Selection for Follow-up

Rachael Caelie (Rocky) Aikens

7/22/2021

Premise

Suppose an analyst is trying to characterize a disease, and that that characterization will be used by clinicians to inform the way they diagnose the disease. Let's imagine the researcher selects a cohort of 100000 individuals without the disease and performs 5 years of follow-up. In those 5 years, some number of individuals develop the disease and some number are diagnosed. The researcher wants to understand the incidence rate of the disease and to characterize the distribution of symptoms among those with the disease. These observations will then direct the future practice of diagnosis.[1]

What happens when doctors decide who gets follow-up

These simulations illustrate some scenarios in which there is a reliable diagnostic method available (a test, referall to an expert, a physical exam), but doctors decide who should have this diagnostic follow-up. This isn't meant to cover all possibilities, just to suggest what might happen in a couple different situations. We look at three cases:

- **Case 1: Follow-up based on representativeness:** Patients whose symptoms are thought to be more representative of the disease are more likely to recieve follow-up.
- **Case 2: Follow-up based on representativeness, demographics (disease risk at parity):** Patients recieve follow-up based on the representativeness of their symptoms and their demographics, but disease risk is equivalent between demographic groups.
- **Case 3: Follow-up based on representativeness, demographics (disease risk differs):** Patients recieve follow-up based on the same characteristics as case 2, but this time there *is* an underlying difference in disease risk between groups
- **Case 4: Follow-up based on representativeness, demographics (representativeness differs):** Similar to case 2, but in this scenario individuals in certain demographic groups present with symptoms thought to be more representative of the disease, while individuals in other demographic groups present with symptoms less canonically associated with the disease.

[1] This set-up is an attempt at better articulating the study design that might have generated these data, but I'm not sure it's exactly right. If the researcher can't observe underlying disease state then they can't select a healthy cohort, so really you run into issues before you even begin.

Case 1: Follow-up based on representativeness:

Set-up

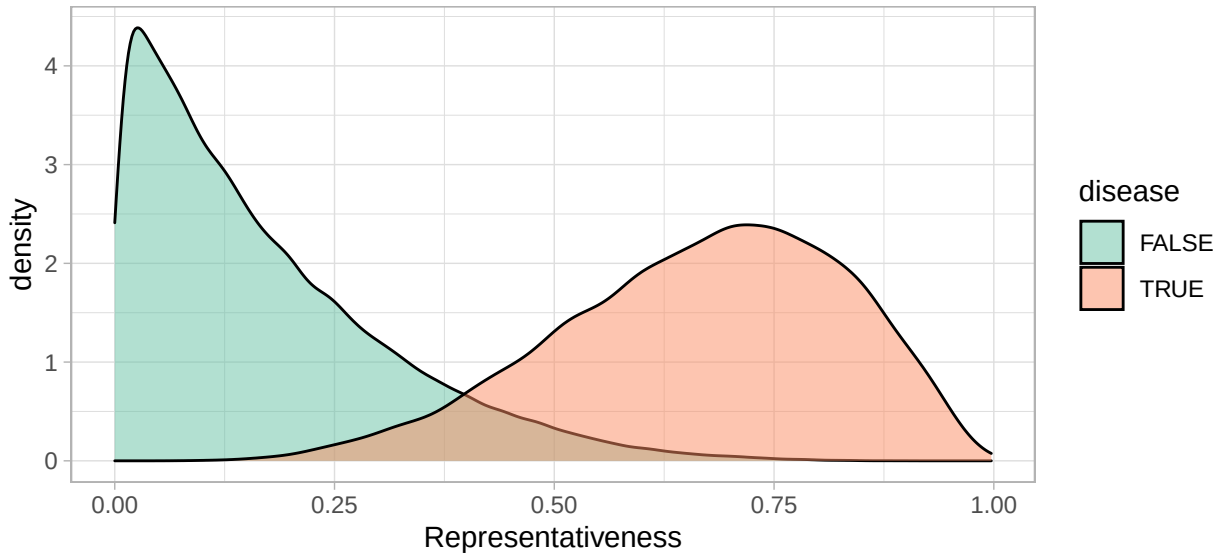
Disease model

We'll start with a single group, and use the notation we've used previously. 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).

