Using sequential target trials to estimate causal effects of treatments from longitudinal observational data

Ruth Keogh

Department of Medical Statistics
London School of Hygiene & Tropical Medicine

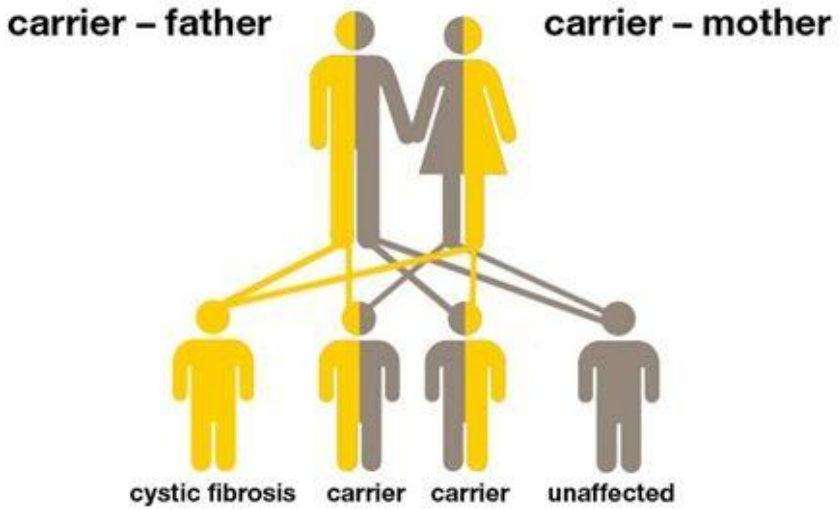
Collaborators

- Jon Michael Gran (University of Oslo)
- Shaun Seaman (MRC Biostatistics Unit, University of Cambridge)
- Stijn Vansteelandt (Ghent University, LSHTM)
- Gwyneth Davies (UCL)
- Siobhan Carr (Royal Brompton Hospital, Imperial College)

Estimating effects of treatments used in cystic fibrosis

Cystic Fibrosis (CF)

- An inherited, chronic, progressive condition
- Affects ~10,000 people in the UK
- Estimated median survival age in the UK is 47
- Key question: What are the effective ways of simplifying the treatment burden of people with CF?



Rowbotham et al. The top 10 research priorities in cystic fibrosis developed by a partnership between people with CF and healthcare providers. Thorax 2018; 73: 388-390.

UK Cystic Fibrosis Registry

- Secure centralized database managed by the Cystic Fibrosis Trust
- Longitudinal data obtained at annual visits
  - clinical measurements
  - treatments used
  - infections
  - hospital stays
- Data on >99% of individuals with CF

Taylor-Robinson et al. Data Resource Profile: The UK Cystic Fibrosis Registry. *Int J Epidemiol* 2017; 47: 9-10e

Treatment: Dornase alfa (Dnase)

- Dnase helps to break down the sticky mucus that builds up in the airways
- Used by about 60% of CF patients in the UK
- Randomized trials:
  - Quite a lot
  - Tend to be short term
  - Restricted to subsets of the patients population

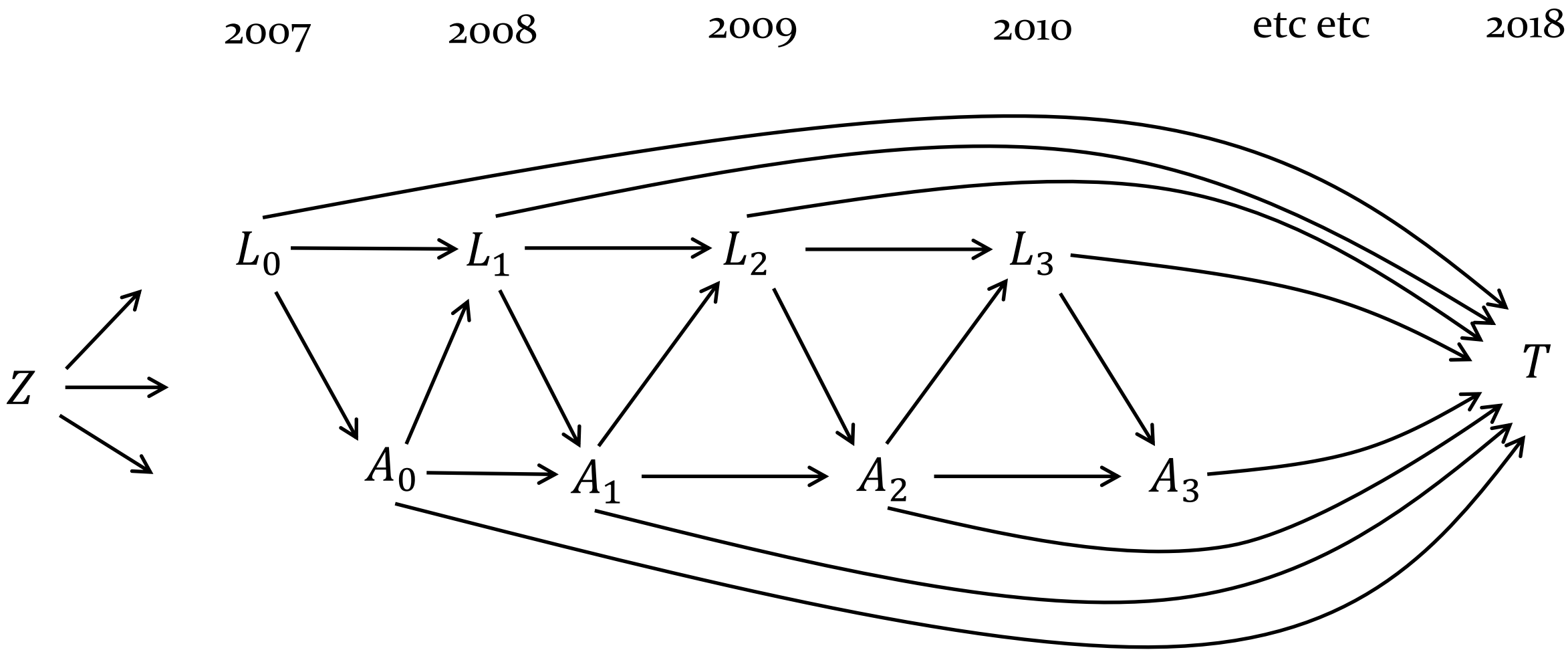
Newsome SJ, Keogh RH, Daniel RM. Estimating long-term treatment effects in observational data: a comparison of the performance of different methods under real-world uncertainty. *Statistics in Medicine* 2018;37:2367-2390.

Newsome SJ, Daniel RM, Carr S, Bilton D, Keogh RH. Investigating the effects of long-term dornase alfa use on lung function using registry data. *Journal of Cystic Fibrosis* 2018.

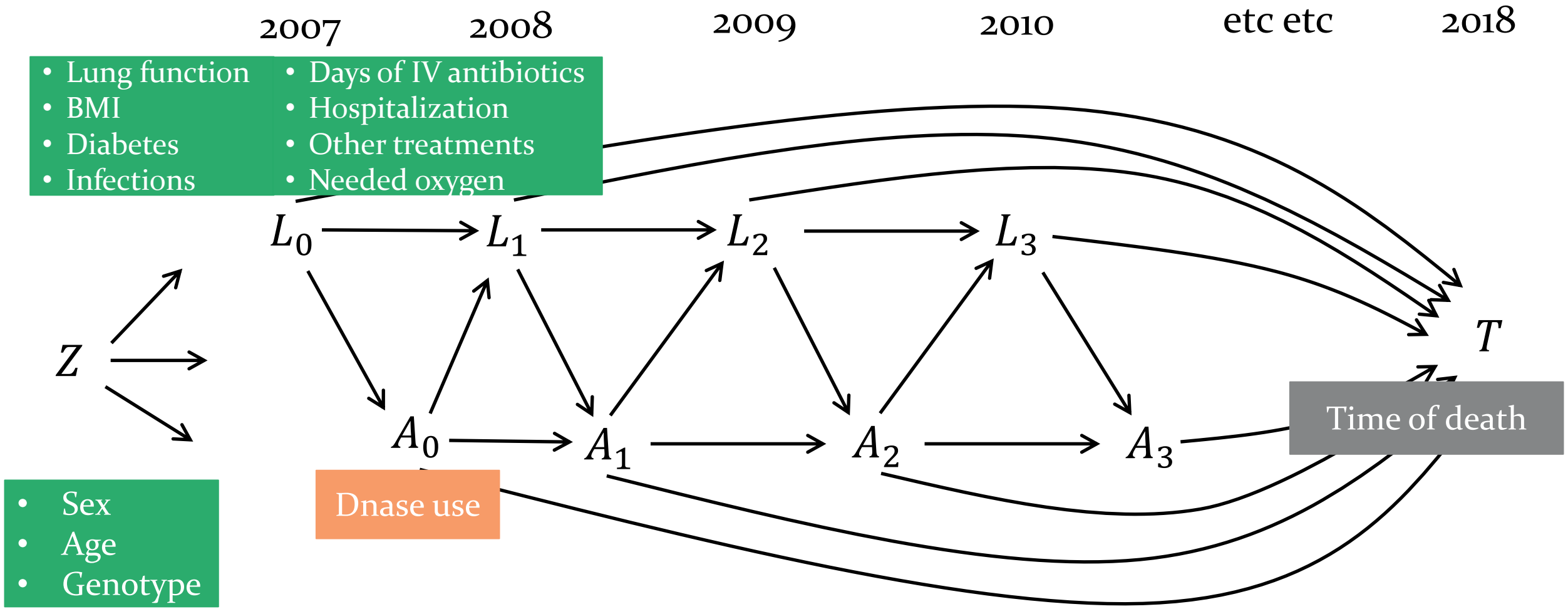
Aims

- What is the impact of the treatment Dnase use on survival?
- How can we go about trying to answer this question using the UK CF Registry?

