Scenarios for feedback loop failure in clinical diagnosis

Keywords: Feedback loop failure, diagnosis, selection bias, feedback loops

Abstract: To a patient, medical diagnosis can be the gateway to essential care. Across a population, the diagnostic process shapes medical research from observational studies to bench science. But the diagnostic process is a selective one. It can depend on a person’s symptoms, their willingness and ability to seek care, and their clinician's judgment synthesizing medical history, risk factors, and presentation with an evolving evidence base. A *feedback loop failure* may occur if misleading “evidence” about disease etiology encourages systematic errors in the diagnostic process in a way that self-perpetuates or self-exacerbates, compromising patient care. This paper defines scenarios for feedback loop failure in medical diagnosis. Through simulated case-studies, we characterize how disease incidence, presentation, and risk factors may be misunderstood when observational data are summarized naive to the diagnostic selection process. We conclude that direct summaries based on diagnosed individuals may be misleading, especially concerning those symptoms and risk factors that influence the diagnostic process itself.

Figures: 6

Tables: 4

Article type: Original Contribution

For a patient, diagnosis can be the gateway to essential care. Across populations, the diagnostic process critically shapes medical research, from epidemiologic studies, to clinical trials, to bench science on biomedical samples. An often-implicit feedback loop connects medical practice to research: studies of diagnosed cases shape our understanding of disease, which then shapes the diagnostic process (Figure 1). While this feedback is intended to advance medical care, even well-intentioned data-driven practices can fail through circular reasoning: selective real-world processes naively summarized into erroneous “evidence” in turn shaping real-world processes.

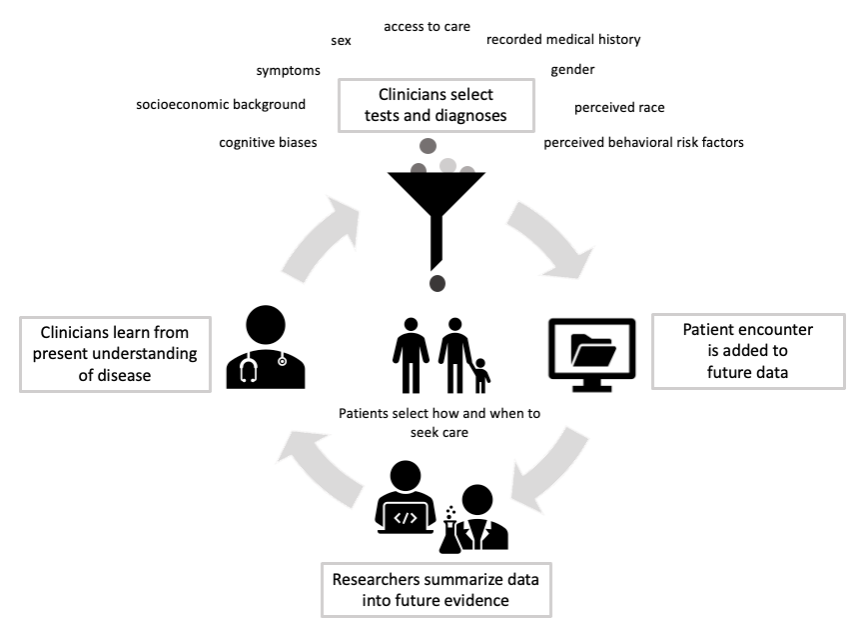
Data-driven feedback loops iteratively estimate some information or *target parameters* (e.g. associations, incidences, or a predictive model) that inform future practice. The feedback mechanism may be explicit – as in an updating predictive model that directly shapes practice – or implicit – as in an evolving evidence base informing human decision-makers (Table 1). We use the term *feedback loop failure* to refer to a scenario in which a data-driven feedback loop gives rise to self-reinforcing or exacerbating patterns of undesirable behavior. Failures may arise because target parameters were incorrectly estimated, because the estimated information is misused, or because the system was built on an incorrect framework altogether.

A learning health system reflects an idealized medical diagnosis feedback loop (Figure 1). Past diagnoses shape the data used in future medical research, which shapes future diagnostic decisions. However, the diagnostic process is selective. Diagnosis depends on the specifics of a person’s symptoms, the individual’s willingness and ability to seek care, and the clinician's judgment synthesizing medical history, risk factors, and presentation. Because of this, direct summaries based on diagnosed individuals may be misleading, especially concerning those symptoms and risk factors that influence the diagnostic process itself. A feedback loop failure may occur if misleading “evidence” about disease etiology encourages systematic errors in medical practice in a way that self-perpetuates or self-exacerbates, compromising patient care.

This paper defines scenarios for self-reinforcing feedback loop failure in medical diagnosis. First, we characterize through simulated illustrations how disease incidence, presentation, and risk factors may be misunderstood when observational data are summarized naive to the diagnostic selection process, and we suggest how those misunderstandings may self-reinforce. Next, we illustrate the impacts of disproportionate underdiagnosis on patient care. Finally, we suggest detection and mitigation strategies.

**Table 1:** Example systems in which feedback loop failure might occur.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Domain** | **Target Parameter(s)** | **Issues contributing to failure** | **Feedback mechanism** | **Consequences** | **Related work** |
| Predictive policing | Localized crime rates | *Selection bias, misclassification bias:* Crime is discovered more often in neighborhoods with more police presence. | Rates of discovered crime inform police presence, which shapes discovered crime rates. | Neighborhoods are subject to disproportionate police presence. | [1,2](https://paperpile.com/c/X7Xn59/FQbVH+5uJE1) |
| Ranking and recommender systems (e.g. media, news) | Popularity ranking model | *Incorrect framework, ‘self-fulfilling prophecy’*: Item popularity may not capture “quality” or “benefit.” Recommendations influence future popularity. | Popularity of an item determines rank or recommendation probability, influencing popularity. | Low-value popular items increase in popularity; High-value unpopular items remain obscure. A “filter bubble” effect[3](https://paperpile.com/c/X7Xn59/s2QIK) may arise. | [4–7](https://paperpile.com/c/X7Xn59/n4zFx+PGyks+cdSP4+agHjv) |
| Risk prediction for medical triage | Predictive model of prognosis | *Incorrect framework,*  *‘self-fulfilling prophecy’:* A causal framework for the effect of care is more appropriate. Predictions directly influence outcomes predicted. | Estimated prognosis determines who is triaged away from care, affecting prognosis. | Individuals with overestimated risk are triaged away from potentially life-saving care. | [8–10](https://paperpile.com/c/X7Xn59/9sb4v+y5JaJ+7d8Lq) |
| Diagnosis | Disease symptoms and risk factors | *Misclassification bias, selection bias:* Diagnosed vs. actual disease cases are not equivalent. | Diagnosed cases shape understanding of prevalence, symptoms, and risk factors used in the diagnostic process. | Disproportionate overdiagnosis or underdiagnosis, with a misunderstanding of disease etiology | This paper. Examples in heart attack[11](https://paperpile.com/c/X7Xn59/w2BED) and autism[12,13](https://paperpile.com/c/X7Xn59/IqmRA+MgSuF) |



**Figure 1** - A schematic of a “learning health care system” feedback loop. Providers synthesize information about a patient with the evidence base to make decisions about testing, specialist referral, and diagnoses. These decisions in turn shape the evidence base. Meanwhile, people select how and when to seek care. Failure to account for a complex and potentially selective diagnostic process can give rise to misunderstandings of disease risk and etiology, to the detriment of scientific knowledge and medical care. In a feedback loop, these misunderstandings may self-reinforce or self-exacerbate.

# Methods

This work begins with simulated case-studies illustrating mechanisms for feedback loop failure in an observational study of a disease. In each case, a researcher selects a cohort of 100,000 individuals and examines newly-diagnosed cases of the disease within a set time period. They may ask: What is the population incidence of the disease? Is a risk factor (i.e. a pre-existing condition associated with the underlying presence of the disease)? What symptoms are representative of people with the disease? Each case considers a “naive analysis,” that fails to account for the diagnostic processes shaping the data. The naive analysis assumes that all diagnosed people have the disease and undiagnosed people do not.

