Longitudinal Disease Progression

Rachael Caelie (Rocky) Aikens

6/24/2020

A model for longitudinal disease progression

Now let's imagine we're dealing with a disease that begins at a low severity and progresses every year. We'll assume that severity progresses from 0 to 1 according to a sigmoid function with the following parameters:

$$S_i(t) = \frac{1}{1 + e^{-\beta_i(t - T_i)}}.$$

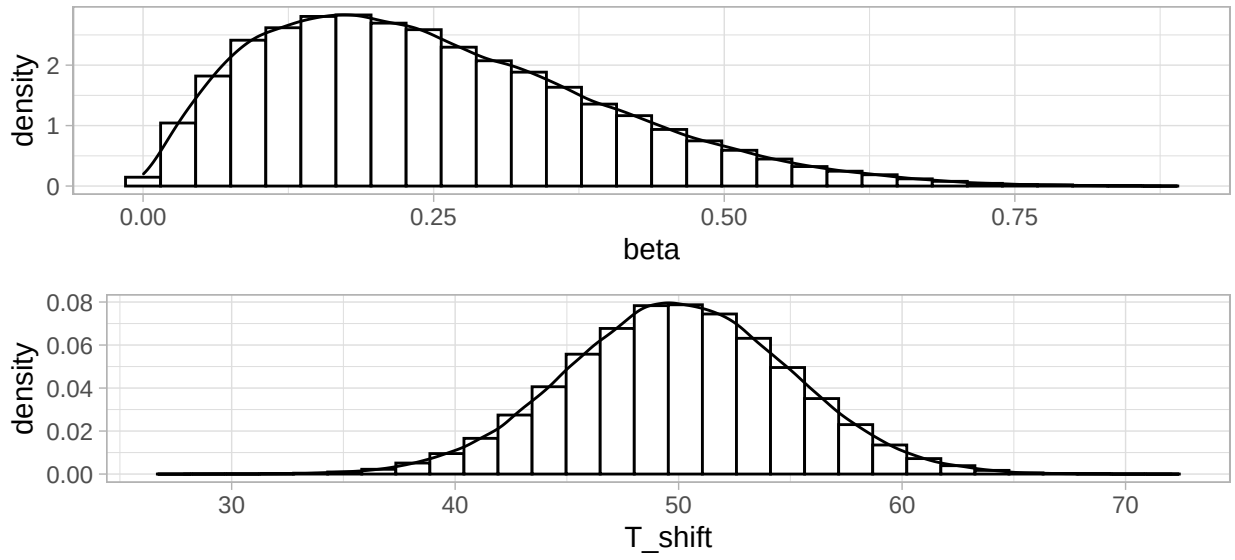
Here, β_i and T_i are intrinsic to the individual:

- β_i is the slope of progression. A larger β_i indicates a quickly progressing disease.
- T_i controlls the time of onset. T_i is the time when severity reaches 0.5

We'll simulate β_i and T_i from the following distributions:

$$\beta_i \sim_{iid} \text{Beta}(\alpha = 2, \beta = 6)$$
$$T_i \sim_{iid} \text{Normal}(\mu = 50, \sigma = 5)$$

The shapes of these distributions is shown below:



Inessense, most people have a gradual disease progression ($\mathbb{E}[\beta_i] = 0.25$), and the disease tends to hit intermediate severity at about age 50 ($\mathbb{E}[T_i] = 50$).

Underlying severity is related to - but not exactly in correspondence with - symptom representativeness, $R_i(t)$. Severity, $S_i(t)$, describes the underlying disease state, while representativeness, $R_i(t)$, describes the observable symptoms and how well they match a "textbook" description of the disease. We'll simulate $R_i(t)$ as a more noisy version of $S_i(t)$:

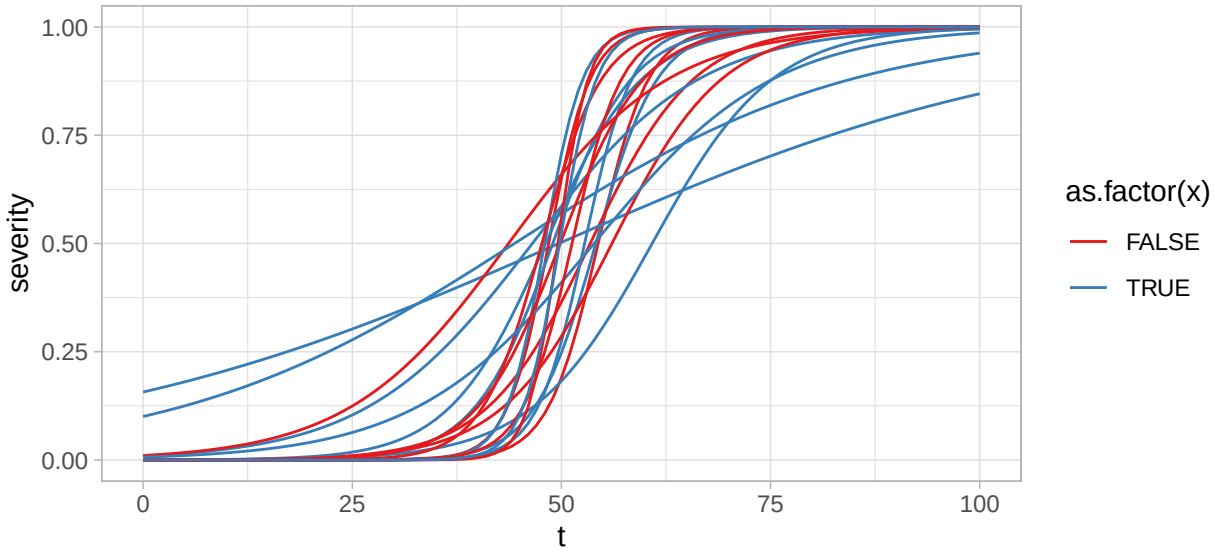
$$R_i(t) = \frac{1}{1 + e^{-\beta_i(t+\epsilon_{ti}-T_i)}}.$$

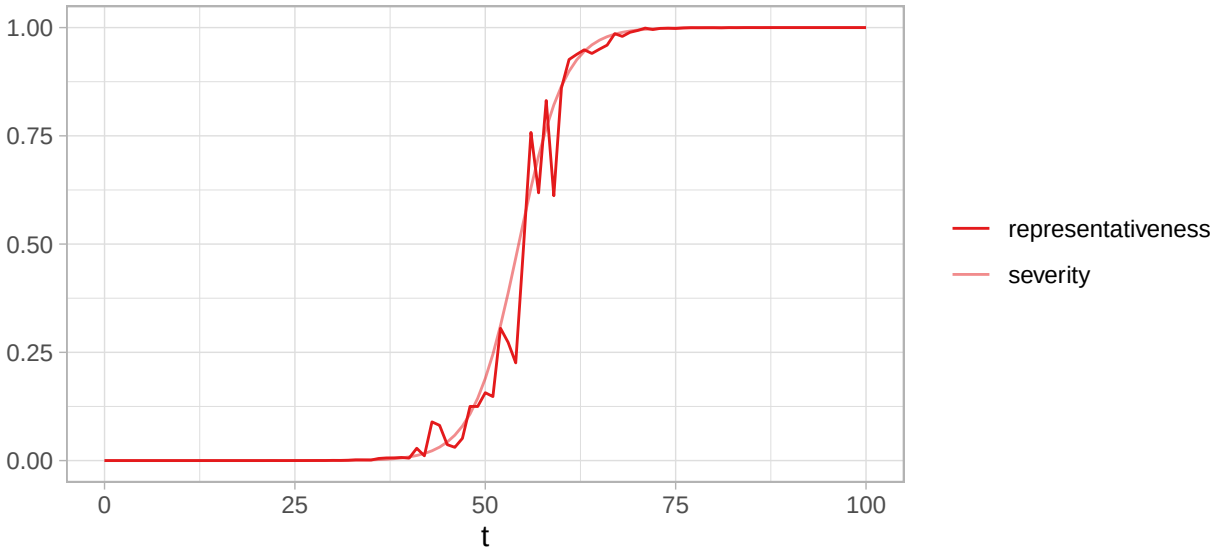
Where

$$\epsilon_{ti} \sim_{iid} N(0, 1).$$

Illustrative example

Here's a quick example dataset of 20 individuals. The plot below shows disease severity over time. Each individual has their own disease trajectory, but the disease trajectories for individuals with $x = 0$ and $x = 1$ are simulated with the same hyperparameters.





