Registry Data, Target Trials, and `ltmle`

Zeyi Wang
wangzeyi@berkeley.edu

Division of Biostatistics, UC Berkeley School of Public Health

December 14, 2023

Hypothetical trials and registry data: example of time zero

What is special about registry data?

- Statistical Analysis Plan (SAP) first, or data first?

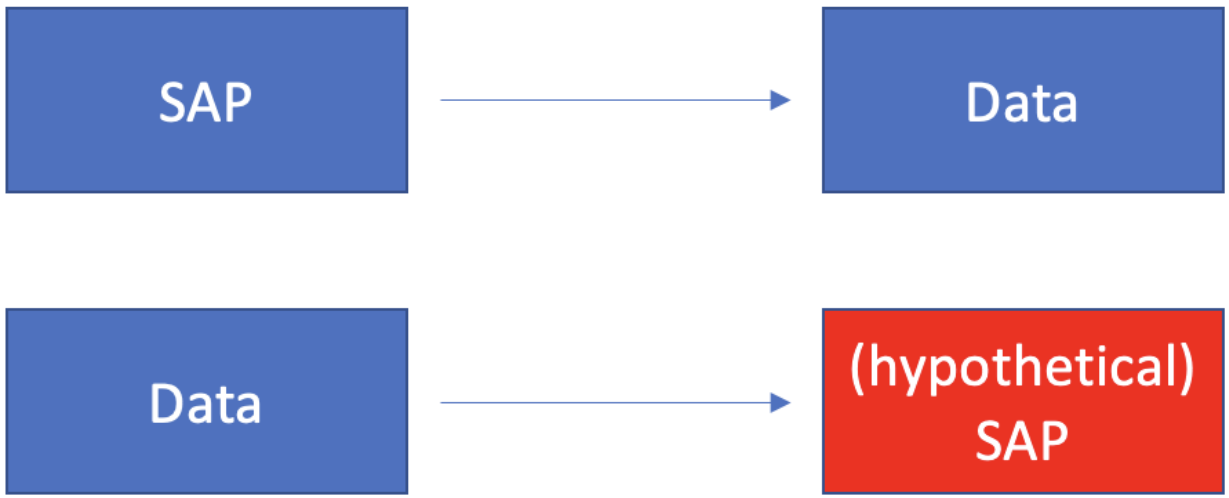


Figure: Randomized trials versus hypothetical trials with registry data.

Why hypothetical trials with registry data?

Pros:

- huge **amount** of data (cost savings),
- answering a **wider range of questions**.

Cons:

- NOT fully randomized, essentially **observational**.

