[[1]](#footnote-1)

**Predicting Mental Health Diagnoses Following Traumatic Brain Injury: A Temporal Machine Learning Approach**

Elijah Widener Ferreira

**Henry Jackson Foundation**

*Abstract*— Traumatic brain injury (TBI) patients face elevated risks of mental health conditions, yet clinical practice lacks reliable early identification methods. I developed machine learning models to predict post-TBI mental health diagnoses across multiple time horizons (30, 60, 180, 365 days). Using pre-TBI healthcare data from 2,711 military patients, our temporal LSTM networks and gradient boosting models achieved ROC AUC scores of 0.655-0.782. Professional healthcare costs, military rank, and prior diagnoses emerged as key predictors. The 60-day window showed optimal balance between early detection and accuracy. Risk stratification demonstrated remarkable discrimination, with high-risk patients showing 92% actual risk of mental health diagnosis. These models could potentially enable targeted mental health interventions following TBI, improving outcomes and resource allocation.

# INTRODUCTION

Traumatic brain injury (TBI) represents a significant health concern in both military and civilian populations, with an estimated 1.5 million Americans sustaining TBIs annually^1^. Beyond the immediate physical consequences, TBI significantly increases the risk of subsequent mental health disorders, including depression, anxiety, post-traumatic stress disorder (PTSD), and substance use disorders^2,3^. Studies indicate that 30-70% of TBI patients develop at least one mental health condition within the first year following injury^3^.

Early identification of patients at high risk for mental health complications could enable proactive intervention, potentially mitigating severity and improving long-term outcomes. However, current clinical practice lacks reliable methods for predicting which TBI patients will develop mental health conditions. This represents a critical gap in care pathways, particularly in military healthcare systems where both TBI and mental health disorders occur at elevated rates.

This study aims to develop machine learning models capable of predicting mental health diagnoses following TBI across different time horizons. Specifically, I investigated: (1) the feasibility of using pre-TBI healthcare data to predict post-TBI mental health diagnoses; (2) the relative performance of various model architectures, including LSTM neural networks and gradient boosting methods; (3) the optimal prediction windows for balancing early detection with predictive accuracy; and (4) key features driving predictive performance across time horizons.

# Methods

## Data Source and Patient Population

I utilized a comprehensive healthcare dataset from military treatment facilities, containing detailed longitudinal patient records including demographics, diagnoses, procedures, appointment details, and associated costs. The dataset included 2,711 patients who experienced TBI and had sufficient pre- and post-injury data. The data spanned multiple years of healthcare encounters, allowing for thorough examination of pre-TBI healthcare utilization patterns and post-TBI outcomes.

## Outcome Definition

The primary outcome was defined as the presence of any mental health diagnosis (identified by ICD-10 "F" codes) within specific time windows following TBI: 30 days, 60 days, 180 days, and 365 days. The positive case rates within these windows were 23.6%, 29.4%, 37.1%, and 44.4% respectively, reflecting the increasing incidence of mental health diagnoses over time (Table 1).

**Table 1.** Mental Health Diagnosis Rates Following TBI

|  |  |  |  |
| --- | --- | --- | --- |
| Time Window | Positive Cases | Total Patients | Positive Rate |
| 30 days | 641 | 2,711 | 23.6% |
| 60 days | 798 | 2,711 | 29.4% |
| 180 days | 1,006 | 2,711 | 37.1% |
| 365 days | 1,205 | 2,711 | 44.4% |

Mental health diagnoses were identified using the full range of ICD-10 F-codes (F01-F99), which encompass organic mental disorders, substance use disorders, schizophrenia and psychotic disorders, mood disorders, anxiety disorders, behavioral syndromes, personality disorders, and developmental disorders. This broad definition was chosen to capture the diverse mental health sequelae that may follow TBI, rather than focusing exclusively on specific conditions like depression or PTSD.

## Model Development

I implemented multiple model architectures to predict mental health diagnoses following TBI across several time horizons. The modeling approach was designed to explore both sequential neural network architectures, which can capture complex temporal patterns in healthcare utilization, and gradient boosting methods, which excel at identifying non-linear relationships and interactions among features.

