Study design and analysis for time-dependent exposures during pregnancy

SPER Advanced Methods Workshop
June 13, 2022

Louisa H. Smith

Today's plan

Louisa:

- Introduction to the problem and data
  - Lab 1
- Basic analysis of time-varying exposures
  - Lab 2
- Dealing with confounding
  - Lab 3
- Additional topics

Chelsea: Time-Dependent Exposures and Selective Testing in Pregnancy

Dr Ellie Murray, ScD
@EpiEllie

Replying to @JulieMOPetersen and @AmJEpi

So many potential sources of bias!!

5:56 PM · Jul 2, 2019 · Twitter for iPhone

4 Retweets **10** Likes

Elías Eyþórsson @eliaseythorsson · Jul 3, 2019

Replying to @EpiEllie @JulieMOPetersen and @AmJEpi

Vaccine epi is pretty hard. Variable uptake and coverage. Exposure causes direct effects among vaccinated and indirect among unvaccinated. The onset of effect on the individual and population level unknown.

1 1

Dr Ellie Murray, ScD @EpiEllie · Jul 3, 2019

True true, vaccine epi is very hard too, but what about vaccines-in-pregnancy epi! 😄

2

An exposure at some point during pregnancy, and an outcome that depends on time…

An exposure at some point during pregnancy, and an outcome that depends on time…

COVID-19 (vaccination) and spontaneous abortion

An exposure at some point during pregnancy, and an outcome that depends on time…

COVID-19 (vaccination) and spontaneous abortion

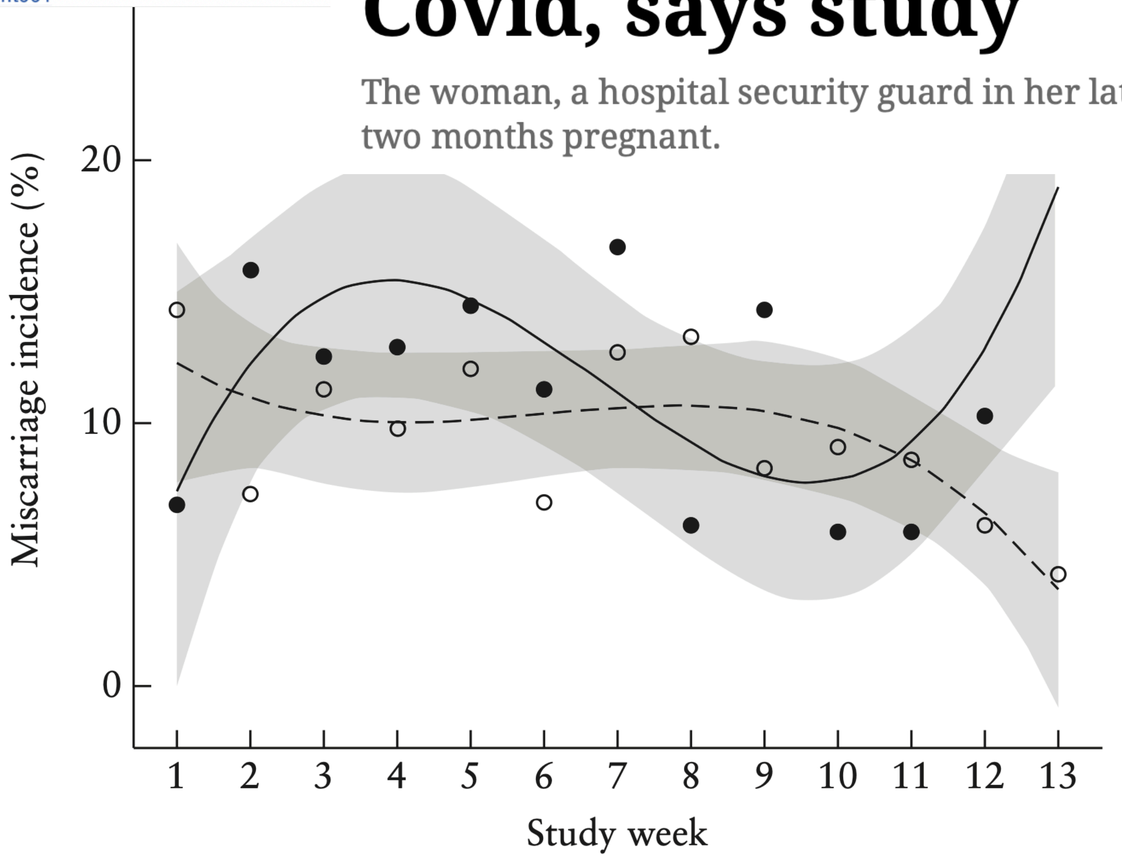
- COVID (vaccination) & preterm birth
- COVID (vaccination) & preeclampsia
- Non-COVID-related events (!) that can occur at varying times during pregnancy
- Pregnancy/birth outcomes that depend on/are affected by pregnancy length

References	Country	Type of study	No. of mothers with COVID-19	No. of abortion	Age	COVID-19 detection method
Fang, Nz. [41]	USA	Case report	1	1	33	PCR
Rana, M. S. [27]	Pakistan	Case report	1	1	30	PCR
Hachem, R. [42]	France	Case report	1	1	21	PCR
Baud, D. [29]	Switzerland	Case report	1	1	28	PCR
Shojaei, S. [43]	Iran	Case report	1	1	-	PCR
Wong, T. C. [39]	Malaysia	Case series	2	2	34	PCR
Yan, J. [44]	China	Case series	116	1	Mean age: 30	PCR
Buonsenso, D. [45]	Italy	Case series	7	1	-	PCR
Richtmann, R. [25]	Brazil	Case series				
Mayeur, A. [28]	France	Case series	1			
Sentilhes, L. [46]	France	Case series	5			

https://doi.org/10.1371/journal.pone.0255994.t001

Mumbai woman suffers miscarriage due to Covid, says study

The woman, a hospital security guard in her late twenties, tested positive for Covid-19 when she was two months pregnant.



Sacinti KG, Kalafat E, Sukur YE, et al. Increased incidence of first-trimester miscarriage during the COVID-19 pandemic. Ultrasound Obstet Gynecol. 2021;57:1013–1014.
Magnus MC, Gjessing HK, Eide HN, et al. Covid-19 Vaccination during Pregnancy and First-Trimester Miscarriage. N Engl J Med. 2021;385:2008–2010.

l.smith@northeastern.ed

In a case series study of 116 patients, Yan et al. reported the miscarriage rate in pregnant women with COVID-19 for the first time. Before the 20th week of gestation, 8 out of 116 pregnant women were tested positive for COVID-19. The miscarriage rate among these eight pregnant women was 12.5 % (n = 1/8, 95 % CI 0.32–52.65) (Yan et al., 2020). Other authors have reported the following miscarriage rates (<22 weeks of gestation): 18.2 % (n = 4/22, 95 % CI 5.19–40.28) (Knight et al., 2020), and 14.3 % (n = 1/7, 95 % CI 0.36–57.87) (Mattar et al., 2020).

Early miscarriage rates (<12 weeks) in pregnant women with COVID-19 diagnosed in the first trimester were 100 % (n = 2/2, 95 %CI 15.81–100) (Wong et al., 2020), 0% (n = 0/2, 95 %CI 0–84.19) (Curi et al., 2020), 19.4 % (n = 6/31, 95 %CI 7.45–37.47) (WAPM (World Association of Perinatal Medicine) Working Group on COVID-19, 2021), 18.2 % (n = 2/11, 95 %CI 2.28–51.78) (Grechukhina et al., 2020), 16.7 % (n = 1/6, 95 %CI 0.42–64.12) (Mattar et al., 2020), 9.2 % (n = 12/130, 95 %CI 4.86–15.57) (Sahin et al., 2021), 40 % (n = 2/5, 95 %CI 5.27–85.34) (Shmakov et al., 2020) and 60 % (n = 3/5, 95 %CI 14.66–94.73) (Singh et al., 2021).

Questions we want to answer

- What is the risk of spontaneous abortion after COVID-19 in pregnancy?
  - Descriptive question
- Does COVID-19 in pregnancy increase the risk of spontaneous abortion?
  - Causal question
  - Compared to…? Never getting COVID-19 in pregnancy? Some timing of exposure more harmful than another?

A little bit of notation

To help clarify the question…

X: gestational age at COVID-19 exposure (weeks + days)

A: indicator of COVID-19 exposure at some point in pregnancy (0/1)

T: gestational age at end of pregnancy (weeks + days)

Y: indicator of spontaneous abortion (0/1)

$$X < T \implies A = 1$$

$$T < 20 \implies Y = 1$$

If no COVID exposure during pregnancy, we can say $X = \infty$ or NA or some large number…

Example of data

id	X	A	T	Y
712	–	0	12 + 4	1
4603	12 + 5	1	38 + 6	0
8527	12 + 0	1	39 + 6	0
9493	–	0	15 + 4	1

Potential (counterfactual) outcomes

$$T^a \text{ and } Y^a$$

pregnancy outcomes for a participant if, possibly counter to fact, they had been exposed ($a = 1$) or unexposed ($a = 0$) to COVID-19 during pregnancy

$$T^x \text{ and } Y^x$$

the outcomes if the COVID-19 exposure had occurred during week x

Consistency

A person's observed outcome under a certain COVID-19 exposure is assumed to be the same as if, under a hypothetical intervention, it had been assigned to be so.

id	X	A	T	Y	$T^{a=1}$	$Y^{a=1}$	$T^{x=12}$	$Y^{x=12}$
712	–	0	12 + 4	1				
4603	12 + 5	1	38 + 6	0	38 + 6	0	38 + 6	0
8527	12 + 0	1	39 + 6	0	39 + 6	0	39 + 6	0
9493	–	0	15 + 4	1				

Missing data

We only see one potential outcome for each observation, leaving us nothing to compare to

id	X	A	T	Y	$T^{a=1}$	$Y^{a=1}$	$T^{x=12}$	$Y^{x=12}$	$T^{a=0}$	$Y^{a=0}$	$T^{x=6}$	$Y^{x=6}$
712	–	0	12 + 4	1					12 + 4	1		
4603	12 + 5	1	38 + 6	0	38 + 6	0	38 + 6	0				
8527	12 + 0	1	39 + 6	0	39 + 6	0	39 + 6	0				
9493	–	0	15 + 4	1					15 + 4	1		

How can we express our questions more precisely and figure out exactly what we want to answer?

How can we express our questions more precisely and figure out exactly what we want to answer?

What does $\Pr\left(Y = 1 \mid A = 1\right)$ mean?

How can we express our questions more precisely and figure out exactly what we want to answer?

What does $\Pr\left(Y = 1 \mid A = 1\right)$ mean?

- The probability of spontaneous abortion among people with COVID-19 in pregnancy

How can we express our questions more precisely and figure out exactly what we want to answer?

What does $\Pr\left(Y = 1 \mid A = 1\right)$ mean?

- The probability of spontaneous abortion among people with COVID-19 in pregnancy

What about $\Pr\left(Y^{a=1} = 1\right)$ vs. $\Pr\left(Y^{a=0} = 1\right)$?

How can we express our questions more precisely and figure out exactly what we want to answer?

What does $\Pr\left(Y = 1 \mid A = 1\right)$ mean?

- The probability of spontaneous abortion among people with COVID-19 in pregnancy

What about $\Pr\left(Y^{a=1} = 1\right)$ vs. $\Pr\left(Y^{a=0} = 1\right)$?

- The probability of spontaneous abortion had everyone been exposed vs. unexposed to COVID-19 in pregnancy

How can we express our questions more precisely and figure out exactly what we want to answer?

What does $\Pr\left(Y = 1 \mid A = 1\right)$ mean?

- The probability of spontaneous abortion among people with COVID-19 in pregnancy

What about $\Pr\left(Y^{a=1} = 1\right)$ vs. $\Pr\left(Y^{a=0} = 1\right)$?

- The probability of spontaneous abortion had everyone been exposed vs. unexposed to COVID-19 in pregnancy

What about $\Pr\left(Y^{x=5} = 1\right)$ vs. $\Pr\left(Y^{x=19} = 1\right)$?

How can we express our questions more precisely and figure out exactly what we want to answer?

What does $\Pr\left(Y = 1 \mid A = 1\right)$ mean?

- The probability of spontaneous abortion among people with COVID-19 in pregnancy

What about $\Pr\left(Y^{a=1} = 1\right)$ vs. $\Pr\left(Y^{a=0} = 1\right)$?

- The probability of spontaneous abortion had everyone been exposed vs. unexposed to COVID-19 in pregnancy

What about $\Pr\left(Y^{x=5} = 1\right)$ vs. $\Pr\left(Y^{x=19} = 1\right)$?

- The probability of spontaneous abortion after getting COVID-19 at 5 weeks' gestation vs. at 19 weeks' gestation

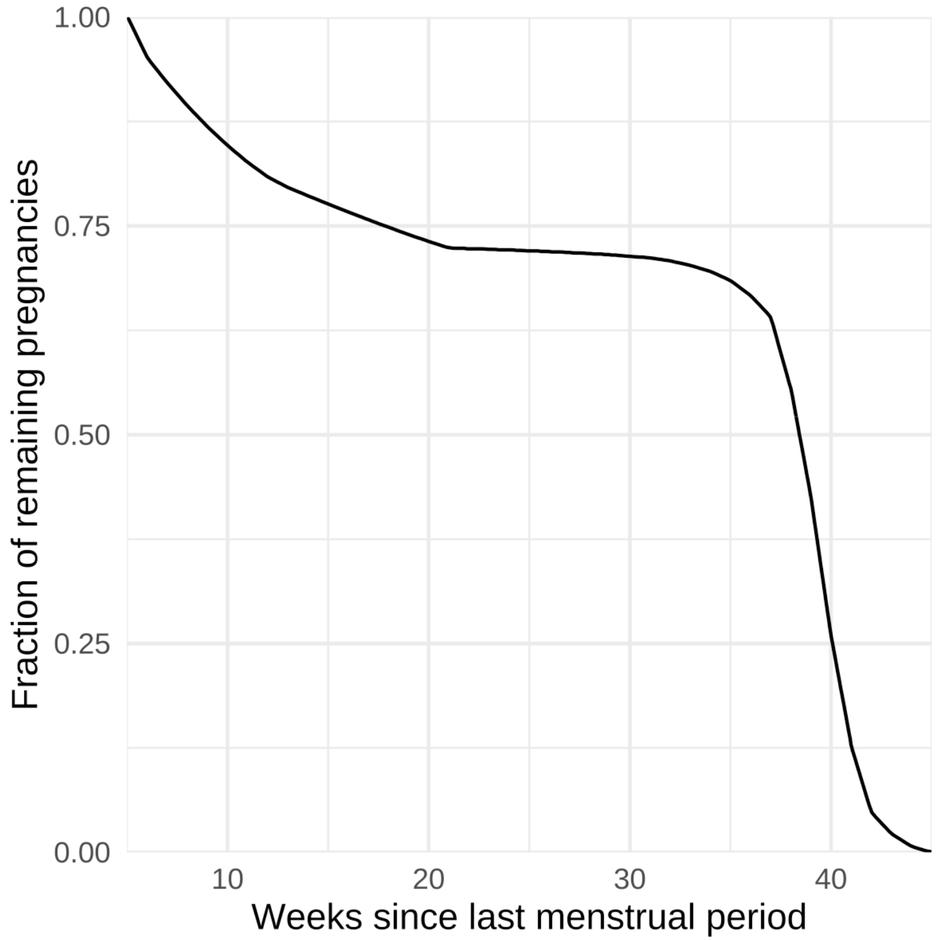
Simulated data

- I drew T (`time_ended`) from a cumulative distribution function of pregnancy lengths that I drew based on data from several papers
- I drew X (`time_exposed`) randomly and uniformly from 5 weeks of gestation through 45 (even if the person was no longer pregnant)
- `id` ranges from 1 to 10000.

There is *no association* between T and X.

Goldhaber MK, Fireman BH. The fetal life table revisited: Spontaneous abortion rates in three kaiser permanente cohorts. Epidemiology. 1991;2:33–39.
Avalos LA, Galindo C, Li D-K. A systematic review to calculate background miscarriage rates using life table analysis. Birth Defects Research Part A: Clinical and Molecular Teratology. 2012;94:417–423.
Mukherjee S, Velez Edwards DR, Baird DD, et al. Risk of miscarriage among black women and white women in a US prospective cohort study. American Journal of Epidemiology. 2013;177:1271–1278.