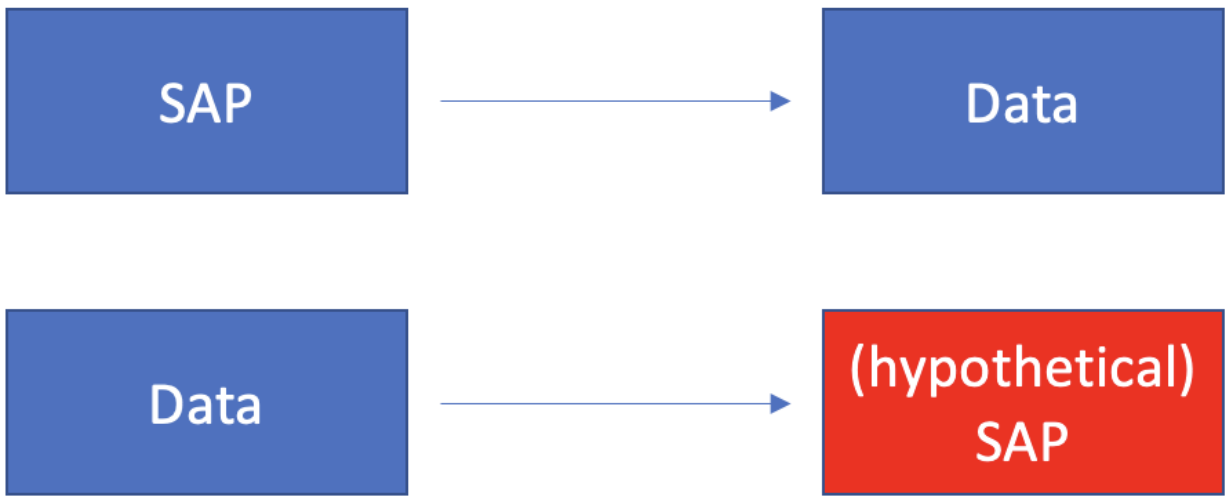
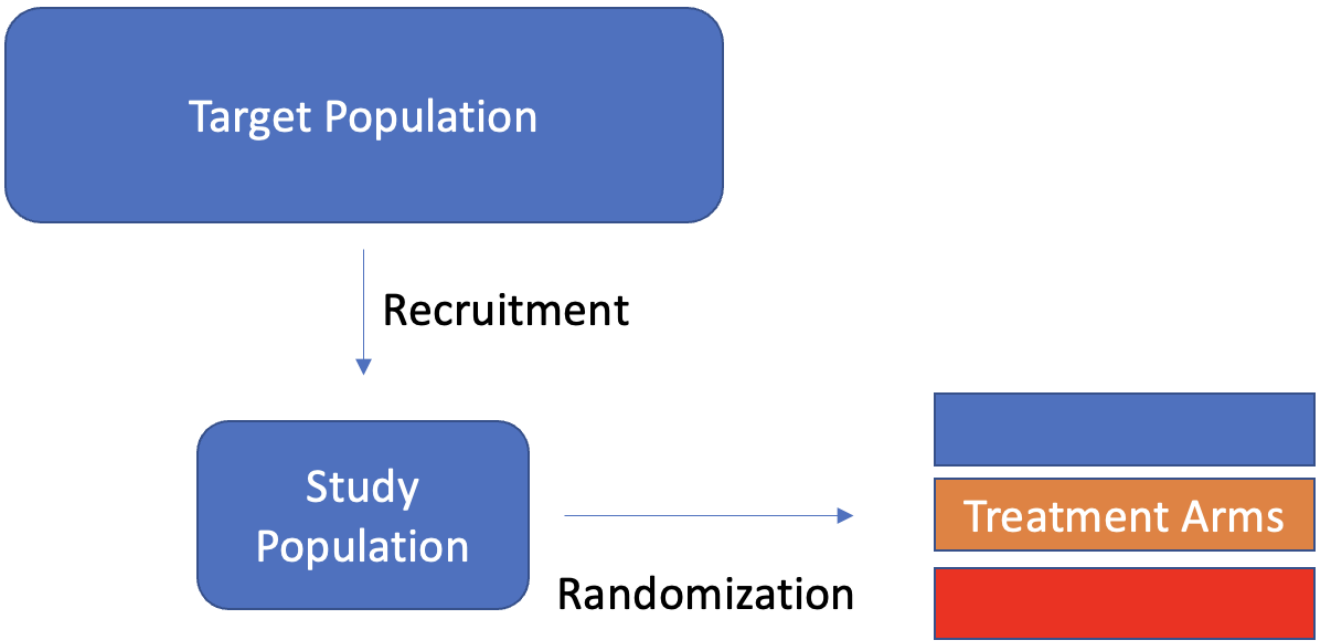
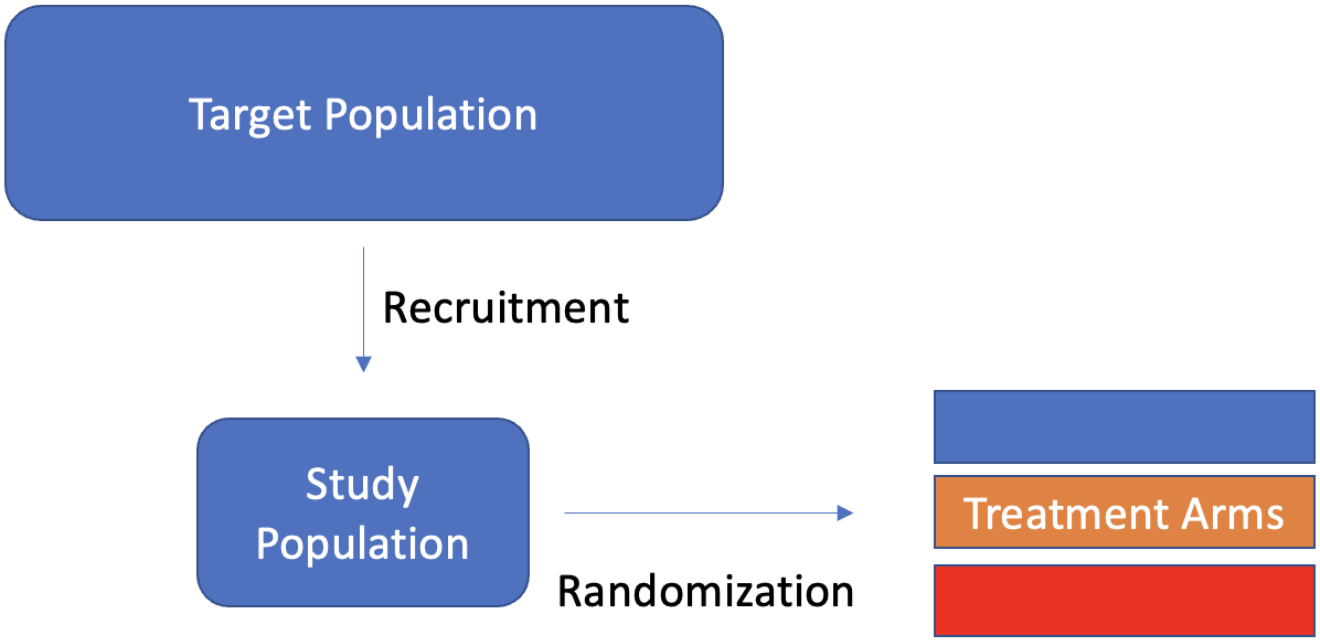


Figure: Randomized trials versus hypothetical trials with registry data.

