Stor 390 Final

Samuel Reid

2024-05-03

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. The CDC reports that opioid deaths have increased by six times since the turn of the century, and synthetic opioid deaths have increased twenty-fold since 2013. However, since 2000 75% of those seeking treatment for opioid abuse reported that their first usage of opioids was a legal prescription of pain medication, meant to treat chronic or severe pain. After the prescription opioids are cut off, these patients turn to off-market or illegal opioids to deal with their pain, leading to substance abuse. The eventual worsening in patient health due to opioid dependency, withdrawals, or overdose, as well as surrounding effects on the patient's family and community have already cost the US trillions. This motivates the need to accurately predict whether or not a patient will abuse postoperative opioids based on prior information, as opioids are not a necessary part of a pain management plan. One strong indicator of postoperative abuse is prior opioid use, legally or not. However, patients are not always willing to disclose prior use, especially when they believe their answers will be used to withhold pain relief medicine 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, while sacrificing minimal accuracy. However, there is no perfect solution to this problem, or else it would have been solved, and this paper will examine both the model's benefits and its shortcomings, mainly questionable 'interpretability' and potential for moral misuse.

The central innovation of the paper is to combine a deep neural network (DNN) with a logistic regression (LR) to produce what the researchers term 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. The benefit of this model structure is to give us some level of interpretability, by understanding 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. Central to this plan is the assumption that these groups actually are different, and as such can be treated in different ways once knowledge of their demographics are learned.

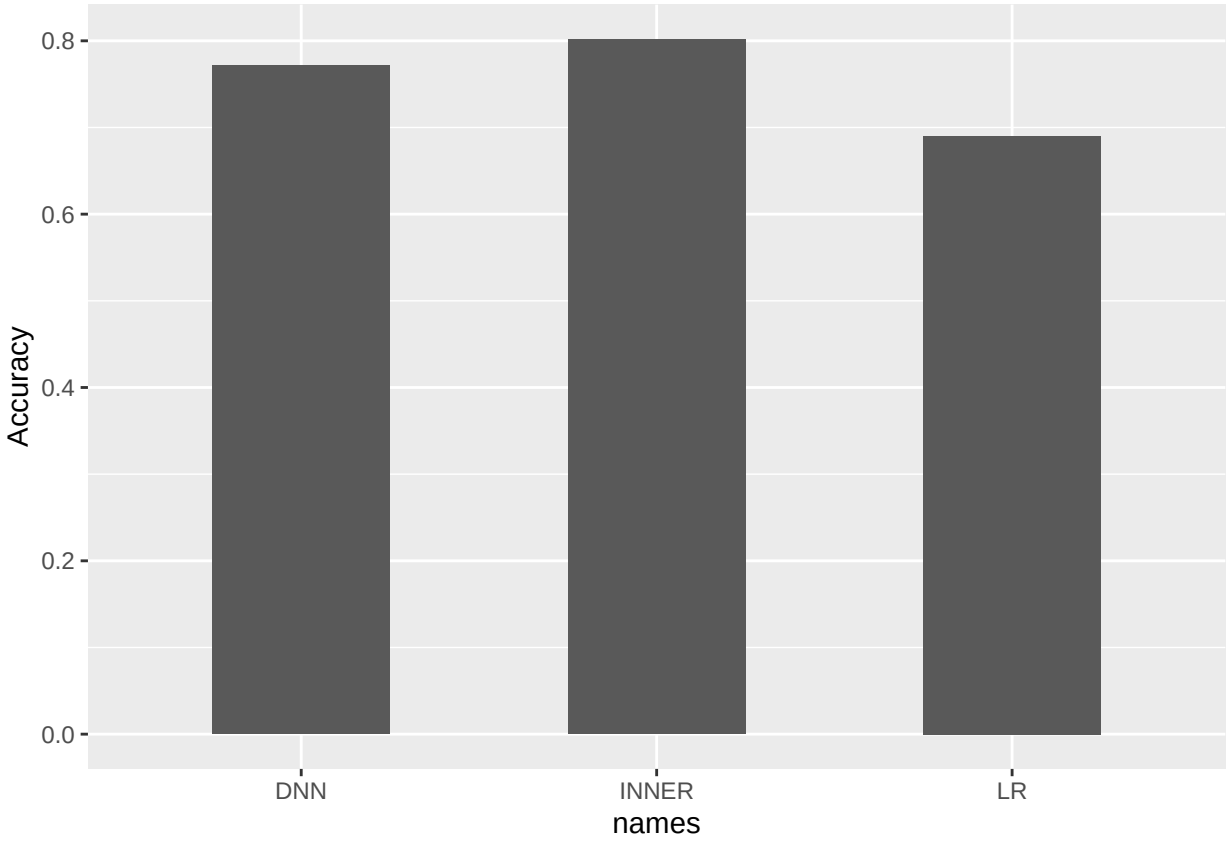
The first section of the paper uses simulated data to compare the performance of INNER with other models in order to test its robustness on data with varying degrees of complexity, noise, and dimensionality. INNER is compared with a standard DNN, a standard LR with multiplicative interaction between the two inputs