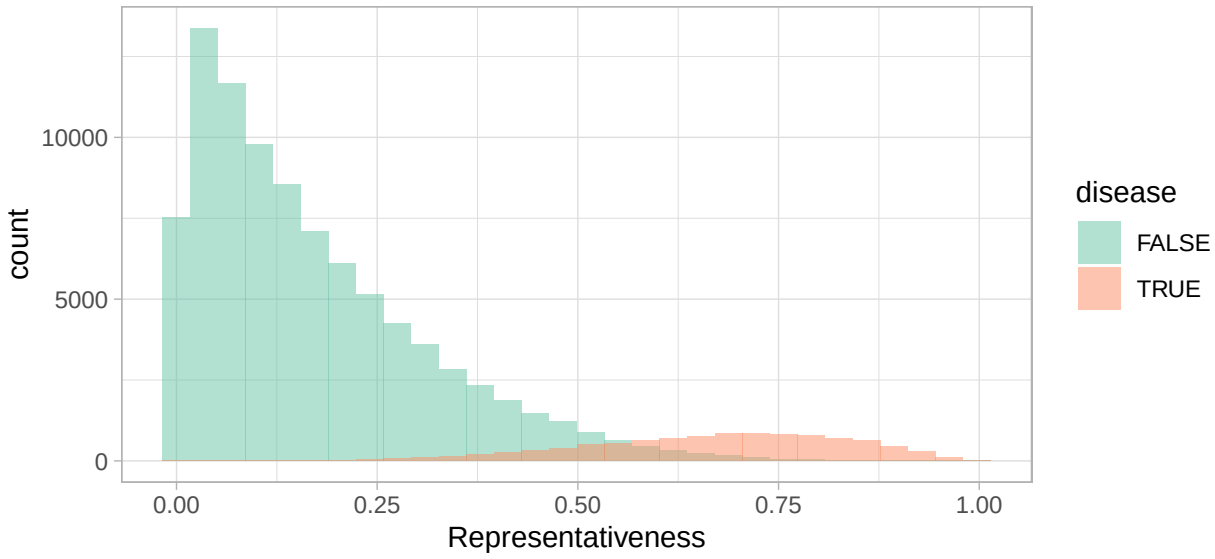
Here's a probabilistic model for one way we could generate these data:

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

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). Here's a histogram of the distribution of representativeness in individuals with and without the disease:



An important thing to remember in this scenario is that individuals with the disease generally greatly outnumber those without. This means that even though it is relatively rare for an individual without the disease to have look-alike symptoms, there may be a substantial number of them overall:



Clinical selection process

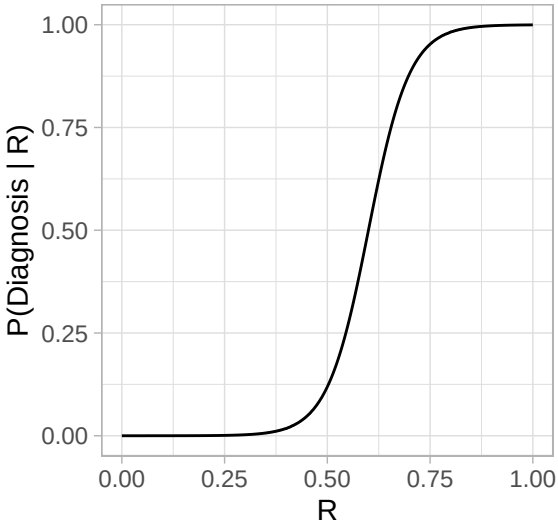
Even though in this case diagnostic follow-up is perfectly reliable, doctors don't have the resources to perform this follow-up for all people. Instead, suppose that a person's likelihood of recieving diagnostic followup is a function of R_i, namely

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

$\phi(R_i)$ could be many possible functions, here I use a sigmoid function:

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

Here, $c = \frac{1}{1+exp(-\beta_0)}$ is a corrective constant to ensure that $\phi(0) = 0$. Let $\beta_1 = 20$ and $\beta_0 = -12$. This gives the following probability of diagnosis as a function of representativeness.

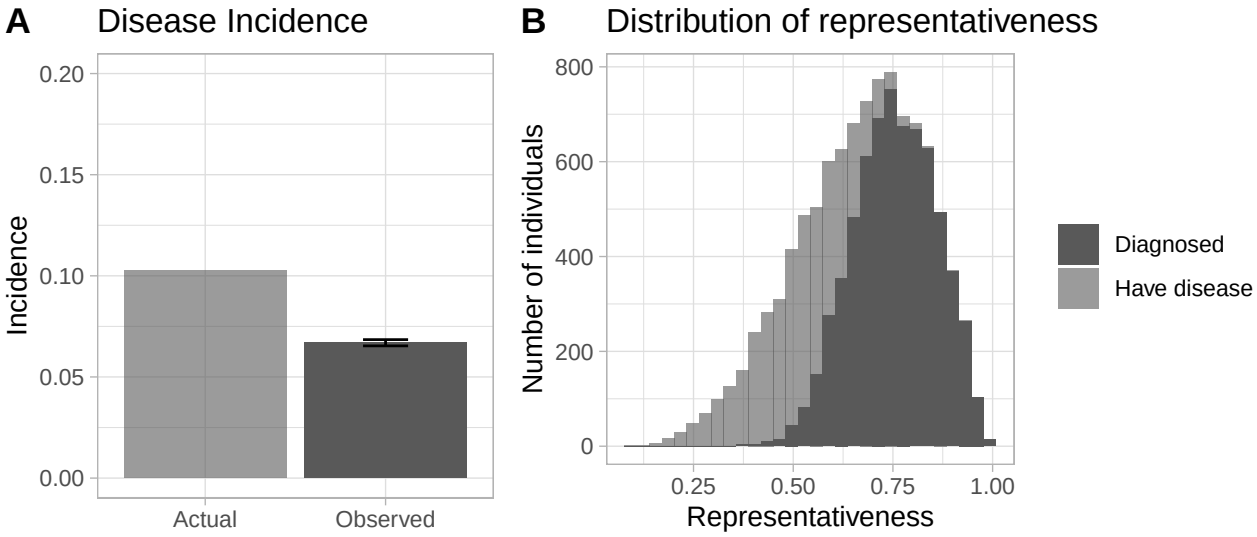


Results

First, we can see that follow-up based on high R values in this case does one thing well: individuals with the disease recieve a diagnostic followup much more often than those without. The table below shows that diagnostic followup is much more common for people who do have the underlying disease than those who don't.

disease	Rate of diagnostic followup
FALSE	0.0160009
TRUE	0.6525597

But what does the analyst learn about the incidence of this disease and the distribution of symptoms? The plot below shows the actual and observed disease rates and the actual vs observed distribution of R.

