# Main Figures, revision 3

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).

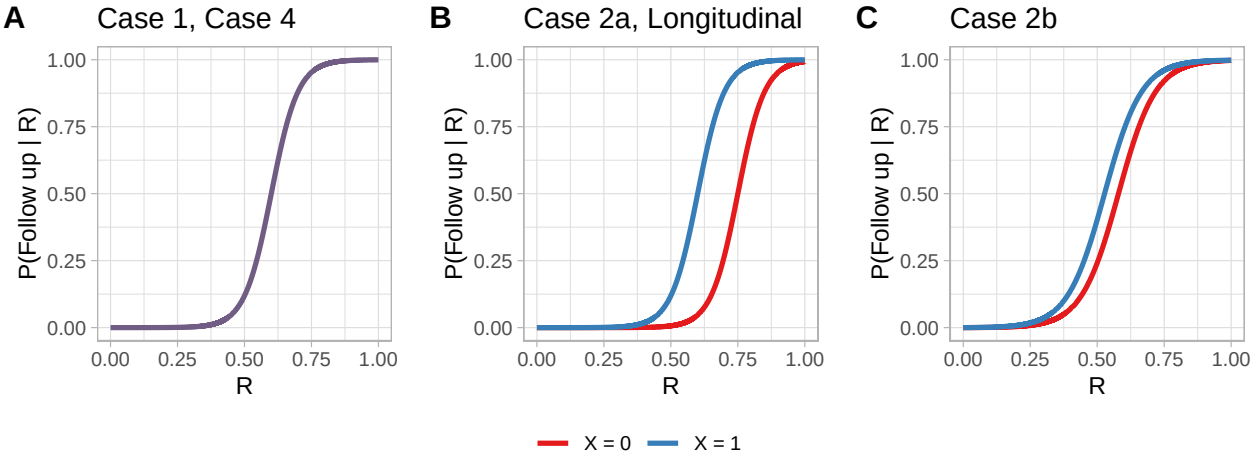
Here's a probabilistic model for one way we could generate these data:

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

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 just shows the clinical selection process 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:

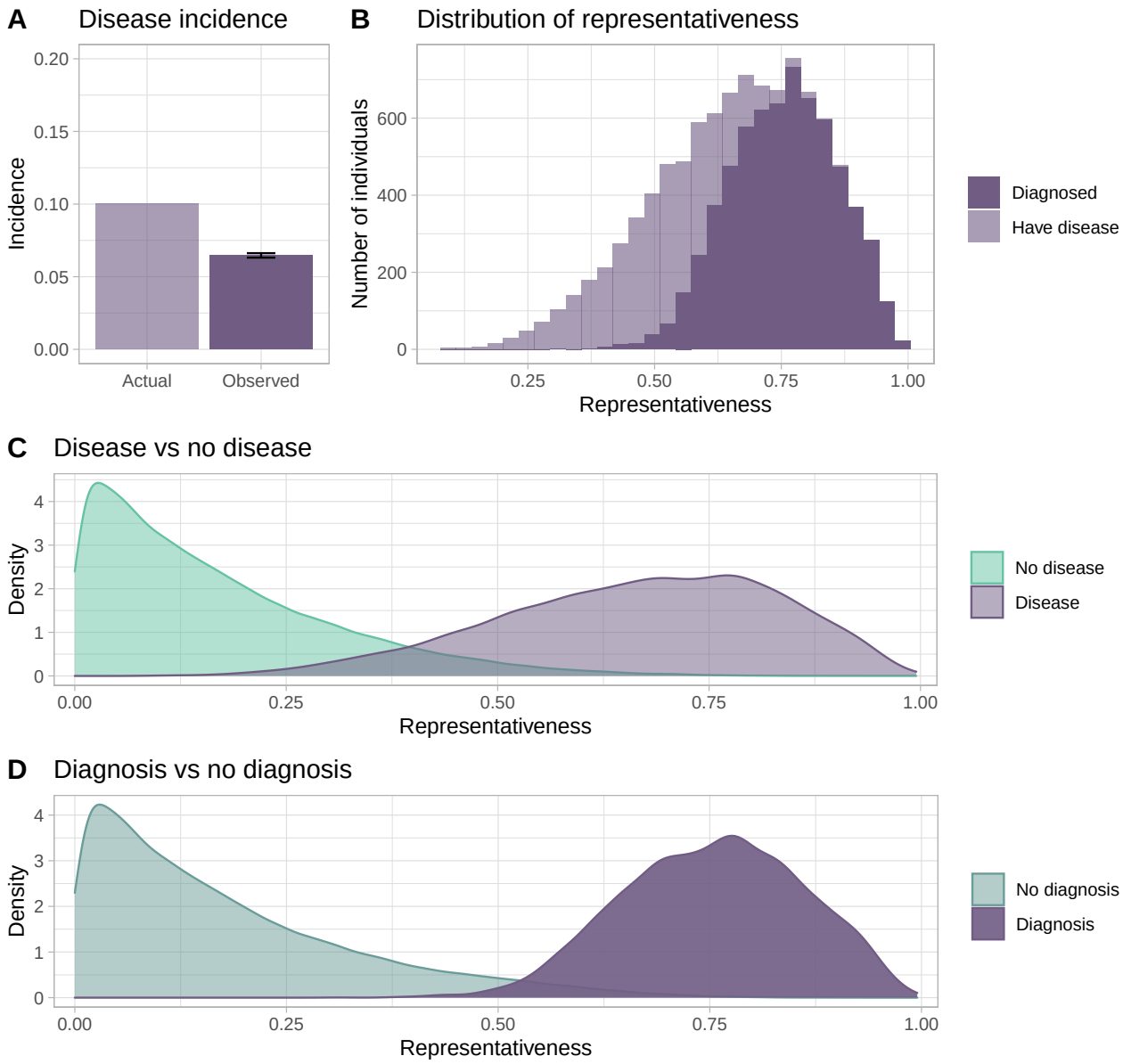
- **Case 1: Selection for followup depends on representativeness, $R$**
- **Case 2: Selection for followup depends on representativeness, $R$, and a demographic characteristic $X$.**
  - **2a: $X$ is not a risk factor for the disease**
  - **2b: $X$ is a risk factor for the disease**
- **Case 3: Selection 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.