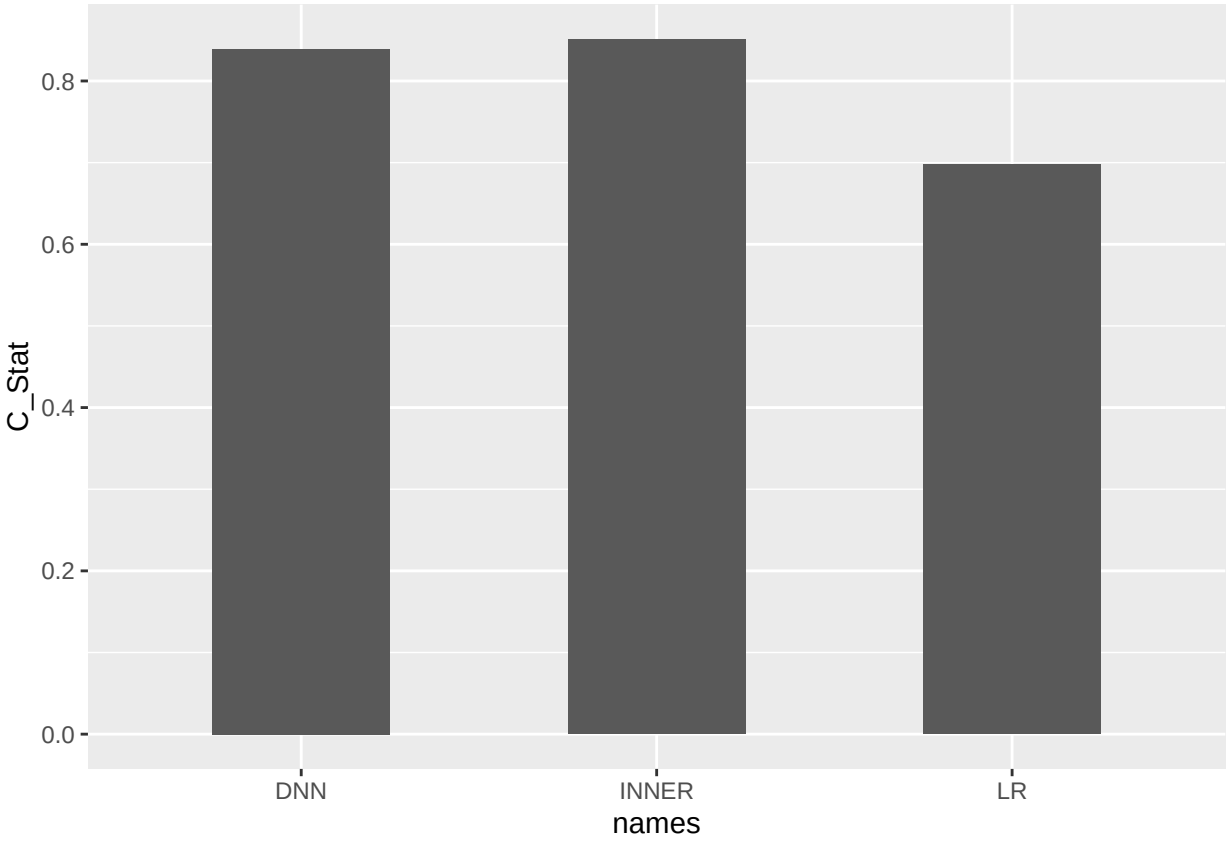
(meant to represent the patient characteristics and pain score in the real data), a random-forest model, a decision tree model, a SVM classifier, and a BART (Bayesian Additive Random Trees, a conglomerative model that uses multiple random trees) model. Each of the models are cross-validated on the data and evaluated based on testing C-statistic. In this simulated data, the INNER model consistently outperforms every model except the DNN, including far outperforming the standard 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 the case of the opioid dataset the number of predictors is small enough for this to not be a concern. This section compares each model on 3 signal-to-noise ratios, and then 3 amounts each of total samples and total covariates on the same ratio. Although I understand the constraints of attempting to display multi-dimensional tables on a 2-dimensional chart, I would have liked some justification as to why the constant variables were chosen when varying the other factors. I would also like some justification that the StN ratios covered a full spread of possibilities that may be seen in real-world data, particularly below the lowest tested value of .2. That is not really a problem in this case, as we are only dealing with a single collection of noise and size descriptors, but it would be something to test before implementation on more datasets. I do commend the prior testing for learning rate included in the supplemental materials

