Longitudinal analysis of Heterogeneous Treatment Effects in Danish Register Data

Thomas Alexander Gerds

Section of Biostatistics, University of Copenhagen

Motivation

Example 1

- Data/setting: Patients with monopolar and bipolar depression are identified via Danish nation-wide population-based longitudinal register linkage
- We analyse study the comparative effectiveness of antidepressants and anticonvulsants.

Example 2

- Data/setting: Patients with diabetes are identified in the Danish nation-wide population-based longitudinal register data.
- We analyse the effects of polypharmacy on a multivariate outcome.

Methods

Trial emulation and cluster randomized trials are relatively new concepts which have received a lot of attention in the epidemiological literature recently.

In Danish register data they allows us to

- estimate treatment effects beyond intention-to-treat
- study populations that would normally not be enrolled in clinical trials
- estimate treatment heterogeneity (causal prediction)

The longitudinal causal inference for electronic health records can be used to analyse the data, but is

- not fully developed (lack of methods)
- not implemented (lack of software)
- not broadly accessible (lack of know-how and experience)

Avoidable flaws in observational analyses[1]

The increasing availability of large healthcare databases is fueling an intense debate on whether real-world data should play a role in the assessment of the benefit-risk of medical treatments.

In many observational studies, for example, statin users were found to have a substantially lower risk of cancer than in meta-analyses of randomized trials. While such discrepancies are often attributed to a lack of randomization in the observational studies, they may be explained by flaws that can be avoided by

explicitly emulating a target trial.

[1] Dickerman, . . . , Hernan. Nature medicine, 25(10):1601–1606, 2019.

Emulating a target trial in register data[2]

Step 1 enrollment date (time zero for survival analysis, inclusion/exclusion)

Step 2 the target trial protocols dictate the treatment at any time during followup

Step 3 define the target parameters as contrasts of the expected outcomes – had all subjects ~~been randomized~~ adhered to the protocols

Step 4 estimation of target parameters

Step 5 communication of results and limitations

[2] Following the roadmap of targeted learning (van der Laan et al.)

The register data look like this:

id	sex	age at enrollment	date of enrollment
1	woman	74	2019-07-02
2	man	69	2016-01-13
3	woman	88	2021-11-27

Treatment:

id	date	drug
1	2019-07-02	GLP1-RA
1	2019-10-08	GLP1-RA
1	2021-01-07	GLP1-RA
2	2016-01-13	SGLT2i
2	2016-08-27	SGLT2i
3	2021-11-28	GLP1-RA
3	2021-12-03	SGLT2i

Hospital diagnoses:

id	hospital admission date	diagnosis
2	2018-02-15	hypertension
2	2018-09-03	hypertension
3	2022-01-02	arterial fibrillation

Concomitant medical treatment:

id	redemption date	drug
2	2018-12-24	ACE/ARP
3	2022-02-11	NOAC
3	2022-08-10	NOAC

Outcome

id	cardiovascular disease	date of death	end of followup
1	NA	NA	2024-09-17
2	NA	2019-03-03	2024-09-17
3	2023-02-07	2023-02-07	2024-09-17

Emulating trials in register data

In register studies there is often no natural control group of patients who are not exposed to one of the drugs.

- Matching can be used (e.g., exposure density sampling).
- A medical diagnosis (biomarker above threshold) can be used.
- Comparative effectiveness study: active comparator arms.

Emulating trials in register data

In register studies there is often no natural control group of patients who are not exposed to one of the drugs.

- Matching can be used (e.g., exposure density sampling).
- A medical diagnosis (biomarker above threshold) can be used.
- Comparative effectiveness study: active comparator arms.

Example: We include all diabetes patients in Denmark who initiated medical treatment with one of the following anti-diabetic drugs: GLP1-RA, SGLT2i between 2015 and 2022:

- Time zero is at (or 30 days after) the first purchase of the drug
- We follow the patients through the registers until comorbidity, death, emigration, or 2024, whatever comes first.

Treatment assignment in the target trial

The protocols of the target trial dictate the treatment(s) given at any time (doctor visit) during the target trial period.

