Combating the Opioid Crisis:

Insights from Machine Learning

Texas A&M University

Spring 2024

Joshua Gabino, Amrit Gill, Alex Jacobs, Riley Rizzo, Regan Schulman, and Burke St. Claire

Abstract

This study examines Janssen and Zhang's 2023 paper *Retail Pharmacies and Drug Diversion during the Opioid Epidemic* on the impact of transitioning from independent to chain pharmacies on opioid distribution in the U.S. The authors found a significant reduction in opioid dispensing linked to enhanced regulatory compliance. Using gradient boosting, our analysis identifies additional key variables affecting distribution. The findings advocate for adaptive regulatory frameworks and highlight the effectiveness of data analytics in public health interventions, offering crucial insights for managing the opioid epidemic through a data-driven approach.

I. Introduction

The comprehensive study *Retail Pharmacies and Drug Diversion During the Opioid Epidemic* examines the critical role that pharmacy ownership plays in the context of the opioid crisis. This focus is particularly pertinent given the widespread concern among law enforcement and regulatory bodies that pharmacies are implicated in nearly 80% of all cases of drug diversion. Drug diversion, in this context, is defined as the illegal procurement or utilization of prescription medications, which constitutes a significant public health concern.

The research highlights a notable disparity in opioid distribution between different types of pharmacies. It was observed that independent pharmacies distributed on average 39.1% more opioids than their chain-operated counterparts. This finding is significant as it suggests that the management and operational practices at independent pharmacies might contribute to higher dispensing rates. Furthermore, the study demonstrates a marked decrease in opioid dispensing when these independent pharmacies transition to chain ownership. This transition often brings about stricter regulatory compliance and enhanced monitoring capabilities, which are likely factors in the reduced dispensing figures.

To deepen the analysis, our study incorporated machine learning techniques, specifically employing a gradient boosting framework. Gradient boosting is a robust method that combines several weak predictive models to create a strong predictive model. This technique is particularly suited for this kind of analysis because of its ability to manage complex nonlinear interactions between multiple variables without falling into the trap of overfitting. Overfitting can skew results and lead to inaccurate conclusions, making robust methods like gradient boosting invaluable for ensuring the reliability of our findings.

A critical component of the gradient boosting model is its focus on feature importance, which assigns weights to various variables based on their relative significance in influencing the predictions. This aspect of the model is crucial as it illuminates which factors are most impactful in predicting opioid dispensing practices, thereby providing insights into potential areas for intervention.

By integrating gradient boosting into our research methodology, we enhanced the sophistication and accuracy of our analysis. This approach allowed us to uncover nuanced insights into the dynamics of opioid dispensing across different types of pharmacies and highlighted the potential regulatory and operational interventions that could help mitigate the opioid epidemic. The use of advanced analytical techniques ensures a deeper understanding of the complex factors at play in drug diversion scenarios, facilitating more informed decisions by policymakers and industry stakeholders aimed at curbing this critical public health issue.

II. Literature Review

Gradient boosting is an advanced machine learning technique renowned for its efficacy in handling large and complex datasets, making it particularly well-suited for the voluminous data commonly associated with pharmacy records. This method excels in enhancing the performance of simple predictive models by methodically focusing on correcting errors. Specifically, each iteration of the model concentrates on the inaccuracies of its predecessors, making incremental adjustments to minimize these errors. This iterative correction process allows gradient boosting to refine its predictions progressively, resulting in robust predictive capabilities.

By synthesizing multiple simple models, gradient boosting constructs a comprehensive model that is both fast and accurate, enabling rapid and informed decision-making based on extensive data. This feature is crucial in environments like pharmacies where timely and precise data analysis is essential for effective management.

Despite its strengths, the application of gradient boosting to areas such as drug misuse and prescription monitoring is not extensively documented. This represents a critical research gap, given the potential benefits of rapid and effective data analysis in these areas. The opioid crisis could benefit significantly from the application of gradient boosting. By analyzing pharmacy transactions, patient histories, and other economic indicators swiftly, this technique could potentially identify patterns indicative of misuse early in the process.

Implementing gradient boosting could provide a powerful tool for pharmacies to monitor and prevent prescription abuse proactively. It could also assist in curtailing the opioid epidemic by enabling early detection of problematic prescription patterns before they escalate into more significant issues.

In conclusion, while gradient boosting has proven highly effective in processing large datasets in pharmacy settings, there is a pressing need for further research into its application for direct interventions in drug abuse scenarios. Such studies could be pivotal in leveraging the full potential of gradient boosting to combat the opioid crisis, providing guidance to preventative strategies that could help mitigate this public health challenge.