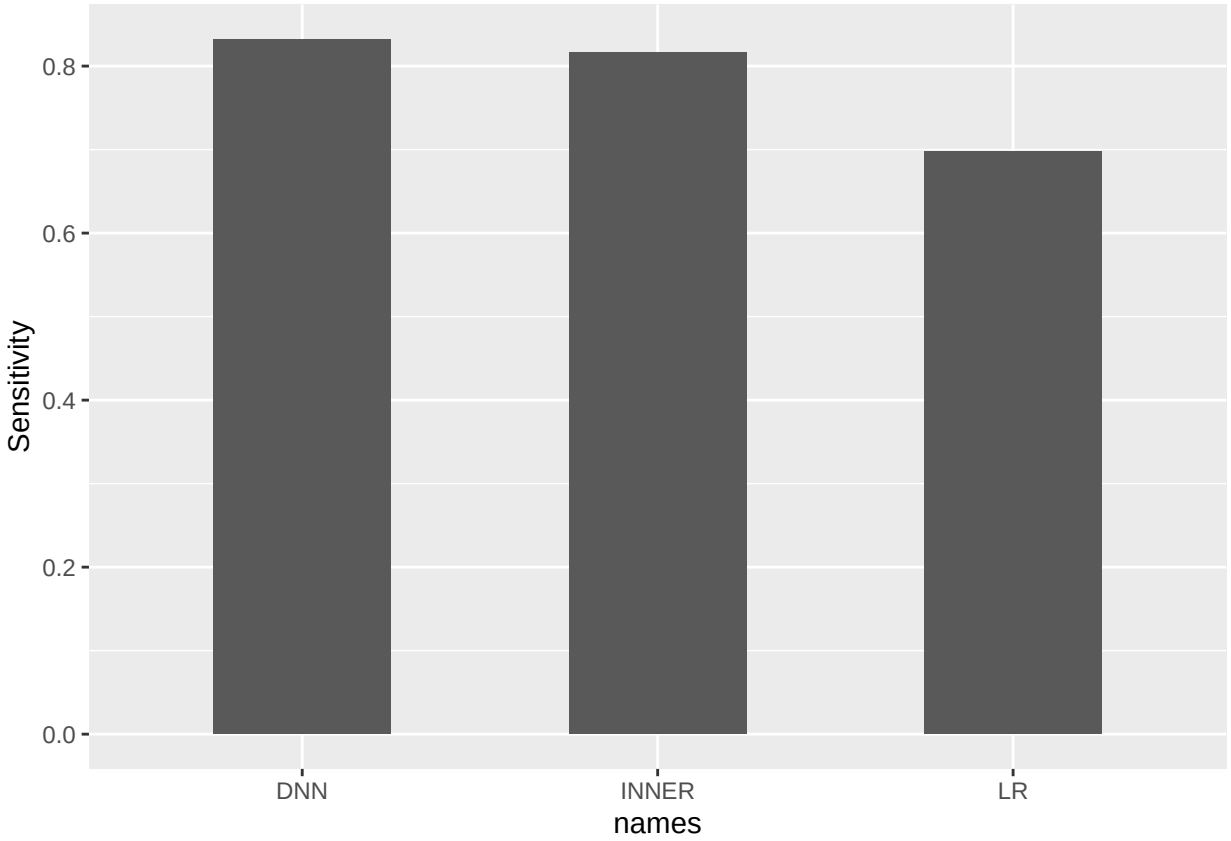
In applying the INNER model to the opioid use data, the researchers first make decisions to clean and balance the data. As only 23%~ of the total data comes from opioid users, the researchers balanced the model by excluding a random subset of the nonuser data, in order to balance the ratio to 50/50. 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. 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.

