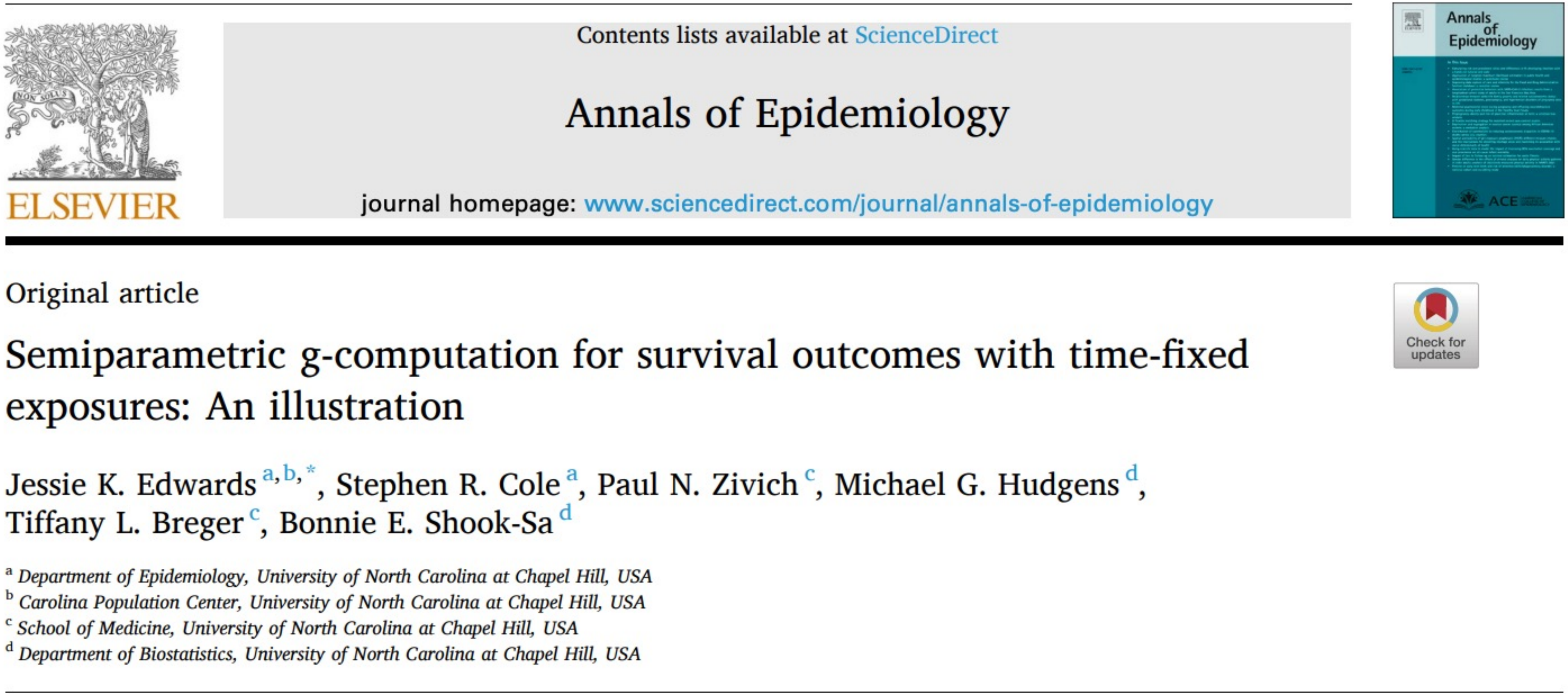
Semiparametric g-computation to account for baseline confounding in observational studies with survival endpoints

Jessie K. Edwards

University of North Carolina at Chapel Hill

jessedwards@unc.edu



Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.sciencedirect.com/journal/annals-of-epidemiology

Original article

Semiparametric g-computation for survival outcomes with time-fixed exposures: An illustration

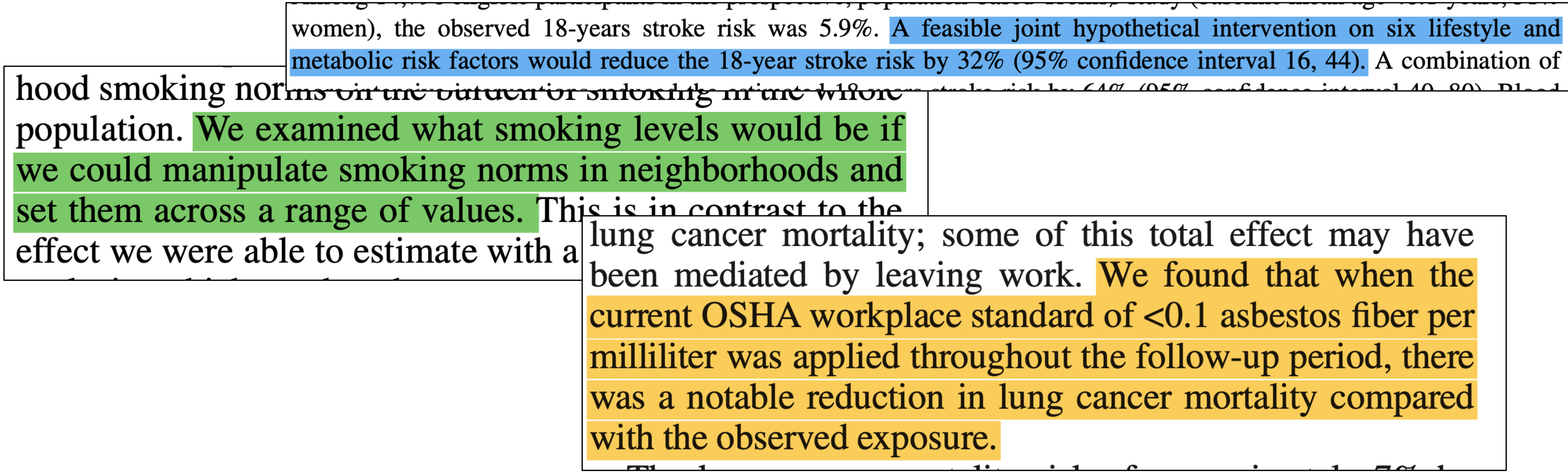
Jessie K. Edwards [a,b,*], Stephen R. Cole [a], Paul N. Zivich [c], Michael G. Hudgens [d], Tiffany L. Breger [c], Bonnie E. Shook-Sa [d]

[a] *Department of Epidemiology, University of North Carolina at Chapel Hill, USA*
[b] *Carolina Population Center, University of North Carolina at Chapel Hill, USA*
[c] *School of Medicine, University of North Carolina at Chapel Hill, USA*
[d] *Department of Biostatistics, University of North Carolina at Chapel Hill, USA*

Edwards JK, Cole SR, Zivich PN, Hudgens MG, Breger TL, Shook-Sa BE. Semiparametric g-computation for survival outcomes with time-fixed exposures: an illustration. Annals of Epidemiology. 2024 Jun 3.

