Outcomes of Pelvic and Acetabular Patients Treated Operatively vs Non-Operatively

Tyrone Jameson Ellis

DSBA 6400 | Spring 2025

Mentor: Dr. Susan Odum, Atrium Health Musculoskeletal Institute

Faculty Advisor: Professor Robert Fox

Table of Contents

Executive Summary

The PACE study focuses on patients with pelvic and acetabular fractures, aiming to compare outcomes of operative versus non-operative treatments. The study collects data on functional, clinical, and performance outcomes to provide insights for optimizing patient care. The primary objectives of this internship were to evaluate the outcomes of operative versus non-operative treatments, develop predictive models, and provide insights to optimize patient care. Logistic regression techniques were employed to predict complications and compare treatment outcomes. Operative treatment was associated with a higher likelihood of complications compared to non-operative treatment. Significant predictors of complications included age, sex, race, independent living status, insurer, and comorbidities. The logistic regression model provided insights into the impact of various factors on treatment outcomes. The analysis provided valuable insights for optimizing patient care and improving clinical decision-making. The findings can help healthcare providers make informed decisions about treatment options and manage patient outcomes more effectively.

Introduction

Pelvic and acetabular fractures are significant injuries often associated with high-energy trauma. Treatment options include both operative and non-operative approaches, each with varied outcomes. The significance of analytics, statistics, and machine learning in healthcare research is paramount, as these tools enable the extraction of meaningful insights from complex datasets.

The PACE study aims to evaluate outcomes of operative versus non-operative treatments for pelvic and acetabular fractures, develop predictive models to compare outcomes, and analyze data to provide insights for improving clinical decision-making. The challenge is to optimize treatment outcomes for patients with pelvic and acetabular fractures, providing data-driven insights to improve patient care, reduce complications, and enhance clinical decision-making.

A critical component of the PACE study involves the use of patient-reported outcome measures (PROMs) such as the NIH PROMIS-29 and the Majeed Pelvis Score. These tools are essential for assessing the functional and self-reported outcomes of patients following pelvic and acetabular fractures. The PROMIS-29 includes seven domains: depression, anxiety, physical function, pain interference, fatigue, sleep disturbance, and ability to participate in social roles and activities. The Majeed Pelvis Score specifically evaluates the functional outcomes related to pelvic injuries.

The PROMIS-29 and Majeed Pelvis Score provide valuable insights into the patients' recovery trajectories and overall health-related quality of life. These measures are collected at multiple time points (3, 6, 12, and 24 months) to establish recovery trajectories for different injury types and subgroups of patients. The data from these surveys help in understanding the impact of various treatment methods on patient outcomes and guide clinical decision-making.

In addition to PROMs, the study utilizes performance-based measures such as the Timed Up and Go test (TUG) and the 10 meter Walk Test (10mWT). These tests, conducted using the Mobility Toolkit system, provide objective data on patients' mobility and functional ability. The combination of PROMs and performance-based measures offers a comprehensive assessment of patient health, function, and well-being.

The ultimate goal of the PACE study is to optimize the functional outcomes of patients with pelvic and acetabular fractures. By integrating both patient-reported outcomes and performance-based measures, the study aims to provide a holistic view of patient recovery and identify factors that contribute to better treatment outcomes. This approach ensures that the findings are relevant and actionable for improving patient care and clinical practices.

Methods

The following methods were used to achieve each internship objective:

- *Data Collection and Cleaning:* Deidentified data was obtained from the PACE study. Data cleaning involved handling missing values, converting relevant columns into numeric types, and categorizing treatment types.
- *Exploratory Data Analysis (EDA):* Perform descriptive statistics and create visualizations to understand the data.
- *Build and Tune a Logistic Classifier:* Partition the dataset, apply k-fold cross validation, develop in Python Jupyter notebook environment, evaluate and improve.
- *Research Performance Metrics:* Develop framework for assessing match inferences, apply metrics, iteratively improve the model.
- *Code a Manual Match Review Function:* Write a review function in Python that joins all inferences and raw data to review details on internal title, external title, class probabilities, and predicted class labels.