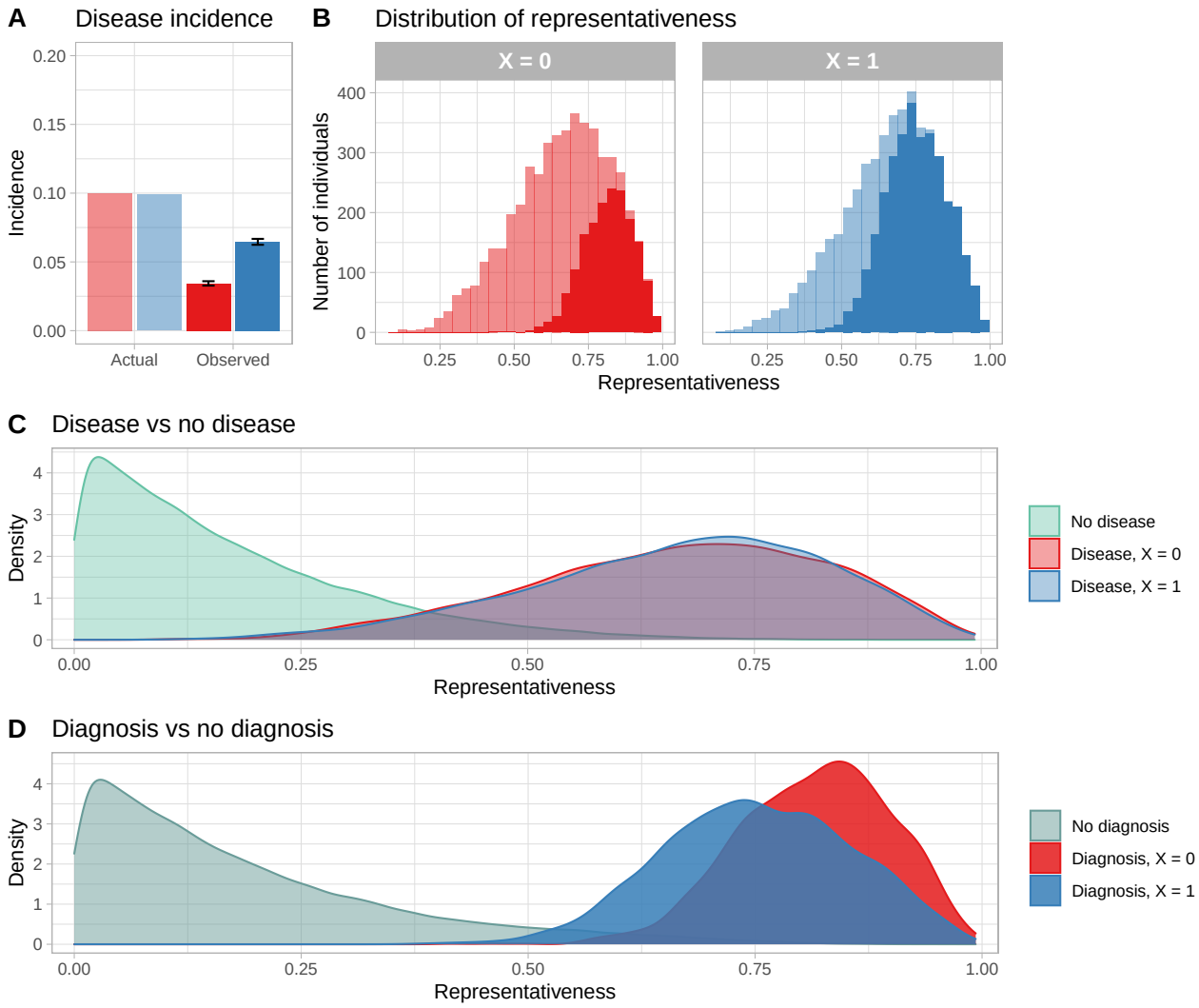
## Case 1

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

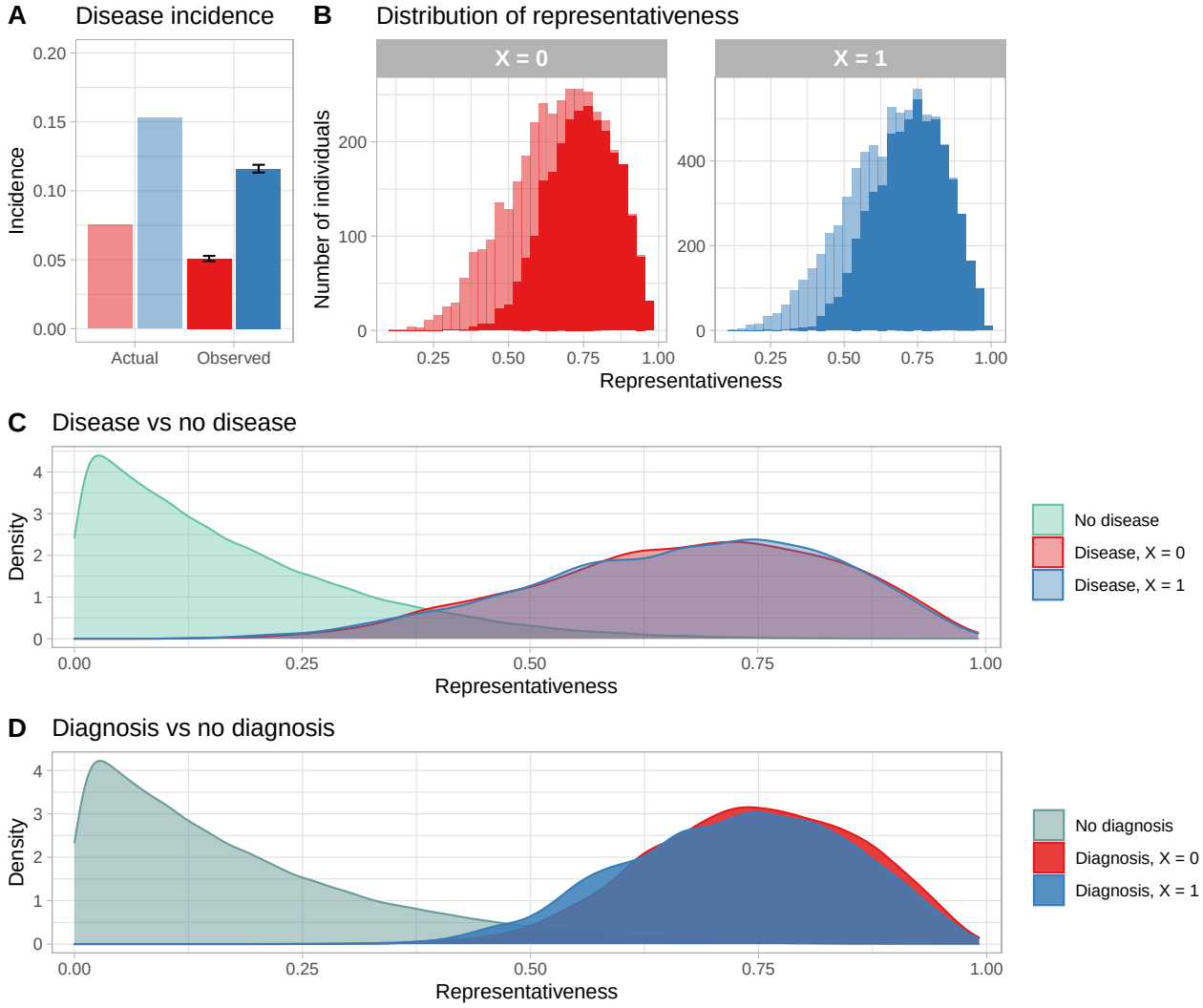
| disease | Rate of diagnostic followup |
|---|---|
| FALSE | 0.0162045 |
| TRUE | 0.6450873 |


A
Disease incidence
Incidence
0.20
0.15
0.10
0.05
0.00
Actual
Observed
B
Distribution of representativeness
Number of individuals
600
400
200
0
0.25
0.50
0.75
1.00
Representativeness
Diagnosed
Have disease
C
Disease vs no disease
Density
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Representativeness
No disease
Disease
D
Diagnosis vs no diagnosis
Density
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Representativeness
No diagnosis
Diagnosis

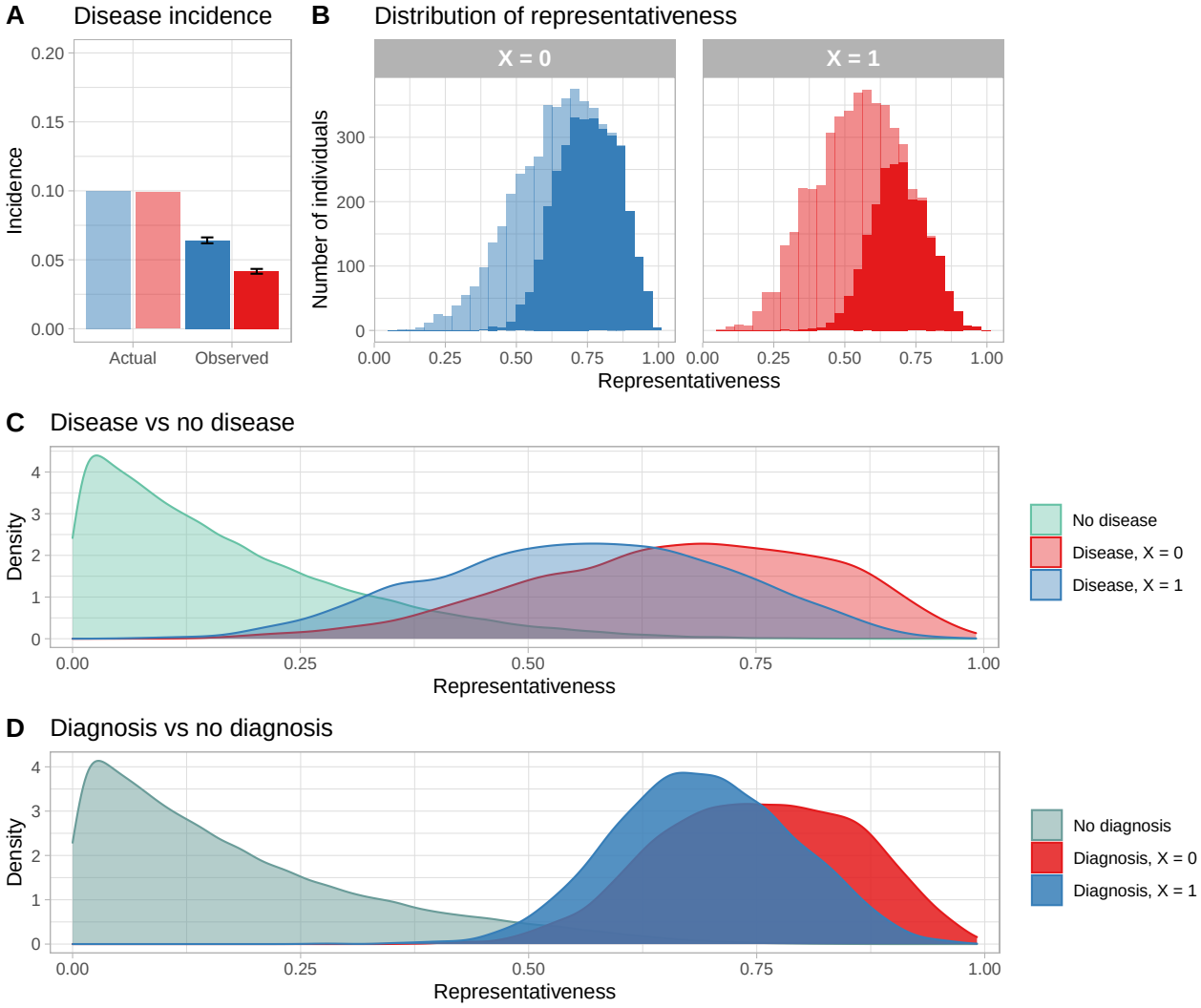
## Case 2a


A
Disease incidence
Incidence
0.20
0.15
0.10
0.05
0.00
Actual
Observed
B
Distribution of representativeness
X = 0
X = 1
Number of individuals
400
300
200
100
0
0.25
0.50
0.75
1.00
Representativeness
C
Disease vs no disease
Density
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Representativeness
No disease
Disease, X = 0
Disease, X = 1
D
Diagnosis vs no diagnosis
Density
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Representativeness
No diagnosis
Diagnosis, X = 0
Diagnosis, X = 1

## Case 2b

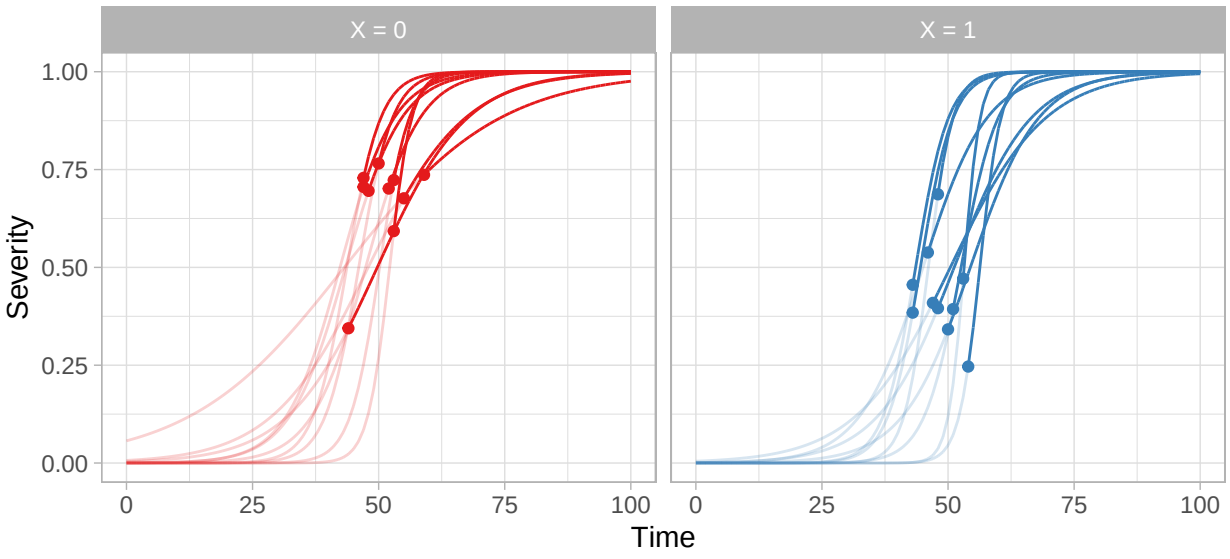

A
Disease incidence
Incidence
0.20
0.15
0.10
0.05
0.00
Actual
Observed
B
Distribution of representativeness
X = 0
X = 1
Number of individuals
200
100
0
400
200
0
0.25
0.50
0.75
1.00
Representativeness
C
Disease vs no disease
Density
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Representativeness
No disease
Disease, X = 0
Disease, X = 1
D
Diagnosis vs no diagnosis
Density
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Representativeness
No diagnosis
Diagnosis, X = 0
Diagnosis, X = 1

## Case 3




# Longitudinal disease process

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


X = 0
X = 1
Severity
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Time