Advanced Epidemiologic Methods

EPID 722
Spring 2021
UNC – Chapel Hill
jessedwards@unc.edu

Instructors

Jess Edwards

Steve Cole

TAs

Linnea Olson

Rachael Ross

What are we doing here?

(What is the point of epidemiology?)

One goal of epidemiology is to describe the world as it is.

Where does disease occur? What is the burden of a specific disease or condition?
Who gets sick? Are there health disparities? Where is need greatest?
Are there associations between health status and specific substances, treatments, or behaviors?

Another goal is to describe the world as it could be under some intervention

How might health outcomes change under more rapid treatment?
Would disease be reduced by banning a specific chemical?
Would behavior modification lead to a reduction in incidence?

Why train in advanced epidemiologic methods?

1. INTRODUCTION

The subject-specific data from a longitudinal study consist of a string of numbers. These numbers represent a series of empirical measurements. Calculations are performed on these strings and causal inferences are drawn. For example, an investigator might conclude that the analysis provides strong evidence for "a direct effect of AZT on the survival of AIDS patients controlling for the intermediate variable – therapy with aerosolized pentamidine". The nature of the relationship between the sentence expressing these causal conclusions and the computer calculations performed on the strings of numbers has been obscure. Since the computer algorithms are well-defined mathematical objects, it is useful to provide formal mathematical definitions for the English sentences expressing the investigator's causal inferences, In Robins (1986, 1987), I proposed a

Describing our data is not the same as describing the world

Our data only inform us about the world to the extent that we believe key assumptions.

In this class, we will discuss these assumptions and methods used to relax some of them.

If we have “perfect” data, epidemiology is easy.

What makes data “perfect”?

An incomplete list:

- No measurement error
- No missing data
- Complete follow-up
- Includes a census of the target population
- No confounding
- etc

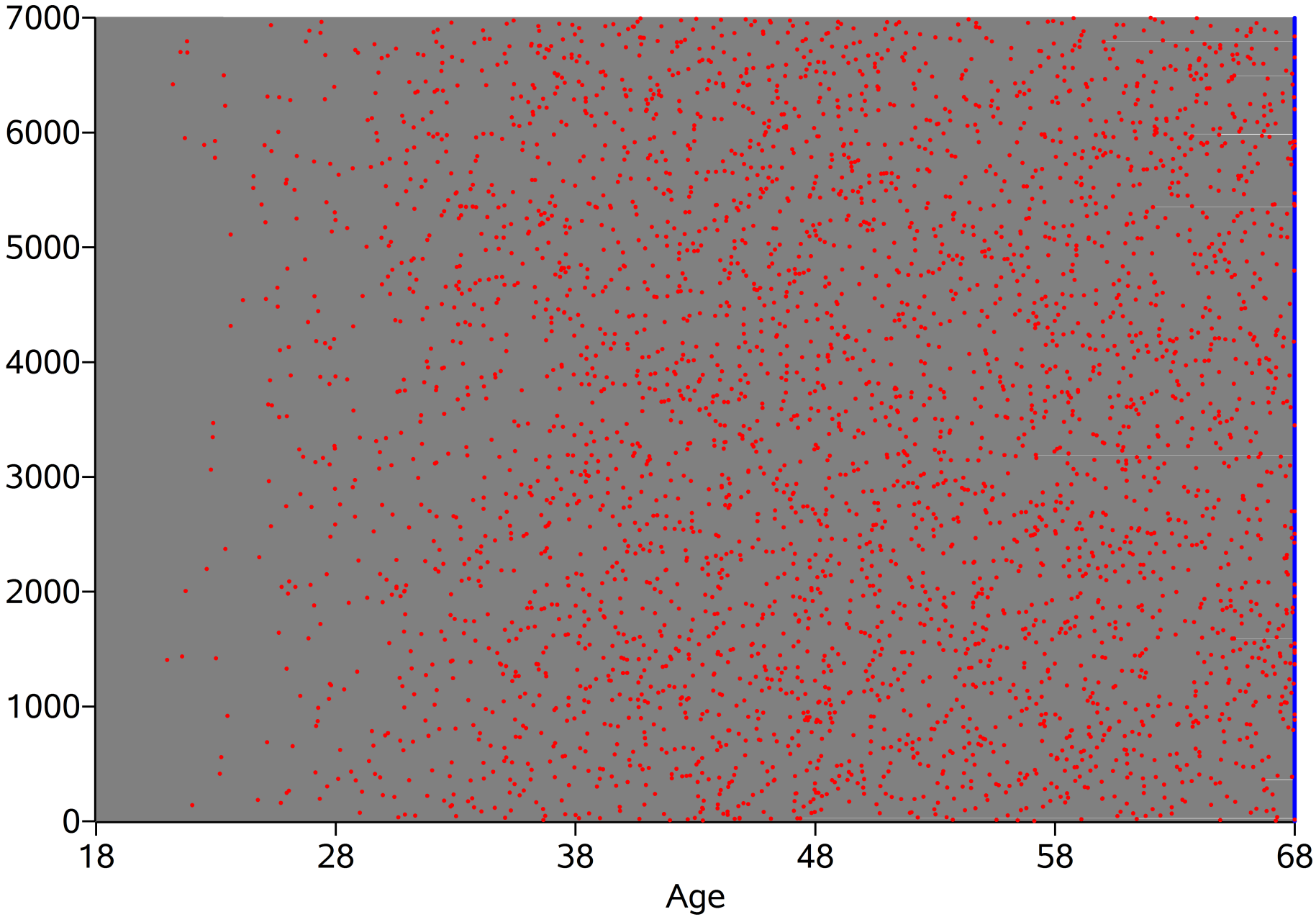
But we can’t wait for perfect data

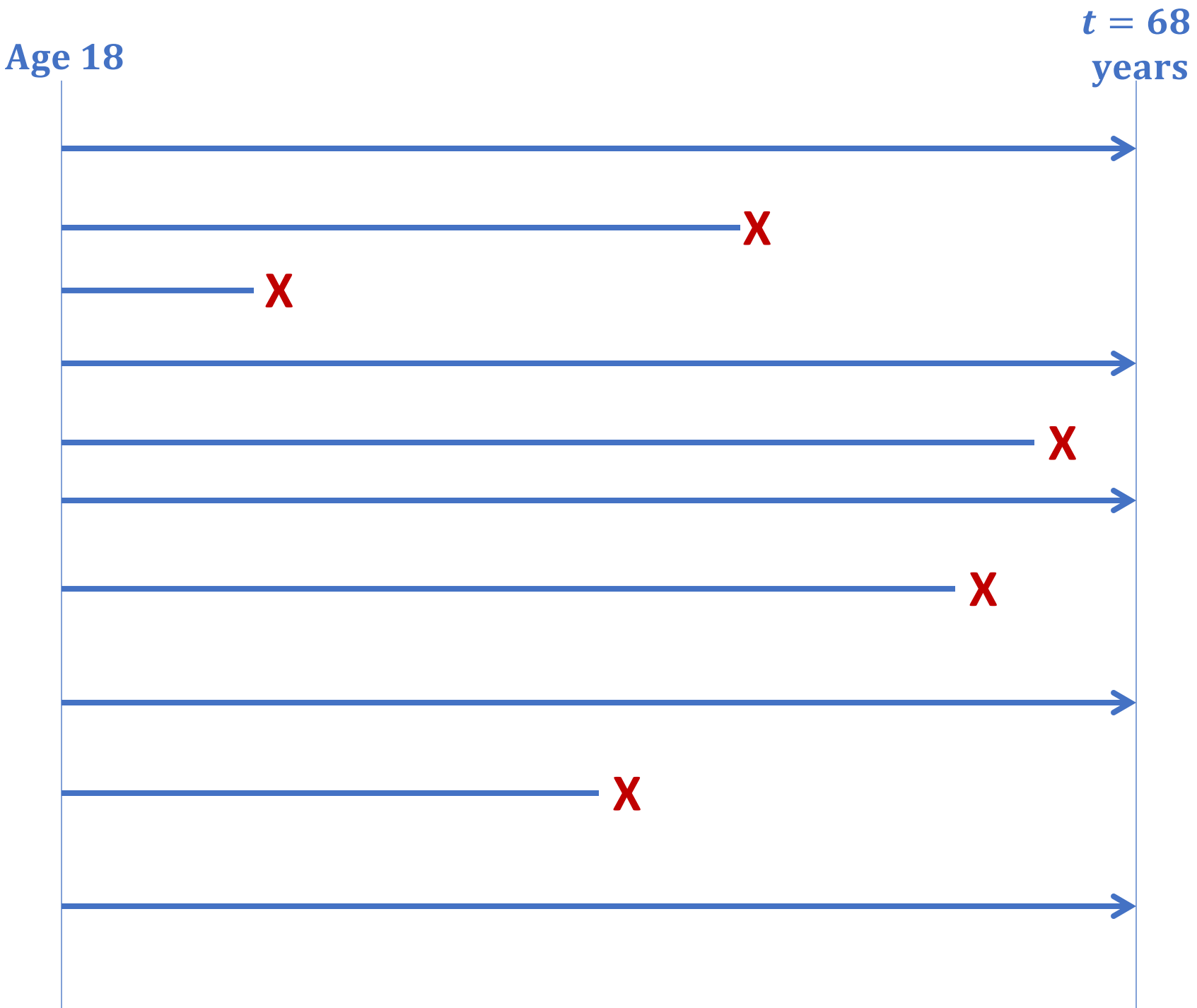
Results from epidemiologic studies power decision making.

Decisions don’t wait. (Not to decide is to decide).

So, the ability to learn from our (almost always) imperfect data is critical.

An example.

