EBM Basic Simulation

Rachael Caelie (Rocky) Aikens

7/22/2020

Set up

We'll start with the basic cross-sectional model, with a single baseline characteristic, X_i. Here, X_i is binary and about half the population has $X_i = 1$, half has $X_i = 0$. (This could be sex, naturally, but we might consider some other discrete group like race, some comorbidity, sexuality, etc.)

$$\begin{aligned} X_i &\sim_{iid} \text{Bernoulli}(0.5) \\ Z_i &\sim_{iid} \text{Bernoulli}(0.15) \\ S_i | Z_i = 1 &\sim_{iid} \text{Uniform}(0, 1) \end{aligned}$$

Where $S_i = 0$ whenever $Z_i = 0$.

Additionally let's consider that a person's probability of diagnosis depends on not only the severity of their disease but their baseline characteristics:

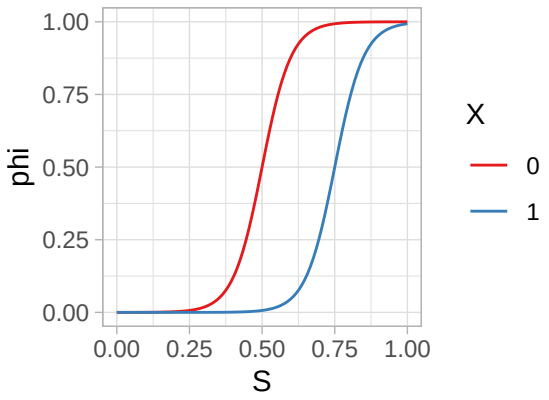
$$D_i \sim_{iid} \text{Bernoulli}\left(\phi(S_i, X_i)\right)$$

Without loss of generality, let's assume that people with $X_i = 0$ are more likely to be correctly diagnosed than people with $X_i = 1$. Explicitly, let's suppose:

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

where again, $c_{x_i} = \frac{1}{1+exp(-(\beta_0+\beta_2 X_i))}$ is a corrective constant to ensure that $\phi(0, X_i) = 0$. Let Let $\beta_0 = -10$, $\beta_1 = 20$, and $\beta_2 = -5$.

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



A Naive Predictive Modeling Study

Again, we collect a dataset of 5000, recording the baseline characteristic $\{S_i, D_i, X_i\}_{i=1}^n$. Now, our researcher wants to fit a logistic model for predicting disease status from X_i and S_i. However, since the true disease status is unkown, they use diagnosis as a proxy for disease. In regression shorthand they want to model:

$$D_i \sim S_i + X_i$$

The model result of such a study is shown below.

```
##
## Call:
## glm(formula = diagnosed ~ x + severity, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4929  -0.0158  -0.0158  -0.0016   4.0332
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.9948     0.7557 -11.902   <2e-16 ***
## xTRUE        -4.5133     0.4907  -9.198   <2e-16 ***
## severity     17.8851     1.4827  12.062   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1996.91  on 4999  degrees of freedom
## Residual deviance:  271.68  on 4997  degrees of freedom
## AIC: 277.68
##
## Number of Fisher Scoring iterations: 10
```

	OR	2.5 %	97.5 %
(Intercept)	0.000	0.000	0.000000e+00
xTRUE	0.011	0.004	2.700000e-02
severity	58530978.201	4221736.753	1.434444e+09

The researcher records a very high AIC and reports that this model is quite effective at predicting who has the disease. They write a nice paper, talking about how predictive modeling is the gateway to evidence-based personalized medicine. Maybe they suggest that future work might leverage the Awesome Power of Machine Learning. When they interpret their results, they might say that $X_i = 0$ significantly reduces predicted disease risk.

What's wrong with this

The researcher above imagines they are fitting a model for disease risk when in fact they are fitting a model for diagnosis probability. In fact, the model shown above is actually quite a good fit for $\phi(S_i, X_i)$.

If we actually ran a logistic regression on disease status rather than diagnosis, we would find that X is not at all related to who gets the disease and who does not.