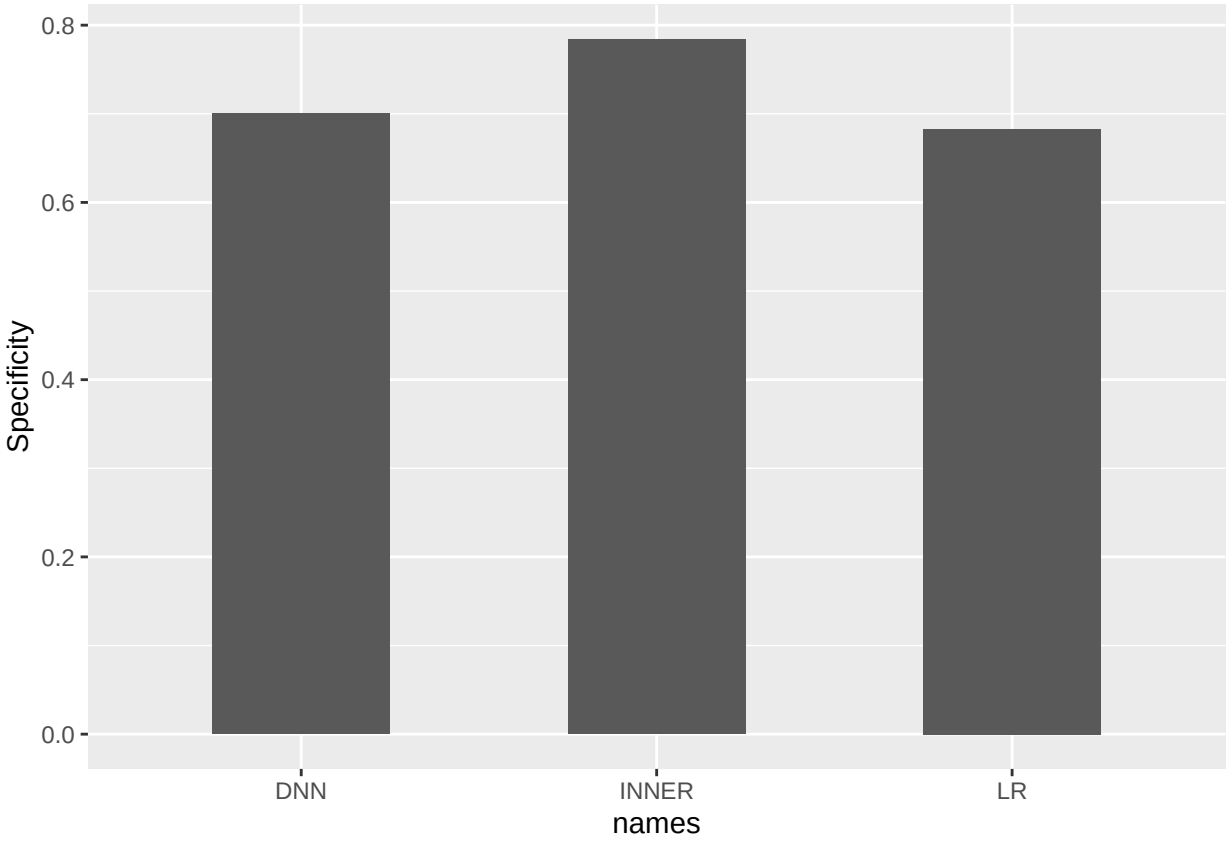
I have applied the INNER model to a dataset predicting heart disease based on 12 different health-related factors, and have compared it with the standard DNN and logistic regression models for comparison. The dataset comprises both immutable and changeable characteristics, and can be separated into 'exercise- and non-exercise-related' predictors, which will be the two groups in the INNER and LR models.