Distribution of pregnancy lengths

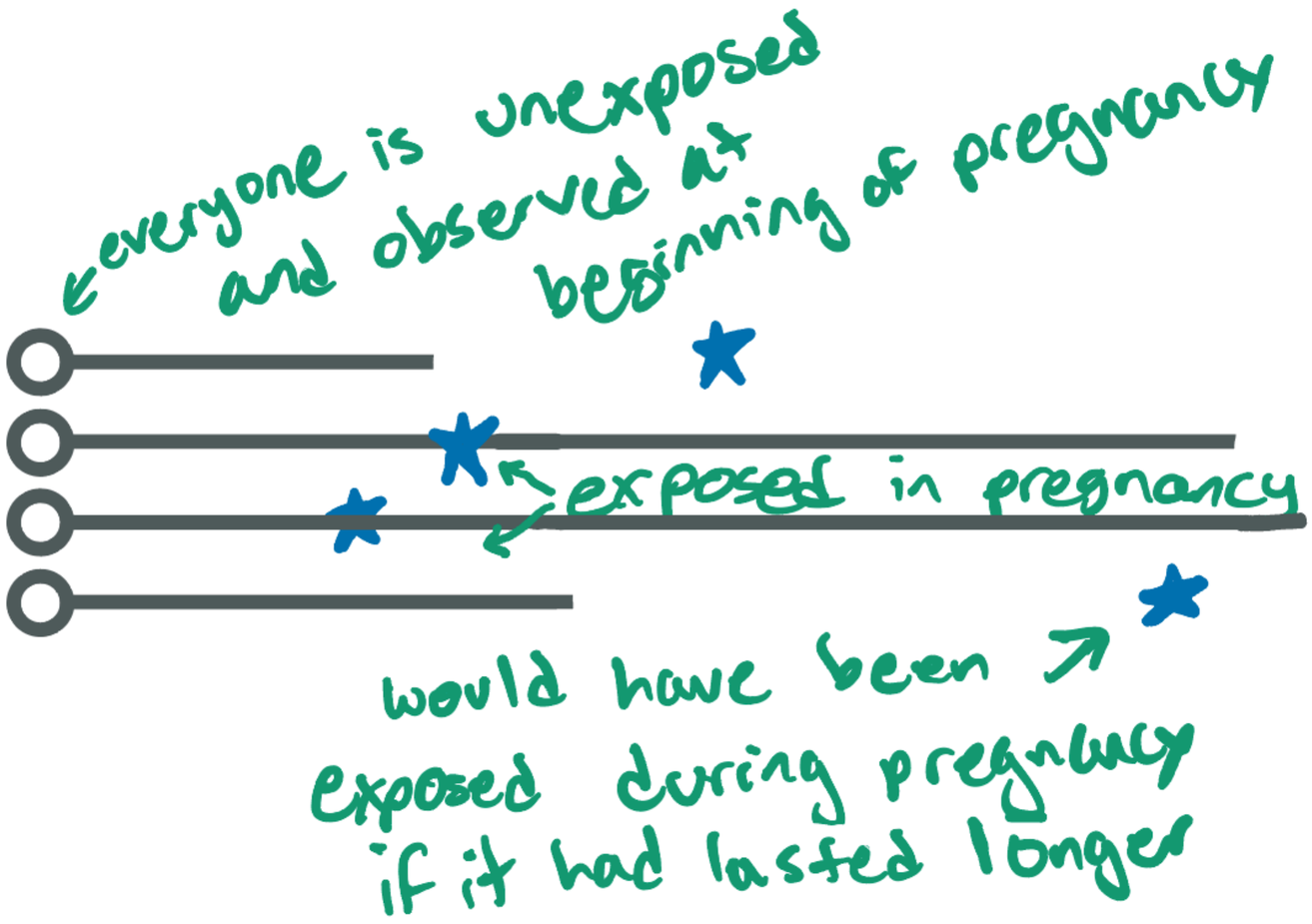



```
n <- 10000

dat <- tibble(
  id = 1:n,
  time_ended = cdf_inverse(runif(n, 0, 1)),
  time_potentially_exposed = runif(n, 5, 45),
  time_exposed = case_when(
      time_potentially_exposed <
          time_ended ~ time_potentially_exposed,
      TRUE ~ NA_real_))
```

Simulated data

id	time_ ended	time_ potentially _exposed	time_ exposed
712	12.57	22.00	–
4603	38.86	12.71	12.71
8527	39.86	12.00	12.00
9493	15.57	36.71	–



Simulated data

A: exposed_while_pregnant

X: time_exposed

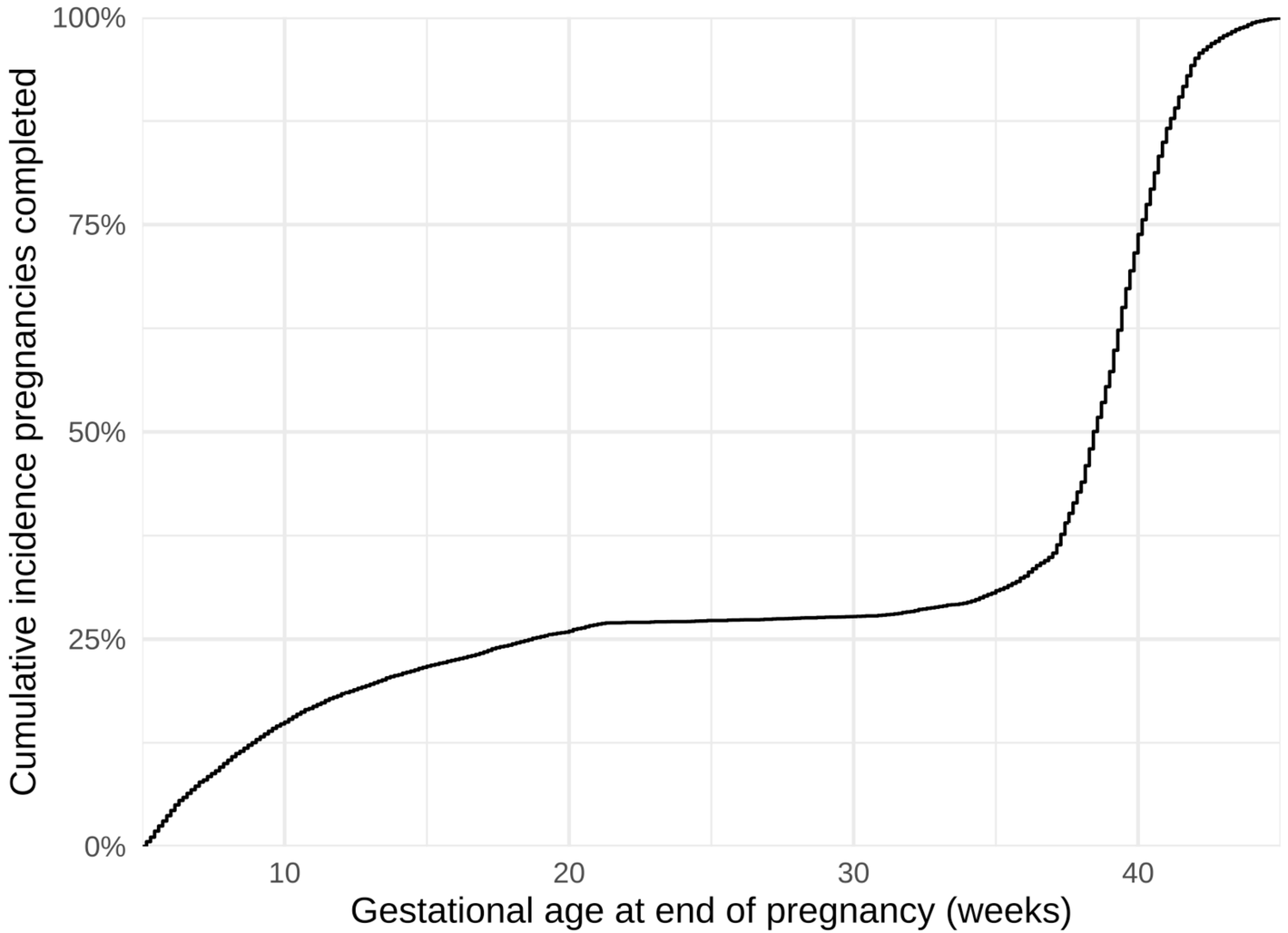
Y: sab

T: time_ended

id	time_ended	time_potentially_exposed	time_exposed	exposed_while_pregnant	sab
712	12.57	22.00	–	0	0
4603	38.86	12.71	12.71	1	1
8527	39.86	12.00	12.00	1	1
9493	15.57	36.71	–	0	0

The pregnancy cumulative incidence curve is as expected

This is just 1 - survival: what fraction of pregnancies have ended by a certain gestational age?



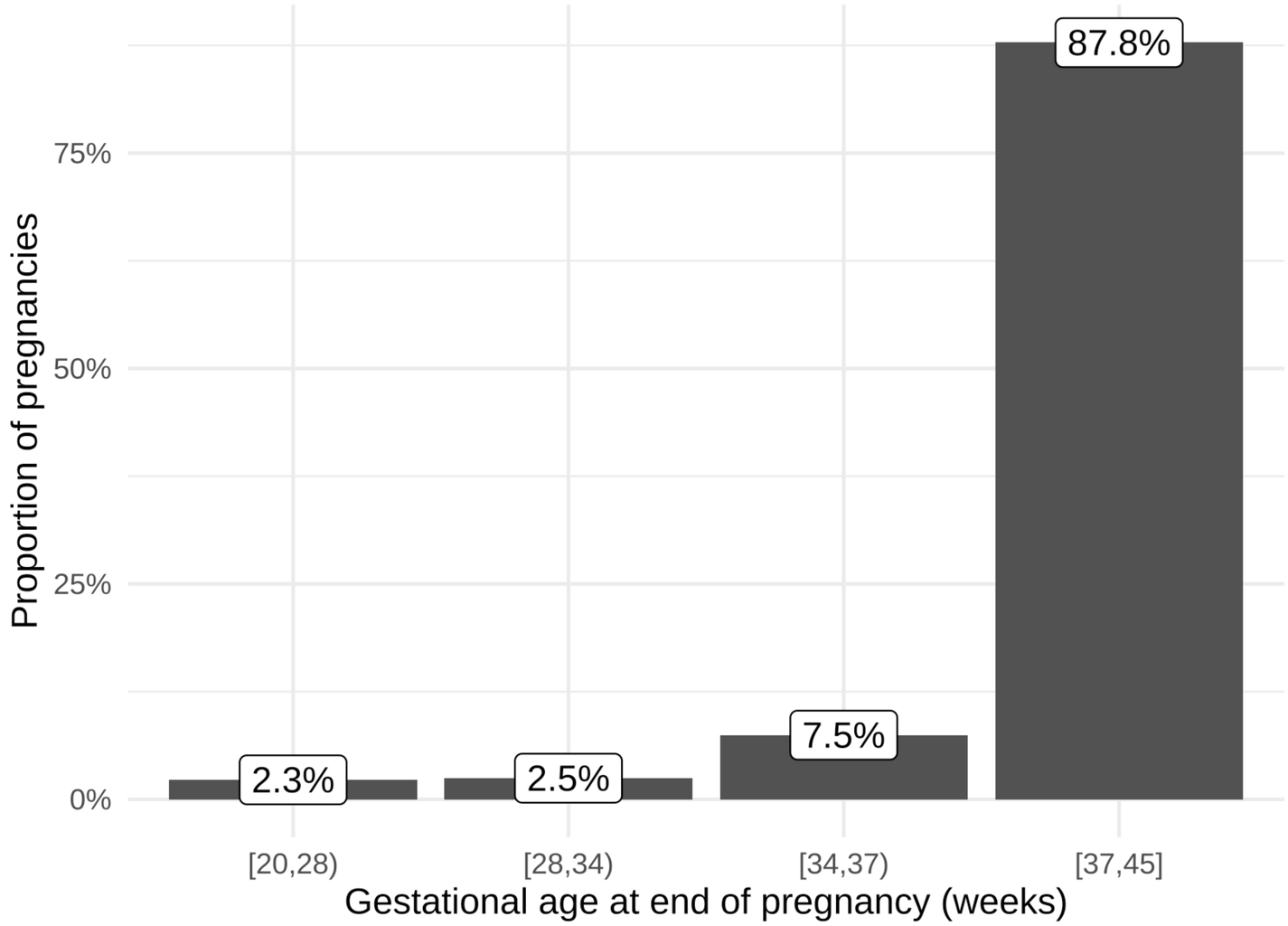
Risk of spontaneous abortion

What is 1 - survival at 20 weeks?

Deliveries over time

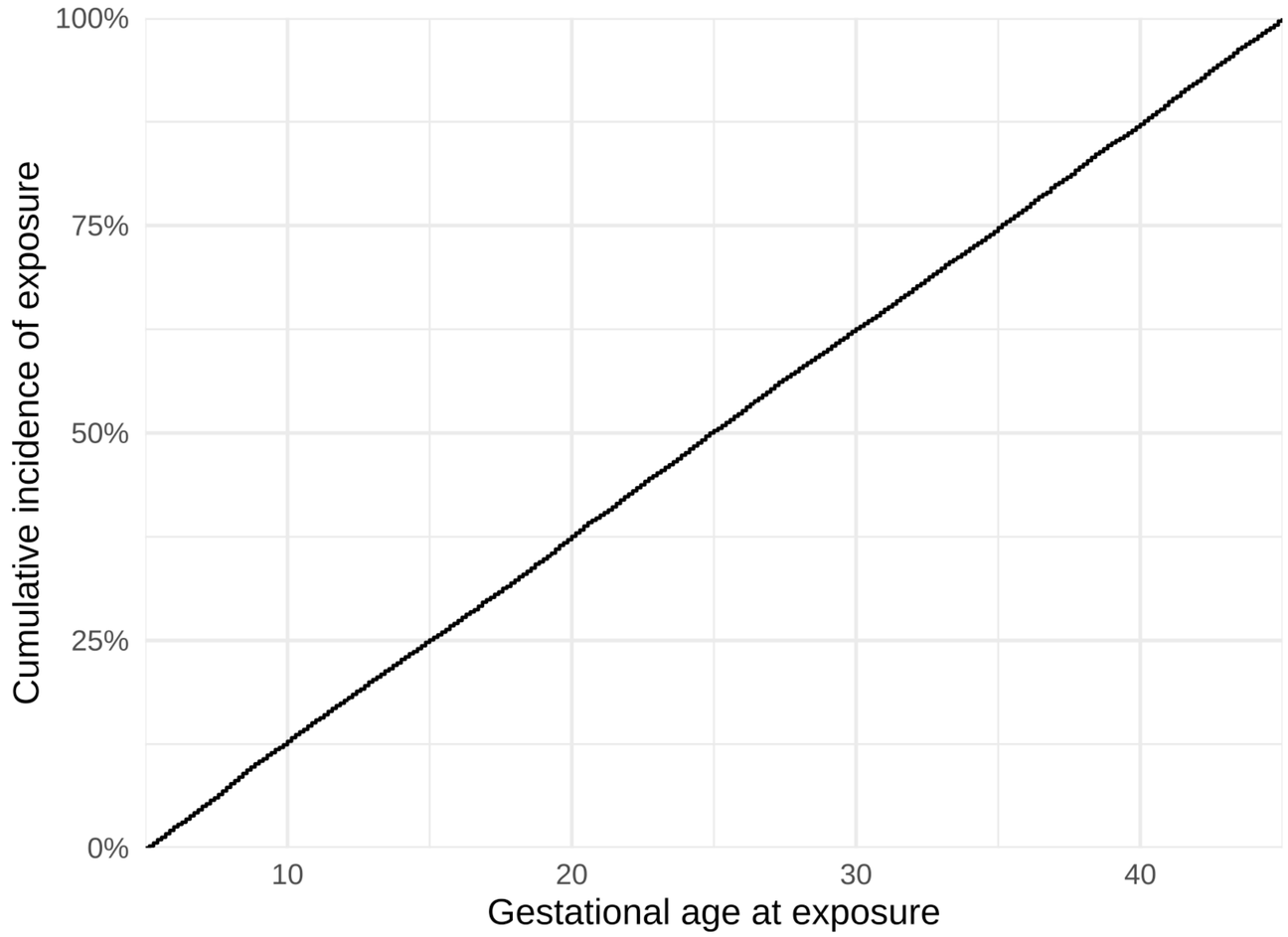
Conditional on surviving 20 weeks

Preterm deliveries



Exposure risk is constant across the population throughout pregnancy

Everyone is exposed at some point (possibly after pregnancy)



What's the risk of SAB among the exposed?

$$\Pr(Y = 1 \mid A = 1)$$

```
dat %>%
  filter(exposed_while_pregnant == 1) %>%
  summarise(risk_in_exposed = mean(sab))
```

```
# A tibble: 1 × 1
  risk_in_exposed
            <dbl>
1          0.0503
```

How do we interpret this?

$$\Pr\left(Y = 1 \mid A = 1\right) = 0.05$$

Hmmm, this sounds really low given what we know about spontaneous abortion (and what we've seen in the data overall).

Why? We are including people who were exposed long after they were at risk for spontaneous abortion.

This seems obvious but I have seen this mistake!