As a simplified model, we suppose a diagnostic evaluation for this disease exists (e.g. a radiological scan, laboratory test, specialist evaluation, or physical examination based on well-defined criteria), but to access that evaluation, a person must appear in the clinic and be selected for workup (e.g. by a primary care provider).

* **Case 1 [*Symptoms-only*]:** Selection for evaluation depends on *,* a summary of how representative a person’s symptoms are of this disease.
* **Case 2 [*Symptoms and background*]:** Selection for evaluation depends on symptom representativeness, , and a background characteristic (e.g. sex, gender, perceived race).
  + **Case 2A [*False risk factor*]:** *is not* associated with disease risk.
  + **Case 2B [*True risk factor*]:** *is* associated with disease risk.
* **Case 3 [*Divergent symptoms*]:** Selection for evaluation depends only on symptom representativeness, , but determines the distribution of *.*

In the second part of this study, we’ll shift to a longitudinal perspective, examining how differential rates of diagnosis affect the timing of diagnosis between groups of people as their diseases progress.

The following subsections summarize the simulation disease models and clinical selection processes, with additional detail in the supplement. All figures of individual simulation results were selected as representative of a typical simulation run. To summarize typical behavior for the *symptoms and background* subcases in Table 2, we calculated the mean and variance of each quantity across 1000 simulations. All code is available on github (https://github.com/[url removed for blinding]).

## Disease model

Each simulated individual is characterized by an underlying disease state (true presence or absence of the disease), a summary of how representative their symptoms are of a “classic presentation” of the disease (denoted by ), and (except in the *symptoms-only* case) a background characteristic (denoted by ). varies continuously from 0 to 1, with a value of 1representing a “textbook” presentation and 0signifying no resemblance to a “textbook” presentation for this disease.

In the *symptoms-only* case, approximately 10% of the cohort develops the disease. is simulated from a different beta distribution depending on the individual’s underlying disease state (Figure 2), with the parameters selected so that individuals *with* the disease tend towards higher values of *,* and those *without* the disease tend towards lower values, with some exceptions on both sides (e.g., disease cases with mild or atypical symptoms, and non-cases with look-alike symptoms arising from some other condition Figure 2A).

The remaining cases additionally include the background characteristic, . For simplicity in our simulations, is a binary characteristic equal to 1 in roughly half the population. In the *false risk factor* case, is *not associated with the underlying disease state*: Individuals with different values of have the same probability of having the disease, and they have the same distributions of conditional on their disease state (Figure 2A). In the *true risk factor* case, having *doubles* disease risk compared to having (Figure 2A). The disease arises in 15% of individuals with and only 7.5% of those with .

In the *divergent symptoms* case, is not a risk factor, but individuals in different groups have different common presentations (different distributions of ). Individuals with tend to have lower , signifying symptoms that are milder, harder to identify, or simply less-well recognized as indicators of this disease (Figure 2A).

The longitudinal simulations consider a disease that progresses over time. In this setting we additionally model the underlying severity of the condition, summarized by , which varies continuously from 0 to 1. Over time, an individual’s disease shifts from no severity (mild or asymptomatic, ) to high severity (highly severe disease, ), following a sigmoid curve. Each individual has a unique disease trajectory with timing and rate of progression determined by randomly generated parameters. These trajectory parameters do not differ systematically based on . Here, representativeness, , is simulated as a noisy version of .

## Diagnostic Process Model

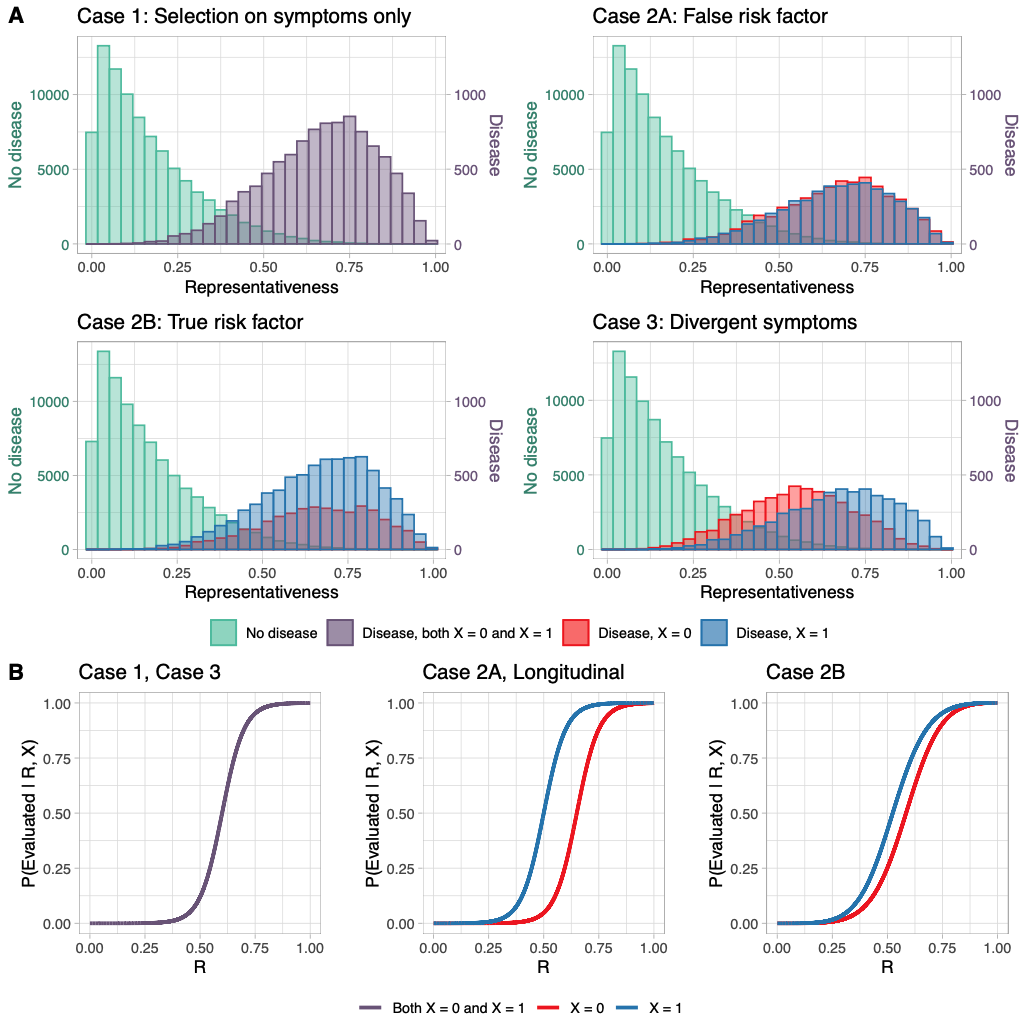
In this model, patients must be selected for diagnostic evaluation before they can be diagnosed. For simplicity, the evaluation is perfectly accurate: all individuals who are evaluated receive a correct diagnosis. However, all individuals who are not evaluated remain undiagnosed, even if they have the disease. A diagnostic test with imperfect sensitivity or specificity is considered as an extension to the *symptoms and background* cases (Case 2).

In each case, a person’s probability of selection for evaluation increases with the representativeness of their symptoms, , potentially with adjustment based on the background characteristic, (Figure 2B). Except in the *true risk factor* case, this is done using a parametrized sigmoid curve (additional simulation detail, Supplemental Digital Content). Our goal in the *true risk factor* case is to understand how data from diagnosed cases may be misleading even when selection for diagnostic evaluation is consistent with the true underlying differences in disease risk between groups. To this end, we suppose the probability of evaluation for a person with background and representativeness is the underlying conditional probability that a person with those characteristics has the disease:

,

where the right hand side is calculated directly from the *true risk factor* disease model (Figure 2B, additional simulation detail in Supplemental Digital Content). To characterize the ramifications of an imperfect test, we ran batches of 1,000 simulations of the *symptoms and background* cases in which the evaluation had some nonzero probability of false positives or false negatives.

The clinical selection process in the longitudinal simulation is identical in form to that in the *false risk factor* case (Figure 2B), except that individuals have an opportunity to be selected for diagnostic evaluation *at each timepoint* based on their background characteristic and the representativeness of their symptoms.