III. Methodology

Janssen and Zhang analyzed data from 84,000 pharmacies spanning the years 2006 to 2012, deliberately excluding those based within hospitals. They highlighted the growing predominance of chain pharmacies, which constituted 53% of the market during this period. An increasing trend was observed where numerous independent pharmacies were transitioning to chain operations, signaling a significant shift in the pharmacy industry landscape.

This study contextualizes its findings within the broader pharmaceutical and healthcare environment influenced markedly by a 1999 change in medical guidelines. These guidelines encouraged physicians to adopt a more proactive approach in managing pain, including the increased prescription of opioids for chronic nonmalignant pain—conditions not associated with cancer, such as headaches, back pain, arthritis, and muscle pain. The change facilitated a dramatic increase in the ease of obtaining prescriptions for powerful painkillers like OxyContin. This period also saw a rise in illicit methods of acquiring such medications, including doctor shopping and prescription forgery, further complicating the opioid crisis.

The researchers employed a rigorous methodological framework, utilizing a Difference-in-Differences (DiD) analysis to analyze the impacts of pharmacies transitioning from independent to chain status over the seven years of data. This transition is particularly relevant because chain pharmacies typically operate under stricter regulatory compliance and possess more sophisticated systems to monitor and mitigate prescription abuse, largely due to greater liability concerns. The analysis revealed a significant decrease of 52.8% in OxyContin dispensing rates post-transition, underscoring the potential regulatory and operational benefits of chain pharmacies in controlling opioid distribution. The study's robustness was further ensured

by controlling for various fixed effects—pharmacy, time, geographic location, and year-month—and conducting thorough checks on different model specifications to validate the results. The model with the strongest fit was that with facility-fixed effects, implying specific pharmacies were the leading driver in Opioid misuse.

We replicated the original archival data from the ARCOS database provided by the Washing Post in SAS then exported our cleaned data to examine feature analysis in Python. Our replicated data is not a perfect match to the original authors'; comments in their provided R code explained the final dataset was compiled after manually reviewing each observation in Excel. However, we cleaned the data to 5,071,787 observations compared to the authors' 5,055,761, only a 0.32% difference.

Further, leveraging advanced analytical techniques, we utilized the Gradient Boosting Regressor algorithm from the scikit-learn library to assess feature importance of both supply drivers, pharmacy ownership change and demand drivers, labor force participation rate and unemployment rate. This addressed a critical gap in the original research, which primarily focused on supply-side factors. By integrating additional variables, we aimed to capture a more comprehensive view of the distribution dynamics for opioids. The inclusion of these socioeconomic indicators provided a deeper understanding of the drivers behind opioid prescriptions. The enhanced model, termed the "boosted plus," demonstrated superior performance with the lower mean squared error than the replicated OLS, indicating its effectiveness in capturing the complex interplay of factors influencing opioid distribution in a more nuanced manner.

Overall, this study offers significant insights into the dynamics of pharmacy operations and their role in the opioid crisis, emphasizing the need for continued research and targeted policy interventions to manage and mitigate the far-reaching impacts of opioid misuse.

IV. Findings

The regression results summarized in the table below provide a clear comparison of the original study's findings (Table 3, column 6) with the replication effort plus our replication with added demand variables, as well as error comparison between the models. The data from these tables offers valuable insights into the impact of different variables and model adjustments on the study's outcomes. The appendix contains visualizations and explanations on feature analysis.

	Original (1)	Replication (2)	Boosted+ (3)
D^{Pre}	5.009 (6.886)	3.110* (0.635)	2.544* (0.670)
D^{Post}	-9.303* (6.886)	-8.824* (0.617)	-10.860* (0.648)
Chain	-8.362* (0.578)	-6.229* (0.063)	-6.345* (0.068)
Constant	32.036* (0.554)	29.860* (0.046)	14.766* (0.975)
Labor Force Participation Rate			-0.00000241* ($3.702e^{-8}$)
Unemployment Rate			2.079* (0.011)
Mean Outcome	27.14	26.79	25.804
Root MSE		73.873	73.479
Observations	5,055,761	5,071,787	4,695,416
R^2	0.003	0.0018	0.0093

$p < 0.1$

We observe the comparison of coefficients and other statistics across three different models: Original, Replication, and Boosted+. The D^{pre} variable shows an initial coefficient of 5.009 in the Original model reducing to 3.110 in the Replication model and further to 2.544 in the Boosted+ model, both with significant p-values. D^{post} also shows a decrease in the coefficient from -9.303 in the Original to -8.824 in Replication and further to -10.860 in Boosted+. The 'Chain' variable indicates a reduction in its negative impact from -8.362 in the Original to -6.229 in Replication, and a slight increase to -6.345 in Boosted+. All variables in every regression except D^{pre} in Original are statistically significant at a 0.01 level.