Model: Diagnosis rate depends on representativeness

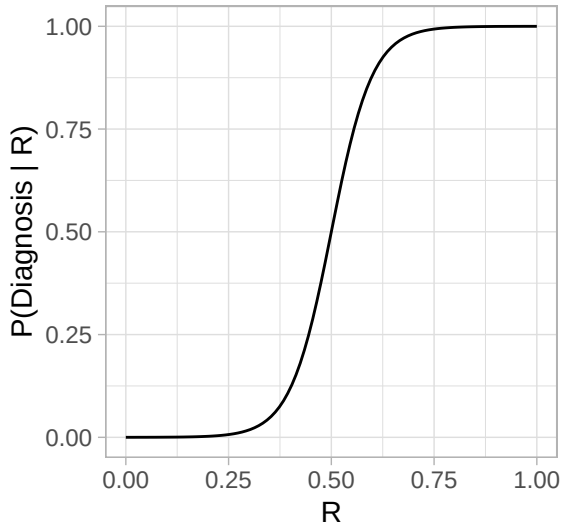
Below is a straightforward example in which diagnosis rate depends on representativeness, but not baseline characteristics. As in the single-time-point models, let

$$D_i \sim_{iid} \text{Bernoulli}\left(\phi(R_i)\right)$$

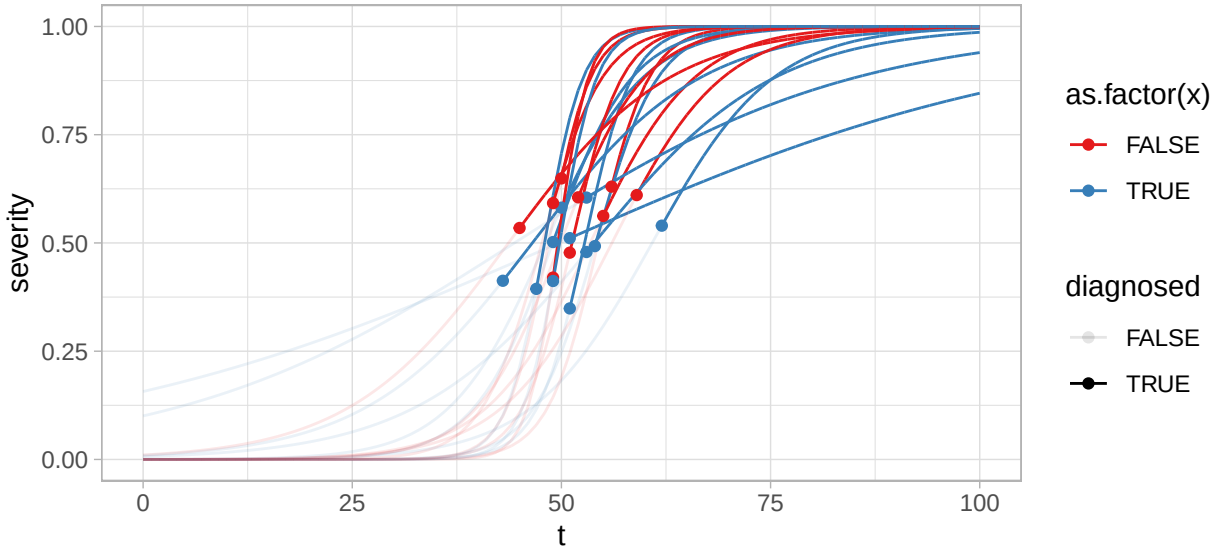
Where,

$$\phi(R_i) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 R_i))} - c$$

Here, $c = \frac{1}{1+exp(-\beta_0)}$ is a corrective constant to ensure that $\phi(0) = 0$. Let $\beta_1 = 20$ and $\beta_0 = -10$. This gives the following probability of diagnosis as a function of representativeness.



The way this plays out is that most people are diagnosed at some point early in the upswing of their diease progression, and their diagnosis probability does not depend on their value of X.