**Figure 2:** Illustrations of the simulation set-up. Explicit formulas and parameters are given in the additional simulation details in the supplemental digital content. (A) Histograms of representativeness, , for a typical simulation from the disease models for each case study. Note that because there are many more individuals without the disease than with the disease, the histograms for individuals with the disease (purple, blue, or red) are plotted using a different y-axis scale (right y-axis) than those for individuals without the disease (green, left y-axis). (A, top left) In the *symptoms-only* case, individuals with the disease tend to have more representative symptoms than those without. (A, top right) In the *false risk factor* case, the disease model is identical to the *symptoms-only* case, except that there is an additional demographic characteristic, , unassociated with disease state or representativeness. (A, bottom left) In the *true risk factor* case, individuals with have double the disease risk of those with . (A, bottom right) In the *divergent symptoms* case, individuals with have less representative symptoms but equivalent disease risk. (B) Plots of the probabilistic functions defining how patients are selected for diagnostic evaluation (B, left) For the *symptoms-only* and *divergent symptoms* cases, probability of evaluation increases with representativeness of symptoms, and does not depend on . (B, center) In the *false risk factor* case and the longitudinal simulations, individuals in the group have a higher probability of being selected for evaluation than those who are not. (B, right) In the *true risk factor* case, the probability a person is selected for evaluation is equal to the true conditional probability of disease given their values of and .

# Results

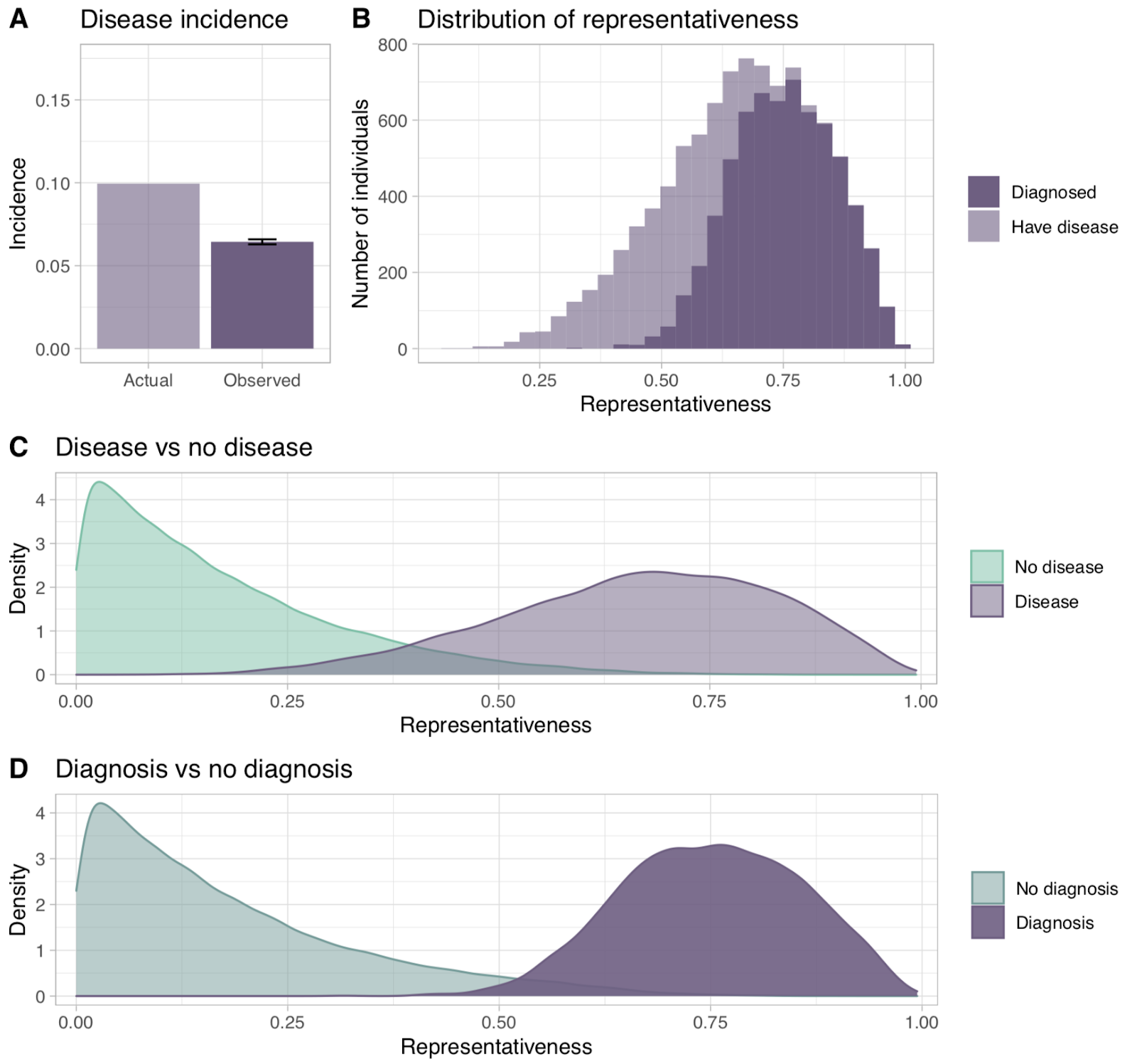
## Case 1: Selection on Symptoms Only

In this scenario, individuals are selected for evaluation based solely on how representative their symptoms are thought to be of the disease (Figure 2B). Figure 3 shows the findings from a naive analysis. The evaluation selection process means that disease cases that are less representative (lower ) are less likely to be diagnosed. Ignoring this by considering diagnosis and actual disease to be equivalent yields two primary misunderstandings:

* The disease appears less common than it really is (Figure 3A).
* Less-representative cases appear less common than they really are (Figure 3B).

Additionally, representativeness appears to do a better job at separating individuals with and without the disease than it really does (Figure 3C/3D). Moreover, the large sample size gives statistical certainty (i.e. narrow confidence intervals) to a highly *biased* incidence rate estimate (Figure 3A).

These misunderstandings could be perpetuated or exacerbated in a feedback loop. The more strongly the disease is believed to be rare and homogenous, the more difficult it will become for “atypical” cases to be detected. This could result in a highly stereotyped disease: one that clinicians believe arises only as "textbook" presentations, allowing an array of “atypical” disease presentations to go undetected.