Notation:

$$\pi^*(t \mid L(t-), A(t-))$$

Examples:

Protocol	Type of intervention	$\pi^*(t \mid L(t-), A(t-))$
Never treat	Static	1
Always treat	Static	0
Treat for 2 years	Static	$1\{t \leq 2\}$
Treat if $L(t-) > \xi$	Dynamic	$1\{L(t-) > \xi\}$
Treat with probability 0.8	Stochastic	0.8
If $L(t-) > \xi$ treat with probability 0.8	Stochastic, Dynamic	$1\{L(t-) > \xi\}0.8$

Treatment assignment in the target trial

Example of treatment regimens:

- **Protocol 1** Patients should use GLP1-RA continuously for 3 years and not intensify with SGLT2i
- **Protocol 2** Patients should use SGLT2i continuously for 3 years and not intensify with GLP1-RA

Protocol 1 assigns 100% probability for GLP1-RA and 0% probability for SGLT2i:

$$\pi^{1*}(t) = (1, 0).$$

Protocol 2 is defined similarly:

$$\pi^{2*}(t) = (0, 1)$$

The target parameter (aka the estimand)

The analysis of the emulated target trial estimates the absolute risks of the outcome(s) if hypothetically all patients had followed the treatment protocols (per-protocol effects).

Example: 3-year risk of cardiovascular disease under π^{j*} and differences thereof:

$$\mathrm{P}_{\pi^{1*}}(Y(3) = 1) - \mathrm{P}_{\pi^{2*}}(Y(3) = 1)$$

The analyst uses the information of the time-varying covariates (comorbidity, co-medicine) to achieve a good compromise between:

- the available data
- the desire of the investigators
- the causal assumptions: *positivity*, *sequential coarsening at random (NUC)*, *consistency*.

The register data forced onto a discretized time scale

```
       id    sex   age hypertension_0  af_0 GLP1RA_0 SGLT2i_0 Censored_1 cvd_death_1 Dead_
   <char> <char> <num>          <num> <num>    <num>    <num>     <fctr>       <num>  <nu
1:      1  woman    74              0     0        1        0 uncensored           0
2:      2    man    69              0     0        0        1 uncensored           0
3:      3  woman    88              0     0        0        0 uncensored           0
   hypertension_1  af_1 GLP1RA_1 SGLT2i_1 Censored_2 cvd_death_2 Dead_2 hypertension_2  af
            <num> <num>    <num>    <num>     <fctr>       <num>  <num>          <num> <nu
1:              0     0        1        0 uncensored           0      0              0
2:              0     0        0        0 uncensored           0      0              0
3:              0     1        0        1 uncensored           0      0              0
   GLP1RA_2 SGLT2i_2 Censored_3 cvd_death_3 Dead_3 hypertension_3  af_3 GLP1RA_3 SGLT2i_3
      <num>    <num>     <fctr>       <num>  <num>          <num> <num>    <num>    <num>
1:        0        0 uncensored           0      0              0     0        0        0
2:        0        0 uncensored           0      0              0     0        0        0
3:        0        0 uncensored           0      1             NA    NA       NA       NA
   Censored_4 cvd_death_4 Dead_4 hypertension_4  af_4 GLP1RA_4 SGLT2i_4 Censored_5 cvd_dea
       <fctr>       <num>  <num>          <num> <num>    <num>    <num>     <fctr>       <
1: uncensored           0      0              0     0        1        0 uncensored
2: uncensored           0      0              0     0        0        0 uncensored
3:       <NA>           0     NA             NA    NA       NA       NA       <NA>
   Dead_5 hypertension_5  af_5 GLP1RA_5 SGLT2i_5 Censored_6 cvd_death_6
    <num>          <num> <num>    <num>    <num>     <fctr>       <num>
1:     0              0     0        1        0 uncensored           0
2:     0              1     0        0        0 uncensored           0
3:    NA             NA    NA       NA       NA       <NA>           0
```

Longitudinal setting [3]