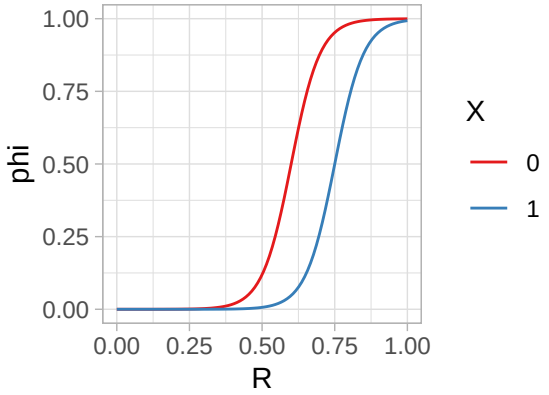


Model: Diagnosis rate depends on representativeness and baseline covariates

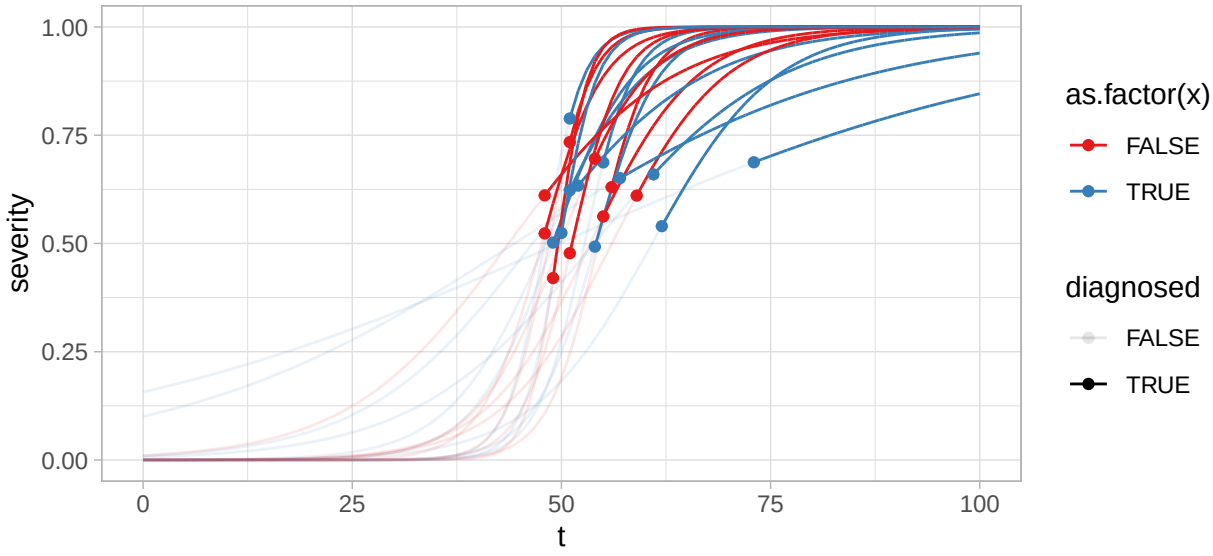
Now, let's suppose that the probability of diagnosis depends on representativeness and baseline characteristics:

$$\phi(R_i, X_i) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 R_i + \beta_2 X_i))} - c_{x_i},$$

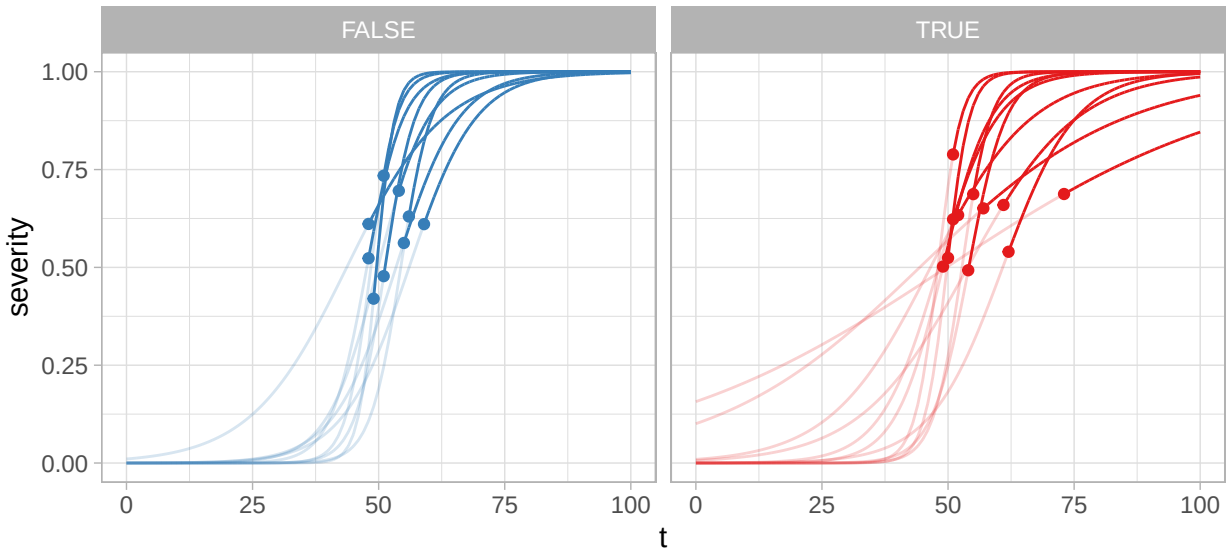
where again, $c_{x_i} = \frac{1}{1+exp(-(\beta_0+\beta_2 X_i))}$ is a corrective constant to ensure that $\phi(0, X_i) = 0$. Also, let $\beta_2 = -3$. Inessence, this means that people with $X_i = 0$ are diagnosed in the same way as above, but people with $X_i = 1$ are diagnosed with a lower probability, shown below:



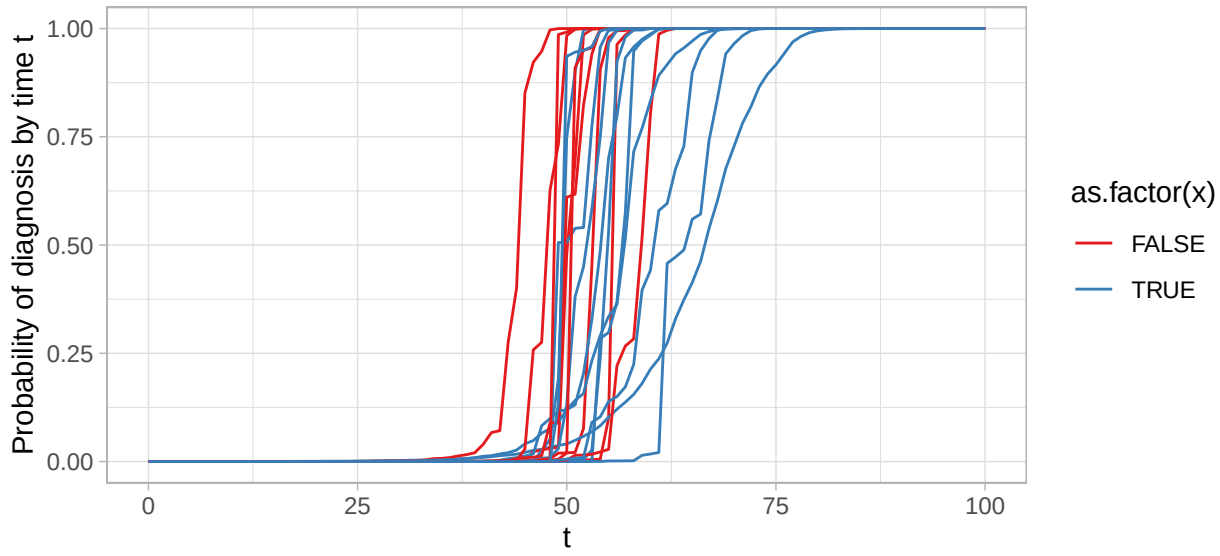
An example of how this plays out is shown below on a subset of the data. People with $X = 0$ tend to be diagnosed earlier in their disease progression than people with $X = 1$.



Here's another representation of the same data where the X groups are shown side-by-side:



Viewed another way, here's the probability of each patient being diagnosed by time t.