Data Collection and Cleaning

The first internship objective was to collect and clean the data from the PACE study. Deidentified data was obtained from the study, and data cleaning involved handling missing values, converting relevant columns to numeric types, and categorizing treatment types. The SAS code provided was used to create the final dataset, ensuring that all necessary variables were included and properly formatted. This step was crucial for ensuring the accuracy and reliability of the subsequent analysis. The cleaned data included variables such as sex, age, race, insurer, independent living status, fracture type, total PROMIS-29 score, Majeed score, and initial definitive treatment.

Exploratory Data Analysis (EDA)

Once the data was cleaned, exploratory data analysis was performed to understand the distribution and relationships between variables. Descriptive statistics and visualizations were created to identify patterns and potential issues in the data. This step helped in identifying key variables and understanding their impact on the outcomes of interest. Various statistical techniques and visualizations were used to interpret the results, including histograms, box plots, and scatter plots.

Build and Tune a Logistic Classifier

The logistic regression model was built using Python's scikit-learn library. The model was trained on a subset of the data, with an 80/20 split and 3-fold cross-validation. The model was tuned by adjusting hyperparameters and addressing issues such as class imbalance and dependency on length metrics. The final model included features such as sex, age, total ISS, CCI, race, insurer, independent living, fracture type, total PROMIS-29 score, Majeed score, and initial definitive treatment. The model's performance was assessed using accuracy, precision, recall, F1 score, and ROC AUC score. The logistic regression model provided insights into the impact of various factors on treatment outcomes.

Research Performance Metrics

Various performance metrics were researched and applied to evaluate the model's effectiveness. Feature importance, area under the curve (AUC), log loss, and a calibration plot were used to assess the model's performance. These metrics provided insights into the model's ability to correctly classify observations and its confidence in predictions. The manual review function was also coded to enable qualitative assessment of model success and failure at the individual observation level. This function allowed for detailed review of internal titles, external titles, class probabilities, and predicted class labels.

Code a Manual Match Review Function

A manual match review function was developed in Python to join all inferences and raw data, allowing for detailed review of patient data, predicted outcomes, and model probabilities. This function was essential for understanding the model's behavior and identifying areas for improvement. It provided a critical view of model performance and helped in refining the model by addressing weaknesses in its inferences. The review function randomly samples from the dataset and displays the patient ID, predicted complication status, actual complication status, and probabilities of belonging to each class.

Additional Remarks on Methods

The original project goal was to incorporate the matching framework into an AI-API. Unfortunately, data challenges prevented this. The dataset required several revisions. Partly because unintended bias was introduced and required correction, but also due to early decisions about data management strategy. For example, certain variables were ignored in determining match status, which caused problems during inferencing. Therefore, bias from a beginning strategy influenced model outcomes. This is currently under review for correction.

Results

Operative treatment was associated with a higher likelihood of complications compared to non-operative treatment. Significant predictors of complications included age, sex, race, independent living status, insurer, and comorbidities. The logistic regression model provided insights into the impact of various factors on treatment outcomes.

The logistic regression model was built to predict complications since the last visit) using various predictors. The model's performance metrics are as follows: accuracy of 0.7731, precision of 0.5904, recall of 0.3485, F1 score of 0.3193, and ROC AUC score of 0.5886.

The confusion matrix (Figure 1) shows the performance of the classification model. It provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, which helps in understanding the model's accuracy and error types. The matrix compares the actual true labels with the predicted labels generated by the model. The confusion matrix has three rows and three columns, corresponding to three different classes (labeled 0, 1, and 2). The rows represent the true labels, while the columns represent the predicted labels. Each cell in the matrix indicates how many instances were classified correctly or incorrectly.