G computation is **useful**

Method to estimate marginal counterfactual outcome distributions while accounting for confounding.



women), the observed 18-years stroke risk was 5.9%. A feasible joint hypothetical intervention on six lifestyle and metabolic risk factors would reduce the 18-year stroke risk by 32% (95% confidence interval 16, 44). A combination of

hood smoking norms on the burden of smoking in the whole population. We examined what smoking levels would be if we could manipulate smoking norms in neighborhoods and set them across a range of values. This is in contrast to the effect we were able to estimate with a

lung cancer mortality; some of this total effect may have been mediated by leaving work. We found that when the current OSHA workplace standard of <0.1 asbestos fiber per milliliter was applied throughout the follow-up period, there was a notable reduction in lung cancer mortality compared with the observed exposure.

Ahern J, Hubbard A, Galea S. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. American journal of epidemiology. 2009 May 1;169(9):1140-7.
Cole SR, Richardson DB, Chu H, Naimi AI. Analysis of occupational asbestos exposure and lung cancer mortality using the g formula. American journal of epidemiology. 2013 May 1;177(9):989-96.
Vangen-Lønne AM, Ueda P, Gulayin P, Wilsgaard T, Mathiesen EB, Danaei G. Hypothetical interventions to prevent stroke: an application of the parametric g-formula to a healthy middle-aged population. European journal of epidemiology. 2018 Jun;33:557-66.

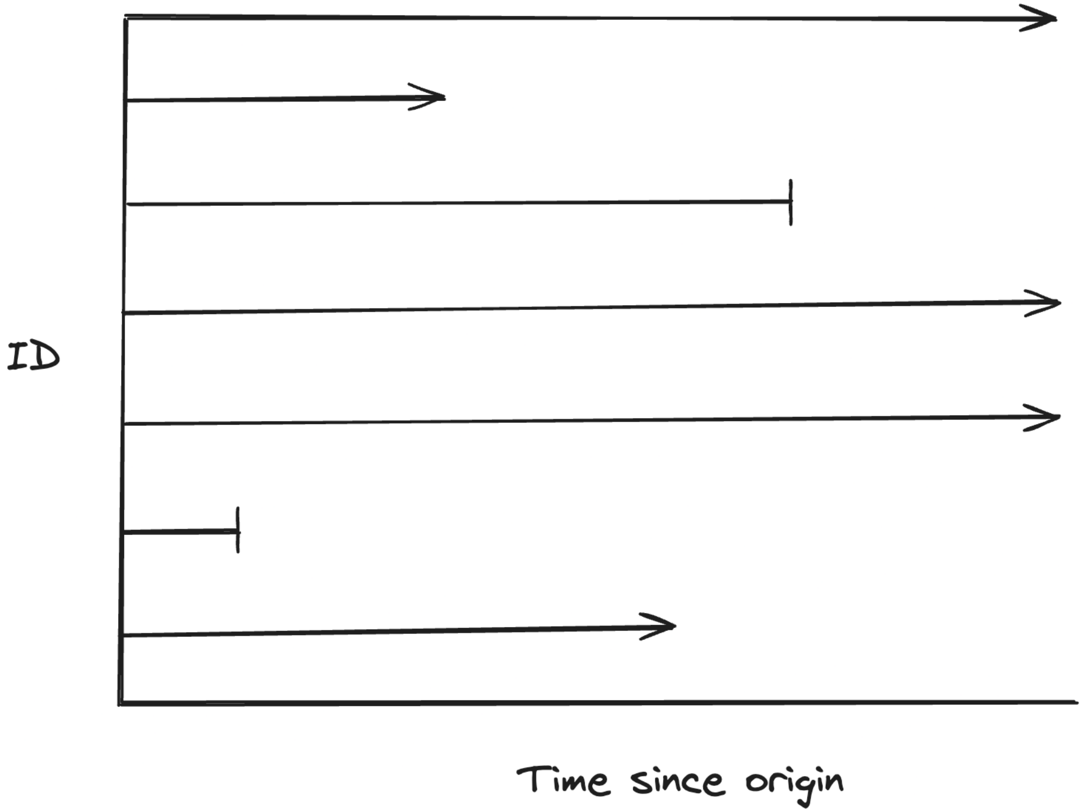
G computation with single time-point outcomes with no missing data is **straightforward**

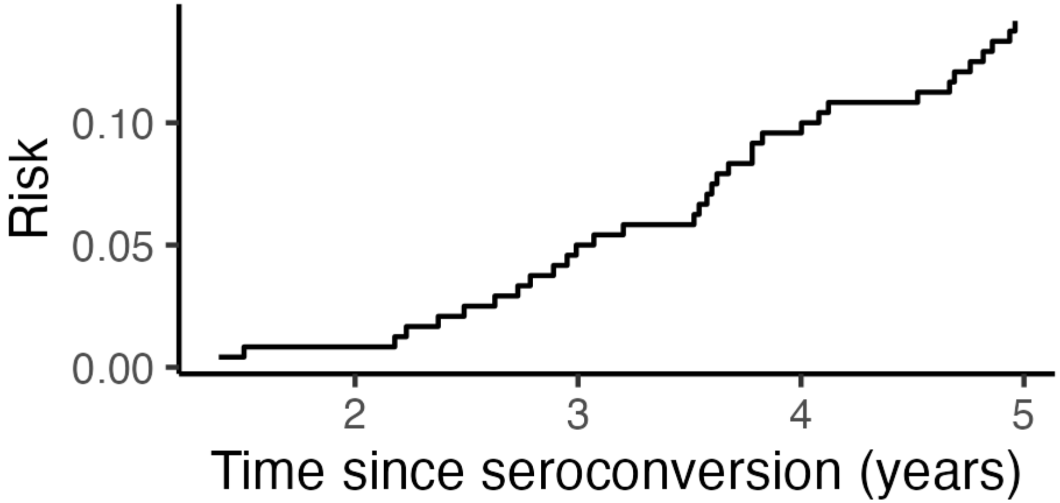
1. Fit a parametric model for the outcome conditional on exposure and covariates