A Naive Study

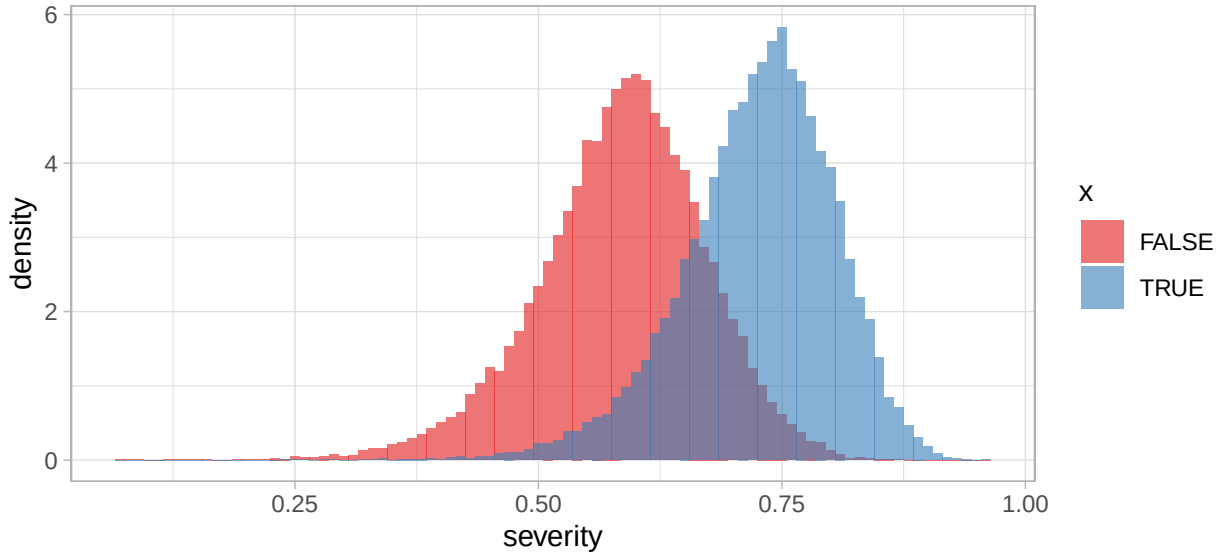
Under this model (where diagnosis probability depends on X), let's see what we glean from a naive study of these data.

Let's suppose there are 5×10^4 individuals with this disease in the EHR. We don't observe their underlying disease state; just their diagnosis. We also don't observe the severity of their illness prior to diagnosis.

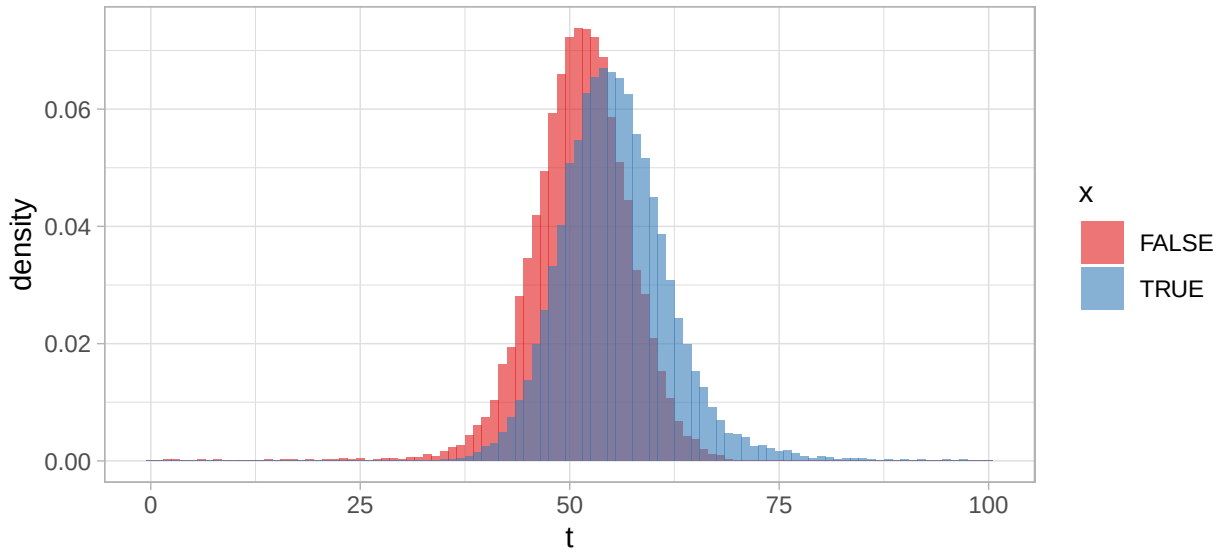
```
## # A tibble: 2 x 2
##   x         n
##   <lgl> <int>
## 1 FALSE 25295
## 2 TRUE  24676
```

```
## # A tibble: 2 x 2
##   x         n
##   <lgl> <int>
## 1 FALSE 25295
## 2 TRUE  24705
```