****

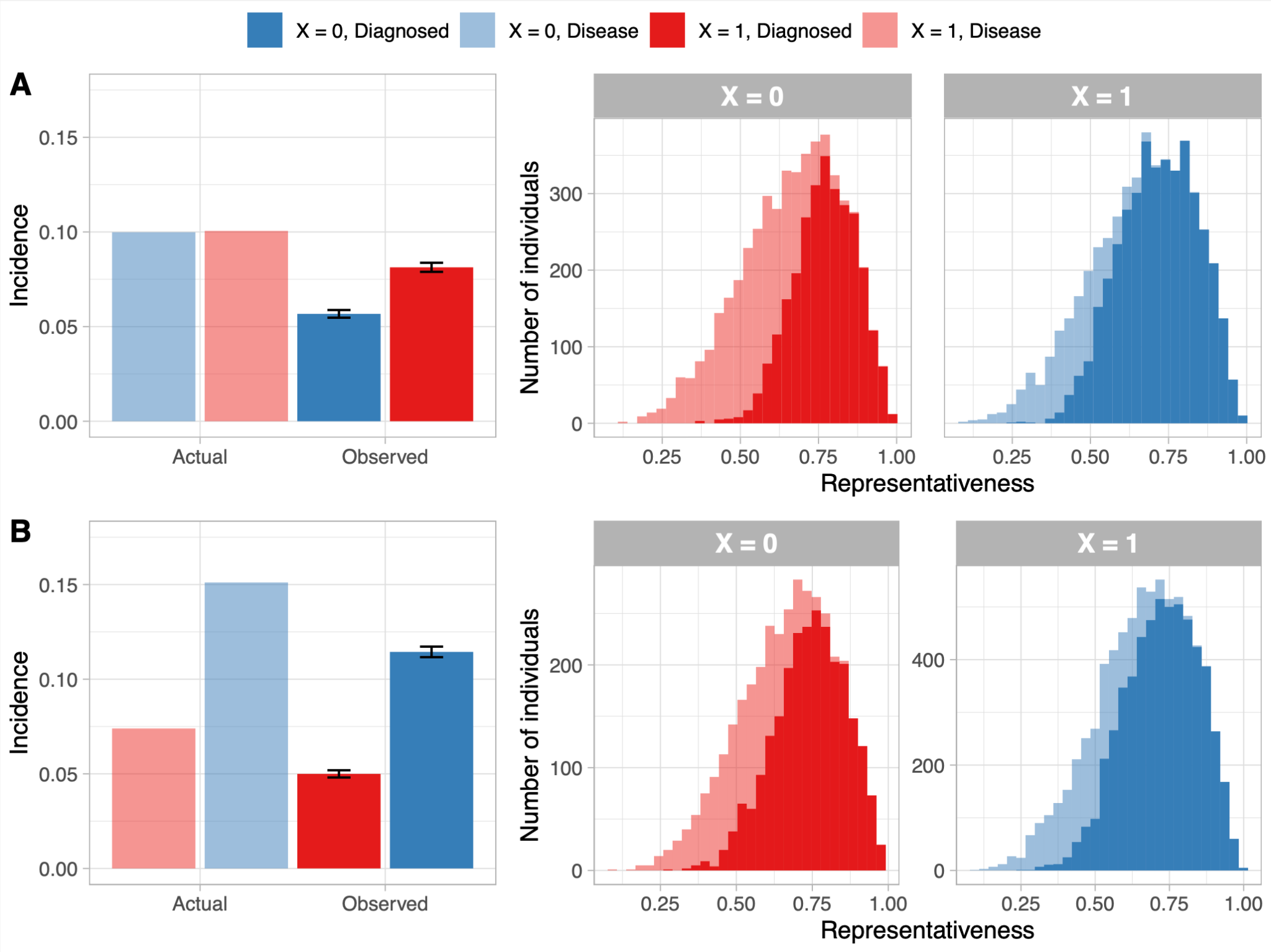
**Figure 3:** Ground truth disease characteristics versus a naive analysis in the *symptoms-only* case. (A) The naive analysis tends to underestimate disease incidence, yet large sample sizes result in high statistical certainty (black bar represents 95% confidence interval) (B) Successfully diagnosed cases tend to be those in which the symptoms were more representative of how the disease is thought to typically present. (C) True distributions of symptom representativeness. (D) The distributions of representativeness in diagnosed and undiagnosed individuals. When diagnosis is treated as equivalent to underlying disease, representativeness appears to be a better separator of diseased and non-diseased individuals than it really is.

## Case 2: Selection on Symptoms and Background

What happens when the diagnostic process also depends on a background characteristic, (e.g. sex, gender, perceived race)? We first consider the consequences when is *not* a true risk factor for the disease (Case 2A). Then consider the consequences when is a risk factor (Case 2B).

### Case 2A: False risk factor

Here, affects the probability that a person is evaluated, but has no association with actual disease risk (Figure 2). This might arise from differences in care-seeking behavior across patient groups or because is *incorrectly perceived* as a risk factor. Figure 4A shows findings from a naive analysis.



**Figure 4:** Ground truth disease characteristics versus a naive analysis in the *false risk factor* and *true risk factor* cases. (A, right) In the *false risk factor* case, the naive analysis tends to underestimate disease incidence more severely in the group that is less likely to be diagnostically evaluated (the group, red). This causes to falsely appear as a risk factor for the disease. (A, Left) Stronger selective pressure on the group means that disease cases must be more representative (and perhaps more severe) to be diagnosed. (B, right) In the *true risk factor* case, underlying disease incidences differ between groups, but the naive analysis tends to underestimate disease incidence more severely in the group (also see Table 2). (B, left) As in the *false risk factor* case, stronger selective pressure on the group means that disease cases must be more representative (and perhaps more severe) to be diagnosed.

Within groups defined by , the analysis yields misleading conclusions analogous to those in the *symptoms-only* case: incidence is underestimated, and less-representative cases appear less common than they really are. Moreover, individuals with the disease in the group are less likely to be selected for evaluation (Figure 2B), and thus disproportionately less likely to be diagnosed with the disease (Table 2). This means that appears to be associated with the disease even though it is not (Figure 4A, left), and diagnosed cases among those with appear to have more “textbook” representative symptoms than those with (Figure 4A, right). Thus is mistaken as a risk factor for the disease when it actually affects the probability of *evaluation for* *diagnosis* rather than the disease itself. As before, the large sample sizes lead to highly statistically significant *incorrect* results.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Percent of disease cases correctly diagnosed | | Naive estimate of cumulative incidence | | Naive estimate of relative risk |
|  |  |  |  |  | ( vs ) |
| Case 2A:  *False risk factor* | 56%  (+/-0.7% ) | 81%  (+/- 0.6%) | 5.6%  (+/-0.1% )  *Truth: 10%* | 8.1%  (+/-0.1% )  *Truth: 10%* | 1.45  (+/- 0.03)  *Truth: 1* |
| Case 2B:  *True risk factor* | 67%  (+/-0.8% ) | 75%  (+/- 0.5%) | 5.0%  (+/- 0.1%)  *Truth: 7.5%* | 11%  (+/- 0.1%)  *Truth 15%* | 2.26  (+/- 0.05)  *Truth 2* |

**Table 2:** Naive estimates of cumulative disease incidence and relative risk in Case 2A and Case 2B scenarios. Values are presented as mean (+/- standard deviation) across 1,000 simulations. Naive use of diagnosis rates as disease rates in observational data can result in highly confident, but incorrect, estimates of relative disease risk across groups. This may reinforce or exacerbate disparities in diagnosis rates.

These summary statistics are consistent with incorrect conclusions that the disease is rarer in people with and that disease cases in this group will have “classical” presentations. In a feedback loop, this could perpetuate and exacerbate a situation wherein people from the group are less likely to seek care when ill and less likely to be selected for evaluation, especially when their presentations are mild or atypical.

### Case 2B: True risk factor

If *is* truly associated with disease risk, similar patterns still arise, even if the evaluation selection process is calibrated with underlying disease risk.

In this setting, the probability of evaluation in the *true risk factor* case is the true conditional probability of disease based on and (Methods). The shape of this curve (Figure 2B) shows that although is associated with *double* the disease risk, a patient with symptom representativeness and does *not* in general have twice the likelihood of having the disease as a person with the same representativeness and X = 0 (Figure S1, Supplemental Digital Content, additional plots of Case2B diagnostic selection process). Failure to grasp this counterintuitive point may lead to cognitive error overweighting the importance of X in judging disease likelihood.

The naive analysis results are qualitatively similar to those from the *false risk factor* case. In both groups, disease incidence is underestimated (Figure 4B, right), and less representative cases appear less common than reality (Figure 4B, left). These biases in the data are stronger for group. As a result, the incidence is *underestimated* more in the low-risk () group relative to the other, causing to appear *more strongly* associated with disease than it really is (Table 2). In real world applications, Cases 2A and 2B are difficult to distinguish between. Without adjusting for the diagnostic process, apparent differences in disease incidence between groups may arise from true differences in risk, an artifact of selective diagnosis, or a combination of the two.

Again, this pattern may be self-reinforcing or -exacerbating. The more heavily is weighted in the diagnostic process, the harder it becomes for a disease case in the groups to be diagnosed, and the wider the apparent gap will be in disease risk between groups.