Discretized time scale:

$$[0 \cdots\cdots\cdots t_1 \cdots\cdots\cdots t_2]$$

Data for two time intervals:

$$X = (L_0, A_0, Y_1, L_1, A_1, Y_2).$$

The joint probability distribution:

$$\mathrm{P}X = \mathrm{P}_{Y_2|A_1,L_1,Y_1,A_0,L_0}\mathrm{P}_{A_1|L_1,Y_1,A_0,L_0}\mathrm{P}_{L_1|Y_1,A_0,L_0}\mathrm{P}_{Y_1|A_0,L_0}\mathrm{P}_{A_0|L_0}\mathrm{P}_{L_0}$$

[3] no censoring, no competing risks, $Y_1 = 1\{T \leq t_1\}$, $Y_2 = 1\{T \leq t_2\}$

Longitudinal setting [3]

Discretized time scale:

$$[0 \cdots\cdots\cdots t_1 \cdots\cdots\cdots t_2]$$

Data for two time intervals:

$$X = (L_0, A_0, Y_1, L_1, A_1, Y_2).$$

The joint probability distribution:

$$\mathrm{P}X = \underbrace{\mathrm{P}_{Y_2|A_1,L_1,Y_1,A_0,L_0}}_{F_2}\underbrace{\mathrm{P}_{A_1|L_1,Y_1,A_0,L_0}}_{\pi_1}\underbrace{\mathrm{P}_{L_1|Y_1,A_0,L_0}}_{H_1}\underbrace{\mathrm{P}_{Y_1|A_0,L_0}}_{F_1}\underbrace{\mathrm{P}_{A_0|L_0}}_{\pi_0}\underbrace{\mathrm{P}_{L_0}}_{H_0}$$

[3] no censoring, no competing risks, $Y_1 = 1\{T \leq t_1\}$, $Y_2 = 1\{T \leq t_2\}$

Longitudinal causal inference on discretized time scale

Uncensored data, two intervals [4]

$$X = (L_0, A_0, Y_1, L_1, A_1, Y_2)$$

Observed likelihood

$$\mathrm{P}X = \underbrace{\mathrm{P}_{Y_2|A_1,L_1,Y_1,A_0,L_0}}_{F_2}\underbrace{\mathrm{P}_{A_1|L_1,Y_1,A_0,L_0}}_{\pi_1}\underbrace{\mathrm{P}_{L_1|Y_1,A_0,L_0}}_{H_1}\underbrace{\mathrm{P}_{Y_1|A_0,L_0}}_{F_1}\underbrace{\mathrm{P}_{A_0|L_0}}_{\pi_0}\underbrace{\mathrm{P}_{L_0}}_{H_0}$$

Likelihood in the target trial

$$\mathrm{P}^* = F_2\,\pi_1^*\,H_1\,F_1\,\pi_0^*\,H_0$$

[4] Changes of A(t) and L(t) in last interval $(t_{1,t2}]$ are ignored.

Robins g-methods

Estimator 1: g-formula

$$\hat{\mathrm{P}}^*_g = \hat{F}_2 \pi^*_1 \hat{H}_1 \hat{F}_1 \pi^*_0 \hat{H}_0$$

Estimator 2: Inverse probability weighting

$$\hat{\mathrm{P}}^*_{\mathsf{IPTW}} = \frac{\mathbb{P}_n \, \pi^*_0 \pi^*_1}{\hat{\pi}_0 \hat{\pi}_1}$$

Robins g-methods

Estimator 1: g-formula

$$\hat{\mathrm{P}}^*_g = \hat{F}_2 \pi^*_1 \hat{H}_1 \hat{F}_1 \pi^*_0 \hat{H}_0$$

Estimator 2: Inverse probability weighting

$$\hat{\mathrm{P}}^*_{\mathrm{IPTW}} = \frac{\mathbb{P}_n \, \pi^*_0 \pi^*_1}{\hat{\pi}_0 \hat{\pi}_1}$$