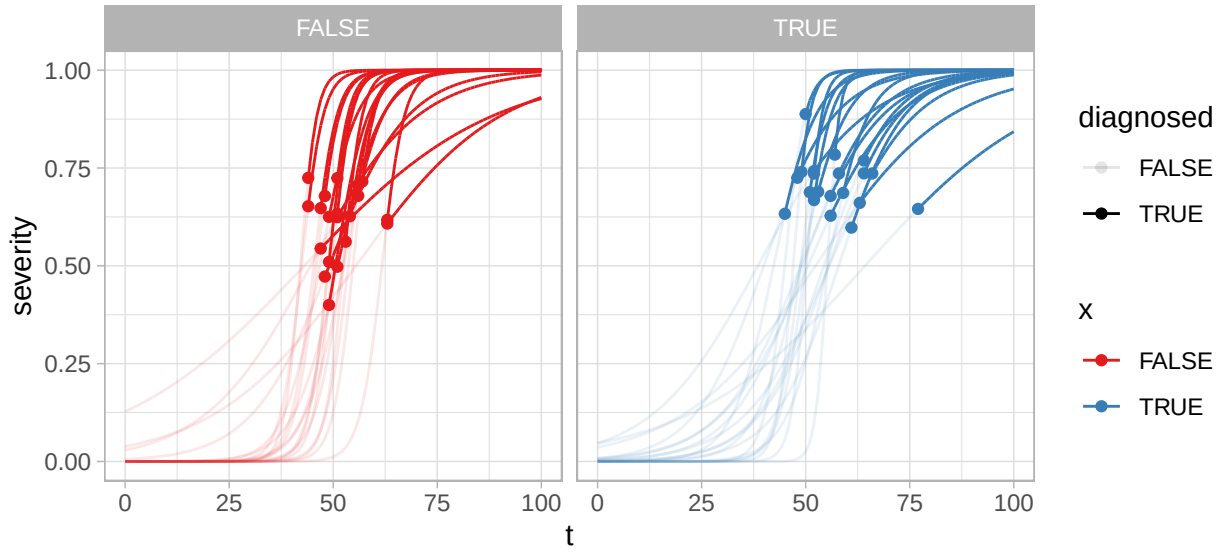
In this set-up, all 49971 individuals out of 5×10^4 with the disease are diagnosed. All of the undiagnosed individuals have $X = 1$. The plot below shows the severity at onset for individuals with $x = 0$ compared to $x = 1$. This probably is indicative of the fact that the people with $x = 0$ tend to be diagnosed at the start of their disease onset, while the people with $x = 1$ tend to be diagnosed later in their disease progression.



There's also a difference in the age of diagnosis between groups. Most people are diagnosed around age 50, when their disease onset tends to begin, but the diagnoses for people with $x = 1$ lag behind the diagnoses for people with $x = 0$.



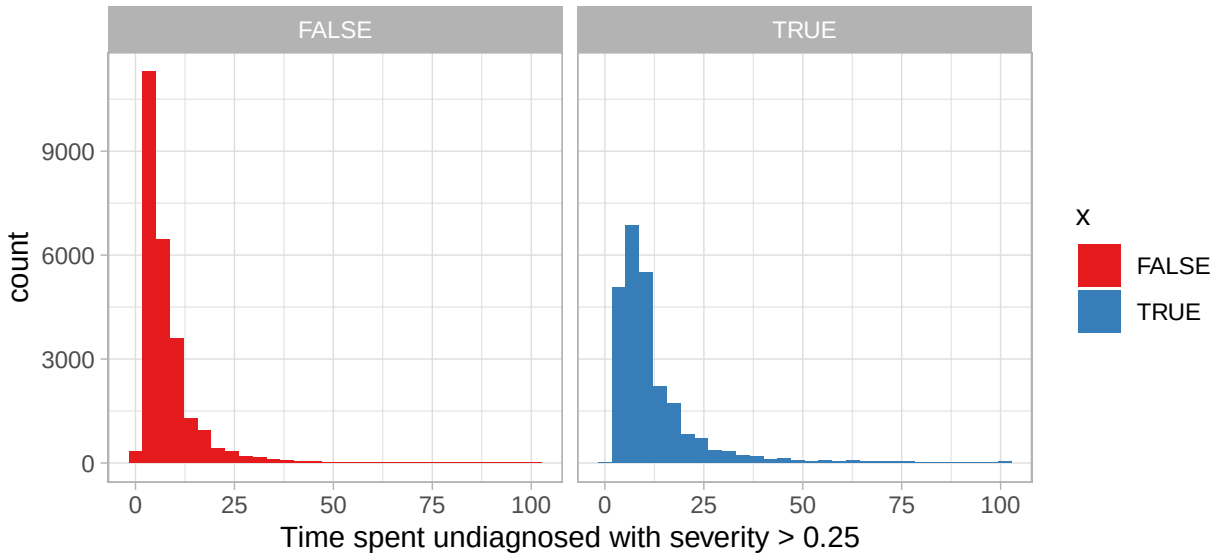
In fact, what is happening is that people with $X_i = 1$ are diagnosed later: with higher disease severities at older ages.



