Predicting TV Virulence: Employing Deep IV for Identifying Virulence Causative Genes

Jack Andraka
jandraka@stanford.edu

Billy Ferguson
billyf@stanford.edu

Charlie Walker
cwalker4@stanford.edu

1 Introduction

Tuberculosis (TB) is a major public health issue. Despite the development of potentially curative chemotherapy, TB continues to cause increasing worldwide morbidity and is a leading cause of human mortality: 1.7 million individuals succumb to TB annually, and over 10 million more individuals contracting the disease each year.[1] Recent advances in bacterial molecular genetics, immunology, and human genetics have yielded insight into the molecular determinants of virulence; however, a large knowledge gap still exists that limits the development and testing of new interventions, including novel drugs and efficacious vaccines.[2]

The genome-wide association study (GWAS) is an experimental design used to detect associations between genetic variants and traits in samples from populations and has served as a main driver of this deeper understanding of TB virulence.[3, 4] However, traditional GWAS study designs are limited in that they run several thousand to million t-tests simultaneously, meaning that only gene variants with large effect size can be identified. This is in contrast with the current hypothesis of genetic architecture that posits the existence of many gene variants acting in tandem to produce a phenotypic trait, each with a small effect size.[5] To address these shortcomings we will employ a two-stage deep neural network on the Tuberculosis Gene Expression Dataset from the Khatri Lab to deduce the size of variant effect on tuberculosis virulence.

2 Methods

We will implement the method described in "Deep IV: A Flexible Approach for Counterfactual Prediction" to identify the effect of gene expression on virulence of tuberculosis.[6] GWA studies have historically been unable to recover significant relationships because of the endogeneity between gene expression and virulence and the enormous quantity of genes (on the order of 4000). Instrumental variables (IV) are a well-developed tool for remedying endogeneity, but require a strong prior understanding of the data generating process and are not well equipped to deal with a large number of covariates. Deep IV promises to marry the best qualities of DNNs and IV, and we believe GWAS are a prime use case.

To perform IV analysis one needs to find an exogenous variable that affects the outcome variable only through the endogenous covariate of interest. More specifically, the instrument z must be conditionally independent of the error (Figure 1).

Traditional IV can be estimated through a procedure called two-stage least squares (2SLS): in the first stage you regress your endogenous variable of interest, p, on the exogenous instrument, z, to create a predicted $\hat{p}$ constructed only with the exogenous variation of z. In the second stage, you regress your outcome variable, y, on the predicted $\hat{p}$ from the first stage.[1]

[1] Chapter 4 of Angrist and Pischke's *Mostly Harmless Econometrics: an Empiricist's Companion* provides a good introduction to instrumental variables.

Figure 1: Generalized Deep IV

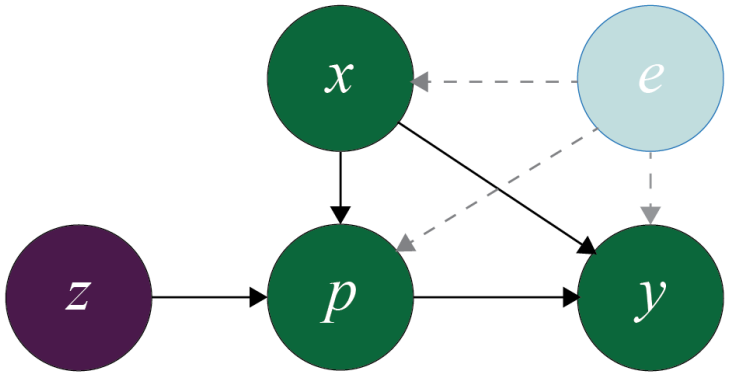
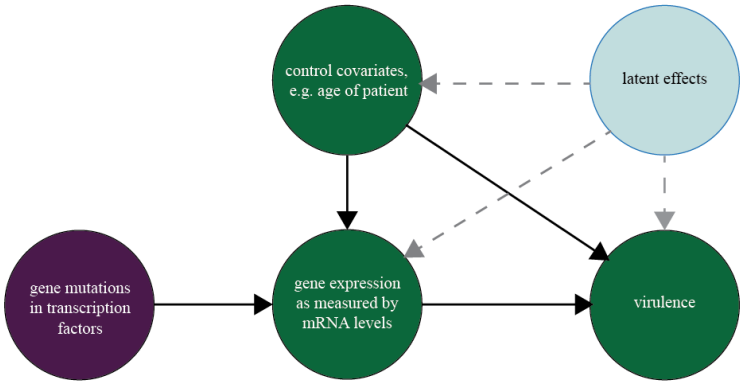


Figure 2 shows the general IV formulation applied to our question. As described in the introduction, the effect of gene expression on observed outcomes (in our case, virulence) is of enormous interest, but the direct estimation of these effects has thus far been unsuccessful. By instrumenting for gene expression with mutations in transcription factors we hope to more accurately predict virulence, as well as develop a more robust understanding of causal effects.

Figure 2: Deep IV Application



3 Challenges

Given that the Deep IV methodology was just developed this past year, there are no other published articles performing this analysis. The Hartford et al. (2017) paper will therefore be our only reference describing the method and we will have to carry out the application of this mostly theoretical research. Moreover, GWAS are notoriously difficult given the heavy interaction between genes. Hopefully DNNs can handle this challenge, but not much can be said about the effectiveness of DNNs for our problem given that no one has carried out this analysis.

4 References

[1] World Health Organization. *Global Tuberculosis Report 2017*. 2017. Print.

[2] Forrellad, Marina A. et al. "Virulence Factors of the Mycobacterium Tuberculosis Complex." *Virulence*, vol. 4, no.1, 2013, pp. 3-66., doi:10.4161/viru.22329.

[3] Uren, Caitlin, et al. "A Post-GWAS Analysis of Predicted Regulatory Variants and Tuberculosis Susceptibility." *PLoS One*, vol. 12, no. 4, June 2017, doi:10.1371/journal.pone.0174738.

[4] Bermingham, M L, et al. "Genome-Wide Association Study Identifies Novel Loci Associated with Resistance to Bovine Tuberculosis." *Heredity*, vol. 112, no. 5, May 2014, pp. 543?551., doi:10.1038/hdy.2013.137.

[5] Korte, Arthur, and Ashley Farlow. "The Advantages and Limitations of Trait Analysis with GWAS: a Review." *Plant Methods*, vol. 9, no. 29, 22 July 2013.

[6] Hartford, Jason, et al. "Deep IV: A Flexible Approach for Counterfactual Prediction." *International Conference on Machine Learning*. 2017.