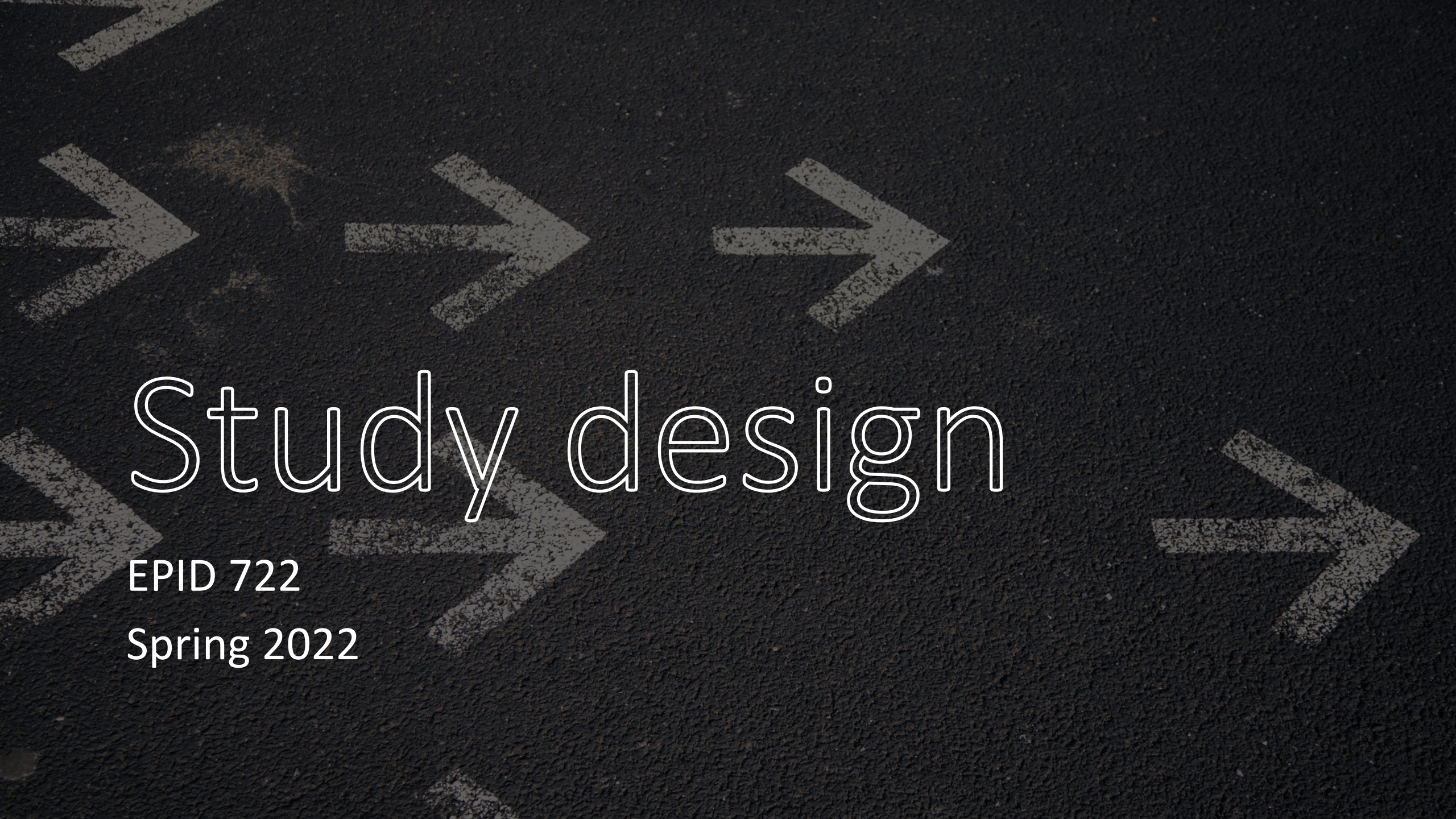

Study design
EPID 722
Spring 2022

Notation

T_i: Time from origin to outcome of interest or censoring

δ_i: Event indicator (1=Event; 0=Censoring)

W_i: Time from origin to study entry

$F(t)$: risk at time t

t indexes time from origin

k: rank of the event times

R_k: Event time of rank k

Estimators

Risk [a]

$$\hat{F}(t) = 1 - \prod_{k:R_k \leq t} \left[1 - \frac{d_k}{n_k}\right],$$

where $d_k = \sum_{i=1}^{N} I(T_i = R_k)\delta_i$ and $n_k = \sum_{i=1}^{N} I(R_k \leq T_i)$

[a] With no competing events

Simple example

Source:

Cole SR, Hudgens MG. Survival analysis in infectious disease research: describing events in time: *AIDS*. 2010;24(16):2423–2431

Say we'd like to estimate the risk of death after AIDS diagnosis in a cohort of men with HIV in the US.

Origin: AIDS diagnosis

T_i: time from AIDS diagnosis to death or censoring

W_i: time from AIDS diagnosis until study entry

Return to the example data

Design:

Enroll 42 men alive on 1 January 1995 with a prior clinical AIDS diagnosis and enroll 36 additional men with a clinical AIDS diagnosis between 1 January 1995 and 1 January 1998

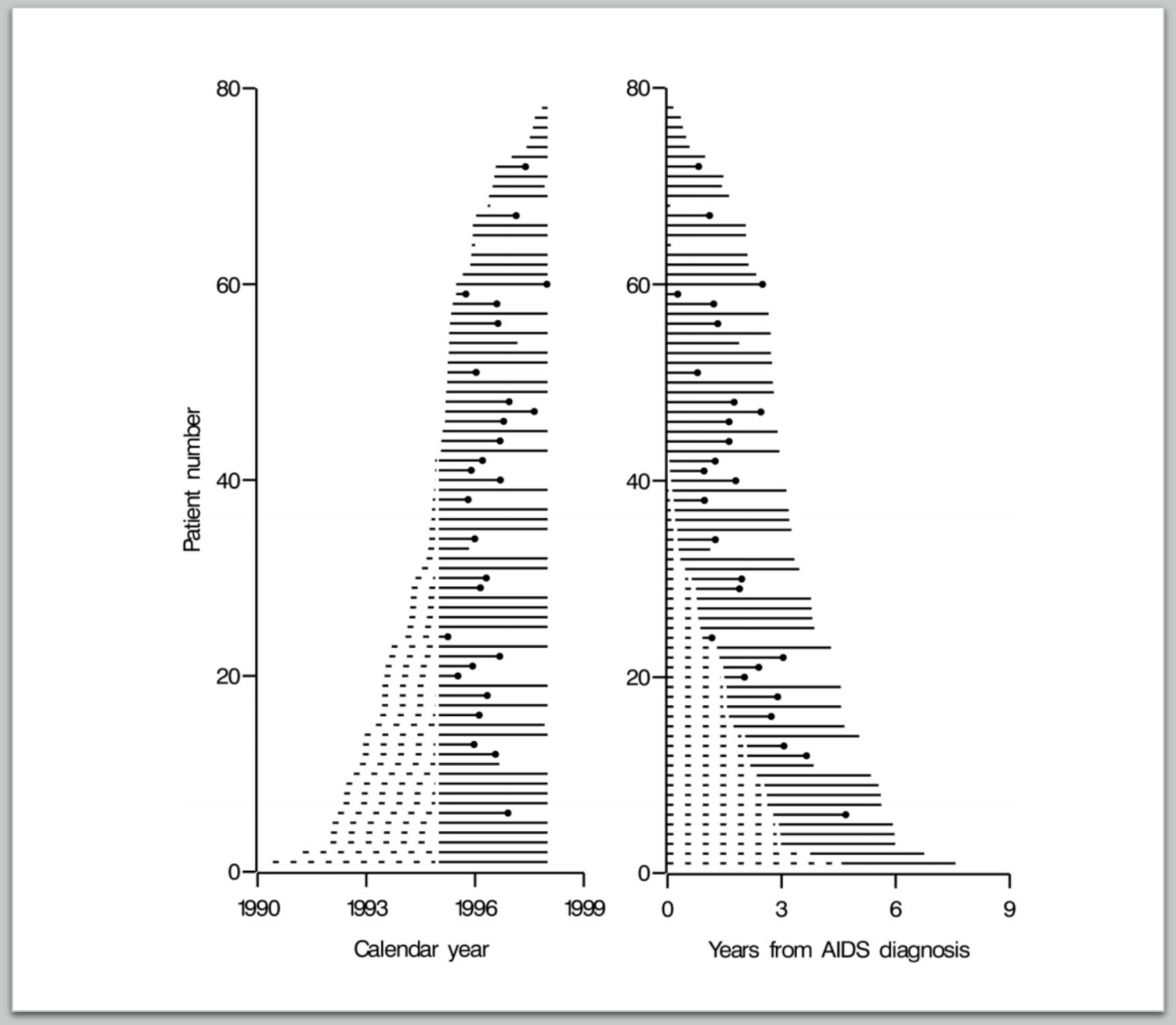
Follow all 78 (=42+36) men for all-cause mortality through 1 January 1998, the date of study completion.