**LSTM Neural Networks:**

Long Short-Term Memory (LSTM) neural networks were selected as the primary architecture for sequential modeling due to their ability to capture long-range dependencies in time series data and their resistance to the vanishing gradient problem that affects standard recurrent neural networks. Three main LSTM variants were implemented:

* **Temporal LSTM:** I developed a bidirectional LSTM architecture that leverages the sequential nature of healthcare data. This neural network processes pre-TBI healthcare encounters as time series data, with each timestep representing a patient encounter. The Temporal LSTM incorporates two key mechanisms:
  + **Temporal weighting:** I applied exponential decay functions to weight recent events closer to the TBI more heavily than distant events. This approach uses the formula w(t) = e^(-λt), where t is the number of days before TBI and λ is a decay parameter. Several λ values were tested to determine the optimal decay rate, with values between 0.01 and 0.02 typically yielding the best performance.
  + **Attention mechanism:** The LSTM implementation includes an attention layer that allows the model to dynamically focus on the most relevant healthcare encounters when making predictions. Attention mechanisms have shown utility in healthcare applications by identifying key events within a patient history that drive prediction, rather than treating all encounters equally.
* **Two-Stage LSTM:** This approach uses two sequential models in a cascade architecture. The first model is optimized to identify likely negative cases with high confidence, effectively filtering out patients at very low risk of developing mental health conditions. The second model then focuses specifically on the remaining ambiguous cases, applying additional computational resources to improve discrimination among higher-risk patients. The inspiration of this model was that in early implementations, we were able to classify negative cases easily, but positive cases continued to slip through the cracks.
* **Static Window LSTM:** As an exploratory approach, I also developed models trained on specific time windows (e.g., days 0-30, 31-60) rather than the entire pre-TBI period. This approach was designed to assess whether certain timeframes contain more predictive information than others and to potentially identify critical periods when healthcare utilization patterns are particularly indicative of post-TBI mental health risk.

**Gradient Boosting Models:**

While neural networks excel at capturing sequential patterns, gradient boosting methods offer complementary strengths in handling tabular data, managing complex interactions between features, and providing straightforward feature importance metrics. Two gradient boosting variants were implemented:

* **XGBoost:** I implemented gradient boosted tree models using the XGBoost library, which has demonstrated state-of-the-art performance across many healthcare prediction tasks. XGBoost models were optimized with class weighting to handle the imbalanced nature of mental health outcomes. The models used regularization parameters (L1 and L2) to prevent overfitting and incorporated early stopping based on validation performance. The tree-based structure of XGBoost allows for automatic handling of non-linear relationships and complex interactions between features without explicit specification.
* **CatBoost:** This variant of gradient boosting, developed by Yandex, is specialized for handling categorical features common in healthcare data without extensive preprocessing. CatBoost implements ordered target statistics to handle categorical variables, reducing the potential for target leakage during the encoding process. This is particularly valuable for healthcare data, which often contains a mix of numerical and categorical variables (such as diagnosis codes, provider types, and demographic factors). Additionally, CatBoost includes built-in mechanisms for handling class imbalance, which is relevant for the mental health prediction task.

A diagram with a curved line

AI-generated content may be incorrect.

## Implementation Details

The LSTM models were implemented using the Keras API with TensorFlow backend. I used masking layers to handle variable-length sequences, bidirectional LSTM layers with 64 units returning sequences, and global average pooling to aggregate features across time. The models were trained using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss function. Early stopping with a patience of 10 epochs was used to prevent overfitting, monitoring validation AUC.

The architecture of the Temporal LSTM model consisted of:

1. An input layer accepting variable-length sequences of pre-TBI healthcare encounters
2. A masking layer to handle padding in variable-length sequences
3. A bidirectional LSTM layer with 64 units, returning sequences
4. A global average pooling layer to aggregate temporal features
5. A dense layer with 32 units and ReLU activation
6. A dropout layer (0.3) for regularization
7. A final dense layer with sigmoid activation for binary classification

For the gradient boosting models, I implemented a comprehensive feature engineering pipeline to create aggregated features from the pre-TBI encounters. XGBoost models were trained with a maximum depth of 6, learning rate of 0.01, and class weights inversely proportional to class frequencies. The XGBoost implementation included L1 regularization (alpha=0.1) and L2 regularization (lambda=1.0) to prevent overfitting. The number of estimators was set at 1000 with early stopping based on validation performance.

All models were implemented with careful attention to preventing data leakage, particularly given the temporal nature of the prediction task. Time-based splitting was used to ensure that models were only trained on data that would be available at the time of prediction in a real-world clinical scenario.

