$$\prod_{1}^{5} P_0(L(t) = l(t) | \bar{A}(t-1) = \bar{a}(t-1), \bar{L}(t-1) = \bar{l}(t-1), \bar{C}(t) = 0, W) P_0(W = w)$$

Question 9: Briefly review implementation of "traditional" longitudinal parametric G-Computation for this estimand. What are some possible pros/cons to this approach?

"Traditional" longitudinal parametric G-Computation requires:

- estimation of distribution of each time-varying covariate given the past, and
- evaluation through simulation.

The pros of this are that the traditional longitudinal parametric G-comp:

- tends to be less sensitive to positivity violations than other options, and

- is efficient, if the model is correct.

However, the cons are that:

- estimates are susceptible to bias if model not correctly specified, and
- in normal circumstances, this requires estimating lots of conditional densities.

Question 10: What is the sequential representation of the statistical estimand?

By sequentially breaking up the expectations into nested conditional expectations, the marginal expectation of the outcome assuming no censoring can be represented like this:

$$\begin{aligned}\Psi_{\bar{a}} &= E(Y_{\bar{a}}) = E(Y_{\bar{a},\bar{c}=0} \\ &= E\{E(Y|C_K = 0, \bar{A}_{K-1} = \bar{a}_{K-1}, \bar{L}_{K-1}, W)\} \\ &= E[E\{E(Y|C_K = 0, \bar{A}_{K-1} = \bar{a}_{K-1}, \bar{L}_{K-1}, W)|C_{K-1} = 0, \bar{A}_{K-2} = \bar{a}_{K-2}, \bar{L}_{K-2}, W\}]\end{aligned}$$

(Note that this nesting would continue until reaching the point that E(Y) is only conditional on W, not represented here.)

Question 11: Give an overview of implementation of the sequential G computation estimator. What are some advantages over the traditional parametric G comp estimator?

To implement the sequential G-comp estimator, you fit a model for each level of conditioning, beginning with the innermost expectation (i.e. first fit Q_K and then the rest of the Q_ts). In this case, that means regressing the outcome Y ($t = 6$) on all the covariate history up to timepoint $t = 5$, for only those completely uncensored subjects with observed breastfeeding regime ($\bar{A}(5) = \bar{a}(5)$, $\bar{C}(6) = 0$). Then, predict the conditional outcome for each individual moving backward in time, eventually taking a mean of Q_1 over all observations, conditional only on the baseline covariates (i.e. including those who were censored) to obtain the target causal parameter.

This method provides a few advantages over the traditional parametric G-comp estimator described in Question 9, namely:

- No longer have to estimate a series of conditional densities, but instead a series of conditional expectations (easier to estimate well, though in this situation with only one binary L covariate that isn't much of a problem).
- Many fewer dimensions (Q factors), which means we don't need to estimate all the conditional densities in the same way.

Question 12: Give an overview of implementation of the sequential TMLE.

To implement sequential TMLE, we go through the following steps:

1. Estimate $E_0(Y|\bar{A}, \bar{L}, W) \equiv \bar{Q}_0(\bar{A}, \bar{L}, W)$, using SuperLearner or similar options.
2. Generate predicted values for Y for each individual, given that individual's $\bar{A}_i, \bar{L}_i, W_i$.
3. Estimate the treatment mechanism (the probability associated with obtaining a given history of breastfeeding $\bar{a}$ up until time $t - 1$), again using SuperLearner or similar ($g_0(\bar{A}|\bar{L}, W)$).
4. Use this estimate to create a new "clever covariate" ($H_n(\bar{A}, \bar{L}, W)$) for each individual.
5. Update the initial estimate of $E_0(Y|\bar{A}, \bar{L}, W)$ using that clever covariate, by running a logistic regression of Y_i on $H_n(\bar{A}_i, \bar{L}_i, W_i)$ using $\log it(\bar{Q}_n^0(\bar{A}_i, \bar{L}_i, W_i))$ as an offset, and obtaining an MLE estimate of the coefficient ϵ on each clever covariate.
6. Calculate the predicted values for each individual under each treatment regimen using the updated estimate.
7. Update $\Psi(P_0)$ as the empirical mean of the predicted values of Y based on the updated fit, for each treatment regimen of interest.