Dissertation Work Formulation

Dominic DiSanto; Tianxi Cai & Junwei Lu

Department of Biostatistics, Harvard University

August 19, 2024

Today's Goals/Questions

- Review our set-up (observables, data-generating processes)
  - Longitudinal modified treatment policy (LMTP) framework
- Understand most relevant extensions of treatment assignment
- Understand PheWAS-relevant formulation of outcomes in Y

Outline

Set-Up and Observables

- For times $t = 1, \dots, \tau$, observe $Z = (L_1, A_1, \cdots L_\tau, A_\tau, Y) \sim \mathbb{P}$
  - Permit the short-hand $\overline{A}_t = (A_t, \cdots, A_1)$, $\overline{L}_t$ similarly defined, and history as $H_t = (\overline{A}_{t-1}, \overline{L}_t)$
  - Covariates

    $$L_t, \; e.g. \begin{cases} \text{EHR Codes Accumulation or general Hawkes process} \\ \text{Demographic/static information} \end{cases}$$
  - Treatment A_t: binary, multinomial, continuous (e.g. dose)
    - More detailed discussion under Treatment/Policy Evaluation section
  - Outcome Y_t
    - For first discussion, assume univariate Y
    - More detailed discussion under PheWAS section

Data Generating Process & Treatment Mechanism

- Assume generating functions with noise $U = (U_A, U_L, U_Y)$
  - $A_t = f_{A_t}(H_t, U_{A_t})$
  - $L_t = f_{L_t}(A_t, H_t, U_{L_t})$
  - $Y = f_Y(A_\tau, H_\tau, U_Y)$
- Assume a new assignment function $\mathbb{d}(a_t, h_t^{\mathbb{d}}) \rightarrow a_t^{\mathbb{d}}$
- Typical non-parametric SEM set-up (similar, somewhat more general to Aaron Sonabend's SSRL set-up)

Example $\mathbb{d}$ functions I

- Thresholding, e.g. patient's with SBP>130 receive drug H and otherwise drug G (here $SBP = L_t$)
  - $A_t^{\mathbb{d}} = \mathbb{d}(a_t, h_t) = \mathbb{1}(SBP_t > 130)H + \mathbb{1}(SBP_t \leq 130)G$
- Shifting for continuous treatments, e.g. drug dose
  - If $P(A_t < u_t(h_t)|H_t = h_t) = 1$, then

$$\mathbb{d}(a_t, h_t) = \begin{cases} a_t + \delta & a_t \leqslant u_t(h_t) - \delta \\ a_t & a_t > u_t(h_t) - \delta \end{cases}$$

Example $\mathbb{d}$ functions II

- Can induce stochastic interventions using noise $\varepsilon \perp\!\!\!\perp \mathbb{P}$, now $\mathbb{d}(a_t, h_t, \varepsilon)$
-
  - Ex: Shifted/Incremental Propensity Score. For a binary treatment with density $g(a|h_t)$ (and user defined $\delta > 0$),

$$g_t^{\mathbb{d}}(1|h_t) = \frac{\delta g_t(1|h_t)}{\delta g_t(1|h_t) + 1 - g_t(1|h_t)}$$

$$\mathbb{d}(a_t, h_t) = \mathbb{1}(\epsilon_t \leq g_t^{\mathbb{d}}(1|h_t)) \text{ for } \epsilon_t \sim U(0,1)$$