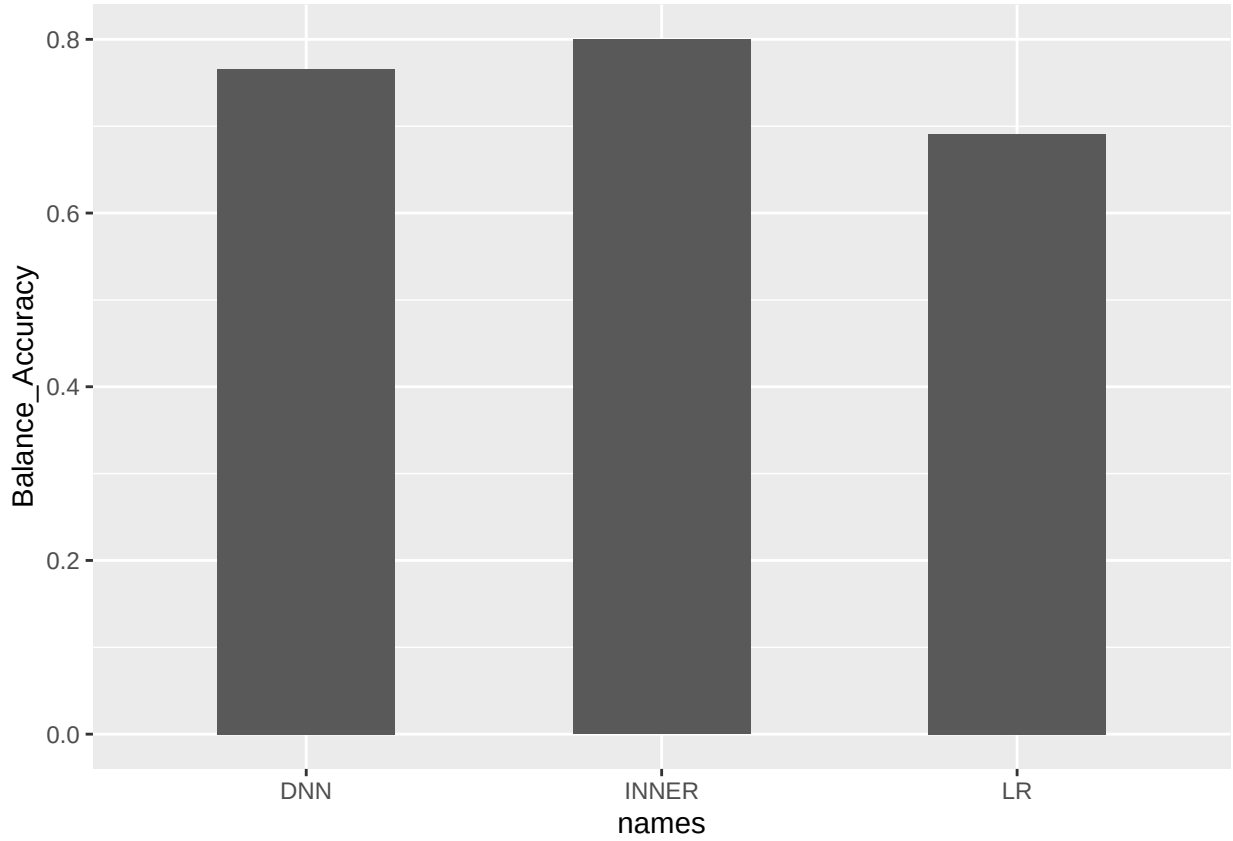
```
## Warning: package 'ggplot2' was built under R version 4.3.2
```