Parameter of interest is risk of death after AIDS diagnosis in this cohort.

Simple example

Source:

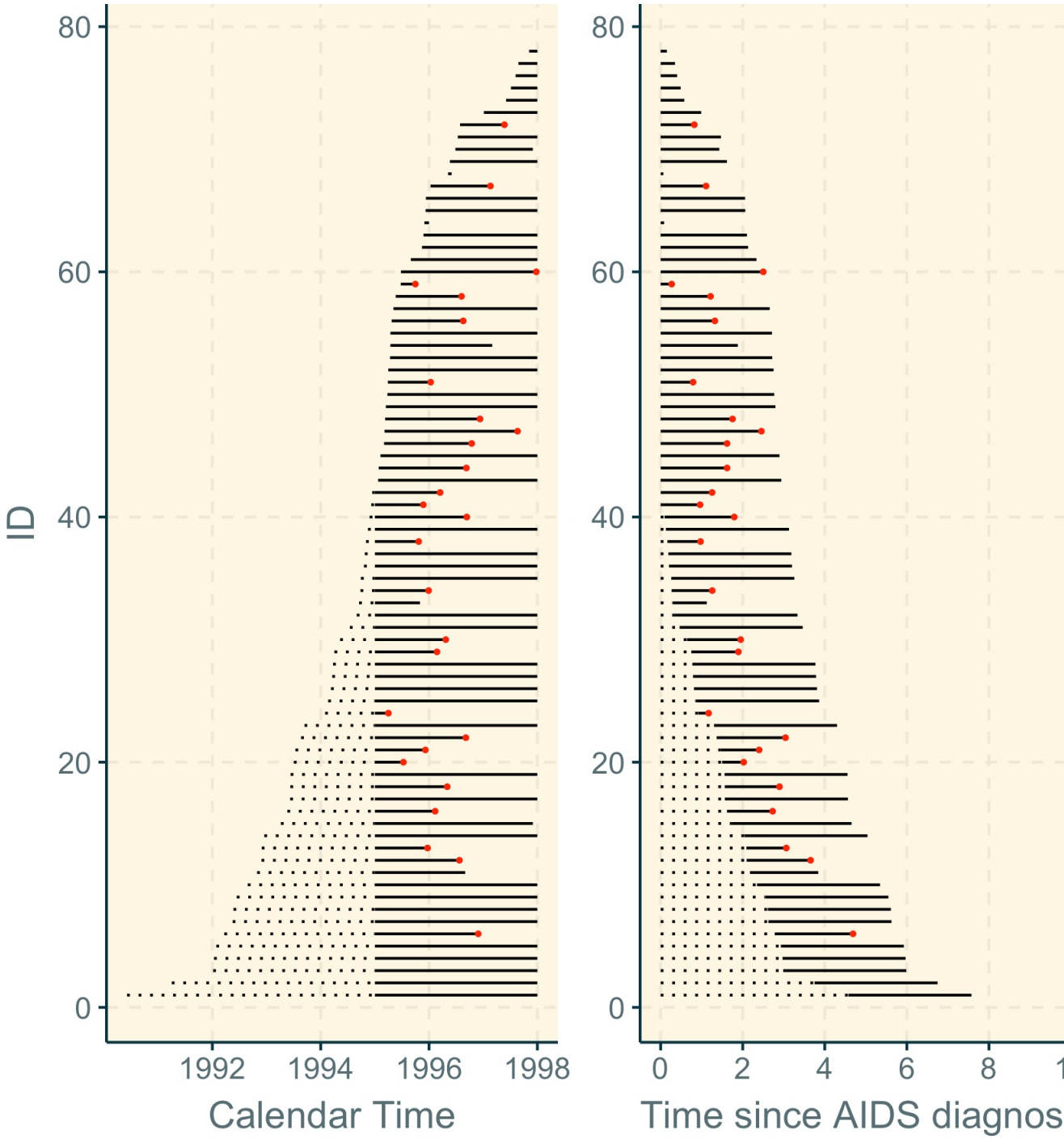
Cole SR, Hudgens MG. Survival analysis in infectious disease research: describing events in time: *AIDS*. 2010;24(16):2423–2431.



Simple example

Source:

Cole SR, Hudgens MG. Survival analysis in infectious disease research: describing events in time: *AIDS*. 2010;24(16):2423–2431.



4 possibilities

1. Use only the 36 who entered at the origin (i.e., incident AIDS cases, or those with $W_i = 0$)
2. Ignore immortal person time between the origin and W_i
3. Use time on study as the timescale (define $V_i = T_i - W_i$)
4. Late enter the people with prevalent AIDS at time W_i

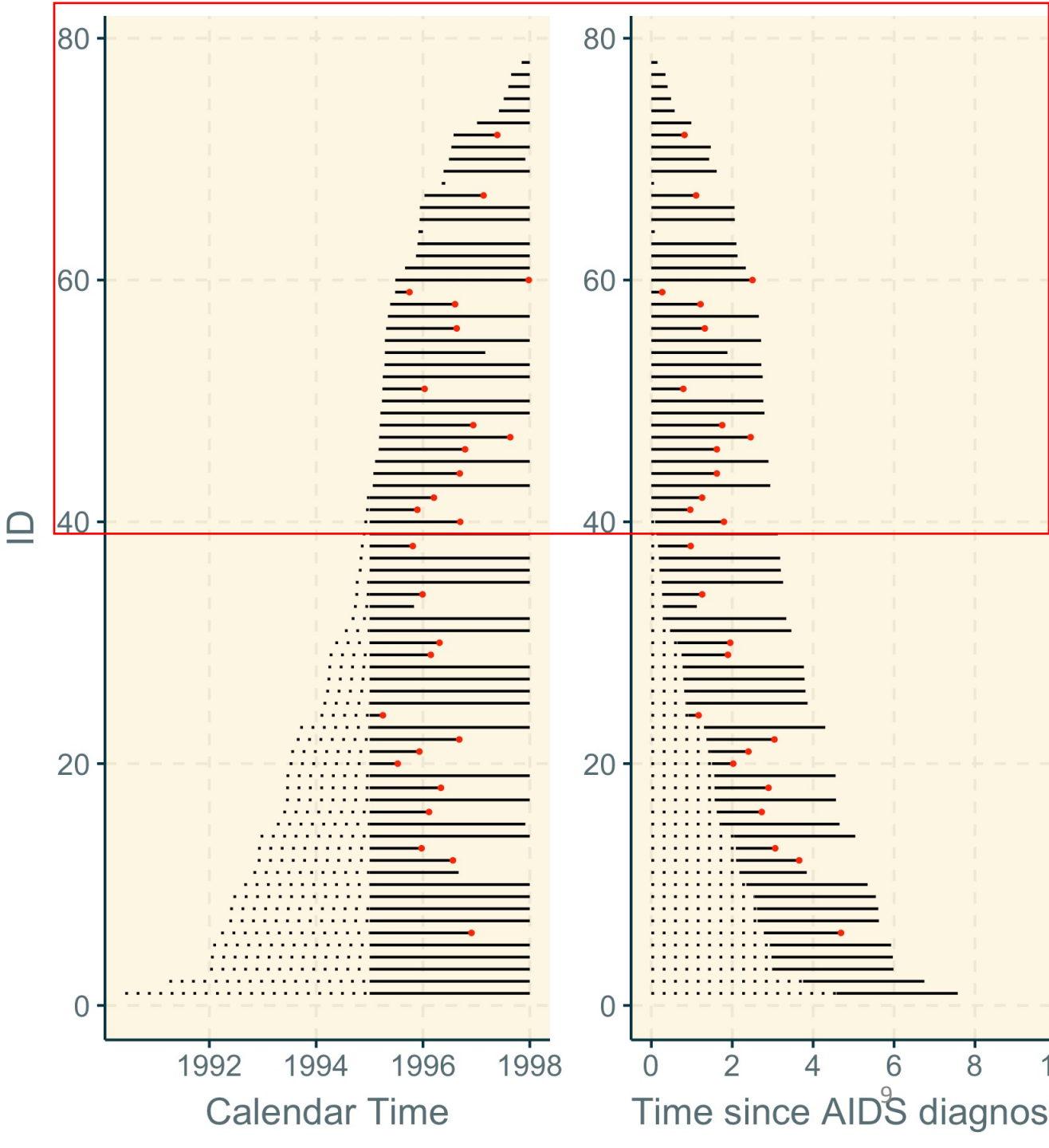
Option 1: Conduct the study among the 36 who entered at the origin