Can hypothetical trials be as good as randomized trials?



Yes.. in the ideal world?



Ideal trial emulation where a same causal roadmap applies

- if we know the **inclusion-exclusion criteria**,
- if we know the **covariates and probability** (even with no actual randomization).

Hypothetical trial analysis in the real world

Less causal than RCT, but ...

- Possible doubly robust estimation $\rightarrow$ super learning and TMLE.
- Large sample size: e.g. 200k vs 10k.
- More trial options (emulating a real trial, trials with no economic incentives, etc.).

Example:

- GLP-1 Receptor Agonist — Blood Glucose — Cardiovascular Disease.

Why focusing on a target/hypothetical trial

- A target trial is an assumption about **how data would have been generated** (recruitment, randomization, etc.).
  - A target trial may or may not be practical for implementation in the real-world (ethical reasons, economic incentives).
- But a target trial must be specified for evaluating the causal validity.
  - Does the hypothetical recruitment reflect the target population?
  - How likely can observed covariates explain randomness in treatment?
- Once specified, a target trial can be statistically analyzed with a same causal roadmap as if it were implemented in the real world.
  - Except that one is **actually randomized**, one is **hypothesized to follow the same treatment distribution**.
  - This is **as close as we can get to an actual RCT**.

Example: time zero definition with or without a target trial

In longitudinal and survival analysis, subjects are repeatedly measured, where measurements are labeled by a **time** axis.

The sequence of (hypothetically) assigned treatment:

$$A_0, A_1, \ldots, A_K.$$

Time-varying measurements:

$$L_0, L_1, \ldots, L_K.$$

But, what is the **time zero** of a participant?

Example: time zero definition with or without a target trial

Consider a target trial for diabetes research based on registry data from 2009 to 2019. The primary outcome can be HbA1C or death.

Define the time zero of a participant as:

1. the moment of age 60;
2. the moment of 0am on January 1, 2009;
3. the day they starts to take ANY second-line drugs (a predefined set of treatments of interest).

Definition by age 60

What if . . .

- someone is completely healthy at 60 and never takes diabetes drug?

Is it a good control group (no treatment) sample?
What are possible consequences?

Definition by year 2009

What if . . .

- someone had started receiving second-line diabetes treatment, whose blood glocose level was well under control on January 1, 2009?

Is it a good treatment group (actually taking a treatment of intrest) sample?
What are possible consequences?

- When the primary outcome is the change of HbA1C since time zero.
- When the primary outcome is five-year survival.

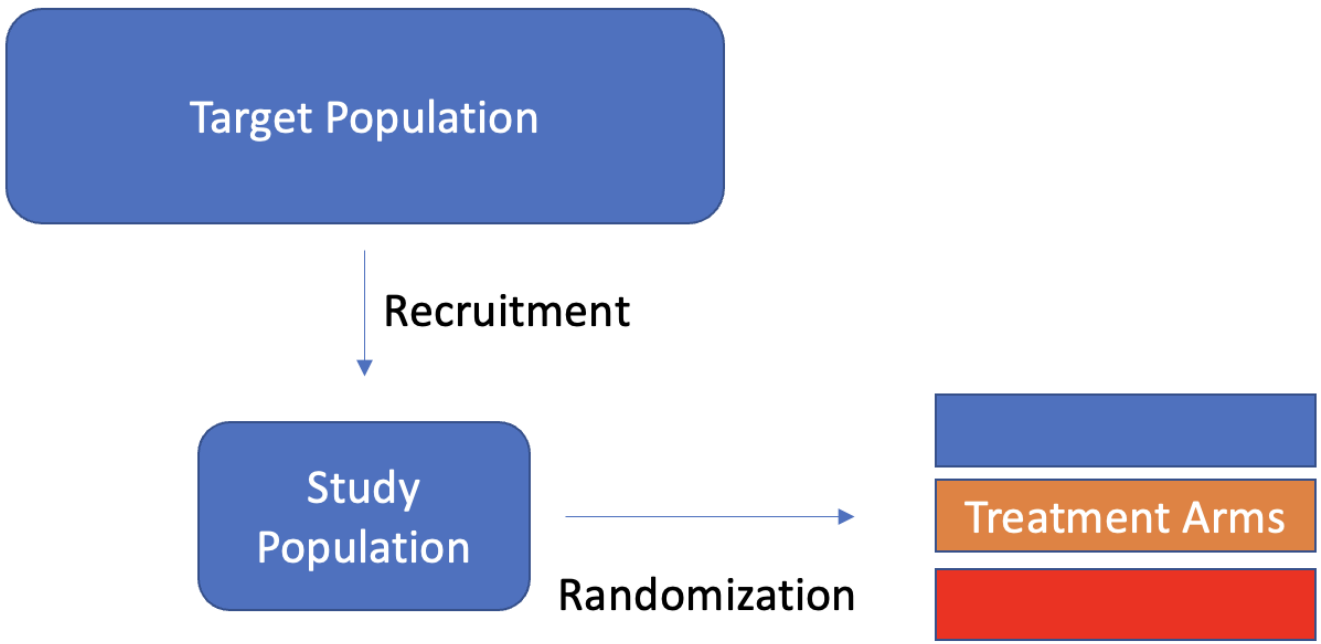
Hint: people usually start taking second-line drugs when the first-line treatment fails to control the blood glucose level.