0.8
0.6
0.4
0.2
0.0
Accuracy
DNN
INNER
LR
names


0.8
0.6
0.4
0.2
0.0
C_Stat
DNN
INNER
LR
names


0.8
0.6
0.4
0.2
0.0
Sensitivity
DNN
INNER
LR
names


0.8
0.6
0.4
0.2
0.0
Specificity
DNN
INNER
LR
names



Shockingly, the INNER model was the best-performing model of the three in almost every metric, including raw accuracy and C-statistic, and was only beaten out by the DNN in sensitivity. These statistics were averaged over 25 trials, so it's not a fluke. The only reason I can reasonably attribute is that the exercise-non split is exactly the kind of interaction this model was made to capture, and so it had the exact right amount of model specification to excel. Unfortunately, race was not included in this model, so I wasn't able to compare that angle of the results.

All that said, 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. 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 rates at which opioids are released to the public has many important consequences, but one that can be statistically analyzed in the short term is disparate impact. 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. On the surface, this is exactly the kind of problem INNER seeks to treat: if expressions of pain are too subjective or cultural, a more objective metric is needed to remove that bias. 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 is 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 itself is also far less subjective than it may seem, as it factors in these subjective reported measures of pain. There is a benefit in having these reports be one step removed from patient-doctor communications, as it's easier to put aside biases when reading a number off a screen than when hearing it directly from someone's mouth, even if they come from the same source. 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 issue at hand is how to effectively minimize harm, both due to improper prescription of opiates and improper withholding of opiates, and so we will

