Web Appendix

Scenarios for feedback loop failure in medical diagnosis

Contents

Simulation Methods: Additional Detail

Simulated Case Studies

This section gives further detail on the simulation settings for the simulated case studies which consider naive analyses of medical datasets.

Disease model

Let $i = 1, ...100,000$ index over all individuals in a cohort. We describe each individual by an underlying disease state, Z_i, a quantity, R_i, summarizing the representativeness of their symptoms, and (except in Case 1), a background characteristic X_i. (Note that the notation for the underlying disease state, Z_i, is dropped in the main text for simplicity.) Let $Z_i = 1$ denote that individual i has the disease in question, and otherwise let $Z_i = 0$. As summarized in the main text, R_i varies continuously from 0 to 1, with a value of 0 indicating no resemblence to a "classic presentation" of the disease and 1 representing a "classic presentation." Where applicable, X_i is binary.

In Case 1, individual variables are simulated as follows:

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

That is, the disease arises in 10% of individuals, and individuals with the disease tend to have more representative symptoms than those who do not, with some exceptions in both directions.

In Cases 2 and 3, we additionally simulatethe background characteristic, X:

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

In Case 2A, X_i is independent of Z_i and R_i, which are simulated in the same way as in the Case 1 set-up, above. In Case 2B, we further adjust the simulation so that $X_i = 1$ doubles disease risk compared to $X_i = 0$:

$$\begin{aligned} Z_i|X_i = 0 &\sim_{iid} \text{Bernoulli}(0.075) \\ Z_i|X_i = 1 &\sim_{iid} \text{Bernoulli}(0.15) \end{aligned}$$

Finally, in Case 3, X_i has no influence on the risk of disease (X_i and Z_i are independent), but each individual's value of X_i determines the distribution of symptoms. The setting is as follows:

$$\begin{aligned} Z_i &\sim_{iid} \text{Bernoulli}(0.1) \\ R_i|Z_i = 1, X_i = 0 &\sim_{iid} \text{Beta}(5, 2.5) \\ R_i|Z_i = 1, X_i = 1 &\sim_{iid} \text{Beta}(5, 2) \\ R_i|Z_i = 0 &\sim_{iid} \text{Beta}(1, 5). \end{aligned}$$

Clinical Selection Process

The models for the probability of evaluation for each case are shown visually in figure 2 from the main text. Here we give the exact formulae. Except in Case 2B, the probability of evaluation given representativeness, R_i (and possibly X_i) is given based on a sigmoid function. For Case 1 and Case 3, this function is:

$$P(\text{evaluation}|R_i) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 R_i))} - c,$$

Where $\beta_1 = 20$ and $\beta_0 = -12$. Here, $c = \frac{1}{1+exp(-\beta_0)}$ is a corrective constant to ensure that $P(\text{evaluation}|R_i = 0) = 0$.

In Case 2, probability of evaluation additionally depends on X_i. For Case 2A, this is also a sigmoid curve:

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

In Case 2A, $\beta_0 = -13$, $\beta_1 = 20$, and $\beta_2 = -3$. The constant, $c_{x_i} = \frac{1}{1+exp(-(\beta_0+\beta_2 x_i))}$ is again a corrective constant to ensure that $P(\text{evaluated}|R_i(t) = 0, X_i = x_i) = 0$. Note that c_{x_i} depends on β_0, β_1, β_2, and X_i.

In Case 2B, we suppose that the probability of evaluation is precisely the "disease risk" of the patient based on their values of R and X. That is,

$$P(\text{evaluation}|R_i, X_i) = P(Z_i = 1|R_i, X_i),$$

Where the right hand side is derived directly from the simulation set-up. For the Case 2B set-up this is:

$$P(\text{evaluation}|R_i, X_i) = \frac{Beta_{5,2.5}(R_i)P(Z_i = 1|X_i)}{Beta_{5,2.5}(R_i)P(Z_i = 1|X_i) + Beta_{1,5}(R_i)P(Z_i = 0|X_i)},$$

where $Beta_{a,b}$ represents the p.d.f of a Beta distribution with parameters a and b. Note that $P(Z_i = 1|X_i)$ is precisely the true incidence of disease in the X_i group, and $P(Z_i = 0|X_i)$ is one minus that incidence.

