Diagnosis with Polya Urns

Rachael Caelie (Rocky) Aikens

12/21/2020

Introduction

This analysis is going to be mathematically identical to the Polya Urn results presented by Ensign et al.. I'm going to implement the simulations with a structure that emphasizes the diagnostic process format, rather than the Polya Urn format, but all of the mathematical results derived under the Polya Urn format still hold.

In many ways, the doctors in this simulation will do something very clearly foolish: they will neglect to account for selection bias as they learn from past data. In the field of reinforcement learning, this would be very quickly identified as a bad algorithm for learning from data. In the field of medicine, this kind of mistake happens perhaps much more often than we would like to admit. When we build models - either with machine learning or with more classical statistics - to identify risk factors for a disease or the symptoms of that disease, we seldom correct for the processes that direct who is *considered* for the diagnoses we use as our "ground truth."

Case 1: Underlying rates do not differ

Suppose that a doctor is trying to decide whom to test for a disease. Maybe this is a literal test like a blood test, or maybe this is something more metaphorical like a physical exam or a referal to a specialist. We'll assume for the moment that the test is perfect - a person gets a positive test result if and only if they have the disease.

Suppose also that there are two different types of people seeking the test - for example men and women - but the underlying rate of disease is not known in either group. Let λ_A and λ_B represent the rates of disease in each group. For now, we'll assume $\lambda_A = \lambda_B = \lambda$.

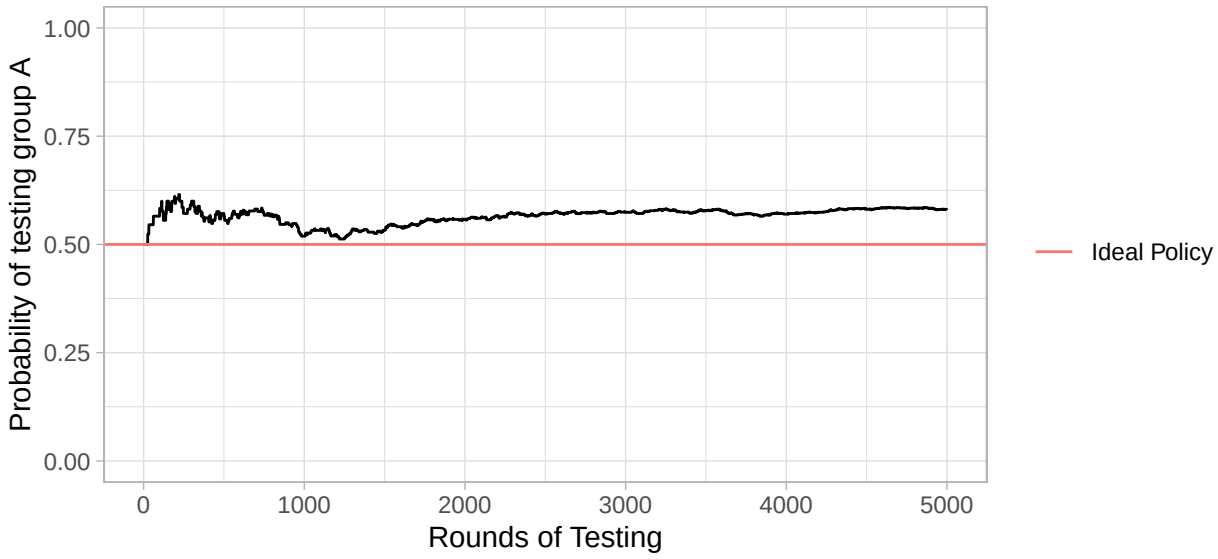
Suppose also that the doctor would like to choose a testing policy based on the following goal:

Goal: If a group has Λ percent of the case burden, they people from that group should recieve Λ percent of the tests.

