Pre-read: Healthcare (Part 2)

COS 597E/SOC 555, Princeton University, Fall 2020

Thursday's readings are about limits to prediction in healthcare: ethical concerns about predictive technologies, pitfalls of performance evaluation, and scientific limits to prediction.

The first two readings are about social suicide prediction, which is the use of the digital traces that people leave online in order to predict suicide risk, in the hope of intervening to prevent suicide. We chose this application in part because it uses non-medical data. This type of cross-domain repurposing of data is both scientifically exciting — social suicide prediction works much better than in-domain prediction based on questionnaires administered by medical professionals — but also raises a slew of ethical concerns.

The next reading is about critically evaluating claims regarding deep learning in healthcare (specifically, medical imaging). Claimed breakthroughs in prediction in any domain should be subjected to this kind of scrutiny. How does the deep learning paper from Tuesday fare w.r.t. the criteria used in this paper? On a side note, make sure to check out Appendix 3 for a practical example of pre-registration and what sorts of deviations from the pre-registered protocol are justifiable. You may also find it interesting to see the paper's entire prepublication history, including draft versions and peer reviews. How does this compare to the peer review practices in your field?

Now let's turn to the final pair of readings. To see the relevance to limits to prediction, we need some background. The possibility of predicting disease risk and other health outcomes from our genomes was opened up following the completion of the Human Genome Project in the early 2000s and the rapidly advancing ability to genotype people. This led to a proliferation in Genome-Wide Association Studies (GWAS) in the 2000s. Such studies aim to identify associations between genotype and phenotype. But the results of GWAS were almost uniformly disappointing: the genetic variants they identified could explain only a small proportion of the variance in outcomes, not just for diseases but also for physical traits like height.

Is the disappointing performance of GWAS because diseases and traits are fundamentally unpredictable from genetic data, or because current methods fall far short of what is possible? Unlike most of the other domains we've discussed — ads, life trajectories, geopolitical events — there is a well known way to answer this, and that's what makes this domain so interesting to us. This is the notion of heritability. Heritability estimation has been intensely studied for over a century and doesn't require genotyping at all. Recall our running theme that measuring unpredictability is very different from measuring predictability.

A historical interlude: understanding inheritance was one of the major motivating applications that led to the birth of modern statistics. Foundational concepts such as correlation, and variance and techniques such as regression and ANOVA all seem to have been pioneered in this context. Francis Galton and Ronald Fisher were two people who made enormous contributions to statistics and to understanding heredity. It is important to know that both were eugenicists and this motivated their work in statistical genetics. Today, research on heredity continues to be misused in scientifically dubious ways toward reprehensible purposes.

Heritability is defined as the fraction of variation in the phenotype that is due to variation in genotype. This is a fuzzy definition, and different ways of estimating it will give different results. But interestingly, different estimates broadly agree. For our purposes, we will interpret heritability as a measure of predictability of phenotype from genotype (more precisely, 1-heritability as a measure of unpredictability). For example, identical twins reared apart tend to have highly correlated heights. If we estimate a heritability of 0.9 for height using this method, it suggests that predicting height from genotype alone will have an accuracy of at most 90% (measured by R^2), since 10% of the variance in height is due to the environment which is unknown. But it doesn't give us a way to actually predict height from genotype with 90% accuracy. The gap between the known limits to prediction and the known predictability is called the missing heritability problem, and the Nature news article is about this.

Identical twins reared apart are rare, so to estimate heritability in practice, we need methods that use easier-to-obtain data. A common method is to compare identical (monozygotic) and fraternal (dizygotic) twins based on the assumption that the former share all their DNA and the latter share half. The Krasby et al. reading is about this method.

Our interest in heritability is not primarily about genetic prediction itself and more about what we can learn from it that might apply to other domains. Here are two observations. First, the existence of both "lower bounds" and "upper bounds" on prediction error, and the gap between them, seems to have spurred a prodigious volume of research and served as a reality check for researchers. Second, nature creates "experiments" in the form of twins that allow heritability estimation. Last week we saw that the weather people were inspired by this. Can we find analogous experiments in yet other domains?