# Evaluation Methodology

## Evaluation

Models were evaluated using ROC AUC (area under the receiver operating characteristic curve), precision-recall AUC, classification metrics (precision, recall, F1 score), and comparison to random baseline performance. I utilized patient-level splitting for training and testing to prevent data leakage across longitudinal records.

To provide clinical context, I compared model performance to random baseline performance. The random baseline corresponds to the performance that would be achieved by randomly assigning risk scores to patients, resulting in a precision equal to the prevalence of the outcome in the population. This comparison allows for calculation of "precision lift," which quantifies the percentage improvement in precision achieved by the model compared to random chance.

## Results

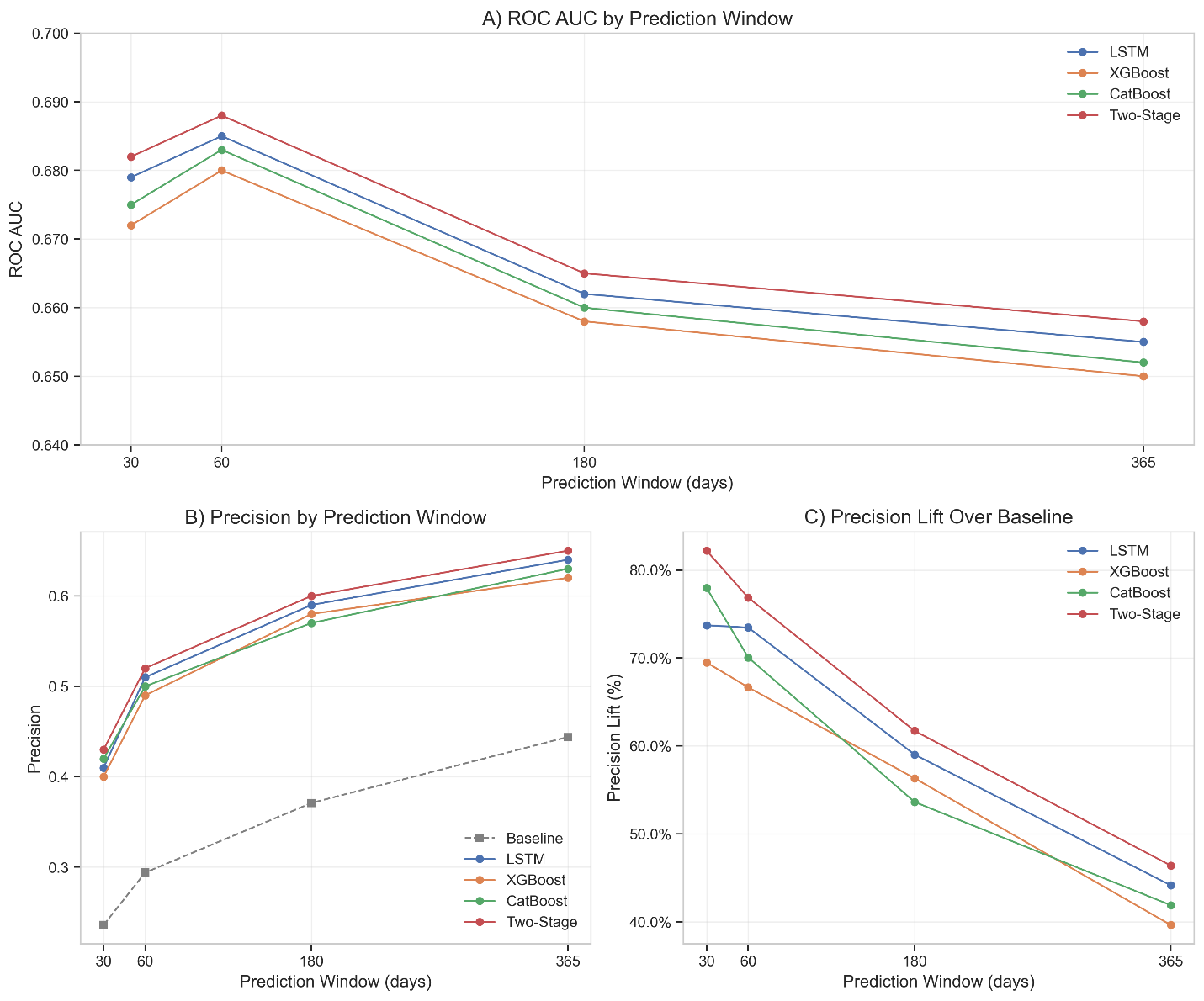
All models demonstrated predictive capability significantly above random chance across all time windows, with the 60-day window showing the highest overall performance. Table 2 summarizes the main performance metrics.