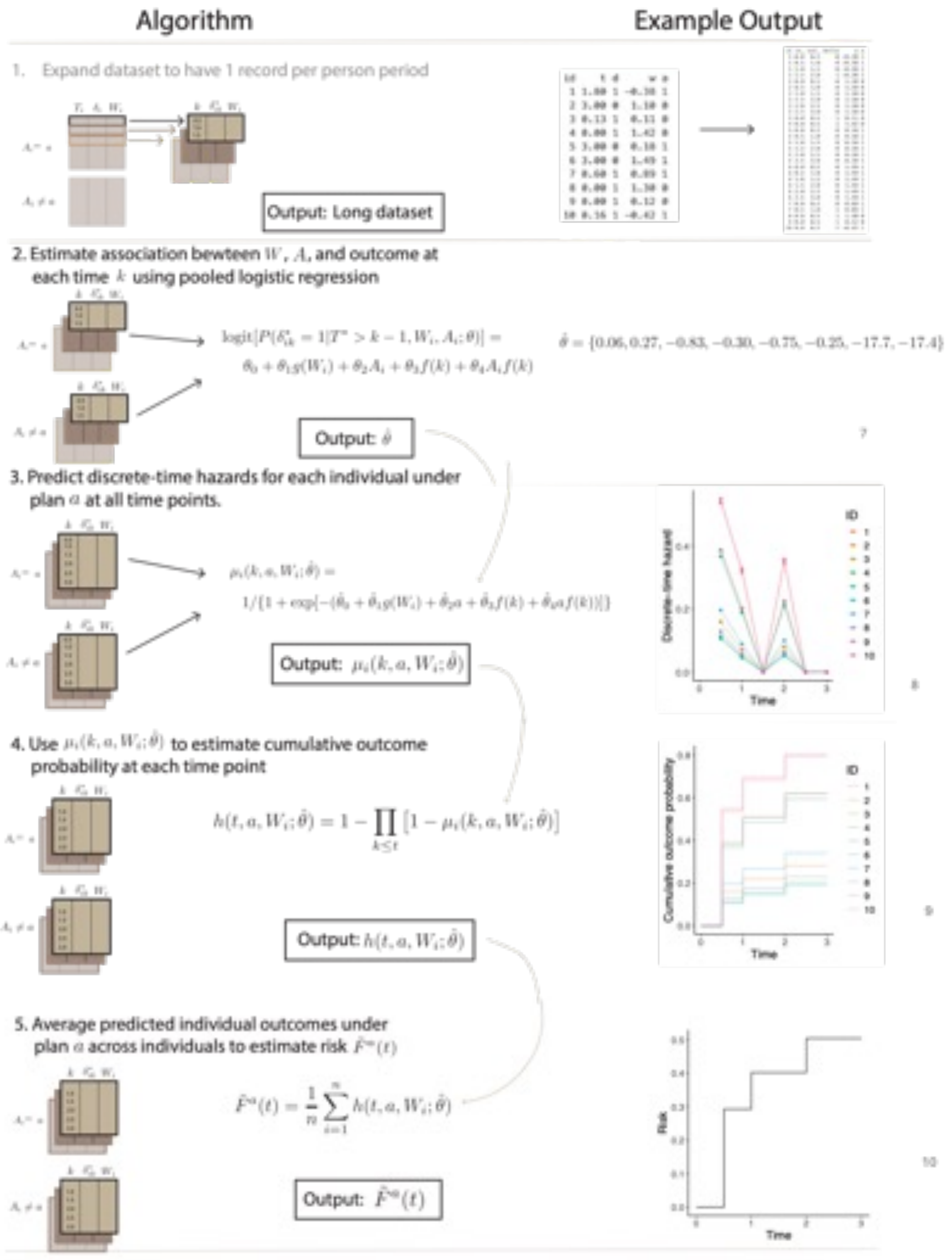
2. Set exposure to desired level

3. Use estimated coefficients from step 1 to predict counterfactual outcomes under each exposure level of interest.

G computation requires more steps when outcome is a **survival time** with **right censoring.**







1. Expand dataset to have 1 record per person period

Output: Long dataset

```
id     t d     w a
 1 1.00 1 -0.30 1
 2 3.00 0  1.10 0
 3 0.13 1  0.11 0
 4 0.00 1  1.42 0
 5 3.00 0  0.10 1
 6 3.00 0  1.49 1
 7 0.60 1  0.89 1
 8 0.00 1  1.30 0
 9 0.00 1  0.12 0
10 0.16 1 -0.42 1
```

2. Estimate association bewteen W, A, and outcome at each time k using pooled logistic regression

$$\text{logit}[P(E_{ik}^{*} = 1 | T^{*} > k-1, W_i, A_i; \theta)] = \theta_0 + \theta_1 g(W_i) + \theta_2 A_i + \theta_3 f(k) + \theta_4 A_i f(k)$$

Output: $\hat{\theta}$

$$\hat{\theta} = \{0.06, 0.27, -0.83, -0.30, -0.75, -0.25, -17.7, -17.4\}$$

3. Predict discrete-time hazards for each individual under plan a at all time points.

$$\mu_i(k, a, W_i; \hat{\theta}) = 1/\{1 + \exp[-(\hat{\theta}_0 + \hat{\theta}_1 g(W_i) + \hat{\theta}_2 a + \hat{\theta}_3 f(k) + \hat{\theta}_4 a f(k))]\}$$

Output: $\mu_i(k, a, W_i; \hat{\theta})$

4. Use $\mu_i(k, a, W_i; \hat{\theta})$ to estimate cumulative outcome probability at each time point

$$h(t, a, W_i; \hat{\theta}) = 1 - \prod_{k \leq t} [1 - \mu_i(k, a, W_i; \hat{\theta})]$$

Output: $h(t, a, W_i; \hat{\theta})$

5. Average predicted individual outcomes under plan a across individuals to estimate risk $\hat{F}^a(t)$

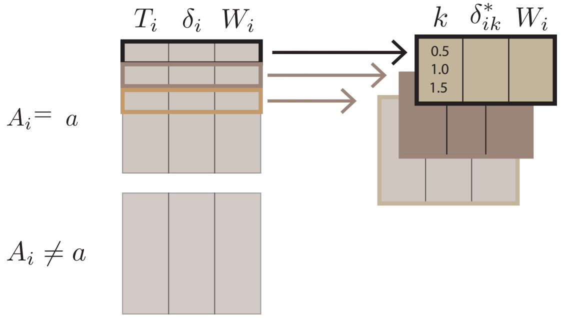
$$\hat{F}^a(t) = \frac{1}{n} \sum_{i=1}^{n} h(t, a, W_i; \hat{\theta})$$

Output: $\hat{F}^a(t)$

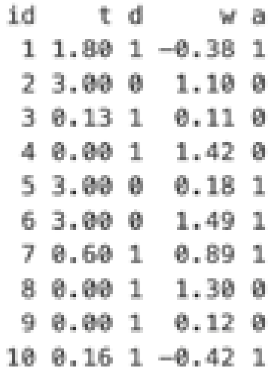
Algorithm

Example Output

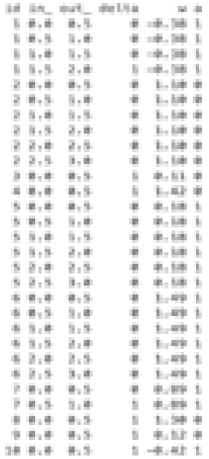
1. Expand dataset to have 1 record per person period