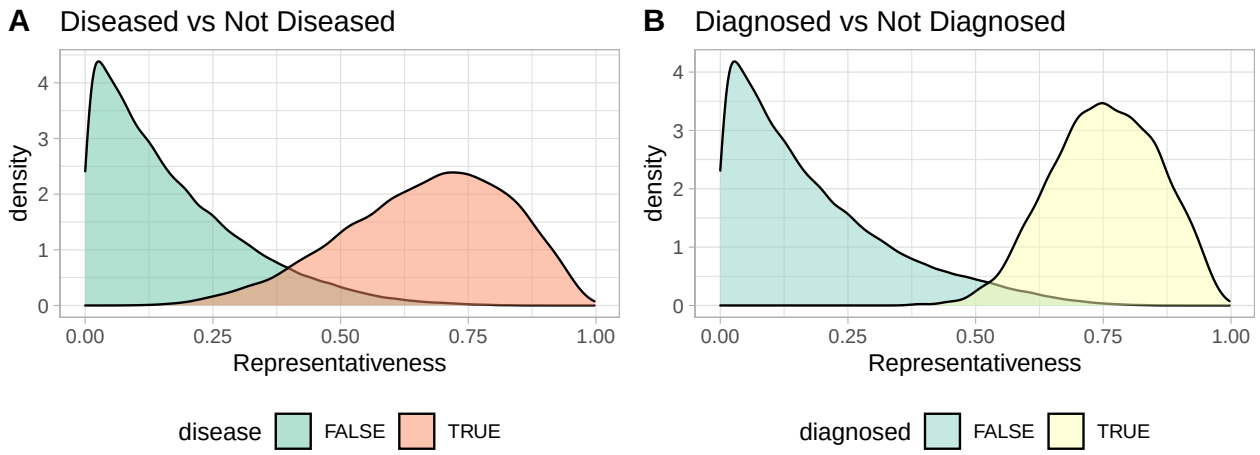


Takeaways

What does this mean? First, the analyst is going to underestimate the incidence of the disease. Second, clinicians are systematically missing cases with lower R values. This leads us to believe that the disease is rarer than in really is, and that the distribution of symptoms is different from what it really is. In particular, we are lead to believe that non-representative presenatations (low R) are less common than they really are.

What happens when we use this information to inform future rounds of diagnosis? First, if we believe that the incidence is lower than it really is, this might justify lower levels of diagnostic follow-up. If we believe that most cases tend to have higher values of R, this might further justify restricting testing to only those with a high value of R (very representative cases). In fact, we begin to believe that R does a better job of separating individuals than it actually does (see below). This justifies weighting R *more* heavily when considering whether to do diagnostic follow-up.

This force can act outside of the clinical setting as well. If the public believes this disease is very uncommon and that mild or atypical (low R) presentations of the disease are rare, they may not seek out the right clinical care when they have this disease (e.g. asymptomatic cases of COVID-19 early in the pandemic).



In a feedback loop, incorrect beliefs about this disease could not only be perpetuated but exacerbated. If the clinicians respond to the data by testing fewer people with low representativeness, then naive research on the resulting data will suggest that less-representative cases are more rare, justifying future clinical practice which denies follow-up for less representative cases. Likewise, a perception that this disease is rarer or more homogenous than it is might cause diseased people with a mild or atypical presentation to not seek the right clinical care. These people may go undiagnosed, leading to fewer less-representative cases recorded in clinical research. This could result in a highly stereotyped disease: a condition which clinical practice considers to occur only in "textbook" presentations, but which actually exhibits a wider array of presentations which go undiagnosed and unnoticed.

Example

This situation somewhat resembles the misunderstanding of COVID-19 in early 2020. Both the public and the clinical community did not realize that there were large numbers of mild or asymptomatic cases of COVID-19. This lead to the belief that COVID-19 was both rarer and more severe than it really was. This perpetuated an environment in which harder-to-identify cases went unnoticed because they did not recieve testing or follow-up. Only later studies with more widespread testing helped the community to realize this misconception.

Case 2: Follow-up based on representativeness, demographics (disease risk at parity):

Set Up

Disease model

We'll build upon the set-up in case 1 by adding a binary demographic characteristic, X_i:

$$X_i \sim_{iid} Bernoulli(0.5)$$

For now, we'll assume that X_i has no association with the underlying incidence of disease or how representative one's symptoms are. We'll think about what happens when those conditions are not true in cases 3 and 4.

Clinical Selection Model

Now let's consider that a person's probability of follow-up depends on not only the representativeness of their symptoms but their baseline characteristics:

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

Without loss of generality, let's assume that people with $X_i = 0$ are more likely to be recieve follow-up than people with $X_i = 1$. There are a couple reasons this could be

- 1. Doctors have a preconception that the disease is less common in people with $X_i = 1$.
- 2. People with $X_i = 1$ are less likely to see a doctor when they experience symptoms of this disease.