Problem without a target trial

Without a clearly defined target trial, we don't even have a meaningful definition of data!

Time zero definition in a target trial

How would you define the time zero of a participant in a randomized trial evaluating GLP-1RA vs DPP4 (Dipeptidyl peptidase-4) inhibitors?



When time-zero is defined by a target trial

- Do we have more or less available data?
- Will we get stronger or weaker effects? (Consider both sample size and possible effect dilution.)

Summary: interpret a hypothetical trial

- Under ideal conditions, a hypothetical trial can be as strong as a gold standard, randomized trial.
- Relying on DR, a hypothetical trial can converge to a same scientific target in a RCT - with even larger sample sizes.
- Problems for data analysis without defining a target trial:
  - possibly uninterpretable effects even with perfect model fitting,
  - sometimes can't even properly define data (time zero).

Data defined by a target trial

Suppose that the observed data is

$$O = (L_0, A_1, L_1, \ldots, A_K, L_K).$$

- Baseline covariate L_0 involves treatment and disease history.
- Time-varying covariate L_1 may contain lab tests (later on missing data) and other status trackers.
- Treatment A_t is randomized according to $L_0, A_1, \ldots, L_{t-1}$.
- A treatment arm can characterized by a vector $(a_1, \ldots, a_K)$ such as $(1, 1, \ldots, 1)$ or a sequence of functions $d_t(L_0, A_1, \ldots, L_{t-1})$, $t = 1, \ldots, K$ (dynamic treatment).
- Outcome $Y \in L_K$ can be measured in the end.

Longitudinal data structure and g-computation

Introduction: longitudinal data

Assume a Structural Causal Model (SCM) for $O = (L_0, A_1, L_1, \ldots, A_K, L_K)$:

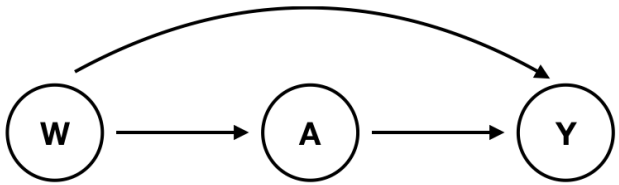
$$\begin{aligned} L_0 &= f_{L_0}(U_{L_0}) \\ A_1 &= f_{A_1}(L_0, U_{A_1}) \\ L_1 &= f_{L_1}(L_0, A_1, U_{L_1}) \\ &\ldots \\ L_K &= f_{L_K}(L_0, A_1, L_1, \ldots, A_K, U_{L_K}). \end{aligned}$$

Suppose we care about an intervention over variables as $A_1 = a_1, \ldots, A_K = a_K$, which defines **counterfactuals**

$$L_k(\bar{a}) = f_{L_k}(L_0, a_1, L_1, \ldots, a_k, U_{L_k}), \text{ for } k = 1, \ldots, K$$

and satisfies certain identification assumptions, then we can identify and analyze targets such as $E[L_K(\bar{a})]$ or $E[Y(\bar{a})]$ with `ltmle`: ...

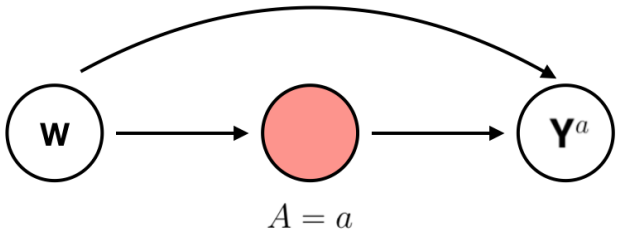
Structural Causal Model (SCM), cross-sectional



$$
\begin{aligned}
W &= f_W(U_W) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}
$$

- U variables are Exogenous, random error;
- f functions are deterministic;

Structural Causal Model (SCM), cross-sectional



$$W = f_W(U_W)$$
$$A = a$$
$$Y(a) = f_Y(W, a, U_Y)$$

- An intervention $A = a$ manipulates endogenous variable A directly;
- other exogenous U variables and deterministic f functions are unchanged;
- this intervention defines counterfactual $Y(a)$.

g-computation, cross-sectional

Identification assumptions:

1. $Y = Y(a)$ when $A = a$ (verified by SCM),
2. $Y(a) \perp A|W$,
   - note that $Y(a)$ is a function of U_Y and U_W through f_Y and f_W.
   - satisfied when W captures all confounders of A and Y so that U_A and U_Y are completely exogenous, independent noises,
   - violated when there exists unmeasured confounder U such that $A = f_A(W, U, U_A)$ and $Y(a) = f_Y(W, U, a, U_Y)$,
3. $P(A = a|W) > 0$ (positivity; will be discussed later).
   - $E(Y|A = a, W)$

Verification of conditions 1 and 2 are conceptual.

g-computation, cross-sectional

When identification assumptions are satisfied, we have **g-computation** formula,

$$E[Y(a)] = E[E(Y|A = a, W)].$$

- $Y(a)$ is **NOT fully observed**;
- $E(Y|A = a, W)$ is **fully observed**; $E[Y(a)]$ is its average.
- Via identification, the analysis of a **causal** target $E[Y(a)]$ becomes fully **statistical**.

ltmle

What does `ltmle` do exactly?

```
ltmle(data, Anodes = "A", Ynodes = "Y", abar = 1)
```

- It fits an outcome model for $E(Y|A = 1, W)$.
- It fits a propensity score model for $E(A|W)$.
  - By default, both models are main-term `glm`, but can be specified as other models, including super learners.
- It finds a targeted update for the outcome model, utilizing the information in the propensity score model.

SCM with longitudinal data

$O = (L_0, A_1, L_1, \ldots, A_K, L_K)$.

Example (Static intervention with longitudinal data, $K = 2$)

$$\begin{aligned} L_0 &= f_{L_0}(U_{L_0}) \\ A_k &= a_k \\ L_k(\bar{a}) &= f_{L_k}(L_0, a_1, L_1, \ldots, a_k, U_{L_k}). \end{aligned}$$

$K = 2$.

Identification assumptions with longitudinal data

1. Consistency, $Y(\bar{a}) = Y$ when $\bar{A} = \bar{a}$ (verified in SCM).
2. Sequential ignorability: $L_s(\bar{a}) \perp A_t | L_0, A_1, \ldots, L_{k-1}$ for $s \geq t$.
   - Again, conceptually verified as "no unmeasured confounding".
3. Positivity: $P_0(A_t = a_t | L_0 = l_0, a_1, \ldots, L_{k-1} = l_{k-1}) > 0$ for all $P_0(L_0 = l_0, A_1 = a_1, \ldots, L_{k-1} = l_{k-1}) > 0$ at the true distribution P_0.

Conditions 1 and 2 are verified conceptually.

g-computation with $K = 2$

Only the g-computation formula becomes more challenging.

$$\begin{aligned} E[Y(\bar{a})] =& E[E[E[Y(\bar{a})|L_1(\bar{a}), L_0]|L_0]] \text{ by iterated conditional expectation} \\ =& E[E[E[Y|A_2 = a_2, L_1, A_1 = a_1, L_0]|A_1 = a_1, L_0]] \end{aligned}$$

- We need a model for $E[Y|A_2 = a_2, L_1, A_1 = a_1, L_0]$. Denote this term as Q_2.
- We need a model for $E[Q_2|A_1 = a_1, L_0]$. Denote this term as Q_1.
- Nested/iterated models:
  - In Q_2, regressers are parents of L_2.
  - In Q_1, regressers are parents of L_1, but the independent variable is Q_2.

What are we modeling in `ltmle`?

- In `ltmle`, by default, nested outcomes models with main-term `glm`.
- These outcome models (Q models) can be super learners, specified by a library of base learners.
- When we specify the regression terms, only need one right after each intervention nodes.

Suppose `L0, A1, L1, Y1, A2, L2, Y2` are the column names. How to add an interaction between `A1` and `L0` for Q_2?

```
Qform=c(L2="Q.kplus1 ~ L0 + A1 + L1 + Y1 + A2 + A1:L0")
```

Note that even the independent variable is Y_2, `Qform` takes `L2` (the one right after `A2`) as the name.

Summary, longitudinal data g-computation formula

- Longitudinal treatment effects have similar SCM and identification assumptions as cross-sectional ones.
- Longitudinal g-computation formula iteratively constructs K **nested** outcome models.
- `ltmle` fit K (nested) outcome models and (NOT nested) K propensity score models, and then find targeted updates for the outcome models.

Censoring, missingness, confounding, competing risks

Survival outcome

Baseline: L_0.

At each (hypothetical) follow-up visit, treatment, covariates, outcome:

$$A_k, L_k, Y_k.$$

What does $\bar{Y} = (Y_1, \ldots, Y_K)$ look like? Suppose $K = 3$, $Y_1 = 0$, $Y_2 = 1$.

