Evaluating the Impact of Differential Privacy on ICU Mortality Prediction

Ioana-Andreea Cristescu Harvard University
AJ Baily Harvard University
Alyssa Mia Taliotis Harvard University

Abstract

1 Introduction

Machine learning models are increasingly used in healthcare to support clinical decision-making, including predicting patient outcomes such as ICU mortality. At the same time, these models often rely on sensitive patient data, which raises serious privacy concerns and motivates the need for privacy-preserving machine learning methods.

In this project, we study how enforcing differential privacy (DP) during model training affects the utility, fairness, and interpretability of ICU mortality prediction models built using the MIMIC-IV-ED clinical database. Differential privacy provides formal guarantees that limit the exposure of individual patient data, but it can also reduce model utility, particularly in terms of predictive performance and calibration. Understanding this privacy-utility tradeoff is especially important in clinical settings, where reliable predictions and equitable treatment across patient subgroups are critical. In addition, it remains unclear how DP impacts the interpretability of machine learning models, which is essential for building trust in clinical decision support systems. To address these gaps, we focus on two research questions: (1) How does training with differentially private stochastic gradient descent (DP-SGD) affect the utility and fairness of ICU mortality prediction models across clinically and demographically relevant subgroups? (2) How does the use of DP influence model interpretability, specifically in terms of the stability of feature importance rankings produced by explanation methods such as SHAP and LIME?

2 Related Works

2.1 Ficek et al. (2021) [1]

In their paper, Ficek et al. provide a broad overview of differential privacy applications in health research, highlighting that as of 2021, most existing work focused on genomic data with few instances of real world implementation, and even fewer with conventional machine learning techniques. Their project highlights the relevance of our project, as we look to implement DP with a real clinical dataset and assess how privacy preserving methods impact utility, fairness, and interpretability for ICU mortality prediction. With a mere eight papers discussing the privacy-utility tradeoff, and none of those papers implementing conventional machine learning techniques, we seek to implement DP-SGD on a real-world dataset to evaluate how well it preserves model utility and fairness under formal privacy constraints.

2.2 Fang et al. (2024) [2]

Fang et al. propose **De**centralised, **C**ollaborative, **a**nd **P**rivacy-preserving ML for Multi-**H**ospital Data (DeCaPH) to allow hospitals to train ML models without exchanging their information directly. They apply their model to three tasks, predicting ICU mortality from EHR data, classifying single cell RNA transcripts, and identifying pathologies in chest X-rays with varying ϵ values from .62 to 5.65. They use a variety of techniques, including a Multi-Layer Perceptron, Deep Convolutional Neural Net, Logistic Regression, and Support Vector Classifier. They report that applying their

privacy-preserving methods typically reduces accuracy by only a few percentage points (often under 3%) compared to non-private approaches, and they argue this minor drop is worth the substantial gains in patient confidentiality. Overall, the authors view this small accuracy tradeoff favorably, demonstrating that privacy protections can be integrated into multi-hospital collaborations without crippling model performance. Their work offers a valuable comparison point for our project, noting that our project will focus on the privacy utility tradeoff for a single comprehensive dataset while analyzing interpretability and fairness.

2.3 Pang et al. (2022) [3]

Pang et al. demonstrate the robustness of MIMIC-IV for predicting ICU mortality by combining subscores commonly used to help clinicians assess prognosis in ICUs. They compare four classifiers - XGBoost, Logistic Regression, SVM, and Decision Trees. Their results show AUCs ranging from .852 (Decision Tree) to .918 (XGBoost). These will be helpful for our benchmarks. They also used SHAP to highlight the most useful predictive features from the MIMIC dataset. More importantly, their work highlights the benefit of nonlinear decision boundaries in mid-risk ranges for ICU mortality, while emphasizing that high-risk mortality ranges are well classified with linear boundaries. This suggests that architectures capable of classifying more complex relationships may generalize better as the authors also explain that most individuals in the MIMIC dataset were classified as mid-risk.