There is a *third* possibility:

- 3. People with $X_i = 1$ have a different presentation with the disease than people with $X_i = 0$.

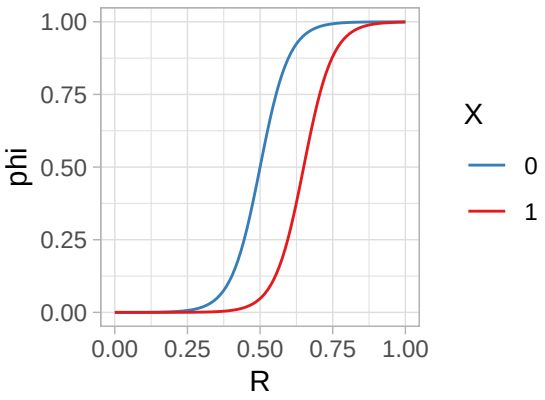
We'll think about this scenario more explicitly in case 4.

To model who gets follow-up, we'll use a modified version of the model in case 1, this time including the demographic characteristic, X:

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

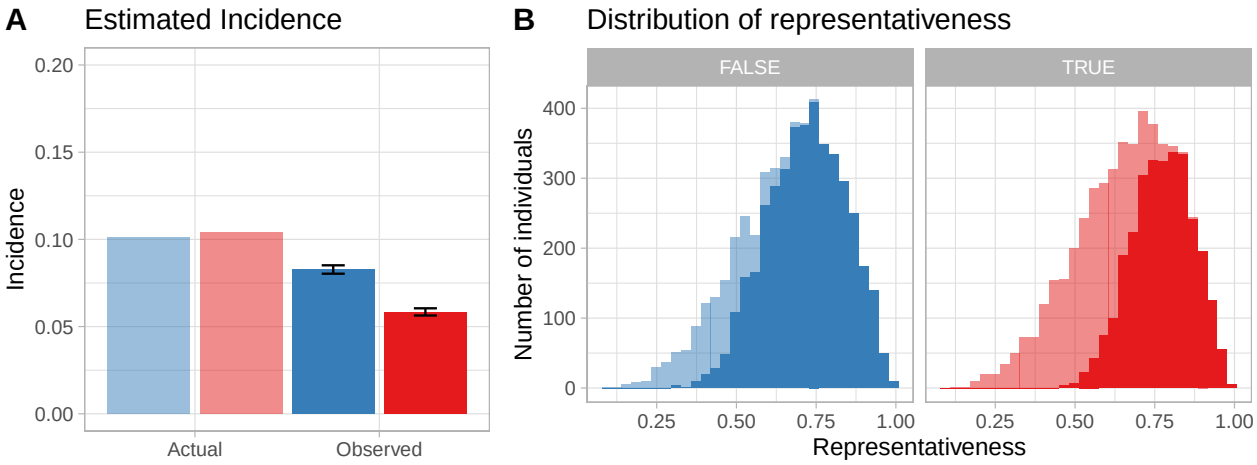
where again, $c_{x_i} = \frac{1}{1+exp(-(\beta_0+\beta_2 X_i))}$ is a corrective constant to ensure that $\phi(0, X_i) = 0$. Let Let $\beta_0 = -10$, $\beta_1 = 20$, and $\beta_2 = -3$.

Inessence, this means that people are selected for followup according to a sigmoid function, with people with $X_i = 1$ needing a higher representativeness to be selected with the same probability as people with $X_i = 0$, shown below:



Results

Now what does the researcher learn about incidence and presentation in these groups? The incidence and distribution of severities is shown below.



Takeaways

This scenario looks a lot like what we observe in Case 1, except that the distortion caused by treating diagnosis as equivalent to disease acts differently on the two groups. In both groups, incidence is underestimated, and the disease appears to contain more highly representative cases than it actually does. However, this distortion is stronger in the group which recieved follow-up less frequently. This leads to the perception that the disease is rarer in this group than it actually is, and that cases in that group are especially representative of the disease. One particularly important misconception that could result from these data is that the clinical community might come to believe that X is a risk factor for the disease, when in fact it is entirely unrelated to the disease state. In reality X is affects the probability of *diagnosis* with the disease, not the probability of having the disease itself.

How might clinicians or the public respond to these observations? They may be lead to believe that the disease is rare in people with $X_i = 1$, and that when these cases arise they are usually very strong classical presentations. This preconception could perpetuate a situation in which people in the $X_i = 1$ group are less likely to seek the correct clinical care when ill, and those who seek care will be less likely to be selected for follow-up - especially those whose presentations are mild or atypical (low R). This is very similar to the scenario in Case 1, except that it affects one group more severely than the other. In particular, one result

may be that people with $X_i = 1$ will only be diagnosed reliably if they have an especially severe or "textbook" presentation.

Examples

In this example, we suppose that the underlying rate of disease is at parity between groups, and clinical preconception or patient self-selection results in incorrect beliefs about disease incidence between groups. In practice, this is very hard to separate from a scenario in which there *is* some biological or socioeconomic reason that X_i is a risk factor for disease, but (we will see) this observed difference in disease risk is exacerbated by a data misunderstanding (Case 3). We'll discuss some examples in Case 3.