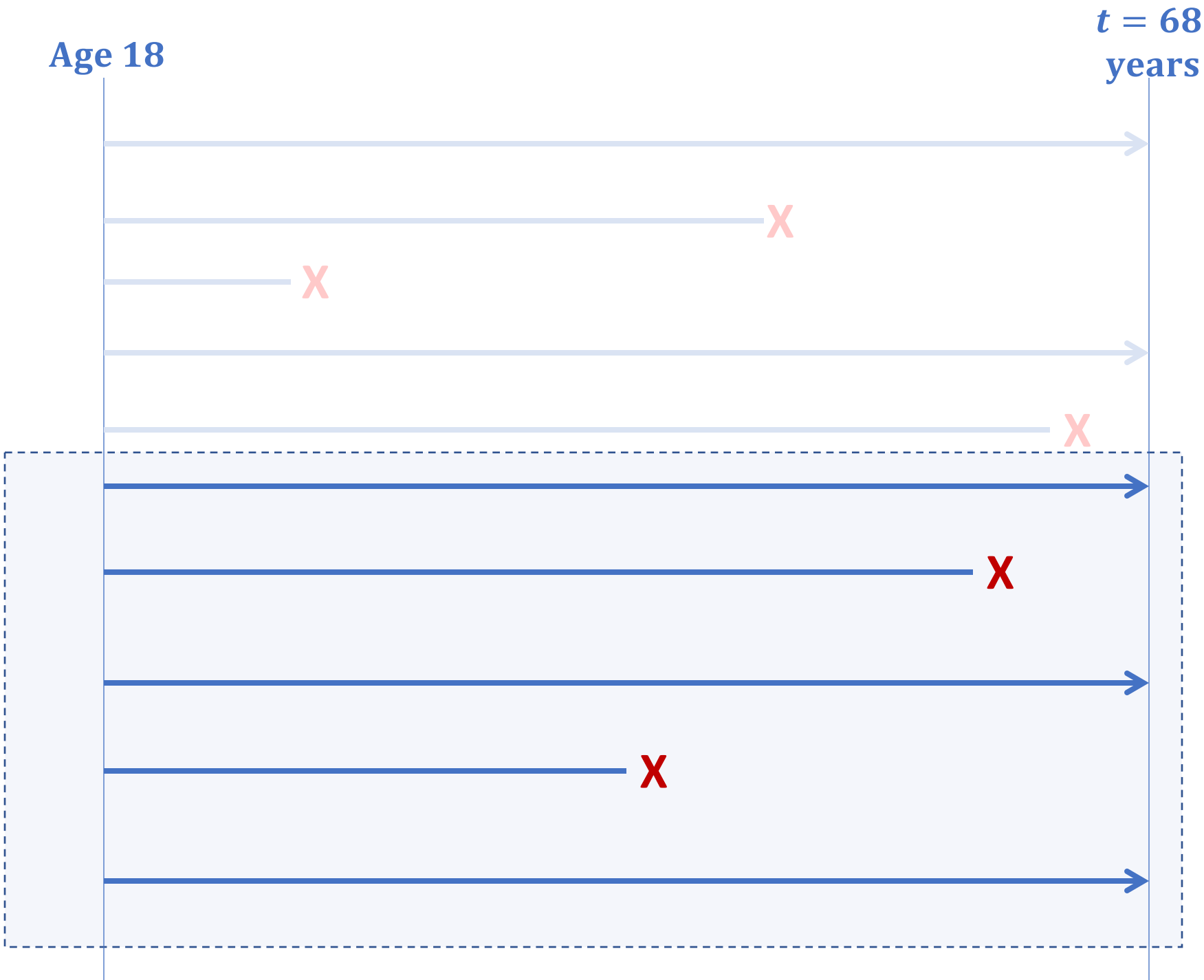


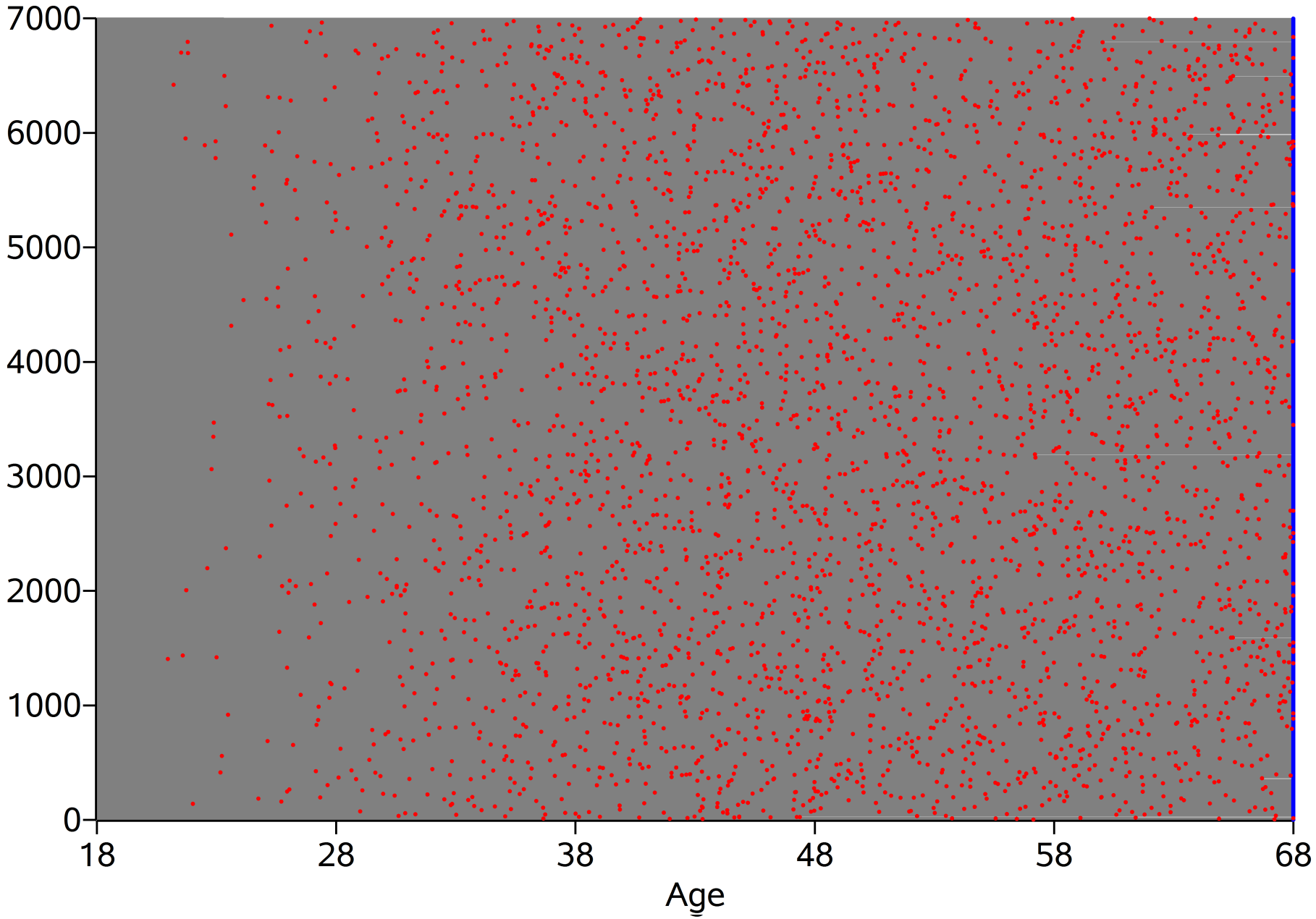

Age 18
$t = 68$ years
X
X
X
X
X

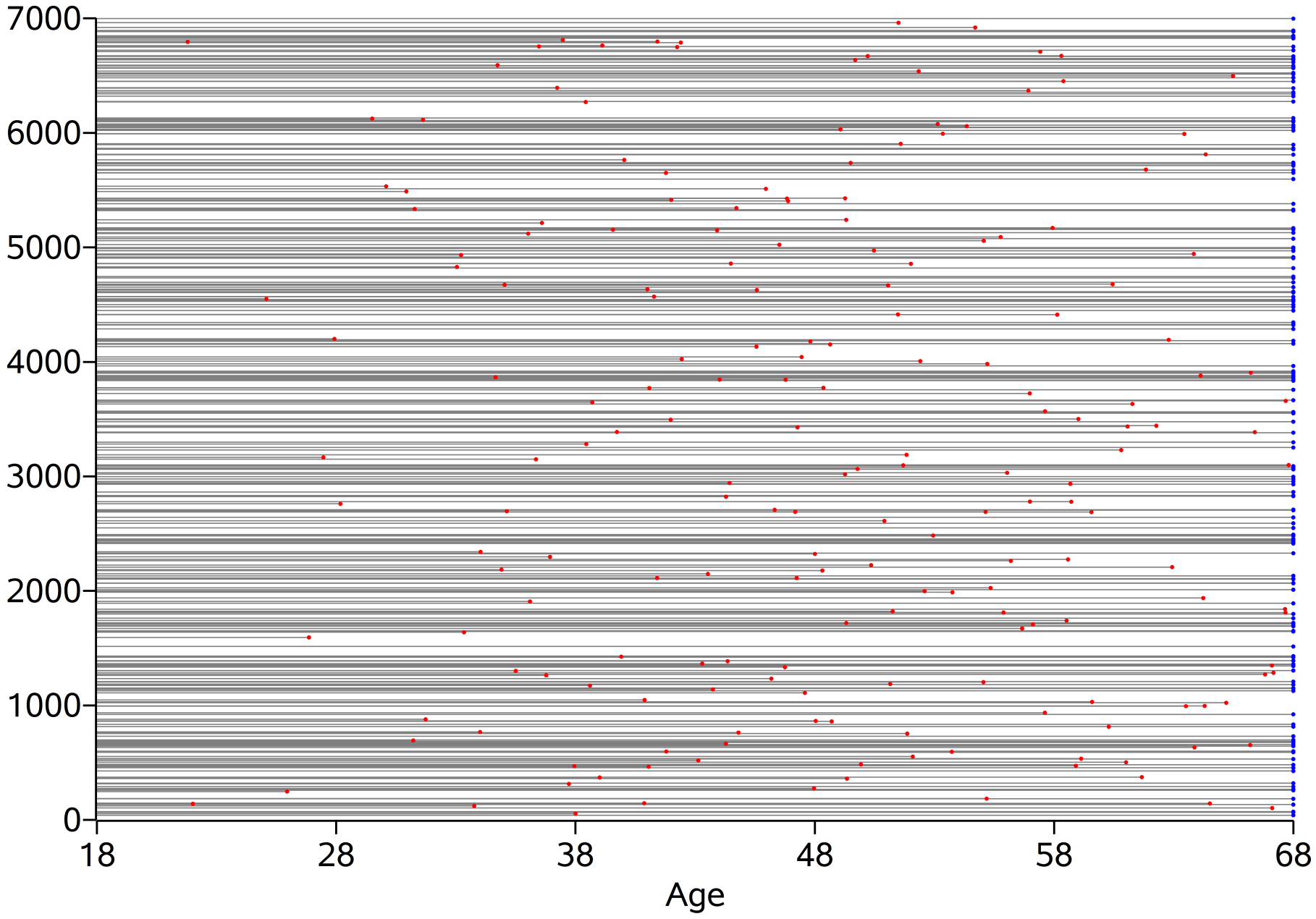
But what if we don't observe $\boldsymbol{i = 1, \ldots 7000}$?

We could fail to observe some subjects due to

- Sampling from the population
- Refusal to participate
- Selective recruitment into the study
- Other reasons?


Age 18
$t = 68$ years


7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age

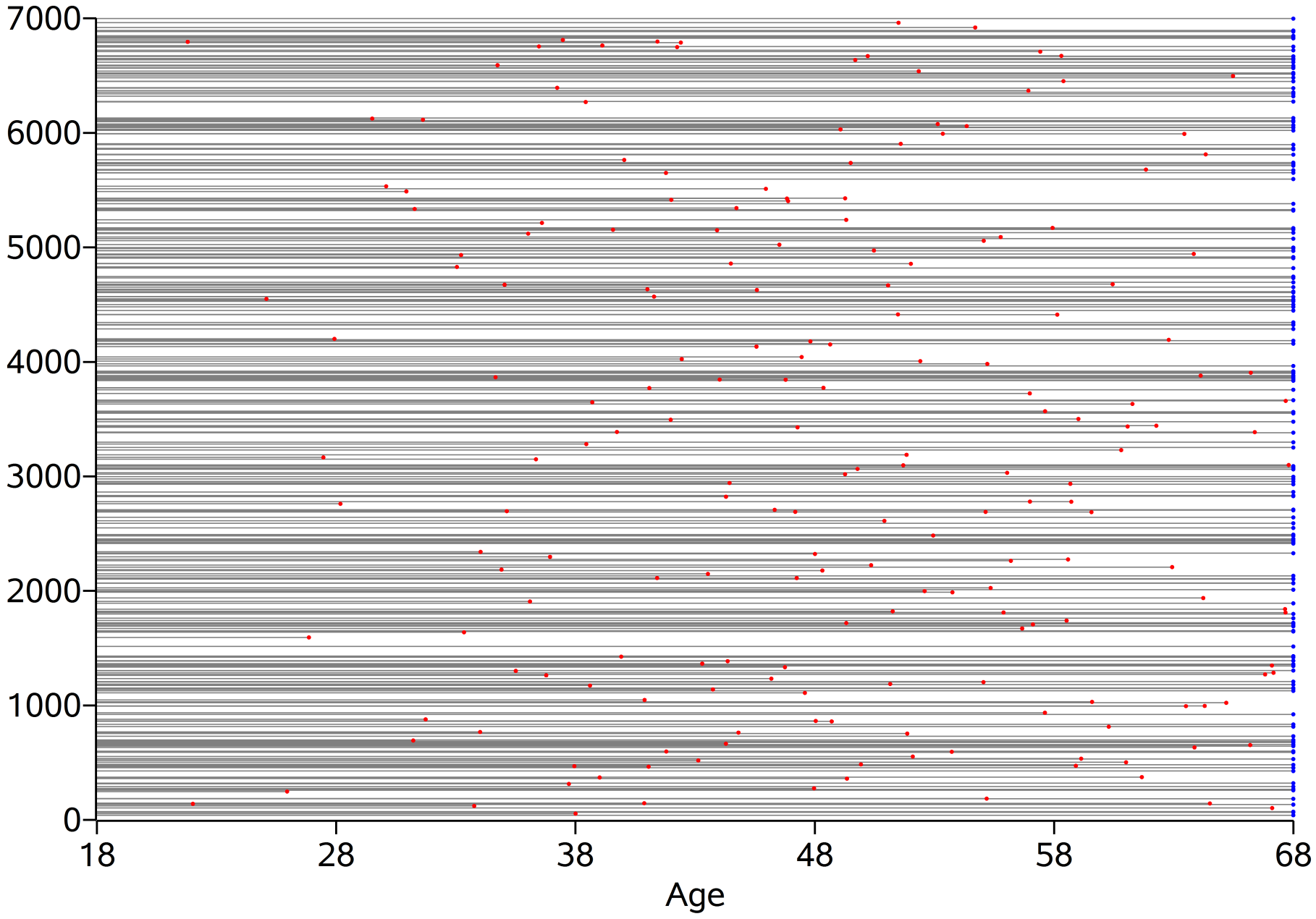

7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age

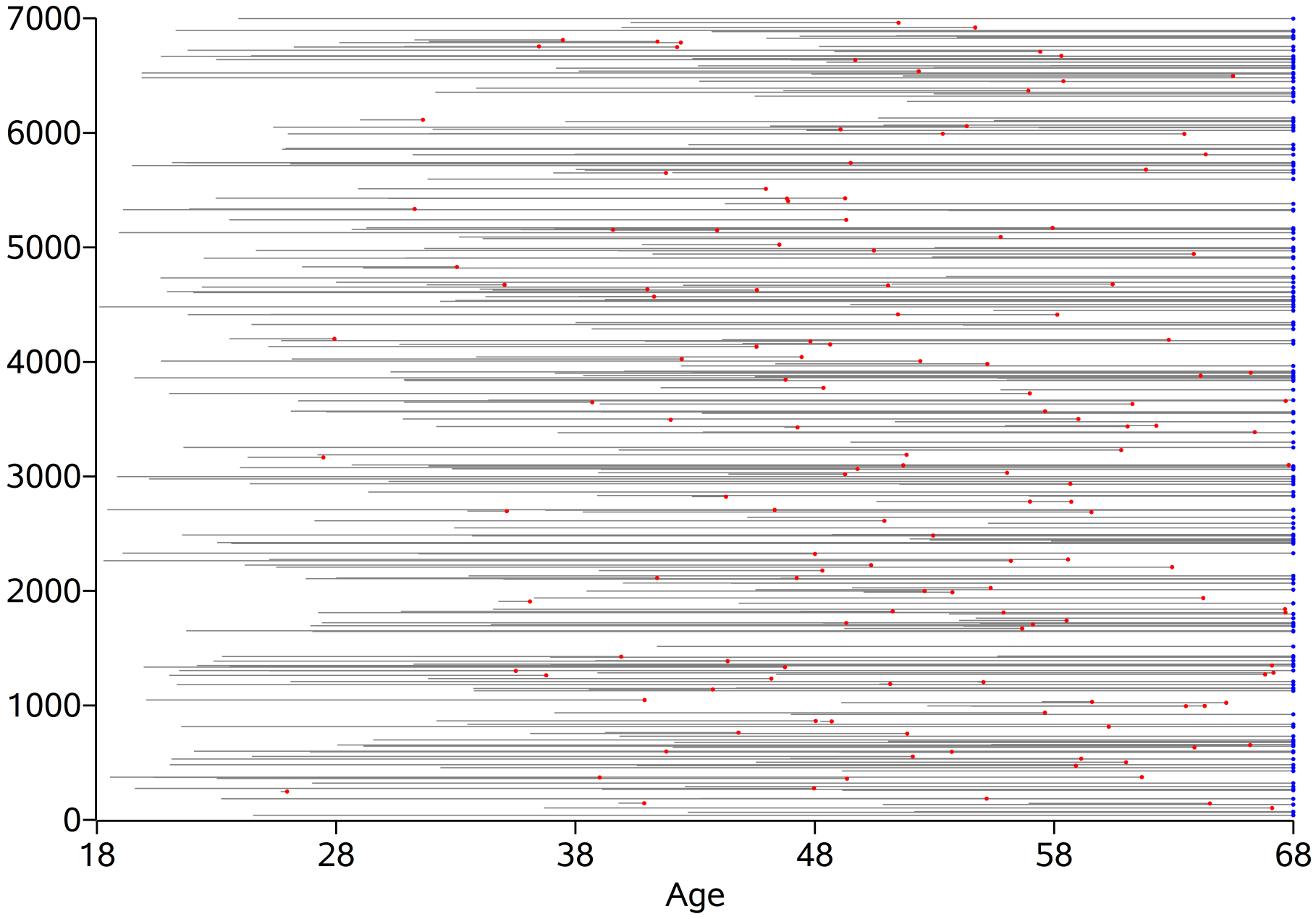
What if some participants are not recruited into the study until after freshman year?

We may enroll subjects into the study late due to

- Delayed identification of eligible subjects
- Migration
- Others?