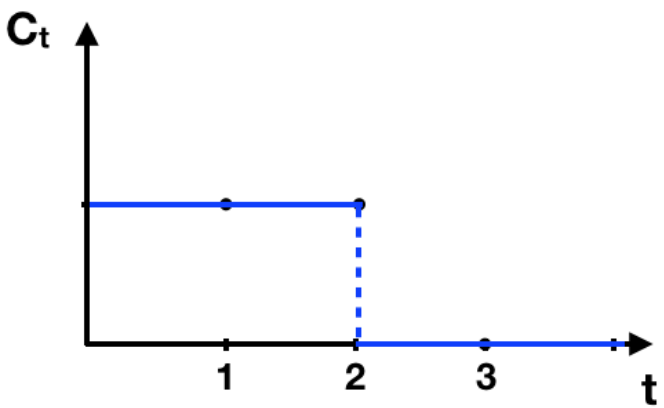
Right-censoring

Suppose one more variable is collected at the beginning of each time point, indicating whether someone remains in the study.

$$C_k, A_k, L_k, Y_k.$$

Example (Administrative censoring)

Suppose a participant join the target trial in 2017, with scheduled yearly visits. However, the registry data stops at year 2019.



Target parameter with censoring

Data censoring/missing is common even in the analysis of actual trials in the real world (end of study, drop-outs, etc.).

Analyzing **treatment effects among uncensored subjects** is essentially defining a hypothetical trial that combine the actual trial with an additional intervention on censoring indicators.

Example (Average treatment effect among the uncensored)

$E[Y(\text{treated, uncensored}) - Y(\text{untreated, uncensored})]$

Missingness of covariates

Example (Missingness in lab testing)

Suppose that someone joined the target trial in 2010. Everyone is invited to (hypothetical) lab tests of blood samples. This subject missed the tests in 2011 but not thereafter.

- This is not right-censoring.
- One may assume that a bivariate summary is enough.
  - At each time point t, we only have access to the last available observation and if the observation is updated.
  - But one may **assume** (!!) that in combination with other observed confounders, such information (whether it is carried forward or actually observed) can fully capture confounding.
- Subsetting can change the target trial (!!).
  - People who frequently visit hospitals by and large may be different from the target population.

Conceptual evaluation: no unmeasured confounding?

Suppose that L_t involves

1. a last-observation-carried-forward blood glucose level,
2. an indicator whether it is actually tested at time point t.

When does this L_t make sense?

Conceptual evaluation: no unmeasured confounding?

Suppose that L_t involves

1. a last-observation-carried-forward blood glucose level,
2. an indicator whether it is actually tested at time point t.

When does this L_t make sense?

- Until someone goes to a hospital and get a formal lab testing, the drug prescription will be unchanged.
- When some gets a formal lab testing, the drug prescription may change, depending on the actual test results.

Conceptual evaluation: no unmeasured confounding?

Suppose that L_t involves

1. a last-observation-carried-forward blood glucose level,
2. an indicator whether it is actually tested at time point t.

When does this L_t make sense?

- Until someone goes to a hospital and get a formal lab testing, the drug prescription will be unchanged.
- When some gets a formal lab testing, the drug prescription may change, depending on the actual test results.

Decide if this is the case in your data.

- **Statistical analysis** can always be conducted with software; however,
- **causal validity** depends on plausibility of the assumptions.

Competing risks

Example: disease and death.

- Both are survival outcomes.
  - "Have you ever been diagnosed . . . ?" will be constant `yes` after the first diagnosis;
- The first diagnosis and death cannot (?) happen at a same moment.
  - In the target trial, we can stop follow-up after either the first diagnosis or death.
  - Need to check the exact time stamp.

Can we treat competing risks as censoring?

The implication of censoring and competing risks

Right-censoring (effect among uncensored) is part of intervention.

Can death-free be part of intervention?

- Consider: a sub-population that will never die before getting a disease.
- Everyone will get this disease, considering everyone will die eventually. This is a population that wait for their lives to get a disease.
- Death-free population can have a **higher risk** of getting a disease, compared with the target population.

Properly handling competing risks

Competing risks are usually NOT intervention and should be part of time-varying observation.

- A competing risk is a knowledge about data.
- Handling competing risks means to respect this knowledge in the outcome models.
- `ltmle` offers `deterministic.Q.function` argument (need to be customized) to incorporate such knowledge. See `https://github.com/joshuaschwab/ltmle/blob/master/vignettes/articles/a04_deterministic-functions.Rmd`.

Summary of a typical survival analysis data

At baseline:

$$L_0.$$

At the k-th follow-up:

$$C_k, A_k, L_k, Y_k.$$

In `ltmle`, column numbers are needed for:

- `Cnodes`,
- `Anodes`,
- `Lnodes`,
- `Ynodes`.