appeal to the harm principle. I see three main questions to address. First, pulling back to the widest level, should highly addictive drugs ever be prescribed? There is a strong argument against this position. Pain by itself is not deadly, and is often more temporary than it appears to the person experiencing it. Opiates are not necessary for treating pain, despite clearly being very useful. Meanwhile, overdoses can kill, as do infected needles, as do drug users who are robbing someone to fund their addiction. The harm spreads outside the user, consuming funds needed to keep families afloat, stagnating communities, funding cartels and their innumerable human rights abuses. The strongest argument to the contrary, and what I will operate under from here on out, is that of scale. How many hundreds of millions of lives have been substantially improved over the past two centuries compared to those affected negatively? The dangers of opiate prescription may be mitigated, but until we find some miracle all-purpose treatment, its benefits simply can't be replaced. I would also claim that addiction and slight financial stress sometimes isn't that bad compared to constant, debilitating pain, and that many opiate users become addicted primarily as a response to outside stressors, rather than an innately addictive personality. A 2018 study found rats voluntarily gave up heroin when instead presented with choices of food and socialization. Outwardly improving society is obviously a task by itself, but this room for improvement is a way forward, and it shows a clear path to a harm-negative outcome, which does not similarly exist for the no-opiate model. Now, should race, or any non-controllable behavior be taken into account when assessing risk for abuse? That depends on whether or not there really are tangible differences between those groups in regards to pain tolerance. If there aren't, then this type of bias is necessarily causing harm by overcorrecting in some areas and undercorrecting in others, as established. If they are, then it still shouldn't be up to the doctor to decide. Fat people can't be charged more at a buffet the second they walk in, so until some court enshrines medical discrimination, that decision shouldn't be left up to the individual doctor or insurance agent on a case-by-case basis. Finally, should algorithms play a role at all in this process? There is an intuitive appeal to having your life be determined by the will of a human. A doctor can be convinced if you are persuasive enough of the truth, an algorithm may only be convinced if you lie. The potential general harm done to society by the depersonalization and lack of direct accountability inherent in this model is significant. However, the potential for good is even greater, directly mitigating the downstream effects of opioid abuse mentioned earlier, and (if done correctly) in a way that may actually increase patient trust in the medical system. Still reliant on the algorithm actually helping, yes, but the possibility is definitely there. We have determined that it is ethical to allow algorithms to prescribe opiates, as long as they are not making discriminating choices based on protected characteristics. INNER, as presented, absolutely does discriminate between inherent characteristics, both via disparate impact, and by design, as these variables are fed into the model itself. INNER still needs to be tested further using statistical fairness models between 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.

On the whole, the INNER model is useful and impressive, but falls short of being revolutionary. It is most applicable when an outcome can be reasonably split into two (or a small number of) competing sets of factors which are more correlated within themselves than between sets. The success of INNER on my heart disease dataset leads me to think that maybe the opioid dataset simply shouldn't be split in that manner at all. The interpretation gained is really not that high, as knowing whether or not someone has a high or low baseline or pain-induced tolerance score is literally about as useful as a political compass. Going forward, if I were to use this model, I would focus on identifying which situations lead to this kind of predictor split, and finding less-invasive methods to test for those patterns. I also think this could have an interesting role as just one of a family of models spanning between the DNN and LR that could be tuned over for the optimal number of sets of predictors. I would also exclude immutable-class-related variables, as I seriously doubt there are improving accuracy by enough to risk a lawsuit.

Citations Yuming Sun, Jian Kang, Chad Brummett, Yi Li "Individualized risk assessment of preoperative opioid use by interpretable neural network regression," The Annals of Applied Statistics, Ann. Appl. Stat. 17(1), 434-453, (March 2023)

NIDA. 2015, October 1. Prescription opioid use is a risk factor for heroin use. Retrieved from https://nida.nih.gov/publications/research-reports/prescription-opioids-heroin/prescription-opioid-use-risk-factor-heroin-use on 2024, May 3

CDC. 2021, August 22. Understanding the Opioid Overdose Epidemic. Retrieved from https://www.cdc.gov/opioids/basics/epidemic.html on 2024, May 3

Beyer, Don, September 28. JEC Analysis Finds Opioid Epidemic Cost U.S. Nearly $1.5 Trillion in 2020. Retrieved from https://beyer.house.gov/news/documentsingle.aspx?DocumentID=5684#:~:text=Adapting%20a%20methodology%20used%20by,likely%20to%20continue%20to%20increase. on 2024, May 3

Venniro, M., Zhang, M., Caprioli, D., Hoots, J. K., Golden, S. A., Heins, C., Morales, M., Epstein, D. H., & Shaham, Y. (2018). Volitional social interaction prevents drug addiction in rat models. Nature neuroscience, 21(11), 1520–1529. https://doi.org/10.1038/s41593-018-0246-6

Fedesoriano, Heart Failure Prediction Dataset, Retrieved from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?resource=download on May, 3