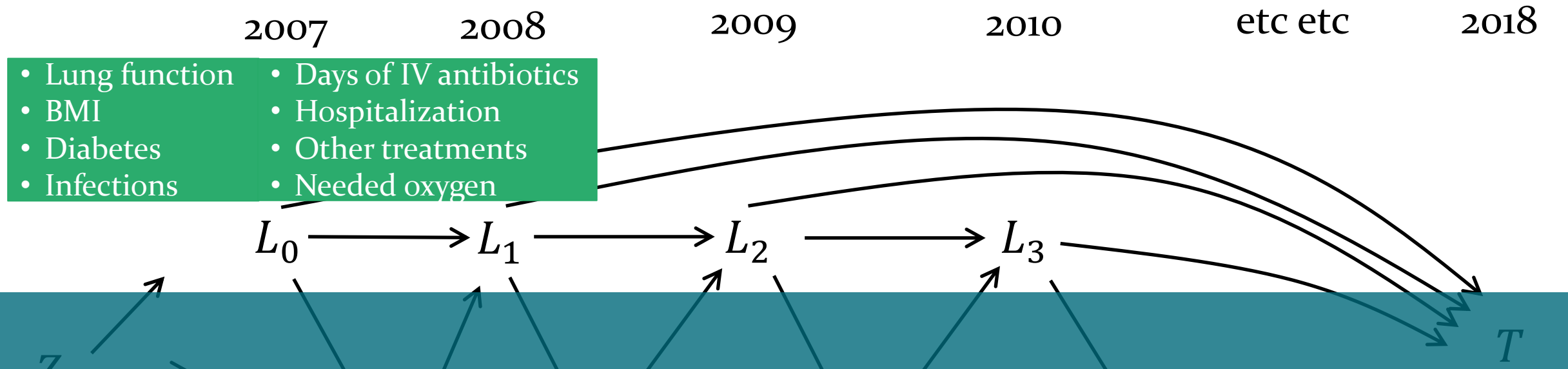
UK CF Registry setting



UK CF Registry setting



UK CF Registry setting

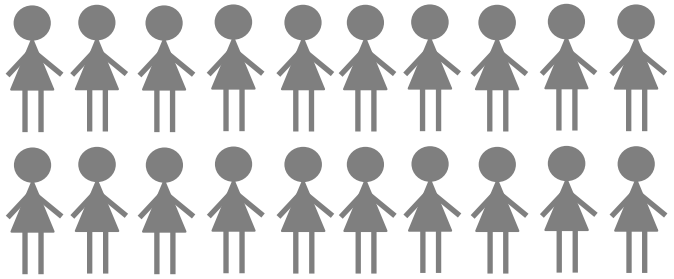


What would be the difference in survival up to t years if everyone received treatment (Dnase) compared with if everyone did not receive treatment?

How can we answer this question?

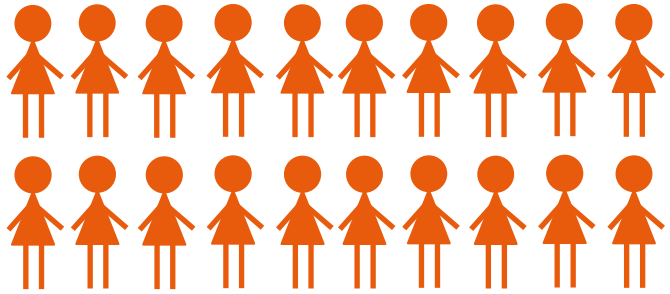
What are we trying to estimate?

Eligible individuals

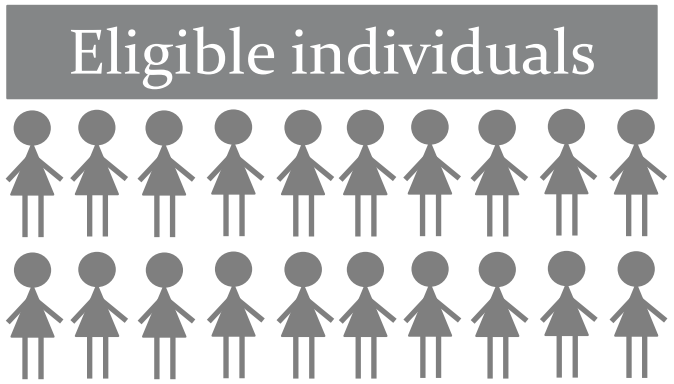


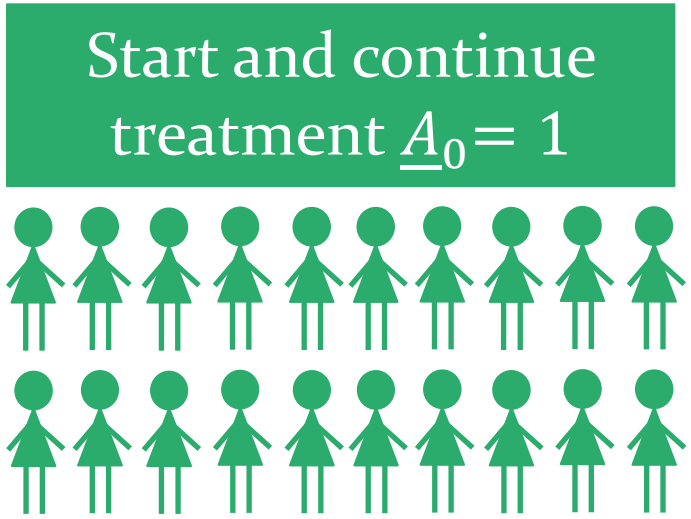
Start and continue treatment $\underline{A}_0 = 1$

Do not use treatment $\underline{A}_0 = 0$



What are we trying to estimate?



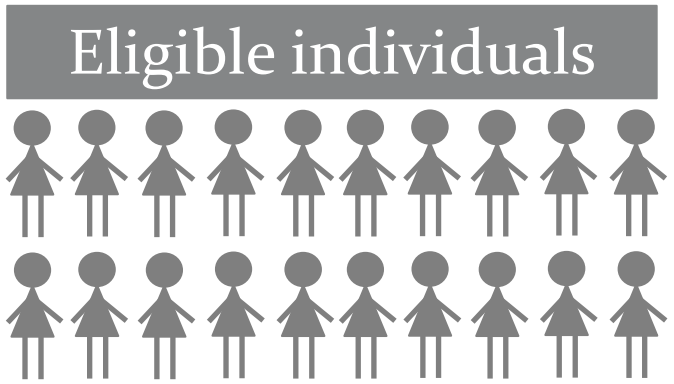


$$T_{\underline{A}_0=1}$$

Counterfactual time of death if an individual receives treatment a from time 0 onwards

$$T_{\underline{A}_0=0}$$

What are we trying to estimate?



Risk difference

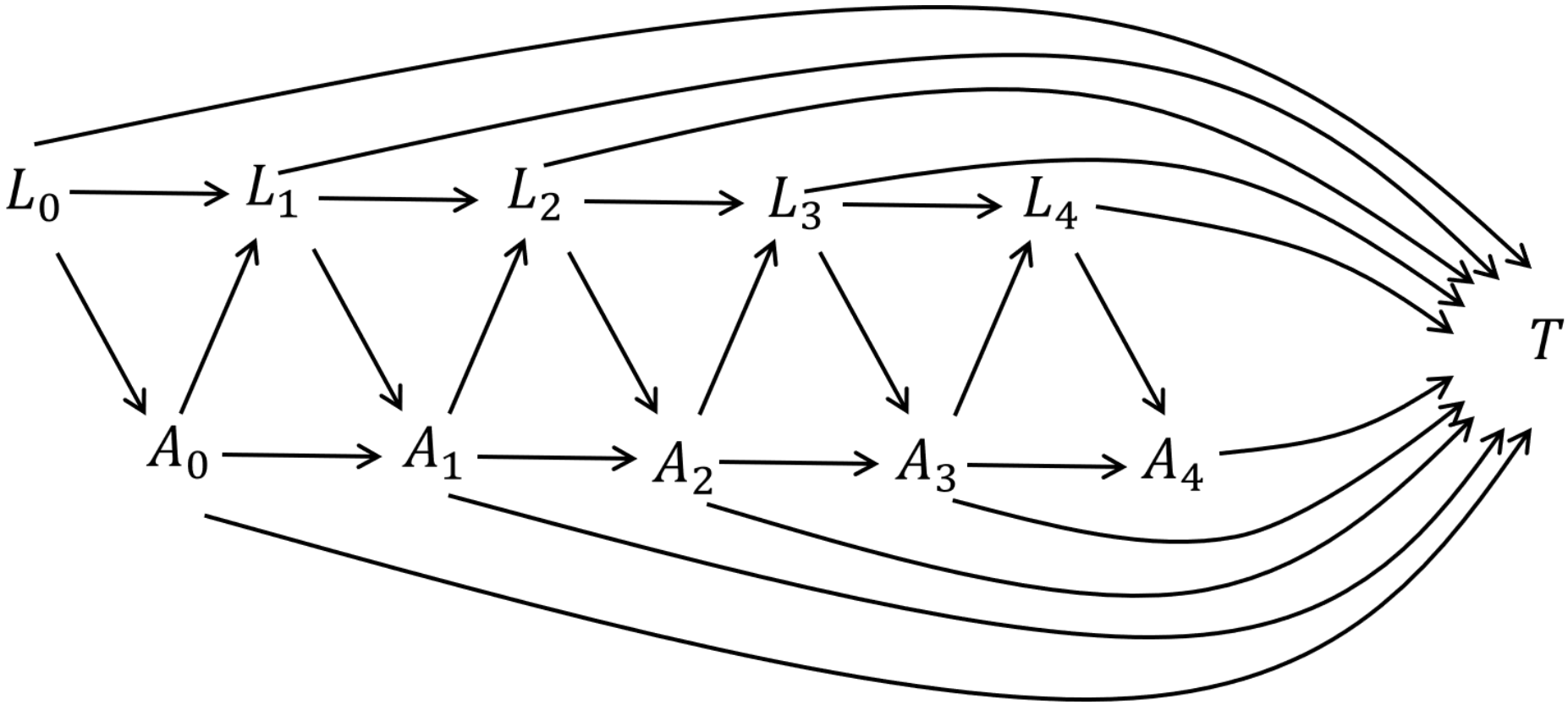
$$\Pr\left(T_{\underline{A}_0=1} > t | \text{Eligible}\right) - \Pr\left(T_{\underline{A}_0=0} > t | \text{Eligible}\right)$$

Methods

1. Marginal structural models (MSM) estimated using inverse probability of treatment weighting (IPTW)
2. Sequential trials approach
3. Several others…

Clare et al. Causal models adjusting for time-varying confounding—a systematic review of the literature. Int J Epi 2018. doi: 10.1093/ije/dyy218

