Main Figures, revision 4

Rachael Caelie (Rocky) Aikens

9/1/2021

Methods

Disease model

Each person, i, has an underlying (unobserved) disease state Z_i. We'll also define a quantity, R_i, which describes how closely the individual's symptoms match a "classical presentation" of the disease, with $R = 0$ representing no resemblance (not representative) and $R = 1$ being a perfect classical presentation (perfectly representative).

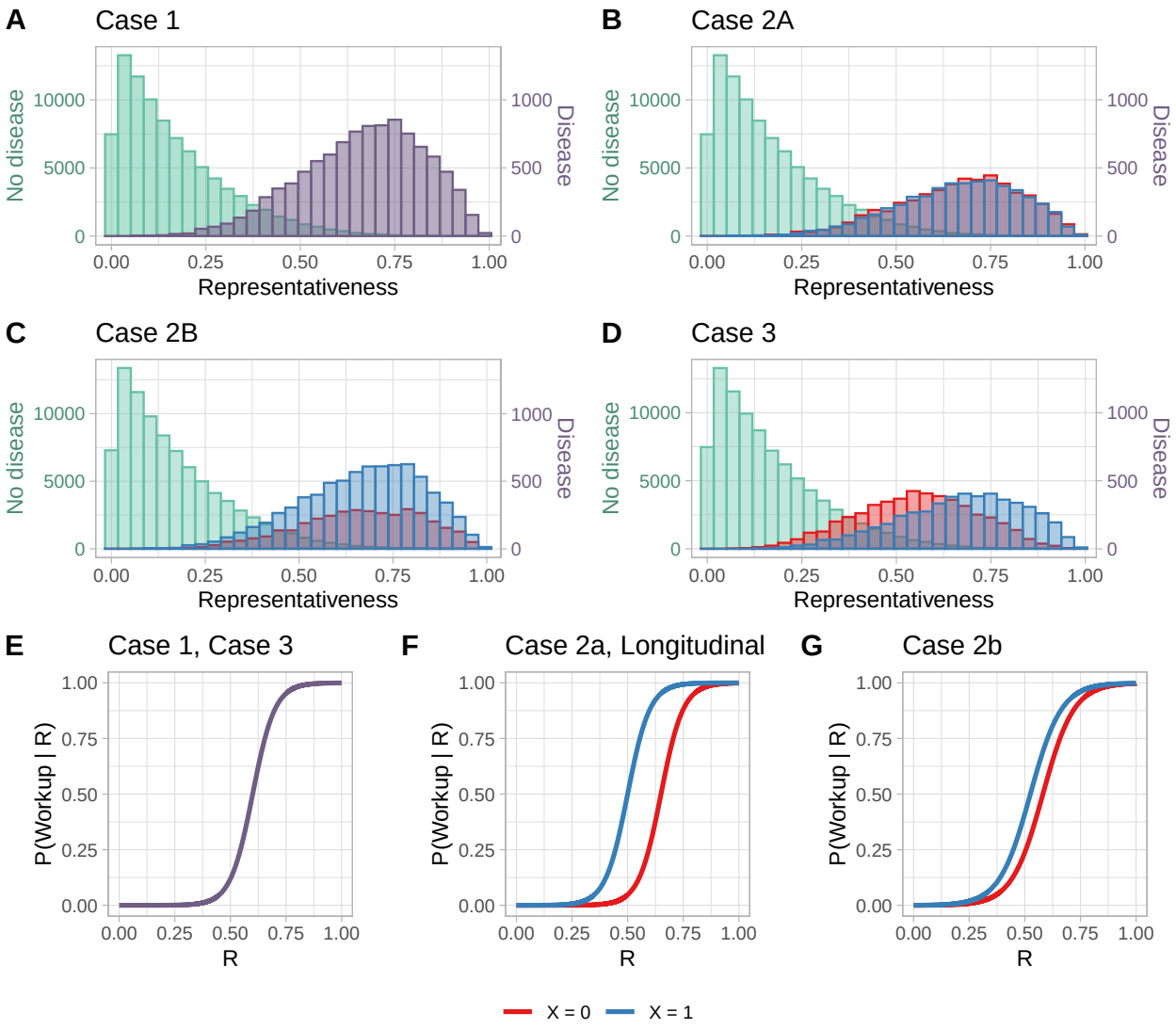
Here's a probabilistic model for one way we could generate these data:

$$\begin{aligned} Z_i &\sim_{iid} \text{Bernoulli}(0.1) \\ R_i | Z_i = 1 &\sim_{iid} \text{Beta}(5, 2.5) \\ R_i | Z_i = 0 &\sim_{iid} \text{Beta}(1, 5) \end{aligned}$$

This means that people *with* the disease tend to have higher values of R_i and people without the disease largely have lower values of R_i, with some exceptions on both sides (for example, individuals with the disease whose presentations are mild or atypical, and individuals without the disease who have some other condition that causes look-alike symptoms).

Clinical selection process

We'll want a figure that shows the clinical selection process and the disease model for each scenario.



Results

We'll suggest some potential consequences of selection bias from the diagnostic process with a set of illustrative simulated examples. Consider a researcher who aims to characterize a disease: it's indicidence, symptoms, and perhaps some risk factors. This researcher's work will then influence the way the disease is considered by the clinical community and perhaps beyond. To do this work, they select a large cohort of 100,000 individuals, and examine who recieved a diagnosis with the disease after 5 years.

We'll discuss the following cases:

- **Case 1: Selection for followup depends on representativeness, R**
- **Case 2: Selection for followup depends on representativeness, R, and a demographic characteristic X.**
  - **2a: X is not a risk factor for the disease**
  - **2b: X is a risk factor for the disease**

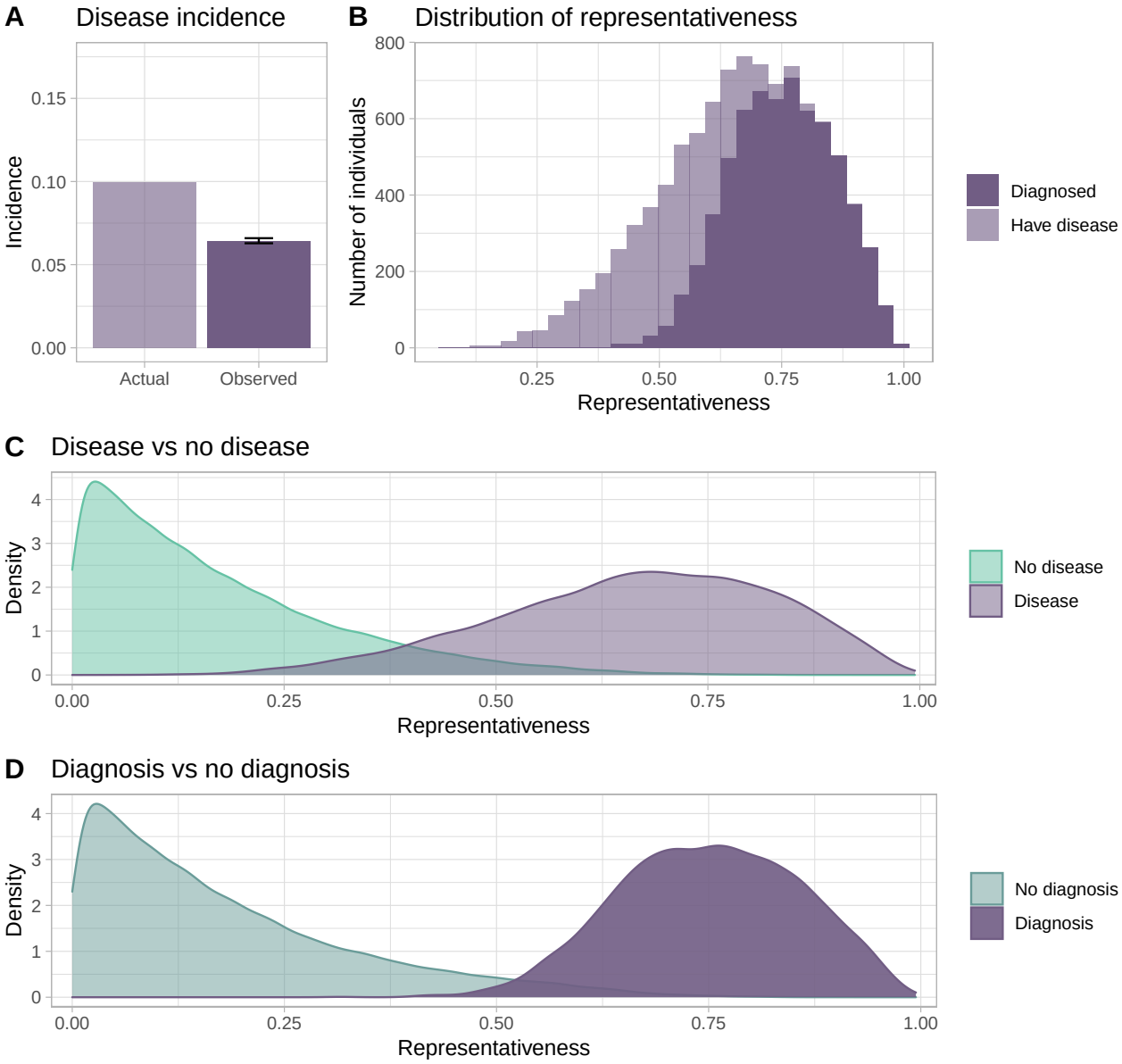
- **Case 3: Selection for followup depends on representativeness, R, but X determines the distribution of R**