3 Description of Data [4]

The MIMIC-IV dataset was sourced by Beth Isreal medical center and de-identified in accordance with the HIPAA Safe Harbor provision. Patients in the dataset received three arbitrary randomized surrogate identifiers, subject_id, hadm_id, and stay_id. stay_id is available in all datasets and will allow us to track patient care across different sectors of care. Dates and times in the dataset were randomly shifted to a year and time between 2100-2200, consistent on a subject_id basis, ensuring consistency in timelines on a per-subject_id basis across all datasets. The dataset we are using is a subset of the larger MIMIC-IV dataset, MIMIC-IV-ED, the emergency department subset deliberately designed to support education initiatives and research studies. Of the ~425,000 stays available in the macro dataset, our analysis will focus on a random subset of 100 from that dataset.

The authors do note the below disclaimer in their paper, which we will factor into our analysis -

"Data contained within MIMIC-IV-ED are collected during routine clinical care, and their use for research is secondary to their use in clinical care. The data may contain implicit biases as a result of local data collection practices, implausible values for measurements, and missing documentation for provided treatments. Many interventions, including major events such as endotracheal intubation, are not documented clearly. Researchers should take care to address these issues in their work"

4 Data Structure

4.1 Hospital Module - All patients are linked by the subject_id field in each csv

- **patients.csv** – De-identified patient-level information
- Key fields include:
  - **anchor_age**: Age, capped at 91 for patients older than 89.
  - **anchor_year_group**: Three-year range in which the patient's admission falls.
  - **gender**: reported gender
  - **date of death**: does not mean the patient died inside of the hospital
- **admissions.csv** - Hospital level admissions records
- Key fields include:
  - **admittime, dischtime**: Times for hospital admission and discharge
  - **deathtime:**: time of death if death occured in the hospital
  - **hospital_expire_flag**: binary flag indicating if the patient died in the hospital or not
- **transfers.csv** - tracks movement of care throughout the hospital

- Key fields include:
  - **eventtype:**: admission, discharge, transfer
  - **careunit, ward**: Name where the patient was transfered (MICU, SICU, etc)
  - **intime, outtime**: Timestamps for leaving or entering a specific ward
- **procedures_icd.csv:** - list of patient procedures
- Key fields include:
  - **chartdate:** date of the procedure - useful for building timelines for the patient (if the patient had complications from previous care)
  - **icd_code/icd_version:** joinable with d_icd_procedures.csv to get human readable procedures
- **labevents.csv** - List of lab work done for a given patient id
- Key fields include:
  - **itemid:** unique ID for the lab test - joinable with d_labitems.csv for human readable lab tests
  - **value:** value of the lab test
  - **ref_range_lower (and upper):** quantities to discern how abnormal a patients lab tests were

From the hospital dataset, we can use gender, age, and comorbidities from the icd_codes to determine fairness across subgroups in our dataset. Noting the de-identification strategy of the authors of the dataset, we will need to build a timeline of events from chart times across each dataset by subject ID to determine a timeline of patient care.

4.2 ICU Module - All patients are linked by the subject_id field in each csv

- **chartevents.csv:** the largest csv in the MIMIC dataset. This includes all events for a patient, can be joined with d_items.csv for human readable item codes
- Key fields include:
  - **itemid:** field linked to d_items.csv, this has codes for things like gender, race, and initial patient status among many other options
  - **charttime:** time the event was logged, can be used to discern trends in patient care
- **icustays.csv:** tracks icu stays for a patient
- Key fields include:
  - **intime, outtime:** tracks the time the patient was in the ICU
- **inputevents.csv:** tracks the medications given to a patient while in the ICU
- Key fields include:
  - **statusdescription:** tracks whether a patient starts a medication, has the medication paused, changes dosage or finishes
  - **itemid:** tracks the name of the medication, linkable to a credentialed database within MIMIC
- **procedureevents.csv:** list of events done while a patient is in the ICU
- Key fields include:
  - **starttime, endtime:** gives the duration of the procedure, potential indicating trivial or non trivial events
  - **ordercategoryname:** has a series of categories that could indicate the procedure, with names like "intubation" and "imaging"

In combination with the information from the hospital database, we can find out trends in patient care to build strategies for determining privacy, interpretability and potential bias. Key fields from this module include itemid in chartevents, as well as statusdescription in inputevents to determine any changes in the status of patients. As the authors note in their paper, there are patients that exist in the ICU module that may not be in the hospital module due to the severity of their care, so the hospital module will likely provide supporting information but may be a source of data imbalance or data missiningness.

5 Implementation Strategy

Our implementation strategy is designed to rigorously investigate the trade-offs between privacy and utility in clinical prediction models, with a focus on ICU mortality using the MIMIC-IV-ED Clinical Database Demo. Building on the original concept, we will systematically compare models trained using standard optimization techniques and differentially private stochastic gradient descent (DP-SGD), examining how the introduction of differential privacy (DP) affects predictive performance, interpretability, and fairness across clinically and demographically relevant subgroups.

We begin by preprocessing the MIMIC-IV-ED dataset to extract relevant features for ICU mortality prediction (e.g., vitals, lab measurements, demographics), ensuring appropriate handling of missing data and variable encoding. Our baseline model will be a regularized logistic regression trained without differential privacy to establish performance benchmarks. We will then retrain this model using DP-SGD, evaluating how privacy impacts classification metrics such as accuracy, AUROC, and calibration (e.g., Brier score, reliability curves).

To explore how model complexity influences the privacy-utility tradeoff, we will extend our experiments to include more flexible architectures such as multi-layer perceptrons and tree-based models (e.g., XGBoost with added DP mechanisms via output perturbation or objective perturbation). We hypothesize that more complex models may suffer greater performance degradation under strong DP constraints, and our implementation will test this hypothesis across different model classes.

We will additionally quantify how differential privacy affects model interpretability using post-hoc explanation techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). This will allow us to assess whether privacy-preserving models preserve the clinical relevance and consistency of feature importance rankings, which is crucial for deployment in high-stakes settings such as ICUs.

To evaluate fairness, we will stratify performance metrics across key subgroups (e.g., age, gender, race, comorbidity profiles) to investigate whether DP-SGD exacerbates or mitigates existing biases. We are particularly interested in whether DP introduces disproportionate utility loss in underrepresented or vulnerable populations.

An important dimension of our study will be experimenting with a range of privacy budgets (ϵ values) to construct privacy-utility curves, helping to identify practical operating points that balance performance and privacy. This sensitivity analysis will inform healthcare practitioners and developers on the trade-offs involved in tuning DP mechanisms for real-world clinical decision support systems.

Libraries and Tools

- **PyTorch / TensorFlow Privacy** – For training neural networks using DP-SGD.
- **Scikit-learn** – For baseline models like logistic regression and evaluation utilities.
- **XGBoost with custom DP wrappers** – For implementing tree-based models under privacy constraints.
- **Opacus** – A PyTorch library for training models with differential privacy.
- **SHAP / LIME** – For generating post-hoc model explanations and visualizations of feature importances.
- **Pandas / NumPy** – For data preprocessing and manipulation.
- **Matplotlib / Seaborn** – For visualizing performance metrics and privacy-utility tradeoffs.