Estimator 3: Iterative conditional expectations AKA Sequential regression

$$\begin{aligned}
\mathrm{E}[Y_2] &= \mathrm{E}[\mathrm{E}[Y_2|L_0]] \\
&= \mathrm{E}[\mathrm{E}[\mathrm{E}[Y_2|L_0, A_0]|L_0]] \\
&= \mathrm{E}[\mathrm{E}[\mathrm{E}[\mathrm{E}[Y_2|L_0, A_0, L_1]|L_0, A_0]|L_0]] \\
&= \mathrm{E}[\mathrm{E}[\mathrm{E}[\mathrm{E}[\mathrm{E}[Y_2|L_0, A_0, L_1, A_1]|L_0, A_0, L_1]|L_0, A_0]|L_0]]
\end{aligned}$$

Iterative conditional expectations: discretized time

Estimator 3: Robins (1999), Bang & Robins (2005)

$$\mathrm{E}[Y_2] = \mathrm{E}[\mathrm{E}[\mathrm{E}[\mathrm{E}[\underbrace{\mathrm{E}[Y_2|L_0, A_0, L_1, A_1]}_{Q_2(L_0,A_0,L_1,A_1)}|L_0, A_0, L_1]|L_0, A_0]\,|L_0]]$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{Q_1(L_0,A_0)}$$

Step 1 $\hat{Q}_2$: Regress Y_2 on L_0, A_0, L_1, A_1

Step 2 Integrate $\hat{Q}_2$ with respect to π_1^*

Step 3 $\hat{Q}_1$: Regress result of Step 2 on L_0, A_0, L_1, A_1

Step 4 Integrate $\hat{Q}_1$ with respect to π_0^*

Step 5 Average with respect to $\hat{H}_0$

Longitudinal targeted minimum loss based estimator

A motivation for the roadmap of targeted learning is the problem that the various nuisance parameter regression models could be misspecified.

The targeted minimum loss based estimator can be consistent even if some of the nuisance parameter models are misspecified.

For longitudinal data analysis, the sequential regression estimator is moved closer to the unknown truth by sequential updating with

"clever covariates"

which depend on the inverse propensity of treatment weights.[5]

[5] van der Laan & Gruber (2012) The international journal of biostatistics

Longitudinal targeted minimum loss based estimator

Under the usual causal assumptions and if the convergence rate of the estimators of the nuisance parameters is sufficiently fast we can estimate the target parameter:

$$\psi : \mathcal{M} \mapsto \mathbb{R},$$

where ψ is a suitably smooth functional defined on a set of probability measures, at the $\sqrt{n}$-rate

$$\sqrt{n}(\psi(\hat{\mathrm{P}}^*_{\mathsf{LTMLE}}) - \psi(\mathrm{P})) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathsf{IF}_{\mathrm{P}}(X_i) + o_{\mathrm{P}}(1)$$

Longitudinal targeted minimum loss based estimator

The efficient influence function for the target parameter (two intervals):

$$\mathrm{IF}_{\mathrm{P}}(X) = (Y_2 - Q_2)\frac{\pi_1^*\pi_0^*}{\pi_1\pi_0} + (Q_2 - Q_1)\frac{\pi_0^*}{\pi_0} + Q_1 - \psi(\mathrm{P})$$

Targeting algorithm:

- Step 1 Initial estimators: $\hat{\pi}_1, \hat{\pi}_0$
- Step 2 $\hat{Q}_2$: Regress Y_2 on L_0, A_0, L_1, A_1
- Step 3 $\hat{Q}_2^*$: TMLE update: Loss function and parametric fluctuation model to solve the current part of the efficient influence function
- Step 4 Integrate $\hat{Q}_2^*$ with respect to π_1^*
- Step 5 Regress result of Step 4 on A_0, L_0
- Step 6 ...

Outlook

We will develop new architectures for semiparametric causal inference of emulated target trials.

We will:

- study block randomization designs and enroll patients into series emulated trials.
- design estimands and predictimands in collaboration with subject matter partners
- derive methods for propensity scores that acknowledge the multivariate treatment options
- develop new super learner algorithms for the longitudinal structure of the data
- work out graphical representations of treatment effects that ease patients decision making with personalized predictions of multivariate outcomes