Age 18
$t = 68$ years
X
X
X
X
X


7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age

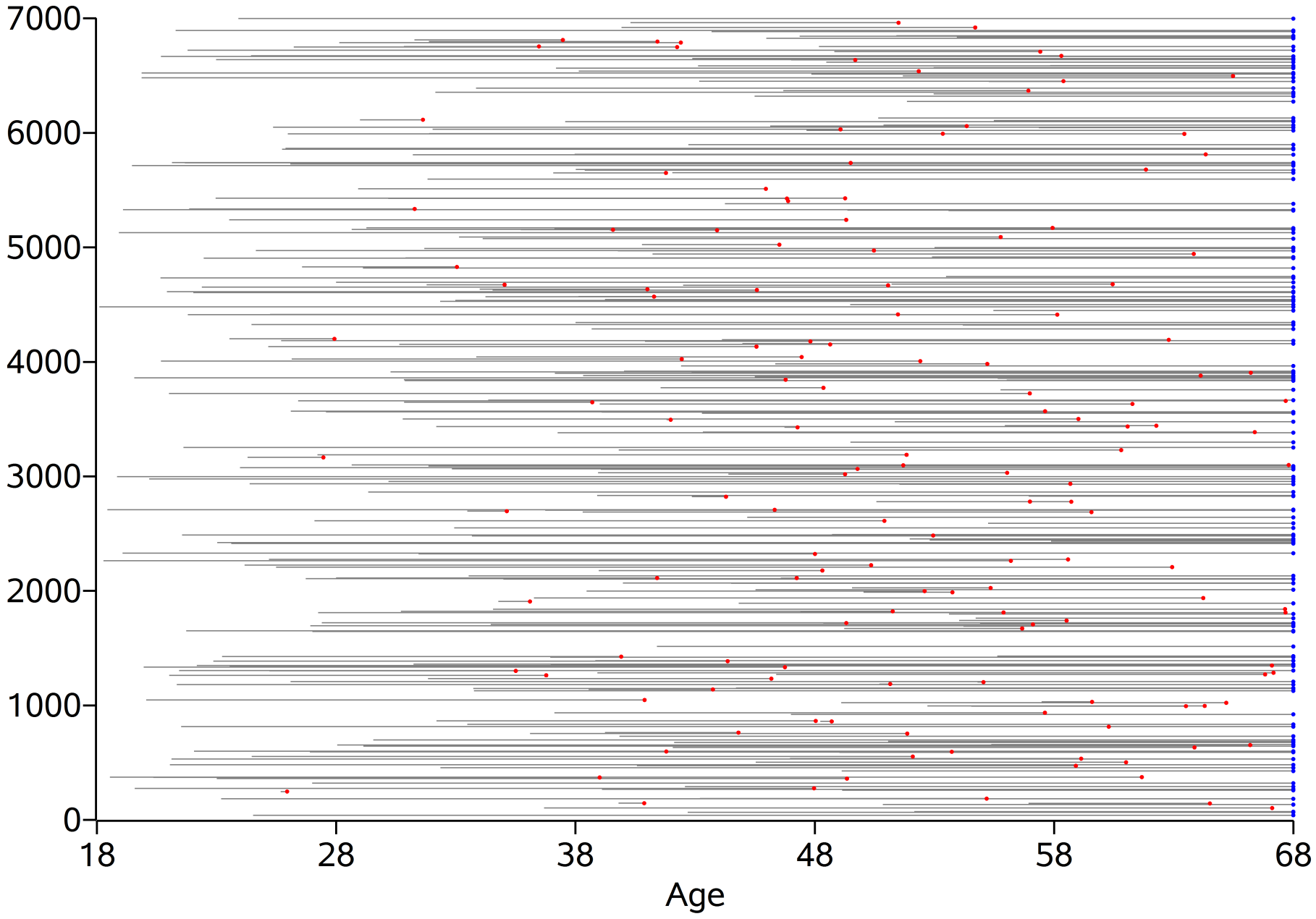

7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age

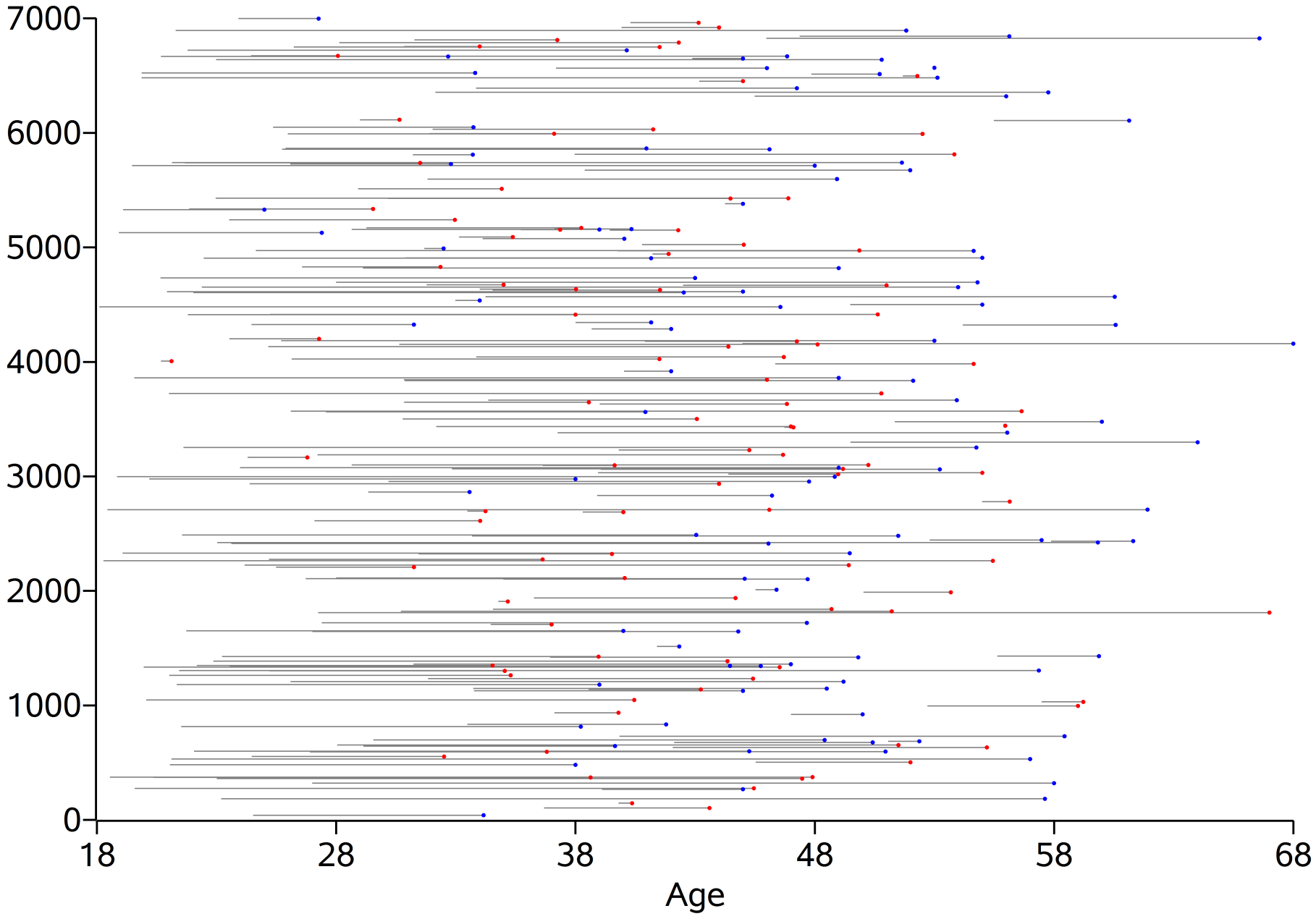
Now, what if we don’t observe the all of the selected participants until age 68?

We could fail to observe some subjects for the full study period due to

- Loss to follow-up
- Administrative end of the study in 2017
- Others?