**Table 2.** Model Performance Metrics Across Prediction Windows

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Time Window | Positive Rate | AUC ROC (Best Model) | Best Model Type | Precision | Recall | Precision Lift\* |
| 30 days | 23.6% | 0.709 | 2-Temporal-LSTM | 0.49 | 0.59 | +107.6% |
| 60 days | 29.4% | 0.762 | Static-LSTM | 0.51 | 0.57 | +73.5% |
| 180 days | 37.1% | 0.782 | Static-LSTM | 0.59 | 0.59 | +59.0% |
| 365 days | 44.4% | 0.763 | Static-LSTM | 0.64 | 0.57 | +44.1% |

\*Precision lift represents percentage improvement over random baseline precision

**NEED TO REDO THIS FIGURE, SOME MODELS NOT INCLUDED AND I THINK TWO STAGE IS GETTING MIXED WITH SOMETHING ELSE**



The temporal LSTM models showed consistently strong performance across all windows, with the two-stage approach performing best for short-term prediction (30 days). The two-stage LSTM's superior performance in the 30-day window likely stems from its ability to first clearly identify low-risk patients before focusing computational resources on ambiguous cases. This approach is particularly valuable when the positive case rate is lower, as in the 30-day window (23.6%).

Interestingly, the static window LSTM models performed particularly well for medium and longer-term predictions (60-365 days), despite being primarily developed as exploratory approaches. These models, which focus exclusively on specific time periods before TBI rather than the entire pre-TBI history, achieved the highest AUC ROC scores for the 60-day (0.762), 180-day (0.782), and 365-day (0.763) windows. This result is most likely because the model has more data to learn from. For example, when predicting the 31-60 day window, the static LSTM knows the data in days 0-30.

Across all models and windows, there was a consistent trade-off between precision and recall. Higher precision (fewer false positives) could be achieved by adjusting classification thresholds, but at the cost of lower recall (more false negatives). This trade-off has important implications for clinical implementation, as different clinical scenarios might prioritize minimizing either false positives (unnecessary interventions) or false negatives (missed opportunities for intervention).

## Feature Importance

Analysis of feature importance across models revealed consistent patterns (Figure 2):

1. **Professional salary costs**: The strongest predictor across all models and time windows, with particularly strong importance in the 180-day window.
2. **Sponsor rank and related interactions**: Military rank-related features showed high importance, especially the interaction between sponsor rank and healthcare utilization (sponsor\_visit\_rate).
3. **Diagnosis history**: Prior diagnoses emerged as increasingly important for longer prediction windows (180, 365 days).
4. **Healthcare utilization patterns**: Frequency and timing of healthcare visits showed varying importance by window, with visit frequency being the most predictive feature in the 60-day window.
5. **Treatment intensity**: A composite measure of healthcare engagement showed consistent importance across all windows.

A screenshot of a graph

AI-generated content may be incorrect.

A notable temporal trend emerged: visit-related features and professional costs dominated short-term prediction (30 days), while diagnosis-related features and sponsor characteristics became increasingly important for longer-term prediction.

## Weighting Approaches

I explored several weighting strategies to address class imbalance and temporal relevance:

1. **Class weighting**: Assigning higher importance to positive cases during training.
2. **Temporal decay variations**: Testing different decay rates (λ) for time weighting.
3. **Combined approaches**: Integrating both class and temporal weighting.

The combined approach showed the best performance, with a 3× class weight and temporal decay rate of λ=0.02 yielding optimal results. This combined approach achieved a 5% relative improvement in AUC ROC compared to unweighted models, with particularly notable gains in recall for positive cases.

A graph and diagram of a graph

AI-generated content may be incorrect.

The effectiveness of the combined weighting approach highlights the complementary nature of these methodological considerations. Class weighting ensures appropriate attention to the minority class (positive cases), while temporal weighting ensures focus on the most relevant time periods. Together, these approaches help the model navigate both the class imbalance and temporal relevance challenges inherent in this prediction task.

These weighting explorations provide methodological insights that extend beyond this specific prediction task, offering guidance for healthcare prediction tasks more broadly. The finding that moderate approaches to both class and temporal weighting typically outperform more extreme approaches suggests that balanced consideration of competing methodological concerns is often optimal in healthcare prediction contexts.