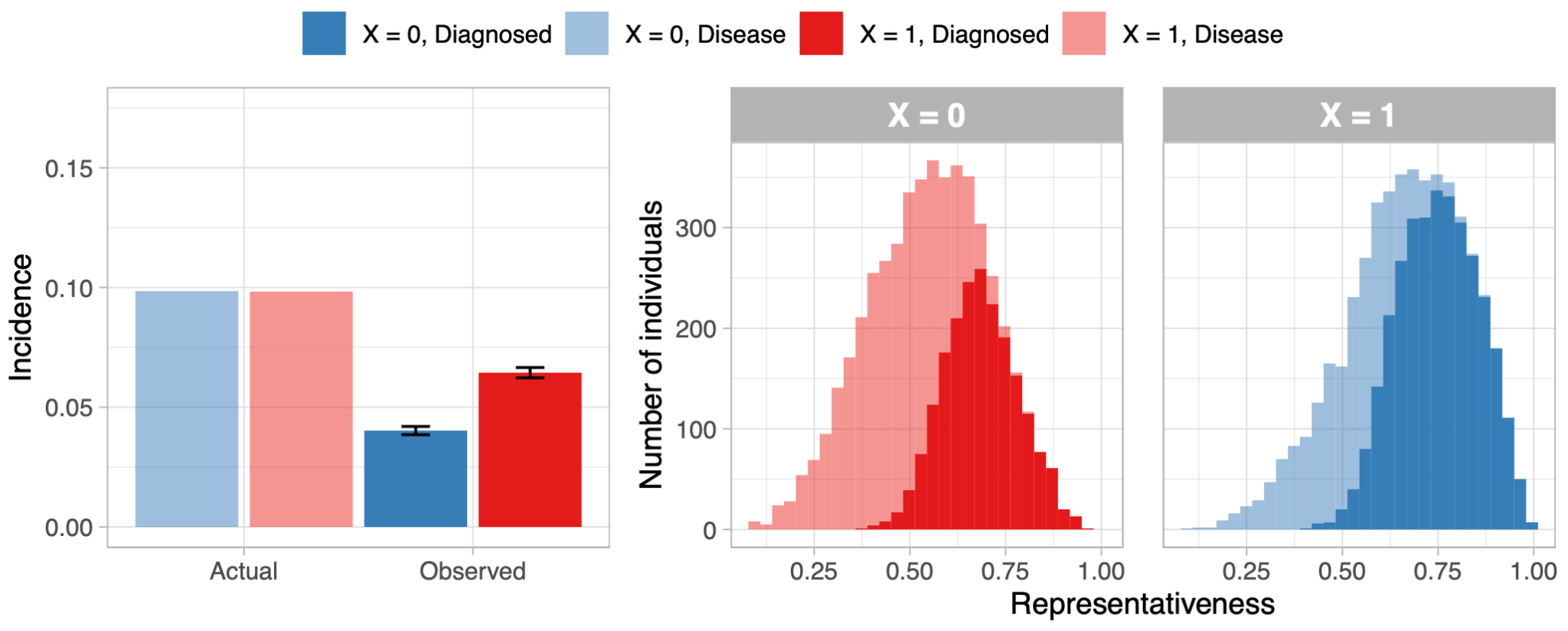
### An imperfect diagnostic evaluation

If the diagnostic evaluations are not perfectly accurate, the situation becomes more complex. However the primary takeaways tend to be robust. As in Table 2, incidences are underestimated and relative risk for the group is overestimated in both the *symptoms and background* sub-cases if either the evaluation sensitivity or specificity (or both) are reduced. The sensitivity and specificity of the evaluation had different implications for a naive analysis depending on the simulation set-up. In some settings, low sensitivity or specificity exacerbated bias in the estimated incidences or relative risk. In others, the misclassification bias from the imperfect evaluation *counteracted* some selection bias from clinical selection on whom to evaluate, although the overall direction of the bias in relative risk estimates remained the same (Figure S2 and Tables S1 and S2, Supplemental Digital Content, summarizing the results from these additional simulations).

## Case 3: Divergent symptoms

Here, we suppose that individuals with tend to have presentations that are thought to be less representative of the disease (Figure 2A). People in this group might have symptoms that are milder, subtler, or more easily mistaken as signs of another disease. Alternatively, common clinical understanding may be centered on the group, so that clinicians and patients are taught to recognize only the symptoms common to that group. A salient example is the diagnosis of heart attack, since sex-related differences in presenting symptoms[14,15](https://paperpile.com/c/X7Xn59/Hz36r+hXp3H) (and low awareness of these differences) has likely played a role in missed and delayed diagnoses in female patients.[11,16–18](https://paperpile.com/c/X7Xn59/xSFBT+IED6f+abaTl+w2BED)

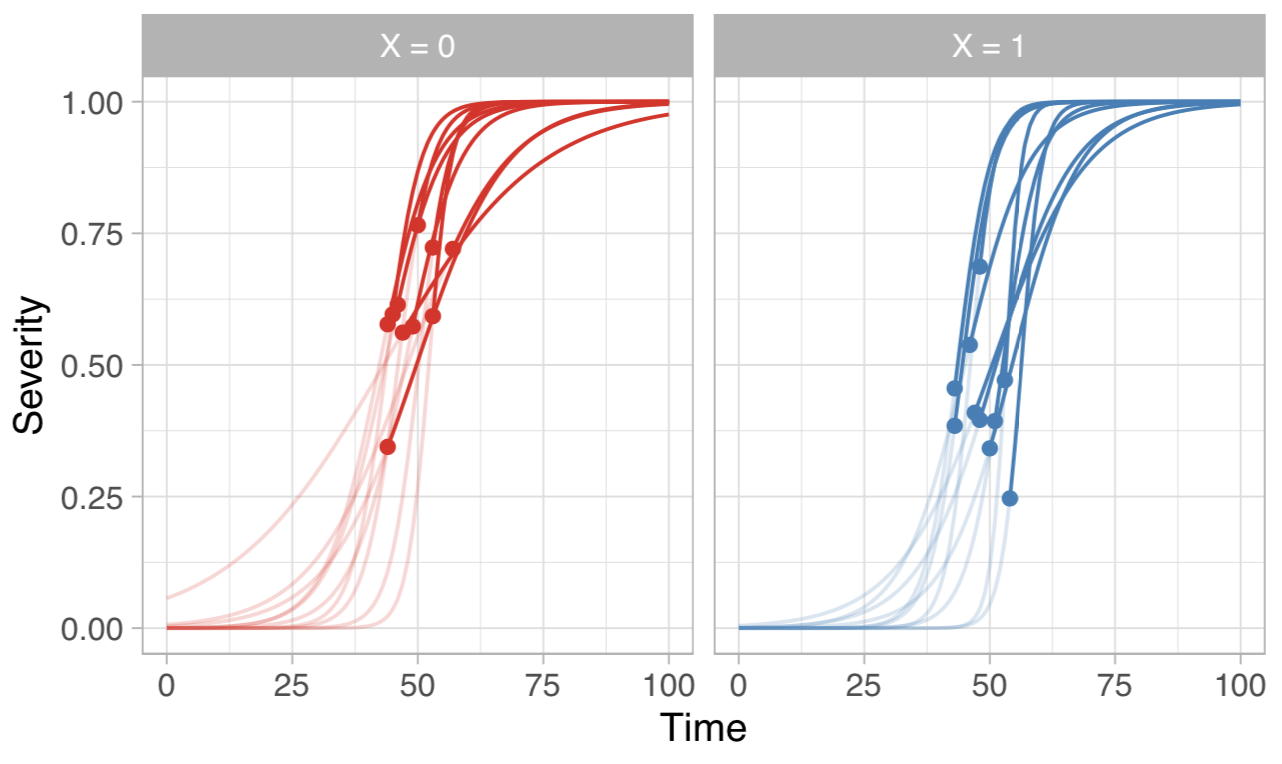
The naive analysis results here are similar to those in the *symptoms and background* sub-cases (Figure 5). Instead of recognizing the differences in presentation between groups, a researcher might conclude from these results that both groups have similar presentations, and the disease is simply less common in the group. Thus, the naive analysis obscures a difference in presentation to appear as a difference in risk.



**Figure 5:** Ground truth disease characteristics compared with a naive analysis in the *divergent symptoms* case. (Left) The naive analysis tends to underestimate disease incidence more severely in the group that has less representative symptoms (the group), causing differences in presentation to be obscured for differences in risk. (Right). Since only the most representative cases in the group receive a diagnosis, there does not appear to be a strong difference in presentation between diagnosed cases.

## Underdiagnosis in a progressive disease

The *symptoms and background* and *divergent symptoms* cases have a commonality: Individuals *with the disease* in the group are less likely to be correctly diagnosed than those in the group (Table 2, Figures 4-5). Even if these differences in diagnosis rates are initially proportional to underlying risk (*true risk factor* case), they may *become* disproportionate over time in a feedback loop with naive analysis of past data (Table 1). To underscore the impacts of disproportionate diagnostic error, we illustrate the ramifications of underdiagnosis for a progressive chronic disease.

We consider a chronic disease increasing in severity over time. At each timepoint, each undiagnosed person with the disease has a chance to be diagnostically evaluated (and thereby diagnosed). However, one patient group, () is less likely to be evaluated (Figure 2B). Individuals in the less-evaluated () group tend to be diagnosed later in their disease progression, with more severe disease at diagnosis (Figure 6). If diagnosis is the gateway to treatment, these diagnostic delays result in more severe disease and worse outcomes for this patient group. Moreover, this may lead to the impression that disease cases in the less-evaluated group are *naturally* more severe when in fact these patients are simply less reliably diagnosed.  
**Figure 6:** Simulated longitudinal disease progressions over time. Each individual’s disease trajectory progresses unobserved (lighter line) until they are diagnosed (colored dot), at which point their disease case becomes “visible” in observational data (bolder line). However, individuals in the group (red) have a lower probability of being selected for diagnostic assessment at each time point (Figure 2B). As a result, these individuals tend to be diagnosed later, when their disease is more severe.

## Mechanisms and interventions

Counteracting feedback loop failure can be challenging, and no one-size-fits-all solution exists. First, it may help to identify specific points where the system is vulnerable to failure (Table 3). We propose four main mechanisms leading to differential misclassification of disease:

1. **Medical “Risk-Adjustment”** - doctors are less likely to consider patients from certain groups for diagnosis due to a belief (correct or incorrect) that their likelihood of having the disease is low. Some risk-adjustment may be disproportionate, especially when influenced by cognitive error (e.g. anchoring and adjustment biases [19–21](https://paperpile.com/c/X7Xn59/3huIa+pxqxE+SFut1)).
2. **Divergent Presentations** - people with milder or less-well recognized presentations of the disease are less likely to be evaluated and diagnosed.
3. **Patient Self-Selection** - individuals may have different care-seeking tendencies or access to care, resulting in different rates of diagnosis.
4. **Differing evaluation efficacy** - A diagnostic evaluation may have different accuracy between patient groups (e.g. diagnosis of melanoma for individuals with dark skin [22,23](https://paperpile.com/c/X7Xn59/o7FwF+k8LQt), glomerular filtration rate calculation in chronic kidney disease diagnosis for African Americans [24,25](https://paperpile.com/c/X7Xn59/v4jMr+O1kHE)).

Note that these may interact simultaneously.

|  |  |  |  |
| --- | --- | --- | --- |
| **Mitigation strategy** | **Point of intervention** | **Description** | **Example(s)** |
| Design robust prospective studies | Data collection | Use data collection schemes robust to different sources of sampling and misclassification bias:  - Prospective surveillance studies  - Online experiments (i.e. bandit approaches)  Random sampling may be central to these designs. Researchers should also consider whether their sampled population represents their target population. | Autism studies must be robust to diagnostic selection to clarify the male-to-female risk ratio[12,13](https://paperpile.com/c/X7Xn59/IqmRA+MgSuF). Population-based disease studies can characterize prevalence and severity (e.g. SARS-CoV2/COVID-19 in Iceland,[26](https://paperpile.com/c/X7Xn59/Gpdii) urologic disease in Rochester, MN[27](https://paperpile.com/c/X7Xn59/H5GGp)) |
| Critically assess how data are collected to inform analysis | Data collection, summarization | Researchers should consider how medical data are generated and potential sources of bias, such as:  - Misclassification of disease status or outcomes  - Systematic data missingness, including which individuals appear as patients in the health system  - Cognitive biases affecting physician-recorded notes and subjective measurements.  Special care should be taken when estimating disease risk based on demographic factors that play a role in diagnosis. | Problem formulation – including understanding measurement error and sampling – is key to designing fair predictive models.[28](https://paperpile.com/c/X7Xn59/3eu3c) Understanding and addressing sources of bias in data is key to epidemiology[29](https://paperpile.com/c/X7Xn59/s1TxK) and sampling[30](https://paperpile.com/c/X7Xn59/IbYWy) methodology. |
| Improve availability and accuracy of diagnostic evaluations | Medical practice | Wider availability and accuracy of diagnostic evaluation can reduce barriers to correct diagnosis, weakening the selective pressures shaping who is diagnosed. Ensure that diagnostic evaluation – either quantitative tests or expert assessments – is accurate across patient groups. Any outstanding differences in accuracy should be well-characterized. | Availability and reliability of tests has been critical for research and monitoring of COVID-19.  New imaging tools may accelerate identification of melanoma,[31](https://paperpile.com/c/X7Xn59/hgh2C) but some such tools have differences in accuracy across skin tones. |
| Medical training against cognitive bias | Medical practice | Medical training examples that over-emphasize perceived race and sex in the diagnostic process may lead to disproportionate use of these characteristics in decision-making. Teaching medical professionals about cognitive biases may help them identify where bias arises in their own decision-making. | Cognitive biases in diagnosis are a known issue.[20,21](https://paperpile.com/c/X7Xn59/pxqxE+SFut1) Practices (including training) that over-emphasize race have come under increased scrutiny.[32,33](https://paperpile.com/c/X7Xn59/82GDn+KOTkb)  Studies explicitly examining diagnostic cognitive biases can indicate where issues arise.[19,34](https://paperpile.com/c/X7Xn59/3huIa+tMkuF) |
| Increase accessibility of care and public disease understanding | Public health, Community awareness. | Work diminishing barriers to care and improving public understanding disease risk and symptoms may diminish selection biases arising from individuals’ differing willingness or ability to seek care. | Initiatives such as the Red Dress Campaign work to increase awareness of heart attack risk among women.[35](https://paperpile.com/c/X7Xn59/ZMIt3) Women’s Heart Clinics aim to improve accessibility of cardiac care for this group.[36](https://paperpile.com/c/X7Xn59/tNzgZ) |

**Table 3:** Intervention strategies against feedback loop failures in medical diagnosis.

# Discussion

Insights from observational data can advance medicine and improve patient care. But unprecedented analytic tools and data collection do not obviate careful scientific judgment. Through simulated case-studies, we characterize scenarios in which analyses naive to a selective diagnostic process produce erroneous results, which may have harmful implications for patient care and medical research, potentially in self-reinforcing ways (Table 4). “Big Data” does not solve these problems: naive analyses of large datasets can give strong statistical credence to erroneous results. “Artificial Intelligence” does not solve these problems: naive predictive modeling on diagnosed observations with systematic errors can formalize biases in future practice.

|  |  |  |
| --- | --- | --- |
| **Implication** | **Description** | **Example** |
| Less-diagnosed groups have poorer outcomes | Individuals from a less-diagnosed group (often, one with lower perceived disease risk) are more likely to go undiagnosed or misdiagnosed - especially if their symptoms are mild or unrepresentative (Figures 4-5). If the disease progresses over time, correct diagnosis may be delayed until the disease is more severe (Figure 6). | Male patients with lupus,[19,37](https://paperpile.com/c/X7Xn59/JPv3l+3huIa) dark-skinned people with melanoma[22](https://paperpile.com/c/X7Xn59/o7FwF) and female patients with heart attacks[11](https://paperpile.com/c/X7Xn59/w2BED) tend to have poorer prognosis than patients more readily identified with the disease. |
| Individuals face disproportionate over- or underdiagnosis. | Biases in the diagnostic process – due to misunderstood disease risk, patient self-selection, clinician cognitive bias, or differences in symptoms – can cause patient populations to face disproportionate over- or under-diagnosis (Figures 4-6, Table 2). This may result in disproportionate rates of false positive tests and overtreatment in some groups and disproportionate rates of underdiagnosis and missed treatment in others. | Dark skinned people with melanoma[22](https://paperpile.com/c/X7Xn59/o7FwF) and female heart attack patients[11](https://paperpile.com/c/X7Xn59/w2BED) may face disproportionate underdiagnosis. Conversely, disproportionate *overdiagnosis* can occur for populations over-screened for certain cancers. |
| Candidate risk factors are difficult to characterize | Inferring the true associations between a candidate risk factor and disease if that risk factor affects the diagnostic process is difficult (Figures 4-5, Table 2). If underlying disease risk *does* differ between groups, these differences may be exaggerated in observational data, even if the diagnostic process itself is free of unnecessary bias (Case 2B: *true risk factor*, Figure 4). Approaches that assume the data came from a random sample of the population can produce false or exaggerated risk factors that appear associated with the disease but for which there is an insufficient causal mechanistic explanation. | Relationships between sex and disease etiology in autism remain unclear.[12,13](https://paperpile.com/c/X7Xn59/IqmRA+MgSuF) |
| Incidence and disease symptoms are mis-characterized | Even within groups with homogenous disease risk, analyses of diagnosed cases may underestimate incidence and may misrepresent the full range of symptoms (Figure 3). In particular, mild cases and cases with symptoms thought to be less representative of the disease will tend to be underdiagnosed (Figures 4-5). | Before reliable diagnostic tests were widely available, diagnosis for COVID-19 was highly selective, leading to uncertainty about the spread and severity of the disease. |