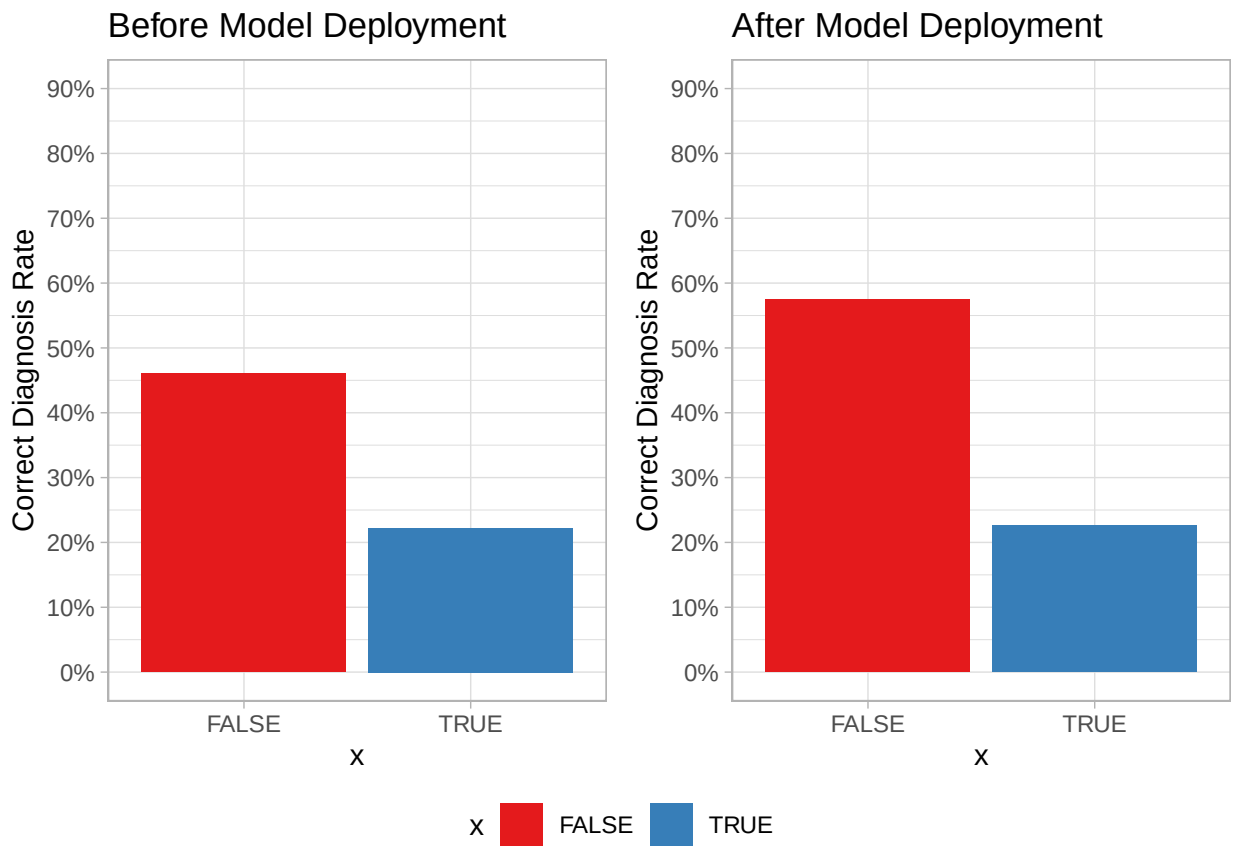
```
##
## Call:
## glm(formula = disease ~ x + severity, family = "binomial", data = df)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -0.0003028  -0.0003028  -0.0001157  -0.0001157   0.0020630
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -18.823    160.285  -0.117    0.907
## xTRUE            1.925    171.711   0.011    0.991
## severity     42089.454 595547.859   0.071    0.944
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.2027e+03  on 4999  degrees of freedom
## Residual deviance: 2.2694e-04  on 4997  degrees of freedom
## AIC: 6.0002
##
## Number of Fisher Scoring iterations: 25
```

	OR	2.5 %	97.5 %
(Intercept)	0.000	0.00	0.000000e+00
xTRUE	6.856	0.02	9.222567e+28
severity	Inf	Inf	Inf