# Risk Stratification

To translate model predictions into actionable clinical insights, I developed a risk stratification approach that categorizes patients into three risk tiers based on their likelihood of developing a mental health condition within 60 days after a TBI. This helps prioritize interventions and allocate resources efficiently.

## How Does the Model Identify High vs Low Risk Individuals

The model assigns each patient a probability score (between 0 and 1) representing their predicted likelihood of developing a mental health condition. These probabilities are based on patterns in pre-TBI healthcare data including:

* Healthcare costs
* Military rank
* Previous diagnoses
* Healthcare visits

Patients are then classified into risk groups using the following probability threshholds:

* **Low-risk (<0.3 probability)** → Unlikely to develop a condition.
* **Medium-risk (0.3–0.7 probability)** → Moderate chance.
* **High-risk (>0.7 probability)** → Very likely to develop a condition.

For example, a patient with frequent medical visits, high healthcare costs, and past mental health issues would receive a high probability score, classifying them as high-risk. Conversely, a patient with few pre-TBI healthcare interactions and no mental health history would likely be low-risk.

**Table 3.** Risk Stratification Results (60-day window, fixed thresholds)

|  |  |  |  |
| --- | --- | --- | --- |
| Risk Tier | Patients (%) | Actual Risk | Number Needed to Screen |
| Low | 40.5% | 0.5% | 202.5 |
| Medium | 45.6% | 23.0% | 4.3 |
| High | 13.9% | 92.1% | 1.1 |

A graph of a bar chart

AI-generated content may be incorrect.

The model demonstrates strong discrimination between risk levels:

* **High-risk patients (13.9%)** are almost certain to develop a mental health condition (92.1% actual risk), meaning resources should be focused on this group for early intervention.
* **Low-risk patients (40.5%)** have an extremely low likelihood (0.5%), allowing healthcare providers to avoid unnecessary monitoring.
* **Medium-risk patients (45.6%)** fall in between, benefiting from regular but less intensive follow-up.

## Number Needed to Screen (NNS)

The **NNS metric** is calculated from precision, and helps measure efficiency:

* **High-risk group (NNS = 1.1)** → Screening 11 patients finds 10 who will develop a condition.
* **Low-risk group (NNS = 202.5)** → Screening over 200 patients finds only 1 true case, making intensive screening impractical.

Patients classified as low-risk (40.5% of the population) had only a 0.5% actual risk of developing a mental health condition within 60 days, with an NNS of 202.5. This indicates that the model is equally effective at identifying truly low-risk individuals who are very unlikely to develop mental health conditions in the short term. The ability to confidently identify low-risk patients is clinically valuable, as it allows healthcare systems to avoid unnecessary intensive follow-up for a substantial portion of the TBI population.

The medium-risk tier (45.6% of the population) showed an intermediate actual risk of 23.0%, with an NNS of 4.3. This group would benefit from enhanced monitoring and standard mental health screening but might not require the intensive intervention appropriate for the high-risk group.

I also explored alternative risk stratification approaches using equal-sized patient groups (terciles and quartiles) rather than fixed probability thresholds. The tercile approach divided patients into three equal groups based on predicted risk, while the quartile approach created four equal groups. These approaches showed similar discrimination between risk tiers but with different trade-offs between tier sizes and risk levels.

The choice between these risk stratification approaches depends on the specific clinical implementation context, available resources, and relative costs of false positives versus false negatives. However, all approaches demonstrated the model's strong ability to discriminate between patients at different levels of mental health risk following TBI.

# DISCUSSION

## Clinical Implications

The study highlights the feasibility of leveraging pre-TBI healthcare data to predict post-TBI mental health diagnoses, offering critical clinical applications:

* **Early Risk Identification:** Our predictive models effectively identify high-risk individuals immediately after TBI, particularly within the crucial 60-day window (AUC ROC 0.762). This enables timely mental health interventions, potentially reducing long-term complications and improving patient outcomes.
* **Precision Risk Stratification:** Patients can be categorized into distinct risk tiers, with high-risk individuals facing a 92% probability of receiving a diagnosis within 60 days, compared to just 0.5% in the low-risk group. This stratification enables targeted, proactive mental health management.
* **Resource Optimization:** The number needed to screen (NNS) in the high-risk group is just 1.1, allowing healthcare systems to efficiently allocate resources to those most at risk, ensuring cost-effective mental health care.