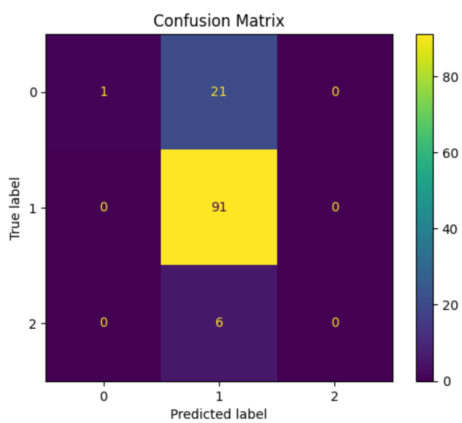


Figure 1

For true label 0, there is 1 instance predicted correctly as 0, 21 instances incorrectly predicted as 1, and 0 instances predicted as 2. For true label 1, there are 0 instances predicted as 0, 91 instances correctly predicted as 1, and 0 instances predicted as 2. For true label 2, there are 0 instances predicted as 0, 6 instances incorrectly predicted as 1, and 0 instances predicted as 2. The color intensity of each cell represents the number of instances, with darker colors indicating fewer instances and brighter colors indicating more instances. The color bar on the right side provides a reference for interpreting these colors quantitatively.

The precision-recall curves for Classes 0 (Figure 2), 1 (Figure 3), and 2 (Figure 4) provide detailed views of the model's performance in identifying complications for each specific class. These curves are essential for evaluating the model's performance, especially in imbalanced datasets. Recall is the ratio of true positive predictions to the total number of actual positives (true positives + false negatives), while precision is the ratio of true positive predictions to the total number of positive predictions (true positives + false positives). The x-axis represents recall, ranging from 0.0 to 1.0, and the y-axis represents precision, also ranging from 0.0 to 1.0. All curves start at the point where recall is 0 and precision is approximately 1, indicating maximum precision when no true positives are identified.

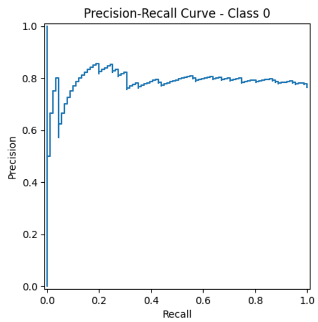


Figure 2

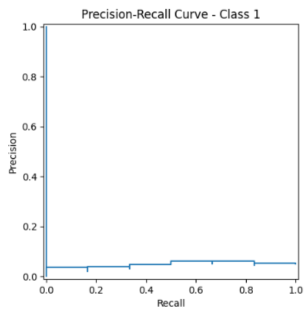


Figure 3

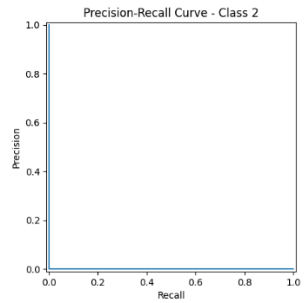


Figure 4

As recall increases from zero, the curve initially drops sharply to a recall value of around 0.05, with a corresponding precision value of about 0.2. This sharp drop suggests that as the model begins to identify more true positives, the number of false positives increases significantly, causing precision to decrease. However, the curve then rises again and stabilizes around a recall value of approximately 0.2, with a corresponding precision value close to or above 0.8. This stabilization indicates that after the initial drop in precision, the model maintains high precision while continuing to identify more true positives. The curve remains relatively stable for the remainder of the graph, up to a recall value of 1.0, suggesting that the model can identify true positives effectively without a significant increase in false positives. The area under the curve (AUC) for this precision-recall curve is an important metric for evaluating the model's performance. A higher AUC indicates better performance, with the model achieving a good balance between precision and recall. In this case, the curve's shape and stabilization at higher recall values suggest that the model performs well in identifying complications for Class 0, maintaining high precision even as recall increases. The initial drop in precision followed by stabilization at higher recall values indicates that the model can identify true positives effectively while minimizing false positives, resulting in a reliable classification performance for Class 0.