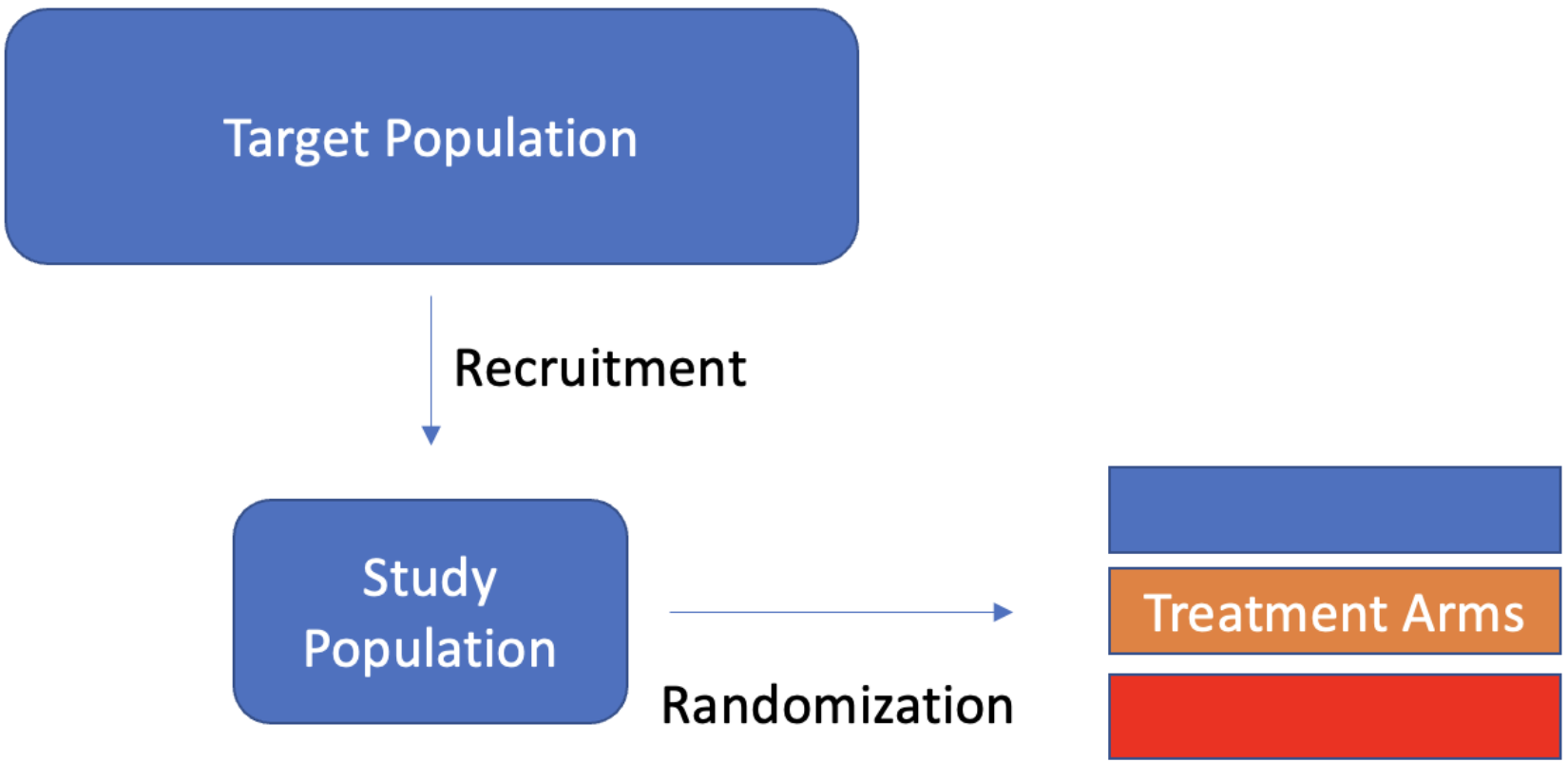
And `survivalOutcome = TRUE`.

Time discretization and positivity

Recall: define target population first

Even for a (fake) recruitment of a hypothetical trial, target population needs to be specified.

$t = 0$ should be the date of (fake) recruitment and treatment allocation as specified in the hypothetical SAP.



Recall: time zero

Example

Consider all type 2 diebetes patients (defined by previous usage of first-line drugs such as metformin) in a healthcare record database who started second-line drugs (different brands/types) between 2010-2019.

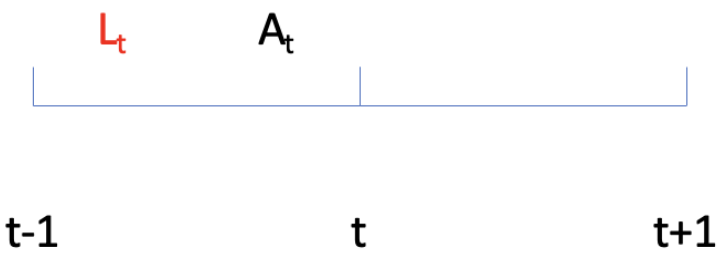
- Suitable for analysis and comparisons of second-line drugs on the market between 2010-2019 without the impact of COVID.
- $t = 0$ is the date of the earliest second-line drug subscription record.
- Randomness in treatment types might be explained by a list of confounders (pre-conditions, education, income, etc.).