The Impacts

Undiagnosed time

One impact of this is that people in the underdiagnosed group tend to spend more time undiagnosed. Suppose that the disease symptoms are negligible until severity reaches at least 0.25, at which point the disease becomes an increasing burden. Below, we calculate the amount of time each group spends experiencing symptoms while undiagnosed.

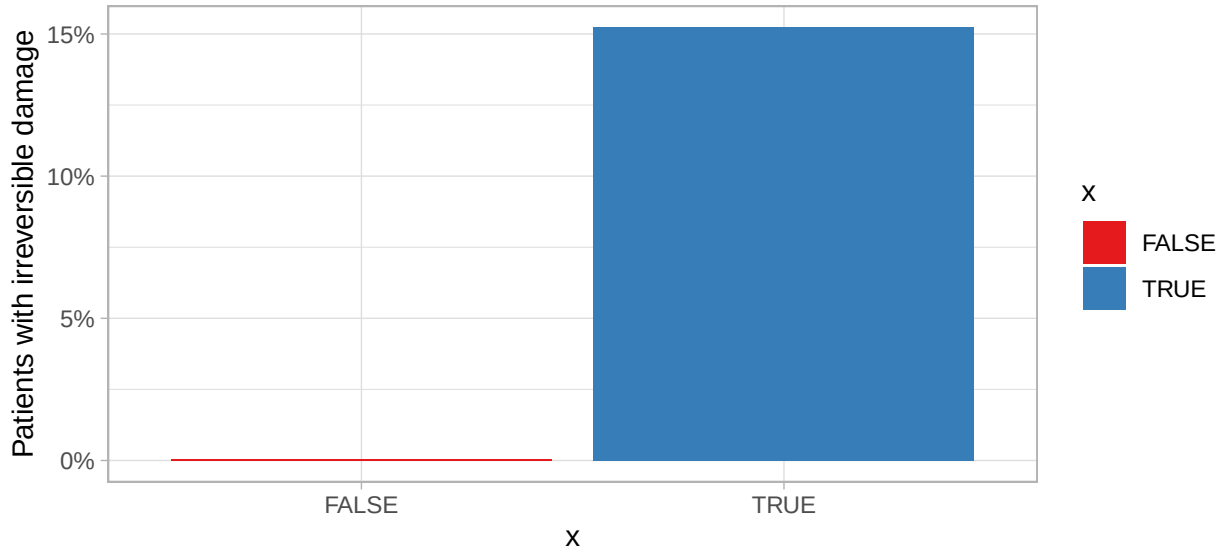


Both distributions are strongly right-tailed, indicating that most people are diagnosed right away, but some people may wait several years for diagnosis. The table below shows the quartiles and means of undiagnosed time for the different groups. People with $X_i = 1$ are likely to spend much longer experiencing burdensome symptoms while undiagnosed.

x	Q1	Q2	Q3	mean
FALSE	4	6	9	7.88
TRUE	6	9	14	12.18

Irreversible damage

An analog to undiagnosed time is irreversible damage. Suppose that, once a person's severity rises above 0.75, irreversible damage begins which cannot be undone my any course of treatment. However, if the person is correctly diagnosed by the time their severity reaches this level, the damage can be undone.



Misdiagnosis

Suppose that, for every year that a patient experiences symptoms and does not have a correct diagnosis, they have a certain fixed probability of *misdiagnosis* (that is, diagnosis with a disease that is not the correct one). People who are misdiagnosed, $M_i = 1$ may recieve incorrect treatments, will appear in the EHR with the incorrect disease label, and may have lower probability of being correctly diagnosed in the future (although I haven't modeled this yet.)

As a simple model, let's suppose

$$M_i \sim_{iid} \text{Bernoulli}\left(\Psi(R_i, D_i)\right),$$

where

$$\psi(R_i, D_i) = I_{D_i=0} \frac{1/10}{1 + exp(-(\beta_0 + \beta_1 R_i))} - c$$

Inessence, a person who is undiagnosed when they reach disease severity R_i has the following diagnosis probabilities: