For a patient, correct diagnosis is 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 which 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.

*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 X 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,” which 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 *R,* 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, *R*, and a background characteristic X (e.g. sex, gender, perceived race).
  + **Case 2A [*False risk factor*]:** X *is not* associated with disease risk.
  + **Case 2B [*True risk factor*]:** X *is* associated with disease risk.
* **Case 3 [*Divergent symptoms*]:** Selection for evaluation depends only on symptom representativeness, *R*, but X determines the distribution of *R.*

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/raikens1/confirmationBias/).

## *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 R), and (except in the *symptoms-only* case) a background characteristic (denoted by X). R 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. *R* 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 *R,* 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, X. For simplicity in our simulations, X is a binary characteristic equal to 1 in roughly half the population. In the *false risk factor* case, X is *not associated with the underlying disease state*: Individuals with different values of X have the same probability of having the disease, and they have the same distributions of *R* conditional on their disease state (Figure 2A).  In the *true risk factor* case, having X= 1 *doubles* disease risk compared to having X=0(Figure 2A). The disease arises in 15% of individuals with X=1 and only 7.5% of those with  X= 0.

In the *divergent symptoms* case, X is not a risk factor, but individuals in different X groups have different common presentations (different distributions of R). Individuals with X= 1 tend to have lower R, 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 which progresses over time.  In this setting we additionally model the underlying severity of the condition, summarized by S, which varies continuously from 0 to 1. Over time, an individual’s disease shifts from no severity (mild or asymptomatic, S= 0) to high severity (highly severe disease, S= 1), 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 X. Here, representativeness, R, is simulated as a noisy version of S.

## *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, R, potentially with adjustment based on the background characteristic, X (Figure 2B).  Except in the *true risk factor* case, this is done using a parametrized sigmoid curve (see supplement). 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 x and representativeness r 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 detail in supplement, Supplementary Figure S1). 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 (Supplementary Figure S2, Supplementary Tables S1 and S2).

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.

*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 R) are less likely to be diagnosed correctly. 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 which clinicians believe arises only as "textbook" presentations, allowing an array of “atypical” disease presentations to go undetected.

## *Case 2: Selection on Symptoms and Background*

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

### *Case 2A: False risk factor*

Here, X 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 X is *incorrectly perceived* as a risk factor. Figure 4A shows findings from a naive analysis.

Within groups defined by X, 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 X= 0 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 X appears to be associated with the disease even though it is not (Figure 4A, left), and diagnosed cases among those with X= 0 appear to have more “textbook” representative symptoms than those with X= 1 (Figure 4A, right).  Thus X 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.

These summary statistics are consistent with incorrect conclusions that the disease is rarer in people with X= 0 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 X= 0 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 X *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 X and R (Methods). The shape of this curve (Figure 2B) shows that although X is associated with *double* the disease risk, a patient with symptom representativeness r andX= 1 does *not* in general have twice the likelihood of having the disease as a person with the same representativeness and X = 0 (Supplementary Figure S1). 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 *true 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 X= 0 group. As a result, the incidence is *underestimated* more in the low-risk (X= 0) group relative to the other, causing X 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 X is weighted in the diagnostic process, the harder it becomes for a disease case in the X= 0 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 X=1 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 (Supplementary Figure S2, Supplementary Tables S1 and S2).

## *Case 3: Divergent symptoms*

Here, we suppose that individuals with X=0 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 X=1 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/p7iIyh/jZOB4+JiFtf) (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/p7iIyh/KtYU3+6hi8n+hsTpw+kXxi6).

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 X=0 group. Thus, the naive analysis obscures a difference in presentation to appear as a difference in risk.

## *Underdiagnosis in a progressive disease*

The *symptoms and background* and *divergent symptoms* cases have a commonality: Individuals *with the disease* in the X= 0 group are less likely to be correctly diagnosed than those in the X= 1 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 if influenced by naive analyses of past diagnoses (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, (X= 0) is less likely to be evaluated (Figure 2B). Individuals in the less-evaluated (X= 0) 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 cared-for.

*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/p7iIyh/SL3JP+yid8Y+awTKy)).
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/p7iIyh/7Grrd+mnPS9), glomerular filtration rate calculation in chronic kidney disease diagnosis for African Americans [(*24*, *25*)](https://paperpile.com/c/p7iIyh/pcaMv+hpDo5)).

Note that these may interact simultaneously.

*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: inappropriate analyses of large datasets can give strong statistical credence to erroneous results.

Our simulations aren’t intended to capture the complexities of any one disease.  Rather, they show 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 which 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/p7iIyh/kXxi6).

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 [(*35*, *36*)](https://paperpile.com/c/p7iIyh/1RKkA+P9d9c).  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/p7iIyh/7Grrd)). 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).

An especially concerning scenario arises if the ‘gold standard’ diagnostic evaluation itself has varying accuracy across groups [(*19*)](https://paperpile.com/c/p7iIyh/SL3JP). This seems especially plausible when diagnostic evaluation includes potentially subjective judgements by a human examiner (e.g. autism, lupus). 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/p7iIyh/TSh4Y+l3dWM).  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 [(*27*, *37*)](https://paperpile.com/c/p7iIyh/qG28y+uOKG5).  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, randomized trial) they can introduce practices which are both misleading 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 apply passively collected 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.