Birth
$t = 68$ years
A
0
1
X
0
X
1
0
X
1
1
X
0
0
X
1

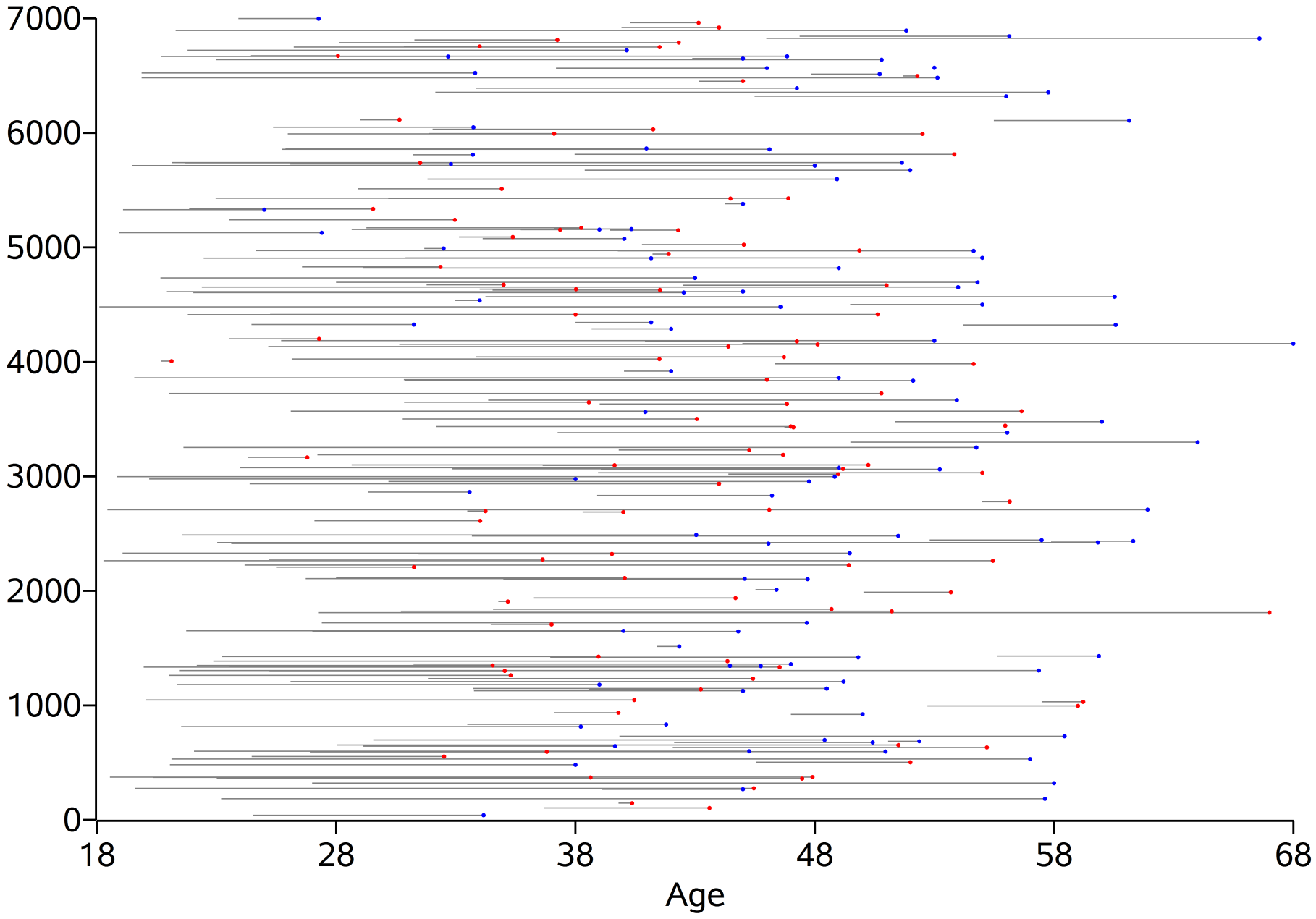

7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age

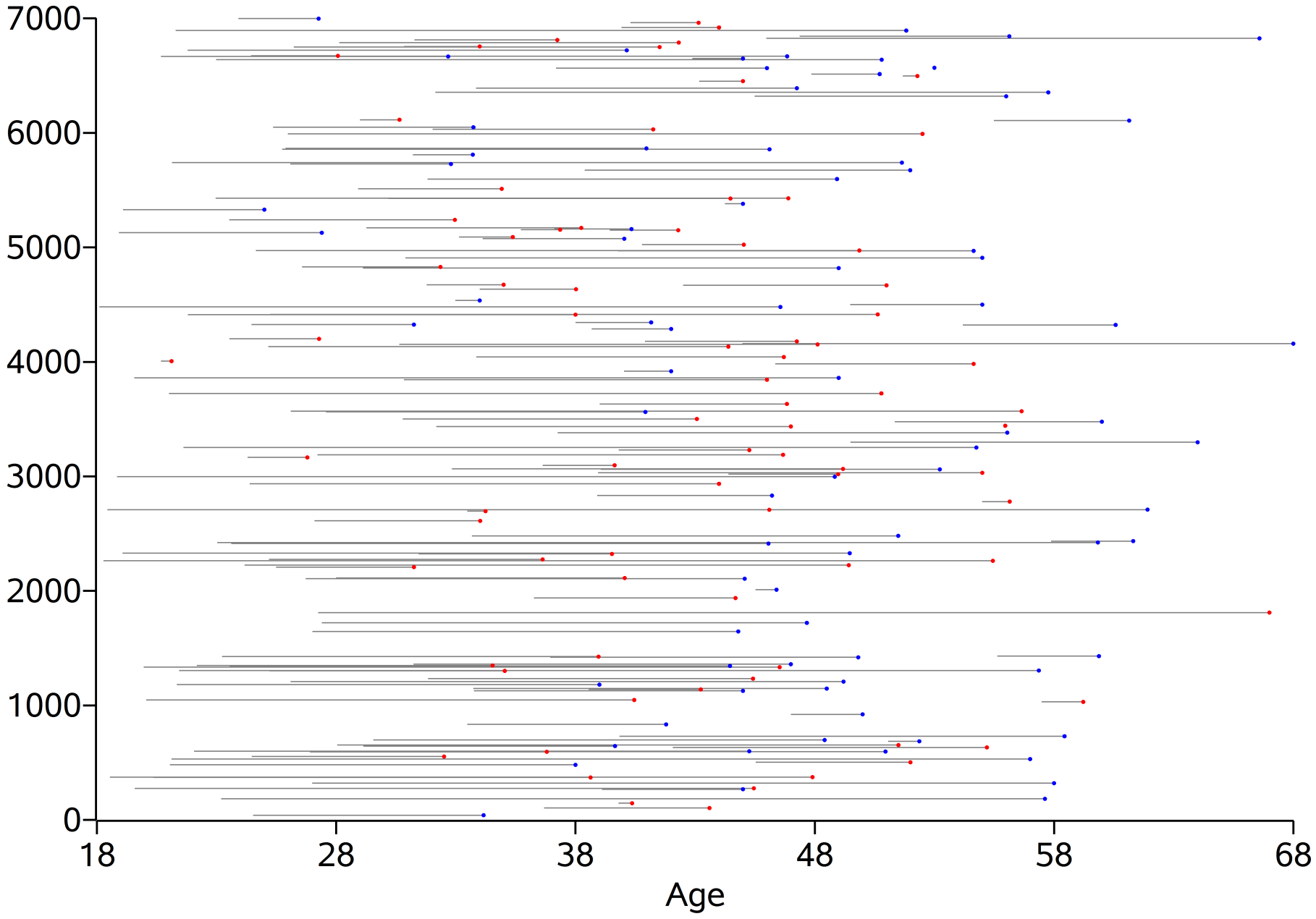

7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age

And subjects in the study may have missing values for exposure or covariates

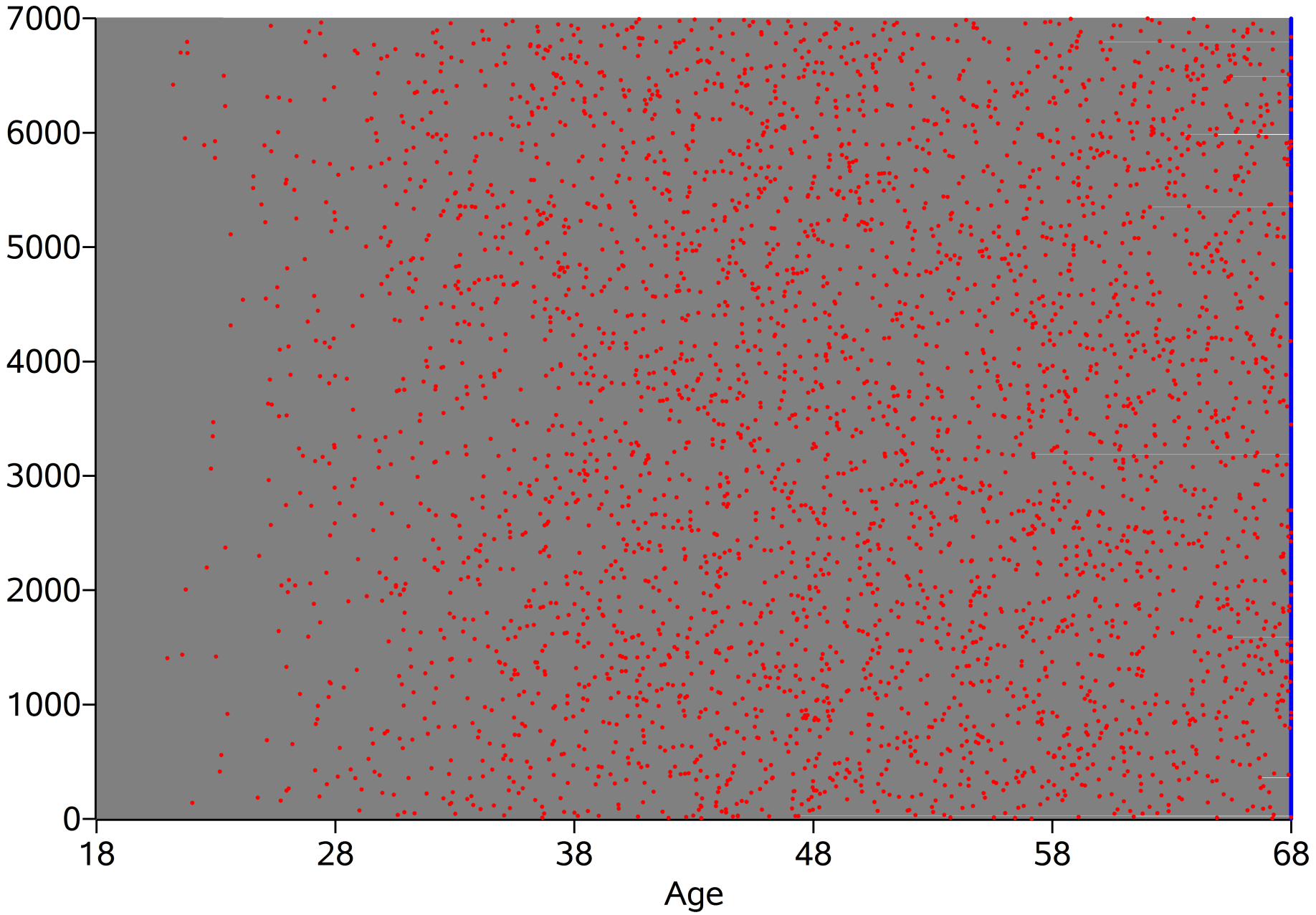
Subjects may have missing values for exposure or covariates due to

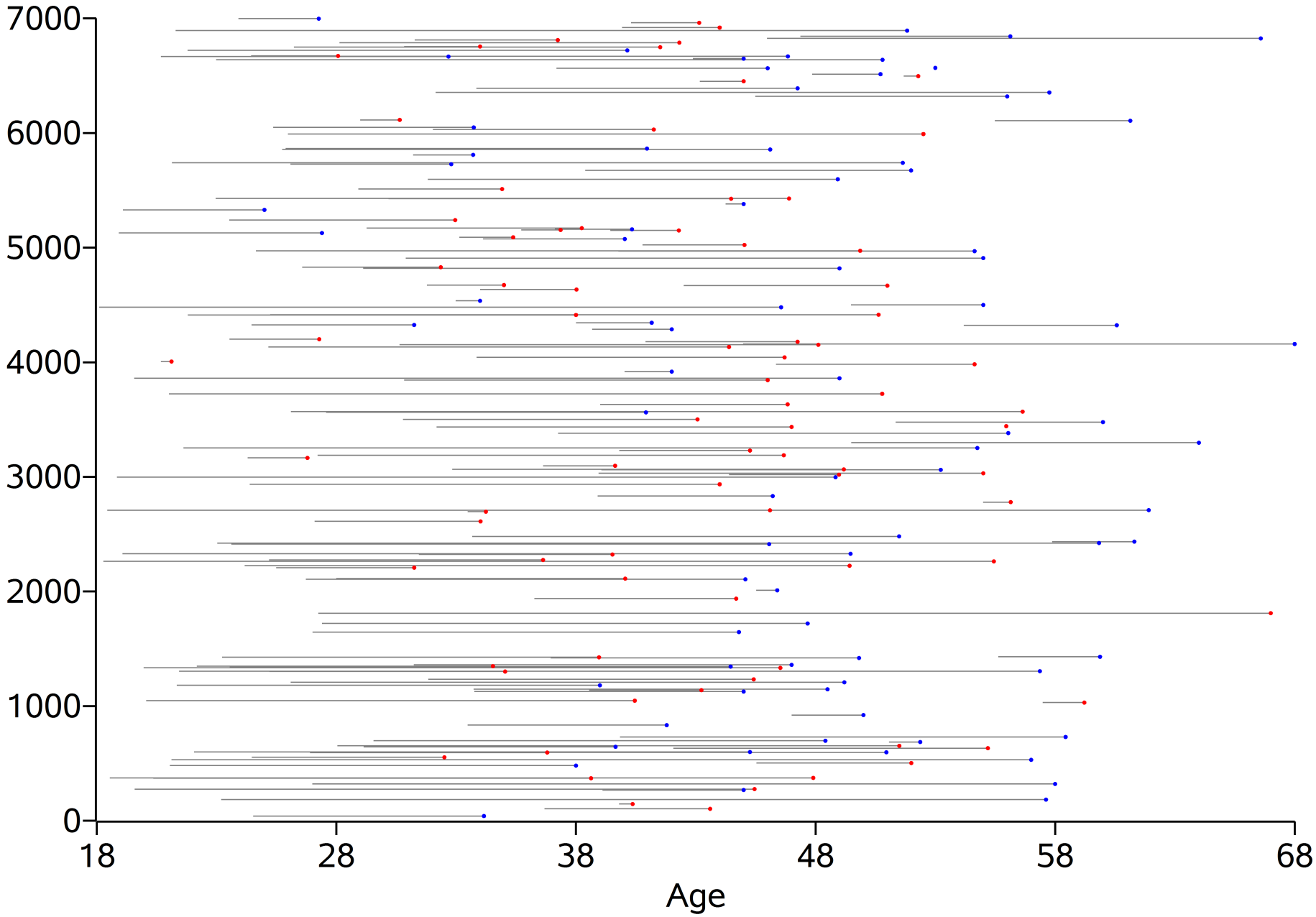
- Errors in data entry
- Subject does not know requested information (e.g., vaccinations)
- Subject refuses to provide information (e.g., drug use)
- Assay failures
- Assay detection limits


7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age


7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age

Just to reiterate the amount of missing information…..


7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age


7000
6000
5000
4000
3000
2000
1000
0
18
28
38
48
58
68
Age

Course Roadmap

A course in 3 acts

Asking and framing epidemiologic questions

Describing the world as it is: Tools for survival analysis

Describing the world under interventions: Tools for causal inference

Asking and framing questions

What are the components of a good question in epidemiology?

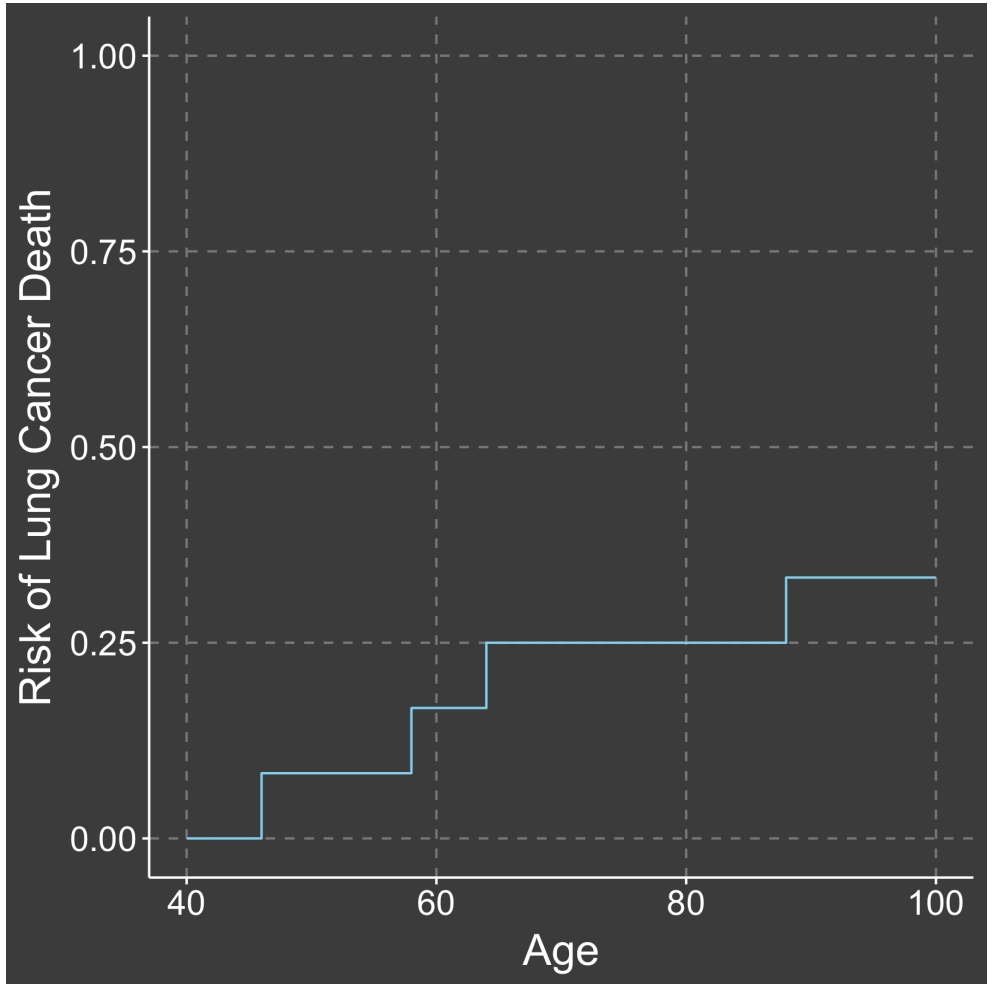
Describing the world as it is

We will focus on estimating risk.

(Why risk?)

We will start with the “single sample” scenario.

(Why 1 sample?)