Output: Long dataset

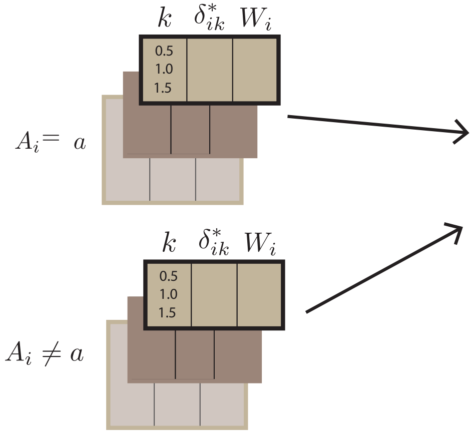


```
id    t d     w a
 1 1.80 1 -0.38 1
 2 3.00 0  1.10 0
 3 0.13 1  0.11 0
 4 0.00 1  1.42 0
 5 3.00 0  0.18 1
 6 3.00 0  1.49 1
 7 0.60 1  0.89 1
 8 0.00 1  1.30 0
 9 0.00 1  0.12 0
10 0.16 1 -0.42 1
```



```
id in_ out_ delta     w a
 1 0.0  0.5     0 -0.38 1
 1 0.5  1.0     0 -0.38 1
 1 1.0  1.5     0 -0.38 1
 1 1.5  2.0     1 -0.38 1
 2 0.0  0.5     0  1.10 0
 2 0.5  1.0     0  1.10 0
 2 1.0  1.5     0  1.10 0
 2 1.5  2.0     0  1.10 0
 2 2.0  2.5     0  1.10 0
 2 2.5  3.0     0  1.10 0
 3 0.0  0.5     1  0.11 0
 4 0.0  0.5     1  1.42 0
 5 0.0  0.5     0  0.18 1
 5 0.5  1.0     0  0.18 1
 5 1.0  1.5     0  0.18 1
 5 1.5  2.0     0  0.18 1
 5 2.0  2.5     0  0.18 1
 5 2.5  3.0     0  0.18 1
 6 0.0  0.5     0  1.49 1
 6 0.5  1.0     0  1.49 1
 6 1.0  1.5     0  1.49 1
 6 1.5  2.0     0  1.49 1
 6 2.0  2.5     0  1.49 1
 6 2.5  3.0     0  1.49 1
 7 0.0  0.5     0  0.89 1
 7 0.5  1.0     1  0.89 1
 8 0.0  0.5     1  1.30 0
 9 0.0  0.5     1  0.12 0
10 0.0  0.5     1 -0.42 1
```

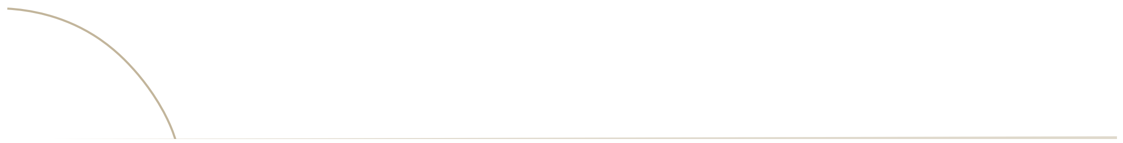
2. Estimate association bewteen W , A, and outcome at each time k using pooled logistic regression



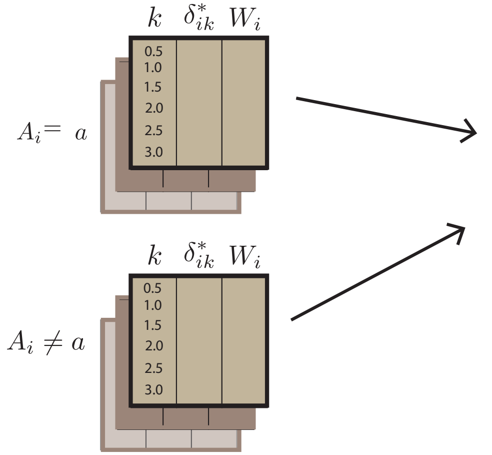
$$\text{logit}[P(\delta^*_{ik} = 1 | T^* > k-1, W_i, A_i; \theta)] = \theta_0 + \theta_1 g(W_i) + \theta_2 A_i + \theta_3 f(k) + \theta_4 A_i f(k)$$

$$\hat{\theta} = \{0.06, 0.27, -0.83, -0.30, -0.75, -0.25, -17.7, -17.4\}$$

Output: $\hat{\theta}$



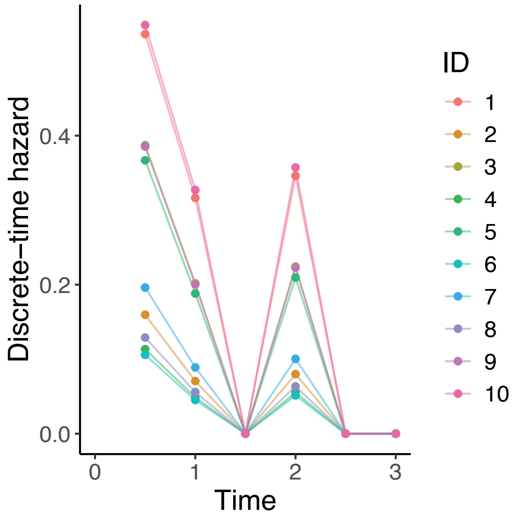
3. Predict discrete-time hazards for each individual under plan a at all time points.



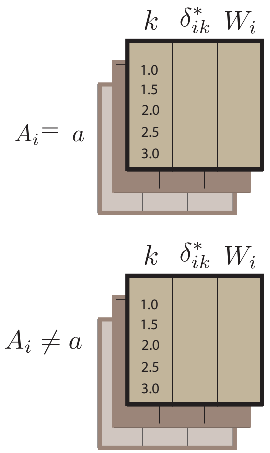
$$\mu_i(k, a, W_i; \hat{\theta}) =$$

$$1/\{1 + \exp[-(\hat{\theta}_0 + \hat{\theta}_1 g(W_i) + \hat{\theta}_2 a + \hat{\theta}_3 f(k) + \hat{\theta}_4 a f(k))]\}$$