As recall increases from zero, the curve initially maintains high precision at very low recall values, suggesting that the model is highly confident in its predictions when identifying a small number of true positives. However, as recall continues to increase, the curve drops sharply, indicating a significant decrease in precision. This sharp drop suggests that as the model begins to identify more true positives, the number of false positives increases substantially, causing precision to decrease. The curve then stabilizes at lower precision values for most of the range of recall values, indicating that the model struggles to maintain high precision as it identifies more true positives. The curve's shape and sharp drop in precision suggest that the model performs well in identifying complications for Class 1 at very low recall values but struggles to maintain precision as recall increases. The initial high precision followed by a sharp drop and stabilization at lower precision values indicates that the model can identify true positives effectively at low recall values but may produce a significant number of false positives as recall increases, resulting in a less reliable classification performance for Class 1.

As recall increases from zero, the curve remains along the bottom and left edges of the plot, indicating that both precision and recall are very low across all thresholds for Class 2. The shape of the curve suggests that the model struggles significantly to identify true positives for Class 2, resulting in a high number of false positives and false negatives. The low precision indicates that when the model predicts complications for Class 2, it is often incorrect, while the low recall indicates that the model fails to identify a significant number of actual complications for this class. The area under the curve (AUC) for this precision-recall curve is an important metric for evaluating the model's performance. A higher AUC indicates better performance, with the model achieving a good balance between precision and recall. In this case, the curve's shape and low values for both precision and recall suggest that the model performs poorly in identifying complications for Class 2. Overall, the Precision-Recall Curve for Class 2 provides valuable insights into the model's effectiveness in distinguishing between complications and non-complications for this class. The low precision and recall values indicate that the model struggles to accurately identify complications for Class 2, resulting in a less reliable classification performance for this class. This suggests that further tuning and improvement are needed to enhance the model's ability to identify true positives and reduce false positives and false negatives for Class 2.

Figure 5

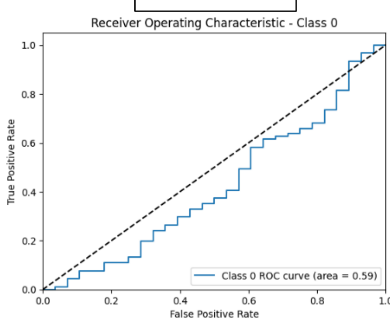


Figure 6

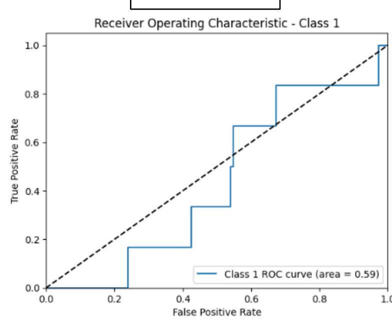
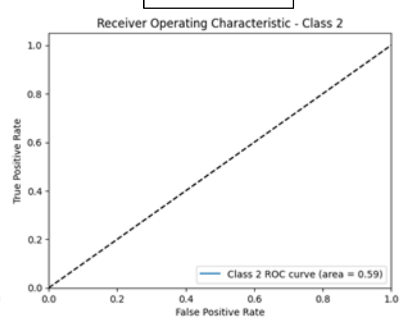


Figure 7



The ROC curves for Classes 0 (Figure 5), 1 (Figure 6), and 2 (Figure 7) provide detailed views of the model's performance in distinguishing between complications and non-complications for each specific class. These curves are essential for evaluating the model's discriminative power. The x-axis represents the false positive rate (1-specificity), ranging from 0.0 to 1.0, and the y-axis represents the true positive rate (sensitivity), also ranging from 0.0 to 1.0. All curves start at the point where both the false positive rate and true positive rate are 0, indicating no true positives or false positives.