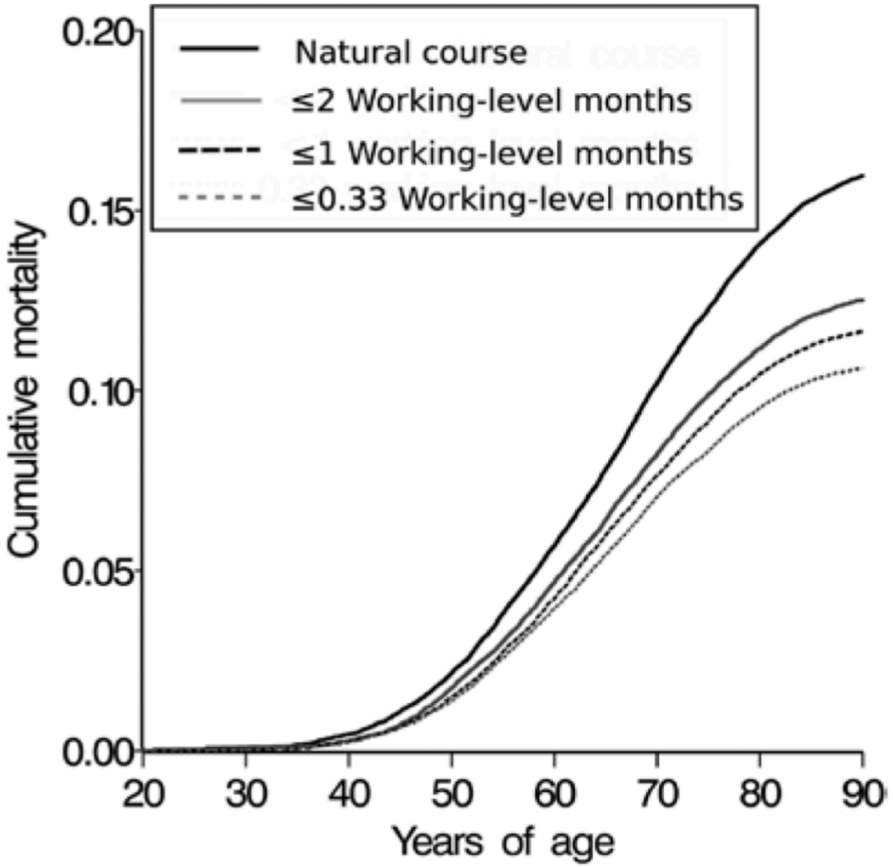


Describing the world under interventions

Tools for causal inference

Potential outcomes framework

Again, focus on risk

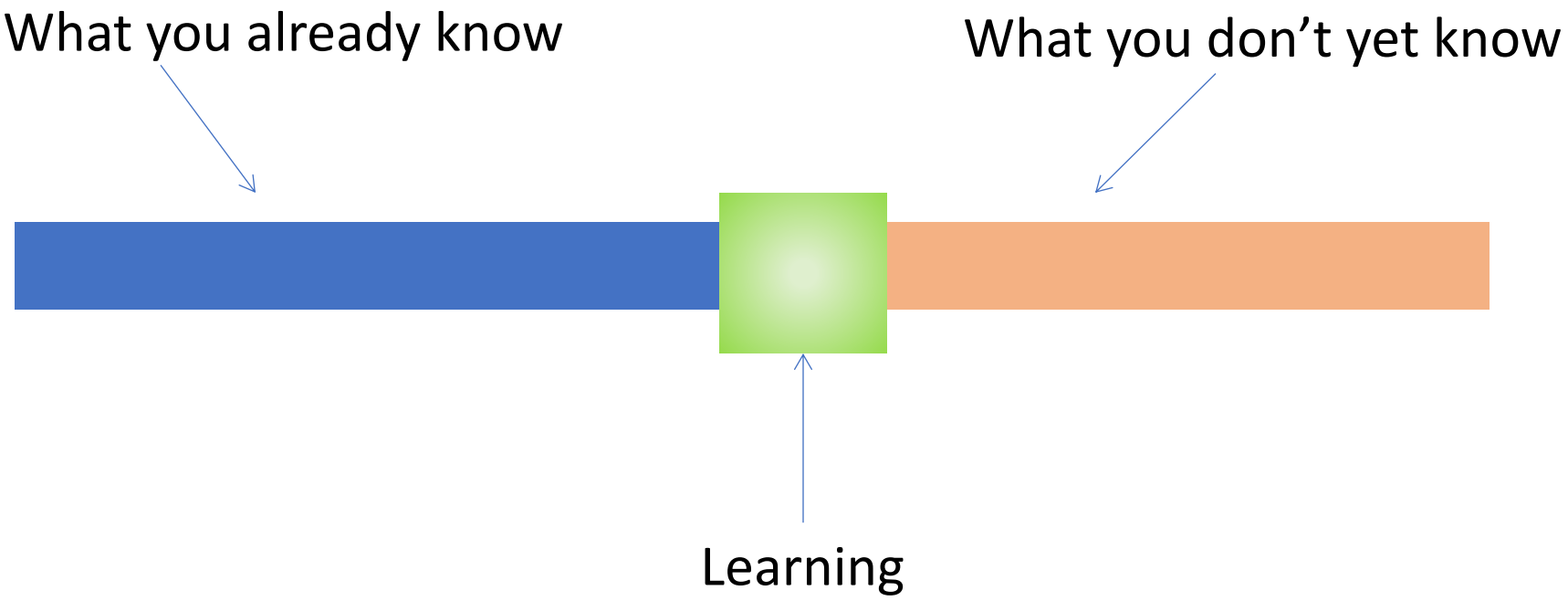


A note on learning methods

Narrow focus on a single parameter (risk) and a specific set of tools.

But learning about risk and this specific set of tools is not the point of the course.

The point is to develop the skills you need to learn *any* new method.


What you already know
What you don’t yet know
Learning

Course Structure

Course website

Feedback

We welcome your feedback throughout the course.

By email to the instructors or the TAs.

Anonymously through the form on the website.

Final course evaluation (through UNC evaluation system).

Closing thoughts

Advanced Epidemiologic Methods

EPID 722
Spring 2021
UNC – Chapel Hill
jessedwards@unc.edu

L1: Questions and time in epidemiologic studies

EPID 722
Spring 2021
UNC – Chapel Hill
jessedwards@unc.edu

Roadmap

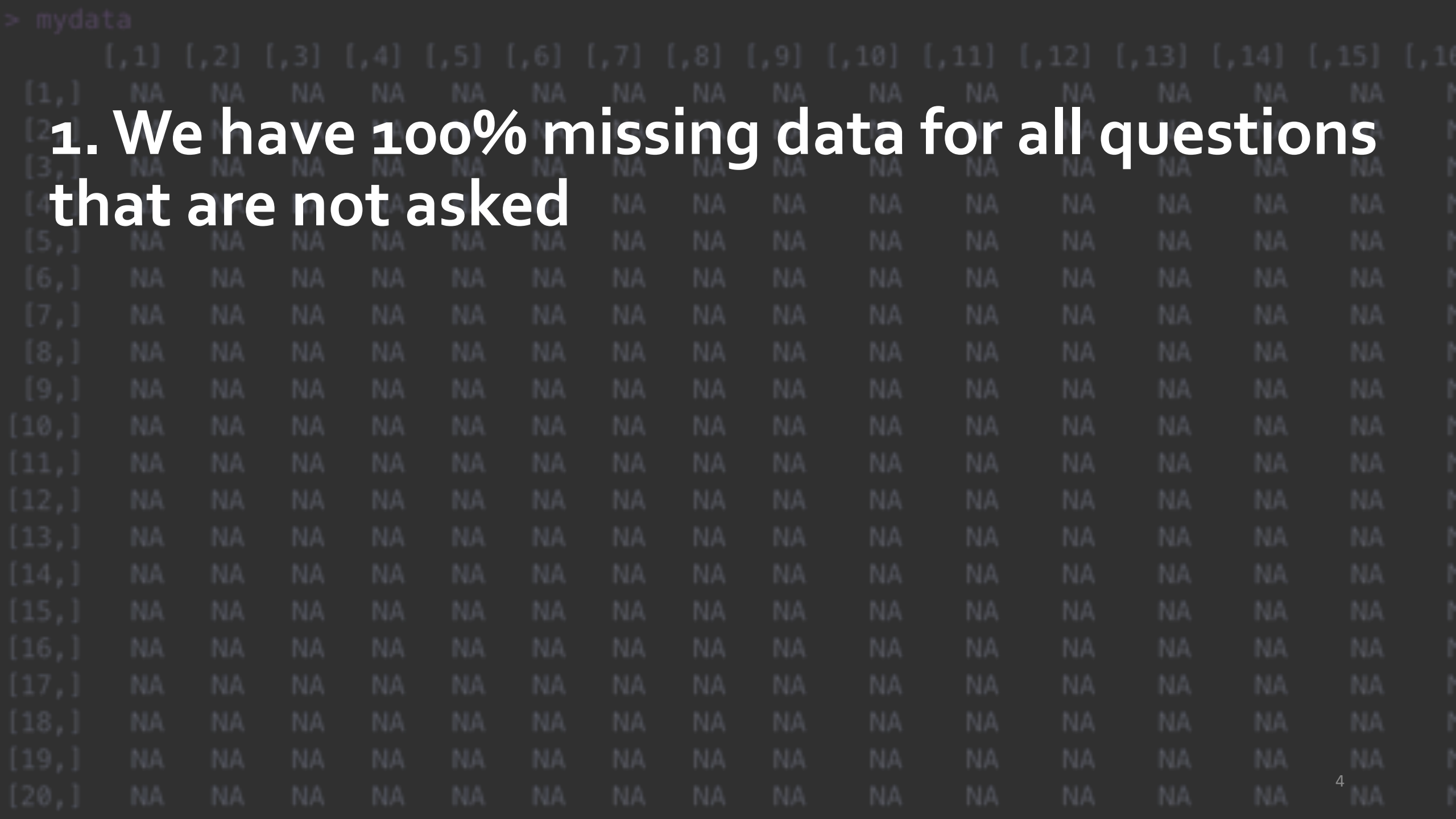
Chapter 1: 10 considerations when asking questions

Chapter 2: Risk

Chapter 3: Why learn methods for survival analysis?

Chapter 4: Line diagrams

Chapter 1: Asking questions



1. We have 100% missing data for all questions that are not asked

2. There are many routes to asking a question, and many sources of influence

How do we decide what to ask?

Who decides which questions we answer?

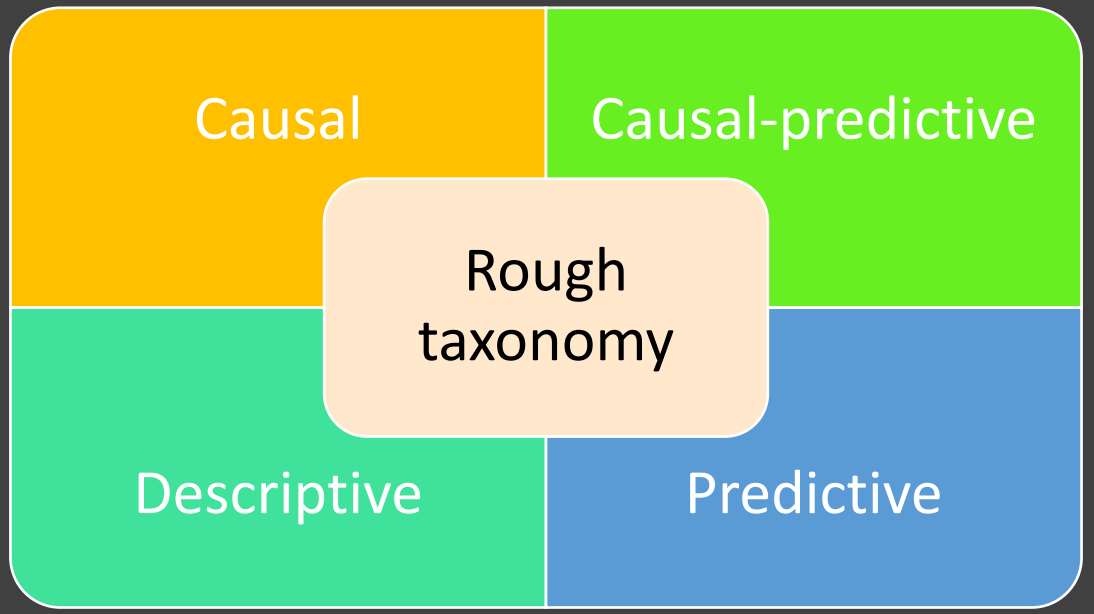
3. “Asking” a question is a 2-step process

What problem needs to be addressed?

[Question]

Frame the question

4. Different types of questions can inform decisions



4. Different types of questions can inform decisions

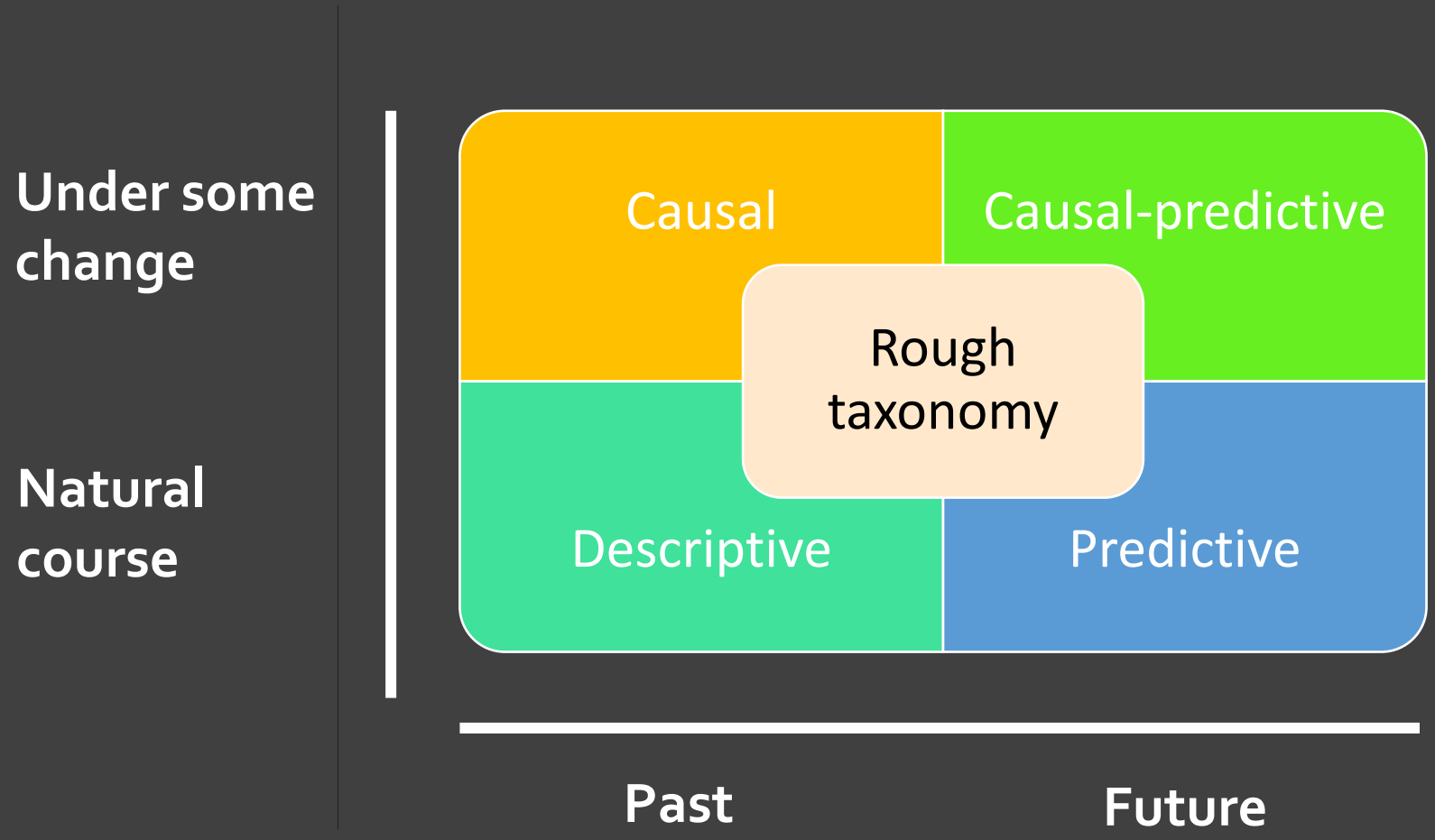
Descriptive: What happened?