The intercept decreases significantly from 32.036 in the Original model to 14.766 in the Boosted+ model, indicating changes in the baseline level of always independent pharmacies across models. The Boosted+ model also introduces additional predictors such as the Labor Force Participation Rate and Unemployment Rate based on feature analysis, both showing significant effects on the dependent variable, suggesting that these factors contribute to the variance explained by the model.

Model fit statistics indicate an improvement across models, with Root MSE slightly decreasing and R^2 increasing significantly in the Boosted+ model to 0.0093 from 0.003 in the Original, demonstrating a better explanation of the variance in the dependent variable. This progression suggests that later models offer a refinement in estimates, reduction in standard errors, and better overall model fit, likely due to improvements in model specifications or data quality. The inclusion of additional variables in the Boosted+ model provides a more robust explanation of the factors influencing the OxyContin dispensing.

One key drawback of our analysis is the discrepancy in the number of observations between the original and replicated studies, which arises from the original authors manually reviewing observations after initial cleaning. This selection could potentially introduce a bias, emphasizing the need for transparency and consistency in data selection processes. Furthermore, despite the minor inconsistencies in the number of observations, the Gradient Boosted Regression model with added labor force and unemployment rate variables emerged as the model providing the best fit. This underscores the importance of including relevant economic indicators to enhance the model's explanatory power and accuracy in contexts like these, where economic factors may significantly influence the outcomes.

Another potential model error we discovered was the disproportionate sample size to the size of the treatment group. The original study analyzed 84,111 pharmacies across the U.S. to create a national model, but only 304 pharmacies were identified as the treatment group. Additionally, the model did not follow a traditional Difference-in-Difference approach but followed the approach of Eliason et al. (2020) to capture pharmacies that were always pre-treatment (Independent) or always post treatment (Chain) over the seven year period. Due to this, our gradient boosted regression did not include an interaction variable so the sheer volume of observations for the control groups versus the number of observations for the treatment group may have influenced the model’s predictions. This may have skewed the relative weights of feature importance. The Chain variable may have been used to split the model more often simply because there were many more observations of it. Figures of the feature analysis on the Replication and Boosted+ OLS regressions are included in the appendix.

Overall, these findings contribute to a nuanced understanding of the variables impacting the original study's results and affirm the robustness and relevance of advanced analytical techniques in refining research outcomes.

V. Discussion

The research paper employed a non-traditional Difference-in-Differences (DID) design, which was innovative in utilizing essentially two control groups to enhance the robustness of the findings. The study analyzed a substantial dataset comprising approximately 5 million observations across 84,000 pharmacies. However, it's important to note that the treatment groups were relatively small, with only around 300 pharmacies included.

One of the standout advantages of employing a Gradient Boosting Regressor in the analysis was addressing the weaknesses in the original model. The initial model's limitations were significantly mitigated, demonstrating the regressor's effectiveness. Despite this improvement, further analysis would have been beneficial, particularly in randomizing the observations for 'always-chain' and 'always-independent' pharmacies to align the number of observations in the treatment groups more closely. This step would likely have led to more definitive conclusions regarding treatment effects.

A crucial aspect of the study was its application of the algorithm for feature analysis, which proved to be its most beneficial use. The initial feature analysis compared data before and after the intervention against a baseline of always-Independent pharmacies (the Chain variable), revealing that the most significant factor influencing the number of dispensed opioids was the dispensing patterns of chain pharmacies. This finding is logical considering that chain pharmacies typically wield greater market power nationally compared to smaller, independent chains.

Further, when demand-related regressors like labor force size and unemployment rate were incorporated into the feature analysis, it became evident that demand factors had a more substantial impact on opioid distribution than supply limitations. This insight underscores the complexity of opioid distribution dynamics, suggesting that addressing the opioid crisis may require focusing more on economic and demand-side factors rather than solely on supply-side restrictions.

VI. Conclusions

The findings from this study are particularly important for policymakers and healthcare regulators. They provide empirical support for the effectiveness of certain regulatory measures and underscore the need for continued surveillance and adaptation of policies to address the evolving landscape of pharmacy operations and opioid distribution. Additionally, the demonstrated value of incorporating machine learning techniques in analyzing complex datasets offers a model that can be replicated in other areas of public health research, potentially providing a pathway to more effective and targeted interventions.