**Table 4:** Implications of feedback loop failure in a medical diagnostic setting.

Our simulations aren’t intended to capture the complexities of any one disease. Future work may consider disease-specific context or model more complex clinical processes where doctors choose between multiple diagnoses. The design of our simulations shows how quickly predictable and problematic patterns emerge even in very simple settings. Our most basic case is especially simple: people are more likely to be diagnostically evaluated if their symptoms more closely resemble a “textbook” presentation (*symptoms-only* case, Figure 2). Since mild or “atypical” cases more frequently go undiagnosed, naïve summaries of diagnosed cases make the disease appear rarer and more homogenous than it really is (Figure 3). This might self-reinforce over time, misguiding clinicians to consider this disease as a diagnosis less frequently and only for more “textbook” cases. Outside the clinic, misconceptions about the rarity and symptoms of this disease might dissuade people from seeking care when they need it.

Our subsequent case studies (*false risk factor*, *true risk factor* and *divergent symptoms*) underscore how the candidate risk factors that influence the diagnostic process can be the most challenging to characterize correctly. When both symptoms and a background characteristic (e.g. sex, gender, perceived race) influenced the diagnostic process, naive analyses of diagnosed cases again portrayed a rarer and more homogenous disease than truth, but with more strongly biased findings in the less-evaluated group (*symptoms and background* cases, Figure 4). Thus, between-group comparisons exaggerated differences in disease risk and symptoms. (Table 2). This occurred even if selection for evaluation corresponded with true underlying risk across groups (*true risk factor* case). One potential impact is that the less-evaluated group may be thought to have more severe presentations with the disease, when in actuality milder cases in this group are simply going underdiagnosed.

When patients’ backgrounds influence how symptoms manifest (*divergent symptoms* case), these background characteristics are again difficult to characterize. If one group’s symptoms are thought to be less representative of disease – either because they are harder to identify or because disease presentation in this group is less well-characterized – a naive analysis obscures differences in presentation as differences in underlying disease risk (Figure 5). For real diseases, the different mechanisms we illustrate can interact in complex ways: for example, differences in risk, differences in presentation, and under-estimation of female heart attack risk likely all play a role in heart attack underdiagnosis in female patients.[11](https://paperpile.com/c/X7Xn59/w2BED)

These misunderstandings can be costly, as failures and delays in diagnosis beget failures and delays in treatment. When the incidence of disease is truly unequal between groups, there are tradeoffs inherent to diagnostic practices directed by estimated disease risk, as suggested by algorithmic fairness research.[38,39](https://paperpile.com/c/X7Xn59/PLoMC+7upFA) Even practices ideally calibrated to risk across groups can force a paradigm in which individuals afflicted with a disease uncommon for their background face later detection and poorer outcomes (e.g. dark-skinned patients with melanoma[22](https://paperpile.com/c/X7Xn59/o7FwF)). This becomes clearly problematic when a demographic characteristic is *over*corrected for due to misunderstanding or cognitive error, needlessly overburdening specific groups with overtreatment or undertreatment. Moreover, if certain patients systematically go undiagnosed until their disease is severe, we may incorrectly conclude that severe disease in these patients is the norm, rather than a result of disparities in care (Figure 6). Delayed diagnosis and treatment of lupus in males may be one example[19](https://paperpile.com/c/X7Xn59/3huIa)).

An especially concerning scenario arises if the ‘gold standard’ diagnostic evaluation itself has varying accuracy across groups.[19](https://paperpile.com/c/X7Xn59/3huIa) This seems especially plausible when diagnostic evaluation includes potentially subjective judgements by a human examiner (e.g. autism). In the worst case, background information may informally become a part of the disease characterization, producing selection biases difficult to disentangle from disease etiology even in a prospective study.[12,13](https://paperpile.com/c/X7Xn59/IqmRA+MgSuF)

Risk modeling and automation seem to be appealing ways to remove the cognitive biases of human practitioners from the diagnostic process, but these too come with challenges. When training disease risk models on observational data, researchers must ask whether the diagnosis ‘labels’ they obtain have varying fidelity across groups.[28,40](https://paperpile.com/c/X7Xn59/3eu3c+nJB2Q) If certain patient groups in training data are more likely to have incorrect or missing diagnoses, a naively constructed model may perpetuate and formalize these errors in practice, rather than ameliorating them.

Data-driven feedback loops face a central conflict across a multitude of systems: Ideal practices for executing decisions that are effective and fair are *almost never* the ideal practices for producing high quality data to inform future decisions. There are few clinical contexts in which testing individuals for a disease at random is an appropriate policy, even though this may be conceptually ideal in a research context. When analysts summarize passively-collected data using approaches intended for researcher-designed data sets (e.g., simple random sample of a population, randomized trial) they can encourage practices that are both erroneous and potentially self-reinforcing. Learning health systems with real world data sources have great potential, but feedback loops must be carefully considered. If we inattentively summarize data without scrutinizing the selective processes that generated it, we invite a risk that the feedback loops we establish to better our decision-making give rise to a self-perpetuating cycle of systematic error.

# Supplemental Digital Content

Supplemental Digital Content. Additional detail on all simulation methods, supplementary figures 1 and 2, and supplementary tables 1 and 2. pdf

# References

1. [Lum K, Isaac W. To predict and serve? *Signif (Oxf)*. 2016;13(5):14-19.](http://paperpile.com/b/X7Xn59/FQbVH)

2. [Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S. Runaway Feedback Loops in Predictive Policing. *arXiv [csCY]*. Published online June 29, 2017.](http://paperpile.com/b/X7Xn59/5uJE1) <http://arxiv.org/abs/1706.09847>

3. [Pariser E. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin; 2011.](http://paperpile.com/b/X7Xn59/s2QIK)

4. [Germano F, Gómez V, Le Mens G. The few-get-richer: a surprising consequence of popularity-based rankings. *arXiv [csIR]*. Published online February 7, 2019.](http://paperpile.com/b/X7Xn59/n4zFx) <http://arxiv.org/abs/1902.02580>

5. [Milano S, Taddeo M, Floridi L. Recommender systems and their ethical challenges. *AI Soc*. 2020;35(4):957-967.](http://paperpile.com/b/X7Xn59/PGyks)

6. [Stemler A. Feedback Loop Failure: Implications for the Self-Regulation of the Sharing Economy. *Minn JL Sci & Tech*. 2017;18(2):673-712.](http://paperpile.com/b/X7Xn59/cdSP4)

7. [Demange G, ,PSE-EHESS, 48 bd Jourdan, Paris, 75014. Collective attention and ranking methods. *J Dyn Games*. 2014;1(1):17-43.](http://paperpile.com/b/X7Xn59/agHjv)

8. [Marafino BJ, Schuler A, Liu VX, Escobar GJ, Baiocchi M. A Causal Machine Learning Framework for Predicting Preventable Hospital Readmissions. *arXiv [statAP]*. Published online May 29, 2020.](http://paperpile.com/b/X7Xn59/9sb4v) <http://arxiv.org/abs/2005.14409>

9. [Adam GA, Chang CHK, Haibe-Kains B, Goldenberg A. Hidden Risks of Machine Learning Applied to Healthcare: Unintended Feedback Loops Between Models and Future Data Causing Model Degradation. Doshi-Velez F, Fackler J, Jung K, et al., eds. 07--08 Aug 2020;126:710-731.](http://paperpile.com/b/X7Xn59/y5JaJ)