Case 3: Follow-up based on representativeness, demographics (disease risk differs)

Set up

Disease Model

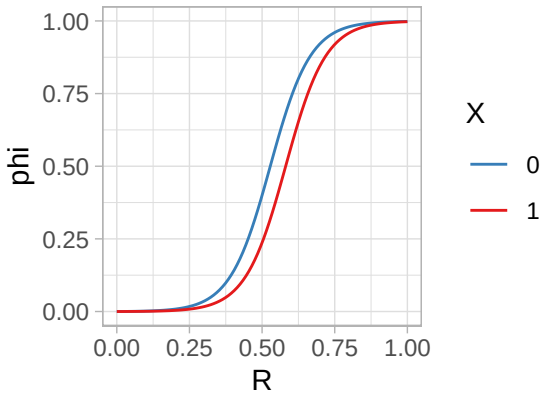
Now lets suppose there is an underlying difference in disease risk between groups. We'll suppose that the disease occurs in 15% of people with $X_i = 0$ and 7.5% of people with $X_i = 1$.

Clinical Selection Process

In order to determine how doctors diagnose, we'll start with the generous assumption that clinical practice is based in a very well-grounded understanding of the disease. Let's imagine that, before we begin, the research community has somehow compiled a dataset of 100,000 individuals with perfectly accurate diagnostic information (e.g. all 100,000 people have been assayed with a 100% accurate diagnostic test). We'll construct a logistic model based on that historical data and use it to inform future practice. Note that while a logistic model is relatively simple compared to potential alternatives from the machine learning field, it is built on perfectly accurate data.

The plot below depicts one such clinical model built using perfectly labeled data. One interesting feature to note is that even though a person's value of X is associated with potentially double the disease risk, X is weighted relatively weakly in the clinical model. This is because in this case the individual's symptoms still contain much more information about disease risk than their demographics. An individual with representativeness r and $X_i = 1$ is *not* in general twice as likely to have the disease than a person with the same representativeness and $X_i = 0$. The information about presentation in this case is much more informative for predicting disease than the demographics. This highlights another way in which this model is generous: demographic characteristics can often be overweighted in clinical decision-making due to cognitive biases, yet here we give an appropriately low weight to demographics.

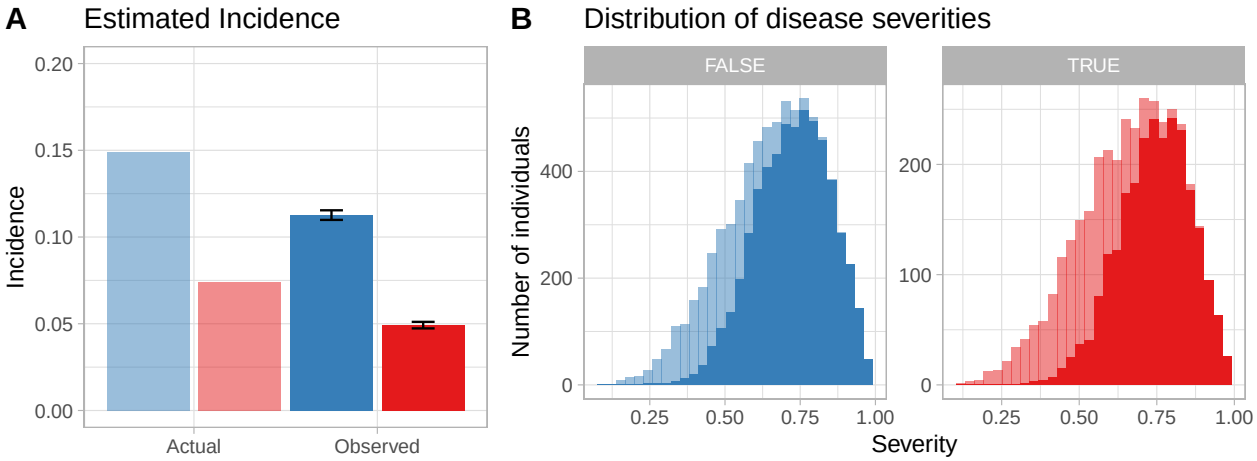
Note that we are also eliminating the potential for patient self-selection. This might be similar to supposing that *everyone* goes to the clinic regardless of their demographics or symptoms, so that the clinician is the only gateway determining who recieves follow-up. In reality, the selection process determining who seeks care is likely to further bias the observed data.



Results

This is the data that results when clinicians select individuals for follow-up based on the model above. Note that the pattern of observed data qualitatively is very similar to case 2 (no difference in disease risk between

groups). This is part of the reason it is difficult to separate these cases in practice without a different study design.



Remember that while clinicians are using a simple model to select people for follow-up (logistic), they are using a model that is built on perfectly labeled prospective data, and patients are not allowed to self-select whether they seek care. Arguably, this is high quality evidence-based practice.