Redefine exposure

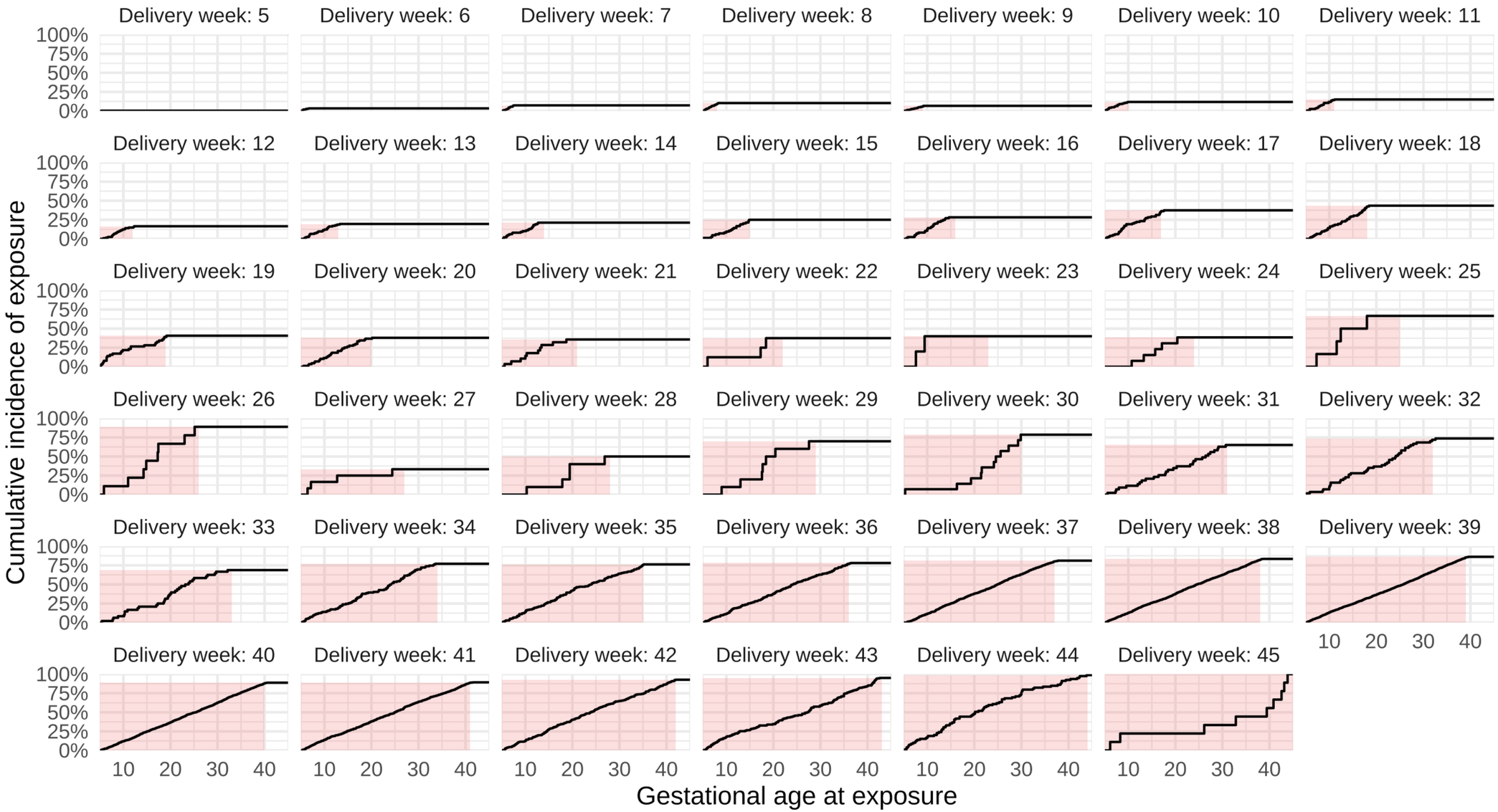
A: COVID before 20 weeks of pregnancy ($X < 20$)

```
dat <- dat %>%
  mutate(exposed_while_pregnant =
     as.numeric(!is.na(time_exposed) & time_exposed < 20))
```

$\Pr\left(Y = 1 \mid A = 1\right) = 0.107$

This also seems too low. What's going on?

The earlier a pregnancy ends, the lower the chance of being exposed



Immortal person-time

Only the longer pregnancies lasted long enough to be exposed…

- Many short pregnancies – those with spontaneous abortions – ended before exposure could occur, so aren't counted as exposed
- This is a common problem in pharmacoepidemiologic studies, when patients have to survive long enough to start taking a drug of interest
- We usually think of the bias it causes when doing comparative studies – e.g., comparing to people who *didn't* take the drug – but it can result in descriptive statistics that aren't meaningful as well
- $P\left(Y = 1 \mid A = 1\right)$ is a quantity that exists, but the extent to which it's meaningful depends on context (did everyone get COVID at week 1? at week 19?)

Immortal person-time matters because the outcome depends on time

- An outcome that doesn't happen over time and isn't affected by time wouldn't have the same problem
- But it's hard to think of a pregnancy/birth outcome that is not related to/ mediated through pregnancy length!

Similar problem: left truncation

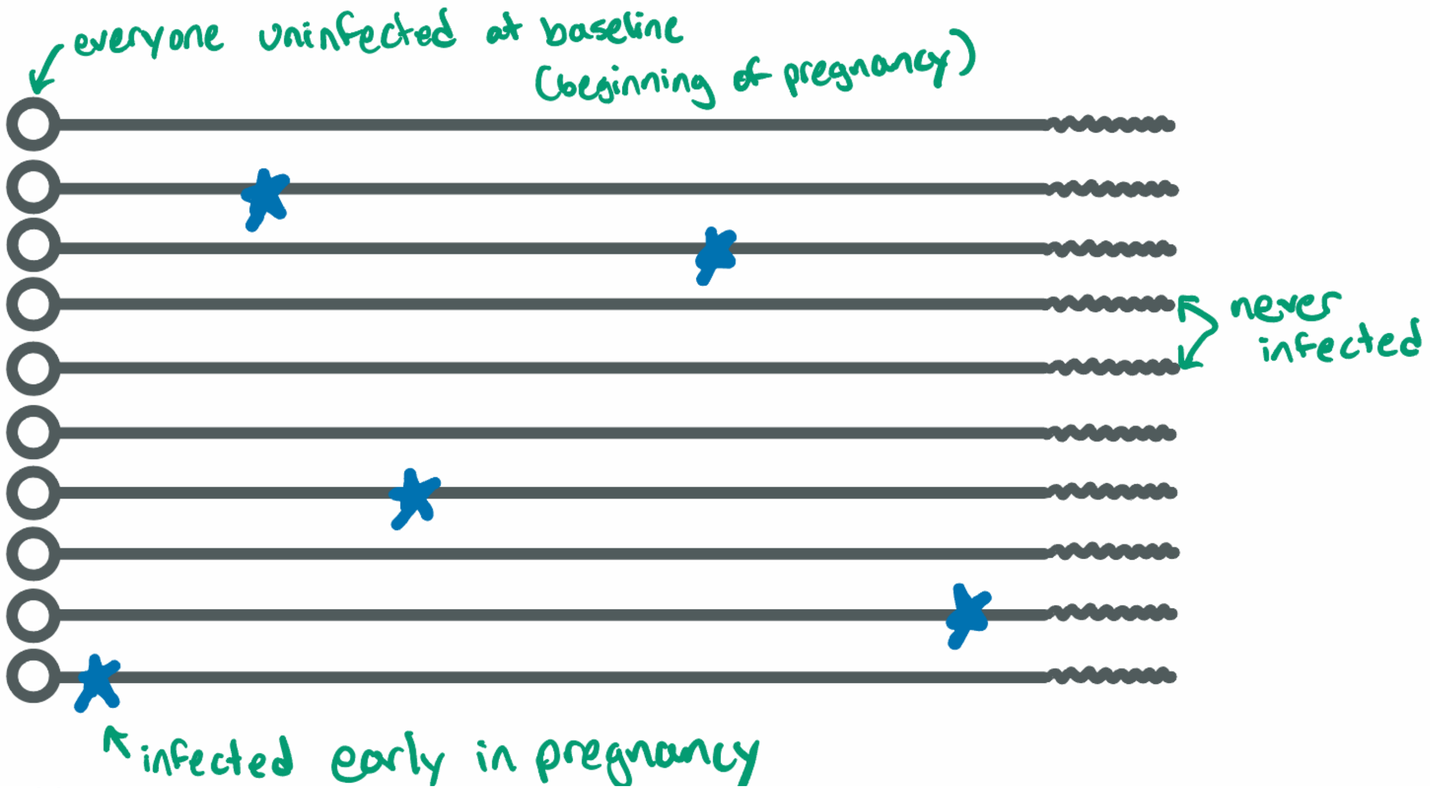
- We don't see all of the early events
- $P\left(Y = 1 \mid A = 1\right)$ is going to depend on when we "start counting"
- We're not necessarily talking about exposure during any pregnancy, but in "recognized pregnancies"
- The time at which they're recognized will of course depend

Conclusion: this quantity is really hard to interpret!

We could estimate rates instead

Rate of SAB per exposed person-month

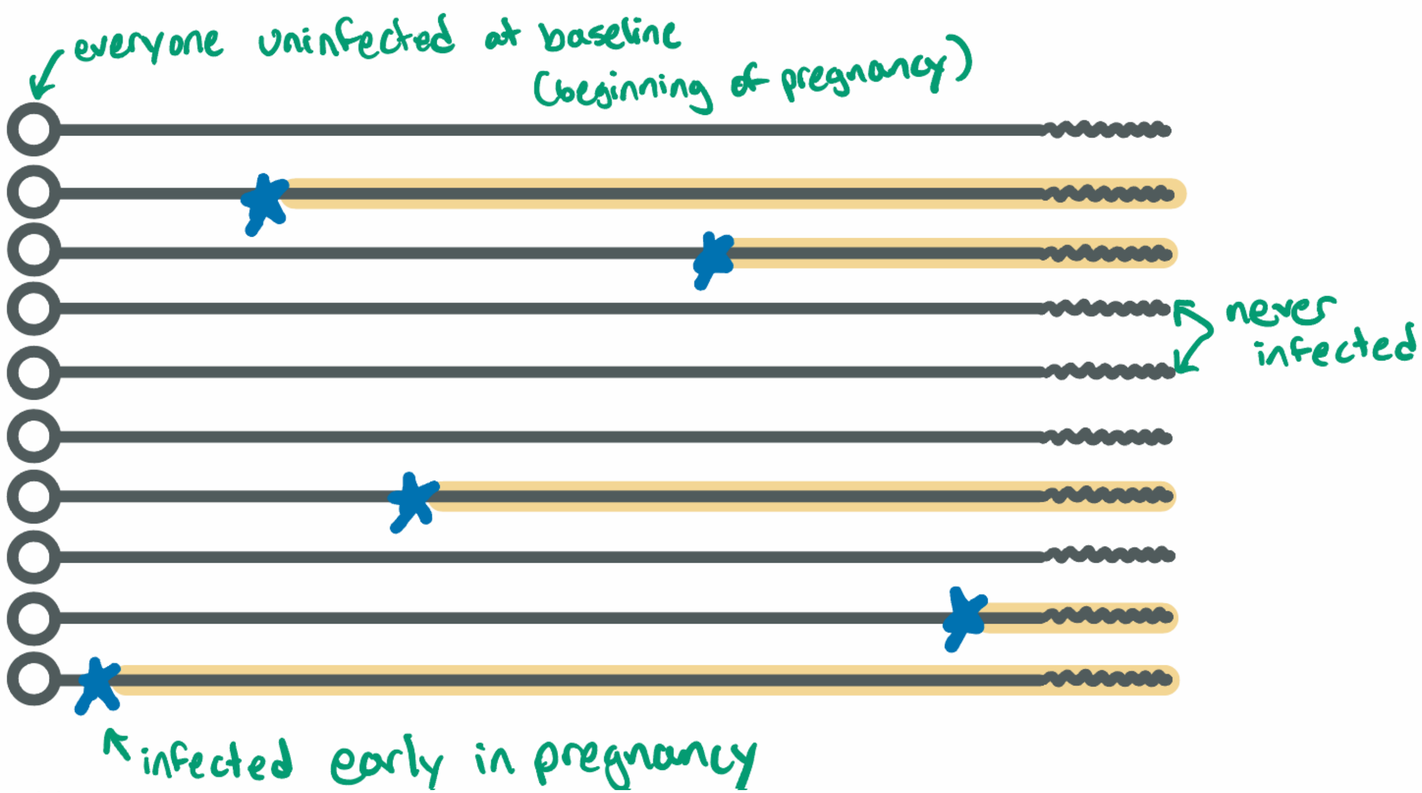
Events	Person-time	Rate
332	22399.29	0.0148219



We could estimate rates instead

Rate of SAB per exposed person-month

Events	Person-time	Rate
332	22399.29	0.0148219



A more interpretable quantity?

$$\Pr(Y = 1 \mid X = \mathrm{x})$$

What's the probability of spontaneous abortion after getting COVID at week *x*?

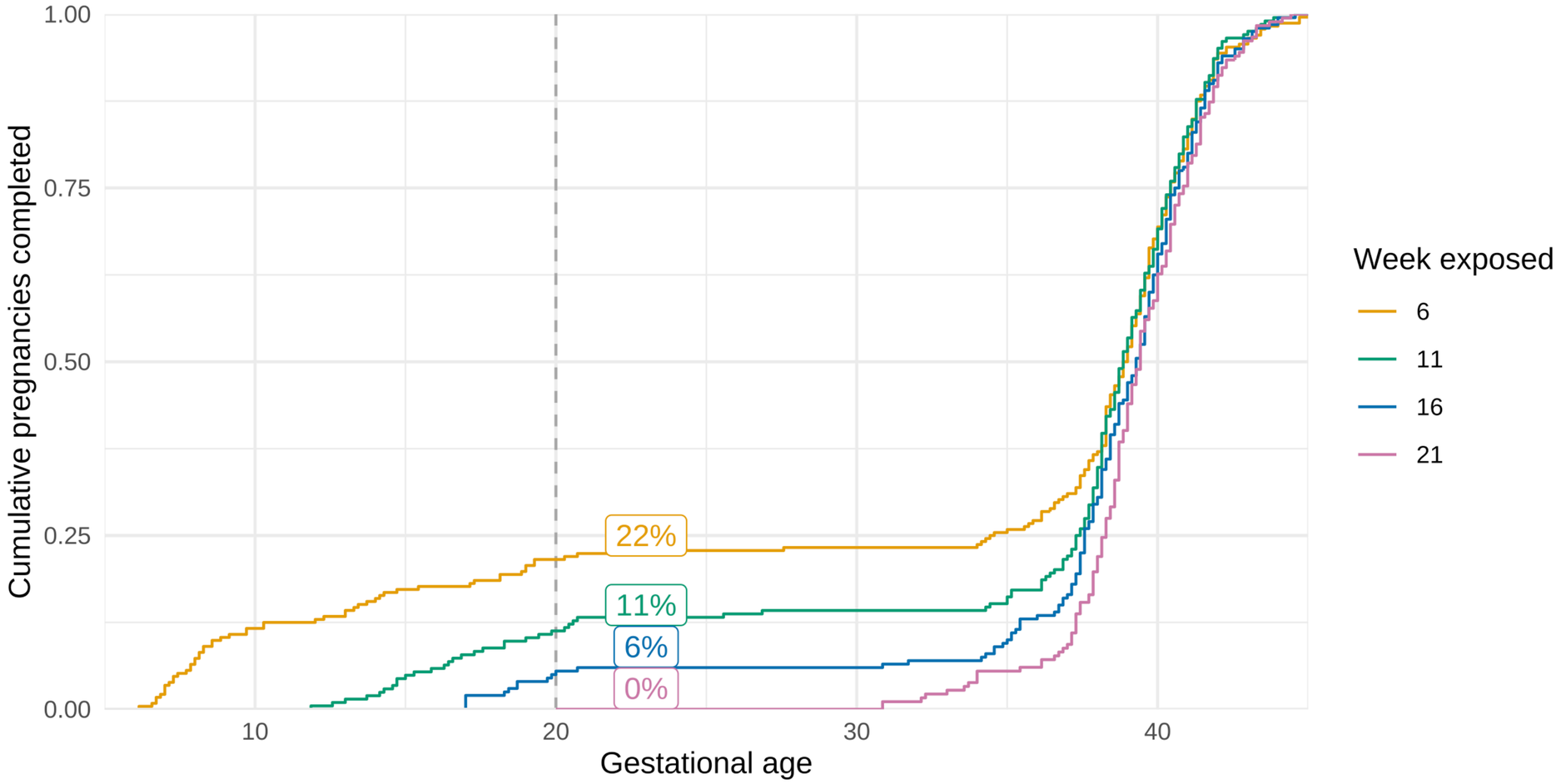
This won't depend on the distribution of exposure over pregnancy.

```
week_6_exposed <- dat %>%
  filter(floor(time_exposed) == 6)
mean(week_6_exposed$sab)
[1] 0.2155172

week_10_exposed <- dat %>%
  filter(floor(time_exposed) == 10)
mean(week_10_exposed$sab)
[1] 0.1090909

week_16_exposed <- dat %>%
  filter(floor(time_exposed) == 16)
mean(week_16_exposed$sab)
[1] 0.05
```

We can’t compare these risks to each other, but they are meaningful



Lab 1

Comparative questions

- We have seen $\Pr\left(Y = 1 \mid A = 1\right)$ is not meaningful without context about the timing of exposure…
- But what about $\Pr\left(Y = 1 \mid A = 0\right)$. When do you not get exposed?
- What we really care about is a causal question: is $\Pr\left(Y^{a=1}\right)$ vs. $\Pr\left(Y^{a=0}\right)$ meaningful?
- Will a ratio or difference measure comparing these be interpretable?

Rate/hazard ratio

Exposure	Events	Person-time	Rate
Yes	332	22399.29	0.0148219
No	2252	152112.86	0.0148048

Rate/hazard ratio

Exposure	Events	Person-time	Rate
Yes	332	22399.29	0.0148219
No	2252	152112.86	0.0148048

We could fit a Cox model using exposure as a time-varying covariate:

	HR	95% CI	p-value
Exposure	1.01	0.90, 1.14	0.8

Interpretation of rate ratio/hazard ratio

Does not map onto a decision-making framework that is of concern in public health

Will be dependent on the pattern of exposure timing and outcomes, but doesn't tell us anything about the time-varying effects of exposure

- If COVID shortened pregnancy length only among those who would have a spontaneous abortion anyway (i.e. led to earlier SAB), the rate would be greater – but it could be argued that that is a "better" outcome
- Late exposures contribute less exposed person-time – smaller denominator

Hernán MA. Counterpoint: Epidemiology to Guide Decision-Making: Moving Away From Practice-Free Research. *Am J Epidemiol*. 2015;182:834–839.
Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010;21:13–15.

Naive risk ratio

	COVID during pregnancy		Total
	Yes	No	
Spontaneous abortion			
Yes	332 (11%)	2,252 (33%)	2,584 (26%)
No	2,760 (89%)	4,656 (67%)	7,416 (74%)
Total	3,092 (100%)	6,908 (100%)	10,000 (100%)

This is a relative risk of 0.33 – in favor of COVID!