Assumed set-up

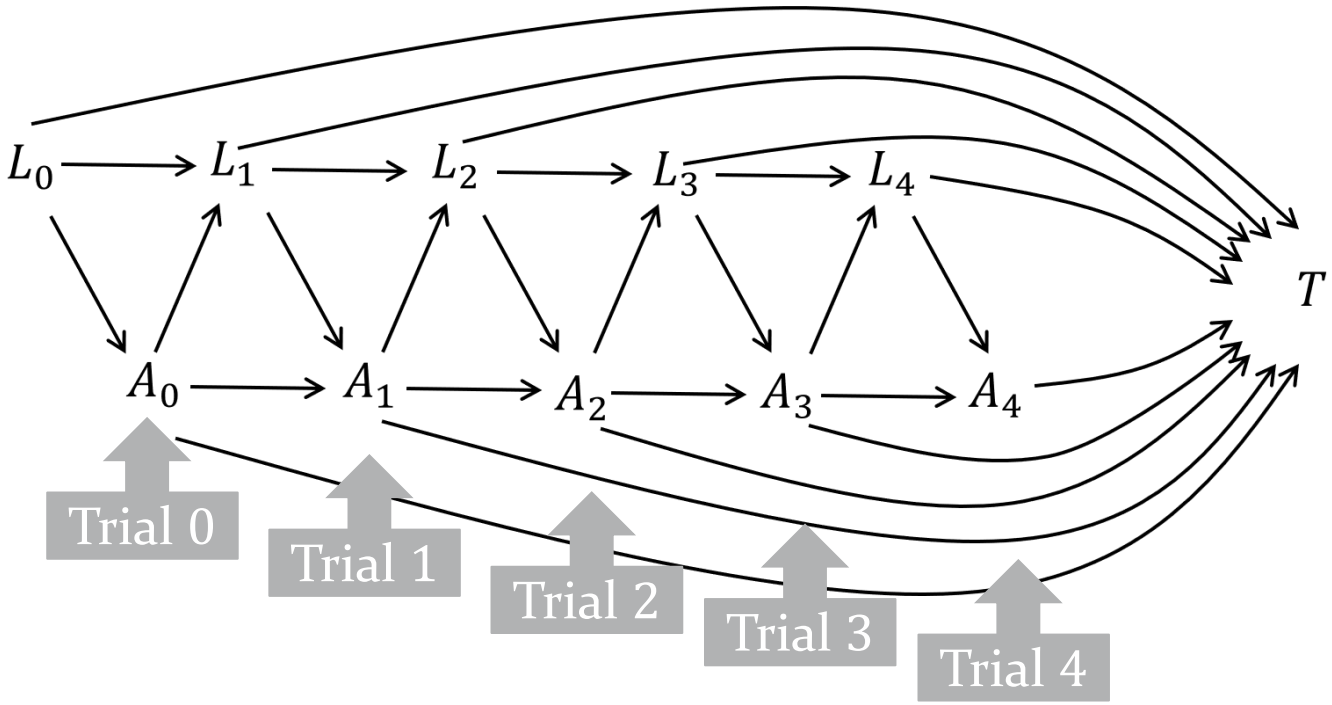


Sequential trials approach

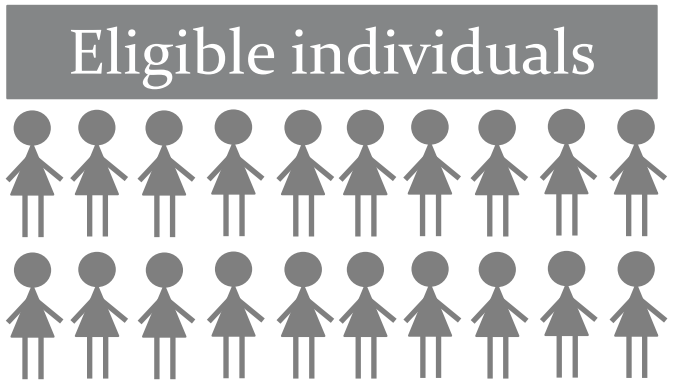
Gran et al. A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. Stat Med 2010.

Hernan et al. Observational studies analysed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 2008.

- Set up 'artificial' trials from times 0,1,2,3,4



What are we trying to estimate?



Risk difference

$$\Pr\left(T_{\underline{A}_0=1} > t | \text{Eligible}\right) - \Pr\left(T_{\underline{A}_0=0} > t | \text{Eligible}\right)$$

A marginal structural hazard model

Counterfactual event time if a person untreated up to time 0 has treatment status a from time 0 onwards

$$T_{\underline{A}_0=a}$$

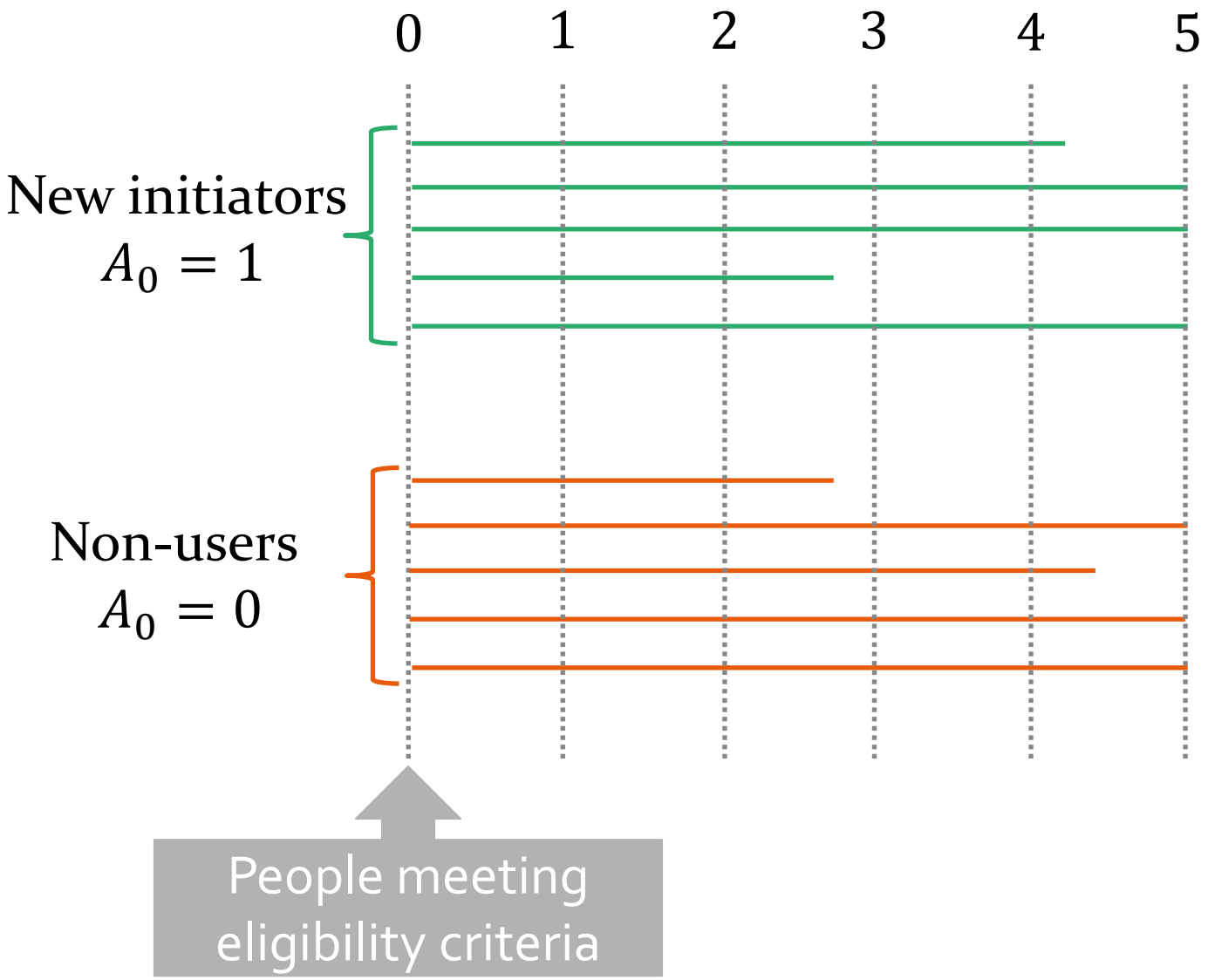
MSM for the hazard:

$$\lambda_{T_{\underline{A}_0=a}}(t) = \alpha(t) + \beta_A(t)a$$

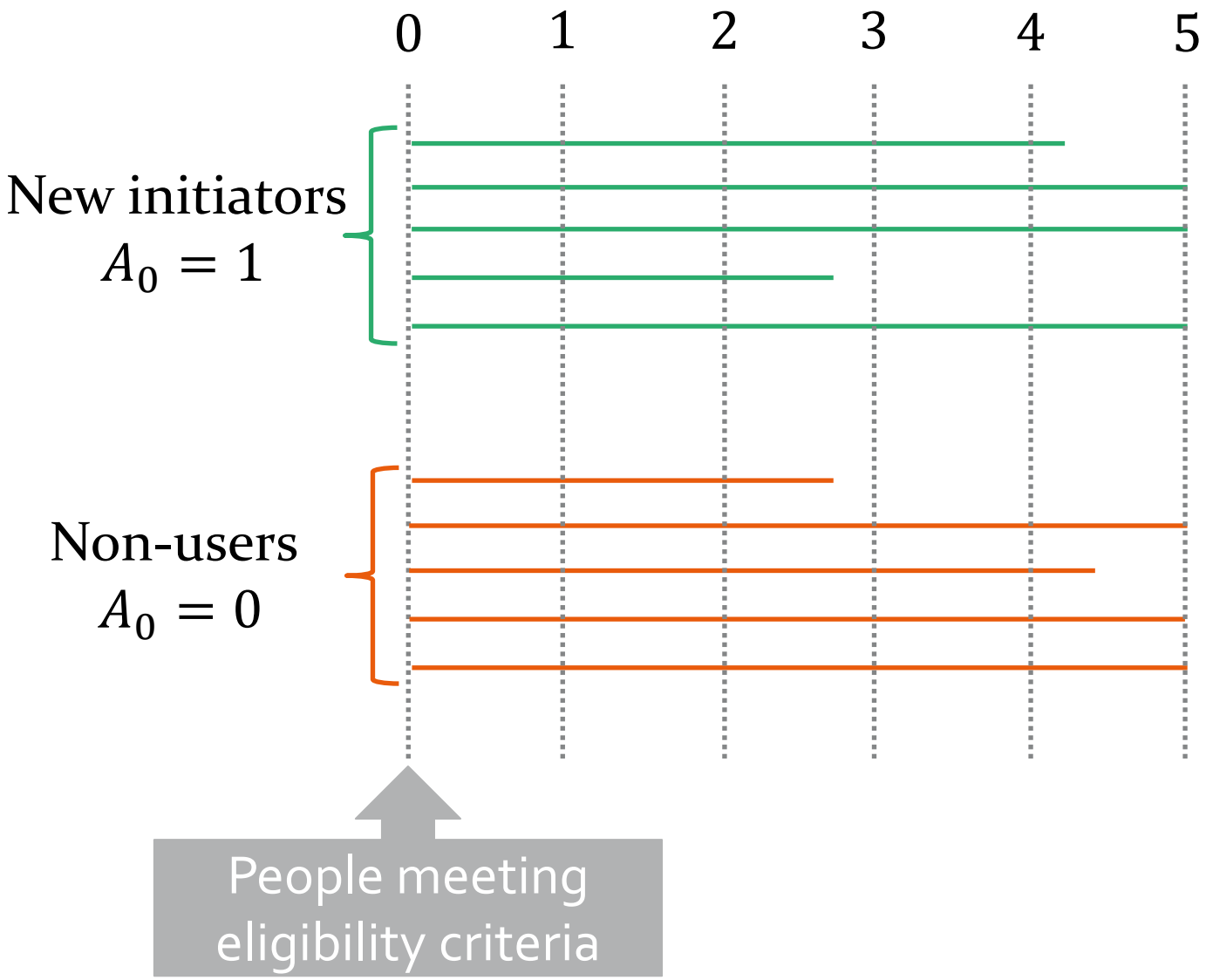
$$\lambda_{T_{\underline{A}_0=a}}(t|L_0) = \alpha(t) + \beta_A(t)a + \beta_L L_0$$

We cannot estimate the parameters of this model directly from the data due to time-dependent confounding

Sequential trials approach



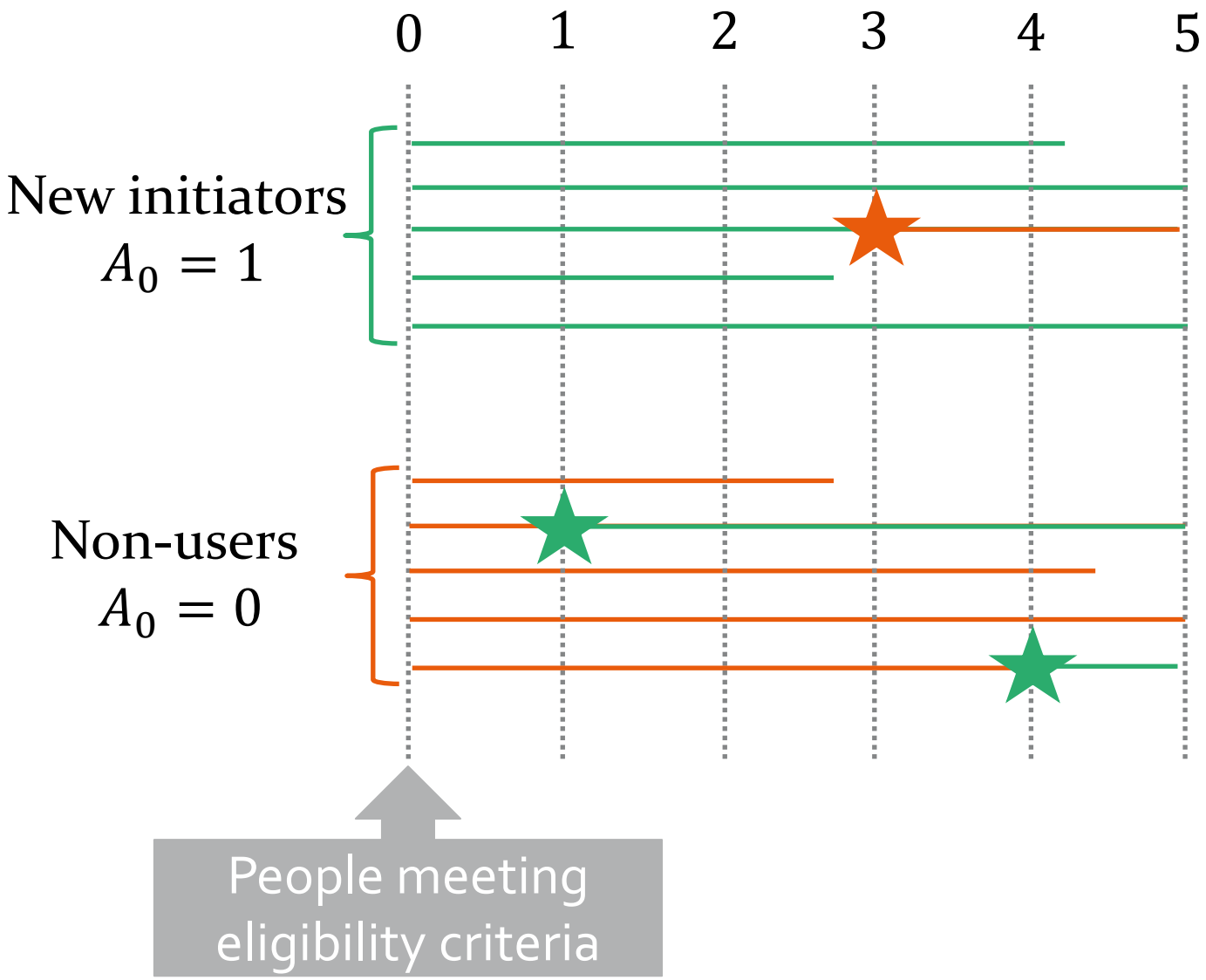
Sequential trials approach



- Who starts to receive treatment at time 0 is not random

- These who start treatment at time 0 may STOP in the future
- These who DO NOT start treatment at time 0 may START in the future

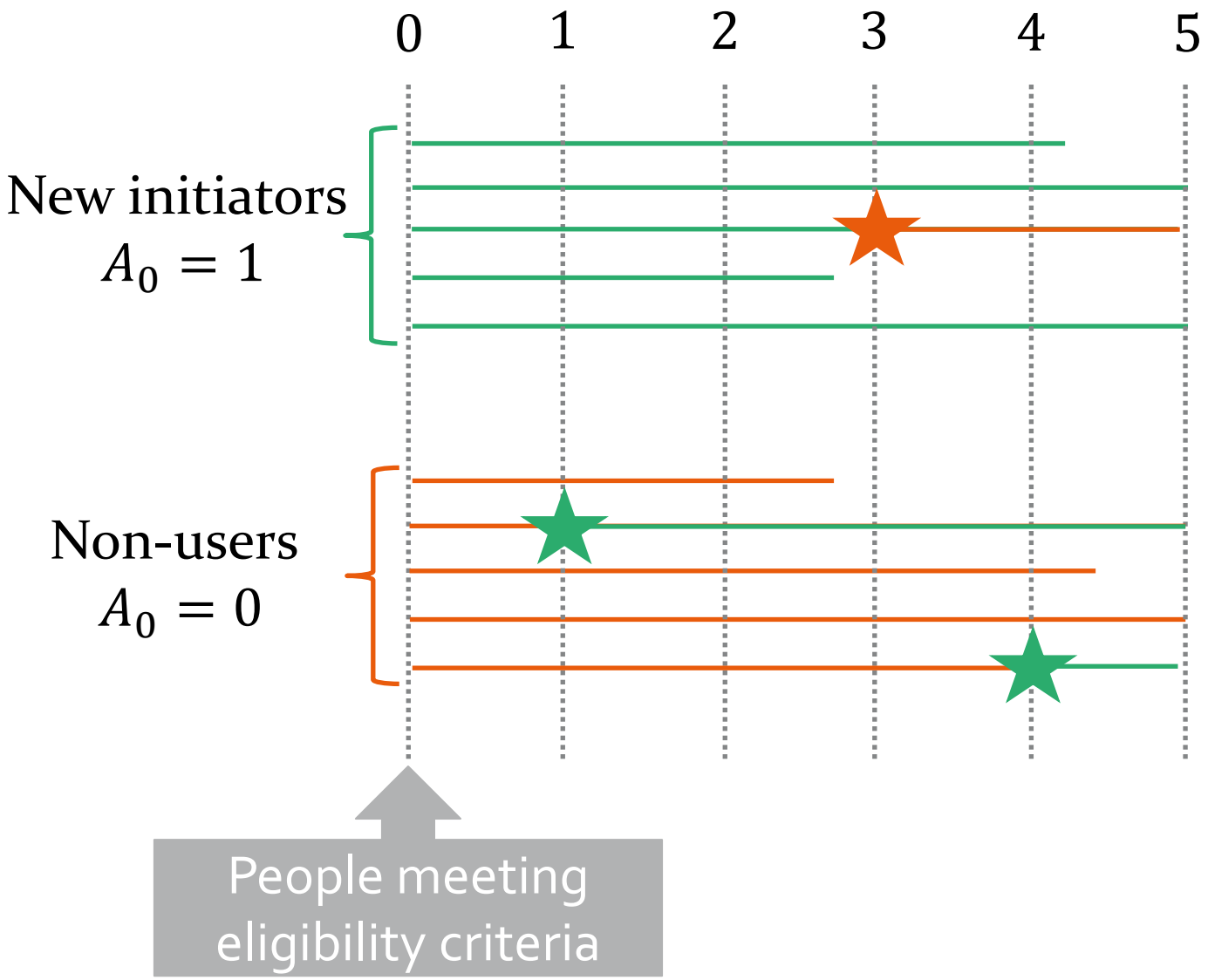
Sequential trials approach



- Who starts to receive treatment at time 0 is not random

- These who start treatment at time 0 may STOP in the future
- These who DO NOT start treatment at time 0 may START in the future

Sequential trials approach



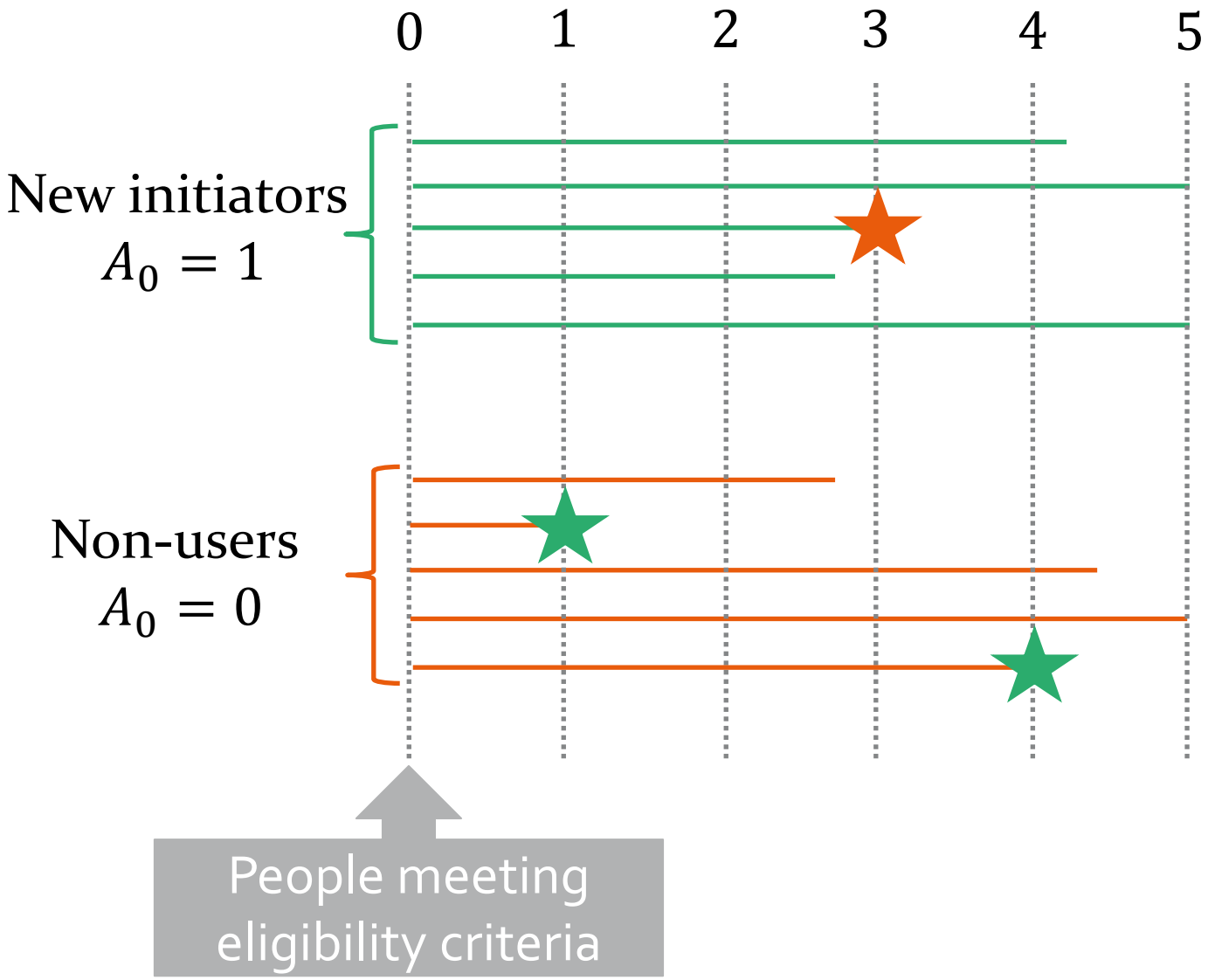
- Who starts to receive treatment at time 0 is not random

- Adjust for baseline confounders

- These who start treatment at time 0 may STOP in the future
- These who DO NOT start treatment at time 0 may START in the future

- Censor people when they deviate from their original treatment group
- And apply weights

Sequential trials approach



- Who starts to receive treatment at time 0 is not random

- Adjust for baseline confounders

- These who start treatment at time 0 may STOP in the future
- These who DO NOT start treatment at time 0 may START in the future

- Censor people when they deviate from their original treatment group
- And apply weights

A marginal structural hazard model

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

MSM for the hazard:

$$\lambda_{T_{\underline{A}_0=a}}(t|L_0) = \alpha(t) + \beta_A(t)a + \beta_L L_0$$

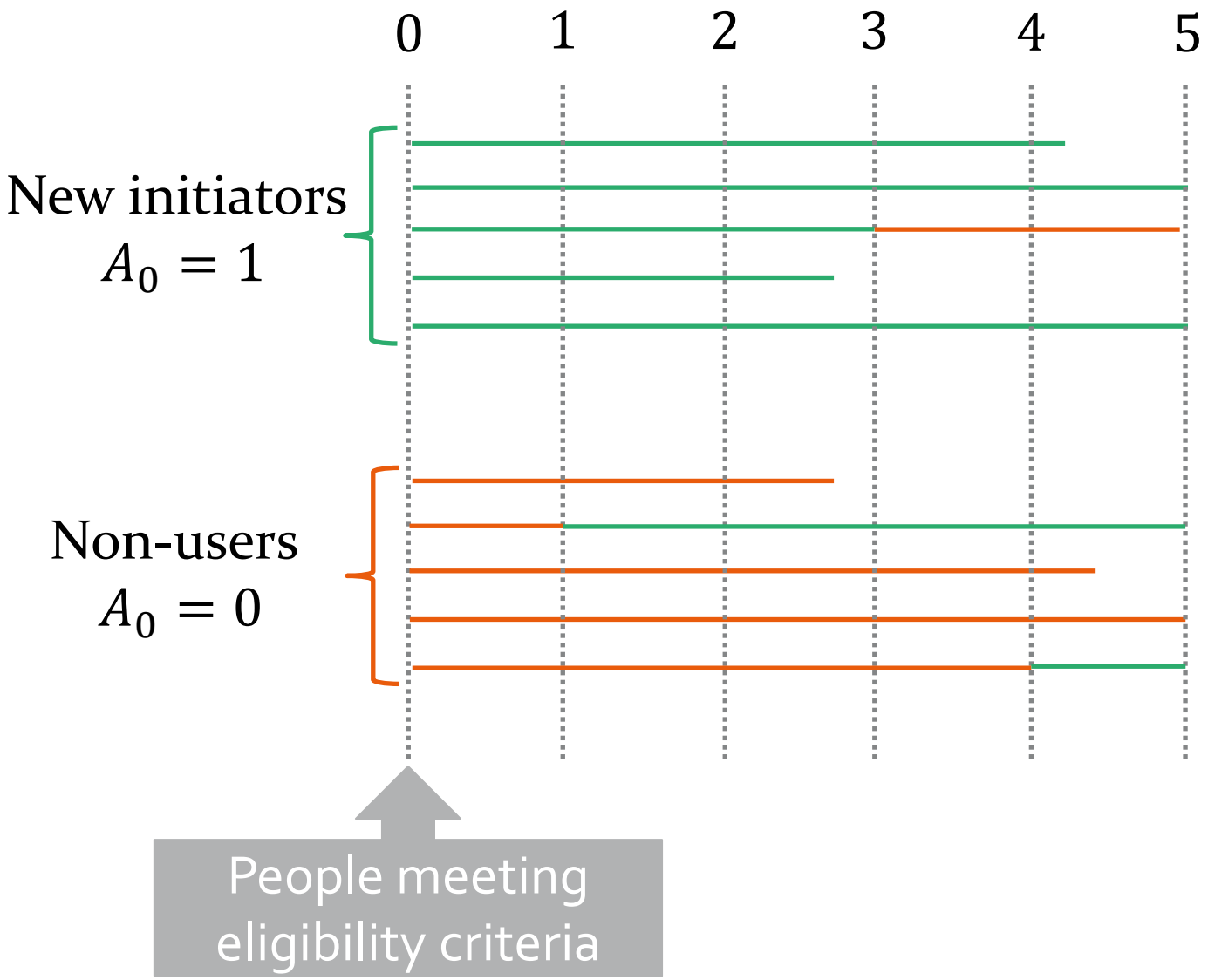
Hazard model for observed event times:

$$\lambda(t|A_0 = a, L_0) = \alpha(t) + \beta_A(t)a + \beta_L L_0$$

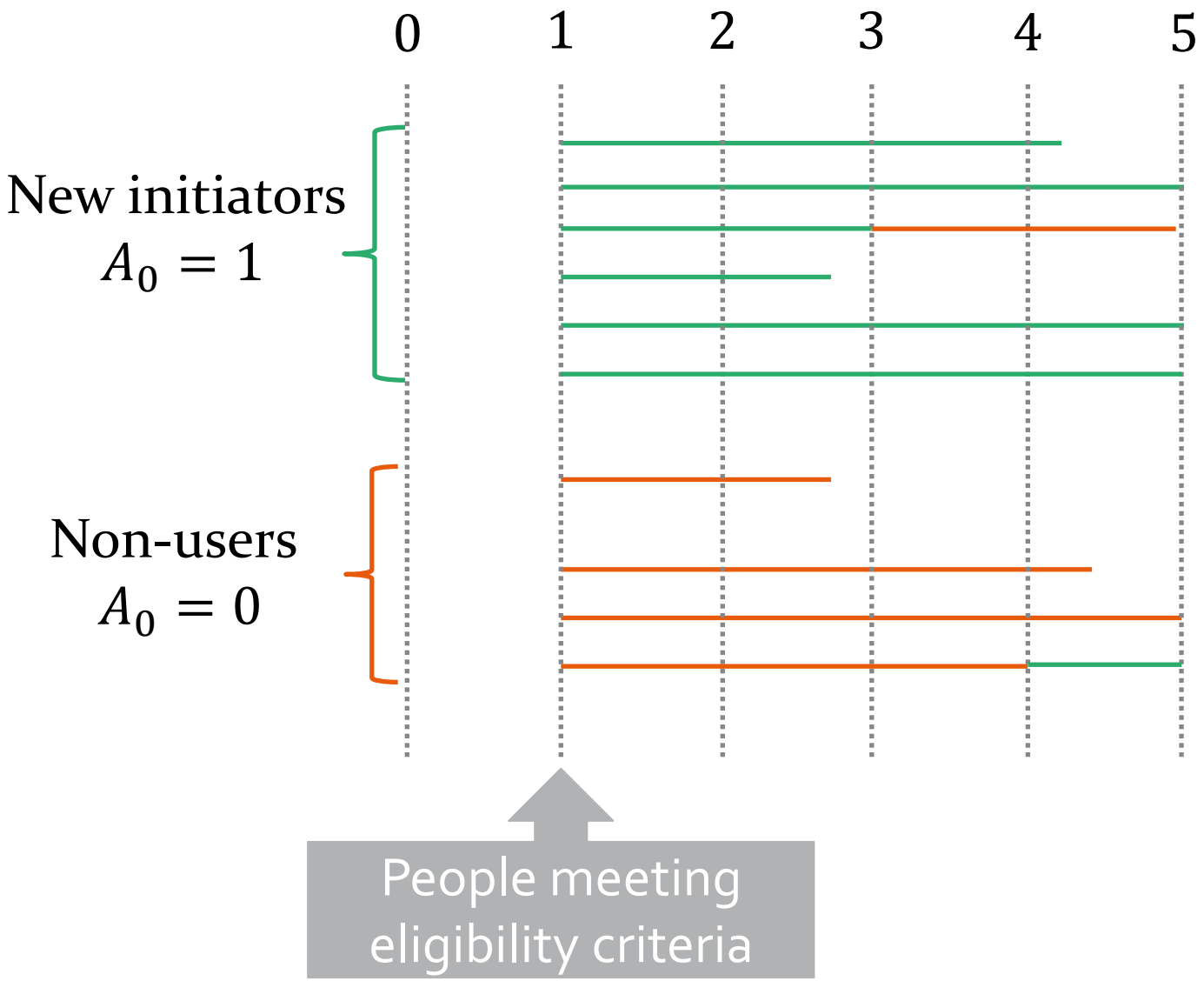
Fitting this model to the observed 'trial' data estimates the parameters of the MSM under some assumptions:

- Consistency
- Positivity
- No unmeasured confounding

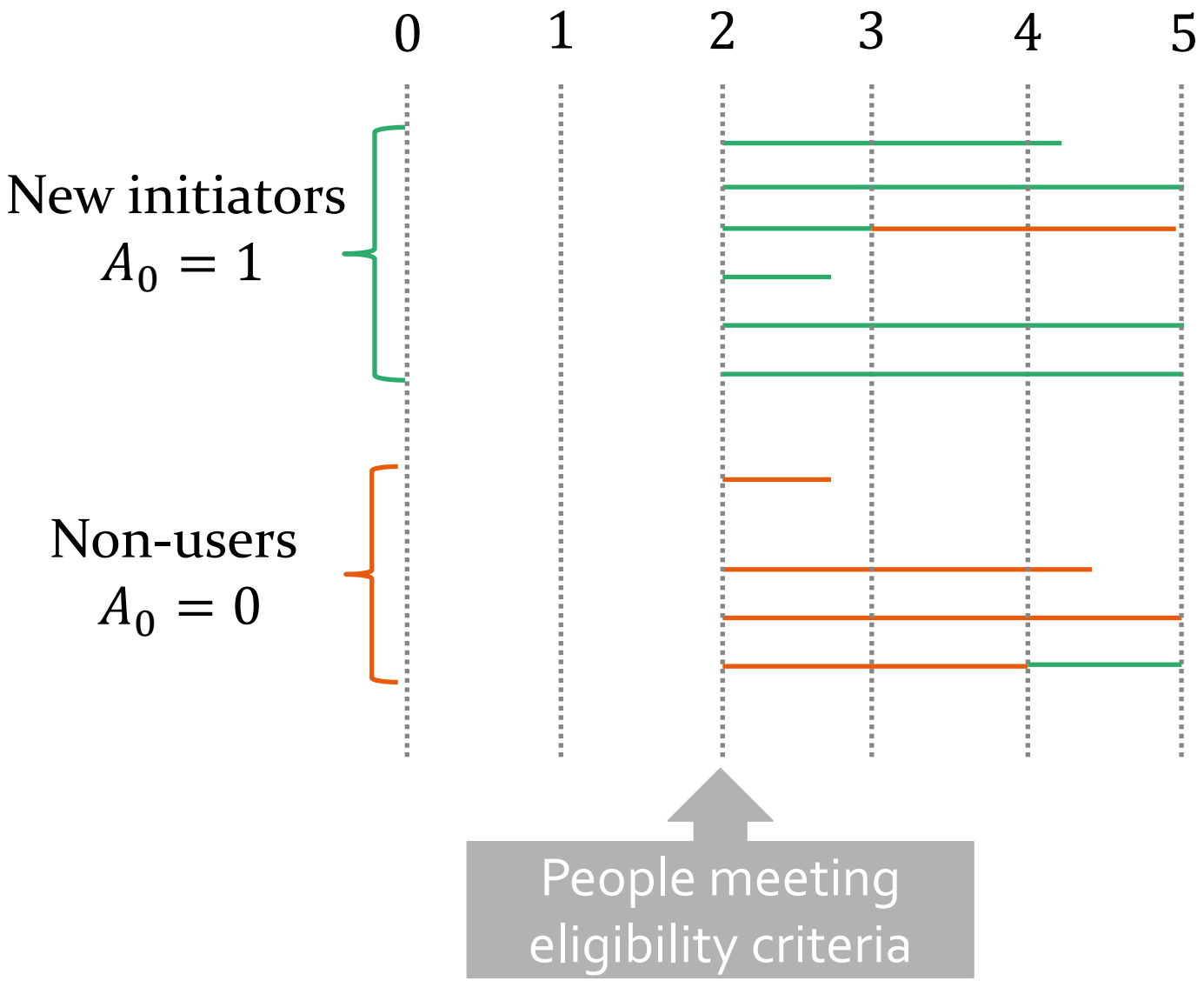
Sequential trials approach



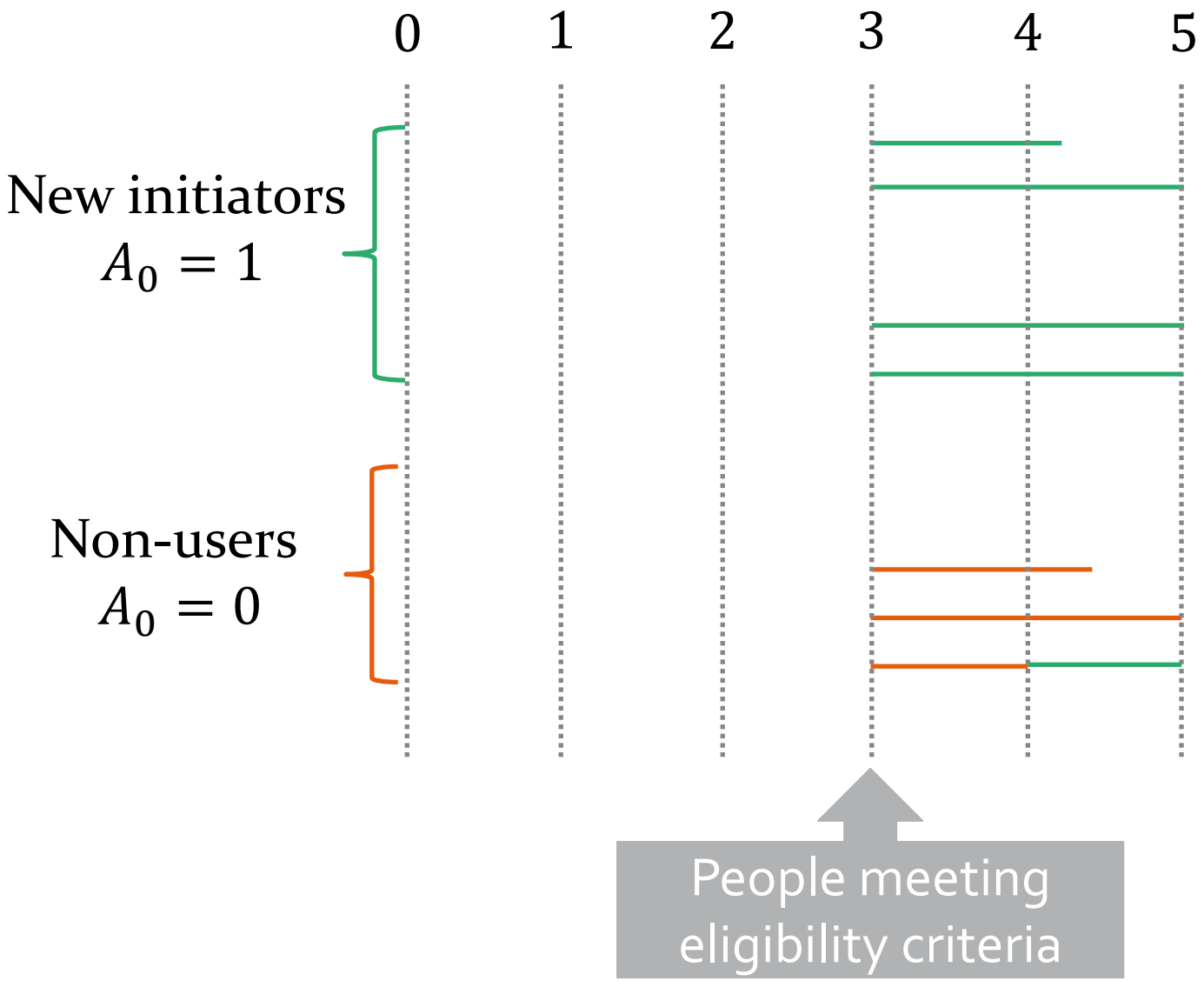
Sequential trials approach



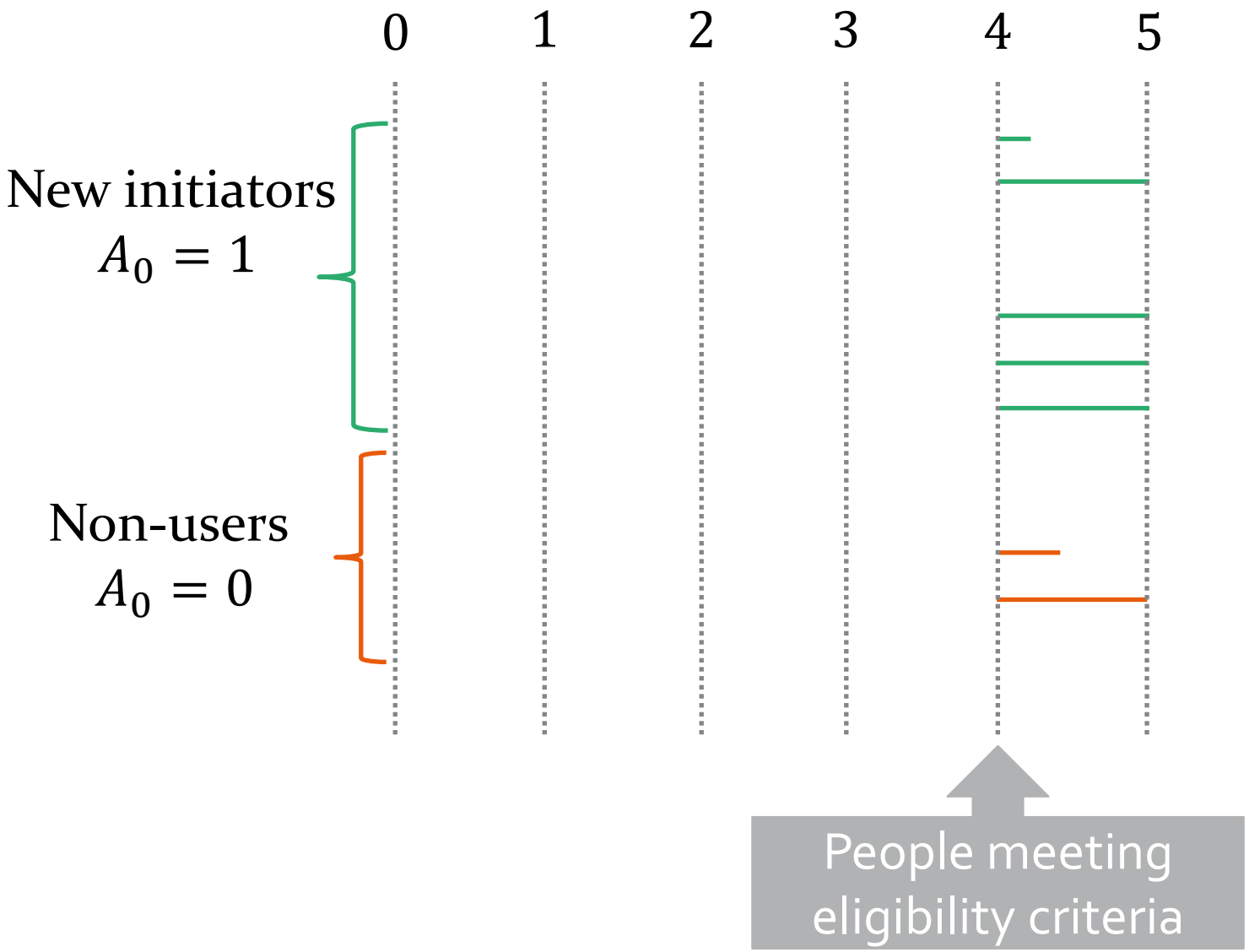
Sequential trials approach



Sequential trials approach



Sequential trials approach



A marginal structural hazard model

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

MSM for the hazard:

$$\lambda_{T_{\underline{A}_0=a}}(t|L_0) = \alpha(t) + \beta_A(t)a + \beta_L L_0$$

Hazard model for observed event times:

$$\lambda(t|A_0 = a, L_0) = \alpha(t) + \beta_A(t)a + \beta_L L_0$$

Fitting this model to the observed 'trial' data estimates the parameters of the MSM under some assumptions:

- Consistency
- Positivity
- No unmeasured confounding

Hazard models for each 'trial'

Trial 0 $\lambda_0(t|A_0 = a, L_0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_0$

Trial 1 $\lambda_1(t|A_1 = a, L_1, \bar{A}_0 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_1$

.
.
.

Trial 4 $\lambda_4(t|A_4 = a, L_4, \bar{A}_3 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_4$

Hazard models for each 'trial'

Time is measured relative to the start of the trial

Trial 0 $\lambda_0(t|A_0 = a, L_0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_0$

Trial 1 $\lambda_1(t|A_1 = a, L_1, \bar{A}_0 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_1$

.
.
.

Trial 4 $\lambda_4(t|A_4 = a, L_4, \bar{A}_3 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_4$

Hazard models for each 'trial'

Time is measured relative to the start of the trial

Untreated in the past

Trial 0 $\lambda_0(t|A_0 = a, L_0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_0$

Trial 1 $\lambda_1(t|A_1 = a, L_1, \bar{A}_0 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_1$

.
.
.

Trial 4 $\lambda_4(t|A_4 = a, L_4, \bar{A}_3 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_4$

Hazard models for each 'trial'

Time is measured relative to the start of the trial

Untreated in the past

Trial 0 $\lambda_0(t|A_0 = a, L_0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_0$

Trial 1 $\lambda_1(t|A_1 = a, L_1, \bar{A}_0 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_1$

.
.
.

Trial 4 $\lambda_4(t|A_4 = a, L_4, \bar{A}_3 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_4$

No time-updated covariates.

Adjusting for L variables at the start of each trial is used to control confounding

Hazard models for each 'trial'

Time is measured relative to the start of the trial

Untreated in the past

Trial 0 $\lambda_0(t|A_0 = a, L_0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_0$

Trial 1 $\lambda_1(t|A_1 = a, L_1, \bar{A}_0 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_1$

.
.
.

Trial 4 $\lambda_4(t|A_4 = a, L_4, \bar{A}_3 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_4$

No time-updated covariates.

Adjusting for L variables at the start of each trial is used to control confounding

Coefficients for A and L are assumed the same across trials

Hazard models for each 'trial'

Time is measured relative to the start of the trial

Untreated in the past

Trial 0 $\lambda_0(t|A_0 = a, L_0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_0$

Trial 1 $\lambda_1(t|A_1 = a, L_1, \bar{A}_0 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_1$

.
.
.

Trial 4 $\lambda_4(t|A_4 = a, L_4, \bar{A}_3 = 0) = \alpha(t) + \beta_A(t)a + \beta_L(t)L_4$

No time-updated covariates.

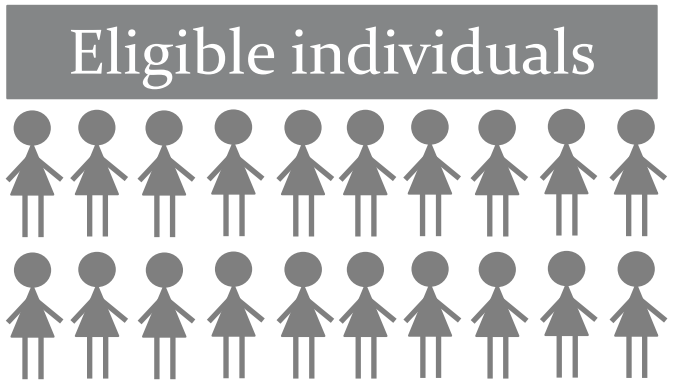
Adjusting for L variables at the start of each trial is used to control confounding

Coefficients for A and L are assumed the same across trials

Hazard MSM $\lambda_{T\underline{A}_k=a}(t|L_k, \bar{A}_{k-1} = 0) = \alpha(t) + \beta_A(t)a + \beta_L L_k$

Estimating survival probabilities

What are we trying to estimate?



Risk difference

$$\Pr\left(T_{\underline{A}_0=1} > t | \text{Eligible}\right) - \Pr\left(T_{\underline{A}_0=0} > t | \text{Eligible}\right)$$

Obtaining survival probabilities

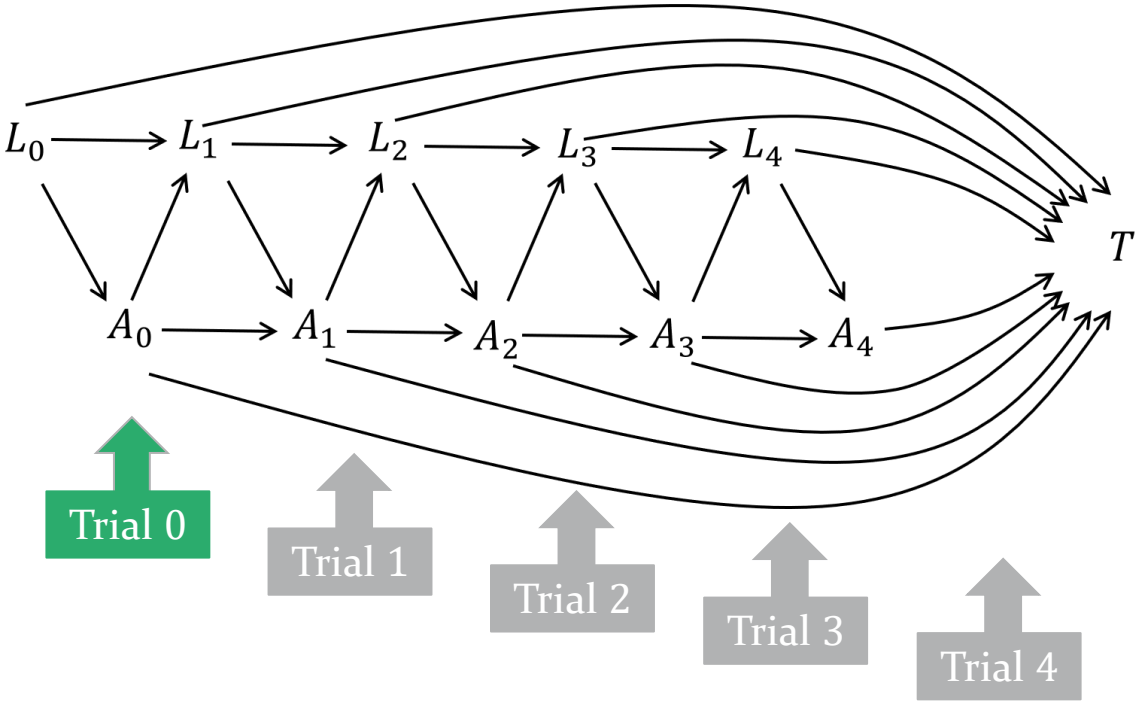
Hazard MSM $$\lambda_{T_{\underline{A}_k=a}}(t|L_k, \bar{A}_{k-1}=0) = \alpha(t) + \beta_A(t)a + \beta_L L_k$$

Counterfactual conditional survival probability

$$\Pr\left(\mathrm{T}_{\underline{A}_k=a} > \mathrm{t}|\bar{L}_k, \bar{A}_{k-1}=0\right) =$$

$$\exp\left\{-\int_0^t \lambda(u|A_k, L_k, \bar{A}_{k-1}=0)du\right\}$$

From conditional to marginal…



From conditional to marginal

1. Create two copies of the data set **for people eligible for trial 0**

Set $a = 0$ for everyone

ID	a	L_0
1	0	27
2	0	25
3	0	23

Set $a = 1$ for everyone

ID	a	L_0
1	1	27
2	1	25
3	1	23

2. Obtain predicted conditional survival probabilities in both data sets

$$\Pr(T_0 > \mathrm{t} | \bar{L}_{0i}, \bar{A}_{-1i} = 0)$$

$$\Pr(T_1 > \mathrm{t} | \bar{L}_{0i}, \bar{A}_{-1i} = 0)$$

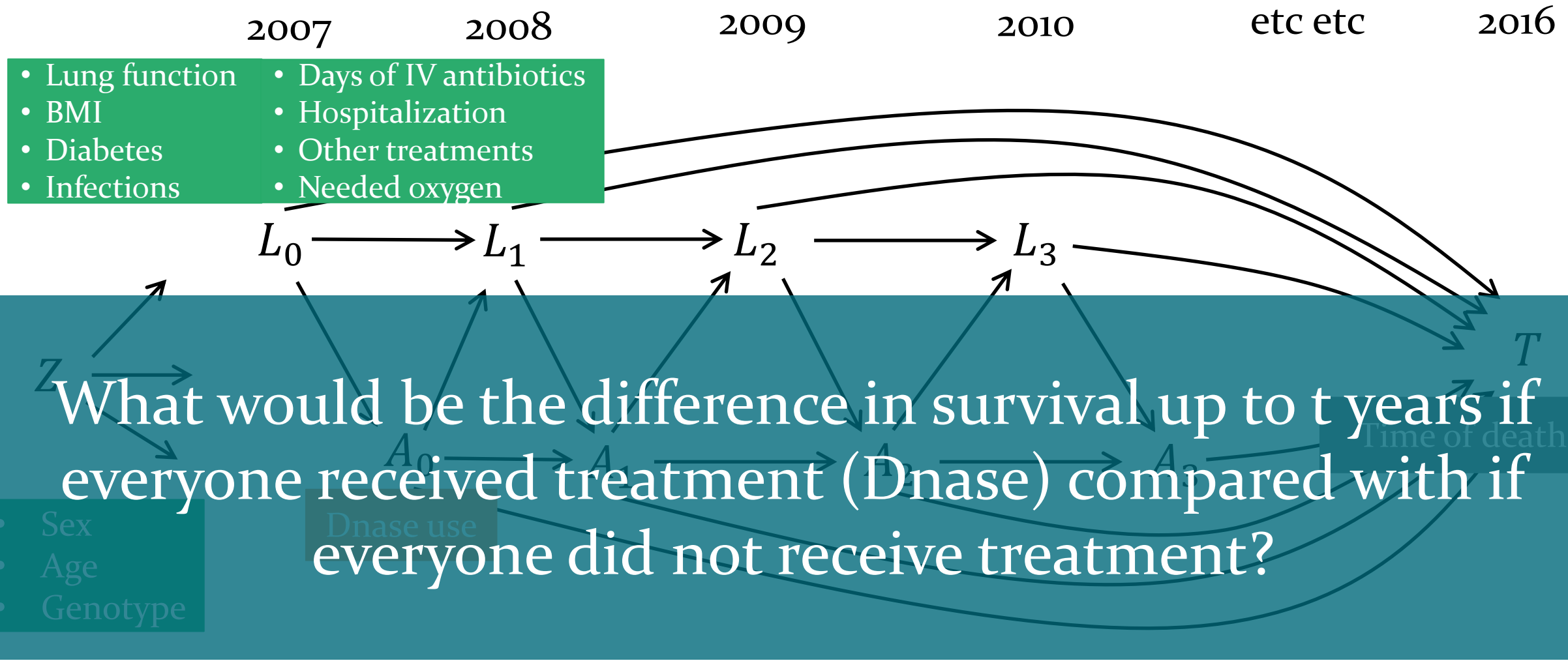
3. Take the empirical average

$$\frac{1}{n}\sum_{i=1}^{n} \Pr(T_0 > \mathrm{t} | \bar{L}_{0i}, \bar{A}_{-1i} = 0)$$

$$\frac{1}{n}\sum_{i=1}^{n} \Pr(T_1 > \mathrm{t} | \bar{L}_{0i}, \bar{A}_{-1i} = 0)$$

Application using CF Registry data

UK CF Registry setting



What would be the difference in survival up to t years if everyone received treatment (Dnase) compared with if everyone did not receive treatment?

Target trial protocol

- Eligibility criteria
- Treatment strategies
- Outcome
- Follow-up
- Causal contrasts
- Analysis plan

Hernan, Robins. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epi* 2016; 183:758–764

Hernan et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epi* 2016; 79: 70-75

García-Albéniz, Hsu, Hernán. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur J Epidemiol* 2017;32:495-500.