10. [Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med*. 2018;378(11):981-983.](http://paperpile.com/b/X7Xn59/7d8Lq)

11. [Aggarwal NR, Patel HN, Mehta LS, et al. Sex Differences in Ischemic Heart Disease. *Circ Cardiovasc Qual Outcomes*. 2018;11(2):e004437.](http://paperpile.com/b/X7Xn59/w2BED)

12. [Loomes R, Hull L, Mandy WPL. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *J Am Acad Child Adolesc Psychiatry*. 2017;56(6):466-474.](http://paperpile.com/b/X7Xn59/IqmRA)

13. [Fombonne E, MacFarlane H, Salem AC. Epidemiological surveys of ASD: advances and remaining challenges. *J Autism Dev Disord*. Published online April 17, 2021. doi:](http://paperpile.com/b/X7Xn59/MgSuF)[10.1007/s10803-021-05005-9](http://dx.doi.org/10.1007/s10803-021-05005-9)

14. [Khan NA, Daskalopoulou SS, Karp I, et al. Sex differences in acute coronary syndrome symptom presentation in young patients. *JAMA Intern Med*. 2013;173(20):1863-1871.](http://paperpile.com/b/X7Xn59/Hz36r)

15. [Canto JG, Goldberg RJ, Hand MM, et al. Symptom presentation of women with acute coronary syndromes: myth vs reality. *Arch Intern Med*. 2007;167(22):2405-2413.](http://paperpile.com/b/X7Xn59/hXp3H)

16. [D’Onofrio G, Safdar B, Lichtman JH, et al. Sex differences in reperfusion in young patients with ST-segment-elevation myocardial infarction: results from the VIRGO study. *Circulation*. 2015;131(15):1324-1332.](http://paperpile.com/b/X7Xn59/xSFBT)

17. [Bugiardini R, Ricci B, Cenko E, et al. Delayed Care and Mortality Among Women and Men With Myocardial Infarction. *J Am Heart Assoc*. 2017;6(8). doi:](http://paperpile.com/b/X7Xn59/IED6f)[10.1161/JAHA.117.005968](http://dx.doi.org/10.1161/JAHA.117.005968)

18. [Roswell RO, Kunkes J, Chen AY, et al. Impact of Sex and Contact-to-Device Time on Clinical Outcomes in Acute ST-Segment Elevation Myocardial Infarction-Findings From the National Cardiovascular Data Registry. *J Am Heart Assoc*. 2017;6(1). doi:](http://paperpile.com/b/X7Xn59/abaTl)[10.1161/JAHA.116.004521](http://dx.doi.org/10.1161/JAHA.116.004521)

19. [Simard JF, Chaichian Y, Rizk N, Rector A, Feldman CH, Falasinnu TO. Are we missing Lupus in Males? Evidence of cognitive bias from a randomized experiment in the United States. *Am J Epidemiol*.](http://paperpile.com/b/X7Xn59/3huIa)

20. [Scott IA. Errors in clinical reasoning: causes and remedial strategies. *BMJ*. 2009;338:b1860.](http://paperpile.com/b/X7Xn59/pxqxE)

21. [Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med*. 2008;121(5 Suppl):S2-S23.](http://paperpile.com/b/X7Xn59/SFut1)

22. [Gupta AK, Bharadwaj M, Mehrotra R. Skin Cancer Concerns in People of Color: Risk Factors and Prevention. *Asian Pac J Cancer Prev*. 2016;17(12):5257-5264.](http://paperpile.com/b/X7Xn59/o7FwF)

23. [Kabigting FD, Nelson FP, Kauffman CL, Popoveniuc G, Dasanu CA, Alexandrescu DT. Malignant melanoma in African-Americans. *Dermatol Online J*. 2009;15(2):3.](http://paperpile.com/b/X7Xn59/k8LQt)

24. [Delanaye P, Mariat C, Maillard N, Krzesinski JM, Cavalier E. Are the creatinine-based equations accurate to estimate glomerular filtration rate in African American populations? *Clin J Am Soc Nephrol*. 2011;6(4):906-912.](http://paperpile.com/b/X7Xn59/v4jMr)

25. [Peralta CA, Lin F, Shlipak MG, et al. Race differences in prevalence of chronic kidney disease among young adults using creatinine-based glomerular filtration rate-estimating equations. *Nephrol Dial Transplant*. 2010;25(12):3934-3939.](http://paperpile.com/b/X7Xn59/O1kHE)

26. [Gudbjartsson DF, Helgason A, Jonsson H, et al. Spread of SARS-CoV-2 in the Icelandic Population. *N Engl J Med*. 2020;382(24):2302-2315.](http://paperpile.com/b/X7Xn59/Gpdii)

27. [Mohr DN, Offord KP, Owen RA, Melton LJ 3rd. Asymptomatic microhematuria and urologic disease. A population-based study. *JAMA*. 1986;256(2):224-229.](http://paperpile.com/b/X7Xn59/H5GGp)

28. [Passi S, Barocas S. Problem Formulation and Fairness. *arXiv [csCY]*. Published online January 8, 2019. doi:](http://paperpile.com/b/X7Xn59/3eu3c)[10.1145/3287560.3287567](http://dx.doi.org/10.1145/3287560.3287567)

29. [Szklo M, Javier Nieto F. *Epidemiology: Beyond the Basics*. Jones & Bartlett Learning; 2004.](http://paperpile.com/b/X7Xn59/s1TxK)

30. [Lohr SL. *Sampling: Design and Analysis (Advanced Series)*. 2nd ed. Cengage Learning; 2009.](http://paperpile.com/b/X7Xn59/IbYWy)

31. [Giuffrida R, Conforti C, Di Meo N, Deinlein T, Guida S, Zalaudek I. Use of noninvasive imaging in the management of skin cancer. *Curr Opin Oncol*. 2020;32(2):98-105.](http://paperpile.com/b/X7Xn59/hgh2C)

32. [Garcia RS. The misuse of race in medical diagnosis. *Pediatrics*. 2004;113(5):1394-1395.](http://paperpile.com/b/X7Xn59/82GDn)

33. [Amutah C, Greenidge K, Mante A, et al. Misrepresenting Race - The Role of Medical Schools in Propagating Physician Bias. *N Engl J Med*. 2021;384(9):872-878.](http://paperpile.com/b/X7Xn59/KOTkb)

34. [Mullainathan S, Obermeyer Z. *Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care*. National Bureau of Economic Research; 2019. doi:](http://paperpile.com/b/X7Xn59/tMkuF)[10.3386/w26168](http://dx.doi.org/10.3386/w26168)

35. [Mosca L, Hammond G, Mochari-Greenberger H, Towfighi A, Albert MA, American Heart Association Cardiovascular Disease and Stroke in Women and Special Populations Committee of the Council on Clinical Cardiology, Council on Epidemiology and Prevention, Council on Cardiovascular Nursing, Council on High Bloo. Fifteen-year trends in awareness of heart disease in women: results of a 2012 American Heart Association national survey. *Circulation*. 2013;127(11):1254-1263, e1-e29.](http://paperpile.com/b/X7Xn59/ZMIt3)

36. [Garcia M, Miller VM, Gulati M, et al. Focused Cardiovascular Care for Women: The Need and Role in Clinical Practice. *Mayo Clin Proc*. 2016;91(2):226-240.](http://paperpile.com/b/X7Xn59/tNzgZ)

37. [Crosslin KL, Wiginton KL. Sex differences in disease severity among patients with systemic lupus erythematosus. *Gend Med*. 2011;8(6):365-371.](http://paperpile.com/b/X7Xn59/JPv3l)

38. [Kleinberg J, Mullainathan S, Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv [csLG]*. Published online September 19, 2016.](http://paperpile.com/b/X7Xn59/PLoMC) <http://arxiv.org/abs/1609.05807>

39. [Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*. 2017;5(2):153-163.](http://paperpile.com/b/X7Xn59/7upFA)

40. [Barocas S, Selbst AD. Big Data’s Disparate Impact. Published online 2016. doi:](http://paperpile.com/b/X7Xn59/nJB2Q)[10.2139/ssrn.2477899](http://dx.doi.org/10.2139/ssrn.2477899)