The table below shows the rates of correct diagnosis between groups. Diseased individuals in the lower-risk group are less likely to be correctly diagnosed.

x	Proportion correctly diagnosed
FALSE	0.7560712
TRUE	0.6640709

The next table shows the mean representativeness of correctly diagnosed cases between groups. In order to be diagnosed, diseased individuals in the lower-risk group have to have more classical presentations on average than those in the higher-risk group. This may mean that people in the lower risk group need to have a more severe case of the disease in order to be diagnosed as consistently as an individual in the higher risk group.

x	Mean representativeness of diagnosed cases
FALSE	0.7242645
TRUE	0.7423318

Each of these patterns is a direct result of the clinical model shown above - one based on high quality data. Although these patterns could lead to concerning health disparities between these groups, there are those who would argue that this is an unfortunate but necessary side-effect of evidence-based medicine. They might make the case that allocating limited clinical resources towards the patients most likely to benefit (in this case, those at higher risk of disease) is part of optimizing the value of care. This becomes more of a discussion on fairness in allocating clinical resources, which is part of a large body of existing literature.

The point we want to make here is not about whether this practice is fair, but about the kind of data it produces and how that might be self-reinforced. Suppose our researcher wants to calculate the ratio of disease incidence between these groups. If they consider only the *diagnosed* people to have the disease, they underestimate the incidence in the high-risk group by -24.3928764166217%, but they would underestimate the incidence in the low-risk group by -33.5929108485499% (see below). As a result, the researcher might conclude that people with in the lower risk group have 44% lower disease risk than people in the high risk group, when the actual value is 50%. **This means the researcher with *overestimate* the difference**

in disease risk between groups. In particular, they will believe that being in the low-risk group is more protective against disease than it actually is!

x	Percent error in incidence estimate (%)
0	-24.4
1	-33.6

Takeaways

Even when doctors begin diagnosis with a well-grounded clinical model, the resulting data yeilds two misleading patterns:

- the difference in disease risk between groups is overestimated
- the observed (i.e. diagnosed) cases in the low-risk group will tend to have more classical presentations than those in the high risk group.

How might the public or the clinical community respond to this misleading information in the next round of diagnosis? Clinicians may be lead to believe that being in the low risk group is more protective than it really is, so they will be less likely than before to perform diagnostic follow-up on individuals in this group, and the individuals selected for follow-up will tend to be the more representative cases. Patients in the low-risk group may develop a perception that their disease risk is lower than it really is, making them less likely to seek the necessary care if they do in fact have the disease. Since only the most representative cases of the disease are correctly diagnosed in the low risk group, individuals in the low risk group may need more severe disease in order to recieve a correct diagnosis.

Over time in a feedback loop, these responses may not only self-perpetuate but self-exacerbate. An overestimate of the protective value of a demographic characteristic will lead people in that low-risk group to seek care less consistently and to be selected for follow-up less often when they do seek care, resulting in data which futher exibits an exaggerated difference between groups.

Examples

Arguably, this kind of phenomenon could affect *any* disease which is not studied in a prospective manner. If a risk factor for the disease also affects the likelihood that a person is correctly diagnosed, then observational studies which fail to account for the diagnostic process may tend to incorrectly estimate the role of that risk factor. This may apply to a variety of risk factors and diseases which are not studied in a prospective way, such as:

- Autism in male vs female children
- Lupus in men vs women
- Breast cancer in men vs women
- Melanoma in dark vs light skinned individuals
- Heart Attack in men vs women

Case 4: Follow-up based on representativeness, demographics (representativeness differs):

Earlier, I proposed three mechanisms that might cause individuals with certain characteristics to have different rates of success at being correctly diagnosed.

- 1. Doctors have a preconception that the disease is less common people with this characteristic
- 2. People with this characteristic are less likely to see a doctor when they experience symptoms of this disease.
- 3. People with this characteristic have a different presentation with the disease than others.

We've discussed mechanisms 1 and 2 in the previous two cases. This case will focus on mechanism 3.

Set up

Here, we'll consider a scenario in which diseased people from different groups have different common presentations. Perhaps people with $X_i = 1$ have symptoms which are milder, harder to identify, or more easily mistaken for being indicative of another disease. Perhaps common clinical understanding is centered on the group with $X_i = 0$, so that the symptoms of this disease that clinicians are taught to recognize as suggestive of this disease are the symptoms more common in the $X_i = 0$.