$$\boxed{\text{Output: } \mu_i(k, a, W_i; \hat{\theta})}$$

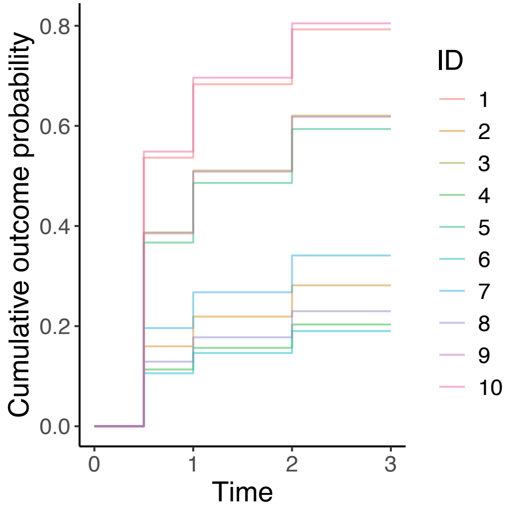


4. Use $\mu_i(k, a, W_i; \hat{\theta})$ to estimate cumulative outcome probability at each time point

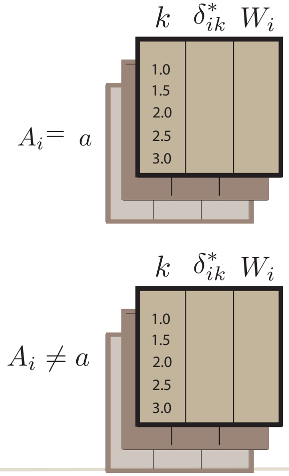


$$h(t, a, W_i; \hat{\theta}) = 1 - \prod_{k \leq t} \left[1 - \mu_i(k, a, W_i; \hat{\theta})\right]$$

Output: $h(t, a, W_i; \hat{\theta})$

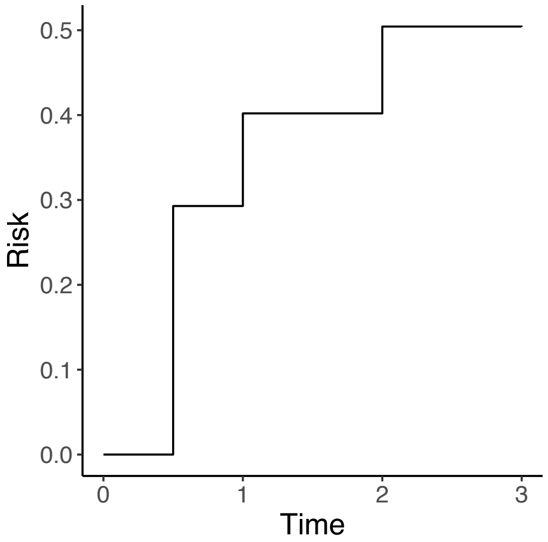


5. Average predicted individual outcomes under plan a across individuals to estimate risk $\tilde{F}^a(t)$



$$\tilde{F}^a(t) = \frac{1}{n} \sum_{i=1}^{n} h(t, a, W_i; \hat{\theta})$$

Output: $\tilde{F}^a(t)$



G-computation with **pooled logistic regression**

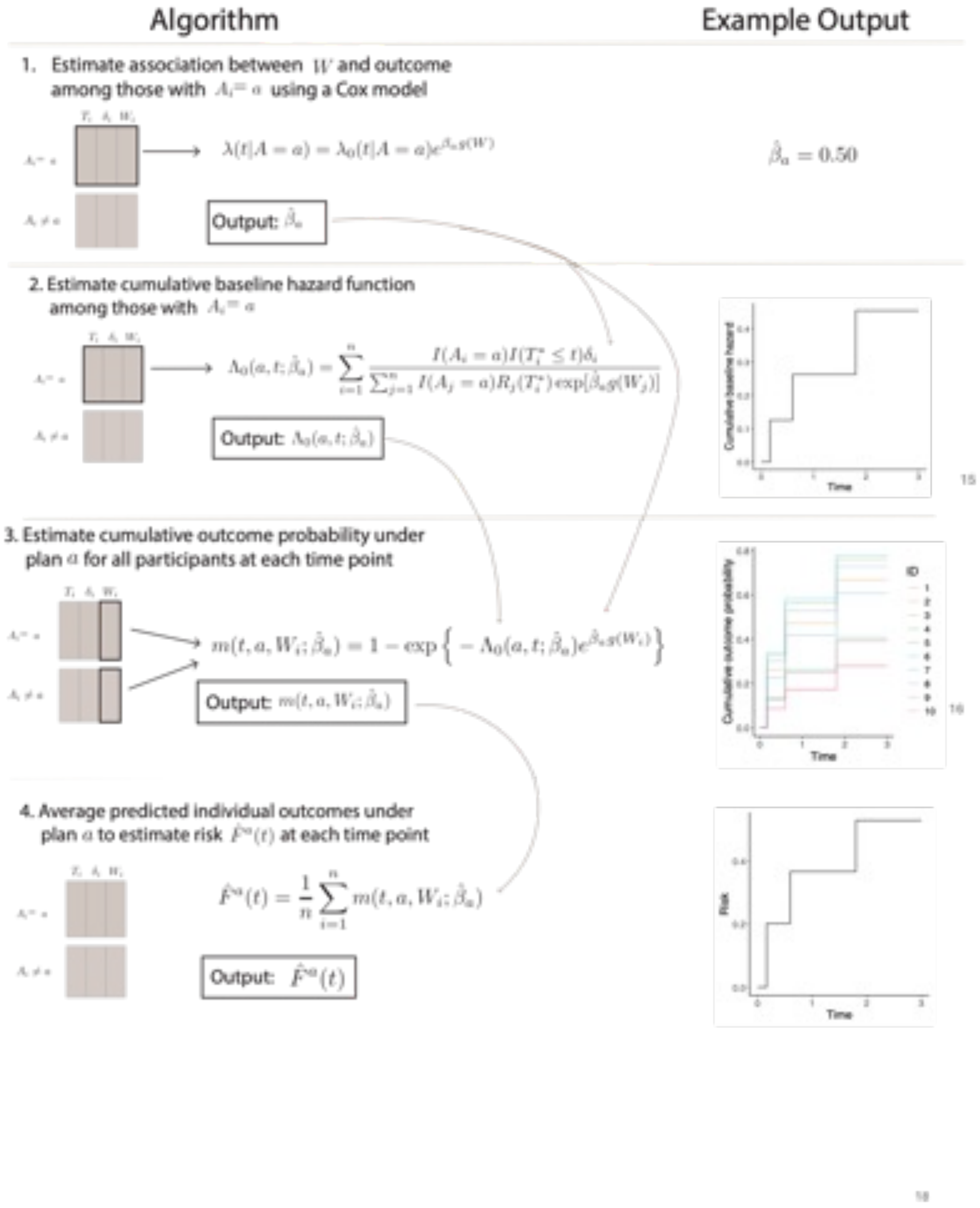
Advantages: easily accounts for confounding and informative censoring affected by time-updated covariates

Disadvantages: relies on choices about the number of time intervals and potentially restrictive parametric models, vulnerable to bias due to discretization of time and model misspecification.

Moreover, requires onerous dataset expansions, which create opportunities for computational issues and user error.

The **Breslow** g-computation estimator avoids these disadvantages.