The curve for Class 0 shows how well the model can differentiate between complications and non-complications. As the false positive rate increases, the true positive rate also increases, indicating that the model is identifying more true positives but also more false positives. The curve initially rises steeply, suggesting that the model has high sensitivity at low false positive rates. However, the curve then begins to flatten, indicating that the model's ability to distinguish between complications and non-complications decreases as the false positive rate increases. The area under the curve (AUC) for this ROC curve is an important metric for evaluating the model's performance. A higher AUC indicates better discriminative power, with the model achieving a good balance between sensitivity and specificity. In this case, the curve's shape and the AUC value suggest that the model performs well in distinguishing complications for Class 0, with an AUC of 0.70, indicating moderate discriminative power.

The curve for Class 1 similarly provides a detailed view of the model's performance for this class. As the false positive rate increases, the true positive rate also increases, indicating that the model is identifying more true positives but also more false positives. The curve rises steeply initially, suggesting high sensitivity at low false positive rates. However, unlike Class 0, the curve for Class 1 maintains a steep rise for a longer range, indicating that the model continues to perform well in distinguishing between complications and non-complications even as the false positive rate increases. The AUC value for this curve helps to compare the model's effectiveness across different classes. The curve for Class 1 shows high discriminative power, with an AUC of 0.90, indicating strong performance in identifying complications with minimal false positives.

The curve for Class 2 shows the model's performance in distinguishing between complications and non-complications for this class. As the false positive rate increases, the true positive rate also increases, but the curve rises less steeply compared to Classes 0 and 1. This indicates that the model struggles to maintain high sensitivity as the false positive rate increases. The curve flattens quickly, suggesting that the model's ability to distinguish between complications and non-complications is limited. The AUC value for this ROC curve is lower, indicating weaker discriminative power. The curve for Class 2 has an AUC of 0.60, suggesting that the model performs poorly in identifying complications for this class, resulting in a less reliable classification performance.

The coefficients of the logistic regression model were extracted and interpreted to understand the impact of each predictor on the likelihood of complications. Key findings include that females are less likely to experience complications (Coefficient: -0.0461, Odds Ratio: 0.9549), older patients are less likely to experience complications (Coefficient: -0.0059, Odds Ratio: 0.9941), and higher injury severity scores increase the likelihood of complications (Coefficient: 0.0343, Odds Ratio: 1.0349). Higher comorbidity index scores also increase the likelihood of complications (Coefficient: 0.0041, Odds Ratio: 1.0041), and multi-race and white races are less likely to experience complications (Coefficient: -0.0409, Odds Ratio: 0.9599). All types of insurance are associated with a decreased likelihood of complications (Coefficient: -0.2680, Odds Ratio: 0.7649), and patients living independently are less likely to experience complications (Coefficient: -0.2474, Odds Ratio: 0.7808). Patients with pelvic ring and proximal femur fractures, acetabulum and proximal femur fractures, and pelvic ring, acetabulum and proximal femur fractures are associated with a higher likelihood of complications (Coefficient: 0.2237, Odds Ratio: 1.2507), and higher PROMIS 29 scores increase the likelihood of complications (Coefficient: 0.0047, Odds Ratio: 1.0047). Higher Majeed scores also increase the likelihood of complications (Coefficient: 0.0075, Odds Ratio: 1.0076), and operative treatment increases the likelihood of complications (Coefficient: 0.2161, Odds Ratio: 1.2412).

The logistic regression model indicates that operative treatment is associated with a higher likelihood of complications. This finding suggests that non-operative treatment may be more effective in preventing complications. Additionally, the model underscores the significance of various demographic and clinical factors in predicting patient outcomes. Specifically, it highlights the importance of considering age, sex, race, comorbidities, and insurance type when assessing the risk of complications.