Predictive: What will happen?

Causal: What would have happened had we done X?

Causal-predictive: What will happen if we do X in the future?

4. Different types of questions can inform decisions



5. Considering the use of the results is important to framing the question

Who will be acting on the results and Whom does the decision affect?

What actions are feasible, now or in the future?

What types of results would be compelling to decision makers?

6. To maximize utility of the answers, questions should have a set of components

1. Target population

2. Time period of interest (origin and timescale)

3. Outcome(s), including WHEN outcomes should be measured

4. If there are a set of **actions** that are being assessed, the set of candidate actions, including WHEN they would occur

5. If there is a **comparison** being drawn **between groups**, the set of groups compared, including WHEN groups are defined.

7. Consider the idealized cohort study to flesh out your question

"Target trials" and beyond

A NEW APPROACH TO CAUSAL INFERENCE IN MORTALITY STUDIES WITH A SUSTAINED EXPOSURE PERIOD—APPLICATION TO CONTROL OF THE HEALTHY WORKER SURVIVOR EFFECT

JAMES ROBINS

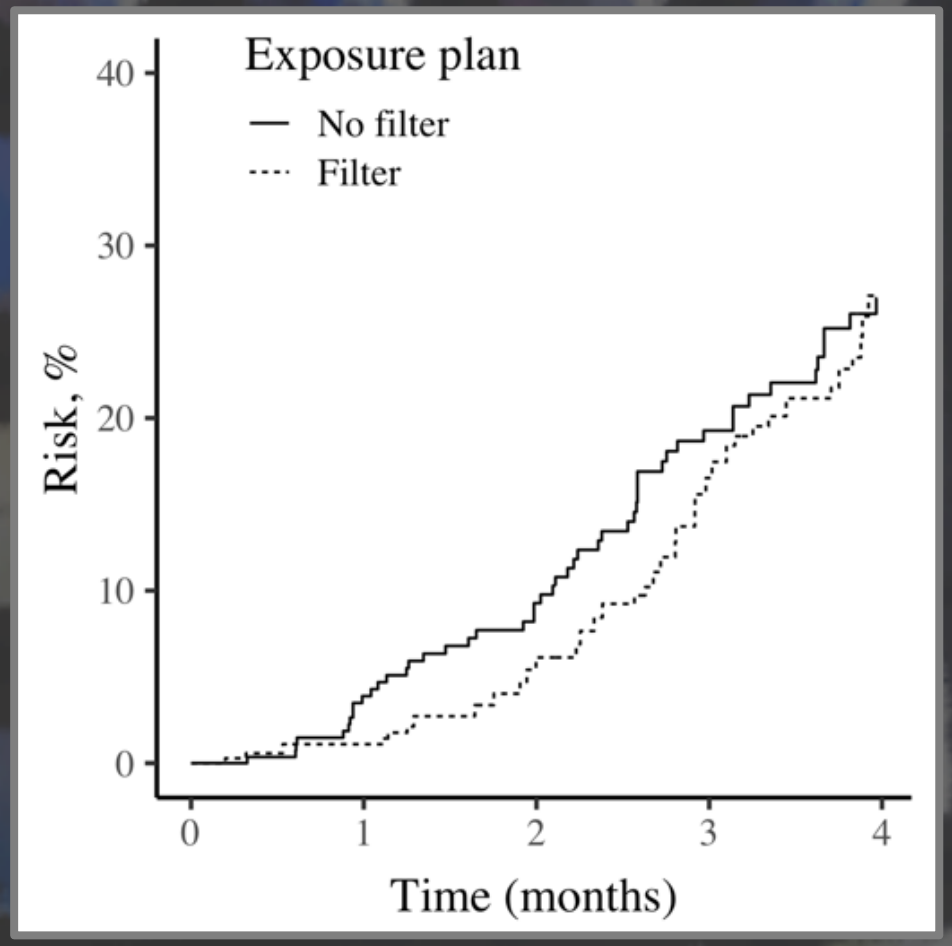
Harvard School of Public Health
665 Huntington Avenue
Boston, MA 02115

The answer to this question often not only is helpful in planning a retrospective study but also may be of assistance in determining

2. OBSERVATIONAL STUDIES AS RANDOMIZED TRIALS MISSING DATA ON TREATMENT PROTOCOL

8. Frame the question without considering existing data or the logistics of data collection

9. Consider absolute measures in addition to contrasts



10. Always ask "compared to what"

WHAT IS A REALISTIC ALTERNATIVE TO ANY ACTION CONSIDERED?

ARE THERE COMPETING EVENTS OR OTHER RISKS/BENEFITS TO BE WEIGHED

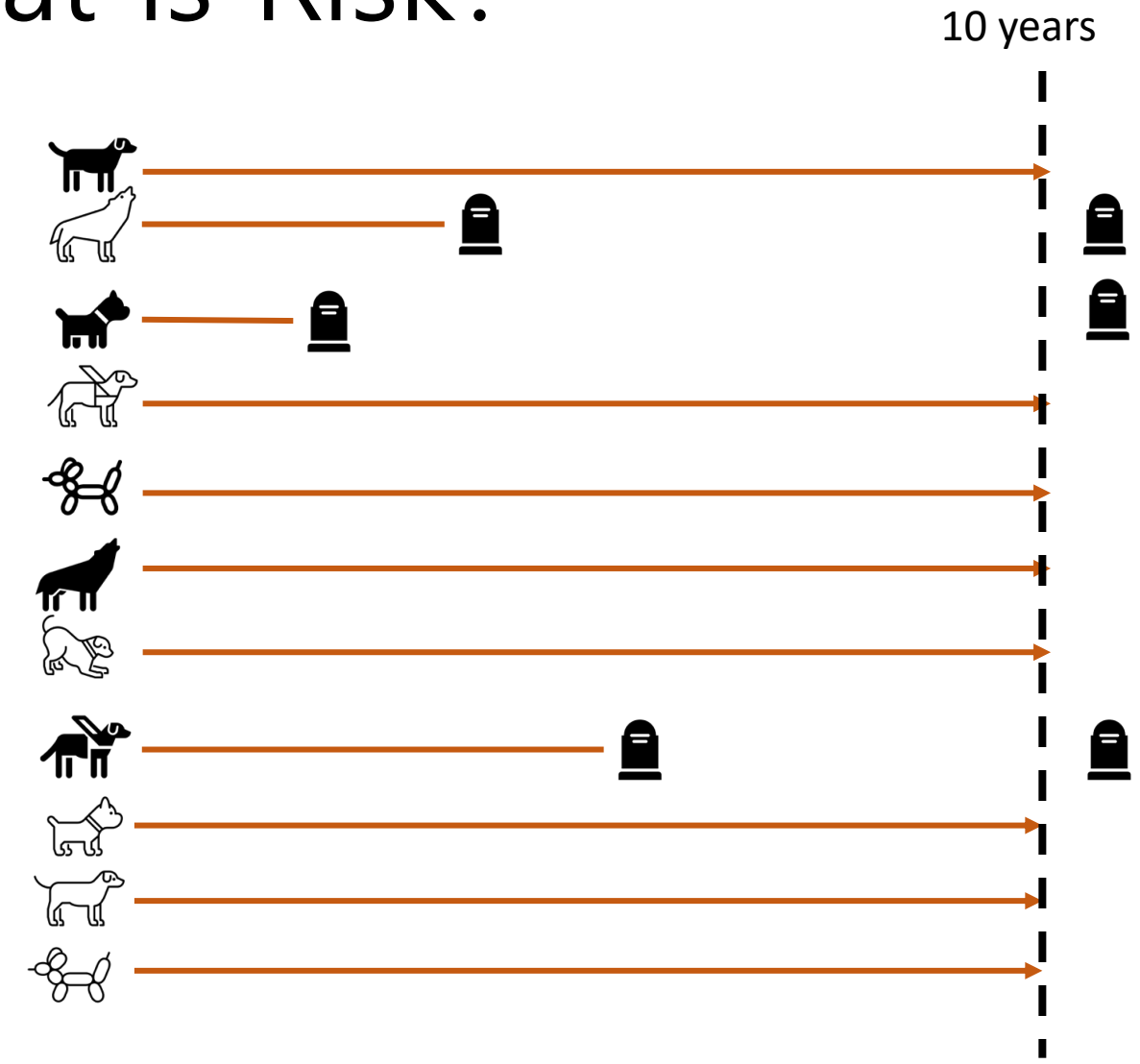
IMPORTANT FOR METHODS, AS WELL AS SUBSTANTIVE QUESTIONS

Chapter 2: Recipe for risk in closed cohorts

What is Risk?

In earlier courses, we defined risk as the proportion of units who experienced the outcome within a defined time period.

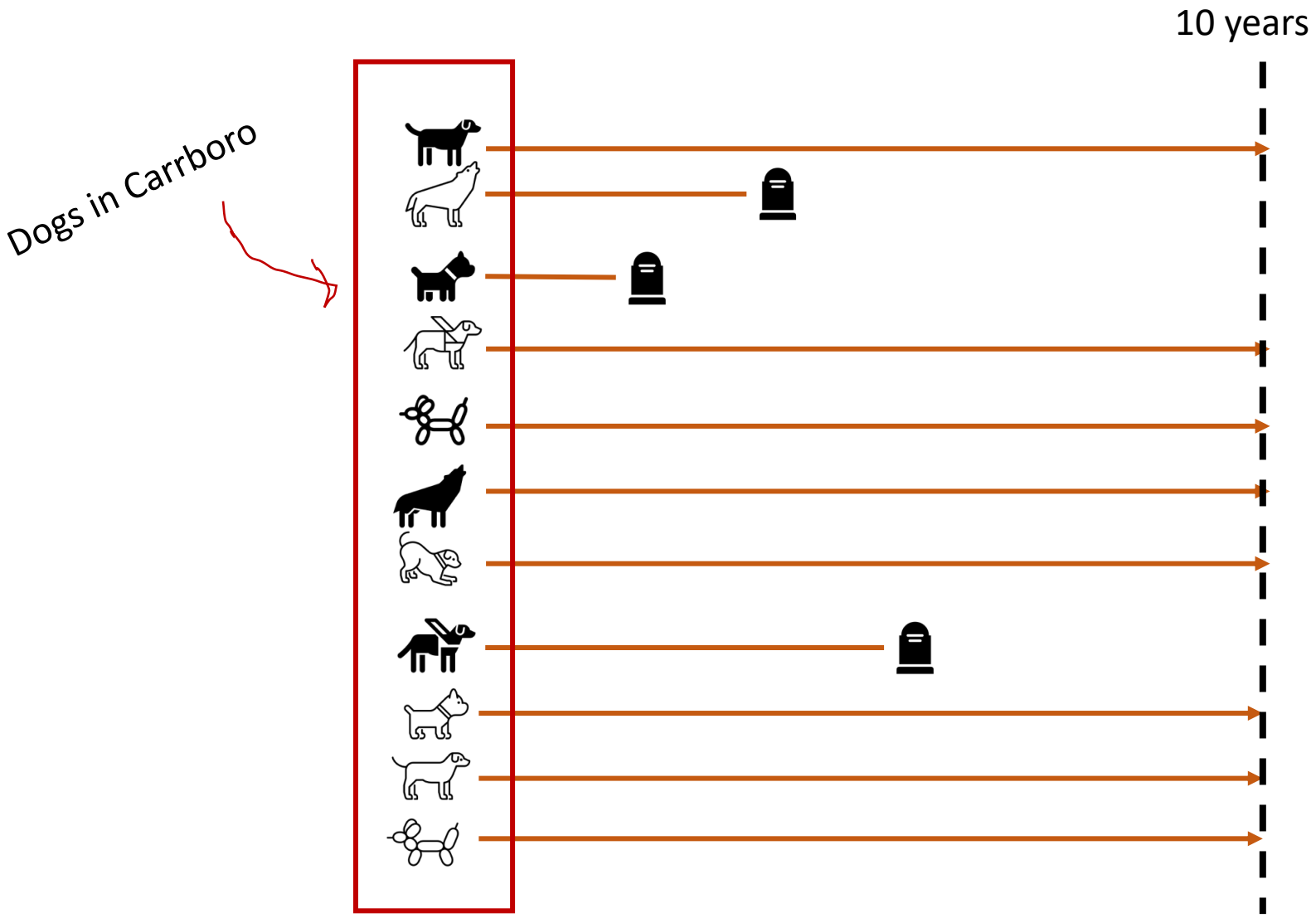
What is Risk?