In conclusion, the study articulates a compelling case for the nuanced application of regulatory oversight and advanced data analytics in tackling the opioid epidemic. The evidence suggests that a combination of enhanced regulatory frameworks for pharmacies, especially encouraging transitions from independent to chain models, as well as addressing demand-side factors, can significantly impact the control of opioid dispensing and potentially reduce drug diversion. This research not only contributes to our understanding of the opioid crisis but also sets a precedent for the application of machine learning in public health, advocating for a more informed and data-driven approach in public health crises management.

VII. References

K., Dhiraj. “Implementing Gradient Boosting Regression in Python.” Paperspace Blog, 13 Dec. 2019, https://blog.paperspace.com/implementing-gradient-boosting-regression-python/.

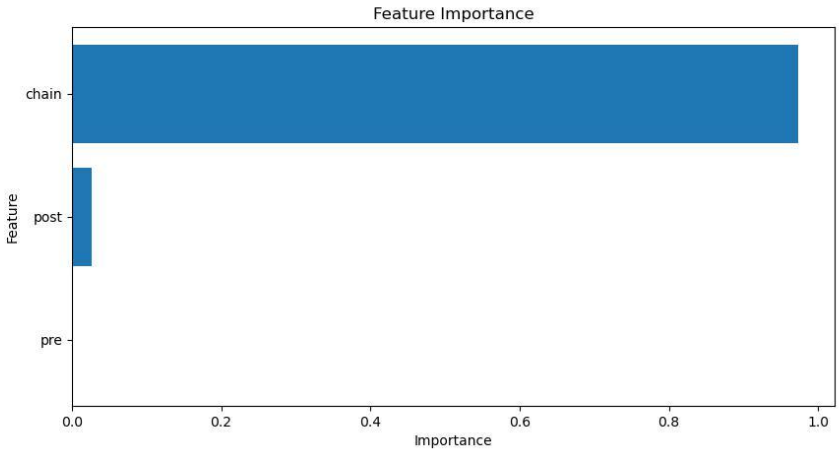
Janssen, Aljoscha and Xuan Zhang. “Retail Pharmacies and Drug Diversion during the Opioid Epidemic.” *American Economic Review.* https://doi.org/10.1257/aer.20210357

Prettenhofer, Peter, and Gilles Louppe. Gradient Boosted Regression Trees. https://orbi.uliege.be/bitstream/2268/163521/1/slides.pdf.

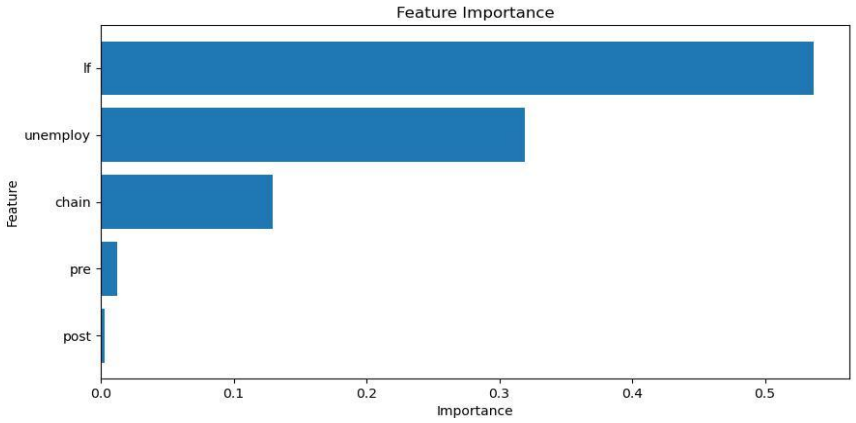
Saini, Anshul. “Gradient Boosting Algorithm: A Complete Guide for Beginners.” Analytics Vidhya, 20 Sept. 2021, www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-acomplete-guide-for-beginners/.

Zemel, Richard, and Toniann Pitassi. A Gradient-Based Boosting Algorithm for Regression Problems. https://proceedings.neurips.cc/paper_files/paper/2000/file/8d9fc2308c8f28d2a7d2f6f48801c705-Paper.pdf.

VIII. Appendix



Comparing feature importance between pre-treatment, post-treatment, and chain variables. Using the Replication OLS regression, his graph illustrates that the largest influence on the amount of distributed opioids are pharmacies that have always been part of a chain. This makes sense as chains have more consistent national market influence than individual pharmacies.



Comparing feature importance between pre-treatment, post-treatment, chain, labor force participation rate, and unemployment rate variables. This illustrates that demand features like labor force participation rate (lf) and unemployment rate (unemploy) have a larger impact than supply features on opioid distribution.