- No need to expand dataset
- Continuous time
- Semiparametric (does not require parametric model for baseline hazard function)



Algorithm

Example Output

1. Estimate association between W and outcome among those with $A_i = a$ using a Cox model

$$\lambda(t|A=a) = \lambda_0(t|A=a)e^{\beta_a g(W)}$$

Output: $\hat{\beta}_a$

$$\hat{\beta}_a = 0.50$$

2. Estimate cumulative baseline hazard function among those with $A_i = a$

$$\Lambda_0(a,t;\hat{\beta}_a) = \sum_{i=1}^{n} \frac{I(A_i = a)I(T_i^* \le t)\delta_i}{\sum_{j=1}^{n} I(A_j = a)R_j(T_i^*)\exp[\hat{\beta}_a g(W_j)]}$$

Output: $\Lambda_0(a,t;\hat{\beta}_a)$

3. Estimate cumulative outcome probability under plan a for all participants at each time point

$$m(t,a,W_i;\hat{\beta}_a) = 1 - \exp\left\{-\Lambda_0(a,t;\hat{\beta}_a)e^{\hat{\beta}_a g(W_i)}\right\}$$

Output: $m(t,a,W_i;\hat{\beta}_a)$

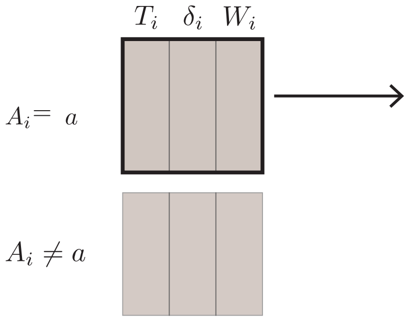
4. Average predicted individual outcomes under plan a to estimate risk $\hat{F}^a(t)$ at each time point

$$\hat{F}^a(t) = \frac{1}{n}\sum_{i=1}^{n} m(t,a,W_i;\hat{\beta}_a)$$

Output: $\hat{F}^a(t)$

Algorithm

1. Estimate association between W and outcome among those with $A_i = a$ using a Cox model



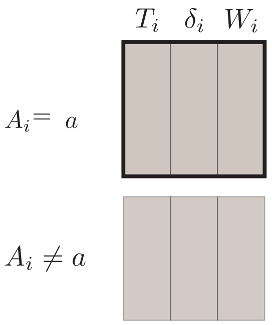
$$\lambda(t|A=a) = \lambda_0(t|A=a)e^{\beta_a g(W)}$$

Output: $\hat{\beta}_a$

Example Output

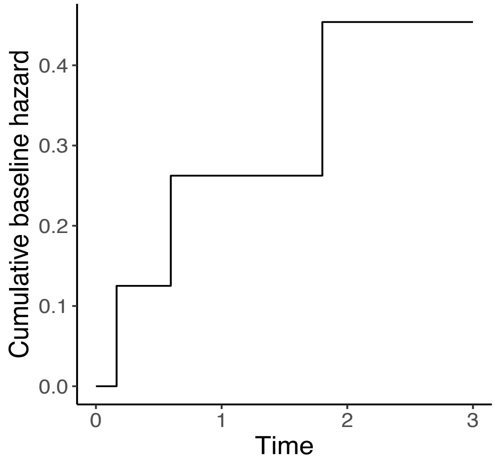
$$\hat{\beta}_a = 0.50$$

2. Estimate cumulative baseline hazard function among those with $A_i = a$

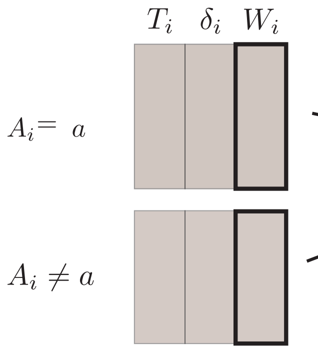


$$\longrightarrow \Lambda_0(a, t; \hat{\beta}_a) = \sum_{i=1}^{n} \frac{I(A_i = a) I(T_i^* \leq t)\delta_i}{\sum_{j=1}^{n} I(A_j = a) R_j(T_i^*) \exp[\hat{\beta}_a g(W_j)]}$$

$\boxed{\text{Output: } \Lambda_0(a, t; \hat{\beta}_a)}$

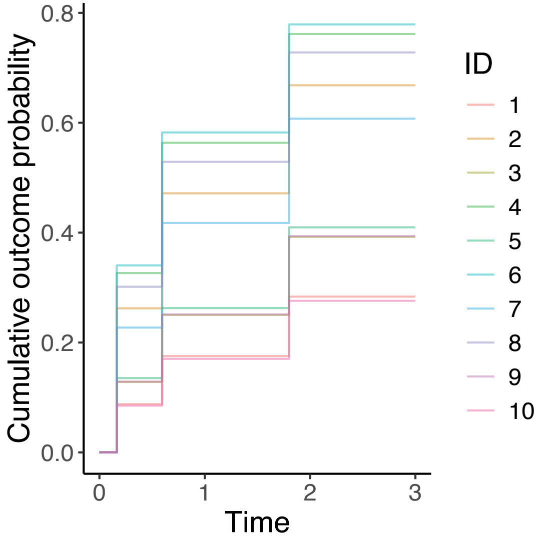


3. Estimate cumulative outcome probability under plan a for all participants at each time point

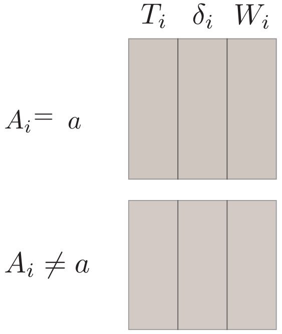


$$m(t, a, W_i; \hat{\beta}_a) = 1 - \exp\left\{ - \Lambda_0(a, t; \hat{\beta}_a) e^{\hat{\beta}_a g(W_i)} \right\}$$