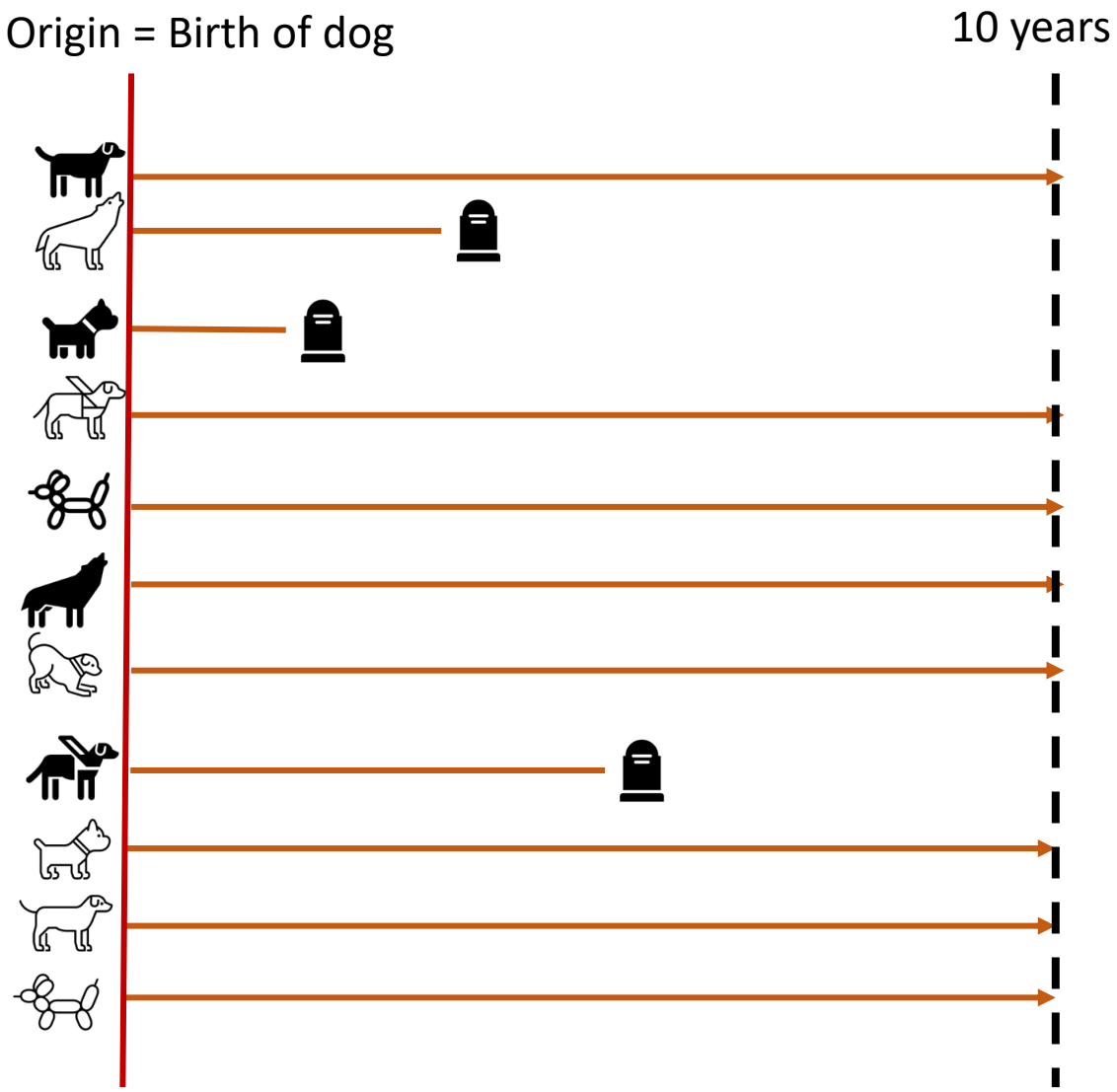
$$\frac{3 \text{ deaths}}{11 \text{ dogs}}$$

10-year risk = 27%

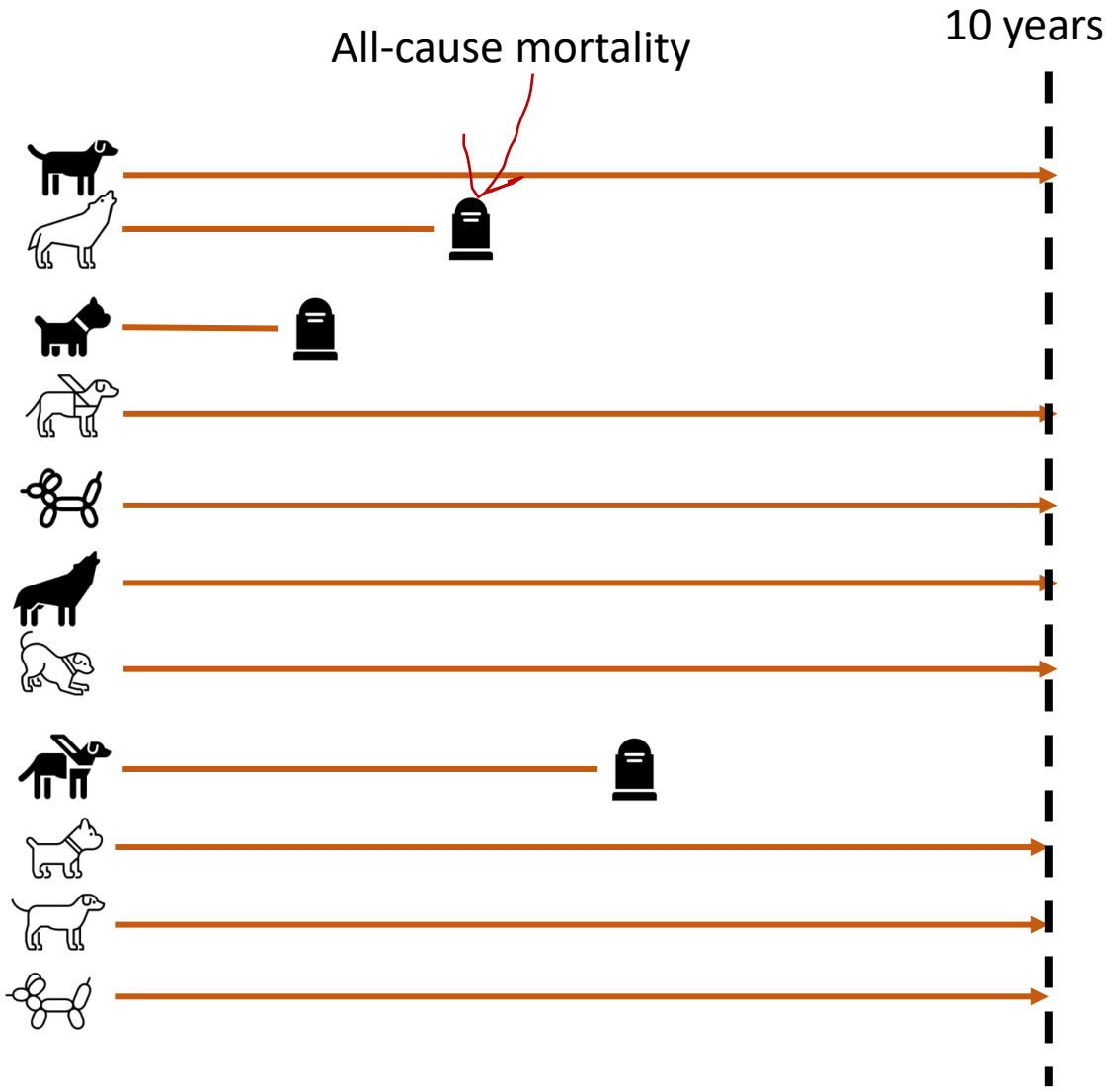
Refining the question: Target population



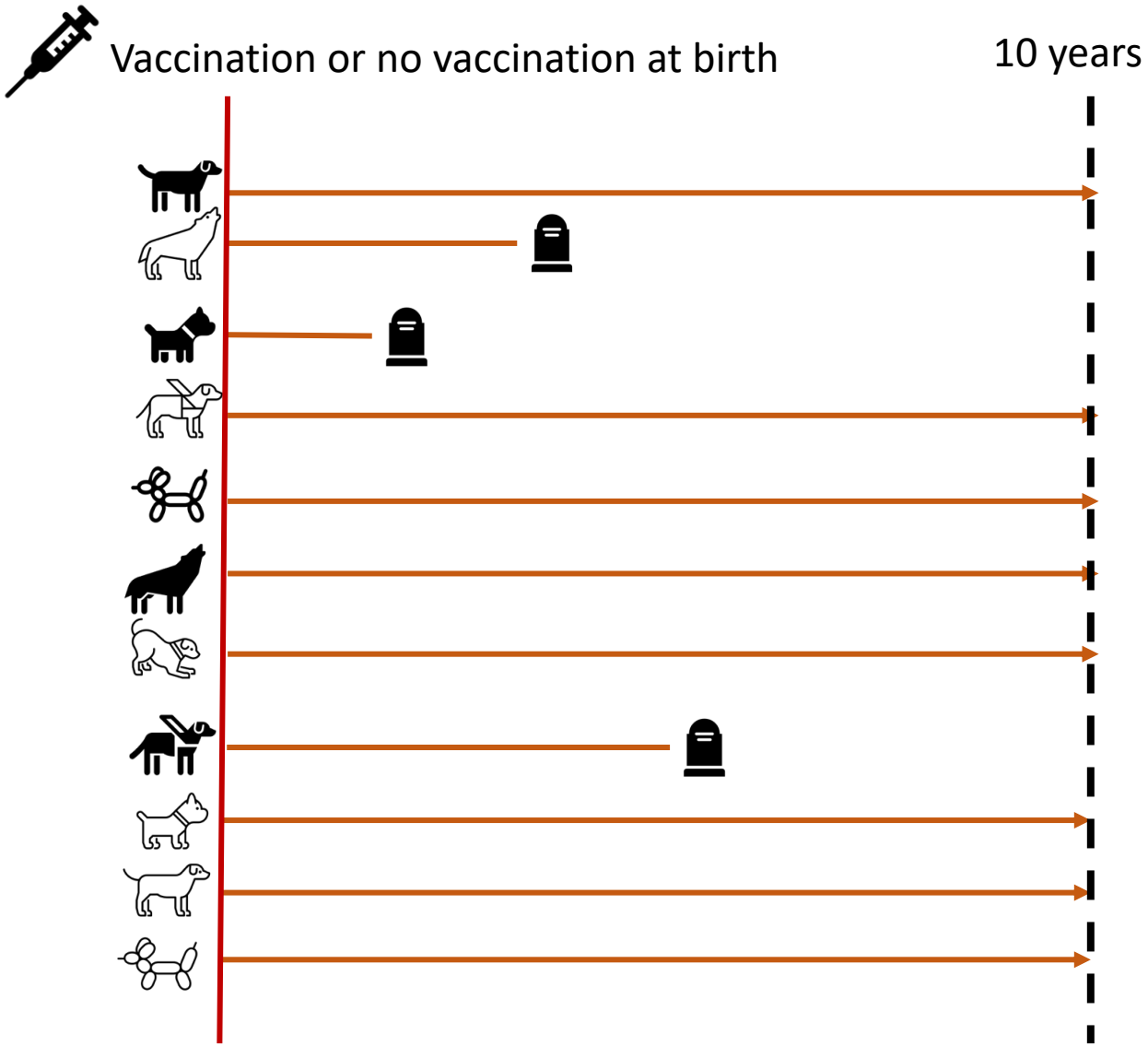
Refining the question: Time period



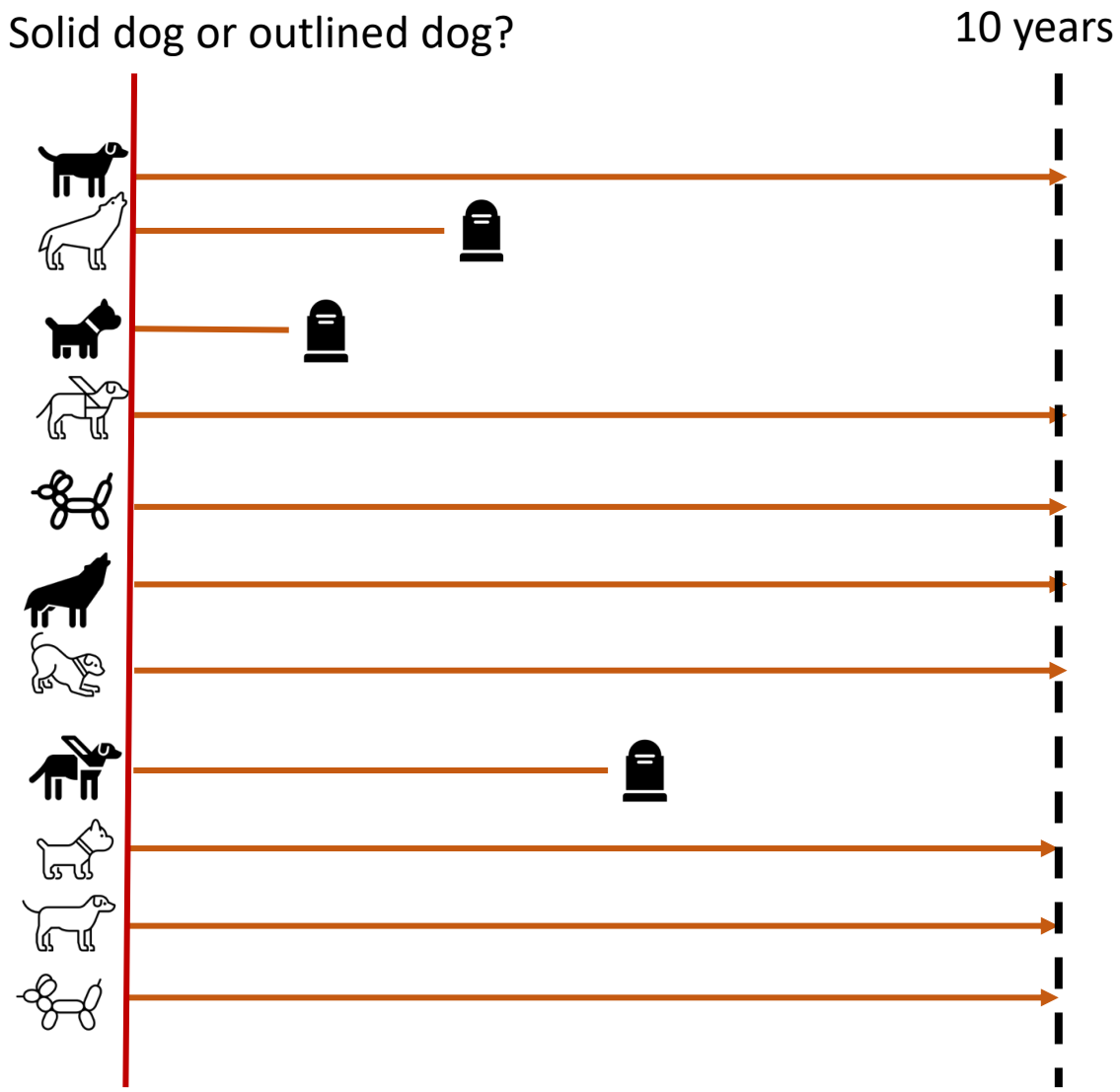
Refining the question: Outcome(s)