A derivation is as follows:

$$\begin{aligned}
P(Z = 1|R = r, X = x) &= \frac{P(R = r|Z = 1, X = x)P(Z = 1|X = x)}{P(R = r|X = x)} \\
&= \frac{P(R = r|Z = 1)P(Z = 1|X = x)}{P(R = r|Z = 1)P(Z = 1|X = x) + P(R = r|Z = 0)P(Z = 0|X = x)} \\
&= \frac{Beta_{5,2.5}(r)P(Z = 1|X = x)}{Beta_{5,2.5}(r)P(Z = 1|X = x) + Beta_{1,5}(r)P(Z = 0|X = x)}.
\end{aligned}$$

Longitudinal simulation setting

This section contains further detail on the data-generation for the longitudinal simulation example.

Disease model

In the longitudinal setting, we simulate a set of disease cases which become more severe over time. In this scenario, we additionally simulate a quantity, S, denoting the severity of the disease. Like R, S varies continuously from 0 to 1, with $S = 0$ meaning no severity and $S = 1$ meaning maximum severity. In this supplement, we use slightly different notation for R and S. Since these two quantities vary over time in the longitudinal simulations, we denote them as functions of time, t. Thus, $R_i(t)$ denotes the representativeness of the symptims of individual i at time t, and likewise for $S_i(t)$. We consider increments of time, t from 0 to 1,00.

Severity progresses from 0 to 1 according to a sigmoid function with the following parameters:

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

Here, β_i and T_i are disease-progression parameters intrinsic to the individual:

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

We simulate β_i and T_i from the following distributions:

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

This means that most people have a gradual disease progression ($\mathbb{E}[\beta_i] = 0.25$), and the disease tends to hit intermediate severity at about $t = 50$ ($\mathbb{E}[T_i] = 50$). Note that T_i and β_i are generated from the same distribution regardless of the background characteristic, X_i. This means that the disease *progression* is not systematically different between X groups (although it does not necessitate that the disease *incidences* are the same between these groups).

Individuals are diagnosed not based on severity but on representativeness of their symptoms. In this setting, we denote representativeness as a function of time, $R_i(t)$ (although the function notation is dropped in the main text). Underlying severity is related to – but not the same as – symptom representativeness. Severity, $S_i(t)$, describes the underlying disease state, while representativeness, $R_i(t)$, describes the observable symptoms and how well they match a "textbook" description of the disease. For the longitudinal simulation *only*, we simulate $R_i(t)$ as a more noisy version of $S_i(t)$:

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

Where

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

Inessense, this means that representativeness at any time corresponds roughly to severity, but that this relationship is staggered with some noise.

Clinical selection process

At each time-point, t, each individual has an opportunity to be selected for diagnostic evaluation based on the representativeness of their symptoms. As in Cases 1-3, if the individual is selected for diagnostic evaluation, they recieve a diagnosis; otherwise they remain diagnosed (at least until the next time point). The probability of being selected for diagnostic evaluation at any given time point is the same as in Case 2A:

$$P(\text{evaluation}|R_i(t), X_i) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 R_i(t) + \beta_2 X_i))} - c_{x_i},$$

where $\beta_0 = -13$, $\beta_1 = 20$, and $\beta_2 = -3$. As before $c_{x_i} = \frac{1}{1+exp(-(\beta_0+\beta_2 X_i))}$ is a corrective constant to ensure that $P(\text{evaluated}|R_i(t), X_i) = 0$.