Counterfactuals

- Observe $\overline{A}_t$, with counterfactuals generated by the assignment functions (defined recursively through $t = 1$)
  - Write the counterfactual $A_t(A^{\mathbb{d}}_{t-1}) = f_{A,t}(H_t(\overline{A}^{\mathbb{d}}_{t-1}), U_{A,t})$
    - "What treatment would we observe at A_t, had we applied our intervention policy $\mathbb{d}$ through time $t-1$?"
    - "Natural value of treatment"
  - $L_t(\overline{A}^{\mathbb{d}}_{t-1}) = f_{L,t}(A^{\mathbb{d}}_{t-1}, H^{\mathbb{d}}_{t-1}, U_{L,t})$
  - $Y(\overline{A}^{\mathbb{d}}) = f_Y(\overline{A}^{\mathbb{d}}_{\tau}, H_{\tau}(\overline{A}^{\mathbb{d}}_{\tau-1}), U_Y)$
- Estimand under a given policy $\mathbb{d}$ is $\theta := \mathbb{E}\left[Y(\overline{A}^{\mathbb{d}})\right]$

Extensions/Problem Formulation

- Extend the framework's treatment regime (for a simple/univariate outcome)
  - Via assignment function(s) $\mathbb{d}$
  - Via treatment vector A_t
- For simple treatment/policy comparisons, multivariate or high-dimensional $Y \in \mathbb{R}^d$

Outline

$\mathbb{d}$

- We can alter $\mathbb{d}$
  - Conceptually we are then trying to estimate parameters/functions that describe treatment patterns, while comparing $\mathbb{E}[Y^{\mathbb{d}}]$
  - We can write assignment function as complicated/high-dimensional function
    - $\mathbb{d}(\bar{a}_t, \bar{\ell}_t; \alpha, \beta) = \alpha^T \bar{a}_t + \beta^T \bar{\ell}_t$
    - Non-parametric $\mathbb{d} = m(\bar{a}_t, \bar{\ell}_t)$
- We can extend A_t (e.g. A_t is now multivariate, captures treatment more globally)
  - Conceptually we still have "control" over assignment via $\mathbb{d}$ that we can (and must) specify
  - e.g. high-dimensional, categorical $A \in \mathbb{R}^d$ drug choices

Outline

PheWAS

- Can simplify treatment w/in previous framework
  - i.e. binary treatment A, two policies $\mathbb{d}_0(a_t) = a_t$, $\mathbb{d}_1(a_t) = 1 - a_t, t = \{0, 1\}$
- Now interested in settings with outcome diversity, $Y \in \mathbb{R}^d$
- Sub-group analysis?
  - e.g. In RA, how could we identify sub-types of heart failure (preserved vs reduced ejection fraction) for which anti-TNF is effective?
- Disjoint outcomes?
  - Across $Y \in \mathbb{R}^d-$outcomes (possibly dependent, overlapping), how do we identify the subset of Y (if any) for which drug G has non-0 treatment effect?

Outline

Summary/Questions

- Is this (LMTP) set-up the appropriate framework which we can extend?
- What are the relevant extensions of treatment trajectories via $A_t, \mathbb{d}$?
  - Parameterizing $\mathbb{d}(a_t, h_t)$ as a high-dimensional policy
  - Specifying a form for A_t with simple policy $\mathbb{d}$
- What are the relevant extensions outcomes Y for desired PheWAS studies (and how can we do better than naïve multiple comparisons adjustment)?
  - Sub-group identification?
  - High-/Multi-dimensional $Y \in \mathbb{R}^d$?

Appendix

Outline

Identification

1. Positivity: $(a_t, h_t) \in \text{Supp}(A_t, h_t) \Rightarrow \mathbb{d}(a_t, h_t) \in \text{Supp}, \forall t \in [\tau]$
2. Strong Sequential Randomization:
   $U_{A_t} \perp\!\!\!\perp \underline{U}_{L,t+1}, \underline{U}_{A,t+1} \mid H_t, \ \forall t \in [\tau]$

The Strong Sequential Randomization assumption can be slightly weakened for stochastic interventions

Outline

Rough Timeline of G3 ('24-Spring '25) I

- Continue/complete early work
  - **August** Submit Weijing-Zongqi applied paper
  - **Fall-Winter** Complete draft of optimal assortment work with Junwei
- **October** - Finalize committee
  - Possibly Rajarshi, Sebastien, Nima
- **March** - Oral Examination