One of the leading causes of excess death in the United States over the past two decades is the misuse and abuse of opioids, a problem that has gotten worse in recent years via the rise of potent synthetic opioids like fentanyl. In many cases of misuse, the inciting incident is a legitimate surgery to which opioids are prescribed as part of a pain-management regiment. After the prescription opioids are cut off, patients turn to off-market or illegal opioids to deal with chronic or lingering pain, leading to abuse. As such, it is important to accurately predict whether or not a patient will abuse postoperative opioids based on prior information. One strong indicator of postoperative abuse is prior opioid use, legally or not. However, patients are not always honest in regards to prior use, especially when they believe their answers will be used to withhold pain relief from them in the future, and so patient information can be used to predict preoperative opioid use, which is then used to predict postoperative opioid abuse. This is the aim of a team from the University of Michigan, who have combined the benefits of neural networks and logistic regression to create a model that is more easily interpretable in its predictions to both the doctor and patient, while sacrificing minimal accuracy. However, no matter how accurate a model is, there is the risk that it will unfairly stereotype certain groups or demographics, and perpetuate long standing issues regarding the pain of minority groups and women not being accurately recognized by physicians.

The central innovation of the paper is to combine a deep neural network (DNN) with a logistic regression (LR) to produce what they call an Interpretable Neural Network Regression (INNER). Instead of being used to directly produce a classification, the DNN is asked to produce regression coefficients that are then used in a logistic regression, which is then used in conjunction with a constant cut-off to produce a classification. While DNNs are powerful, they are often completely black-boxed and the intermediate steps leading up to the prediction are uninterpretable. The advantage of this method is that it combines the DNNs potential to discover more obscure patterns in the data through its multilayered structure, while keeping the interpretability of a LR in that, for every patient, we have an easier way to interpret the patients needs, as well as being able to take advantage of the LR's ability to give a probability, rather than just a yes-no prediction. The structure of the logistic regression is to take in a constant Baseline Opioid Tendency (BOT) term, which characterizes innate tendency to abuse opioids habitually or without pain, and a Pain-induced Opioid Tendency (POT) term, weighted by a categorical pain score on a 1-10 scale, which characterizes opioid abuse that is specifically spurred on by severe pain. Thus we have a model that is a function of both stable patient characteristics as well as temporary pain at the time of the operation. For example, this model helps us to understand if there are certain groups that are more likely to abuse opioids no matter what level of pain they are experiencing, vs groups that are only likely to abuse if they are experiencing a very high level of pain.

The first section of the paper uses simulated data to compare the performance of INNER with other models in order to test its robustness on data with varying degrees of complexity, noise, and dimensionality. INNER is compared with a standard DNN, a standard LR with multiplicative interaction between the two inputs (meant to represent the patient characteristics and pain score in the real data), a random-forest model, a decision tree model, a SVM classifier, and a BART (Bayesian Additive Random Trees, a conglomerative model that uses multiple random trees) model. Each of the models are cross-validated on the data and evaluated based on testing C-statistic. The C-statistic measures the area under the ROC (Receiver Operating

Characteristic) curve, which plots the classification error of a model as a function of the threshold or cut off point used to make a classification. The statistic ranges between .5 (random chance) and 1 (perfect classification). In this simulated data, the INNER model consistently outperforms every model except the DNN, including far outperforming the logistic regression. It is also able to obtain impressively high C-statistics (> .95) even with a large number of noise variables, showing it's resilience over the tree-type models, and even outperforms the DNN when there is a high number of true covariates. However, given that the data was constructed with the model of the INNER framework in mind, it should not be surprising that it performs as well or even better than the more general DNN. Notably, the models are not tested on very high dimensional (p>n) data, which would need confirmation if we were to apply INNER to something like chemical or genetic data, however in this case the number of predictors is small enough to not worry about.

In applying the INNER model to the opioid use data, the researchers first make decisions to clean and balance the data. As only 23%~ of the total data comes from opioid users, the researchers balanced the model by excluding a random subset of the nonuser data, in order to balance the ratio to 50/50. Any data with missing values was substituted with the mean (or categorical mode) value instead of being excluded, which is advantageous as it would have led to a 20% reduction in sample size. The researchers use a cutoff line of .5 for the final sigmoid classification for all training, although they acknowledge in the discussion that this could be tweaked based on differing costs for type 1 vs type 2 misclassification. In generating the model, race and sex variables were included. Notably, there was no attempt to normalize the data via exclusion or bootstrapping in regards to race in a similar fashion as opioid use, despite 89% of the respondents being identified as white. On the opioid data, the INNER model achieves comparable (not statistically different) results of C-statistic, overall accuracy, sensitivity (true positive rate), and specificity (true negative rate) to the DNN, with both networks being substantially better than a standard logistic regression. The INNER and DNN networks are optimized via stochastic gradient descent, and both networks have dropout layers to prevent overfitting. The number of layers and nodes within each layer are themselves tuned prior to training, with the researchers settling on a model with 3 layers of 250, 125, and 1 node respectively. The paper does appear to blatantly misquote one of its own charts at one point, wherein the table reports the DNN accuracy to be .76 but a few paragraphs later it is reported as being .72, in comparison to the .72 accuracy of the INNER model. I can't tell whether this is an actual error or just me misreading something beyond my depth. In interpreting the results, the researchers do some subjective clustering analysis, dividing the patients into 3 overall risk groups, and 6 subgroups based on reported levels of BOT and POT. They then use a ANCOVA metric to identify the most important features in determining opioid use for each subgroup. Some of the most important factors in determining risk are measures of general wellbeing such as the Fibromyalgia survey score (measures chronic pain), Charlson Comorbidity Index and ASA index (both summarize various chronic diseases and quality of life factors). Admitted illegal drug consumption and tobacco consumption are also highly correlated with opioid use. Interestingly, alcohol use is negatively correlated with opioid use, suggesting that some patients are able to use different drugs to manage their pain, however they admit that this result is not concordant with previous literature.

Although the response variable in this study is preoperative opioid use, the most obvious implication, which is referenced by the authors in the introduction, is in the prescription of postoperative opioids for pain management. The issue of whether black people's pain is treated as less important is both a historic and current topic, as seen in Advil's Believe My Pain campaign, which highlights the history of racial bias in administration of pain-relief medication. If the INNER model predicts that black people are at worse risk of abusing opioids, then legitimate cases and grievances could be denied and ignored at a higher rate. There may be evidence of this in that "African Americans constitute about 17% of patients in the high risk group, while there are only 5% African Americans in the low risk group." (13). The model structure itself also assumes everyone's reported level of pain is equal, given by the flat 1-10 score. Pain itself is subjective, so if some groups report less pain due to personal pride, they may be misclassified as having a higher POT, due to taking opioids to combat "less pain". The model may need to be tested using statistical fairness models within these groups, and modified if it is discriminatory. Possible ways to encourage fairness may be deliberately blinding the model to race, or using a lasso-type penalty on the resulting regression to encourage zeroing out of race-tied variables. There is also the issue of data privacy in regards to who should have access to this data, and where it should be stored. While the information might be necessary for doctors, what about insurance companies? If the data is compromised and leaked, who is responsible for potential losses incurred by the patients? These questions point to the idea that proper anonymization and data cleanliness would be necessary when using this (or any other similar) model which utilizes private patient data.