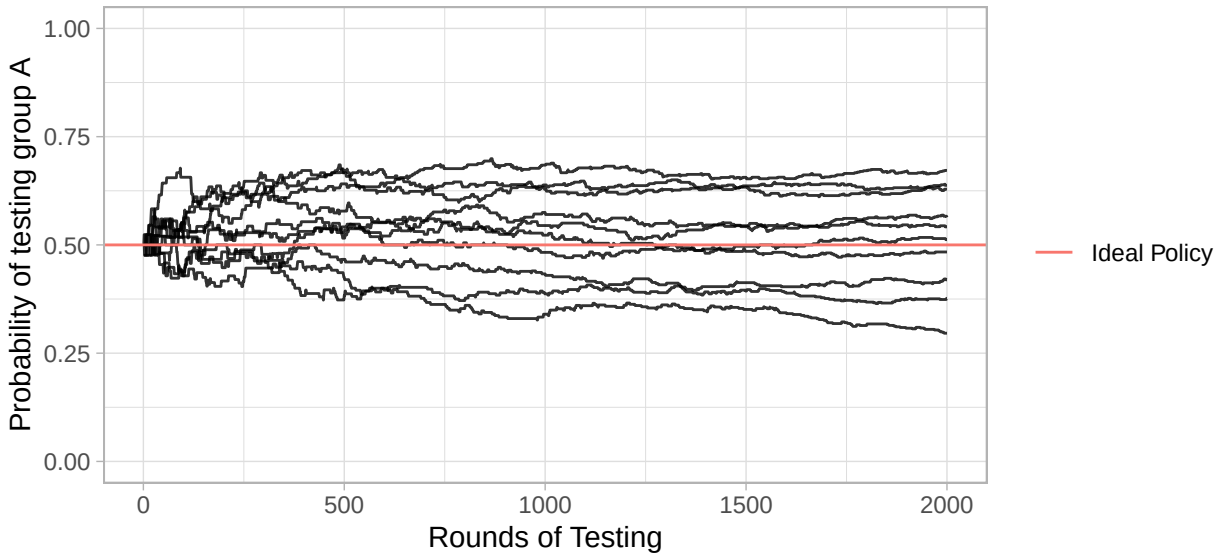
In order to choose whom to test next, the doctor selects based on the existing evidence as follows:

- Let n_A and n_B represent the number of observed cases in groups A and B.
- The doctor tests someone from group a with probability $\frac{n_A}{n_A+n_B}$ and they test someone from group B with probability $\frac{n_B}{n_A+n_B}$. That is, the next test is selected based on the fraction of observed cases attributable to each group.

Let N_A and N_B represent discovered cases when the experiment is started. We'll start with $N_A = N_B = 10$ and $\lambda = 0.1$. The plot below summarizes one example simulation. Even though the seeded data was exactly equal and the underlying rates are the same, this feedback loop settles on a policy of testing one group at a higher rate than another. An ideal policy would test both groups at the same rate, since the underlying disease burden is the same. (We'll assume for simplicity that both groups are the same size.)



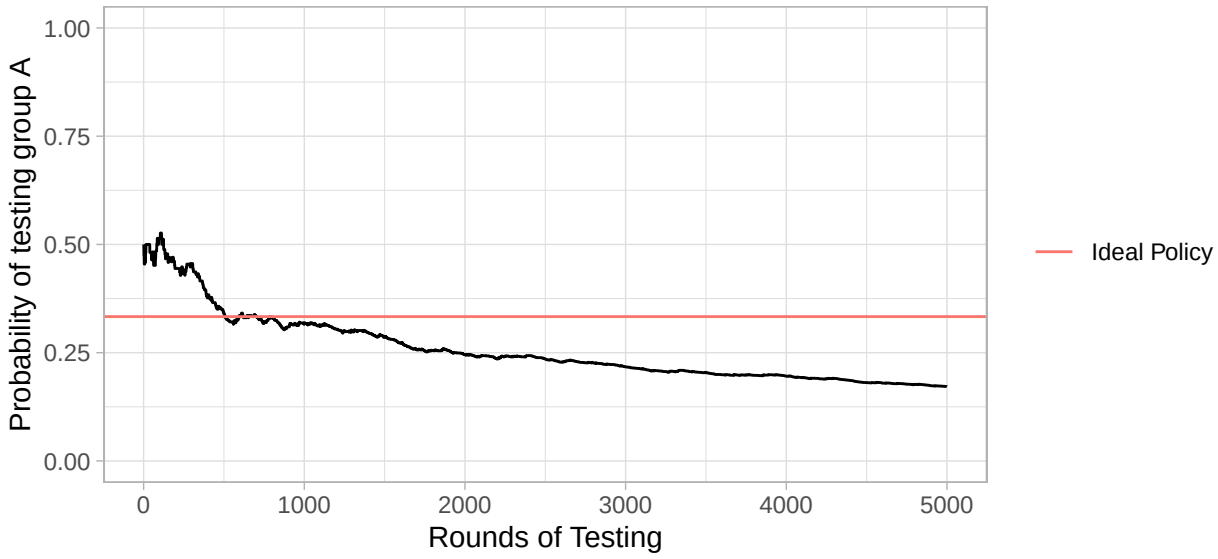
What happens if we run this simulation multiple times? The plot below shows the results of many different simulations with the same starting parameters.