Output: $m(t, a, W_i; \hat{\beta}_a)$

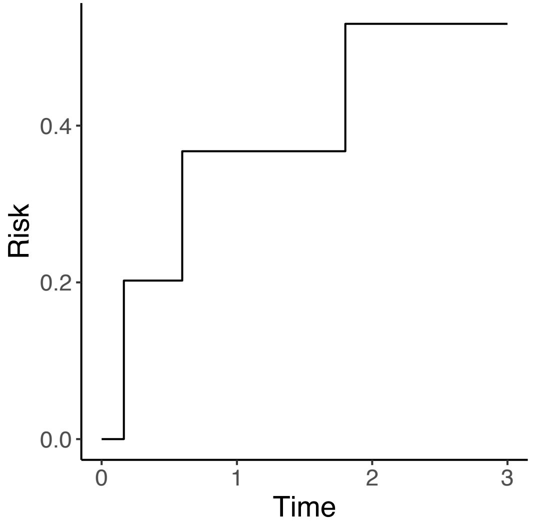


4. Average predicted individual outcomes under plan a to estimate risk $\hat{F}^a(t)$ at each time point

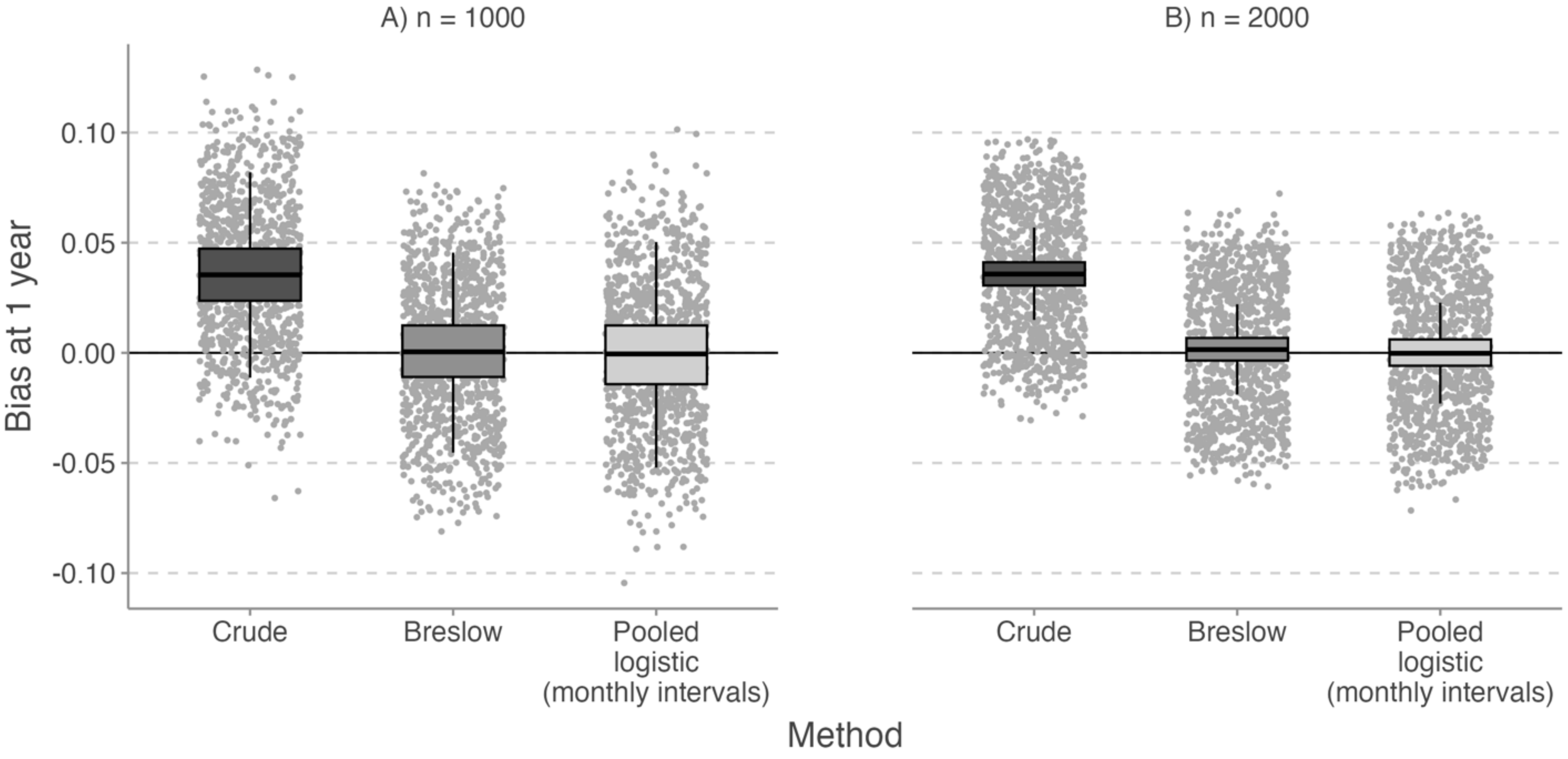


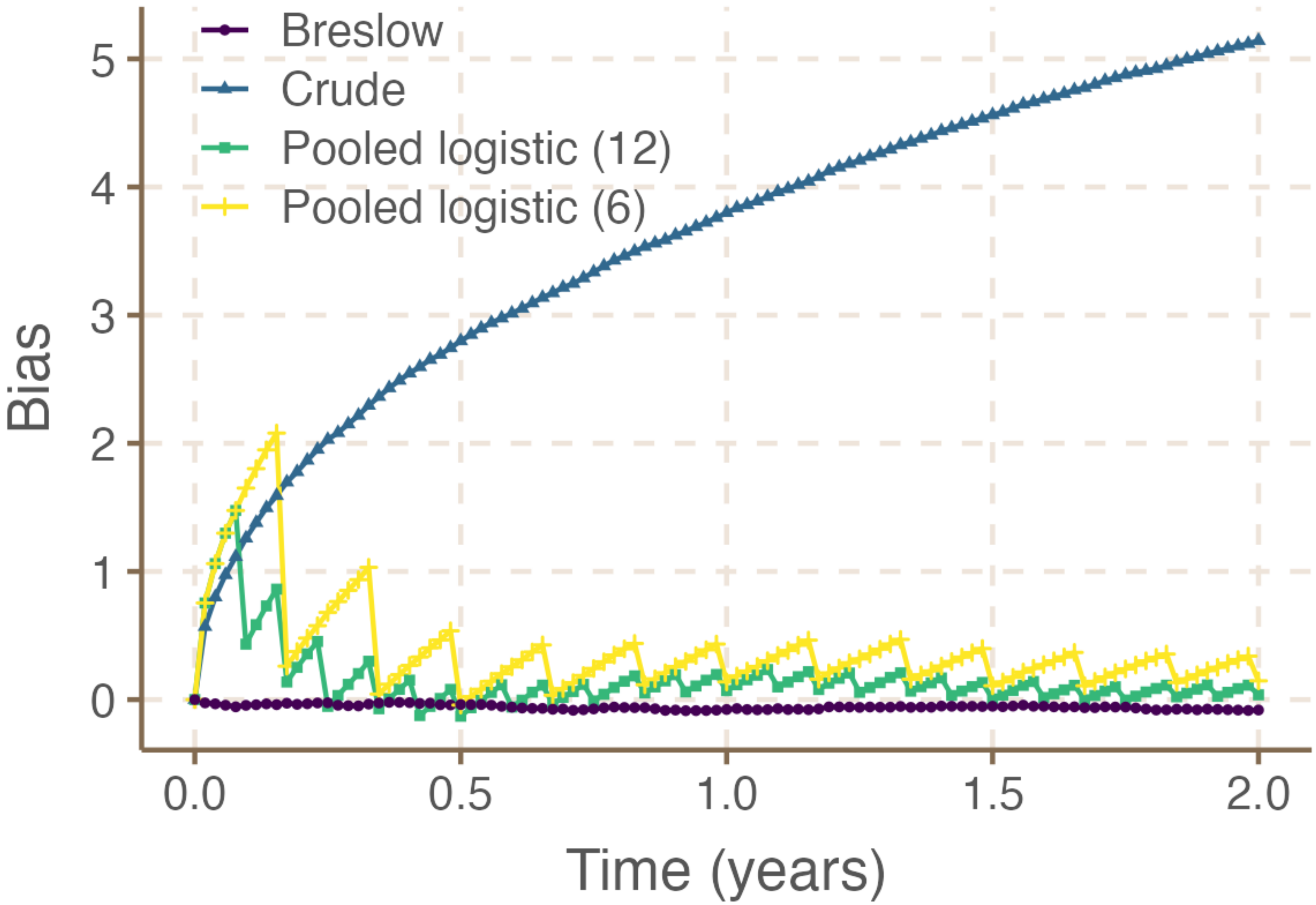
$$\hat{F}^a(t) = \frac{1}{n}\sum_{i=1}^{n} m(t, a, W_i; \hat{\beta}_a)$$