Evaluation Metrics

Utility Metrics

- AUROC (Area Under the Receiver Operating Characteristic Curve)
- Accuracy
- Precision, Recall, F1 Score
- Calibration (Brier Score, Calibration Curves)

Fairness Metrics

- Subgroup AUROC / Accuracy
- Disparity in calibration across demographic groups
- Equality of opportunity / Demographic parity (where applicable)

Privacy Metrics

- Privacy budget (ε) used in DP-SGD
- Sensitivity analysis across multiple ε values to visualize the privacy-utility tradeoff
- Composition accounting to understand cumulative privacy loss across training epochs

6 Timeline

Week	Task
Week 1	Data preprocessing: extract features from MIMIC-IV-ED, handle missing values, encode variables. Train baseline models (logistic regression, MLP, XGBoost) without differential privacy. Evaluate accuracy, calibration, and fairness metrics.
Week 2	Integrate DP-SGD using Opacus. Train models with varying ε values. Compare privacy-preserving models to baselines on utility and fairness. Run SHAP/LIME for interpretability assessment.
Week 3	Complete analysis of fairness impact under DP. Plot privacy-utility curves. Draft and finalize report. Discuss limitations, future work, and prepare poster.

7 Fallback Plan

If our initial experiments reveal that training complex models like neural networks with differential privacy leads to severe performance degradation, we will focus our analysis on simpler models such as logistic regression, which are often more robust under privacy constraints. In the event that SHAP or LIME explanations become unstable or uninformative when applied to DP-trained models, we will instead evaluate interpretability by directly comparing model coefficients or feature weights across private and non-private models. Additionally, if the MIMIC-IV-ED demo dataset proves too small to support reliable subgroup analyses, we will either aggregate similar subgroups to increase sample size or shift our focus toward overall utility and calibration metrics rather than fairness-specific evaluations.

8 Feedback from TF

This is a promising project. Whenever dealing with real data settings and a variety of model comparisons, the threat is that the data handling time can dwarf the work that demonstrates course mastery or in this case builds on differential privacy. So as soon as possible, get data in hand, and get non-privacy-preserving baseline models running. This will iron out any issues. You don't want to discover data issues at the last moment. Another foundational piece will be to nail down what you envision by utility. I assume one form of utility will be some form of loss/precision/recall. But, you've noted the worry about bias, and bias is really a (negative) utility definition, so quickly come up with some working definition of bias that you can measure in your work. With data in hand, and metrics to work against, you should have a good foundation. You have some lovely cites for exemplars of DP in the health modelling space. A broader set can be found in this paper: https://academic.oup.com/jamia/article-abstract/28/10/2269/6333353

References

[1] Ficek, J., Wang, W., Chen, H., Dagne, G., & Daley, E. (2021). Differential privacy in health research: A scoping review. *Journal of the American Medical Informatics Association*, 28(10), 2269–2276. https://doi.org/10.1093/jamia/ocab135

[2] Fang, C., Dziedzic, A., Zhang, L., Oliva, L., Verma, A., Razak, F., Papernot, N., & Wang, B. (2024). Decentralised, collaborative, and privacy-preserving machine learning for multi-hospital data. *eBioMedicine*, 101, 105006. https://doi.org/10.1016/j.ebiom.2024.105006

[3] Pang, L., Li, F., Zhang, R., & Xiong, W. (2022). Combining LODS and APS III for ICU Mortality Prediction: A Comparative Study of Machine Learning Methods using MIMIC-IV. *IEEE Access*, 10, 18532–18542. https://doi.org/10.1109/ACCESS.2022.3163490

[4] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMICIV Clinical Database Demo (version 2.2). PhysioNet. https://doi.org/10.13026/dp1f-ex47

[5] Mehrotra, S., Kilambi, V., Gilroy, R., Ladner, D. P., Klintmalm, G. B., & Kaplan, B. (2015). Modeling the Allocation System. *Transplantation*, 99(2), 278-281. https://doi.org/10.1097/tp.0000000000000656

[6] Gentry, S. E., Segev, D. L., Kasiske, B. L., Mulligan, D. C., & Hirose, R. (2015). Robust Models Support Redistricting Liver Allocation to Reduce Geographic Disparity. *Transplantation*, 99(9). https://doi.org/10.1097/tp.0000000000000834

Appendix

A.1

A.2