

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 $i = 1, \dots, 100,000$ 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 resemblance 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 simulate the 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 $P(\text{evaluation}|R_i(t) = 0, X_i = x_i) = 0$. 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:

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

where $\text{Beta}_{a,b}$ 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:

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

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 symptoms 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 controls 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:

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

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

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

In essence, 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 receive 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 $P(\text{evaluated}|R_i(t), X_i) = 0$.