Output: $\hat{F}^a(t)$



Simulation results



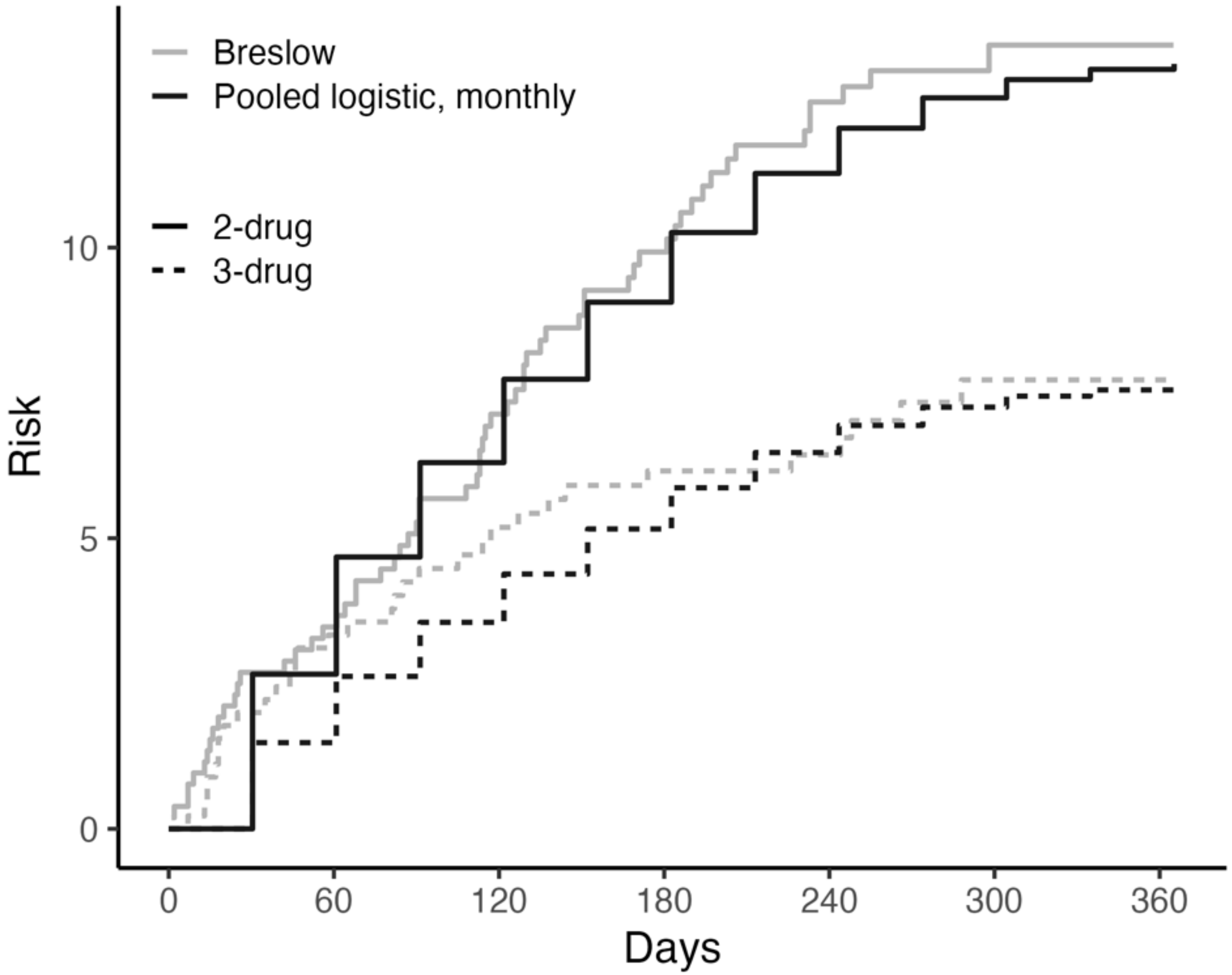

Breslow
Crude
Pooled logistic (12)
Pooled logistic (6)
Bias
5
4
3
2
1
0
0.0
0.5
1.0
1.5
2.0
Time (years)

Example: 3-drug vs 2-drug ART for people with HIV

Table. Modified version of the ACTG 320 trial, distorted to induce confounding.

	Overall (N=978)		2-drug (n = 579)		3-drug (n = 399)	
Characteristic	n	%	n	%	n	%
Age > 40	384	39	232	40	152	38
Black race	286	29	165	28	121	30
Hispanic ethnicity	173	18	106	18	67	17
Injection drug use	156	16	93	16	63	16
CD4<100	699	71	364	63	335	84
Male sex	813	83	485	84	328	82

Figure 5. Estimated risk functions (x 100) under 2-drug (solid lines) and 3-drug (dotted lines) antiretroviral therapy regimens in a modified version of the ACTG 320 data using the pooled logistic g-computation estimator with time discretized to the month (black lines) and the Breslow g-computation estimator (grey lines).



Summary

1. G-computation is useful

2. Standard “pooled logistic” approach is ideal for settings with time-varying covariates, but also subject to disadvantages
    1. Discretization of time
    2. Parametric models
    3. Need to expand dataset

3. Breslow g-computation removes these disadvantages in settings with time-fixed exposure and a survival outcome.

Data and code available at

`https://github.com/edwardsjk/semiparametric_gcomp`

Semiparametric g-computation to account for baseline confounding in observational studies with survival endpoints

Jessie K. Edwards

University of North Carolina at Chapel Hill

jessedwards@unc.edu