In each case, the researcher may ask: What is the incidence of the disease? Is X a risk factor? What symptoms are representative of people with the disease? In each case, the hidden selection process – when ignored – can obscure the answers, potentially in a way that will lead to practice which reinforces or exacerbates these misunderstandings.

Case 1

People with the disease typically get follow-up much more than people without:

disease	Rate of diagnostic followup
FALSE	0.0160905
TRUE	0.6474314

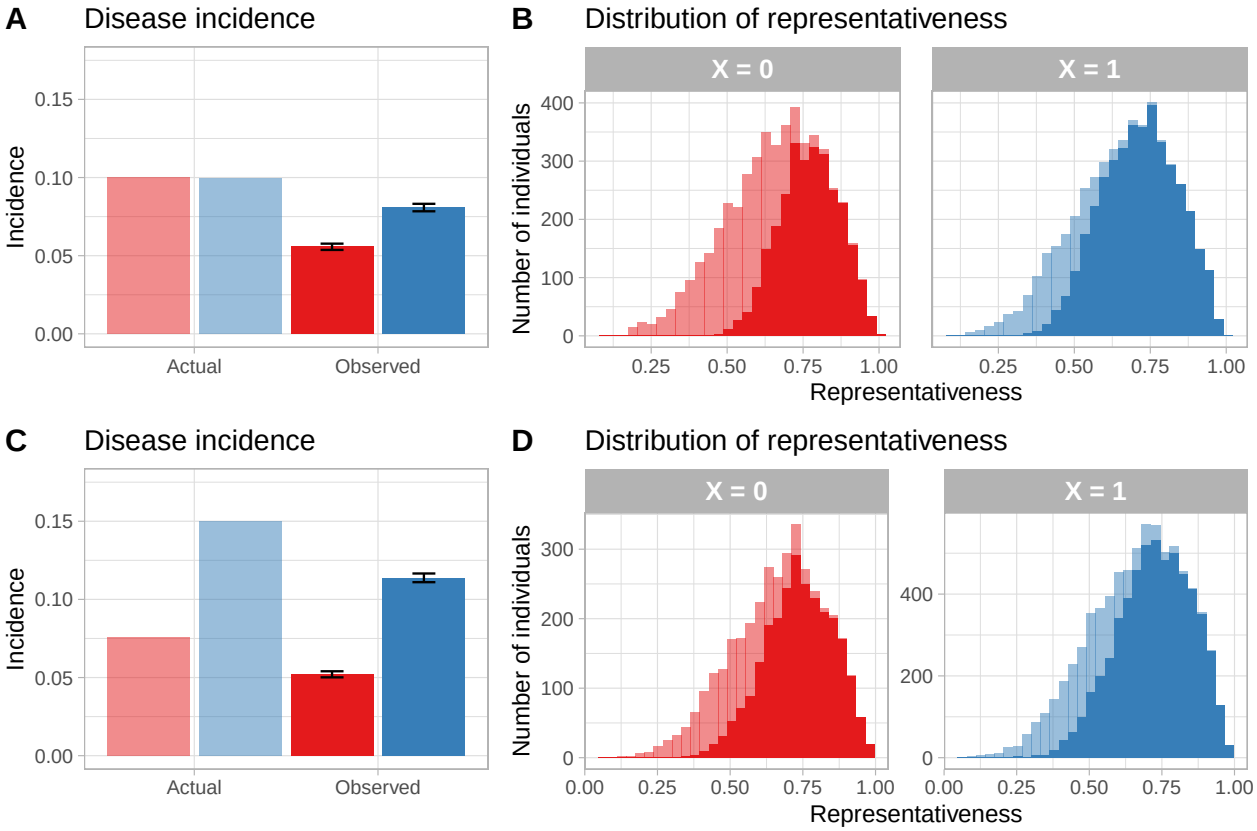


Case 2

Case 2A

Case 2b

head1		head2	
mpg	cyl	disp	hp
21	6	160	110
21	6	160	110
22.8	4	108	93



E

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

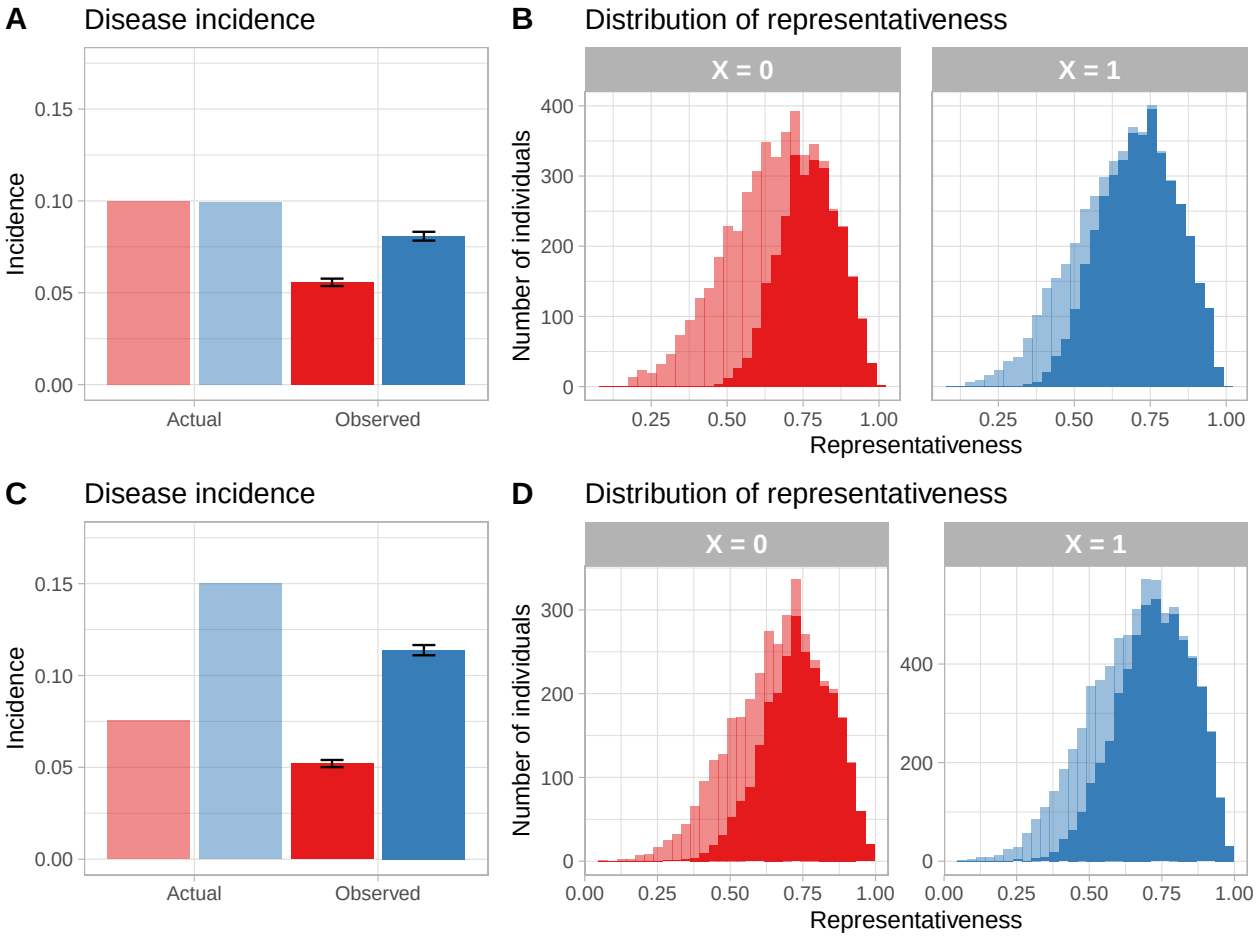
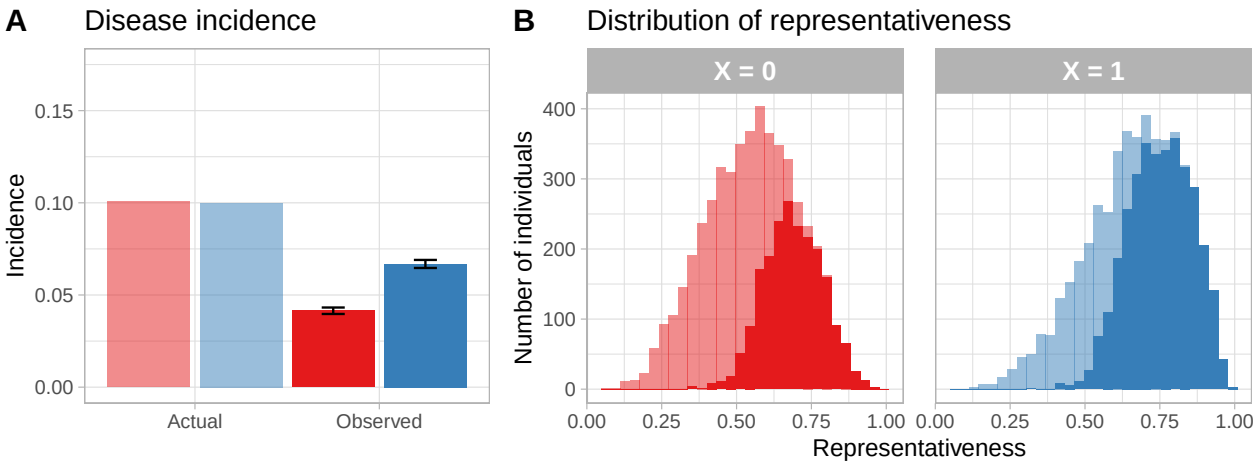


Table 2

percent_cases_diagnosed_X0	percent_cases_diagnosed_X1	est_incidence_X0	est_incidence_X1	est_relative_incidence	scenario
0.559	0.812	0.056	0.081	1.454	Case 1
0.673	0.757	0.050	0.114	2.253	Case 2

percent_cases_diagnosed_X0	percent_cases_diagnosed_X1	est_incidence_X0	est_incidence_X1	est_relative_incidence	scenario
0.0068	0.0055	0.001	0.0012	0.034	Case 1
0.0086	0.0055	0.001	0.0015	0.054	Case 2

Case 3



Longitudinal disease process

For the longitudinal example, we'll use the same simulation set-up as in Case 2.