Limit the cohort to those with $W_i = 0$

Limitations:

a. Scenarios with few new users

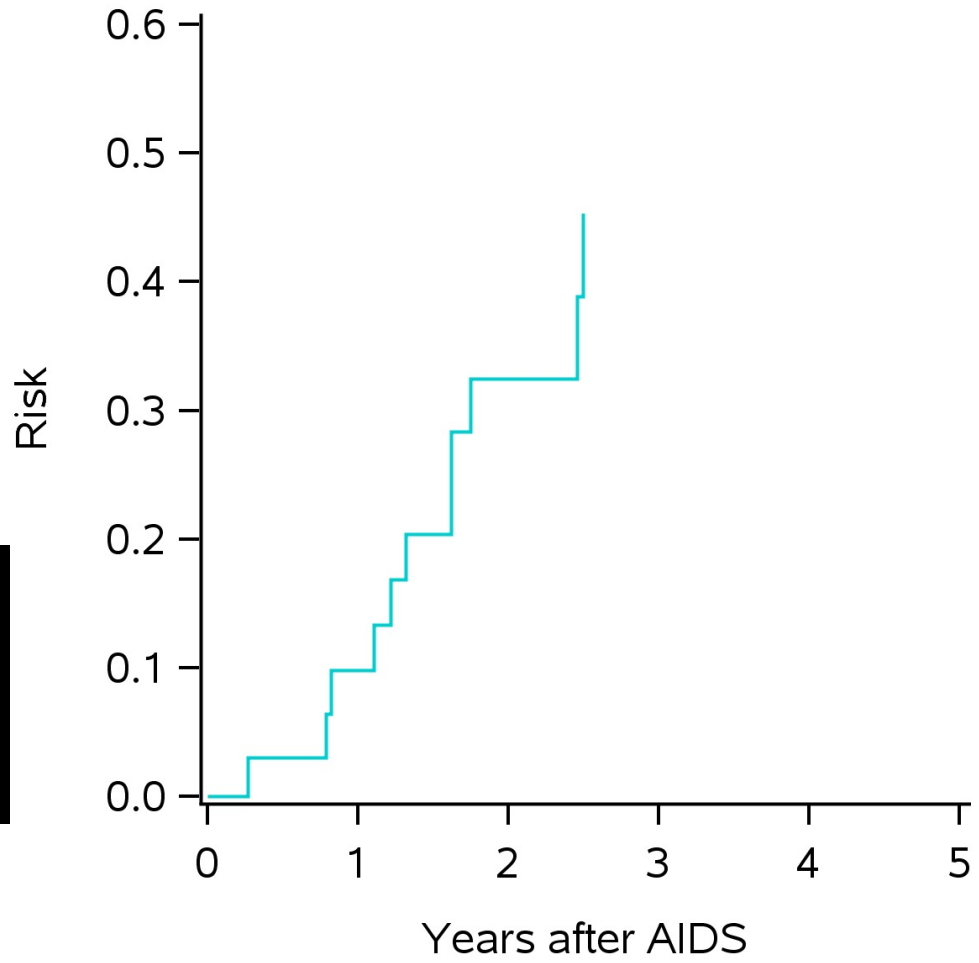
b. Limited follow-up time



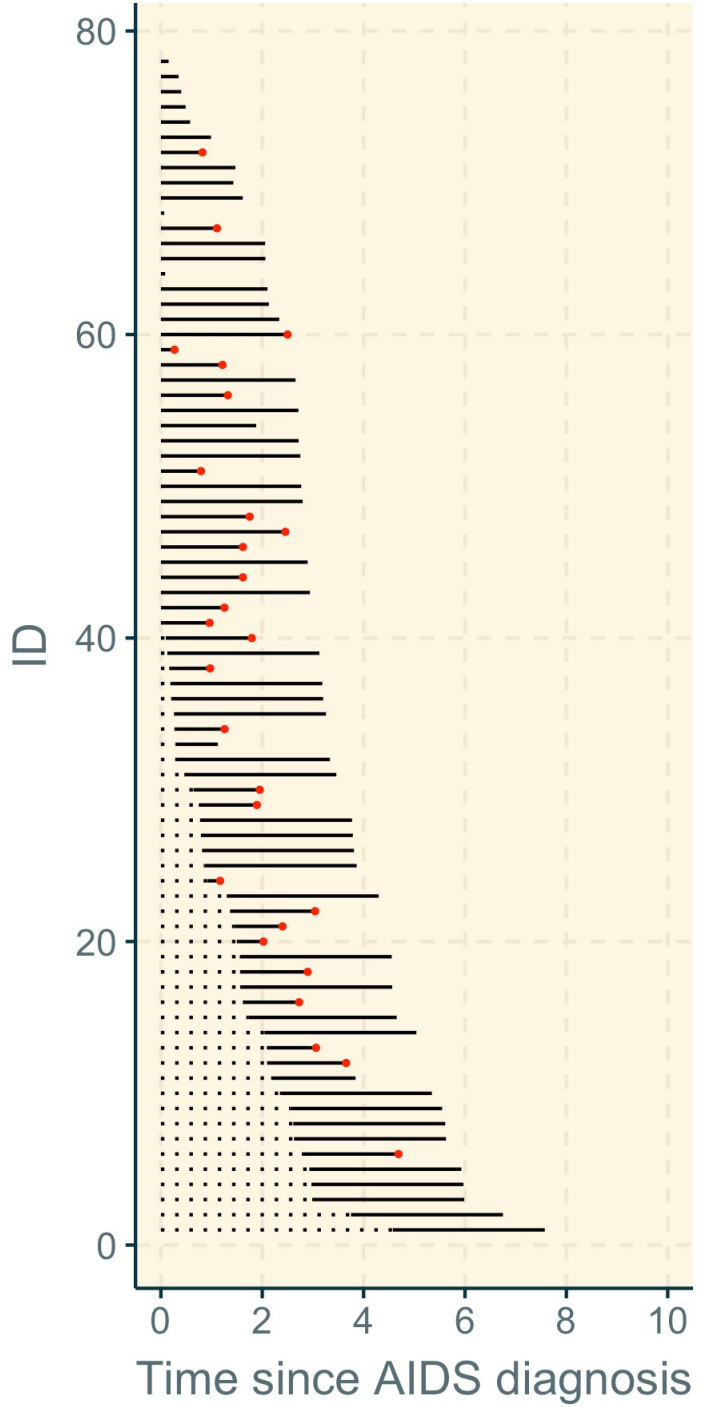
Option 1: Conduct the study among the 36 “New Users”

```
proc phreg data=nu; where w = 0;
       model t*d(0)=;
       baseline out=km survival=s;
run;
data km; set km; r=1-s; run;
```

```
km <- survfit(Surv(t, d) ~ 1,
              data = nu %>% filter(w == 0))
r <- 1-summary(km)$surv
```



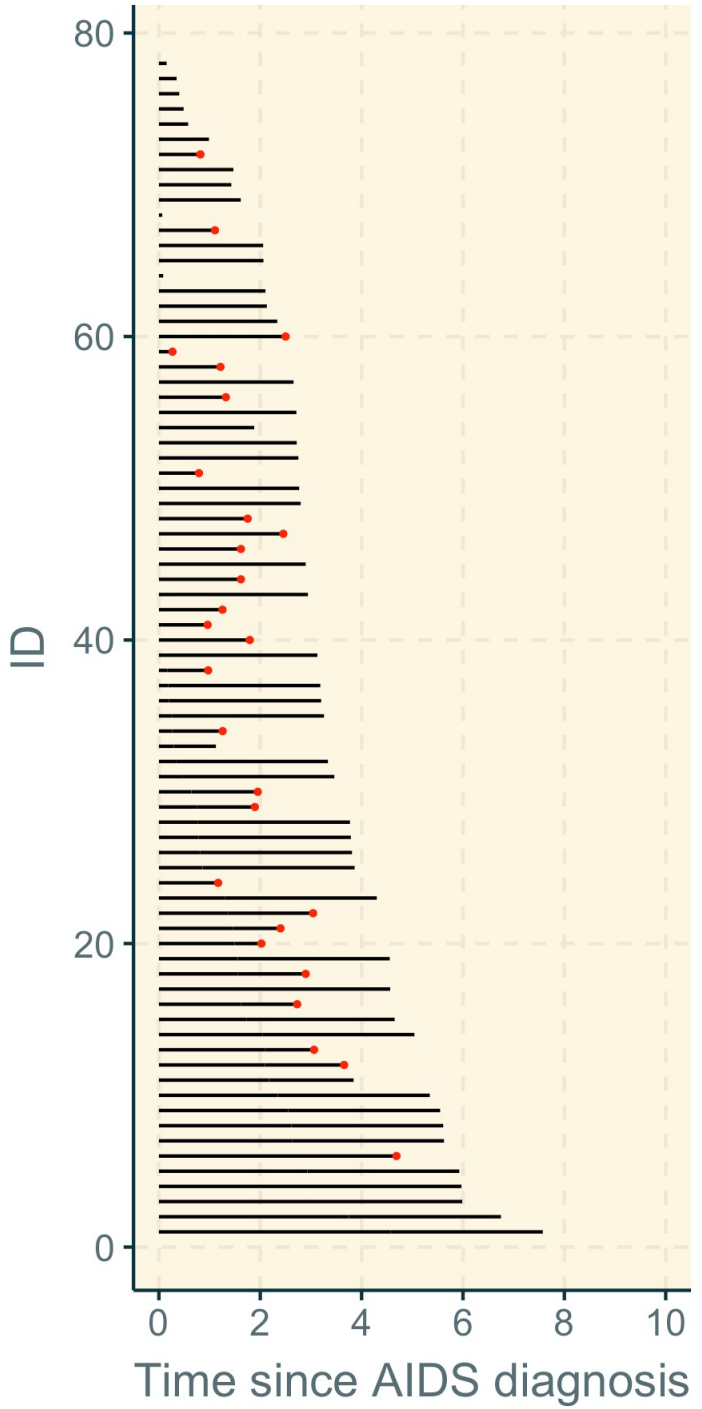
Option 2: Ignore immortal person time



Option 2: Ignore immortal person time

```
proc phreg data=nu;
       model t*d(0)=;
       baseline out=km survival=s;
run;
data km; set km; r=1-s; run;
```