Target trial protocol

- Eligibility criteria
- Treatment strategies
- Outcome
- Follow-up
- Causal contrasts
- Analysis plan

- People with CF in the UK between 2008 and 2017.
- Aged 12+
- Have not used Dnase for at least 3 years
- Not previously received organ transplant
- A person is eligible to be randomized at any point between 2008 and 2017 at which they meet the eligibility criteria

Target trial protocol

- Eligibility criteria
- **Treatment strategies**
- Outcome
- Follow-up
- Causal contrasts
- Analysis plan

1. Assignment to initiate DNase and continue treatment thereafter
2. Assignment not to use DNase

Target trial protocol

- Eligibility criteria
- Treatment strategies
- **Outcome**
  - Death or transplant
- Follow-up
- Causal contrasts
- Analysis plan

Target trial protocol

- Eligibility criteria
- Treatment strategies
- Outcome
- **Follow-up**
- Causal contrasts
- Analysis plan

From randomization to the earliest of:

- Death
- Loss-to follow-up
- Transplant
- 5 years

Target trial protocol

Eligibility criteria

Treatment strategies

Outcome

Follow-up

Causal contrasts

Analysis plan

- Per-protocol effect of treatment use on survival probabilities
- Marginal risk differences

Target trial protocol

Eligibility criteria

Treatment strategies

Outcome

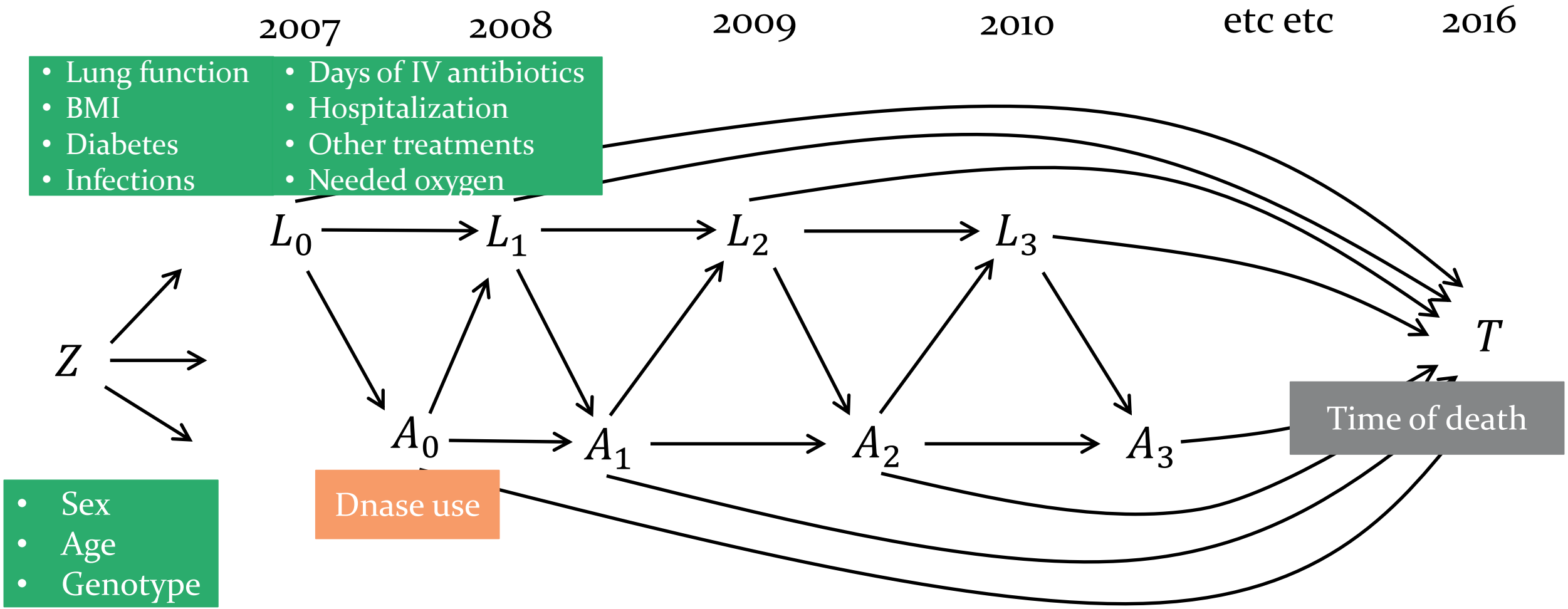
Follow-up

Causal contrasts

Analysis plan

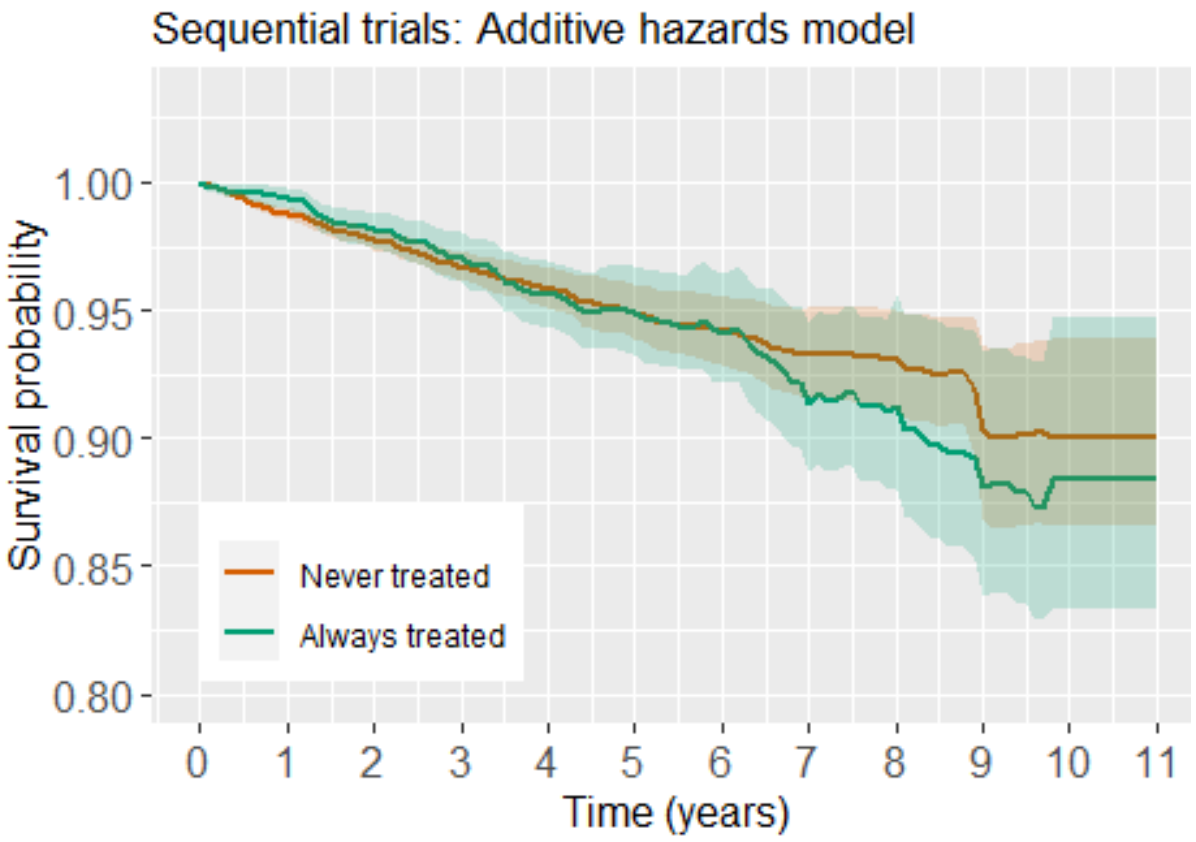
- Kaplan-Meier analysis
- Semi- or fully-parametric survival analysis adjusted for baseline covariates

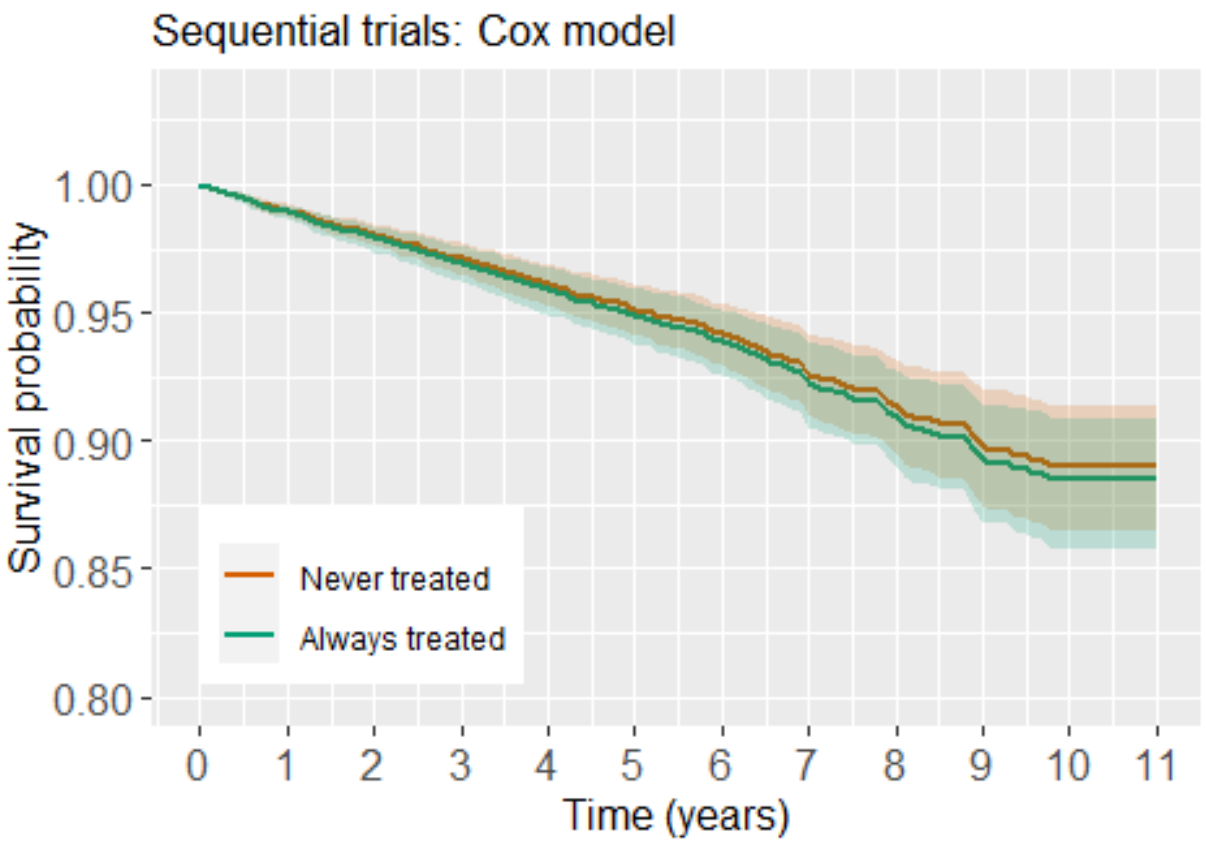
Emulating this target trial



Sequential trials analysis

- 3855 unique individuals met the eligibility criteria at at least one time point.
- There were 338 events (death or transplant)





Summary

- The sequential trials approach is an intuitive way of analysing longitudinal data to estimate long-term treatment effects
- The estimands can be expressed in terms of causal effects in a marginal structural model

  …and it is straightforward to estimate meaningful quantities e.g. survivor curves
- Simulation comparisons suggest this approach performs well relative to the standard MSM approach, and can be more efficient

Keogh, Seaman, Vansteelandt, Gran. Simulating longitudinal data from marginal structural models using the additive hazard model. arxiv.org/abs/2002.03678

https://github.com/ruthkeogh/causal_sim

Funding

UK Research & Innovation Future Leaders Fellowship

Data

UK Cystic Fibrosis Registry and those who provide their data