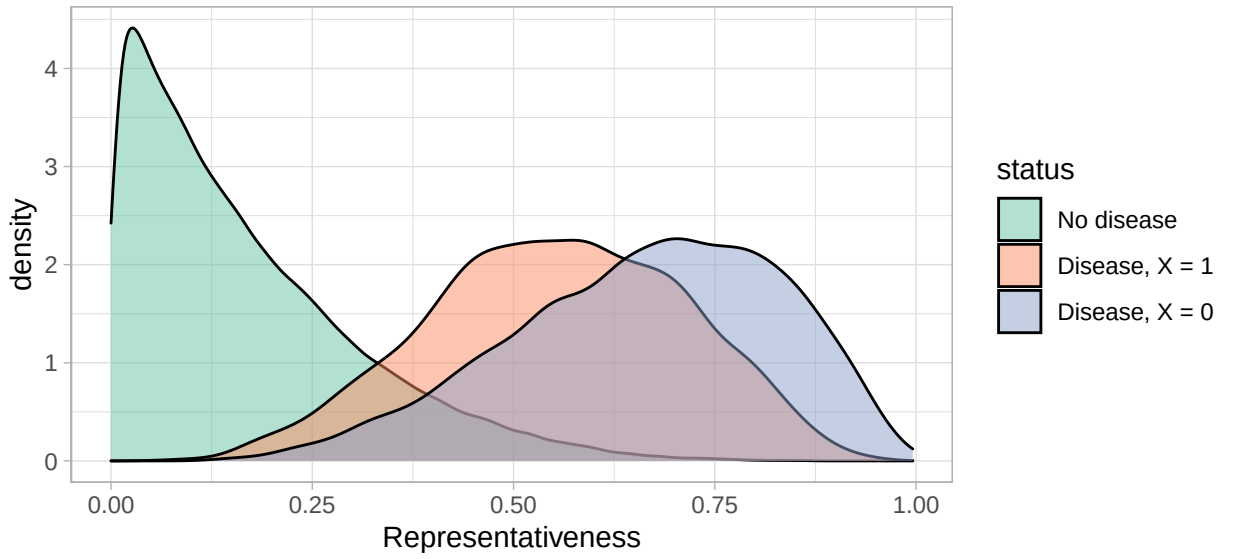
A particularly important example for this scenario is heart attacks in men and women. For many years, the clinical community (and the public) did not recognize that women often presented with different symptoms when they were having a heart attack. Additionally, these symptoms could be more subtle or less specific to heart attack, such as fainting, indigestion, and extreme fatigue. This has resulted in ongoing disparities in heart attack survival between men and women.

Disease Model

For simplicity, we'll remove the possibility that disease rates differ between groups, and we'll suppose that doctors don't believe there is any difference in disease incidence between groups, and they are unaware of or otherwise unable to respond to the difference in presentation between groups. In reality, it may be that a disease has different presentations *and* different incidence rates (or a *perceived* difference in incidence rates) between groups. This is certainly the case for heart attack, for example. We separate these phenomena here in order to characterize the ramifications of each one separately.

We'll build off the scenario in case 1 (group membership does not affect disease incidence, and doctors treat both groups equivalently during diagnosis), but now we'll add the nuance that individuals with $X_i = 1$ tend to have presentations that are thought to be less representative of the disease (lower R_i distribution). We'll use the following rules to simulate the distribution of R_i

$$\begin{aligned} R_i|Z_i = 1, X_i = 0 &\sim_{iid} \text{Beta}(5, 2.5) \\ R_i|Z_i = 1, X_i = 1 &\sim_{iid} \text{Beta}(5, 2) \\ R_i|Z_i = 0 &\sim_{iid} \text{Beta}(1, 5) \end{aligned}$$

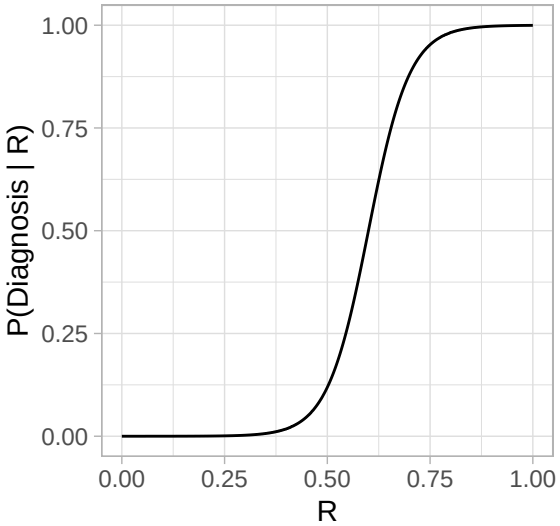


Clinical Diagnostic Model

For simplicity in this case, we'll begin by supposing that doctors are unaware of the difference in presentation between groups, so they diagnose based on representativeness alone, as in Case 1. In specific, they will use the same model as in Case 1 for deciding whom to select for diagnostic followup:

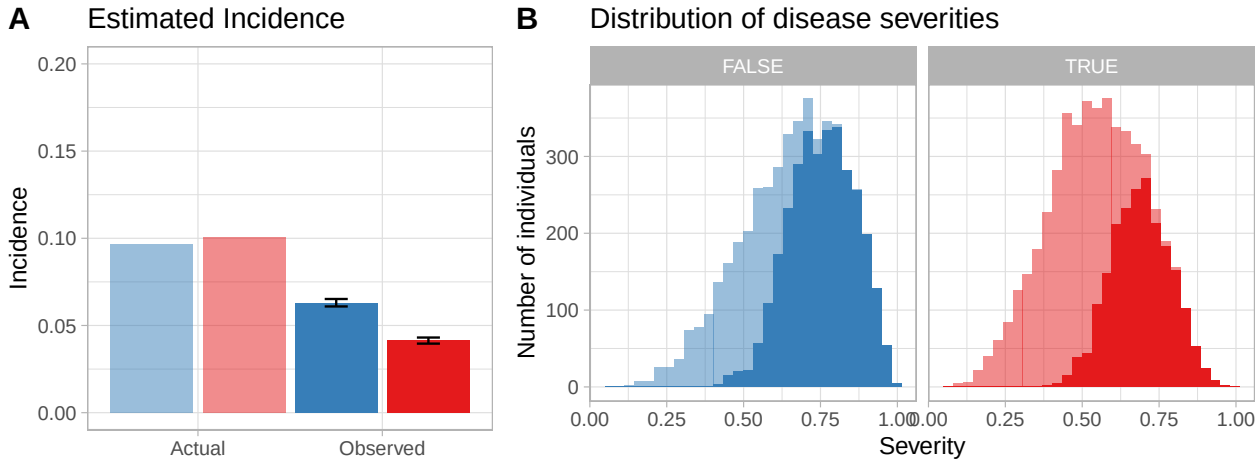
$$\phi(R_i) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 R_i))} - c$$