Refining the question: Actions

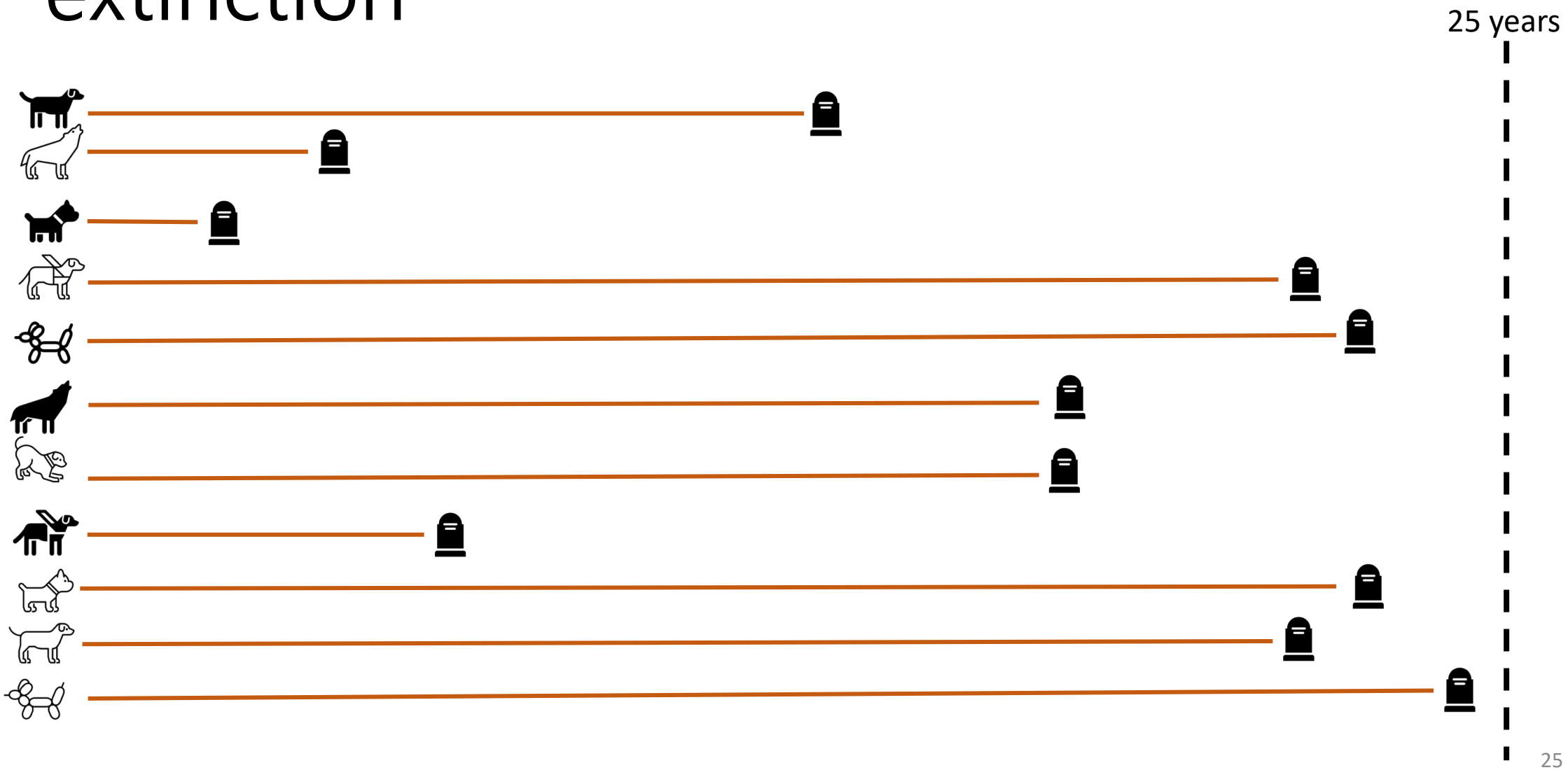


Refining the question: Groups



Chapter 3: Why learn survival analysis?

Learn about cohorts followed to extinction



Risk as a function

First, some notation

Let T_i represent the time from the origin to death for each dog i

Risk as a scalar

Risk at 25 years $= P(T_i \leq 25) = 1$

Or

Risk at 10 years $= P(T_i \leq 10) = 0.27$

Risk as a function

First, some notation

Let T_i represent the time from the origin to death for each dog i

Risk as a function

Risk at t years $= P(T_i \leq t) = F(t)$

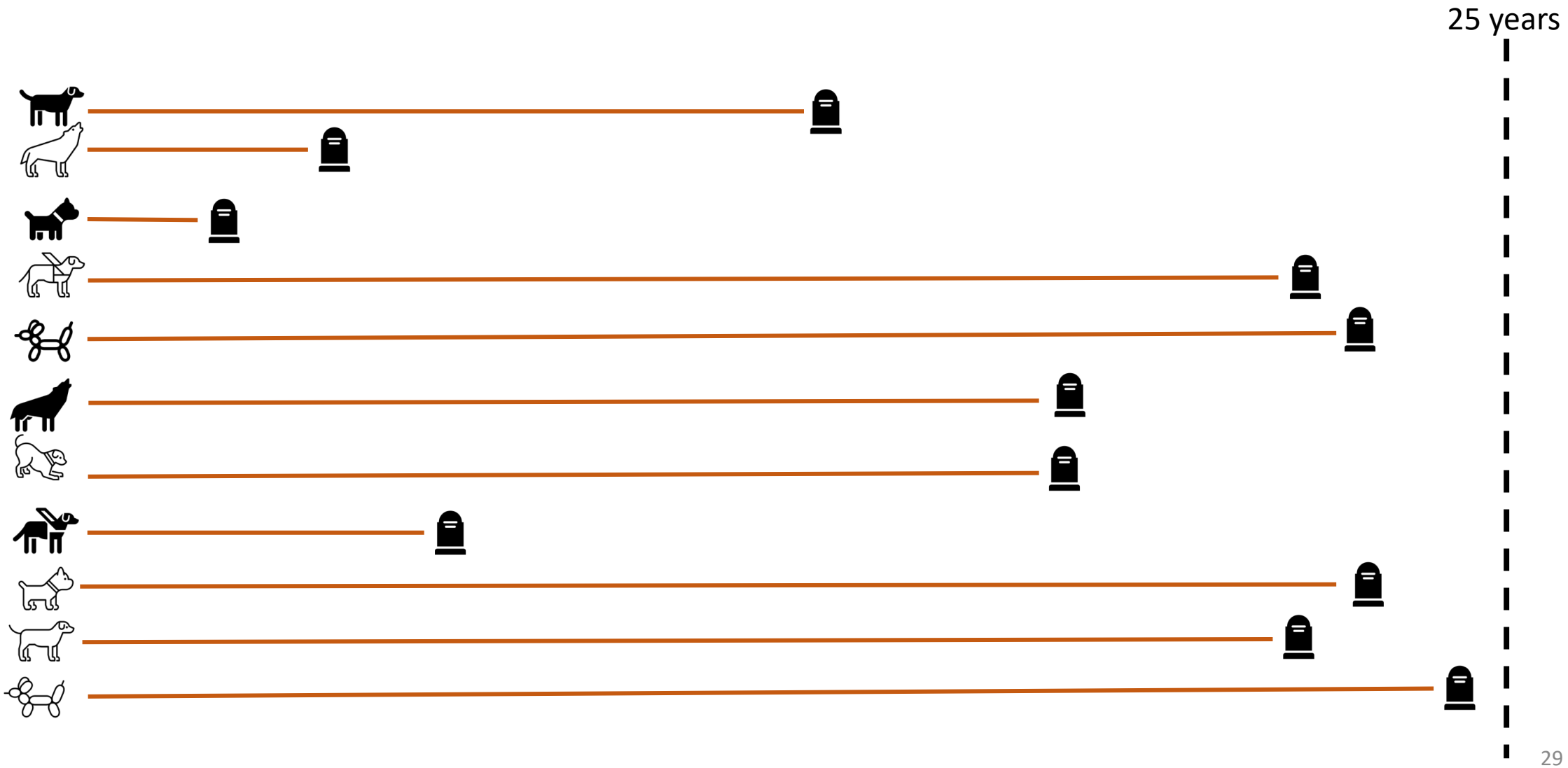
Learn from non-closed cohorts

Cohorts may be *open on the right* or *open on the left*

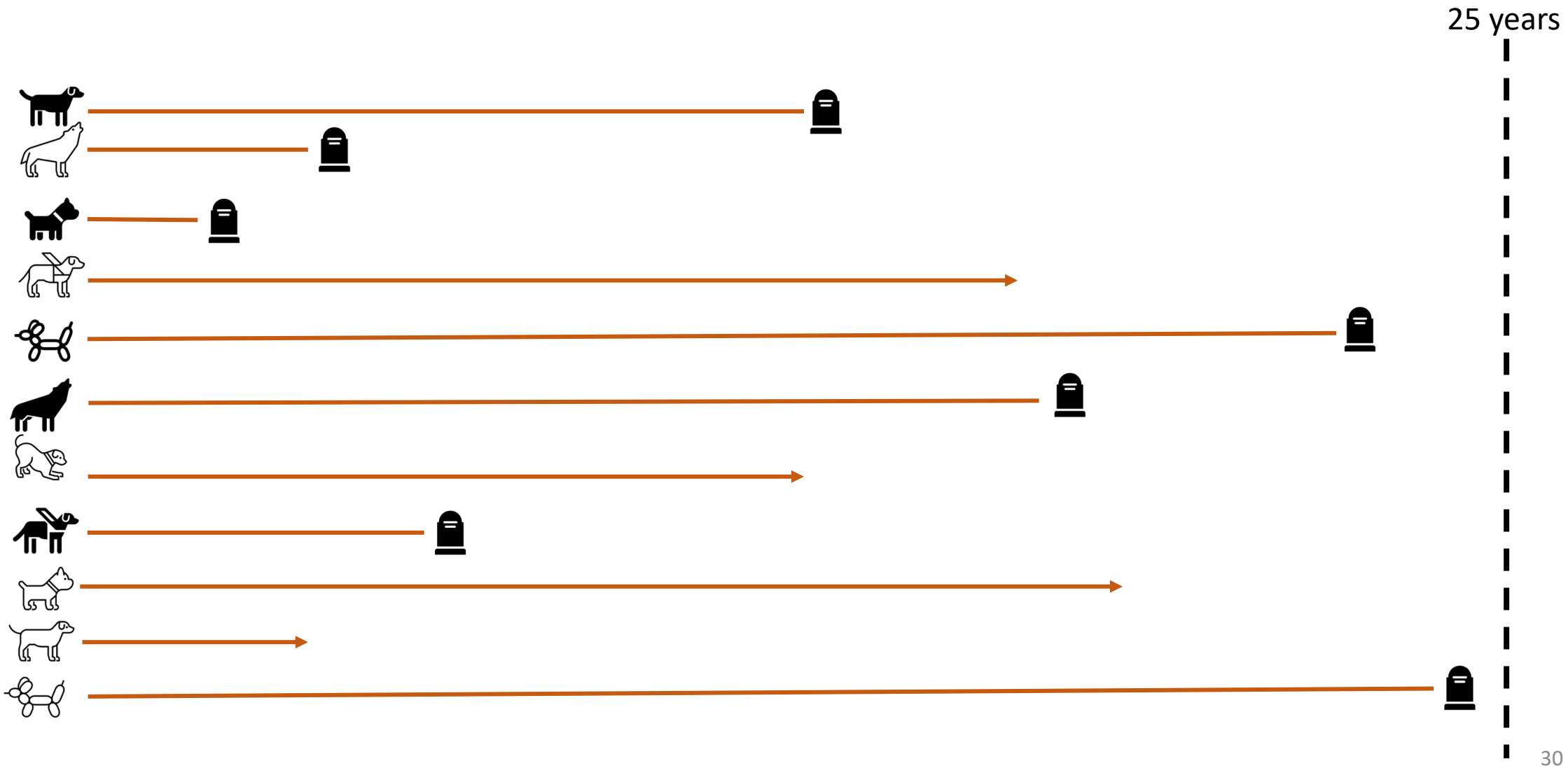
Open on the right: drop out, administrative censoring

Open on the left: late entry

Learn from non-closed cohorts



Learn from non-closed cohorts



Cohorts can be open on the right even without dropout!

Say you wish to estimate the 5-year risk of death among people entering HIV care.

You enroll people in the study at entry into care (origin) between 2012 and 2020.

If your methods require a closed cohort, you require 5 full years of follow-up, so can only use data on those enrolled between 2012 and 2015, essentially throwing away 5 years of data.

But using time-to-event methods, the partially observed person time (and any events) among those enrolled between 2015 and 2020 also contribute to the study.

Consider time-varying exposures

- Nonadherence
- Policy changes

Critical to compare the right people at the right times!

Chapter 4: Line diagrams

Line diagrams

Visualize choices of origin and timescale

See which participants are being compared at a given point in time.

In next week's reading, we will see how choice of timescale results in very different line diagrams and very different risk functions.

Line diagrams, steps

1. Draw and label axes (y axis is usually studyid, x axis is timescale, starting at the origin)
2. Calculate the amount of time after the origin that the first person enters the study. Place an open circle at this timepoint.
3. Draw a line from this open circle until the last available information for that participant.
4. If an event was observed, place a closed circle at that time
5. If no event was observed, place an arrow at that time
6. Repeat for remaining participants

An example, drawing line diagrams.

Say have a (small) cohort study of 5 soldiers selected at random from all people enlisting in the armed services between 2004 and 2014 and followed for mortality up to 10 years.

Complete line diagrams as described by the following 3 slides.

Example data

Draw a line diagram with the x axis as *age (starting at age 18)* and the y axis as soldier id number.

ID number	Age at enlistment	Enlistment date	Last date with available information	Vital status at last info
1	24	1 July 2008	Today()	Alive
2	18	1 July 2008	1 January 2011	Dead
3	21	1 January 2011	15 June 2016	Dead
4	35	1 January 2011	Today()	Alive
5	19	1 July 2013	Today()	Alive

Recall: Say we have a (small) cohort study of 5 soldiers selected at random from all people 18 and over enlisting in the armed services **between 2004 and 2014** and followed for mortality **up to 10 years**.

Example data

Draw a line diagram with the x axis as *calendar time (starting 1 Jan 08)* and the y axis as soldier id number.

ID number	Age at enlistment	Enlistment date	Last date with available information	Vital status at last info
1	24	1 July 2008	Today()	Alive
2	18	1 July 2008	1 January 2011	Dead
3	21	1 January 2011	15 June 2016	Dead
4	35	1 January 2011	Today()	Alive
5	19	1 July 2013	Today()	Alive

Recall: Say we have a (small) cohort study of 5 soldiers selected at random from all people 18 and over enlisting in the armed services **between 2004 and 2014** and followed for mortality **up to 10 years**.

Example data

Draw a line diagram with the x axis as *time since enlistment* and the y axis as soldier id number.

ID number	Age at enlistment	Enlistment date	Last date with available information	Vital status at last info
1	24	1 July 2008	Today()	Alive
2	18	1 July 2008	1 January 2011	Dead
3	21	1 January 2011	15 June 2016	Dead
4	35	1 January 2011	Today()	Alive
5	19	1 July 2013	Today()	Alive

Recall: Say we have a (small) cohort study of 5 soldiers selected at random from all people 18 and over enlisting in the armed services **between 2004 and 2014** and followed for mortality **up to 10 years**.