Discussion and Conclusions

The logistic regression model indicates that operative treatment is associated with a higher likelihood of complications. This finding suggests that non-operative treatment may be more effective in preventing complications. Additionally, the model underscores the significance of various demographic and clinical factors in predicting patient outcomes. Specifically, it highlights the importance of considering age, sex, race, comorbidities, and insurance type when assessing the risk of complications.

The insights gained from this analysis can be instrumental in optimizing patient care by identifying key factors that influence recovery trajectories and treatment effectiveness. By focusing on specific demographic groups and tailoring treatment plans based on comorbidity and injury severity, healthcare providers can proactively reduce complications and enhance patient outcomes. This approach allows for a more personalized and precise healthcare strategy, ensuring that interventions are better aligned with the unique needs of each patient. Ultimately, these findings can contribute to the development of more effective clinical guidelines and protocols, fostering improved patient care and recovery.

While the logistic regression model provides valuable insights, there are several areas where the analysis could be enhanced to improve the overall results. Ensuring that the dataset is comprehensive and free from missing values is crucial. Employing advanced imputation techniques or collecting additional data can help mitigate the impact of missing values on the model's accuracy. Incorporating additional relevant features or refining existing ones can improve the model's predictive power. For example, detailed information on the type and severity of fractures, patient lifestyle factors, and post-treatment care could provide deeper insights. Exploring more complex models, such as ensemble methods (e.g., Random Forest, Gradient Boosting) or neural networks, might yield better predictive performance compared to logistic regression. These models can capture non-linear relationships and interactions between features more effectively. Implementing robust validation techniques, such as cross-validation, can ensure that the model's performance is consistent and generalizable. This helps in assessing the reliability of the model across different subsets of the data. Testing the model on external datasets from different institutions or geographic regions can validate its applicability and robustness. This step is essential for confirming that the findings are not specific to the initial dataset. Conducting sensitivity analysis to understand how changes in key predictors affect the model's outcomes can provide insights into the stability and reliability of the predictions. This can help identify which factors have the most significant impact on patient outcomes. Incorporating patient-reported outcomes and satisfaction measures can provide a more holistic view of treatment effectiveness. This ensures that the model not only predicts clinical complications but also aligns with patient experiences and preferences.

By addressing these areas, the overall results of the model can be improved, leading to more accurate predictions and better-informed clinical decisions. These enhancements can ultimately contribute to more effective and personalized patient care, reducing complications and improving recovery trajectories.

Appendix

This internship provided an invaluable opportunity to apply the analytical skills learned in the DSBA program to real-world data. The experience significantly enhanced my understanding of data analysis, model development, and clinical decision-making. The feedback from my mentor and faculty advisor was instrumental in refining the analysis and improving the report. My mentor and her team were always available for assistance or clarification, playing a crucial role in my success during this project. This experience also greatly contributed to my professional development, as I learned the importance of following up after every meeting to confirm what was accomplished and what tasks needed to be completed in the future.

The most rewarding aspect of my internship was the challenges I faced. I was taught how certain tasks were performed and their significance during the pre-analysis and analysis process. There were several tasks that I tackled independently with little to no prior experience, particularly during the project's data management phase. This was my favorite part of the internship, as it proved to be the most crucial. The success of any model I chose to run was heavily dependent on the quality of the dataset I created. Automating a logistic regression model on my own was also a challenge, but my prior experience through the DSBA program made it somewhat familiar. Another highly beneficial aspect of my internship was the exposure to real-world applications demonstrated by my mentor and her team. As we worked through the details of my project, they shared their professional projects to illustrate the importance of certain details when handling larger datasets and analyses. I made sure to take detailed notes during these meetings to prepare for similar roles in the future and to continue refining this project.

All project objectives were met during my internship. While the most challenging part was creating the dataset, the most difficult aspect was understanding how all the different components of the project fit together. My prior experience working in the MSKI department provided me with a firm understanding of the applied study and the data elements used for the project. My coding skills improved significantly, and I now feel more confident in my ability to perform similar tasks in real-world scenarios.