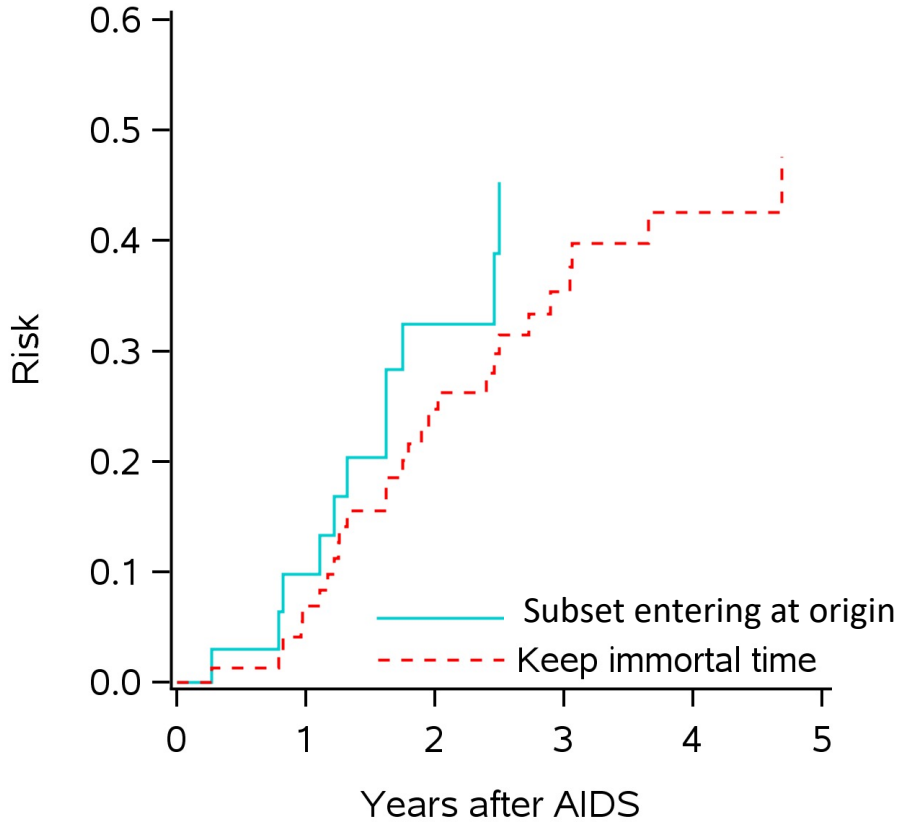
```
km <- survfit(Surv(t, d) ~ 1,
              data = nu)
r <- 1-summary(km)$surv
```

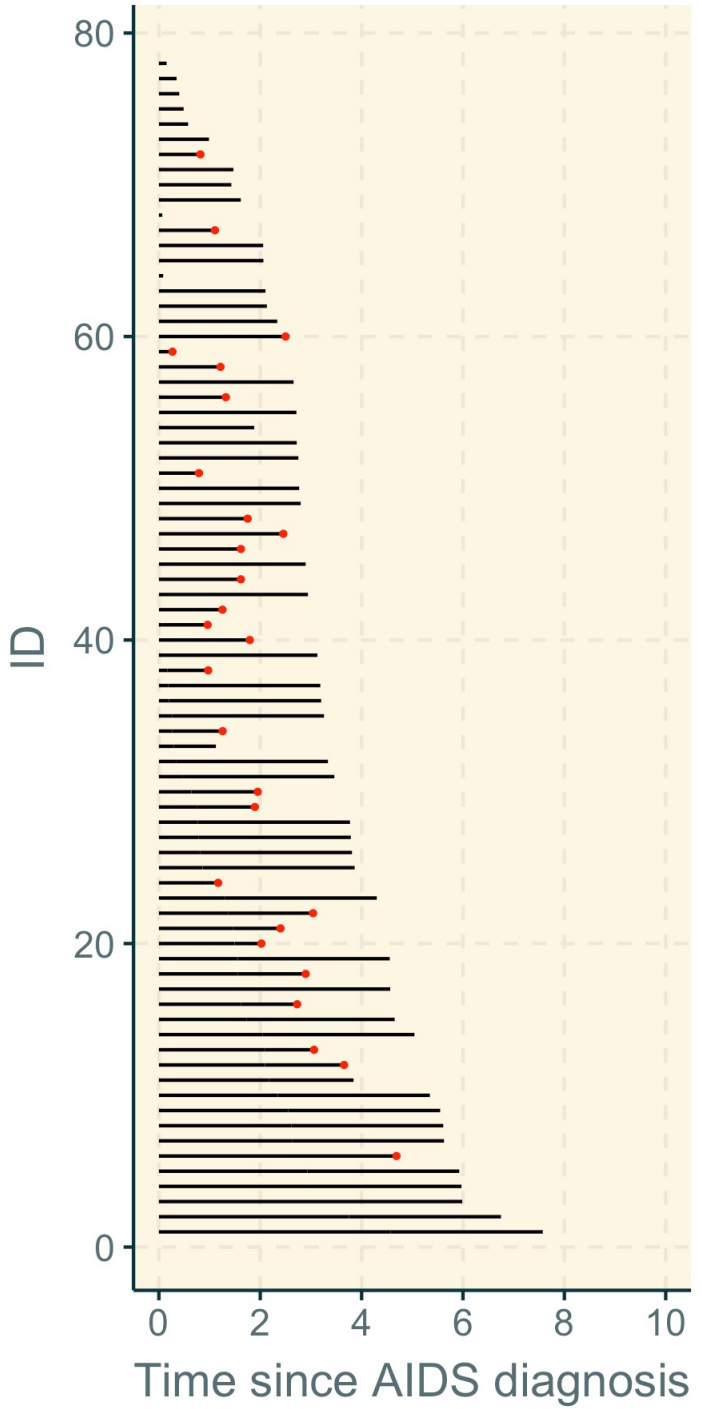


Option 2: Ignore immortal person time

Time between AIDS diagnosis and study entry is immortal

Assumes hazard prior to study entry is 0





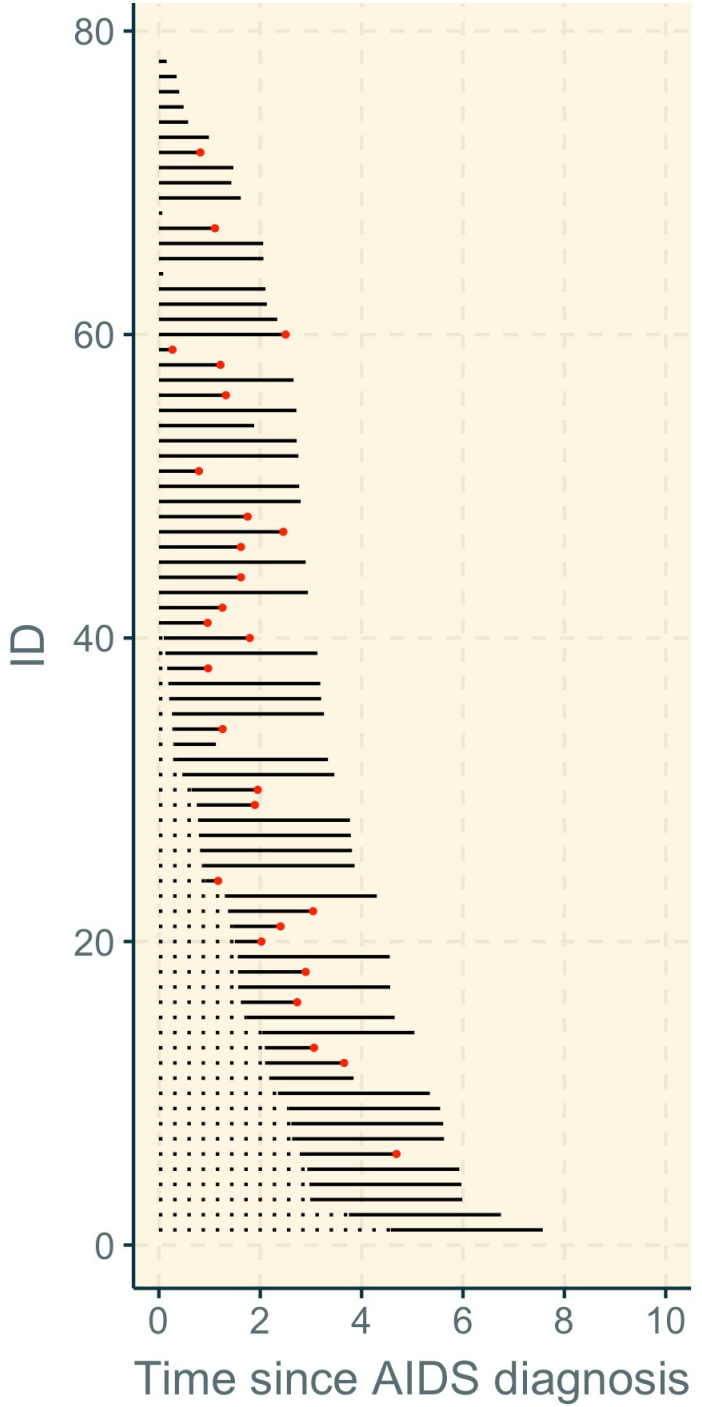
Option 3: Use time on study as the timescale

Reset timescale to be “time on study.”

Makes sense in some settings where study enrollment concurs with an important clinical decision point.

But changing the timescale changes the study question.

i.e., from “What is the risk of death 5 years after AIDS dx?” to “What is the risk of death 5 years after entering the study?



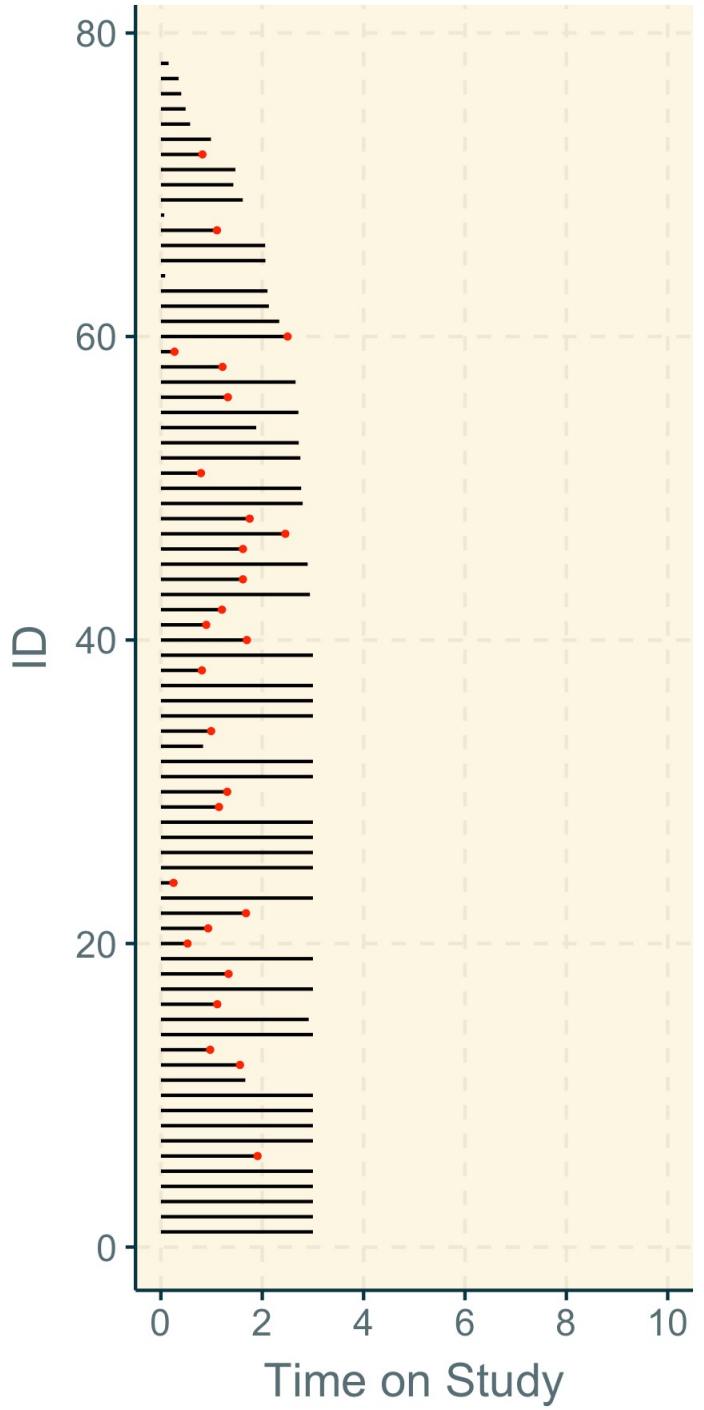
Option 3: Use time on study as the timescale

```
data nu;
   set nu;
   new_t = t - w;
run;


proc phreg data=nu;
        model new_t*d(0)=;
        baseline out=km survival=s;
run;
data km; set km; r=1-s; run;
```

```
nu$new_t <- nu$t - nu$w

km <- survfit(Surv(new_t, d) ~ 1,
              data = nu)
r <- 1-summary(km)$surv
```



Option 4: Late enter those who arrive after the origin at W_i

We commonly allow people who enter the study after the origin to be "late entries" at time W_i.

To estimate risk, we use an "extended Kaplan-Meier estimator"

Why extended?

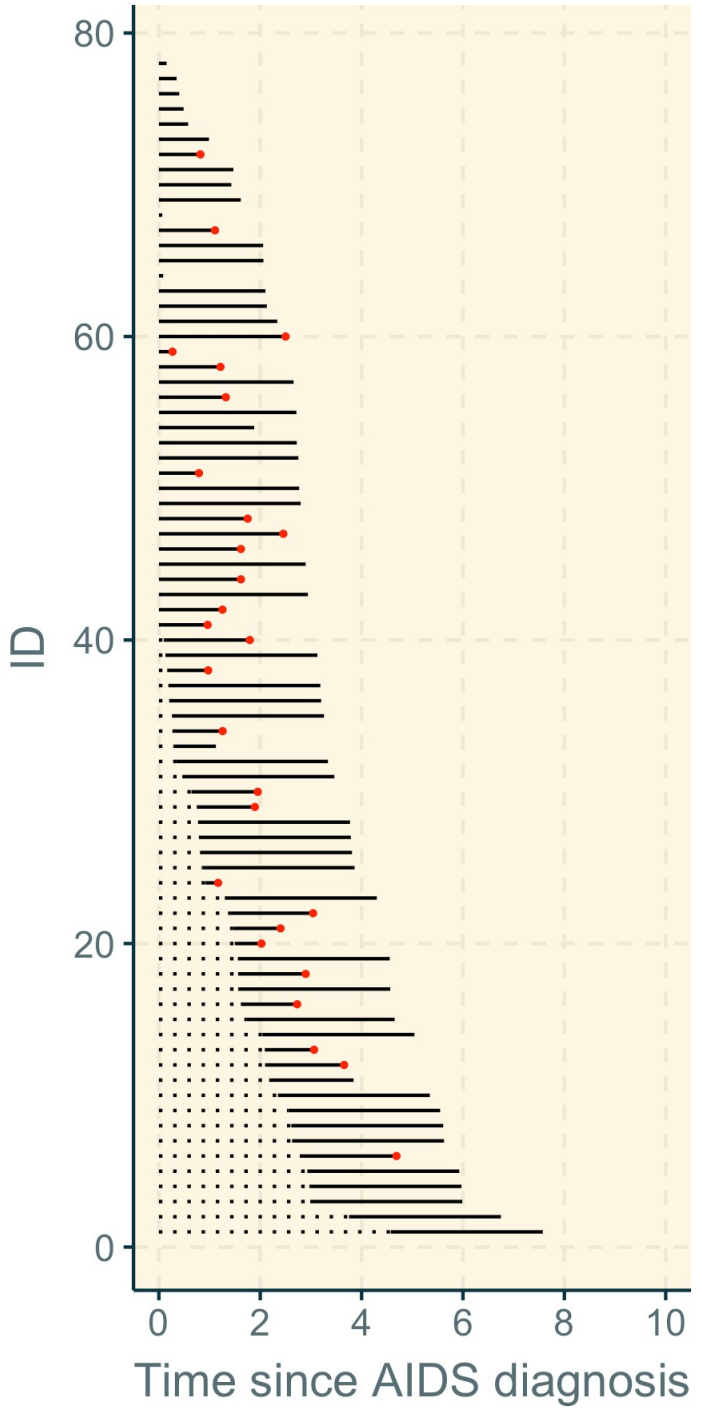
Recall, $\hat{F}(t) = 1 - \prod_{k:R_k \leq t}\left[1 - \frac{d_k}{n_k}\right]$, where $d_k = \sum_{i=1}^{N} I(T_i = R_k)\delta_i$ and $n_k = \sum_{i=1}^{N} I(R_k \leq T_i)$

Extend, $\widehat{F}(t)_2 = 1 - \prod_{k:R_k \leq t}\left[1 - \frac{d_k}{n_k}\right]$, where $d_k = \sum_{i=1}^{N} I(T_i = R_k, W_i < R_k)\delta_i$ and $n_k = \sum_{i=1}^{N} I(W_i < R_k \leq T_i)$

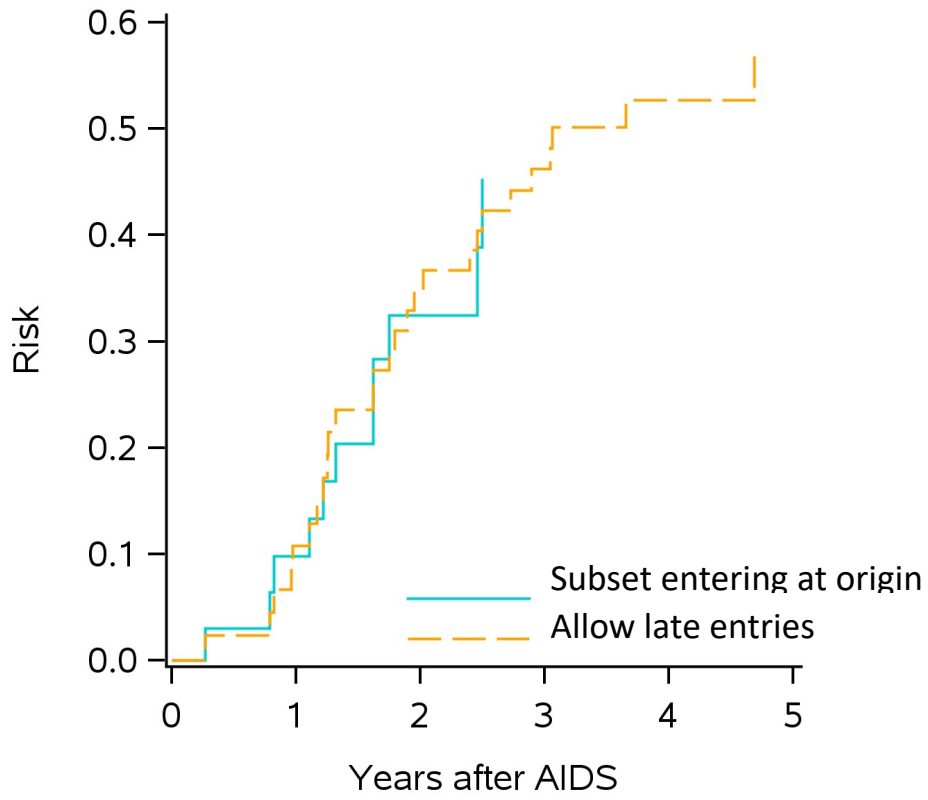
Option 4: Late enter those who arrive after the origin at W_i

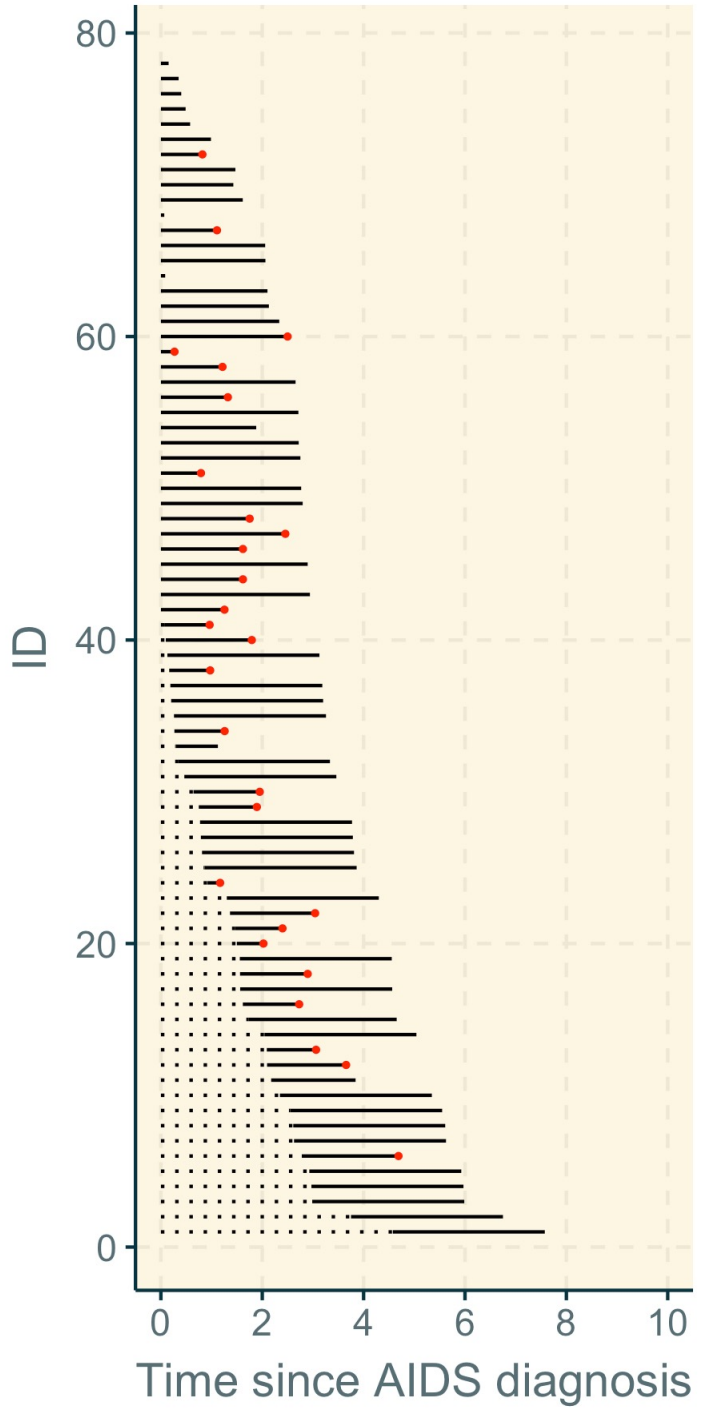
```
proc phreg data=nu;
       model (w, t)*d(0)=;
       baseline out=km survival=s;
run;
data km; set km; r=1-s; run;
```

```
km <- survfit(Surv(w, t, d) ~ 1,
              data = nu)
r <- 1-summary(km)$surv
```



Option 4: Late enter those who arrive after the origin at W_i






Questions and target trials

Study question	Description
Aim	What is the purpose of the study
Significance	Why is it important?
Target population	Who would be affected by any decisions made as a result of the study?
Origin	What is the decision point? That is, what is the indicator that the action or decision being evaluated is relevant to someone in the target population? When does follow-up begin?
Timescale	What are the units of follow-up time? When should follow up end? What is on the x-axis of your line diagram?
Actions (if any)	Are we estimating effects of any interventions, treatments, exposures, or other actions?
Groups (if any)	Are we comparing people with different characteristics?
Outcomes	What endpoint are we describing? What is the primary outcome? Are any secondary outcomes or competing events of interest?
Parameter of interest	How will outcomes be summarized (and compared)? At what timepoint?
Impact	How will results of this study improve public health, clinical practice, or the world?

Idealized studies

Study question
Aim
Significance
Target population
Origin
Timescale
Actions (if any)
Groups (if any)
Outcomes
Parameter of interest
Impact

Study question	Idealized descriptive study
Aim	Aim
Significance	Significance
Target population	Eligibility (person, place, time)
	Sampling
	Recruitment
Origin	When does follow-up begin?
Timescale	What is the maximum length of follow-up?
Actions (if any)	
Groups (if any)	When and how are groups measured?
Outcomes	Which outcomes are of interest?
	When are outcomes measured?
Parameter of interest	Parameter of interest
Impact	How will results be presented?
	How will results be used?

Study question	Idealized descriptive study	Target trial
Aim	Aim	Aim
Significance	Significance	Significance
Target population	Eligibility (person, place, time)	Eligibility (person, place, time)
	Sampling	Sampling
	Recruitment	Recruitment
Origin	When does follow-up begin?	When does follow-up begin?
Timescale	What is the maximum length of follow-up?	What is the maximum length of follow-up?
Actions (if any)		What actions are being considered?
		How are actions assigned?
		When are actions assigned?
Groups (if any)	When and how are groups measured?	When and how are groups measured?
Outcomes	Which outcomes are of interest?	Which outcomes are of interest?
	When are outcomes measured?	When are outcomes measured?
Parameter of interest	Parameter of interest	Parameter of interest
Impact	How will results be presented?	How will results be presented?
	How will results be used?	How will results be used?

Target trial details (for reference)

Study question	Idealized study/target trial	Details
Aim	Aim	What is the purpose of the study?
Significance	Significance	Why is it important?
Target population	Eligibility	Who is in the sampling frame for the study? Are there characteristics that would exclude people from the study? Should include person, place, and time
	Sampling	How will you select individuals to be included in the study? *For the project, assume that it is not possible to include a census of the target population if target is over 10,000 people.*
	Recruitment	How will you enroll people into the study?
Origin	When does follow-up begin?	When, on the timescale of interest, will people be enrolled into the study?
Timescale	What is the maximum length of follow-up of interest?	How long will people be followed in the idealized study?

Study question	Idealized study/target trial	Details
Actions (if any)	What actions are being considered?	What decision is being evaluated? What actions are considered? At what level do actions occur (individual, household, community, etc)?
	How are actions assigned?	How will investigators select which participants are assigned to which actions? Will actions be assigned at the individual or cluster level?
	When are actions assigned?	When will actions be assigned (on the timescale of interest)?
Groups (if any)	When and how are groups measured?	Will results be stratified by individual or cluster level characteristics? When will group membership be defined? How will group membership be measured?
Outcomes	Which outcomes are of interest?	What is the primary outcome? Are any secondary outcomes or competing events of interest? Will tradeoffs between the primary outcome and any competing events/secondary endpoints be considered?
	When are outcomes measured?	Will outcomes be assessed prospectively, at regular intervals, or only at the end of follow-up?
Parameter of interest	Parameter of interest	What statistical parameter will quantify the causal effect of interest?
Impact	How will results be presented?	Include at least one sample figure and one sample table summarizing results of your study
	How will results be used?	How will the results you present serve as inputs into decision making processes?

Examples

We will consider 2 examples:

- What is the risk of COVID-19 among elementary school kids in North Carolina
- What is the effect of an intervention to reduce occupational radon exposure among underground uranium miners?

Example 1

Study question	Example question	Study component	Example idealized study
Aim	To determine risk of COVID-19 among children in elementary school in NC	Aim	
Significance	While kids are at low risk of severe outcomes, high disease incidence implies non-negligible risk for severe outcomes. Moreover, high incidence in kids may pose transmission risk to vulnerable adults.	Significance	
Target population	Children in elementary school in North Carolina	Eligibility	
		Sampling	
		Recruitment	

Study question	**Example question**	**Study component**	**Example idealized study**
Origin	First day of fall semester 2021	When does follow-up begin?	
Timescale	Days since start of fall semester, ending at the end of spring semester, 2022	What is the maximum length of follow-up of interest?	
Actions	None	What actions are being considered?	
		How are actions assigned?	
		When are actions assigned?	
Groups	None	When and how are groups measured?	

Outcomes	First COVID-19 diagnosis	Which outcomes ?	
		When are outcomes measured?	
Parameter of interest	Risk function for COVID-19 diagnosis over time	Parameter of interest	
Impact	Results will assist serve as a baseline for future intervention planning strategies	How will results be presented?	
		How will results be used?	

Example 2

Study question	Example question	Study component	Example idealized study
Aim	To estimate the effects of an intervention to reduce radon exposure on lung cancer mortality among underground uranium miners	Aim	
Significance	Uranium is mined in open pits or underground as fuel for nuclear reactors. Radon is a radioactive gas can build up mine shafts and present health hazards to workers. We need to know how best to protect workers in these mines.	Significance	
Target population	Underground uranium miners in the Colorado Plateau	Eligibility	
		Sampling	
		Recruitment	

Origin	First day of work in the mine.	When does follow-up begin?	
Timescale	Years since hire, up to 50 years.	What is the maximum length of follow-up of interest?	
Actions (if any)	Limit occupational radon exposure to a) 4 WLM/month b) 1 WLM/month c) 0.25 WLM/month	What actions are being considered? How will actions be implemented?	
		How are actions assigned?	
		When are actions assigned?	

Groups (if any)	None	When and how are groups measured?	
Outcomes	Lung cancer mortality All cause mortality	Which outcomes ?	
		When are outcomes measured?	
Parameters of interest	Risk functions for lung cancer death and all-cause death Value of risk functions at 50 years	Parameters of interest	
Impact	Results will be used to guide regulatory limits on occupational radon exposure	How will results be presented?	See next page
		How will results be used?	

Sample Figures

Sample Table

Arm	n	Risk of lung cancer death	RD	95% CI	Risk of death	RD	95% CI
<4 WLM/month			0			0	
<1 WLM/month							
<0.25 WLM/month							

Appendix

See Cole SR, Edwards JK, Naimi AI, Muñoz A. Hidden imputations and the Kaplan-Meier estimator. American journal of epidemiology. 2020 Nov;189(11):1408-11.

The “extended” Kaplan-Meier hides ghosts

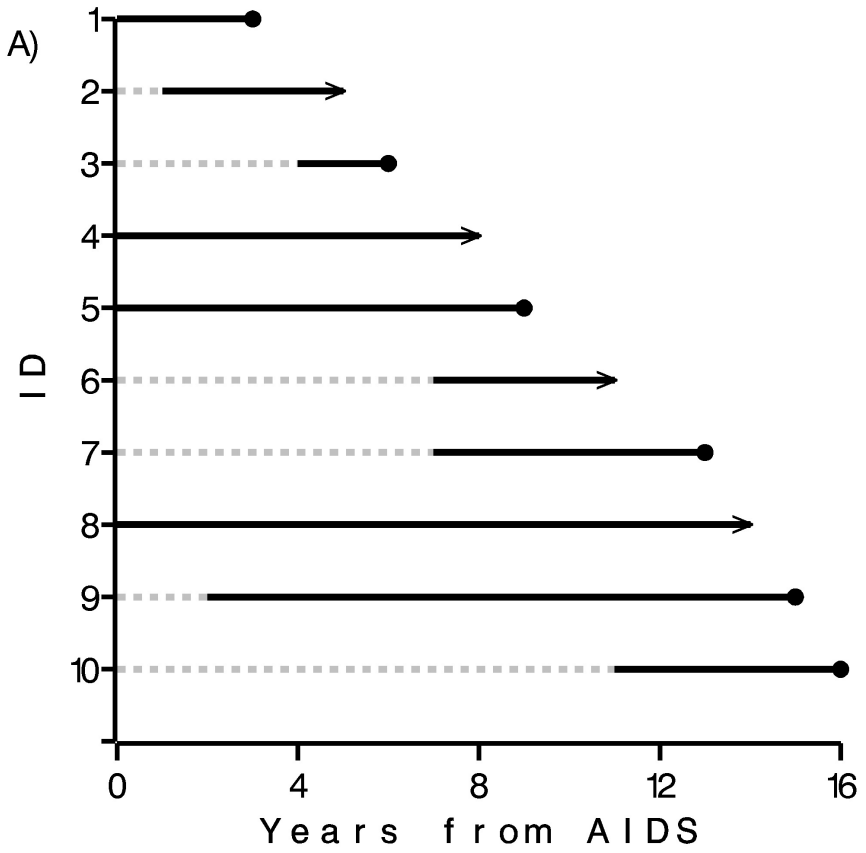
Here, we will illustrate how this works using a simple redistribution algorithm (using a toy example written up by S. Cole and A. Muñoz).

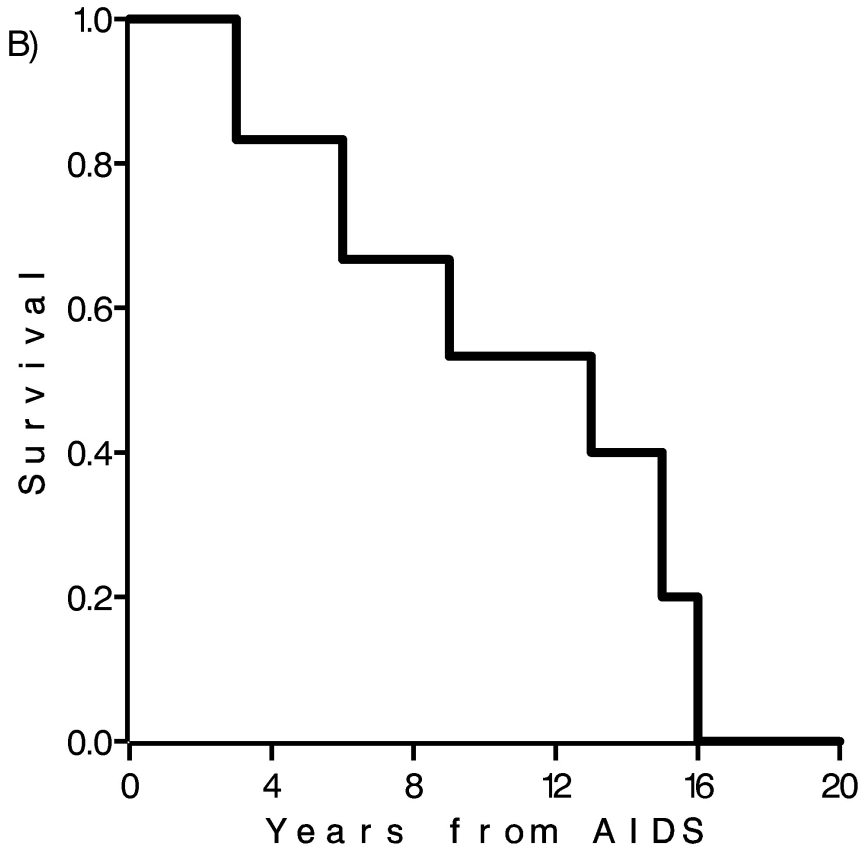
Each person who enters late brings with them $[1 - S(W_i)]/S(W_i)$ truncated events.

These events are distributed to all of the event times that occurred *prior to* W_i.

And thus, we need to be incredibly cautious about the assumptions made when using this foreknowledge.

Toy example


A)
1
2
3
4
5
6
7
8
9
10
I D
0
4
8
12
16
Years from AIDS


B)
1.0
0.8
0.6
0.4
0.2
0.0
Survival
0
4
8
12
16
20
Years from AIDS

Toy example: A redistribution algorithm

			Event times, r_k:	3	6	9	13	15	16	
			No. Events, d_k:	1	1	1	1	1	1	
			Risk set, n_k:	6	5	5	4	2	1	
			$S(t)$:	0.833	0.667	0.533	0.400	0.200	0.000	
			Jumps, $p(r_k)$:	0.167	0.167	0.133	0.133	0.200	0.200	
ID	$\{w, t; \delta\}$	$S(w)$	No. Truncated, U							Total
1	0, 3; 1									
2	1, 5; 0									
3	4, 6; 1									
4	0, 8; 0									
5	0, 9; 1									
6	7, 11; 0									
7	7, 13; 1									
8	0, 14; 0									
9	2, 15; 1									
10	11, 16; 1									
	Total:									
	Jumps:									

Step 1: Assign observed events to times

			Event times, r_k:	3	6	9	13	15	16	
			No. Events, d_k:	1	1	1	1	1	1	
			Risk set, n_k:	6	5	5	4	2	1	
			$S(t)$:	0.833	0.667	0.533	0.400	0.200	0.000	
			Jumps, $p(r_k)$:	0.167	0.167	0.133	0.133	0.200	0.200	
ID	$\{w, t; \delta\}$	$S(w)$	No. Truncated, U							Total
1	0, 3; 1			1	0	0	0	0	0	1
2	1, 5; 0									
3	4, 6; 1				1					
4	0, 8; 0									
5	0, 9; 1			0	0	1	0	0	0	1
6	7, 11; 0									
7	7, 13; 1						1			
8	0, 14; 0									
9	2, 15; 1							1		
10	11, 16; 1								1	
	Total:									
	Jumps:									

Step 2: Handle censoring

Recall, the event for censored participants gets redistributed to events happening after the censoring time.

Event redistribution is proportional to the step size in the survival curve.

E.g., $V_c(k) = 1 \times \frac{[S(k-1) - S(k)]}{[S(C_i) - S(last)]}$

Step 2: Handle censoring

			Event times, r_k:	3	6	9	13	15	16	
			No. Events, d_k:	1	1	1	1	1	1	
			Risk set, n_k:	6	5	5	4	2	1	
			$S(t)$:	0.833	0.667	0.533	0.400	0.200	0.000	
			Jumps, $p(r_k)$:	0.167	0.167	0.133	0.133	0.200	0.200	
ID	$\{w, t; \delta\}$	$S(w)$	No. Truncated, U							Total
1	0, 3; 1			1	0	0	0	0	0	1
2	1, 5; 0			0	0.200	0.160	0.160	0.240	0.240	1
3	4, 6; 1				1					
4	0, 8; 0			**0**	**0**	**0.200**	**0.200**	**0.300**	**0.300**	1
5	0, 9; 1			0	0	1	0	0	0	1
6	7, 11; 0									
7	7, 13; 1						1			
8	0, 14; 0			0	0	0	0	0.500	0.500	1
9	2, 15; 1							1		
10	11, 16; 1								1	
	Total:									
	Jumps:									

Step 3: Handle late entry

Recall, people who enter the study late bring the ghosts of people who would have entered the study at or before W_i, but didn't because they had the event.

The late enter-er carries ghost events in the time before study entry of the form

E.g., $V_t(k) = \frac{\{1-S(W_i)\}}{S(W_i)} \times \frac{\{S(k-1)-S(k)\}}{\{1-S(W_i)\}}$

Total number of ghost events needed

Proportion of ghost events allocated to each event time prior to study entry

Step 3: Handle late entry

Why is the number of truncated events $[1 - S(W_i)]/S(W_i)$?

Consider: we observe 1 person enter at time W_i, indicating that 1 person survived until time W_i.

Ask: What study size would we have needed to start out with to see 1 person survive to time W_i?

Answer: $1/S(W_i)$

But we do get to see the 1 person enter at W_i, so number of truncated events is only $1/S(W_i) - 1 = [1 - S(W_i)]/S(W_i)$

Step 3: Handle late entry

			Event times, r_k:	3	6	9	13	15	16	
			No. Events, d_k:	1	1	1	1	1	1	
			Risk set, n_k:	6	5	5	4	2	1	
			$S(t)$:	0.833	0.667	0.533	0.400	0.200	0.000	
			Jumps, $p(r_k)$:	0.167	0.167	0.133	0.133	0.200	0.200	
ID	$\{w, t; \delta\}$	$S(w)$	No. Truncated, U							Total
1	0, 3; 1	1		1	0	0	0	0	0	1
2	1, 5; 0	1		0	0.200	0.160	0.160	0.240	0.240	
3	4, 6; 1	**0.833**	0.2	***0.200***	**1**	**0**	**0**	**0**	**0**	**1.2**
4	0, 8; 0	1		0	0	0.200	0.200	0.300	0.300	1
5	0, 9; 1	1		0	0	1	0	0	0	1
6	7, 11; 0	0.667	0.5	0.250	0.250	0	0.250	0.375	0.375	1.5
7	7, 13; 1	0.667	0.5	0.250	0.250	0	1	0	0	1.5
8	0, 14; 0	1		0	0	0	0	0.500	0.500	1
9	2, 15; 1	1		0	0	0	0	1	0	1
10	11, 16; 1	0.533	0.875	0.3125	0.3125	0.250	0	0	1	1.875
	Total:									
	Jumps:									

Step 3: Handle late entry

			Event times, r_k:	3	6	9	13	15	16	
			No. Events, d_k:	1	1	1	1	1	1	
			Risk set, n_k:	6	5	5	4	2	1	
			$S(t)$:	0.833	0.667	0.533	0.400	0.200	0.000	
			Jumps, $p(r_k)$:	0.167	0.167	0.133	0.133	0.200	0.200	
ID	$\{w, t; \delta\}$	$S(w)$	No. Truncated, U							Total
1	0, 3; 1	1		1	0	0	0	0	0	1
2	1, 5; 0	1		0	0.200	0.160	0.160	0.240	0.240	
3	4, 6; 1	0.833	0.2	*0.200*	1	0	0	0	0	1.2
4	0, 8; 0	1		0	0	0.200	0.200	0.300	0.300	1
5	0, 9; 1	1		0	0	1	0	0	0	1
6	7, 11; 0	**0.667**	**0.5**	**0.250**	**0.250**	**0**	**0.250**	**0.375**	**0.375**	**1.5**
7	7, 13; 1	0.667	0.5	0.250	0.250	0	1	0	0	1.5
8	0, 14; 0	1		0	0	0	0	0.500	0.500	1
9	2, 15; 1	1		0	0	0	0	1	0	1
10	11, 16; 1	0.533	0.875	0.3125	0.3125	0.250	0	0	1	1.875
	Total:			2.0125	2.0125	1.610	1.610	2.415	2.415	12.075
	Jumps:			0.167	0.167	0.133	0.133	0.200	0.200	1.0

Assumptions hidden in Step 3?

We use $S(W_i)$ *estimated in the study sample* to determine the number of truncated events and a function of $S(W_i)$ and $S(k)$ to distribute the truncated events in time prior to a late entry.

Implies an assumption that the hazard among those in the study is the same as the hazard of those in the target population but not yet in the study.

If $S(W_i) = 1$ then allowing person i to enter at time W_i does not require these assumptions.

Side note

In the algorithm described, we used Kaplan-Meier estimates of $S(t)$ to derive the number of truncated events and censored events.

If we did not know the Kaplan-Meier, we could use an iterative EM algorithm that would converge to the same values.