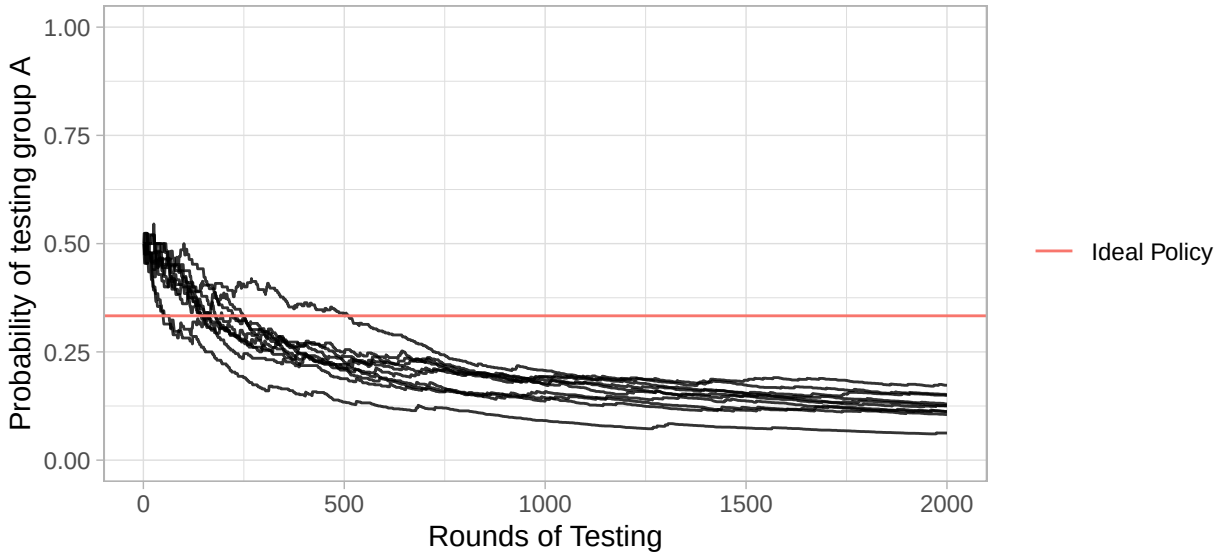
The "doctor" always converges to some policy determining the probabilities of testing one group or the other, but the final policy is different across simulations! Sometimes the policy is close to the ideal (testing both groups at equal rates), but other times it is entirely wrong. This fact is actually well-characterized mathematically from the literature on Polya Urns! Let p_A be the limiting probability of deciding to test a person from group A. If N_A and N_B are the starting counts of observed cases when the simulation begins then p_A follows the distribution $Beta(N_A, N_B)$ (Redlund, 2010).

Case 2: Underlying rates differ

How does this behavior change when the underlying rates of the disease do differ? Let's let $\lambda_A = 0.1$ and $\lambda_B = 0.2$. That is, being in group B doubles one's odds of having the disease. Based on our goal above, an "ideal" policy would allocate $1/3$ of the tests to group B and $2/3$ of the tests to group A.



Even though an ideal policy would allocate 1/3 of the results to group A, the doctor converges to a much lower probability of testing group A. In fact, as we see below, this result happens systematically in every simulation:



Now the doctor will systematically be less and less likely to test people from group A over time. This system is also characterized well mathematically! In fact, if p_A represents the asymptotic probability of testing a person from group A as before, then $p_A = 0$ if $\lambda_A < \lambda_B$ and $p_A = 1$ if $\lambda_B > \lambda_A$. That is, if the prevalence of the disease is just slightly higher in one group than the other, doctors will eventually converge to testing that group *always*.

Conclusions

In practice, the simulations above don't perfectly capture medical decision-making. In reality, doctors don't update their beliefs about a disease after every test, which means there is more delay in the feedback loop. Eventually, they also probably stop looking for new information, settling on ideas about the relative prevalences of disease that aren't continuously updating. But these simulations and mathematical results illustrate how things might go wrong when we fail to think about selection bias influencing who is *considered*

for a diagnosis. In particular, we may come to conclusion that differences in prevalence exist when they do not, and we may *overestimate* differences in disease prevalence when they do exist.