Deployment

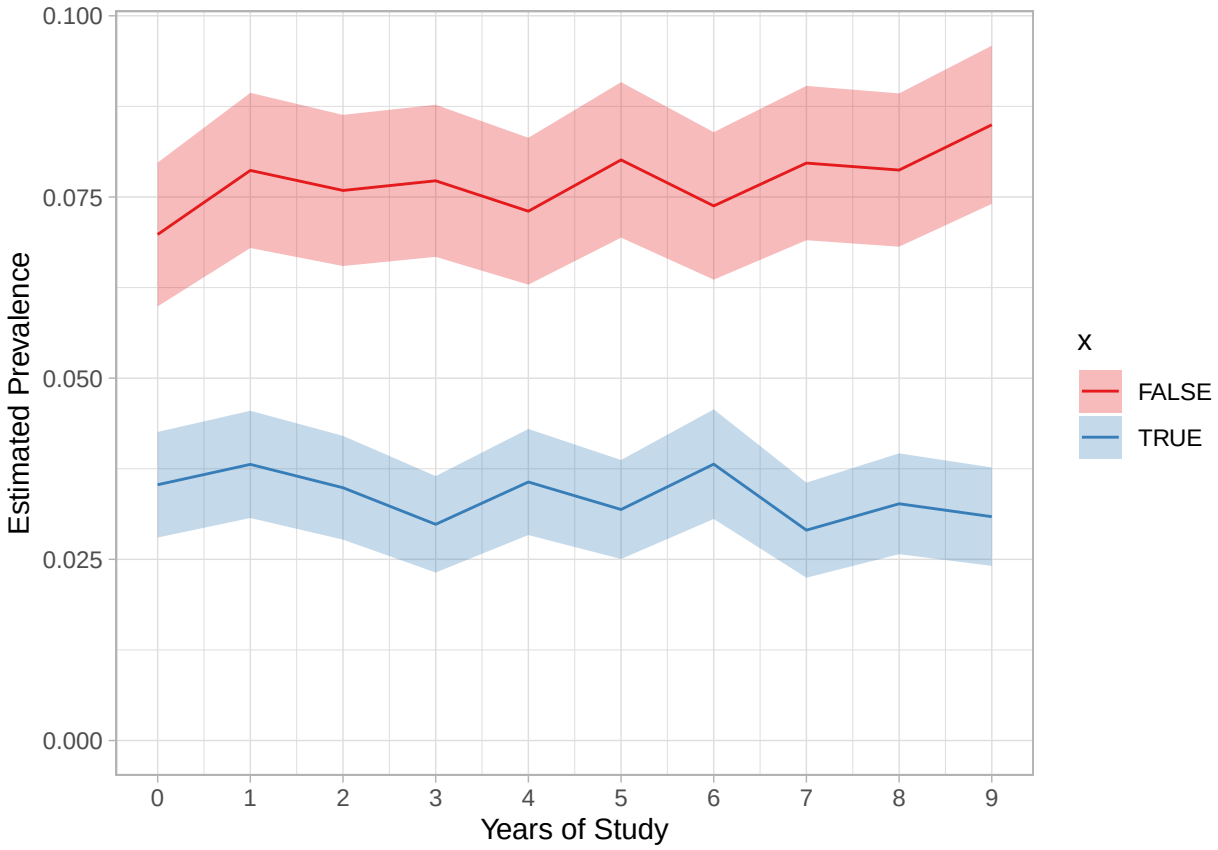
Nonetheless, the naive model is published. Perhaps it is deployed as a "risk score" calculator with the objective of informing doctor's diagnosis. Doctors then read about this model and deploy it. In reality, we know that doctors usualy do not do exact numeric calculations to decide what diagnosis to give, but let's suppose for the moment that they perfectly follow this published model. In reality, some doctors will be less faithful and some more faithful.