Example: time discretization

But registry data are collected in **continuous time** (sometimes with masked time stamps), while hypothetical trials have scheduled visits (**discrete time**).

Example (Unintended post-randomization confounders.)

Suppose that L_t, A_t are collected between $(t-1, t]$.
An online tutorial uses ordering $L_0, A_0, L_1, A_1, \ldots, L_{K+1}$.

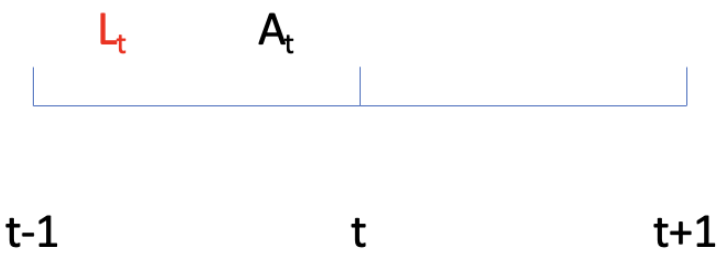


Example: time discretization

But registry data are collected in **continuous time** (sometimes with masked time stamps), while hypothetical trials have scheduled visits (**discrete time**).

Example (Unintended post-randomization confounders.)

Suppose that L_t, A_t are collected between $(t-1, t]$.
An online tutorial uses ordering $L_0, A_0, L_1, A_1, \ldots, L_{K+1}$.



Possible issues?

Issues caused by careless time discretization

- Effects may be wiped out due to post-randomization adjustment.
- Better the confounder L_t, larger the bias! Eg. hypertension and head pain, prescription change and related conditions.

Confounders of A_t must have a time stamp strictly earlier to treatment.

Principle: variable ordering and discretized time

Need to decide between:

- a conservative adjustment possibly biased due to unmeasured confounding;
- a careless adjustment with invalid causal interpretation.

Considerations

- Only adjust for confounders that are guaranteed to be pre-randomization.
- For longitudinal treatment, randomization can happen at multiple time points, and this applies to each time point separately.
- Evaluate the chance that the final confounder list fully explains randomness in treatment allocation.

Different types of time-discretized variables

- Status (treatment/censoring/confounder) at the beginning/end of an interval;
- Summary of measurements collected across an interval;
- Incidence count across an interval;
- Indicator that an event has happens within an interval.

Note that **status** may be a summary of a "block".

If A_t is actually defined by a "block" of time, any confounder definition cannot overlap in the time stamps!

Width of time bin

- Will lose information for wider intervals.
- No need to make the finest time grid either — can cause trouble.
- No need to make a time grid finer than how frequent variables can change.

Remember: SAP is reversible.

- Check with the real data before designing/refining a target trial...
- But it has to be **outcome-blind**!

Positivity violation

Positivity: $P_0(A_t = a_t | L_0 = l_0, a_1, \ldots, L_{k-1} = l_{k-1}) > 0$ for all possible values of $\bar{l}$ where $P_0(L_0 = l_0, A_1 = a_1, \ldots, L_{k-1} = l_{k-1}) > 0$.

- Following $\bar{A}_{t-1} = \bar{a}_{t-1}$, for all possible observations prior to A_t, there must be a non-zero propensity score.

Adjust target trial design because of positivity violation

Example (Off-label use.)

An approved anti-diabetic drug was unofficially used for preventing dementia. Patients were in doubt of the effect; none of them regularly prescribed the drug throughout the whole follow-up period; majority of the users only received one prescription.

- An "always-on" treatment arm would have almost no support in the data.
- Outcome-blind simulation is useful for detecting problems at an early stage and creating a chance to update (reversible) SAP.
- Data only supports a reasonable length of follow-up.

Challenges of outcome-blind simulation under positivity violation

- Replacing outcomes with completely random noise may hide positivity violation.
  - Consider regenerate outcomes with fitted models, but only use those models in positivity problems.
  - At least check with completely outcome-blind simulation.
- Finite sample positivity is dependent on estimation algorithms (e.g. an extremely sparse lasso selects one covariate).
  - In outcome-blind simulations, choose algorithms similar to the actual analysis.
  - Still need to conceptually verify the assumption.

Collecting diagnostics

- `ltmle` object returns `cum.g` which is by default bounded above 0.01.
- `cum.g.unbounded` can also be used to evaluate the finite sample violation of positivity assumptions.

Summary of a target trial analysis

- Always generate time-discretized data according to a target trial.
- The validity of causal interpretation conceptually depends on identification assumptions. When identification assumptions are satisfied, a target/hypothetical trial analysis is as close as one can get with registry data to an expensive, gold-standard RCT.
- `ltmle` integrates super learning and TMLE; following the same causal roadmap for actually randomized data, it applies to longitudinal and survival data generated by a target trial, for DR estimation.
- Outcome-blind simulation is useful for detecting and handling positivity issues caused by target trial design.

Thank You

Thank you!

Questions?

Happy to discuss more!

Email: wangzeyi@berkeley.edu