Supplementary Figures

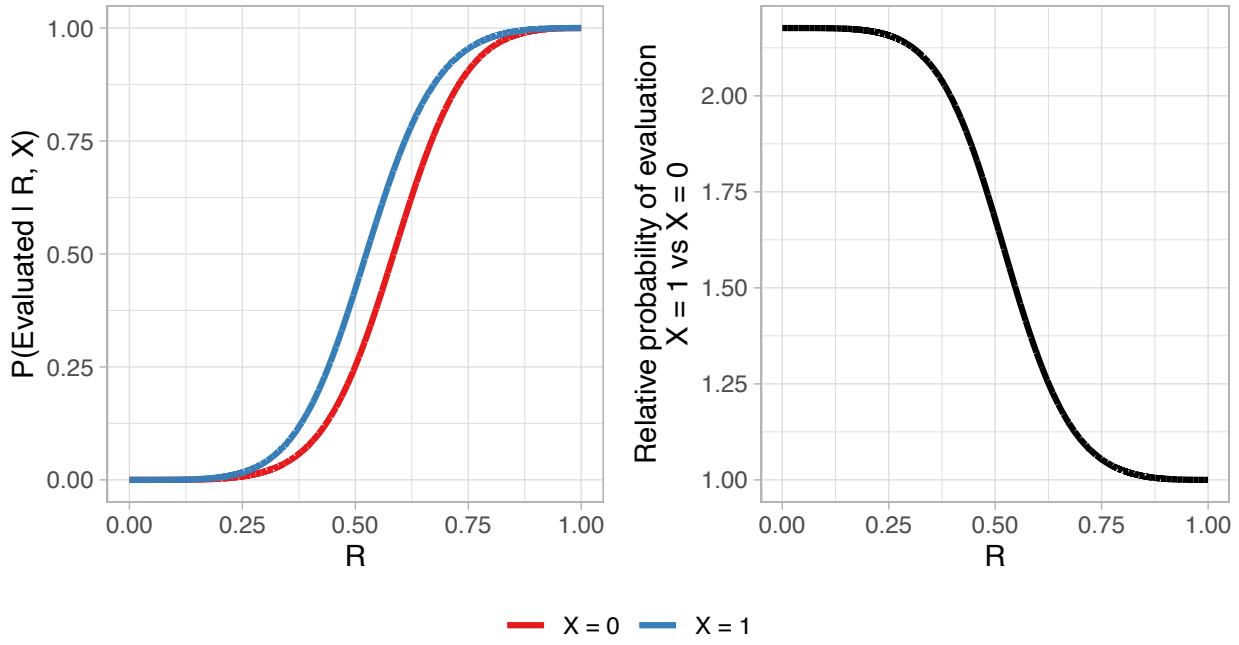

Figure S1: Visualization of clinical selection process model for Case 2B. (Right) Probability of evaulation given R and X for Case 2B ('true risk factor'). Note that while the form of the curve may appear logistic, it is not (see supplementary additional detail for clinical selection process, Cases 1-3). (Left) Relative probability of evaluation in Case 2B model for invividuals with X = 1 vs X = 0 for different levels of R. When R is low (symptoms are not very representative of the disease), people with X = 1 are more than twice as likely to have the disease (and thus more than twice as likely to be evaluated) than people with X = 0. As R increases, the relative rates of evaluation converge. When R is close to 1 (symptoms are highly representative of the disease), the likelihood of disease is close to 1 in both groups, so the relative difference in likelihood of disease converges to 0. That is, for more representative cases, X is less informative to likelihood of disease.

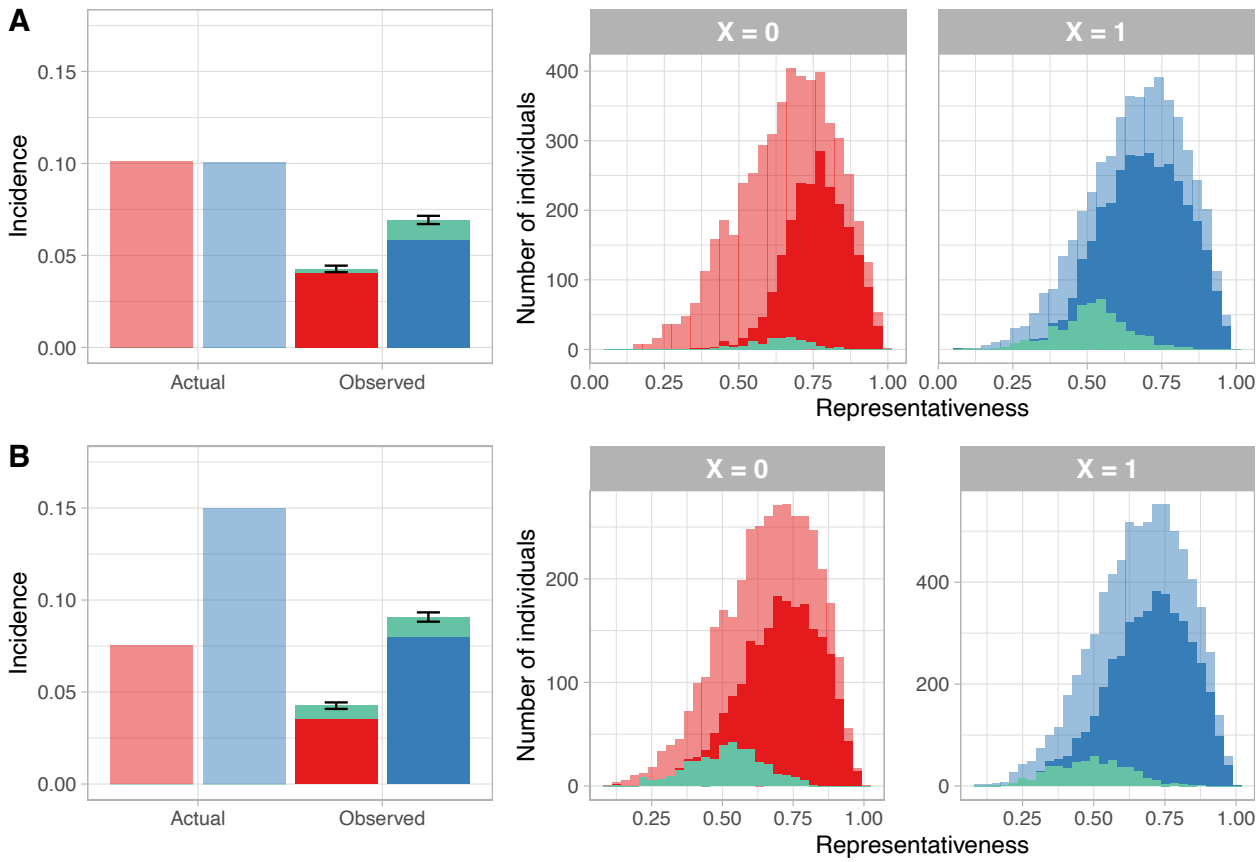

Figure S2: Ground truth disease characteristics compared with a naive analysis in Cases 2A and 2B ('true risk factor' and 'false risk facor', respectively) for a simulation in which the diagnostic evaluation has a sensitivity and specificity of 0.7. Observed incorrect diagnoses of non-diseased individuals with a false-positive evaluation result are shaded in green. In the 'true risk factor' case (A), individuals in the X = 1 group are evaluated disproportionately more often. This means that a larger share of false positive diagnoses are present for the observed data in this group. The false positive diagnoses tend to have less representative symptoms than the true positive diagnoses, distorting the distribution of observed cases. In this case, this distortion is more pronounced in the more-evaluated X = 1 group. (B) In the 'false risk factor' case, the X = 0 group also includes a larger number of false positives, and false positives in both groups tend to fall on the lower end of the distribution of representativeness among diagnosed cases. However, the probability of evaluation corresponds to true disease risk based on representativeness and X.

Supplementary Tables

Table S1: 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	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
Ranking and recommender systems (e.g. media, news)	Popularity ranking model	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] may arise.	4–7
Diagnosis	Disease symptoms and risk factors	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[8] autism[9,10], and risk modeling[11]

References for Table S1

1. Lum K, Isaac W. To predict and serve? *Signif (Oxf)* 2016; 13: 14–19.

2. Ensign D, Friedler SA, Neville S, et al. Runaway Feedback Loops in Predictive Policing. In: Friedler SA, Wilson C (eds) *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 23--24 Feb 2018, pp. 160–171.

3. Pariser E. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin, 2011.

4. Germano F, Gómez V, Le Mens G. The few-get-richer: a surprising consequence of popularity-based rankings? In: *The World Wide Web Conference*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 2764–2770.

5. Milano S, Taddeo M, Floridi L. Recommender systems and their ethical challenges. *AI Soc* 2020; 35: 957–967.

6. Stemler A. Feedback Loop Failure: Implications for the Self-Regulation of the Sharing Economy. *Minn JL Sci & Tech* 2017; 18: 673–712.

7. Demange G. Collective attention and ranking methods. *J Dyn Games* 2014; 1: 17–43.

8. Aggarwal NR, Patel HN, Mehta LS, et al. Sex Differences in Ischemic Heart Disease. *Circ Cardiovasc Qual Outcomes* 2018; 11: e004437.

9. 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: 466–474.

10. Fombonne E, MacFarlane H, Salem AC. Epidemiological surveys of ASD: advances and remaining challenges. *J Autism Dev Disord*. Epub ahead of print 17 April 2021. DOI: 10.1007/s10803-021-05005-9.

11. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…. *J Am Med Inform Assoc* 2019; 26: 1645–1650.

Table S2: Average results from 1000 simulations of the Case 2A scenario with varying levels of evaluation sensitivity and specificity. Values are presented as "mean (standard deviation)" calculated across simulations. Incidences are presented as percents. Note that the true incidences are 10 percent for both groups, and the true relative risk is 1.

Sensitivity	Specificity	Percent cases diagnosed (X = 0)	Percent cases diagnosed (X = 1)	Estimated incidence (X = 0)	Estimated incidence (X = 1)	Estimated relative incidence (X = 1 vs X = 0)
1.0	1.0	55.8 (0.7)	81.1 (0.6)	5.6 (0.1)	8.1 (0.1)	1.45 (0.035)
1.0	0.7	55.9 (0.7)	81.2 (0.6)	5.8 (0.1)	9.3 (0.1)	1.59 (0.036)
0.7	1.0	39.1 (0.7)	56.8 (0.7)	3.9 (0.1)	5.7 (0.1)	1.45 (0.042)
0.7	0.7	39.1 (0.7)	56.8 (0.7)	4.2 (0.1)	6.8 (0.1)	1.64 (0.045)

Table S3: Average results from 1000 simulations of the Case 2B scenario with varying levels of evaluation sensitivity and specificity. Values are presented as "mean (standard deviation)" calculated across simulations. Incidences are presented as percents. Note that the true incidences are 7.5 percent for the X = 0 group and 15 for the X = 1 group, giving a true relative risk of 2.

Sensitivity	Specificity	Percent cases diagnosed (X = 0)	Percent cases diagnosed (X = 1)	Estimated incidence (X = 0)	Estimated incidence (X = 1)	Estimated relative incidence (X = 1 vs X = 0)
1.0	1.0	66.6 (0.8)	75.4 (0.5)	5.0 (0.1)	11.3 (0.1)	2.26 (0.053)
1.0	0.7	66.6 (0.8)	75.4 (0.5)	5.7 (0.1)	12.4 (0.1)	2.16 (0.046)
0.7	1.0	46.6 (0.9)	52.8 (0.6)	3.5 (0.1)	7.9 (0.1)	2.26 (0.066)
0.7	0.7	46.7 (0.8)	52.8 (0.6)	4.2 (0.1)	9.0 (0.1)	2.12 (0.056)