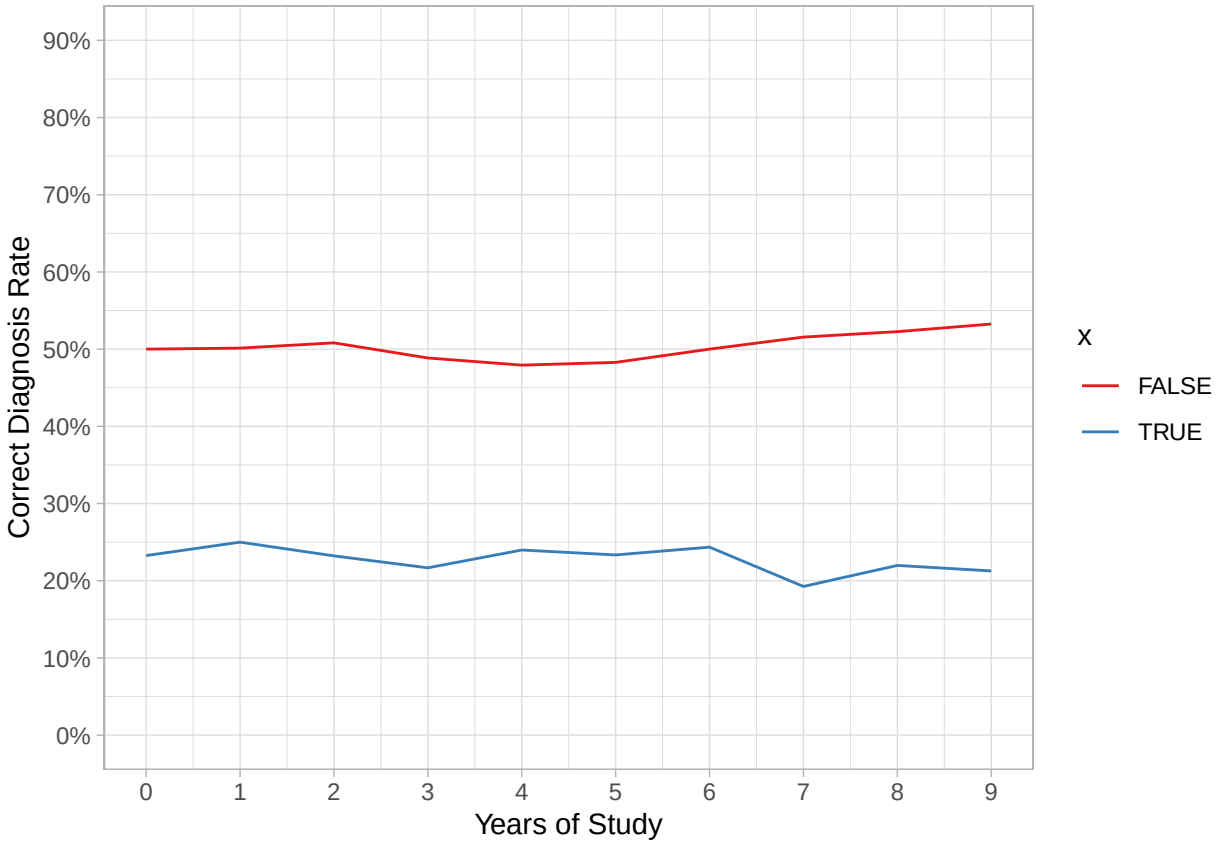
As shown below, the misdiagnosis rates prior to the model being deployed are essentially identical to the misdiagnosis rates when doctors follow the model exactly.



What happens when this process is iterated

Now we imagine that this is part of a learning healthcare system. Every year, a new batch of patients comes to the hospital, and is diagnosed (or isn't). Every year, the researchers retrain their model based on the previous year's diagnoses. The plots below suggest what happens: neither quality of care nor our understanding of the disease improves. Our estimates of disease prevalence are just as incorrect as ever, and the misdiagnosis rates do not improve for either group. In fact, from year-to-year, the estimated model is largely the same; the parameter estimates get closer or farther from the true values at random, but over time the model does not systematically improve or degrade; it merely reinforces what was already being done.


0.100
0.075
0.050
0.025
0.000
Estimated Prevalence
0
1
2
3
4
5
6
7
8
9
Years of Study
X
FALSE
TRUE


90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Correct Diagnosis Rate
0
1
2
3
4
5
6
7
8
9
Years of Study
X
FALSE
TRUE

The multi-center case

Now let's suppose there are multiple hospitals using a data-driven approach. Each one trains a "diagnostic model" on it's own set of data in year 0 and each subsequent year. The diagnostic patterns of each hospital appear as somewhat of a random walk. A hospital which has a particularly accurrate or inaccurate diagnostic year by chance will build a particularly accurate or inaccurate model, respectively, in the following year. Unsurprisingly, if the hospitals are very small, their diagnostic rates will vary more wildly from year to year.