## Performance Trends Across Time Windows

The model’s predictive performance varied across different time windows, offering insights into post-TBI mental health risk dynamics. In the **short-term (30-day window),** the two-stage LSTM approach performed best, achieving an AUC ROC of 0.709. This method effectively filtered out negative cases before refining predictions, but its reliance on immediate post-TBI factors may have overlooked conditions that develop later.

In the **medium-term (60–180 days)**, the model reached peak performance, with AUC ROC scores ranging from 0.762 to 0.782. This time frame struck an optimal balance between early identification and predictive accuracy, with static window LSTM models performing particularly well. The strong results suggest that pre-TBI healthcare patterns play a crucial role in identifying at-risk individuals.

For **long-term predictions (365 days)**, model performance remained strong (AUC ROC **0.763**), though key predictive factors shifted over time. While healthcare utilization patterns dominated earlier predictions, demographic and diagnostic history became increasingly influential at longer time horizons. Additionally, precision lift—which measures the improvement over random guessing—was highest at 30 days (107.6%), but steadily declined as the prediction window lengthened, reaching 44.1% at 365 days. This trend reflects the increasing baseline prevalence of mental health diagnoses over time, making long-term predictions inherently easier but less distinctively powerful.

## Key Predictive Factors and Their Implications

Analysis of predictive features reveals critical insights into risk factors and intervention opportunities:

1. **Healthcare Utilization:** Pre-TBI healthcare costs emerged as one of the strongest predictors, indicating that frequent medical engagement may signal underlying mental health vulnerabilities.
2. **Social Determinants:** Military rank-related factors significantly influenced outcomes, likely reflecting variations in socioeconomic status, occupational stressors, and access to support networks.
3. **Temporal Shifts in Predictive Features:** Short-term predictions relied heavily on recent healthcare visits, while long-term models emphasized diagnostic history, suggesting that different mechanisms drive risk at different stages of post-TBI recovery.
4. **Treatment Intensity:** The stability of treatment intensity metrics across time suggests that healthcare utilization patterns serve as reliable indicators of mental health risk.

Methodological enhancements, including temporal weighting and class balancing, significantly improved model performance, reinforcing their importance in predictive healthcare modeling.

## Limitations and Future Directions

Despite strong predictive capabilities, our study has several limitations that warrant further investigation:

1. **Generalizability:** Our models were developed using military populations, potentially limiting their applicability to civilian healthcare settings with different demographic and medical characteristics.
2. **Outcome Specificity:** We used broad ICD-10 F-code classifications; future research should refine predictions for specific conditions such as PTSD, depression, and anxiety disorders.
3. **Data Constraints:** Our dataset lacked critical variables, including TBI severity, social support structures, and educational background, which may influence mental health trajectories.
4. **Implementation Feasibility:** Prospective clinical validation is needed to assess real-world effectiveness and integration into clinical workflows.

Future research should focus on developing condition-specific predictive models to improve targeted interventions. Validating these models across diverse healthcare populations and settings is essential to ensure generalizability. Incorporating additional data sources, such as medication history, social determinants, and TBI severity, could enhance predictive accuracy. Further, real-world testing of risk stratification frameworks is needed to assess their clinical effectiveness. Lastly, improving the interpretability of neural network models will be crucial for facilitating clinical adoption and increasing trust in AI-driven predictions.

References

1. Centers for Disease Control and Prevention (CDC). Surveillance report of traumatic brain injury-related emergency department visits, hospitalizations, and deaths. Atlanta, GA: CDC; 2019.
2. Jorge RE, Robinson RG, Moser D, Tateno A, Crespo-Facorro B, Arndt S. Major depression following traumatic brain injury. Arch Gen Psychiatry. 2004;61(1):42-50.
3. Bombardier CH, Fann JR, Temkin NR, Esselman PC, Barber J, Dikmen SS. Rates of major depressive disorder and clinical outcomes following traumatic brain injury. JAMA. 2010;303(19):1938-1945.
4. Stein MB, Jain S, Giacino JT, Levin H, Dikmen S, Nelson LD, et al. Risk of posttraumatic stress disorder and major depression in civilian patients after mild traumatic brain injury: A TRACK-TBI study. JAMA Psychiatry. 2019;76(3):249-258.
5. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785-794.
6. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735-1780.
7. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363. 2018.
8. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018;1(1):1-10.

1. [↑](#footnote-ref-1)