As before, $c = \frac{1}{1+exp(-\beta_0)}$ is a corrective constant to ensure that $\phi(0) = 0$. Let $\beta_1 = 20$ and $\beta_0 = -12$. This gives the following probability of diagnosis as a function of representativeness.

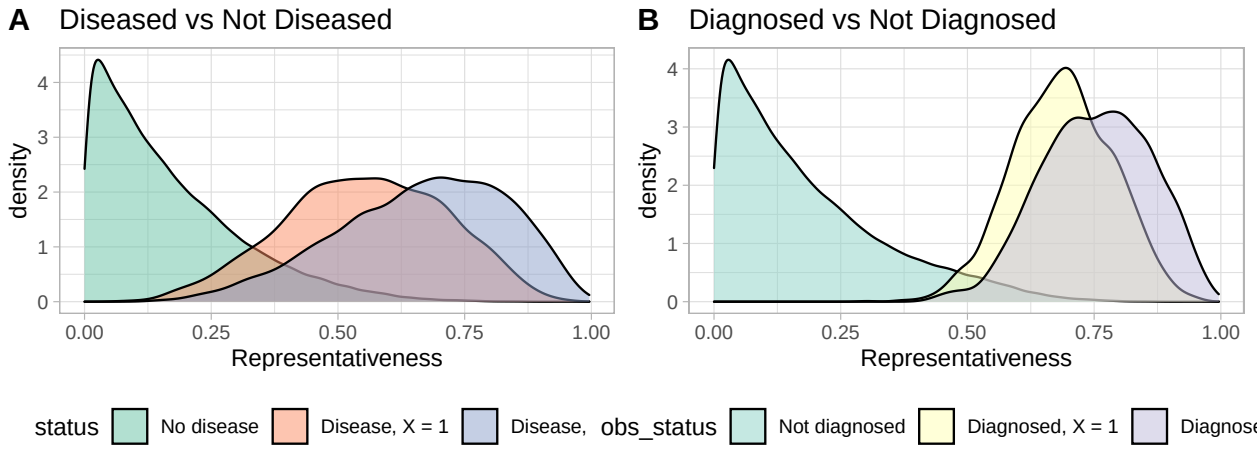


Results

The plots below show the estimated and actual incidence (A) and the estimated and actual disease severities. These visuals show a similar story to what we have seen before.



These plots show the distribution of representativeness among individuals with and without the disease in each group, and individuals with and without diagnosis.



Takeaways

These results paint a similar picture to what we've seen in other cases, but with some modifications. First, even though doctors are not explicitly using different rules to diagnose people from different groups, they have a lower rate of successful diagnosis for people with in the group with less representative symptoms ($X_i = 1$). This differential rate of missed diagnosis results in a more severe underestimate of disease prevalence in these groups. Second, the individuals in both groups who *do* tend to be diagnosed successfully tend to be those whose symptoms are thought to be more representative of the disease (high R_i). This results in the illution that R is a better separator of those with and without the disease than it actually is. This might lead to an important misconception: rather than identifying that people in one group tend to have different presentation than the other, we might instead conclude that both groups have similar presentations, and the disease is simply less common in the group with the less-well-understood presentation.

This self-reinforces in a similar way to other results we've seen. We unintentionally produce data which obscures the differences in presentation between groups and instead makes it appear that there is a strong difference in disease risk between groups. This might clinicians to only follow up on patients in the percieved low-risk group if they have particularly representative symptoms (similar to case 2). *This is particularly unfortunate because it is exactly the opposite of what someone might suggest if they knew the whole story:* Instead, it might be the case that we should be *more* open to following up on someone with low R if they

are in the $X_i = 1$ group, because many diseased people in this group will have presentations that are less representative of the textbook case!

Example

The case of heart attacks in women is an illustration of how different selective mechanisms can interact to cause confusion. In the beginning, doctors were taught to identify the male presentation for heart attack. The medical community and the general public had a preconception that women had a much lower risk of heart attack than they actually did (Cases 2 and 3), and they were ignorant of the differences in presentation between men and women. This presented a self-reinforcing problem inside and outside of the hospital. Outside the healthcare system, women were not given the tools to recognize their own risk of heart attack or to identify the signs that they were having a heart attack. When women with heart attacks did interact with clinical professionals, they were less likely to get appropriate care. Moreover, it is likely that the women who *were* identified as having a heart attack were those whose presentations were similar to the male heart attack presentation of clinical cannon, reinforcing clinical ignorance of common female heart attack symptoms[2]

[2]is there any evidence of this? would be great to have a citation, but I'm guessing this wasn't studied very carefully at the time.