This is the protective effect of immortal person-time.

Target trial framework

- Design the randomized trial you would use to test your hypothesis
  - Doesn't need to be feasible or ethical
    - (the observational study you do has to be ethical, obviously)
- Helps avoid immortal time bias by forcing the researcher to align all observations to the same "time zero"
  - Time zero is when participants are randomized to one of the treatment arms
- Treatment arms in a target trial to test $\Pr\left(Y^{a=1}\right)$ vs. $\Pr\left(Y^{a=0}\right)$ would involve assigning people to get COVID in pregnancy and to *not* get it during pregnancy

What would the target trial look like?

$\Pr\left(Y^{a=1}\right)$ vs. $\Pr\left(Y^{a=0}\right)$

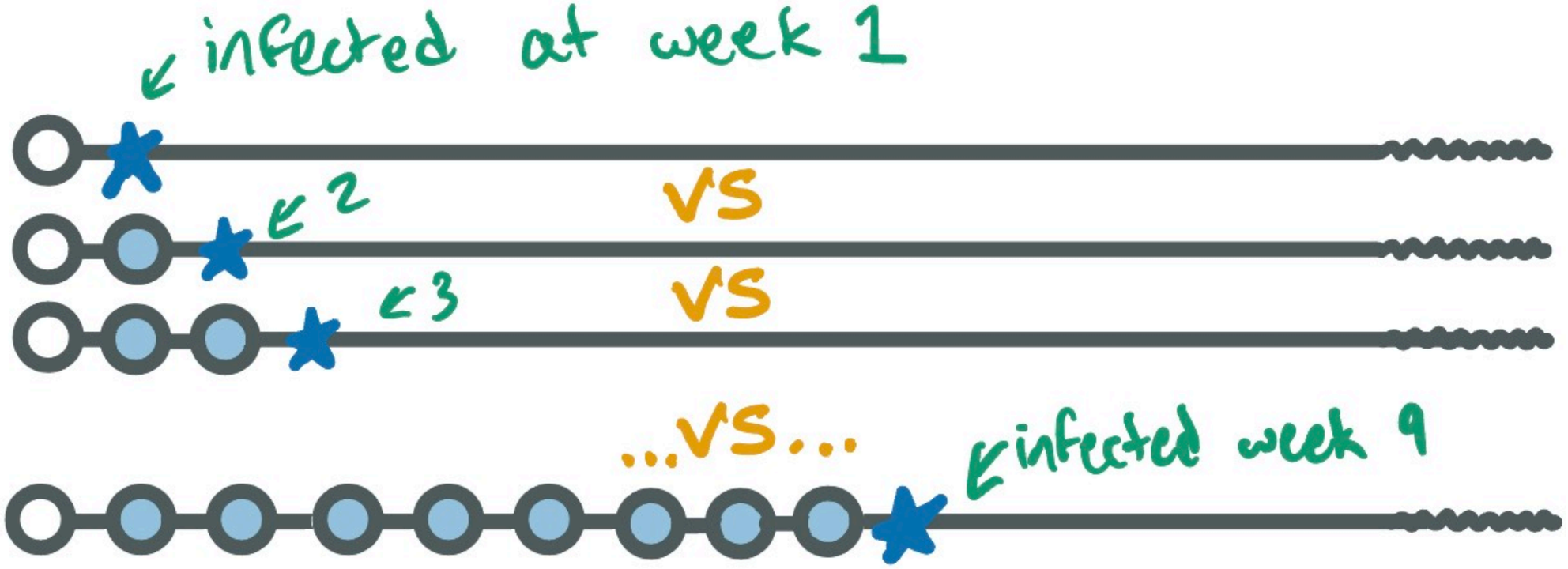
- We randomly assign $a = 1$ at the beginning of pregnancy, then that group has to get COVID in pregnancy (say, before 20 weeks).
- Either they all get it immediately following randomization, or they have to know something about their length of pregnancy in the absence of COVID ($T^{a=0}$ – which is unknown!) in order to make sure they get COVID before their pregnancy ends.
- So we could design a target trial assigning $a = 1$ vs. $a = 0$ only if we forced those assigned to $a = 1$ to get COVID immediately and those with $a = 0$ to quarantine through pregnancy, or else we would have a lot of non-compliance
- This would only test the effect of getting COVID very early in pregnancy

Target trials for time-varying risks

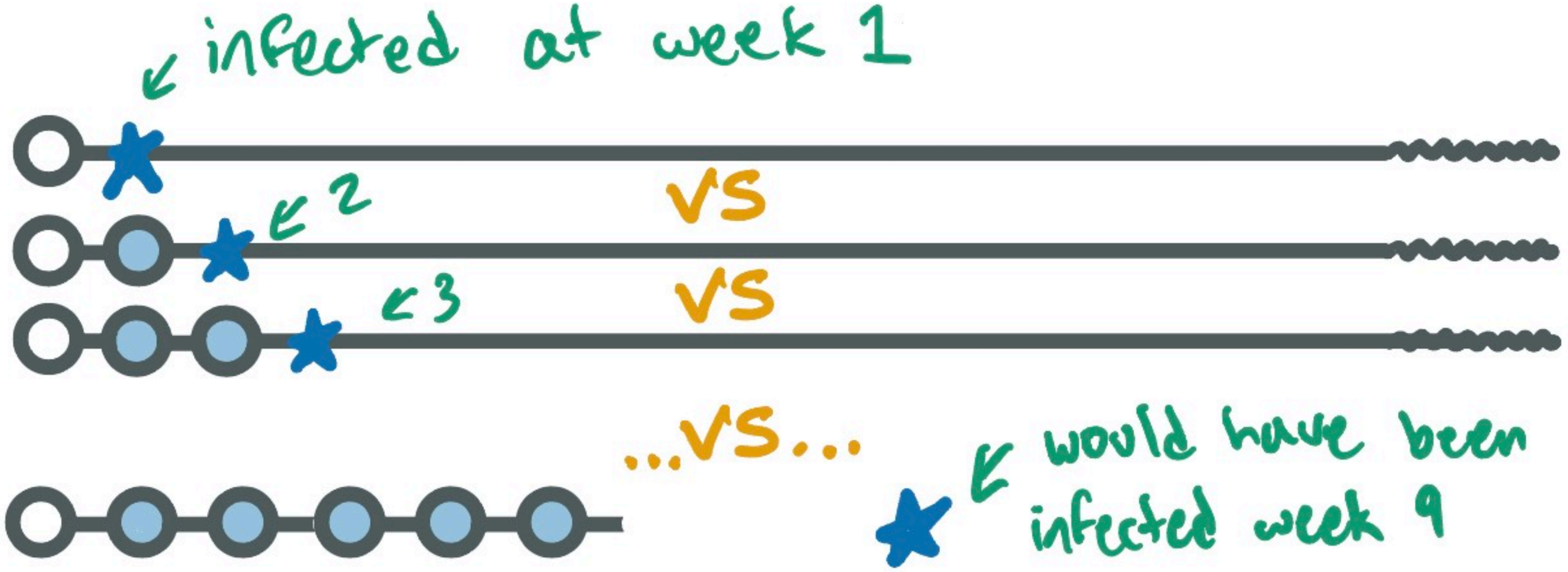
To understand the possibly time-varying risks of COVID across pregnancy, we could:

- At the beginning of pregnancy, randomly assign a gestational age at which to get COVID
  - Can compare $\Pr\left(Y^{x=10} = 1\right)$ vs. $\Pr\left(Y^{x=15} = 1\right)$ vs. $\Pr\left(Y^{x\geq 20} = 1\right)$, etc.
  - $x \geq 20$ (including $x = \infty$) means that it's after risk of spontaneous abortion, so can't affect the outcome (use as a reference strategy)
- Recruit people at varying stages of pregnancy (who haven't already had COVID) and randomize them to get COVID immediately or not
  - Can estimate $\Pr\left(Y^{x=10} = 1 \mid X >= 10\right)$ vs. $\Pr\left(Y^{x>10} = 1 \mid X >= 10\right)$
  - i.e., get COVID now vs. not right now (possibly never)

Target trial A



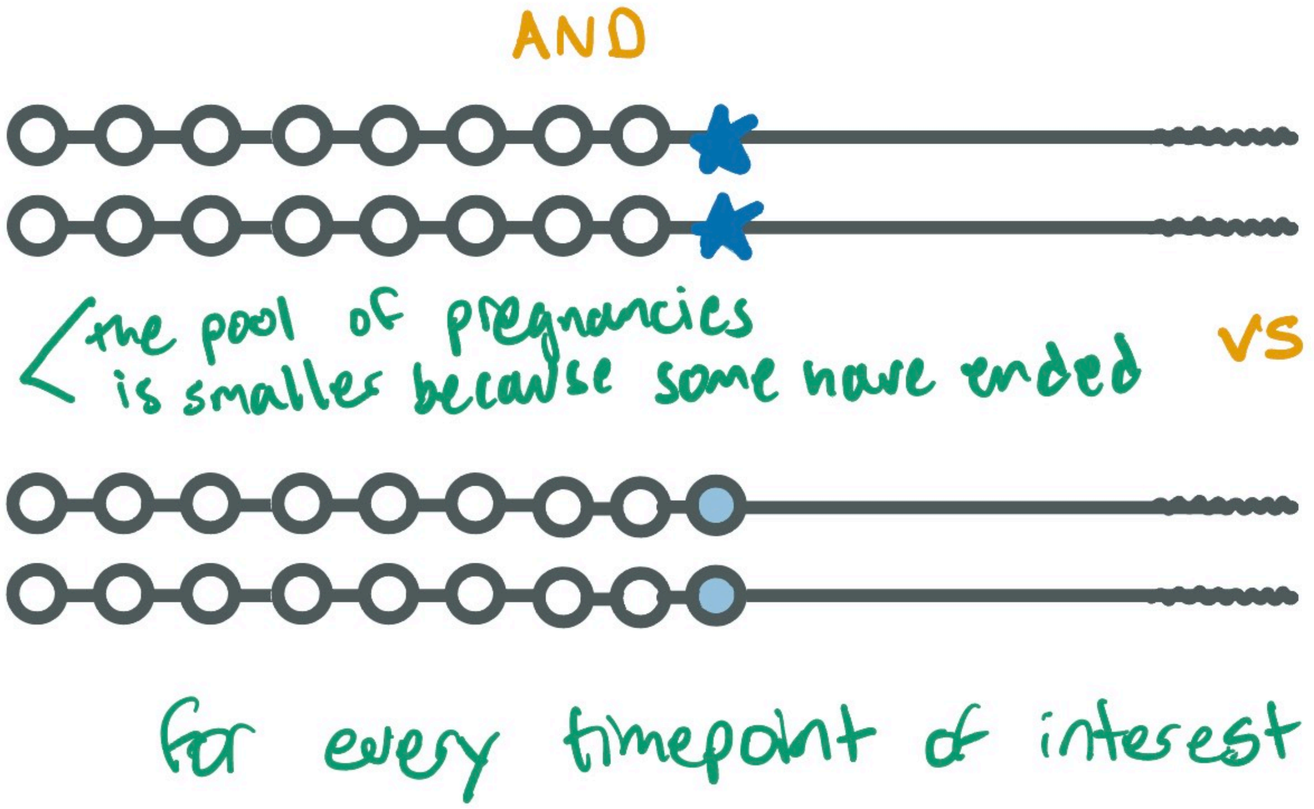
Target trial A



Target trial B, part 1

Target trial B, part 2

Target trial B, part….



Pros and cons

Trial A

When is the "best" time to get COVID in pregnancy, with respect to spontaneous abortion?

- All risks are directly comparable, e.g., $\Pr\left(Y^{x=10}=1\right)$ vs. $\Pr\left(Y^{x=15}=1\right)$
- Risk ratios/differences can all have the same reference strategy
  - $\Pr\left(Y^{x=10}=1\right)/\Pr\left(Y^{x\geq 20}=1\right)$ compared to $\Pr\left(Y^{x=15}=1\right)/\Pr\left(Y^{x\geq 20}=1\right)$
- Most spontaneous abortions will happen before getting COVID for those assigned late weeks
  - $\Pr\left(Y^{x=19}=1\right)/\Pr\left(Y^{x\geq 20}=1\right)$ will be close to 1 no matter what

Trial B

Given that I haven't gotten COVID yet, and am still pregnant, how much will my risk of spontaneous abortion increase if I get COVID now?

- Can't compare $\Pr\left(Y^{x=10}=1 \mid X>=10\right)$ vs. $\Pr\left(Y^{x=15}=1 \mid X>=15\right)$
- This also means that the relative magnitude of risk ratios/differences compared to a reference strategy won't be directly comparable
  - $\Pr\left(Y^{x=10}=1 \mid X>=10\right)/\Pr\left(Y^{x>10}=1 \mid X>=10\right)$ compared to $\Pr\left(Y^{x=15}=1 \mid X>=15\right)/\Pr\left(Y^{x>15}=1 \mid X>=15\right)$
- The risk ratios/differences are more targeted; e.g., $\Pr\left(Y^{x=19}=1 \mid X>=19\right)/\Pr\left(Y^{x>19}=1 \mid X>=19\right)$ isolates an acute effect of COVID on late spontaneous abortion

References for the first type of target trial

(We'll focus on the second)

Schnitzer ME, Guerra SF, Longo C, et al. A potential outcomes approach to defining and estimating gestational age-specific exposure effects during pregnancy. *Stat Methods Med Res*. 2022;:096228022110651.

Cain LE, Robins JM, Lanoy E, et al. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. T*he International Journal of Biostatistics.* 2010;6:1–42.

Young JG, Cain LE, Robins JM, et al. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in Biosciences.* 2011;3:119–143.

Specify the other components of the target trial

- Eligibility: currently pregnant, have not yet had COVID during pregnancy
  - Or if it's common to get COVID (other infection/exposure) multiple times in pregnancy, may not exclude, but stratify by previous infection
- Treatment assignment: stratify on gestational week, and assign with 50% chance to get COVID immediately vs. not
- Follow-up: until end of pregnancy (or at least 20 weeks for SAB)

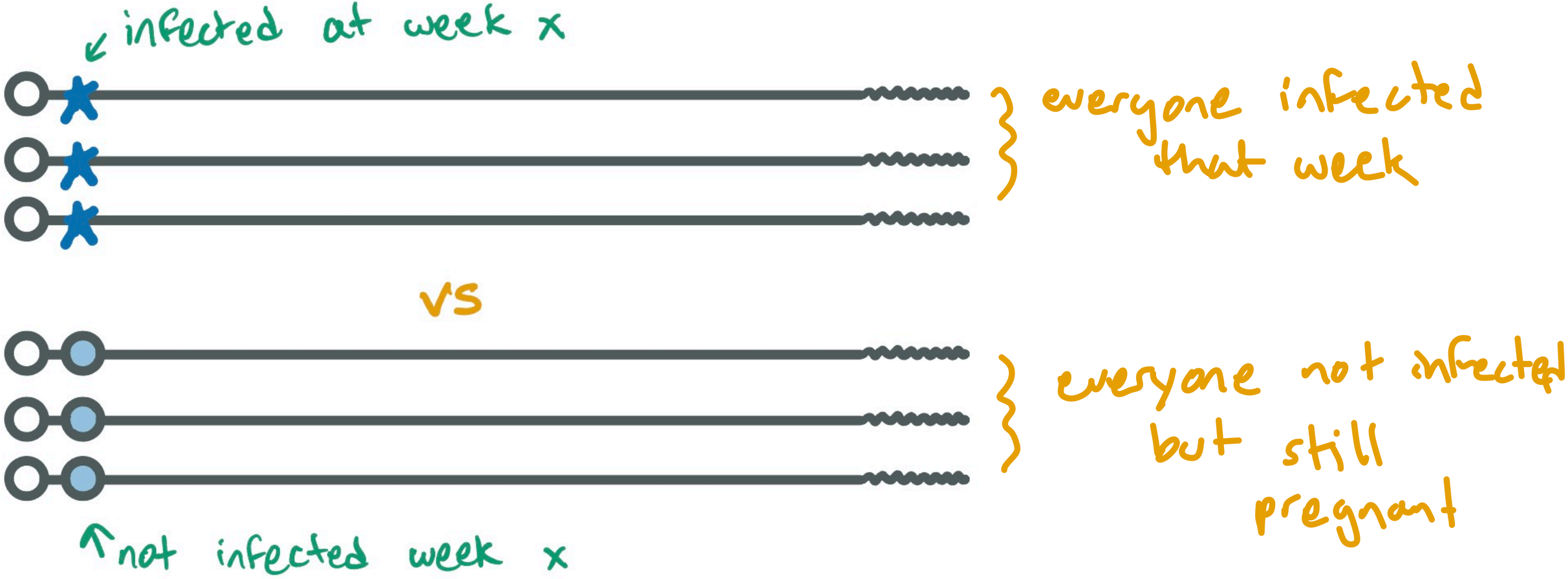
Point treatment vs. time-varying treatment

- Get COVID immediately vs not immediately
  - This is a one-time treatment at that moment (week)
  - You can tell right then whether someone has adhered or not
  - Not saying anything about what should happen next week
- Time-varying treatment might be: get COVID right now and not again, vs. never get COVID during pregnancy
  - Adherence requires following the treatment strategy throughout the rest of pregnancy
  - If there's non-adherence (people get COVID later), need to think about whether there are time-varying confounders
  - For other exposures there may be more complexities (or even with getting Covid again, or getting all shots in a multi-part vaccination)

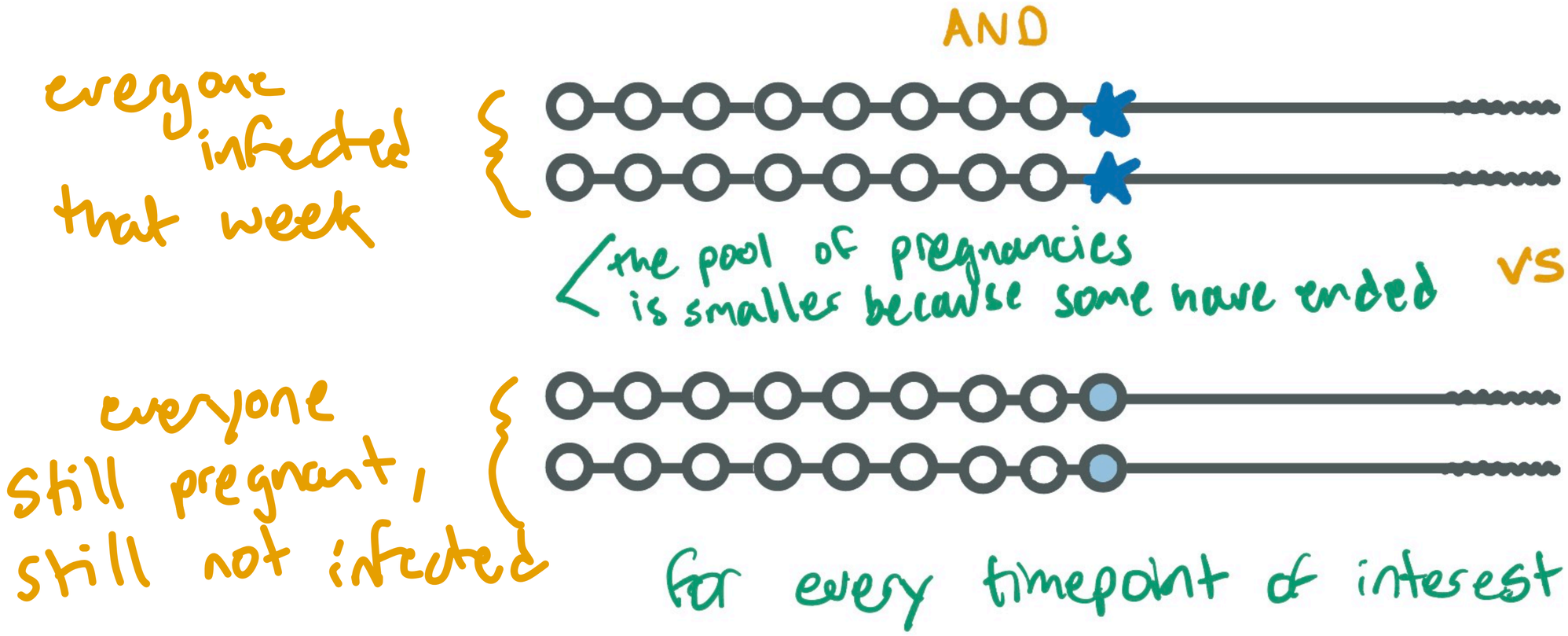
How to emulate in observational data?

- No one was assigned to get COVID or not, at any particular time
- However, we can use those that were exposed at say, week 12, to emulate what would have happened if they had been assigned to be exposed then
- We can use those who were still pregnant but had not yet been exposed at week 12 to emulate what would have happened to those assigned to be unexposed then

Emulation of target trial B, part 1



Emulation of target trial B, part…



Compare those who did vs. did not have COVID but were still pregnant at 12 weeks' gestation

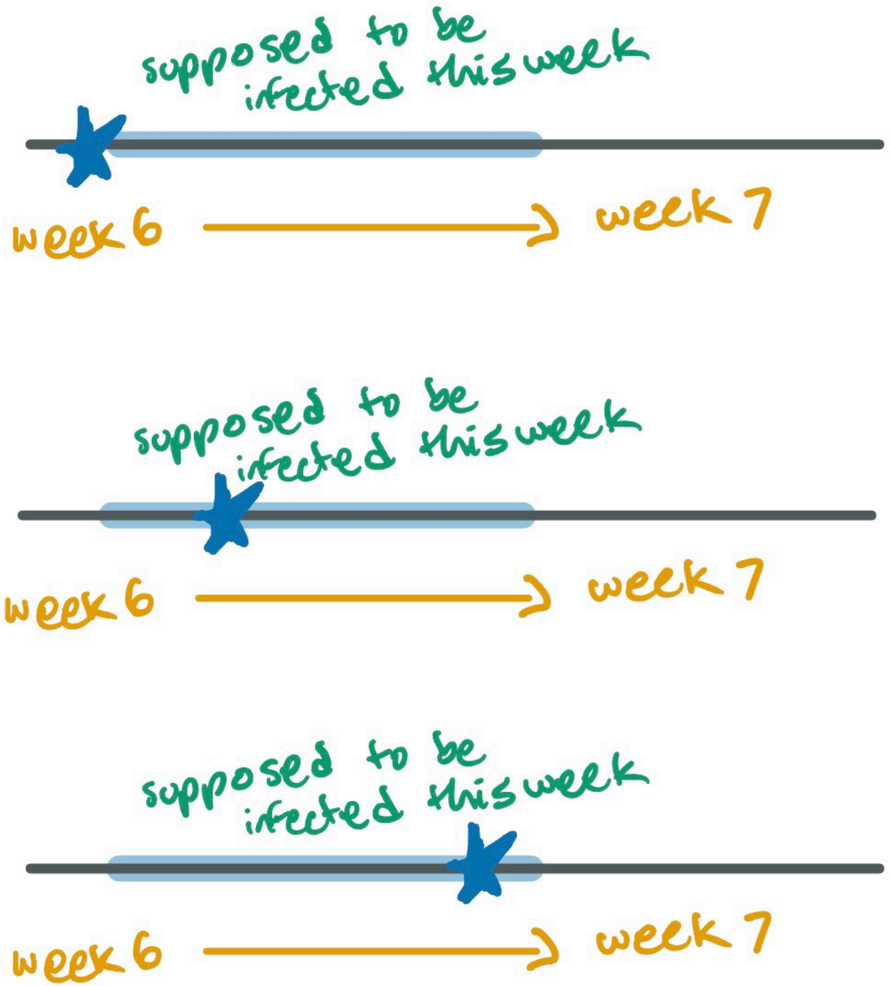
	COVID at 12 weeks		Total
	Yes	No	
Spontaneous abortion			
Yes	17 (8.1%)	532 (10%)	549 (10%)
No	194 (92%)	4,656 (90%)	4,850 (90%)
Total	211 (100%)	5,188 (100%)	5,399 (100%)

A little bit more immortal time bias

- We didn't actually compare those who were infected exactly at 12 weeks to all those still pregnant at 12 weeks
- Like our problem with the "get COVID sometime in pregnancy" trial, not everyone "randomized" at 12 weeks will get COVID right away
- We counted everyone who got COVID at 12 weeks + 1 day, …, 12 weeks + 6 days as exposed at 12 weeks
- We are missing those who were assigned to get COVID at 12 weeks, didn't get it immediately (e.g., would have gotten it at 12 weeks + 4 days), but their pregnancy ended before that happened

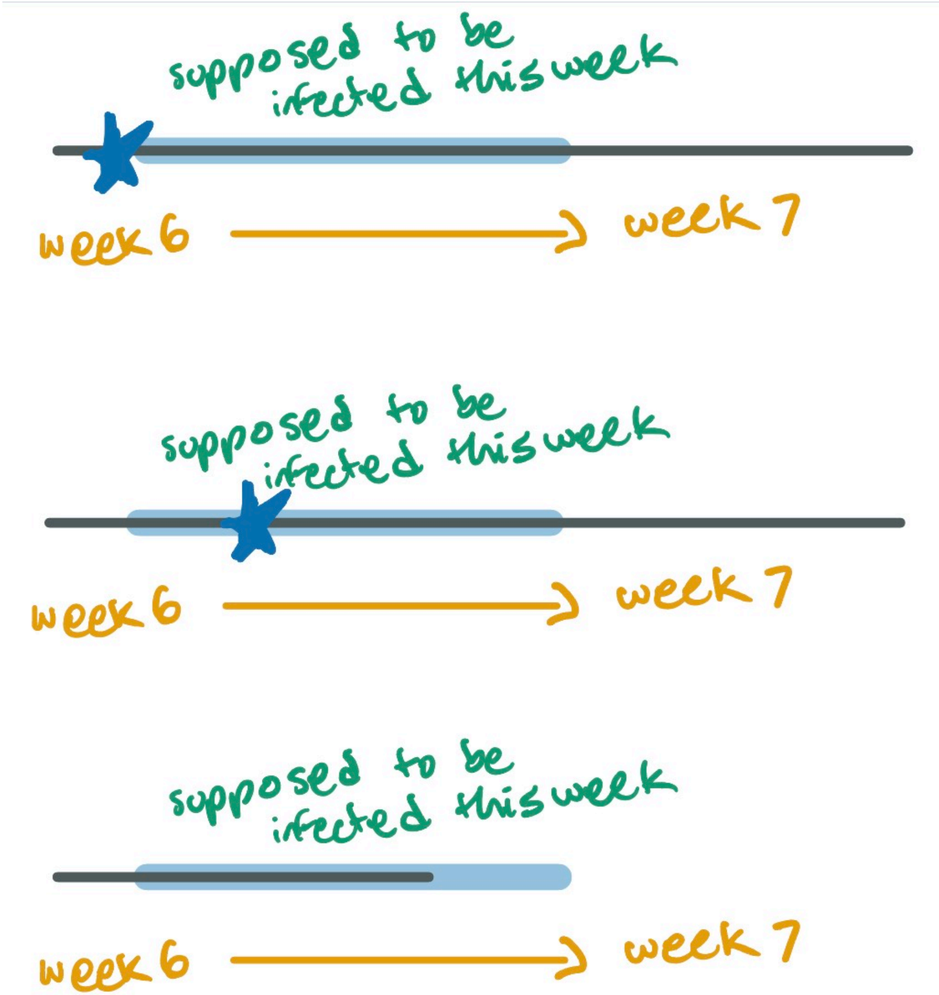
A little bit more immortal time bias

- Those events at 12 weeks + 1 day, …, 12 weeks + 6 days will be counted as unexposed
- Even if those people were “randomized” to be exposed at 12 weeks (but we didn’t know that)
- This is a problem if there are a lot of events in that time period!



A little bit more immortal time bias

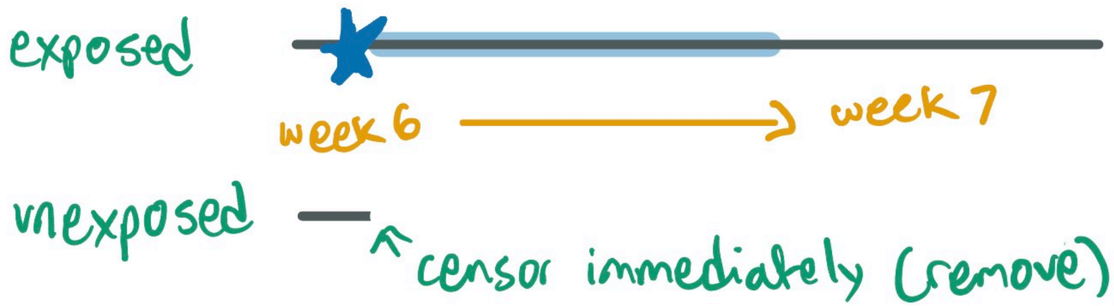
- Those events at 12 weeks + 1 day, …, 12 weeks + 6 days will be counted as unexposed
- Even if those people were “randomized” to be exposed at 12 weeks (but we didn’t know that)
- This is a problem if there are a lot of events in that time period!



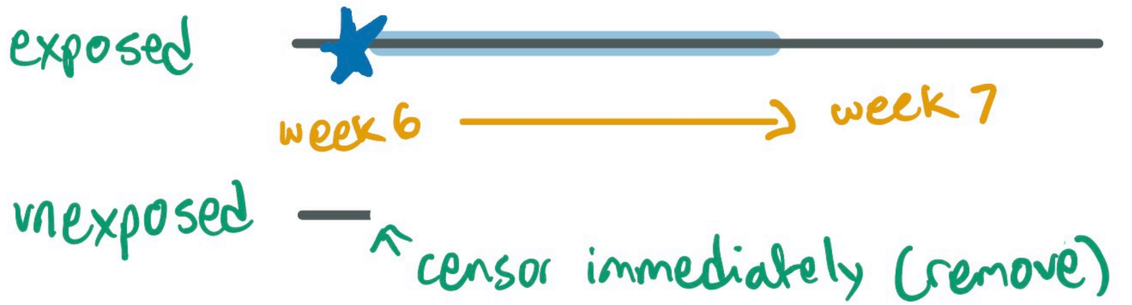
Solutions

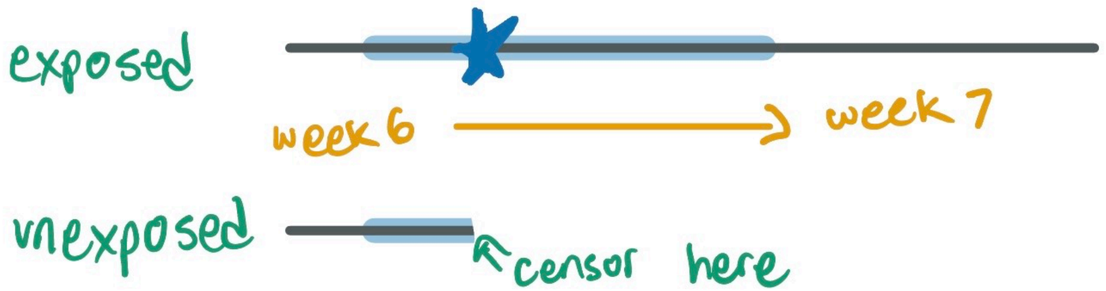
- Redefine the target trial so that infection must happen on a specific day of gestational age
  - This will be difficult if few people are infected on any given day
- Allow for a grace period: randomize at 12 weeks, but tell people they have the whole week to get infected
  - In the observational analysis, events that happen that week among the unexposed will count for *both* exposure groups, since we don't know which they were randomized to
- Use the smallest time scale that is computational feasible, aligns with the data, and doesn't allow for too many events to occur before exposure can take place
  - e.g., we'd never have data to the millisecond on exposure status, and there would be no point in randomizing at every millisecond because it would be almost impossible to have an event before the next millisecond

Grace period

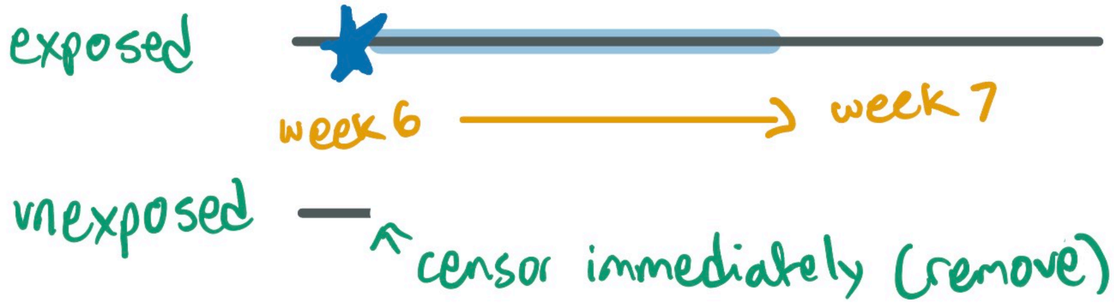


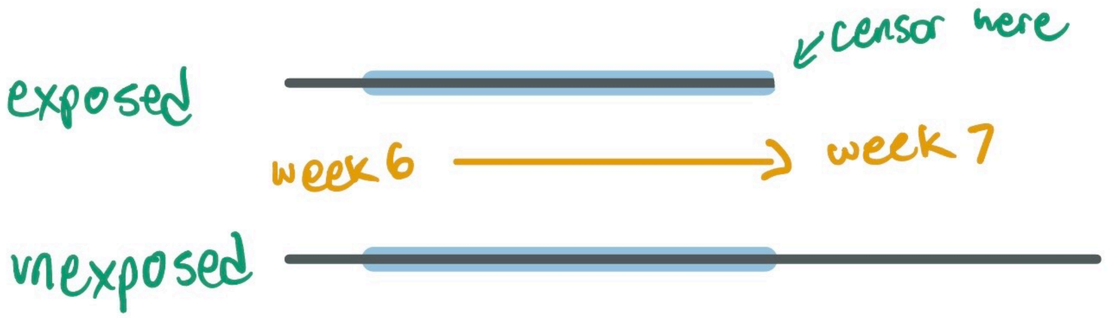
Grace period

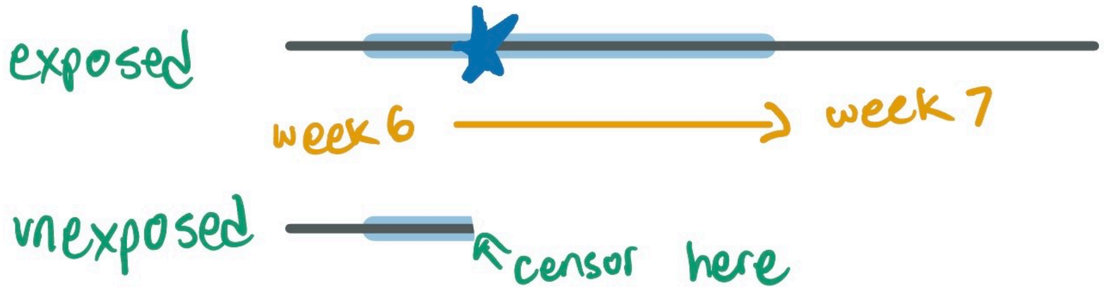




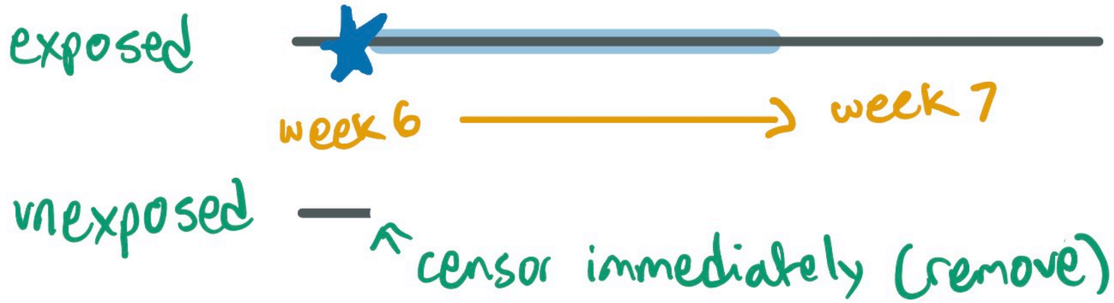
Grace period

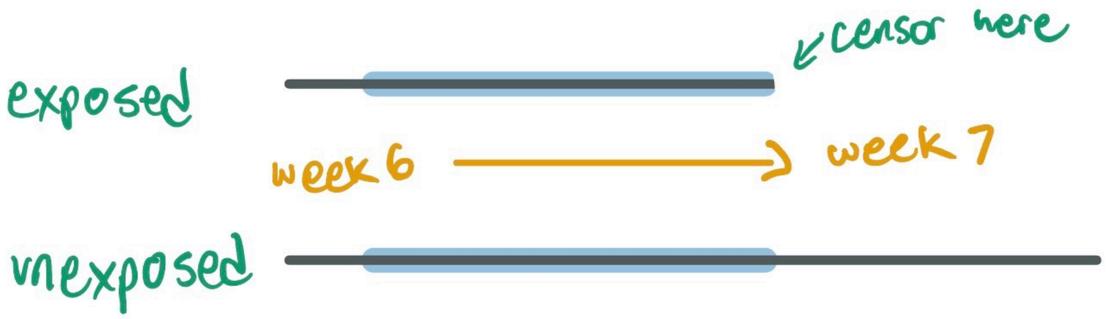


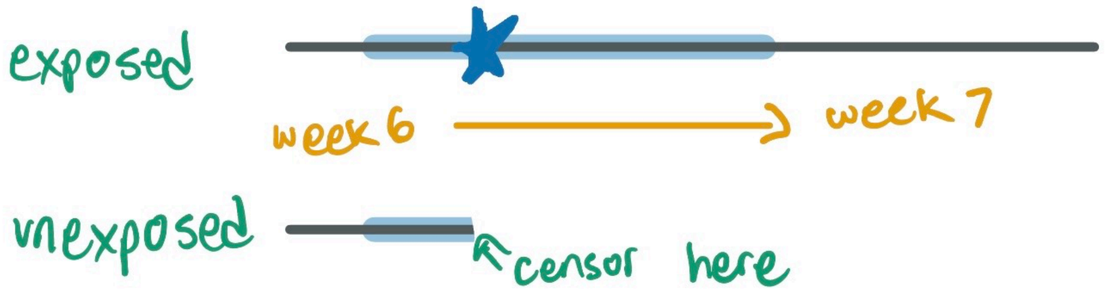


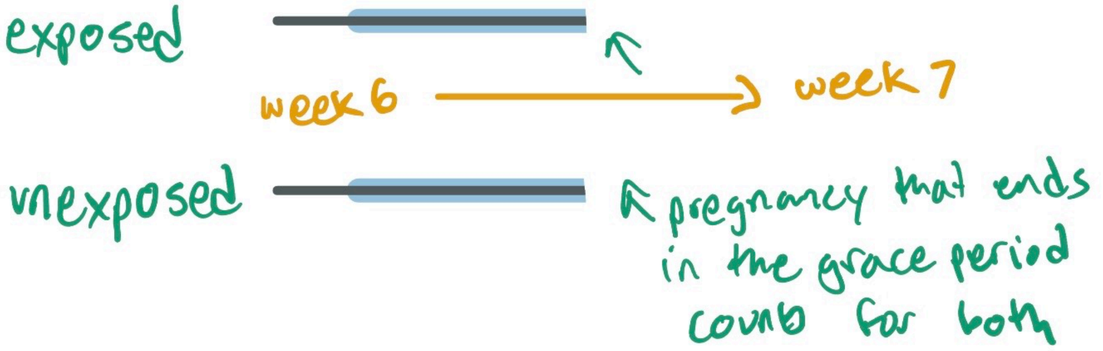


Grace period









Since no one was assigned a specific exposure, we can assign them to both

"Clone" the data

id	time_zero	exposed	time_exposed	time_ended
712	12	0	–	12.57
712	12	1	–	12.57
4603	12	0	12.71	38.86
4603	12	1	12.71	38.86
8527	12	0	12.00	39.86
8527	12	1	12.00	39.86
9493	12	0	–	15.57
9493	12	1	–	15.57

Since no one was assigned a specific exposure, we can assign them to both

"Clone" the data

id	time_zero	exposed	time_exposed	time_ended
712	12	0	–	12.57
712	12	1	–	12.57
4603	12	0	12.71	38.86
4603	12	1	12.71	38.86
8527	12	0	12.00	39.86
8527	12	1	12.00	39.86
9493	12	0	–	15.57
9493	12	1	–	15.57

} both stay! could have been either

← censor when exposed

← great!

← censor immediately (delete)

← great!

← great!

← censor after grace period

Then we censor the observations that don't match their assigned treatment strategy

Censored data

id	time_zero	exposed	time_exposed	time_ended	time_in	time_out	event
712	12	0	–	12.57	12	12.57	1
712	12	1	–	12.57	12	12.57	1
4603	12	0	12.71	38.86	12	12.71	0
4603	12	1	12.71	38.86	12	38.86	1
8527	12	1	12.00	39.86	12	39.86	1
9493	12	0	–	15.57	12	15.57	1
9493	12	1	–	15.57	12	13.00	0

Then we censor the observations that don't match their assigned treatment strategy

Censored data

id	time_zero	exposed	time_exposed	time_ended	time_in	time_out	event
712	12	0	–	12.57	12	12.57	1
712	12	1	–	12.57	12	12.57	1
4603	12	0	12.71	38.86	12	12.71	0
4603	12	1	12.71	38.86	12	38.86	1
8527	12	1	12.00	39.86	12	39.86	1
9493	12	0	–	15.57	12	15.57	1
9493	12	1	–	15.57	12	13.00	0

Then we censor the observations that don't match their assigned treatment strategy

Censored data

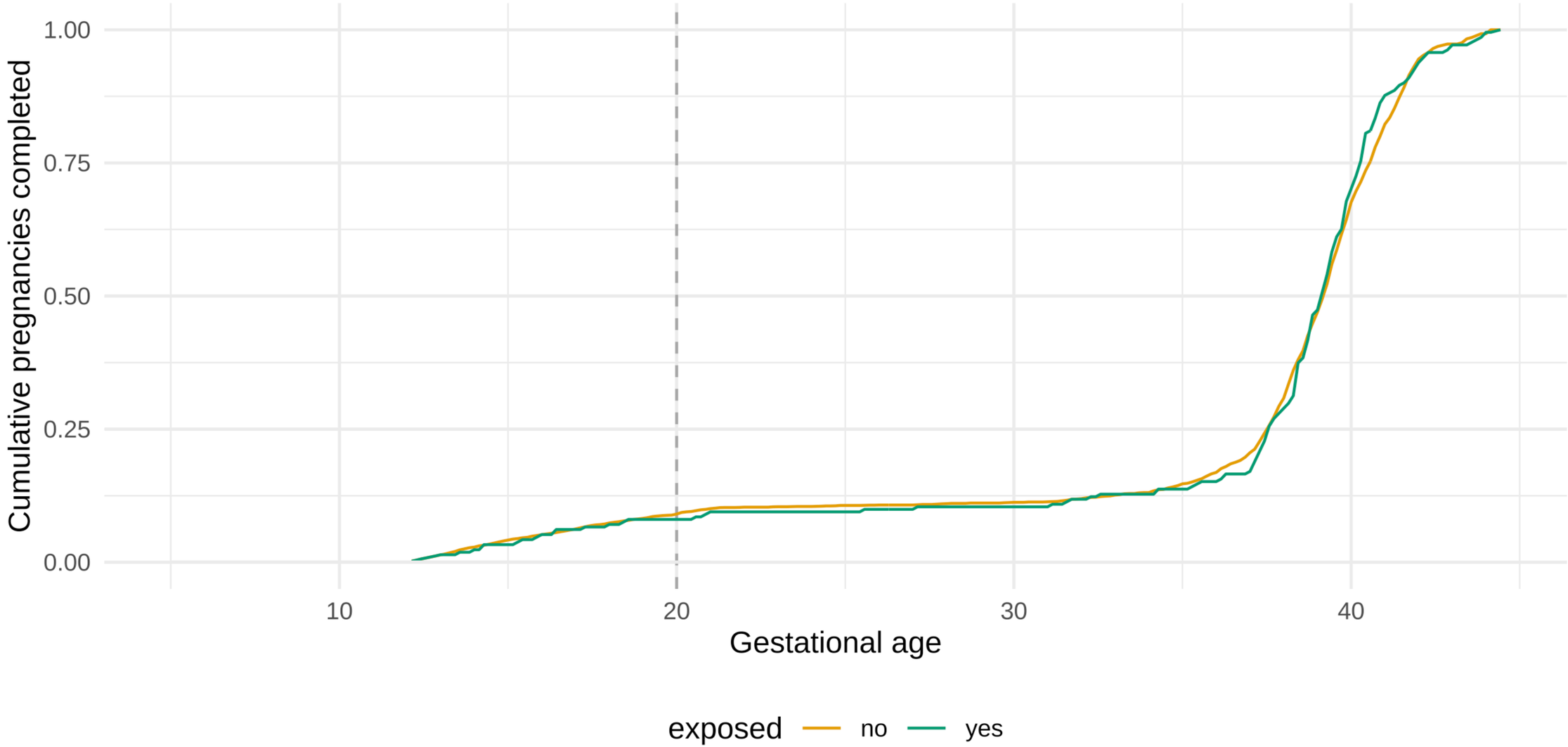
id	time_zero	exposed	time_exposed	time_ended	time_in	time_out	event
712	12	0	–	12.57	12	12.57	1
712	12	1	–	12.57	12	12.57	1
4603	12	0	12.71	38.86	12	12.71	0
4603	12	1	12.71	38.86	12	38.86	1
8527	12	1	12.00	39.86	12	39.86	1
9493	12	0	–	15.57	12	15.57	1
9493	12	1	–	15.57	12	13.00	0

Then we censor the observations that don't match their assigned treatment strategy

Censored data

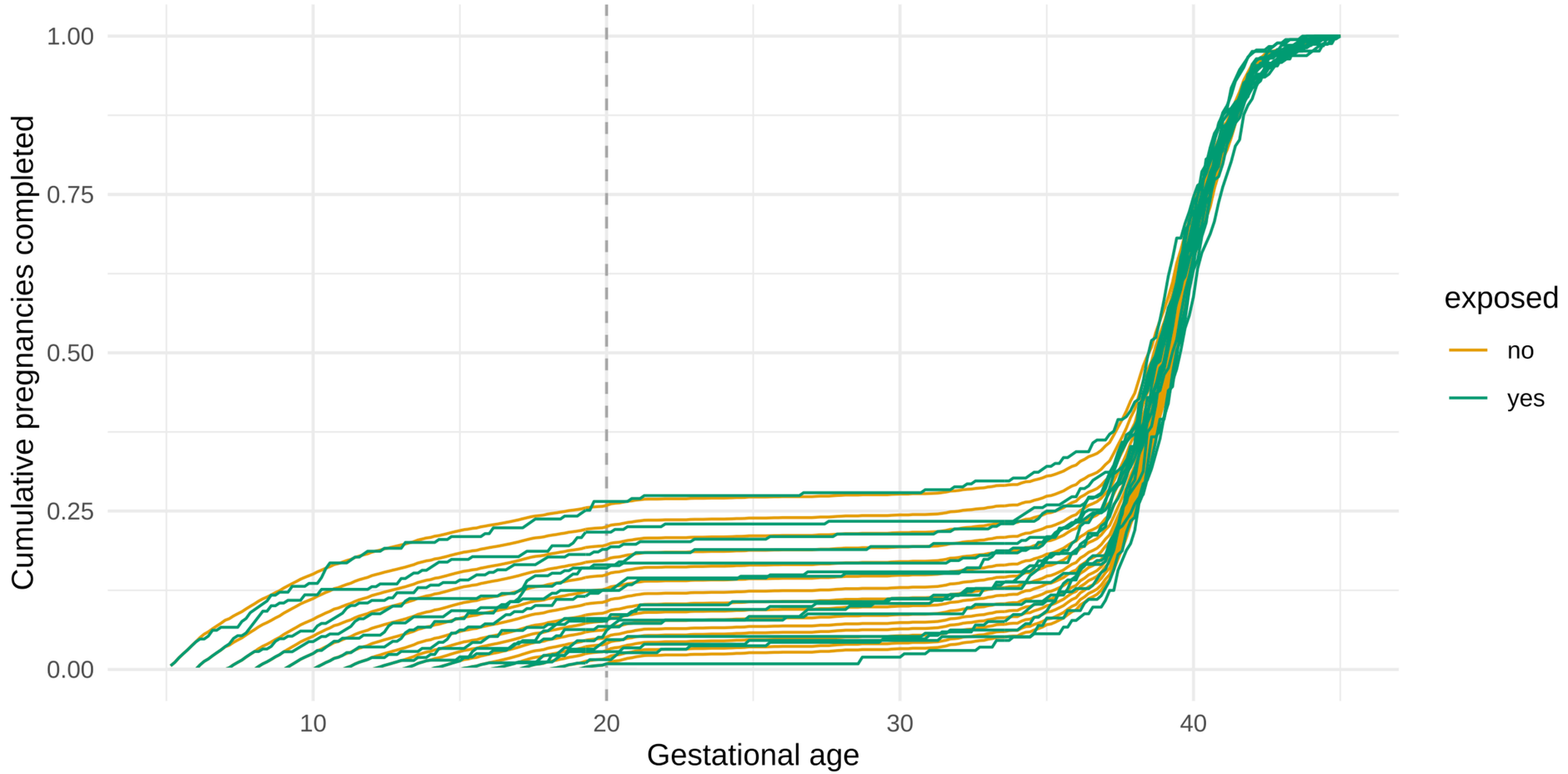
id	time_zero	exposed	time_exposed	time_ended	time_in	time_out	event
712	12	0	–	12.57	12	12.57	1
712	12	1	–	12.57	12	12.57	1
4603	12	0	12.71	38.86	12	12.71	0
4603	12	1	12.71	38.86	12	38.86	1
8527	12	1	12.00	39.86	12	39.86	1
9493	12	0	–	15.57	12	15.57	1
9493	12	1	–	15.57	12	13.00	0

Can use Kaplan-Meier estimator to fit survival curves



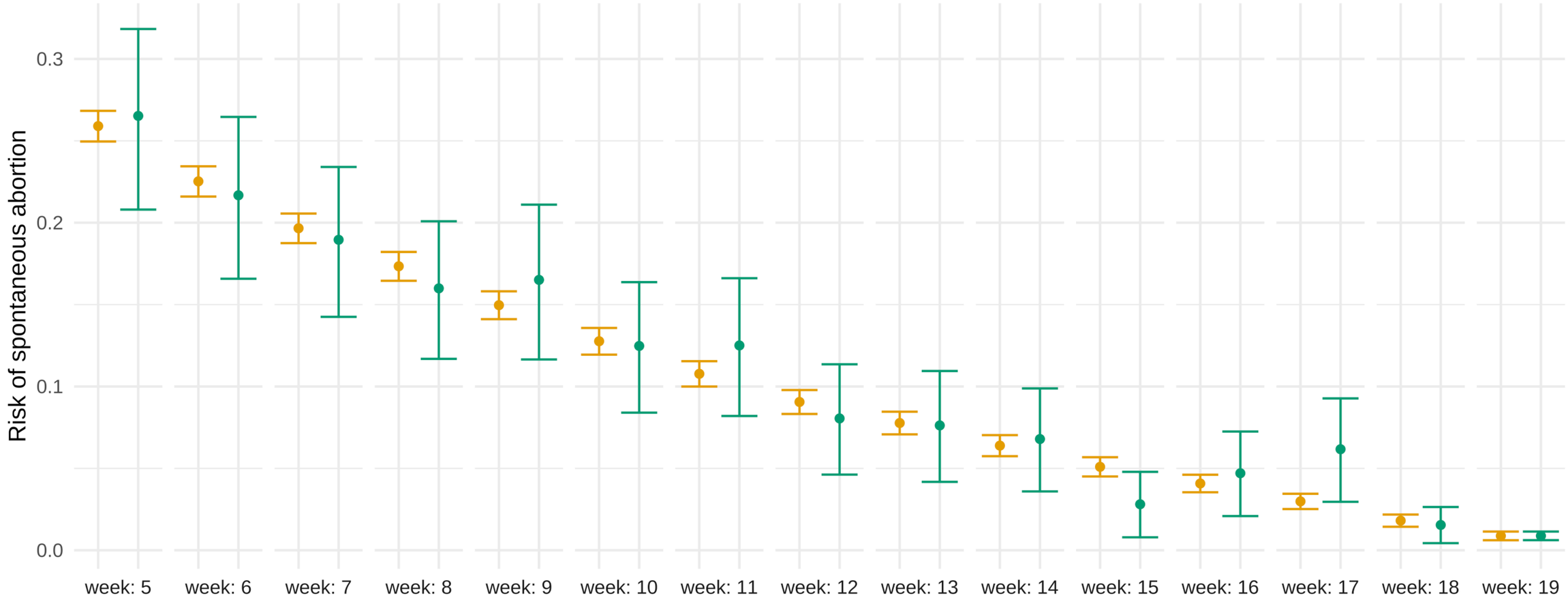
Now do so at each gestational week (before 20)

Survival curves for each time zero



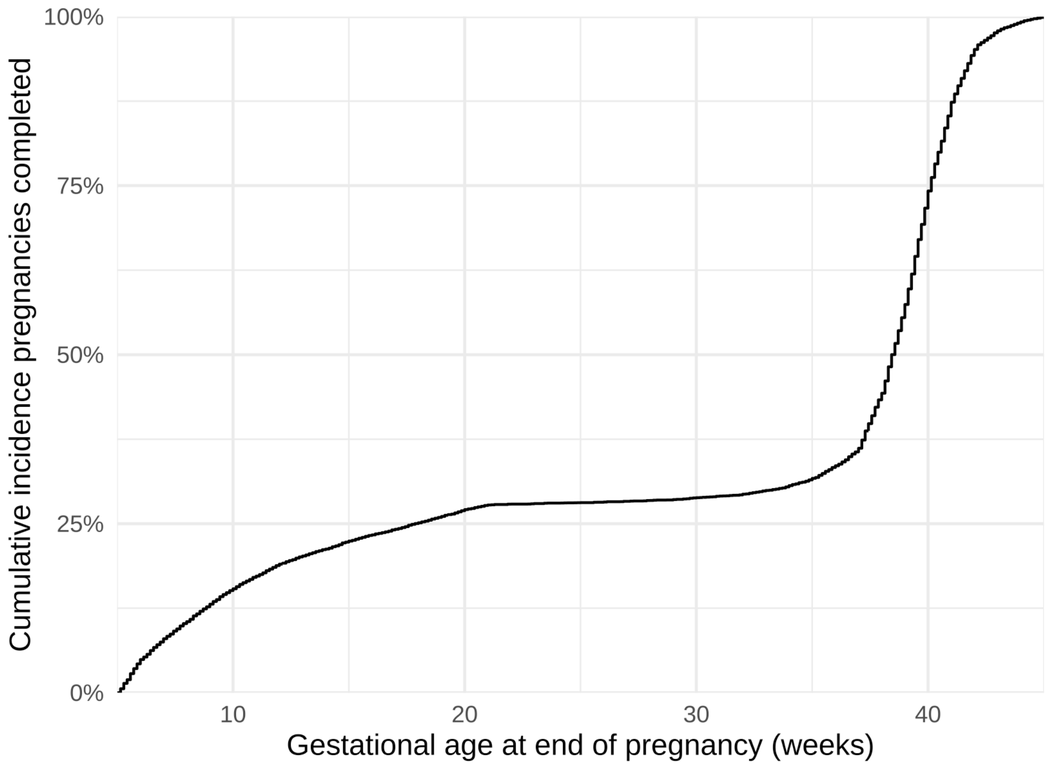
Compare risks across weeks of infection (time zero)

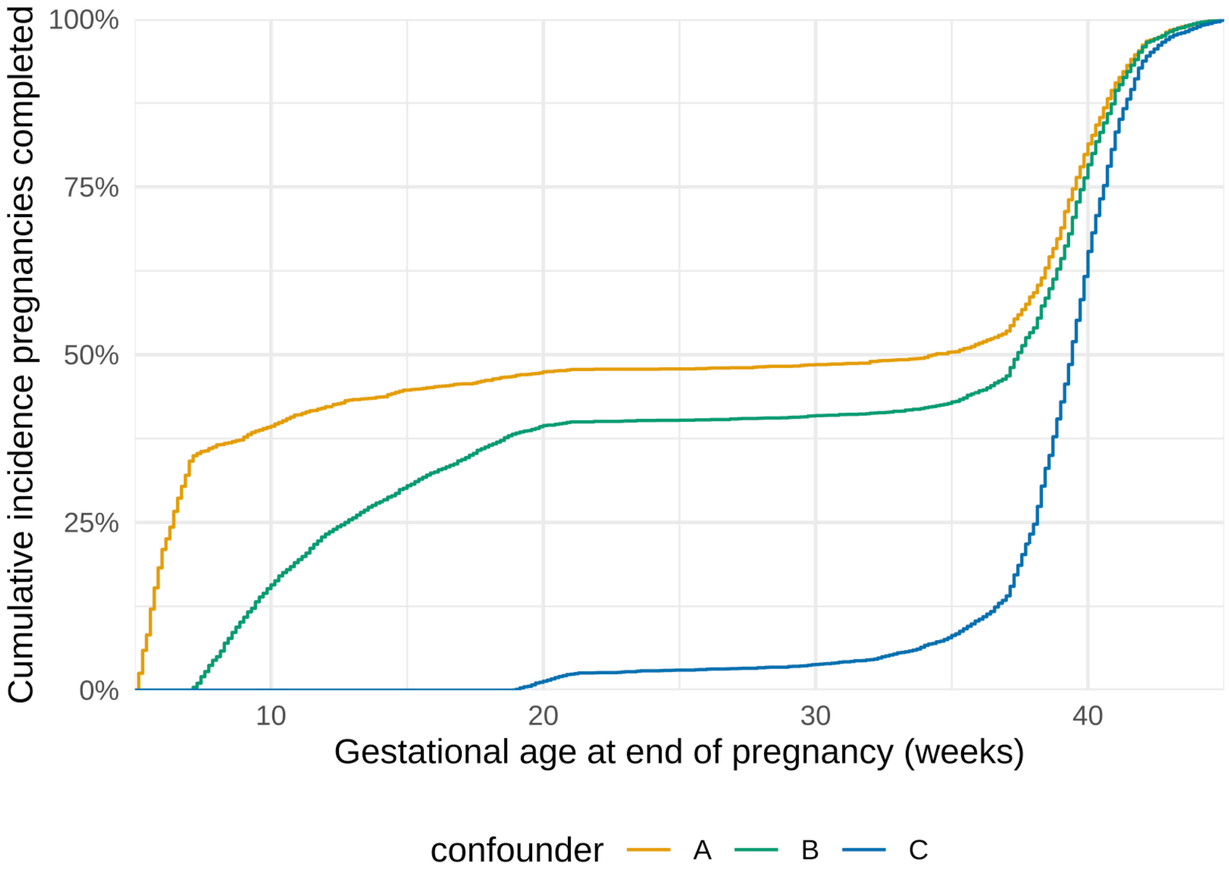
Where is the survival curve at 20 weeks after each time zero?



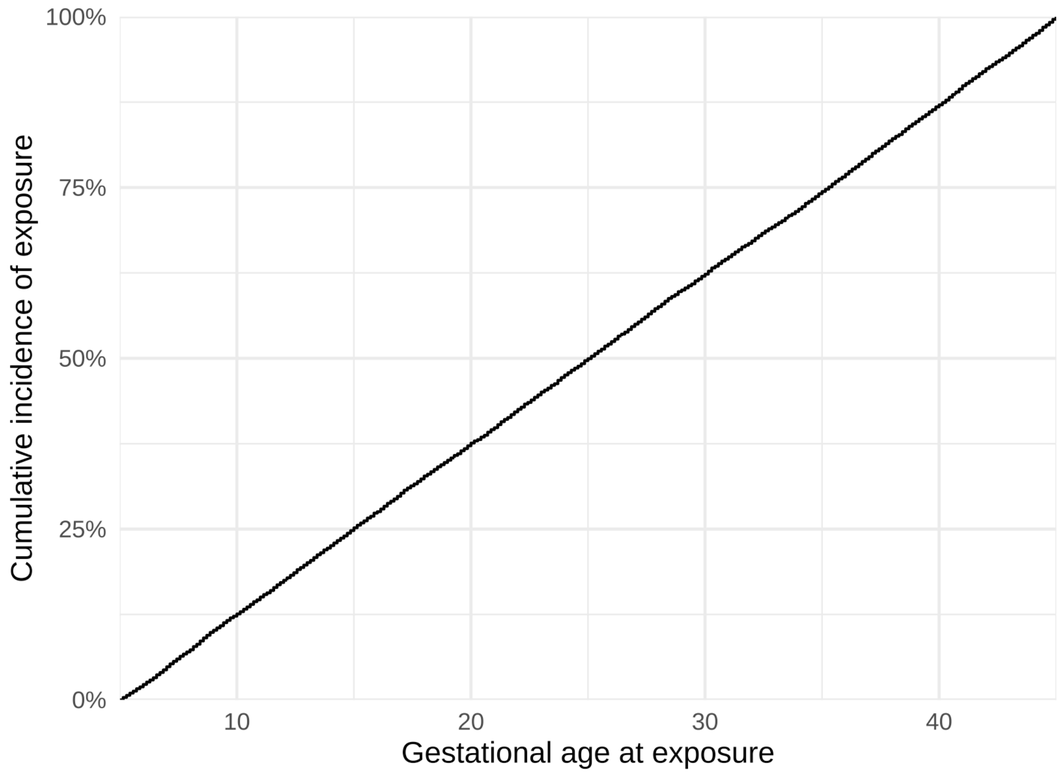
Lab 2

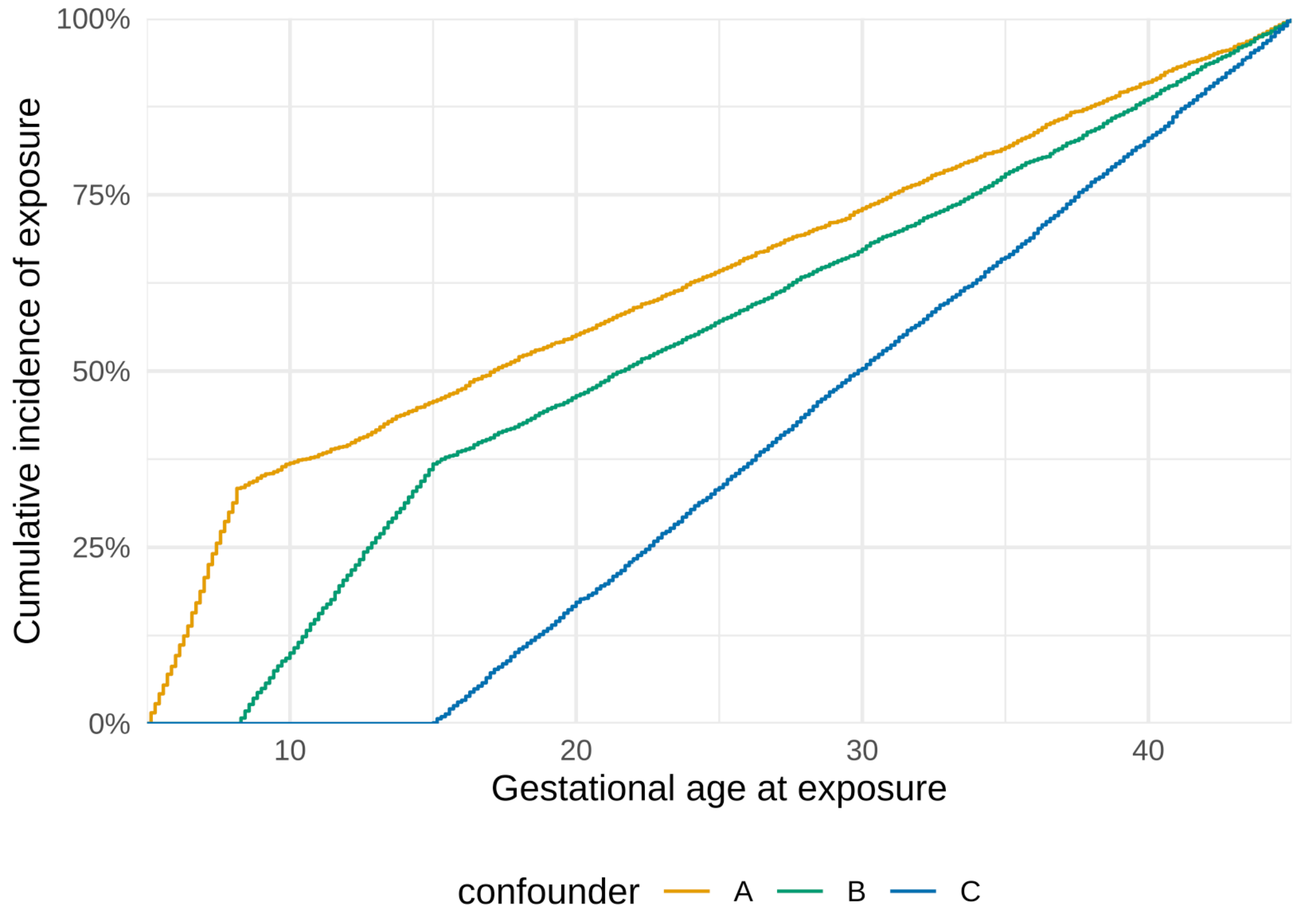
Like any other observational analysis, confounding is an issue





Like any other observational analysis, confounding is an issue

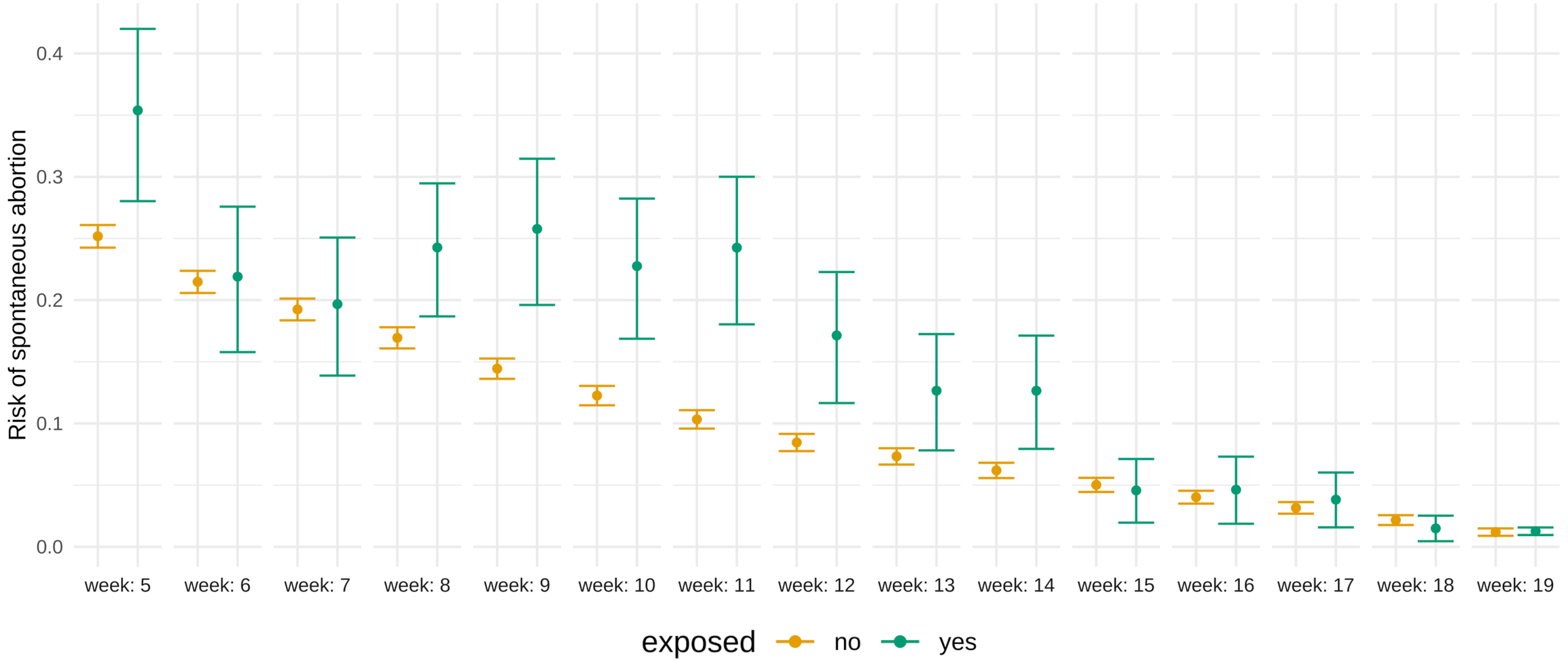




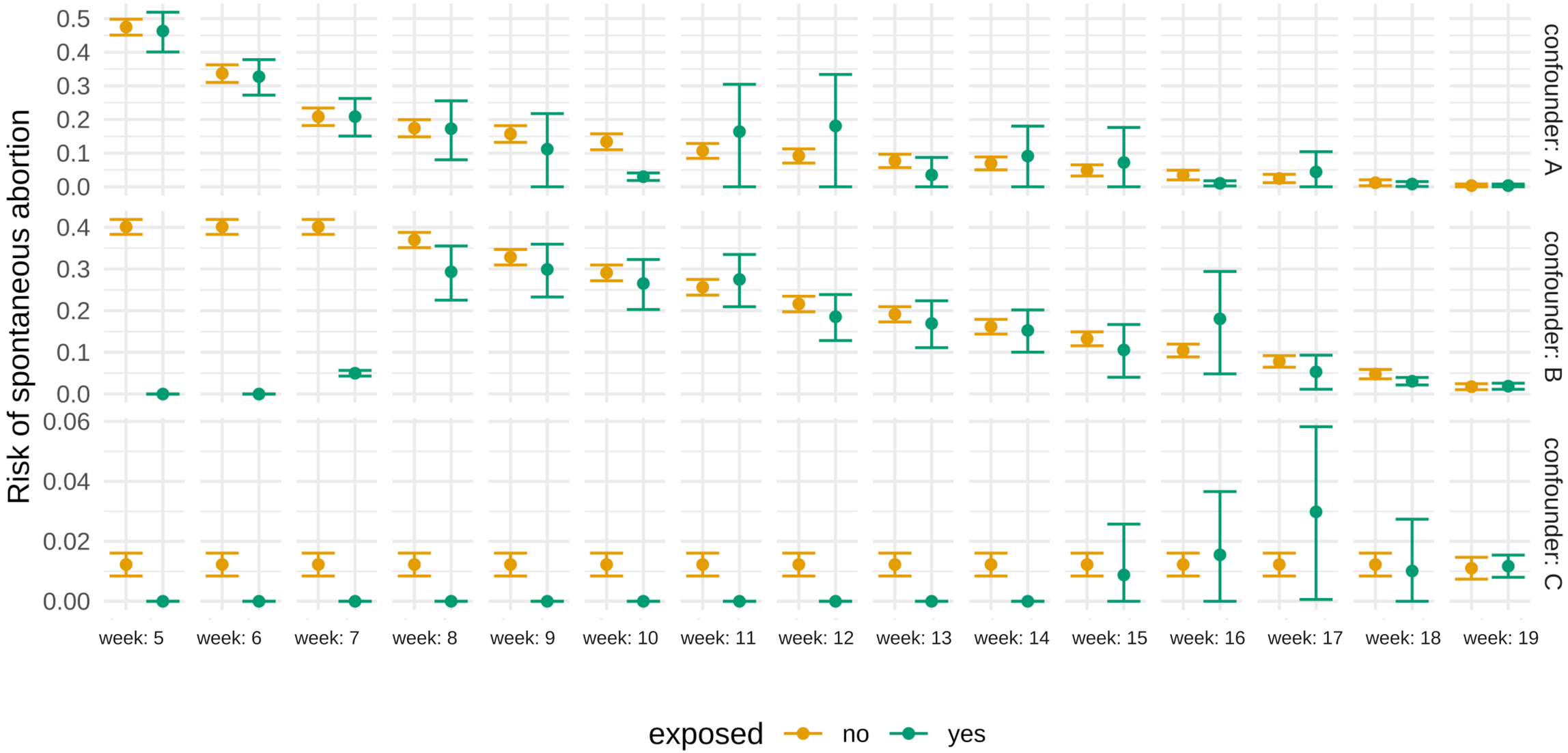
Time-varying confounding?

- These examples are not affected by time-varying confounders
  - We can adjust for the selection bias induced by the censoring of people who are later exposed with the baseline confounders
  - If we assume that there's nothing that affects being exposed later that didn't affect being exposed earlier (and therefore that was considered a baseline covariate)
- Time-varying confounding may come into play with more complex treatment strategies
  - Other methods (e.g., inverse probability weighting, which could be used in the point-treatment case as well)

Unadjusted analysis



K-M risks stratified by (single categorical) confounder

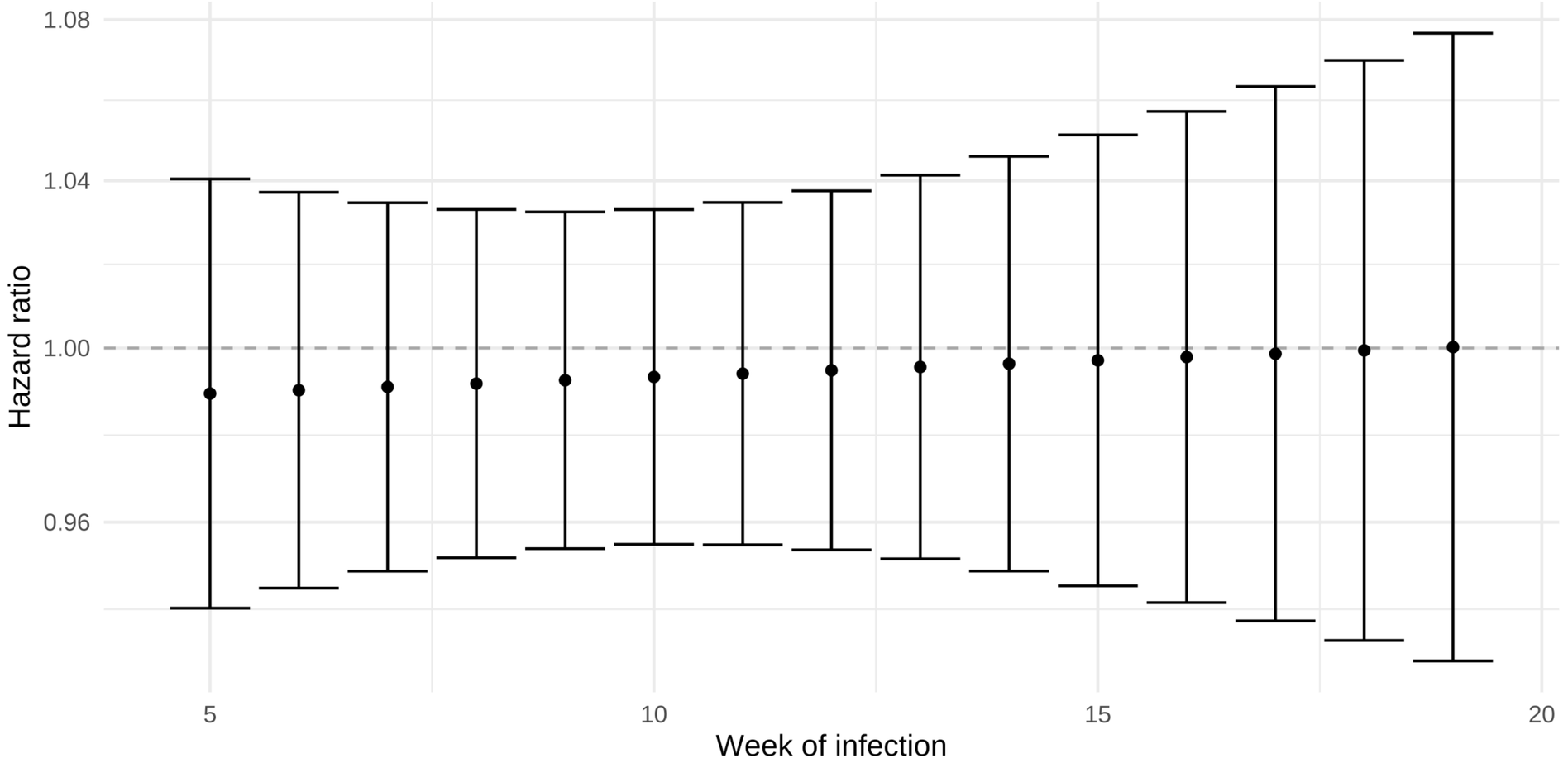


We need a model

```
cox_mod <- coxph(Surv(time_in, time_out, event) ~
                 exposed*time_zero + strata(confounder),
                     data = expanded_data_censored, id = id)
```

	HR	95% CI	p-value
exposed	0.99	0.91, 1.06	0.7
time_zero	1.00	1.00, 1.00	>0.9
exposed * time_zero	1.00	0.99, 1.01	0.8

Hazard ratios



Pooled logistic regression

Expand the cloned and censored data to have one day for every day/week/ whatever is computational feasible (long person-time data)

id	time_zero	exposed	time_in	time_out	week	event
712	5	1	5	6.00	5	0
712	5	0	5	12.57	5	0
712	5	0	5	12.57	6	0
712	5	0	5	12.57	7	0
712	5	0	5	12.57	8	0
712	5	0	5	12.57	9	0
712	5	0	5	12.57	10	0
712	5	0	5	12.57	11	0
712	5	0	5	12.57	12	0
712	5	0	5	12.57	13	1
712	6	1	6	7.00	6	0
712	6	0	6	12.57	6	0
712	6	0	6	12.57	7	0
712	6	0	6	12.57	8	0
712	6	0	6	12.57	9	0
712	6	0	6	12.57	10	0
712	6	0	6	12.57	11	0
712	6	0	6	12.57	12	0
712	6	0	6	12.57	13	1
712	7	1	7	8.00	7	0
712	7	0	7	12.57	7	0
712	7	0	7	12.57	8	0
712	7	0	7	12.57	9	0
712	7	0	7	12.57	10	0
712	7	0	7	12.57	11	0
712	7	0	7	12.57	12	0
712	7	0	7	12.57	13	1
712	8	1	8	9.00	8	0
712	8	0	8	12.57	8	0
712	8	0	8	12.57	9	0

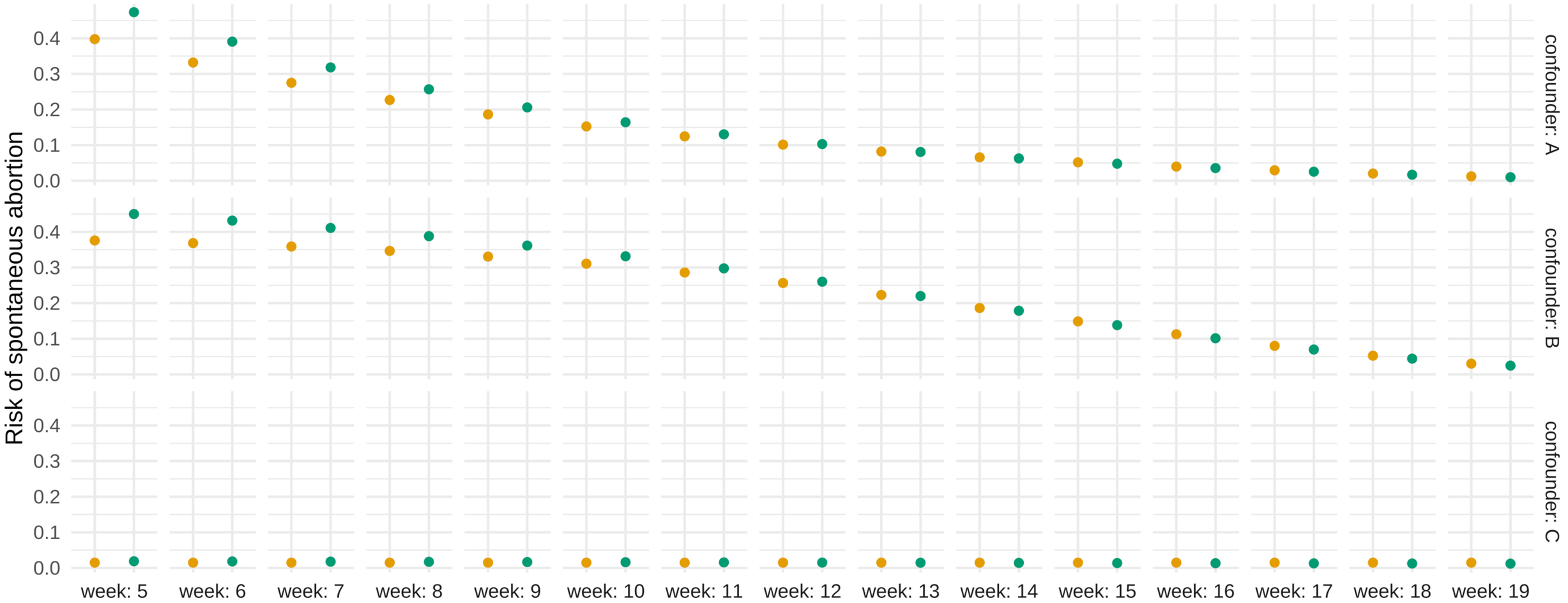
Pooled logistic regression

Have to specify baseline hazard as a function of time, e.g., with splines

```
glm_mod <- glm(event ~ splines::ns(week, 4) + exposed +
                       exposed:time_zero,
               data = long_cloned_and_censored,
               family = binomial())
```

This is a model for the log(hazard). We can easily estimate risks using the fitted values from the model.

Risk estimates from pooled logistic regression

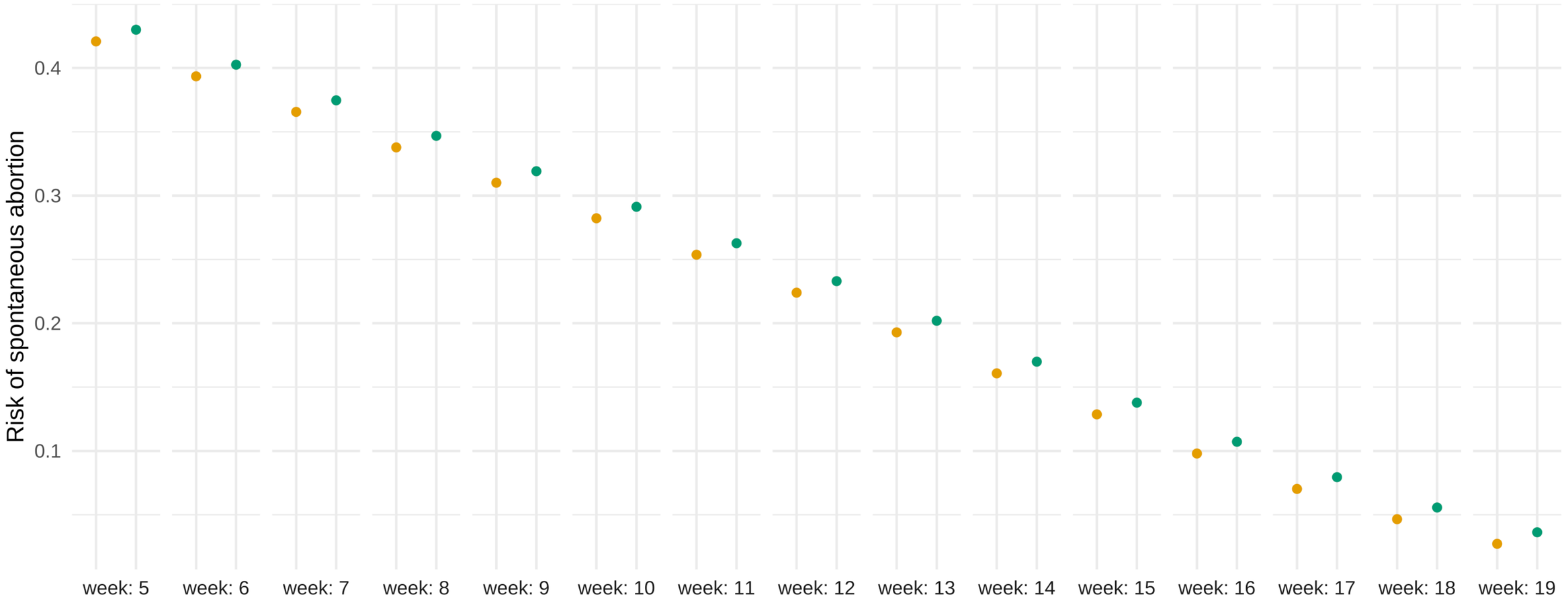


Standardized risk estimates

We can’t estimate separate risks for every confounder stratum

- We might standardize to the covariate distribution of the whole population at baseline.
- Depending on the study design, you may not know the whole population at baseline (with late enrollment) or you may have oversampled exposed participants, and you may want to choose to standardize to some meaningful subset of the population.
- We standardize by estimating the hazard at each timepoint from the model for each person in the population.
- We can then use the hazards to compute the risk at a certain timepoint (here, 20 weeks), and average the risk over the population.

Standardized risk estimates



Lab 3

Problems in actual data!

e.g., enrollment

- We have assumed that participants enroll as soon as they are pregnant, if not before
- Obviously this is very hard in real life!
- Let's consider some problems/solutions/possible biases

Enroll unexposed people anytime in pregnancy

- This is ok as long as enrollment time is independent of event time (conditional on covariates)
- E.g., the unexposed people who enrolled early have the same distribution of pregnancy length as those who enrolled late
  - We can assume that there are “missing” pregnancies among those who enrolled late (because those pregnancies ended early), but we know how many because the early enrollers have the same distribution
  - Literature on “left truncation”
  - If there are few people who enrolled early, any random weirdness/bias in their distribution of events can “infect” the whole survival curve (Tsai et al 1987)
- There has to be some (large) risk of exposure if everyone is unexposed at enrollment!

Enroll exposed after exposure

- The exposed are “oversampled”
  - This is ok because we are of course conditioning on exposure
- If, e.g., a pregnancy loss or delivery happens soon after exposure, will not have a chance to enroll while eligible (i.e., pregnant)
  - These are potentially the causal events, if the exposure has an acute effect!
  - We may miss harms of the exposure (bias toward or beyond the null)

Enroll after pregnancy

- Potentially get back the events that happen soon after exposure, but may have selection bias
- People who have had an adverse event and were exposed may be more likely to enroll than those whose pregnancies went smoothly
  - Bias away from the null

Real-world data

- E.g., claims, medical records
- Depends on what's measured and recorded, and how accurately!

Loss to follow-up

- Luckily we can deal with this the same way we did our induced censoring
- We just need to make sure relevant covariates are measured

Thanks!

Contact me! l.smith@northeastern.edu; @louisahsmith; louisahsmith.com

I'll be looking for a PhD student at Northeastern (Boston) and a postdoc (flexible, Portland ME, Boston) soon!

I'm grateful for helpful conversations with Sonia Hernández-Díaz and Tyler VanderWeele about some of this content