The Potential of AI in Nursing Care: A Multi–Center Evaluation in Fall Risk Assessment

Ivana Nanevski, Sebastian Jäger, Matthias Schulte-Althoff, Eva-Maria Behnke, Daniel Fürstenau, Felix Biessmann

Submitted to: Journal of Medical Internet Research
on: January 08, 2025

Table of Contents

The Potential of AI in Nursing Care: A Multi–Center Evaluation in Fall Risk Assessment

Ivana Nanevski[1*] MSc; Sebastian Jäger[1*] MSc; Matthias Schulte-Althoff[2, 3] PhD; Eva-Maria Behnke[4] BSc; Daniel Fürstenau[2, 3] PhD; Felix Biessmann[1, 5] PhD

[1] Berliner Hochschule für Technik Berlin DE
[2] Institute of Medical Informatics Charité - Universitätsmedizin Berlin Berlin DE
[3] School of Business & Economics Freie Universität Berlin Berlin DE
[4] Medizinische Hochschule Brandenburg Theodor Fontane Neuruppin DE
[5] Einstein Center Digital Future Berlin DE
[*]these authors contributed equally

Corresponding Author:
Ivana Nanevski MSc

Berliner Hochschule für Technik
Luxemburger Str. 10
Berlin
DE

Abstract

Background: With 28%-35% of individuals aged 65 and older experiencing incidents of falling, falls are the second leading cause of unintentional injury-related deaths globally. Limited availability of clinical staff often impedes timely detection and prevention of potential falls. Advances in artificial intelligence (AI) could complement existing fall risk assessment and help to better allocate nursing care resources. Yet, many studies are based on small datasets from a single institution, which can restrict the generalizability of the model, and do not investigate important aspects in AI model development such as fairness across demographic groups.

Objective: This study aims to provide a comprehensive empirical evaluation of the potential of AI in nursing care, focusing on the case of fall risk prediction. To account for demographic and contextual differences in fall incidences, we analyze data from a university and a geriatric hospital in Germany. To the best of our knowledge, these are the largest datasets for fall risk prediction to date with heterogeneous data distributions. We focus on three key objectives: Does AI help in improving fall risk prediction? Which approaches should be considered and how can AI models be trained safely across different hospitals? Are these models fair?

Methods: This study used two datasets for fall risk prediction: one from a university hospital with 931,912 subjects, 3,351 of whom experienced falls, and another from a geriatric hospital with 12,773 subjects, 1,728 of whom have fallen. State of the art AI models were used within three experimental approaches. First, separate models were trained on the data from each hospital; second, models were retrained on the respective other dataset; and Federated Learning (FL) was applied to both datasets for collaborative learning. The performance of these models was compared to the rule-based systems for fall risk prediction. Additional analysis was conducted to test for model fairness.

Results: Our findings demonstrate that AI models consistently outperform rule-based systems across all experimental setups, with AUROC of 0.735 (90% CI 0.727 - 0.744) for the geriatric hospital, and 0.93 (90% CI 0.928 - 0.934) for the university hospital. FL did not improve the fall risk prediction in this setting. Our fairness analysis ruled out disparities in model performance between different gender groups, but we found fairness infringements in age-based performance.

Conclusions: This study demonstrates that AI models consistently outperform traditional rule-based systems across heterogeneous datasets in predicting fall risk. However, it also reveals the challenges related to demographic shifts and label distribution imbalances, which limited the FL models' ability to generalize. While the fairness analysis indicated promising predictive parity and equal opportunity across gender subgroups, age-related disparities emerged. Addressing data imbalances and ensuring broader representation across demographic groups will be crucial for developing more fair and generalizable models.

(JMIR Preprints 08/01/2025:71034)
DOI: https://doi.org/10.2196/preprints.71034

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

The Potential of AI in Nursing Care: A Multi-Center Evaluation in Fall Risk Assessment

Abstract

Background: With 28%-35% of individuals aged 65 and older experiencing incidents of falling, falls are the second leading cause of unintentional injury-related deaths globally. Limited availability of clinical staff often impedes timely detection and prevention of potential falls. Advances in artificial intelligence (AI) could complement existing fall risk assessment and help to better allocate nursing care resources. Yet, many studies are based on small datasets from a single institution, which can restrict the generalizability of the model, and do not investigate important aspects in AI model development such as fairness across demographic groups.

Objectives: This study aims to provide a comprehensive empirical evaluation of the potential of AI in nursing care, focusing on the case of fall risk prediction. To account for demographic and contextual differences in fall incidences, we analyze data from a university and a geriatric hospital in Germany. To the best of our knowledge, these are the largest datasets for fall risk prediction to date with heterogeneous data distributions. We focus on three key objectives: Does AI help in improving fall risk prediction? Which approaches should be considered and how can AI models be trained safely across different hospitals? Are these models fair?

Methods: This study used two datasets for fall risk prediction: one from a university hospital with 931,912 subjects, 3,351 of whom experienced falls, and another from a geriatric hospital with 12,773 subjects, 1,728 of whom have fallen. State of the art AI models were used within three experimental approaches. First, separate models were trained on the data from each hospital; second, models were retrained on the respective other dataset; and Federated Learning (FL) was applied to both datasets for collaborative learning. The performance of these models was compared to the rule-based systems for fall risk prediction. Additional analysis was conducted to test for model fairness.

Results: Our findings demonstrate that AI models consistently outperform rule-based systems across all experimental setups, with AUROC of 0.735 (90% CI 0.727 - 0.744) for the geriatric hospital, and 0.93 (90% CI 0.928 - 0.934) for the university hospital. FL did not improve the fall risk prediction in this setting. Our fairness analysis ruled out disparities in model performance between different gender groups, but we found fairness infringements in age-based performance.

Conclusions: This study demonstrates that AI models consistently outperform traditional rule-based systems across heterogeneous datasets in predicting fall risk. However, it also reveals the challenges related to demographic shifts and label distribution imbalances, which limited the FL models' ability to generalize. While the fairness analysis indicated promising predictive parity and equal opportunity across gender subgroups, age-related disparities emerged. Addressing data imbalances and ensuring broader representation across demographic groups will be crucial for developing more fair and generalizable models.

Keywords: fall risk prediction; machine learning; artificial intelligence; federated learning; clinical decision support; nursing care; fairness

Introduction

Background

Falls are the second most common cause of unintended injury-related deaths worldwide, mostly affecting elderly people. Globally, falls are responsible for approximately 680,000 deaths annually [1,2]. Furthermore, falls have been identified as one of the most prevalent risk factors impacting elderly people, especially inpatients [3,4]. In the United States the fall rate is 3.56 per 1000 patient-days [5]. In Germany, an analysis of data from 55 institutions in 2004 reported that the fall rate in hospitals was 4.2 per 1000 patient-days, while in nursing homes, it was 5.1 per 1000 patient-days [6]. According to [7], the average incidence of falls in nursing homes is 1.5 falls per bed per year, with a range from 0.2 to 3.6 falls per bed annually.

The incidence of falls among the elderly population not only causes significant physical health risks, including fractures and head injuries, but also results in psychological and social consequences, such as fear of falling, which can impair quality of life [1,4]. Moreover, this also leads to increased healthcare costs and longer hospital stays [5]. The limited availability of professional resources such as physical therapists, nurses, and doctors hinders timely detection and prevention of potential falls [8]. In Germany, nursing care wards are understaffed with an estimated shortage of 100,000 to up to 520,000 nurses by 2030 [9]. According to recent reports this will continue to increase until 2049 with 280,000 to 690,000 missing nurses [10]. At the same time, according to [2], 28% [11] to 35% [12] of individuals aged 65 and older experience at least one fall, a rate that increases to 50% for those aged 80 years and older. In addition, it is estimated that by 2030 one in six individuals will be aged 60 or older [4]. Therefore, it is crucial to develop highly effective assessment tools for predicting fall risk in patients to reduce the number of people who experience falls and to enable nurses to make more informed decisions regarding fall risk management [13,14].

Existing rule-based fall risk assessment tools, such as The Expert Standard for Fall Prophylaxis (ESFP) [15] and the World Guidelines for Falls Prevention (WGFP) [16], are built on existing literature and focus on a broad range of indicators. ESFP focuses on demographic parameters, risk indicators such as fear, fractures, lack of mobility and balance, contracture, diabetes, calcium deficiency, overweight, depression, mobility aid such as walking stick, wheelchair or knee tutor, dementia or cognitive impairment, and the Timed "Up & Go" (TUG) test [17]. WFGP is based on a set of rules that cover mobility related risk factors such as the fear of falling, mobility aid such as walking stick, wheelchair or knee tutor and the TUG test, sensory functioning indicators such as dizziness, glasses, acuity, contract perception, hearing aid, then the Barthel index (BI) [18], a measurement for Activities of Daily Life (ADL), the cognitive function rules, such as the Mini-Mental State Examination (MMSE) [19], presence of delirium, behavioral patterns such as excitement, focus, apathy, tendency to stray, then autonomous functions such as orthostatic hypotension, nocturia or incontinence, medical history such as signs of Parkinson or depression, and nutrition history and vitamin D deficiency.

In contrast to these guidelines using a broad range of indicators, others often focus on a smaller set of indicators, for example, the Morse Fall Scale [20], the Hendrich II Fall Risk Model [21], and the St. Thomas Risk Assessment Tool [22]. Often those use specialized tests such as TUG or Tinetti [23] evaluate whether patients have balance or walk impairments. In addition to traditional risk assessment tools, there have been attempts to develop machine learning (ML) and artificial

intelligence (AI) (we use ML and AI interchangeably) models for better fall risk assessment [3–5,8,14,24–32]. Multiple studies leverage data from wearable devices [8,24,25] and camera motion tracking data [29,31] for fall prediction. For example, the authors of [31] use a camcorder to gather data, and they base their analysis on different types of falls. Other studies focus on tabular data [3,4,14,28,30,32,33], such as demographic information, initial diagnosis, risk factors (such as history of falls, Activities of Daily Living (ADL), medication and/or cognitive impairments. The most commonly used models for tabular data in these studies were Support Vector Machines (SVM), Logistic Regression, Decision Trees, Random Forrest and Gradient Boosting Trees. Of these, Random Forests and Gradient Boosting Trees have been shown to perform better on tabular data [34].

However, these studies rely on data from a single hospital, which could restrict the models' ability to generalize to datasets from other hospitals. These generalization issues stem from fundamental demographic differences between datasets and emphasize the importance of inclusive data collection, as different distributions often carry distinct risk factors [27,35]. Achieving broader population representation for training AI models can be challenging, particularly in hospital settings, where data privacy is important. However, there are potential solutions, such as retraining of models on local data [36], Federated Learning (FL) [37] and variations thereof [38]. In such collaborative learning, multiple participants or institutions simultaneously train a common AI model. FL often shows advantages in the medical domain, addressing key challenges such as data privacy, data scarcity, and the need for collaborative research. Numerous studies [39–42] have shown that FL can help in developing joint models by leveraging data from multiple centers.

Objective

In this study, we use two large-scale tabular datasets, coming from one of the largest hospitals of Germany and from a smaller geriatric hospital. We explore the potential of AI in addressing one of the key challenges in modern healthcare systems: developing data-driven AI models which can handle heterogeneous data distributions, such as those arising from age variations. Our study focuses on three primary research objectives. First, we examine whether AI models can provide improvements over traditional rule-based fall risk assessment tools used in nursing care. Second, we explore approaches for training AI models to collaborate on two large-scale datasets. Lastly, we assess if the models are fair across diverse demographic groups. To the best of our knowledge, this study represents the most comprehensive evaluation on the potential of ML for fall risk assessment in terms of data size, and no existing related studies have specifically focused on investigating the fairness aspect of a fall prediction model on such a large-scale dataset in Germany. -

Methods

Datasets

General Data overview

We used two anonymized datasets extracted from a large university hospital and a geriatric hospital in Germany. Both datasets contain central demographic information (age and sex), diagnoses, procedures, and fall risk assessment scores obtained from tools like the TUG test or the Jones Index [43]. The procedures, which are interventions performed by a healthcare professional to diagnose, treat, monitor, or prevent a health condition, follow the standardized *German Operationen- und Prozedurenschlüssel (OPS) [44]*. The diagnoses follow the *International Statistical Classification of*

Diseases, 10. Revision, German Modification (ICD) [45], representing the patients' most common diagnosis in their electronic health records (EHR).

ICD and OPS

ICD and OPS are hierarchical classification systems. For example, ICD I21.4 encodes "acute subendocardial myocardial infarction", which belongs to "acute myocardial infarction" (I24). This broader category is itself part of "ischemic heart diseases" (I20 to I25) that fall under the umbrella of "diseases of the circulatory system" (I00 to I99). Preliminary experiments with different levels showed best results when using the fourth level (i.e., I21.4) if available, otherwise level three (Z11). Similarly, for OPS, we use the third level. For example, "diagnostic catheter examination of the heart and circulation" (1-27) belongs to "examination of individual body systems" (1-20 to 1-33), which is part of "Diagnostic Measures" (1-10 to 1-99).

Fall Risk Assessments

Fall risk assessments are conducted manually by nursing staff during the admission process after hospitalization and updated regularly (at least every five days). The goal is to gather information to determine whether patients have an increased risk of falling to apply preventive measures. Both hospitals' assessments are motivated by ESFP and WGFP described above. Therefore, they similarly ask for past fall incidences, mobility impairments, cognitive impairments, excretions, and certain medications, such as psychotropic drugs or sedatives. However, they use different measurements to determine patients' impairments.

The university hospital adheres to the patient classification system introduced by Jones [43], which assess the patients' independence in different dimensions of daily life, such as, feeding, personal toilet, and walking. On the other hand, the geriatric hospital employs a suite of measurements specialized for specific aspects. The TUG test measures the time (in seconds) required for a patient to stand up from a chair, walk 3 meters, turn, walk back, and sit down again. A TUG score of ≥ 20 seconds is indicative of a mobility impairment. The *Mini-Mental State Examination* (MMSE) [19] test is a series of questions and verbal and written commands to test for cognitive impairments. Patients can achieve up to 30 points (more is better), where MMSE ≤ 23 is considered a cognitive impairment. Finally, the *Tinetti* test assesses patients' balance (15 points) and gait (13 points) to determine their risk of falling. A score of ≤ 20 (out of 28) points is indicative of an increased fall risk.

University Hospital Data

The university hospital dataset comprises of 931,912 Electronic Health Records (EHRs) ranging between 2016 and 2022. The patients range in age from 19 to 124 years, with a mean age of about 58 years and a median age of 59 years. About (496,588/931,912, 53%) of the patients are female. A mere (3351/931,912, 0.36%) of patients have experienced at least one fall incident, which occurs more often for older patients.

Geriatric Hospital Data

This dataset is considerably smaller, with only $12{,}773$ patients between 2019 and 2022. Furthermore, the prevalence of fall incidence is higher, (1,728/12,773, 13.53%), and the patients are older, ranging between 62 and 102 years (mean: 79, median: 80). Additionally, this dataset includes a measure of the patients' independence, the Barthel index [18]. The index considers various daily activities such as, feeding, personal toilet, walking, etc., and computes a score between 0 and 100, where 100 indicates full independence and 0 indicates full dependence on help.

Common Data Schema

In this study, we encountered the challenge of integrating data from two different sources, the geriatric hospital and university hospital, which initially had distinct data schemata. To develop a

comprehensive and effective predictive model, it was crucial to harmonize these datasets into a unified format. This process involved identifying common features between the two hospitals' data schemata and aligning them accordingly. To achieve a unified dataset schema, we mapped the common features across both hospitals, ensuring that each corresponding attribute hold the same type of information and adhered to the same format. For example, the indicator if a person is wearing orthosis in one dataset is listed under "medical items", and in the other under "mobility", both defined as text fields. We mapped both columns to a common binary feature. For the features unique to either geriatric or university hospital, we incorporated them into the combined schema by assigning null values where the data is not available. This approach ensures that no information is lost while preserving the dataset's integrity. The common data schema consisted of 124 columns. By carefully managing the disparities in data schemata, we ensured that our fall prediction model can perform well across both hospital environments.

Ethics Approval

The ethics committee of the Charité – Universitätsmedizin Berlin, Germany approved this data analysis (EA2/184/21). Due to the retrospective design of the study using data from standard care the need for informed consent was waived. The data protection officers of the Charité and the EGZB advised on data protection rules to ensure compliance with those rules. Data from the Charité were anonymized using standard procedures involving their Health Data Platform, while the EGZB anonymized their data according to internal procedures prior to analysis. We consulted the guidelines for developing and reporting machine learning predictive models in biomedical research [46].

Experiments

Baselines

Expert Standard for Fall Prophylaxis

The ESFP is curated by the German Network for Quality in Nursing (DNQP) [15] and the German Board of Trustees for the Elderly (KDA) [47]. The aim of the ESFP is to determine the risk of falls for people in need based on a selection of risk factors. These factors include the history of falls and fractures, fear from falling, mobility impairment (strength, balance, endurance, flexibility) and cognitive impairments. They recognize three additional types of risk factors: personal fall risk factors, like vision impairment or cognitive impairment, medication-related fall risk factors, like psychotropic drugs which affect mental processes, and environmental fall risk factors, e.g., obstacles on the floor. This guideline is used by nurses who should assess patients for all potential risk factors and note their severity. Patients who are at high risk of falling should receive help that prevents falls.

World Guidelines for Falls Prevention

The WGFP [16] is the product of a worldwide collaboration of 96 experts. While developing the guidelines, they incorporated input from elderly people and final decisions were reached through a minimal voter consensus. Initially, the guidelines have two ways of assessing the fall risk. First, there is the opportunistic case finding which includes people who fell in the last 12 months. This approach has low sensitivity as it does not take any risk factors into account but can be used for certain patients, for whom more information is not available. The other, more recommended tool from the guideline, applies to patients who have fallen in the past year, or feel unsteady when standing or walking, or worry about falling. Then, an additional assessment identifies the risk of falling. The overall assessment is rule-based, and the end outcome is the indicator if a patient needs an additional help. Its decisions are based on some risk factors such as injury or loss of consciousness, as well as some directly measured values, like a TUG test resulting in longer than 15 seconds.

Implementation

To implement these rule-based systems, we followed the well-known scikit-learn[48] API for ML models, which offers a *fit*, *predict*, and *predict_proba* method. Since these baselines follow fixed rules, the fit method only exists for interoperability and does not do anything. We implemented these rule-based systems based on the available data, described in Section Datasets, which predict increased risk of falling if any of the rules apply. However, in our implementation *predict_proba* computes how many of the predefined rules apply. For some of the features, e.g., diabetes and depression, we looked up whether the patient had ICD codes associated with them.

AI Model and Experimental Frameworks

AI model

Random Forests and Gradient Boosting Trees are recognized for their performance on tabular data, as evidenced by recent research [34]. Among these techniques, XGBoost has distinguished itself as a particularly effective algorithm for handling tabular datasets and has been used in the specifically the medical domain [3,28]. In our study, initial experiments indicated that tree-based models outperform other model classes (Multimedia Appendix, Table 8). Therefore, in this study, we chose XGBoost as our AI model and its python implementation [49] accordingly.

Federated Learning FL [37] is a collaborative learning approach wherein multiple participants or institutions simultaneously train a machine learning model. Instead of collecting training data in a central repository, the sensitive data stays within the respective institution and instead of the data, model parameters are shared between institutions, thus maintaining data privacy. This approach is particularly advantageous when individual institutions lack sufficient data to independently train a robust model. Each participant then sends their model updates to a central server to aggregate them and create a unified common model. The unified model is then sent back to all participants for additional training. FL often shows advantages in the medical domain, addressing key challenges such as data privacy, data scarcity, and the need for collaborative research. Numerous studies [39–42] have shown that FL can help in developing joint models by leveraging data from multiple medical institutions. For our FL experiments, we used Flower [50], a community-driven python tool that supports XGBoost. Flower facilitates FL by enabling multiple clients to collaboratively train a model without sharing their raw data. We chose this tool because of the community support and the ease of implementing such a complex framework.

Experimental Setup

Baselines

We included the WGFP and ESFP as baselines against which we benchmarked our models. As our first key objective included comparison with these rule-based models, we applied them to each dataset to assess whether AI models can provide improvements over the established rule-based fall risk assessment tools for nursing care.

Separate Models

One model for each dataset was trained exclusively on that data, providing a baseline for hospital-specific performance. The reason behind these sets of experiments was two-fold. First, we wanted to compare each hospital individually with the baselines and conclude the assumptions in the first objective. Second, with this experiment, we partially focused on our second objective: Approaches for training AI models for collaboration of two large scale datasets. These results served as a baseline

comparison for our other approaches in training AI models. It is important to acknowledge that in a setting where hospitals do not collaborate with each other, this is the only viable experiment on a single-institution basis. We will hereinafter refer to these models as *Separate*.

Retrained Models

Initially trained on data from one hospital, each of the two models was subsequently retrained using data from the other hospital, allowing us to assess the benefits of knowledge transfer between hospitals. This approach was particularly useful when notable domain shifts exist between institutions, as it allowed the models to better represent a broader, real-world population. The idea behind this experiment was to gain intuition on how two different large-scale hospital datasets, with considerable demographic shifts within the age distributions, can complement each-others knowledge. With this experiment, we partially focused on our second objective and showed an approach for collaboration of two large scale datasets: Is FL more advantageous than model retraining from one hospital to another? We will refer to these experiments as *Retrained*.

FL Model

This model was trained simultaneously on data from both hospitals. As detailed in Section Federated Learning, FL is a collaborative model training across multiple institutions while preserving privacy, which allows models to benefit from diverse data sources. We performed this experiment to determine whether combining these two different large-scale datasets could contribute to a more effective model by leveraging the collective knowledge from both institutions. We will refer to this model as *Federated*.

Table 1. AI model—range of hyperparameter values.

Parameter	Range
num_parallel_tree	(10, 210)
eta	(5e-3, 0.2)
max_depth	(3, 11)
colsample_bytree	(0.5, 1)
min_child_weight	(1, 6)

Data and Models Environment

The Experiments were conducted using a single machine to simulate a distributed environment. For the retrained experiments, the model was initially trained on one dataset and then saved to disk. Subsequently, the model was loaded from disk, along with a second dataset, and further trained. In a practical distributed setting, instead of saving the model to disk, it would be digitally transferred to another hospital or accessed via a shared infrastructure. For the FL experiments, a central server managed the secure distribution of models to participating clients. Again, all three processes were run on a single machine, with two clients simulating hospitals, each accessing a single dataset and communicating solely with the central server. This single-machine simulation is functionally equivalent to a distributed system with machines communicating over the internet.

Cross Validation

To determine the optimal set of hyperparameters in each experiment, we conducted random search [51] for each dataset, selecting certain ranges for each hyperparameter (Table Error: Reference source not found). We performed a nested 5-fold cross validation to optimize for the best set of hyperparameters, consisting of: a learning rate (eta) of 0.0089, 97 parallel trees (num_parallel_tree), maximum depth (max_depth) of 5 for each tree, the minimum child weight (min_child_weight) was set to 3, and the 57.8% (0.578) of all features were used by each tree (colsample_bytree). These

parameters provided the best results across both institutions. In the outer loop of our 5-fold cross validation we evaluated the models' generalization and robustness by comparing the test set scores over every fold. In contrast to other studies [5,30,52], we trained and tested on stratified splits derived from the original dataset distribution. We did not remove positive samples from the distribution nor performed Synthetic Minority Oversampling Technique (SMOTE) to account for the imbalanced classes.

Evaluation Metrics

In our model evaluation strategy, we aimed at comparability with previous work and evaluated all relevant metrics but focus primarily on the established metrics. We chose AUROC for optimization and model evaluation similarly to previous studies on fall risk assessment [3–5]. Given that our models were optimized for the highest AUROC score, we needed a calibrated threshold and a better suited metric for the binary decisions when directly comparing the models. To perform a calibrated evaluation, we selected a fixed decision threshold based on the F1 score, since it is a combined metric of both precision and recall, which were further used for fairness analysis as well.

Fairness Metrics

Equal Opportunity

While there are a sizeable number of fairness metrics available [53], in this study we focused on the most established ones [54]. Equal Opportunity specifically focuses on ensuring that the False Negative Rate (FNR) is equal across the different groups. The False Negative Rate is the probability that a subject in the positive class is incorrectly predicted to be in the negative class [54]. Equal Opportunity requires:

$$FNR = \frac{FN}{FN + TP}$$

where FN is the number of false negatives and TP is the number of true positives. For Equal Opportunity to be satisfied, we need:

$$FNR_{Group\ A} = FNR_{Group\ B}$$

Mathematically, a classifier with equal FNR will also have equal True Positive Rate (TPR). In this study we used TPR to express equal opportunity as it is directly translated into recall.

To give an example on how equal opportunity is relevant for this study, consider two patients in our geriatric ward: one is 70 years old female with a high risk of falling, and the other is 70 years old male with the same risk level. Suppose the predictive model has missed that the female patient is at high risk of falling and thus this patient does not get timely intervention. Equal opportunity then requires that the model should similarly estimate that the male patient should not be at risk, thus a potential gender inequality would be excluded from the model in this scenario.

Predictive Parity

Predictive Parity examines whether the probability that a subject with a positive predictive value (PPV) truly belongs to the positive class (or precision) is equal across different subgroups [54]. The precision is measured as:

$$PPV = \frac{TP}{TP + FP}$$

where TP is the number of true positives FP is the number of false positives. For Predictive Parity to hold, we need:

$$PPV_{Group\ A} = PPV_{Group\ B}$$

In this work we choose this metric as it is directly translated into Precision.

To explain how predictive parity would be relevant in this setting, consider two patients in our geriatric ward: one is 70 years old with a high risk of falling, and the other is 90 years old with the same risk level. Suppose the predictive model has identified the 90-year-old patient is at high risk of falling and thus this patient needs additional preventive measures. Predictive parity then requires that the model should similarly identify the 70-year-old patient as high risk so this patient would not be excluded from intervention, no matter their age.

Feature importance

To investigate potential causes for difference in predictive performance between the established rule-based systems and AI models, we used SHapley Additive exPlanations (SHAP) to finding the most important features for our models. It is a commonly used approach among researchers for interpreting AI models [3,14]. SHAP calculates the average marginal contribution of each feature by considering all possible combinations of features, providing a comprehensive measure of each feature's influence on the prediction. It is especially important in clinical settings, ensuring that the AI model decisions are explainable. In this study we used shap [55], a widely used python implementation for SHAP analysis.

Results

AI models and FL outperform the rule-based models in fall risk prediction

The Precision-Recall (PR) curves (Figure 1) illustrate the performance of all three experiments conducted on both datasets. For both hospitals, the areas under the PR curves indicate that the AI models achieve a substantial improvement over the baselines, and in case of the university hospital the baseline models do not yield any positive results. For the geriatric hospital (Figure 1a), there is a consistent performance across the different experiments: Separate, Retrained and Federated, even after incorporating data from the university hospital. FL does not improve results in any of the settings explored. In the case of the geriatric hospital, FL does not improve predictive performance. For the university hospital (Figure 1b), there is a notable decline in performance in the FL experiment, suggesting that the Federated approach disadvantages this hospital. Similarly, the Retrained models also fail to improve performance over the Separate models.

To further investigate this, we also compare the models' AUROC scores (Table 2). Based on the distributions of performance scores across the 5 cross-validation folds we computed medians and 90% confidence intervals (CIs) (5th and 95th percentile) for each of the three experiments: Separate, Retrained and Federated. For the geriatric hospital, we observe no improvement over the Separate model from the other models. In contrast, FL results in a decrease in AUROC score for the university hospital, dropping from 0.93 with the Separate model to 0.75. Overall, our findings indicate that retraining and FL can result in models "unlearning" crucial information at the hospital level, thereby failing to outperform the Separate Models. In this setting, FL does not offer advantages over retraining models either. That indicates that FL is unable to effectively leverage the knowledge from the two hospitals, which have considerably different demographic distributions. Consequently, we continue our analysis based on the Separate models.

For a calibrated evaluation and for the fairness analysis, we select a fixed decision threshold based on the F1 score, a combination of both precision and recall. Our analysis confirms that the AI models consistently outperform the rule-based models (Multimedia Appendix, Table 1 and 2).

Figure 1 The AI based models (Separate, Retrained and Federated) substantially outperform the baseline rule-based models (ESFP, WGFP). FL (gray curves) struggles to effectively integrate knowledge from the two hospitals with considerable demographic shifts, leading to reduced performance for the university hospital.

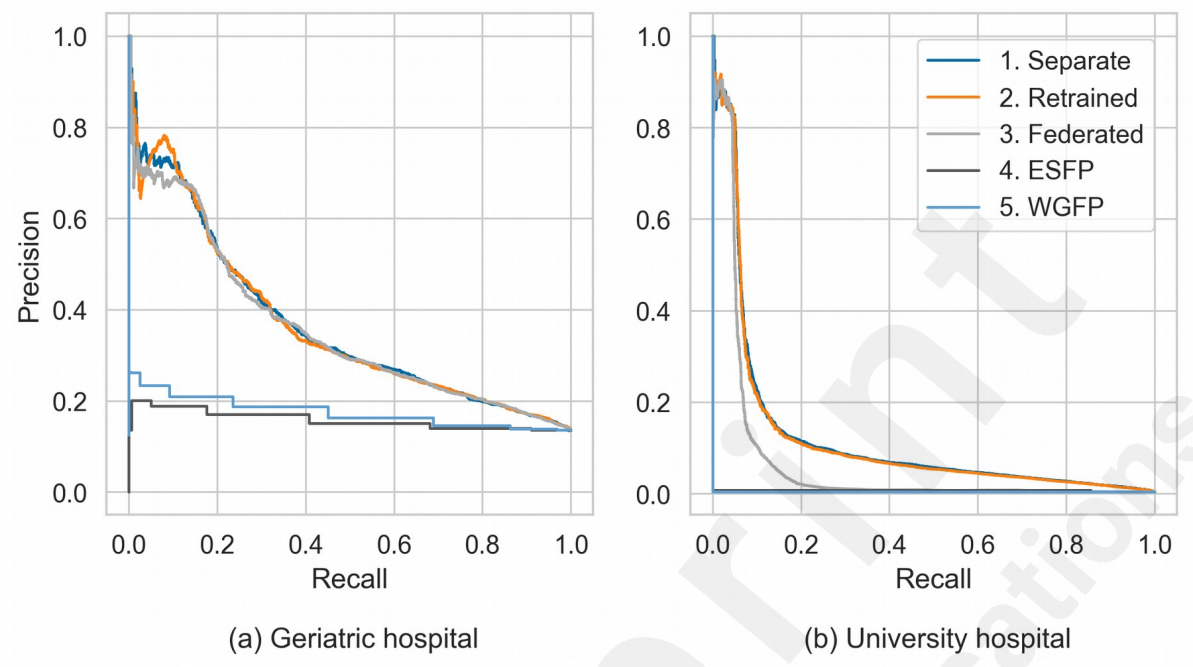


Table 2. Median AUROC scores and 90% CI (5th and 95h percentile): FL fails to enhance performance, with a notable decrease in AUROC scores at the university hospital from 0.93 (Separate and Retrained models) to 0.75, and no improvement over the Separate and Retrained models for the geriatric hospital.

	Geriatric hospital	University hospital
ESFP	0.556 (0.533, 0.598)	0.712 (0.709, 0.717)
WGFP	0.606 (0.572, 0.608)	0.500 (0.500, 0.500)
Separate	**0.735 (0.727, 0.744)**	**0.930 (0.928, 0.934)**
Retrained	0.735 (0.724, 0.744)	0.927 (0.925, 0.930)
Federated	0.735 (0.725, 0.741)	0.750 (0.708, 0.759)

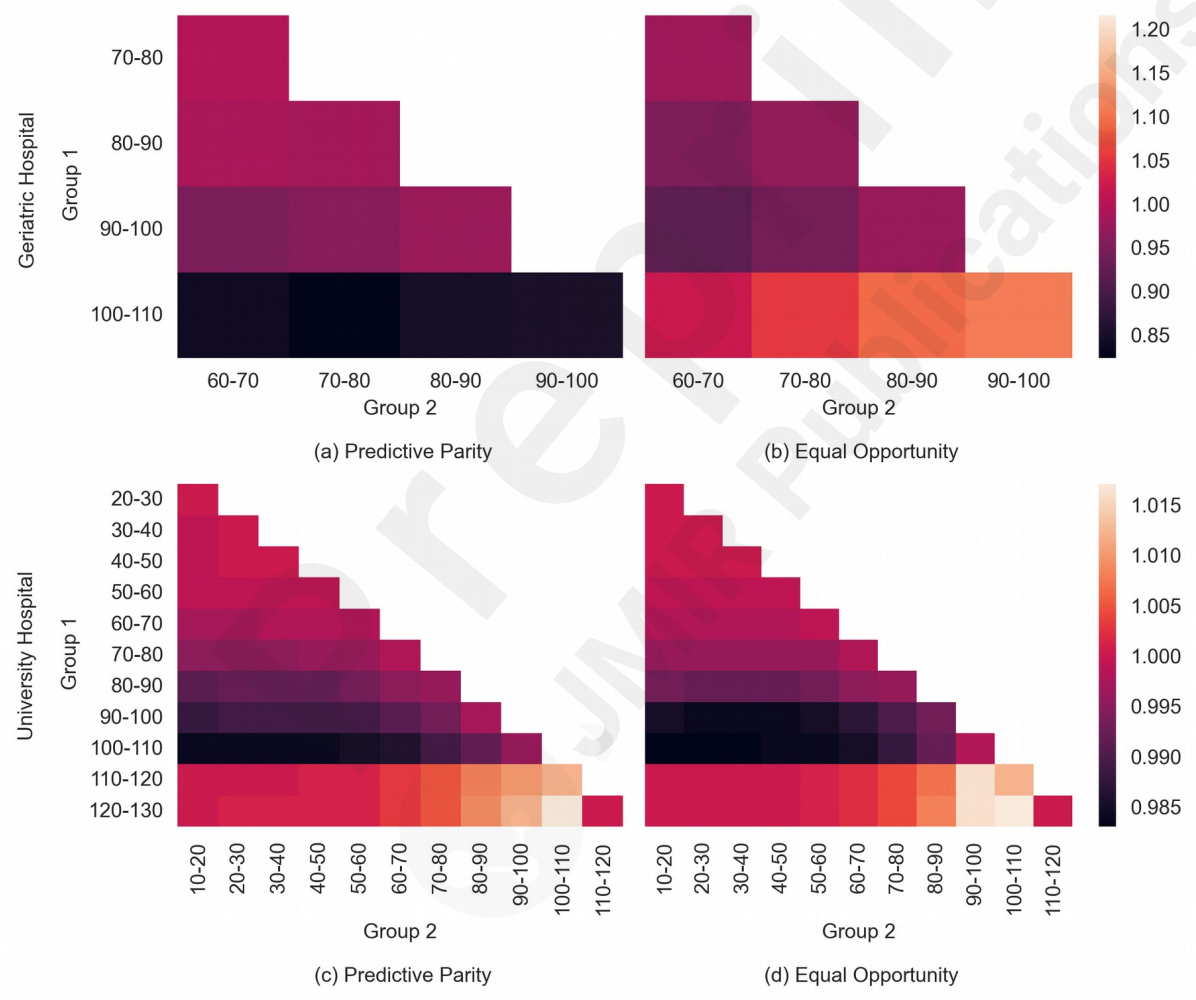
Demographic fairness analysis of fall prediction

Fairness analyses across demographics

To assess fairness across the demographic factors of age and gender, we examine predictive parity (related to precision) and equal opportunity (related to recall) as shown in (Figure 2). For each demographic group within both datasets, we compute five PPV and five TPR scores from 5-fold cross-validation. Fairness is evaluated by calculating the median ratio of scores between two groups within each demographic. A ratio of 1 indicates equal performance between the groups, while a ratio greater than 1 suggests that Group 1 performs better, and a ratio less than 1 suggests Group 1 is disadvantaged. In terms of gender, the models show fairness for predictive parity and equal opportunity on both datasets, in both statistical tests and ratio point estimates (Multimedia Appendix, Table 3). However, when looking at age, there are noticeable disparities in performance between age groups. To support these findings, we apply a Wilcoxon signed-rank test for both age and gender

subgroups (Multimedia Appendix, Table 4, 5, 6, 7). While the test results suggest fairness across all use cases, they do not fully align with our point ratio estimates (Figure 2). We attribute this discrepancy to the tests not reaching significance, likely due to the limited sample sizes.

Figure 2 Fairness analysis indicates infringements of AI fall risk assessment models for some age groups. Models demonstrate lower precision for patients aged 100-110, indicating overestimation of fall risk, particularly in the geriatric hospital (a, c). However, this fairness disparity is less pronounced in the university hospital, where the precision score is closer to parity. In contrast, the geriatric hospital model shows reduced recall for patients aged 90-100 (b), failing to accurately identify those patients who fall, in comparison to patients aged 100-110. In the university hospital model (d), the recall score for patients aged 90-100 and 100-110 is also lower, though still near parity. The ratios represent the score of Group 1 over the score of Group 2. Darker shades indicate that Group 1 has a lower score than Group 2, a lighter shade indicates that Group 1 has a higher score than Group 2, a score of 1 indicates that both groups have equal scores.



Fairness in the Geriatric Hospital

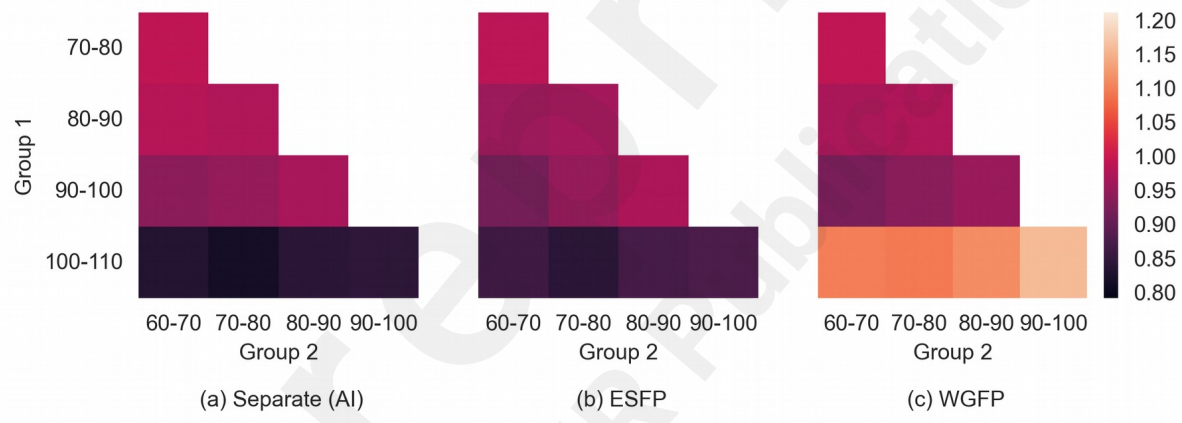
For predictive parity in the geriatric hospital dataset (Figure 2a) patients aged 100-110 show lower precision, meaning there are more false positives in this age group. This could lead to unnecessary interventions, such as prescribing treatments for patients who are not actually at high risk of falling. However, it is important to note that this age group consists of only 12 patients (0.09% of the dataset). For equal opportunity (Figure 2b), patients aged 90-100 are less likely to be correctly identified as high risk compared to those aged 100-110. This suggests that patients in the 90-100 age

range may not receive timely interventions, increasing their fall risk.

Fairness in the University Hospital

In the university hospital dataset, predictive parity (Figure 2c) shows a similar trend, with the 100-110 age group having lower precision, though the difference is less pronounced than in the geriatric hospital and closer to parity. For equal opportunity (Figure 2d) both the 90-100 and 100-110 age groups have lower TPR scores, indicating that these patients would be less likely to be correctly identified as high fall risk. However, here again the disparity is smaller compared to the geriatric hospital dataset and approaches parity more closely.

Figure 3 Predictive parity comparison—geriatric hospital. The AI model and the ESFP demonstrate lower precision for patients aged 100-110, indicating overestimation of fall risk, particularly the AI model (a). However, this fairness disparity is lower for the ESFP (b). In contrast, the WGFP model shows higher precision for patients aged 100-110 (c), having more false positives for the rest of the groups. The ratios represent the score of Group 1 over the score of Group 2. Darker shades indicate that Group 1 has a lower score than Group 2, a lighter shade indicates that Group 1 has a higher score than Group 2, a score of 1 indicates equal scores across both groups.



Fairness comparison between AI models and rule-based models

To provide a better perspective on the AI models' fairness, we compare their performance with the two baseline models, the ESFP and WGFP. In the university hospital dataset, no considerable fairness issues were found when applying either model. However, in the geriatric hospital dataset, we observe differences in the Predictive Parity metric, particularly across age groups (Figure 3). As previously discussed, the AI model (Figure 3a) results in lower precision for patients aged 100-110, although this age group represents only 0.09% of the dataset. This trend is also present, though less pronounced, in the ESFP (Figure 3b). The WGFP (Figure 3c), however, shows a different trend. In this model, patients aged 100-110 have a higher precision score compared to other age groups, suggesting a lower fall risk estimate for this group. This implies that in a real-world scenario, the WGFP would treat patients aged 60-100 as higher risk than those aged 100-110, potentially leading to overtreatment. However, it is important to note that, unlike the AI model, the ESFP and WGFP do not undergo training on our dataset. Their rules are predefined outside of this study, meaning that fairness outcomes are more likely due to the models' inherent design rather than the data distribution.

Feature Importance

Geriatric Hospital

In (Figure 4) we visualize the most important features for fall risk prediction from the geriatric hospital data, highlighting the influence of each feature on the model's output. The most influential features are *procedure, the billable diagnosis code (DRG), has_dementia_or_cognitive_impairment, secondary_diagnosis (as ICD code), age, the tinetti score,* and *MMSE*. If a patient has dementia or cognitive impairment, they have a higher risk for falling. If a patient has a higher tinetti score, they have a lower risk of falling. In comparison to the baselines, this feature importance analysis suggests that the model overlaps with the ESFP in four decisive features, *sex, has_dementia_or_cognitive_impairment, age* and *TUG*, and an overlap of five features with the WGFP, *TUG, MMSE, barthel_index, has_delirium* and *has_parkinson*. This shows that the AI models diverge in certain aspects from the rule-based models when assessing fall risk.

University Hospital

In (Figure 5) present the most influential features for fall risk prediction from the university hospital data. The most influential features are *procedure, has_decubitus_admission, age, the walk jones score, has_decubitus_atm* and *secondary_diagnosis (as ICD code)*. If a patient has decubitus on admission or at the moment when the assessment is done, they are of lower risk for falling. In comparison with the baselines, this feature importance analysis suggests that the model overlaps with three decisive features with the ESFP, *age*, *sex* and *has_cognition_impairment*, and no overlapping features with the WGFP. Again, as for the geriatric hospital, the AI models diverge from the baselines with the most decisive features.

Figure 4 Geriatric hospital—each point along the x-axis represents a single patient and indicates its SHAP value for that feature. Red color means that the feature has higher value, and a blue color means that the feature has lower value. Categorical features are shown as gray points. The most important features include procedure, has_dementia_or_cognitive_impairment, the billable diagnosis code (DRG), secondary_diagnosis, age, the Tinetti score, and MMSE. Individuals with dementia or cognitive impairment face a higher fall risk (red points along the X axis). Patients with higher Tinetti scores are mostly at lower risk of falling.

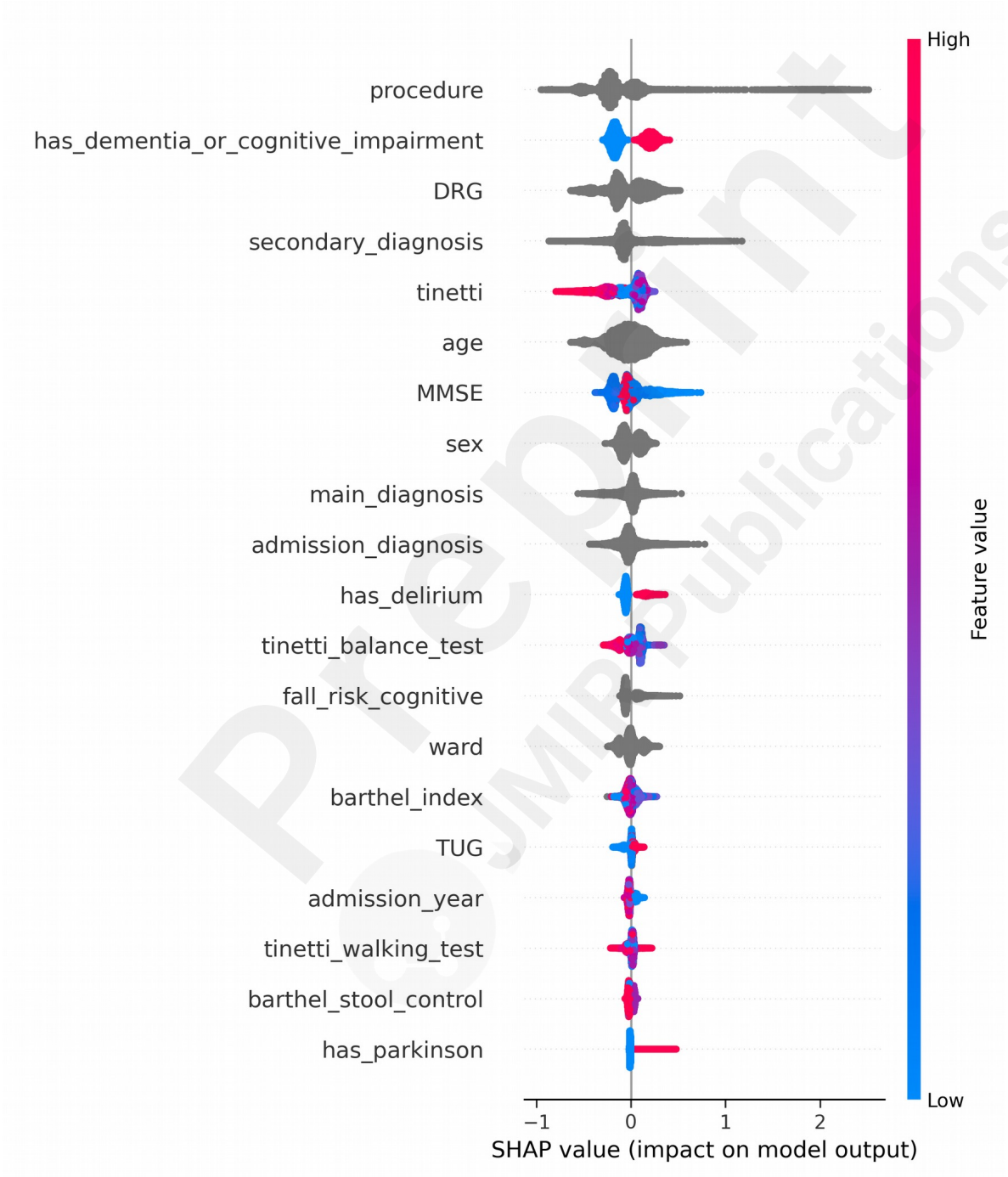
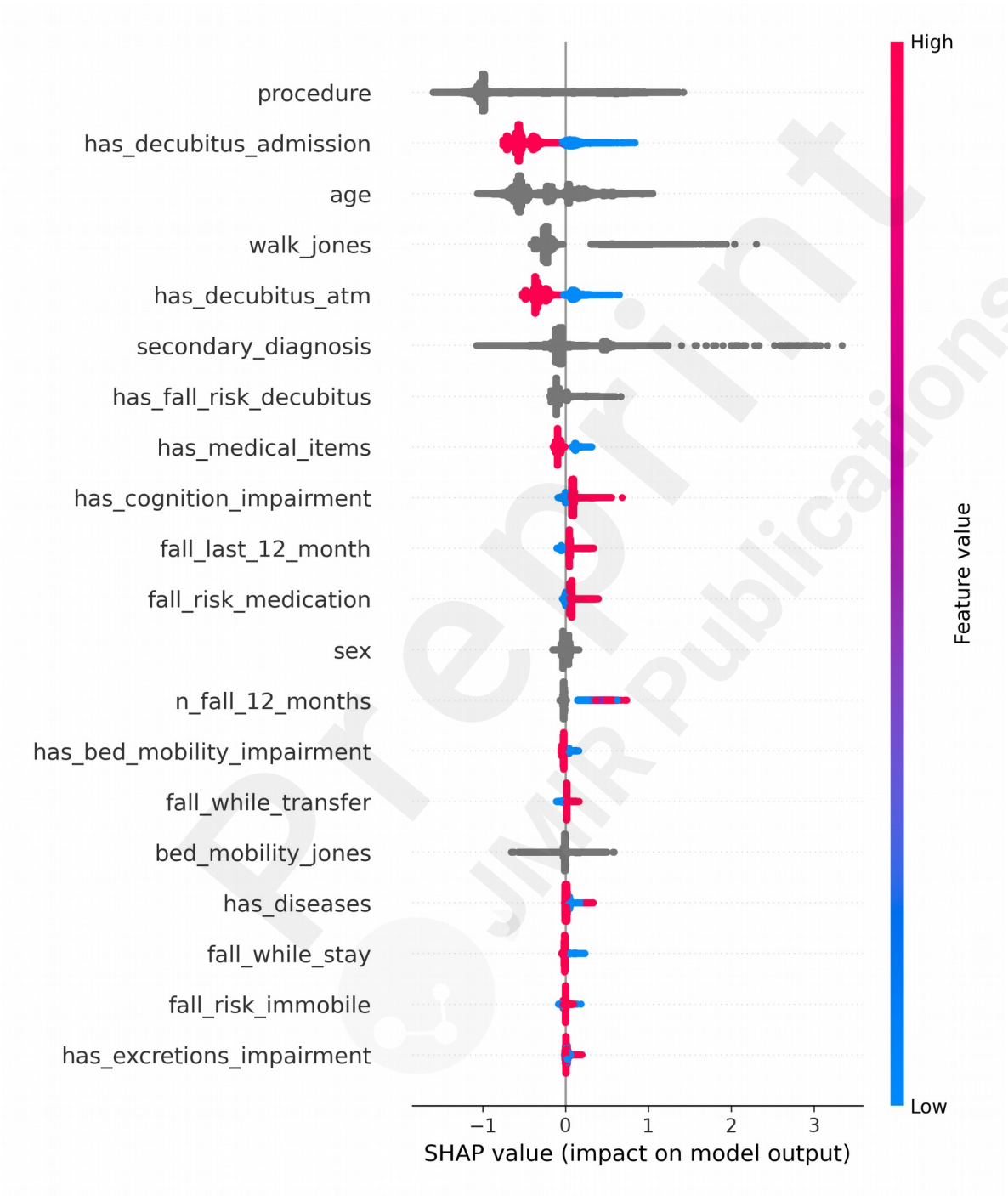


Figure 5 University hospital—each point along the x-axis represents a single patient and indicates its SHAP value for that feature. Red color means that the feature has higher value, and a blue color means that the feature has lower value. Categorical features are shown as gray points. The most influential features are procedure, has_decubitus_admission, age, the jones score, has_decubitus_atm and the secondary_diagnosis. Individuals who have decubitus on admission or at the moment of the assessment are at lower risk of falling (red points along the X axis).



Discussion

Principal Results

In this study, we compared three AI based approaches with traditional rule-based systems, ESFP and WGFP, for predicting fall risk across two hospital datasets with considerable demographic shifts. Our

results show that AI based models consistently outperform rule-based models, with the university hospital dataset achieving an AUC-ROC score of 0.93 compared to 0.712 and 0.5 for the ESFP and WGFP models, respectively. In the geriatric hospital, the AI model similarly outperformed the rule-based models, with an AUC of 0.735 compared to 0.556 for ESFP and 0.606 for WGFP. The improvements in recall over the rule-based models used in practice translates into more accurate identification of patients at risk for falls. The higher precision scores of the AI models mean that they would reduce the number of false positive patients if these models were used in clinical settings. This would lead to an improvement in nurse staffing. Our experiments also included training separate models for each dataset, retraining models across datasets, and applying FL to combine knowledge across both datasets. Neither retraining nor FL offered improvements over the separate models, particularly when faced with large demographic shifts. In fact, the application of FL led to a notable drop in performance for the university hospital dataset, where the AUC dropped from 0.93 to 0.75. Our fairness analysis focused on two key metrics: predictive parity and equal opportunity across demographic subgroups, particularly age and gender. We found that the AI model shows fairness for predictive parity and equal opportunity on both datasets across the gender subgroups, but age-related disparities were more pronounced in certain subgroups of elderly patients. Nonetheless, the Wilcoxon signed-rank tests for age and gender subgroups did not indicate significant fairness disparities, although the test may not have reached significance due to limited sample size.

Limitations and Future Directions

Data Integration challenges

Because our analysis required integration of heterogeneous data schemata, not all information was available for all patients, leading to data quality problems such as missing data. These issues project further towards the experiments and analyses, specifically the FL model and the fairness analysis.

Challenges with FL

While FL is considered effective for collaborative model training, it did not lead to improvements in our case. Many FL algorithms are built under the assumption that the data is independent and identically distributed (IID) across clients [56,57]. When the data distribution differs across institutions, the performance of the FL model can be impacted negatively [57–59]. There are several reasons why data distributions differ across institutions and not all of them can be easily alleviated. For example, there could be demographic shifts across institutions, nursing homes and wards. Also, there could be different pathologies and diseases more frequently expressed in individual patient populations across institutions, leading to shifts in EHR data. Other shifts are due to data infrastructure and data processing and collection. Here one might opt for other solutions than FL, like coordination among participating institutions, or mitigating feature noise while collecting and processing data. If not handled timely, these discrepancies can then propagate when converting EHRs into structured tabular data.

Challenges in Fairness analysis challenges

Finally, although our fairness analysis did not reveal statistically significant differences in performance across gender and age groups, the overestimation of fall risk in the small group of older patients should be reconsidered on a larger scale.

Future directions

To address our limitations, future analysis will require larger sample sizes across different subgroups to conduct a more refined fairness analysis. Addressing the issues with missing data could also lead towards an improved FL model. Moreover, handling the data imbalance could also enhance the model’s ability to collaboratively learn across hospitals, ultimately leading to a more robust and

generalizable model.

Comparison With Prior Work

In addition to traditional risk assessment tools, there have been attempts to develop AI models for fall risk assessment. We find our work mostly related with [3,4,28,32], which we share a common data format with. However, many studies are constrained to datasets comprising only a few hundred data points, which poses significant challenges in developing models that accurately reflect real-world scenarios. In contrast to some studies [5,30,52], we keep the original data distribution. We do not remove positive samples from the distribution nor perform SMOTE to account for the imbalanced classes, in both train and test splits. Additionally, most of these studies focus on data from a single institution, which limits opportunities to improve model robustness and generalizability through multi-institutional data collaboration. While these studies provide valuable insights and risk factors for fall risk prediction, they differ from our study in several key aspects: Firstly, this study focuses on data from two distinct institutions that cannot be shared, aiming to train a shared AI models while retaining privacy across institutions. Secondly, we leverage real-world and large-scale heavily imbalanced datasets which substantially differ in their age distributions, thus providing a more realistic scenario. Finally, none of these studies prioritize fairness metrics across demographics, whereas our goal is to investigate whether the model delivers fair predictions across all demographic groups.

Conclusions

This study demonstrates the potential of AI models in predicting fall risk, consistently outperforming traditional expert systems across two hospital datasets. However, it also reveals the challenges related to demographic shifts and label distribution imbalances within the datasets, which likely limited the FL models' ability to generalize. Additionally, while the fairness analysis indicated promising predictive parity and equal opportunity across gender subgroups, age-related disparities emerged. To enhance model performance and fairness in future research, addressing data imbalance and ensuring broader representation across demographic groups will be crucial for developing more fair and generalizable models.

Acknowledgments

We thank Armin Hauss, Jörg Pohle, Cem Kozcuer and the entire KIP-SDM team for their valuable feedback and discussions. We also acknowledge the support of Felix Balzer as the Head of the Institute of Medical Informatics at Charité – Universitätsmedizin Berlin. This research was supported by the German Federal Ministry of Education and Research, grant number 16SV8856.

Authors contribution

Conceptualization – IN (lead), SJ (equal),
Data curation – MSA (lead), SJ (equal)
Formal analysis – IN (lead), SJ (supporting)
Funding acquisition – FB (lead), DF (equal)
Investigation – IN (lead), SJ (supporting)
Methodology – IN (lead), SJ (equal)
Project administration – FB
Software – SJ (lead), IN (supporting)
Supervision – FB (lead), DF (equal)
Validation – EMB
Visualization – IN

Writing – original draft – IN (lead), SJ (supporting)
Writing – review & editing – IN (lead), SJ (equal), EMB (supporting), MSA (supporting), FB (supporting), DF (supporting)

Conflicts of Interest

None declared.

Abbreviations

ADL: Activities of Daily Living
AI: Artificial Intelligence
AUROC: The area under the receiver operating characteristic
BI: Barthel Index
CI: Confidence Interval
DNQP: German Network for Quality in Nursing
EHR: Electronic Health Record
ESFP: Expert Standard for Fall Prophylaxis
FL: Federated learning
FNR: False Negative Rate
ICD: International Statistical Classification of Diseases
IID: Independent and Identically Distributed
KDA: German Board of Trustees for the Elderly
ML: Machine learning
MMSE: Mini-Mental State Examination
OPS: Operationen und Prozedurenschlüssel (Operation and Procedure Classification System)
PPV: Positive Predictive Value
PR: Precision-Recall
SHAP: SHapley Additive exPlanations
SMOTE: Synthetic Minority Oversampling Technique
SVM: Support vector machine
TPR: True Positive Rate
TUG: Timed Up & Go
WGFP: World Guidelines for Falls Prevention

References

1. Schoberer D, Breimaier HE, Zuschnegg J, Findling T, Schaffer S, Archan T. Fall prevention in hospitals and nursing homes: Clinical practice guideline. Worldviews Ev Based Nurs 2022 Apr;19(2):86–93. doi: 10.1111/wvn.12571
2. WHO global report on falls prevention in older age. Geneva, Switzerland: World Health Organization; 2008. Available from: https://iris.who.int/bitstream/handle/10665/43811/9789241563536_eng.pdf?sequence=1 [accessed Jun 10, 2024]ISBN:978-92-4-156353-6
3. Thapa R, Garikipati A, Shokouhi S, Hurtado M, Barnes G, Hoffman J, Calvert J, Katzmann L, Mao Q, Das R. Predicting Falls in Long-term Care Facilities: Machine Learning Study. JMIR Aging 2022 Apr 1;5(2):e35373. doi: 10.2196/35373
4. Millet A, Madrid A, Alonso-Weber JM, Rodríguez-Mañas L, Pérez-Rodrá-Guez R. Machine Learning Techniques Applied to the Development of a Fall Risk Index for Older Adults. IEEE Access 2023;11:84795–84809. doi: 10.1109/ACCESS.2023.3299489
5. Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C. Predicting Inpatient Falls Using

Natural Language Processing of Nursing Records Obtained From Japanese Electronic Medical Records: Case-Control Study. JMIR Med Inform 2020 Apr 22;8(4):e16970. PMID:32319959
6. Heinze C, Halfens RJ, Dassen T. Falls in German in-patients and residents over 65 years of age. Journal of Clinical Nursing 2007 Mar;16(3):495–501. doi: 10.1111/j.1365-2702.2006.01578.x
7. Rubenstein LZ. Falls in the Nursing Home. Ann Intern Med 1994 Sep 15;121(6):442. doi: 10.7326/0003-4819-121-6-199409150-00009
8. Chen S-H, Lee C-H, Jiang BC, Sun T-L. Using a Stacked Autoencoder for Mobility and Fall Risk Assessment via Time-Frequency Representations of the Timed Up and Go Test. Front Physiol 2021;12:668350. PMID:34122139
9. Reiff E, Gade C, Böhlich S. Handling the shortage of nurses in Germany: Opportunities and challenges of recruiting nursing staff from abroad. 2020 Mar. Available from: https://econstor.eu/bitstream/10419/222921/1/1726039005.pdf [accessed Jun 10, 2024]
10. Sachverständigenrat Zur Begutachtung Der Entwicklung Im Gesundheitswesen Und In Der Pflege. Fachkräfte im Gesundheitswesen - Nachhaltiger Einsatz einer knappen Ressource. PUBLISSO; 2024; doi: 10.4126/FRL01-006473488
11. Prudham D, Evans JG. Factors Associated with Falls in the Elderly: A Community Study. Age and Ageing 1981 Jan 1;10(3):141–146. doi: 10.1093/ageing/10.3.141
12. Blake AJ, Morgan K, Bendall MJ, Dallosso H, Ebrahim SBJ, Arie THD, Fentem PH, Bassey EJ. FALLS BY ELDERLY PEOPLE AT HOME: PREVALENCE AND ASSOCIATED FACTORS. Age Ageing 1988;17(6):365–372. doi: 10.1093/ageing/17.6.365
13. Lindberg DS, Prosperi M, Bjarnadottir RI, Thomas J, Crane M, Chen Z, Shear K, Solberg LM, Snigurska UA, Wu Y, Xia Y, Lucero RJ. Identification of important factors in an inpatient fall risk prediction model to improve the quality of care using EHR and electronic administrative data: A machine-learning approach. Int J Med Inform 2020 Nov;143:104272. PMID:32980667
14. Mishra AK, Skubic M, Despins LA, Popescu M, Keller J, Rantz M, Abbott C, Enayati M, Shalini S, Miller S. Explainable Fall Risk Prediction in Older Adults Using Gait and Geriatric Assessments. Front Digit Health Frontiers; 2022 May 6;4. doi: 10.3389/fdgth.2022.869812
15. Deutsches Netzwerk für Qualitätsentwicklung in der Pflege. Available from: https://www.dnqp.de/ [accessed Jun 10, 2024]
16. Montero-Odasso M, Van Der Velde N, Martin FC, Petrovic M, Tan MP, Ryg J, Aguilar-Navarro S, Alexander NB, Becker C, Blain H, Bourke R, Cameron ID, Camicioli R, Clemson L, Close J, Delbaere K, Duan L, Duque G, Dyer SM, Freiberger E, Ganz DA, Gómez F, Hausdorff JM, Hogan DB, Hunter SMW, Jauregui JR, Kamkar N, Kenny R-A, Lamb SE, Latham NK, Lipsitz LA, Liu-Ambrose T, Logan P, Lord SR, Mallet L, Marsh D, Milisen K, Moctezuma-Gallegos R, Morris ME, Nieuwboer A, Perracini MR, Pieruccini-Faria F, Pighills A, Said C, Sejdic E, Sherrington C, Skelton DA, Dsouza S, Speechley M, Stark S, Todd C, Troen BR, Van Der Cammen T, Verghese J, Vlaeyen E, Watt JA, Masud T, the Task Force on Global Guidelines for Falls in Older Adults, Kaur Ajit Singh D, Aguilar-Navarro SG, Aguilera Caona E, Alexander NB, Allen N, Anweiller C, Avila-Funes A, Barbosa Santos R, Batchelor F, Becker C, Beauchamp M, Birimoglu C, Blain H, Bohlke K, Bourke R, Alonzo Bouzòn C, Bridenbaugh S, Gabriel Buendia P, Cameron I, Camicioli R, Canning C, Alberto Cano-Gutierrez C, Carlos Carbajal J, Cristina Carvalho De Abreu D, Casas-Herrero A, Ceriani A, Cesari M, Chiari L, Clemson L, Close J, Manuel Cornejo Alemán L, Dawson R, Delbaere K, Doody P, Dsouza S, Duan L, Duque G, Dyer S, Ellmers T, Fairhall N, Ferrucci L, Freiberger E, Frith J, Gac Espinola H, Ganz DA, Giber F, Fernando Gómez J, Miguel Gutiérrez-Robledo L, Hartikainen S, Hausdorff J, Hogan DB, Hooi Wong C, Howe S, Hunter S, Perez Jara J, Jauregui R, Jellema A, Jenni S, Jepson D, Kalula S, Kamkar N, Kaur Ajit Singh D, Anne Kenny R, Kerse N, Kobusingye O, Kressig R, Kwok W, Lamb S, Latham N, Ling Lim M, Lipsitz L, Liu-Ambrose T, Logan P, Lord S, Alves Lourenço R, Madden K, Mallet L, Marín-Larraín P, Marsh DR, Martin FC, Martínez Padilla D, Masud T, Mat S, McGarrigle L, McIlroy B, Melgar-Cuellar F, Menant J, Milisen K, Mimenza A, Moctezuma-Gallegos R, Montero-Odasso M, Morris ME, Muneeb I, Negahban H,

Nieuwboer A, Norris M, Ogliari G, Oliveira J, Parodi JF, Perez S, Perracini M, Petrovic M, Ernesto Picado Ovares J, Pieruccini-Faria F, Pighills A, Pinheiro M, Poelgeest E, Ramirez Ulate X, Robinson K, Ryg J, Said C, Sakurai R, Schapira M, Sejdic E, Seppala LJ, Sgaravatti A, Sherrington C, Skelton D, Song Y, Speechley M, Stark S, Sultana M, Suri A, Pin Tan M, Taylor M, Thomsen K, Tiedemann A, Lucia Tito S, Todd C, Troen B, Van Der Cammen T, Van Der Velde N, Verghese J, Vlaeyen E, Watt J, Welmer A-K, Won Won C, Rixt Zijlstra GA. World guidelines for falls prevention and management for older adults: a global initiative. Age and Ageing 2022 Sep 2;51(9):afac205. doi: 10.1093/ageing/afac205

17. Podsiadlo D, Richardson S. The Timed “Up & Go”: A Test of Basic Functional Mobility for Frail Elderly Persons. J American Geriatrics Society 1991 Feb;39(2):142–148. doi: 10.1111/j.1532-5415.1991.tb01616.x

18. Mahoney FI, Barthel DW. Functional Evaluation: The Barthel Index. Maryland State Medical Journal; 1965. p. 61–64. PMID:14258950

19. Folstein MF, Folstein SE, McHugh PR. “Mini-mental state.” Journal of Psychiatric Research 1975 Nov;12(3):189–198. doi: 10.1016/0022-3956(75)90026-6

20. Morse JM, Black C, Oberle K, Donahue P. A prospective study to identify the fall-prone patient. Social Science & Medicine 1989 Jan;28(1):81–86. doi: 10.1016/0277-9536(89)90309-2

21. Hendrich AL, Bender PS, Nyhuis A. Validation of the Hendrich II Fall Risk Model: A large concurrent case/control study of hospitalized patients. Applied Nursing Research 2003 Feb;16(1):9–21. doi: 10.1053/apnr.2003.016009

22. Oliver D, Britton M, Seed P, Martin FC, Hopper AH. Development and evaluation of evidence based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. BMJ 1997 Oct 25;315(7115):1049–1053. doi: 10.1136/bmj.315.7115.1049

23. Tinetti ME, Franklin Williams T, Mayewski R. Fall risk index for elderly patients based on number of chronic disabilities. The American Journal of Medicine 1986 Mar;80(3):429–434. doi: 10.1016/0002-9343(86)90717-5

24. Sun R, Sosnoff JJ. Novel sensing technology in fall risk assessment in older adults: a systematic review. BMC Geriatr 2018 Dec;18(1):14. doi: 10.1186/s12877-018-0706-6

25. Saleh M, Jeannes RLB. Elderly Fall Detection Using Wearable Sensors: A Low Cost Highly Accurate Algorithm. IEEE Sensors J 2019 Apr 15;19(8):3156–3164. doi: 10.1109/JSEN.2019.2891128

26. Greene BR, McManus K, Ader LG, Caulfield B. Unsupervised Assessment of Balance and Falls Risk Using a Smartphone and Machine Learning. Sensors 2021;21(14). doi: 10.3390/s21144770

27. Usmani S, Saboor A, Haris M, Khan MA, Park H. Latest Research Trends in Fall Detection and Prevention Using Machine Learning: A Systematic Review. Sensors (Basel) Switzerland; 2021 Jul 29;21(15). PMID:34372371

28. Chu W-M, Kristiani E, Wang Y-C, Lin Y-R, Lin S-Y, Chan W-C, Yang C-T, Tsan Y-T. A model for predicting fall risks of hospitalized elderly in Taiwan-A machine learning approach based on both electronic health records and comprehensive geriatric assessment. Front Med (Lausanne) 2022;9:937216. PMID:36016999

29. Eichler N, Raz S, Toledano-Shubi A, Livne D, Shimshoni I, Hel-Or H. Automatic and Efficient Fall Risk Assessment Based on Machine Learning. Sensors (Basel) 2022 Feb 17;22(4). PMID:35214471

30. Jahandideh S, Hutchinson AF, Bucknall TK, Considine J, Driscoll A, Manias E, Phillips NM, Rasmussen B, Vos N, Hutchinson AM. Using machine learning models to predict falls in hospitalised adults. International Journal of Medical Informatics 2024 Jul;187:105436. doi: 10.1016/j.ijmedinf.2024.105436

31. Shu F, Shu J. An eight-camera fall detection system using human fall pattern recognition via

machine learning by a low-cost android box. Sci Rep Nature Publishing Group; 2021 Jan 28;11(1):2471. doi: 10.1038/s41598-021-81115-9
32. Tago M, Katsuki NE, Oda Y, Nakatani E, Sugioka T, Yamashita S-I. New predictive models for falls among inpatients using public ADL scale in Japan: A retrospective observational study of 7,858 patients in acute care setting. PLoS One 2020;15(7):e0236130. PMID:32673366
33. Ye C, Li J, Hao S, Liu M, Jin H, Zheng L, Xia M, Jin B, Zhu C, Alfreds ST, Stearns F, Kanov L, Sylvester KG, Widen E, McElhinney D, Ling XB. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. International Journal of Medical Informatics 2020 May 1;137:104105. doi: 10.1016/j.ijmedinf.2020.104105
34. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems Curran Associates Inc.; 2024. Available from: https://dl.acm.org/doi/10.5555/3600270.3600307 [accessed Oct 23, 2024]ISBN:978-1-71387-108-8
35. Dormosh N, Van De Loo B, Heymans MW, Schut MC, Medlock S, Van Schoor NM, Van Der Velde N, Abu-Hanna A. A systematic review of fall prediction models for community-dwelling older adults: comparison between models based on research cohorts and models based on routinely collected data. Age and Ageing 2024 Jul 2;53(7):afae131. doi: 10.1093/ageing/afae131
36. Li Z, Mao F, Wu C. Can we share models if sharing data is not an option? Patterns 2022 Nov;3(11):100603. doi: 10.1016/j.patter.2022.100603
37. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics PMLR; 2017. Available from: https://proceedings.mlr.press/v54/mcmahan17a.html [accessed Oct 23, 2024]
38. Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, Sarveswara R, Händler K, Pickkers P, Aziz NA, Ktena S, Tran F, Bitzer M, Ossowski S, Casadei N, Herr C, Petersheim D, Behrends U, Kern F, Fehlmann T, Schommers P, Lehmann C, Augustin M, Rybniker J, Altmüller J, Mishra N, Bernardes JP, Krämer B, Bonaguro L, Schulte-Schrepping J, De Domenico E, Siever C, Kraut M, Desai M, Monnet B, Saridaki M, Siegel CM, Drews A, Nuesch-Germano M, Theis H, Heyckendorf J, Schreiber S, Kim-Hellmuth S, COVID-19 Aachen Study (COVAS), Balfanz P, Eggermann T, Boor P, Hausmann R, Kuhn H, Isfort S, Stingl JC, Schmalzing G, Kuhl CK, Röhrig R, Marx G, Uhlig S, Dahl E, Müller-Wieland D, Dreher M, Marx N, Nattermann J, Skowasch D, Kurth I, Keller A, Bals R, Nürnberg P, Rieß O, Rosenstiel P, Netea MG, Theis F, Mukherjee S, Backes M, Aschenbrenner AC, Ulas T, Deutsche COVID-19 Omics Initiative (DeCOI), Angelov A, Bartholomäus A, Becker A, Bezdan D, Blumert C, Bonifacio E, Bork P, Boyke B, Blum H, Clavel T, Colome-Tatche M, Cornberg M, De La Rosa Velázquez IA, Diefenbach A, Dilthey A, Fischer N, Förstner K, Franzenburg S, Frick J-S, Gabernet G, Gagneur J, Ganzenmueller T, Gauder M, Geißert J, Goesmann A, Göpel S, Grundhoff A, Grundmann H, Hain T, Hanses F, Hehr U, Heimbach A, Hoeper M, Horn F, Hübschmann D, Hummel M, Iftner T, Iftner A, Illig T, Janssen S, Kalinowski J, Kallies R, Kehr B, Keppler OT, Klein C, Knop M, Kohlbacher O, Köhrer K, Korbel J, Kremsner PG, Kühnert D, Landthaler M, Li Y, Ludwig KU, Makarewicz O, Marz M, McHardy AC, Mertes C, Münchhoff M, Nahnsen S, Nöthen M, Ntoumi F, Overmann J, Peter S, Pfeffer K, Pink I, Poetsch AR, Protzer U, Pühler A, Rajewsky N, Ralser M, Reiche K, Ripke S, Da Rocha UN, Saliba A-E, Sander LE, Sawitzki B, Scheithauer S, Schiffer P, Schmid-Burgk J, Schneider W, Schulte E-C, Sczyrba A, Sharaf ML, Singh Y, Sonnabend M, Stegle O, Stoye J, Vehreschild J, Velavan TP, Vogel J, Volland S, Von Kleist M, Walker A, Walter J, Wieczorek D, Winkler S, Ziebuhr J, Breteler MMB, Giamarellos-Bourboulis EJ, Kox M, Becker M, Cheran S, Woodacre MS, Goh EL, Schultze JL. Swarm Learning for decentralized and confidential clinical machine learning. Nature 2021 Jun 10;594(7862):265–270. doi: 10.1038/s41586-021-03583-3
39. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, Milchenko M, Xu W, Marcus D, Colen RR, Bakas S. Federated learning in medicine: facilitating multi-institutional

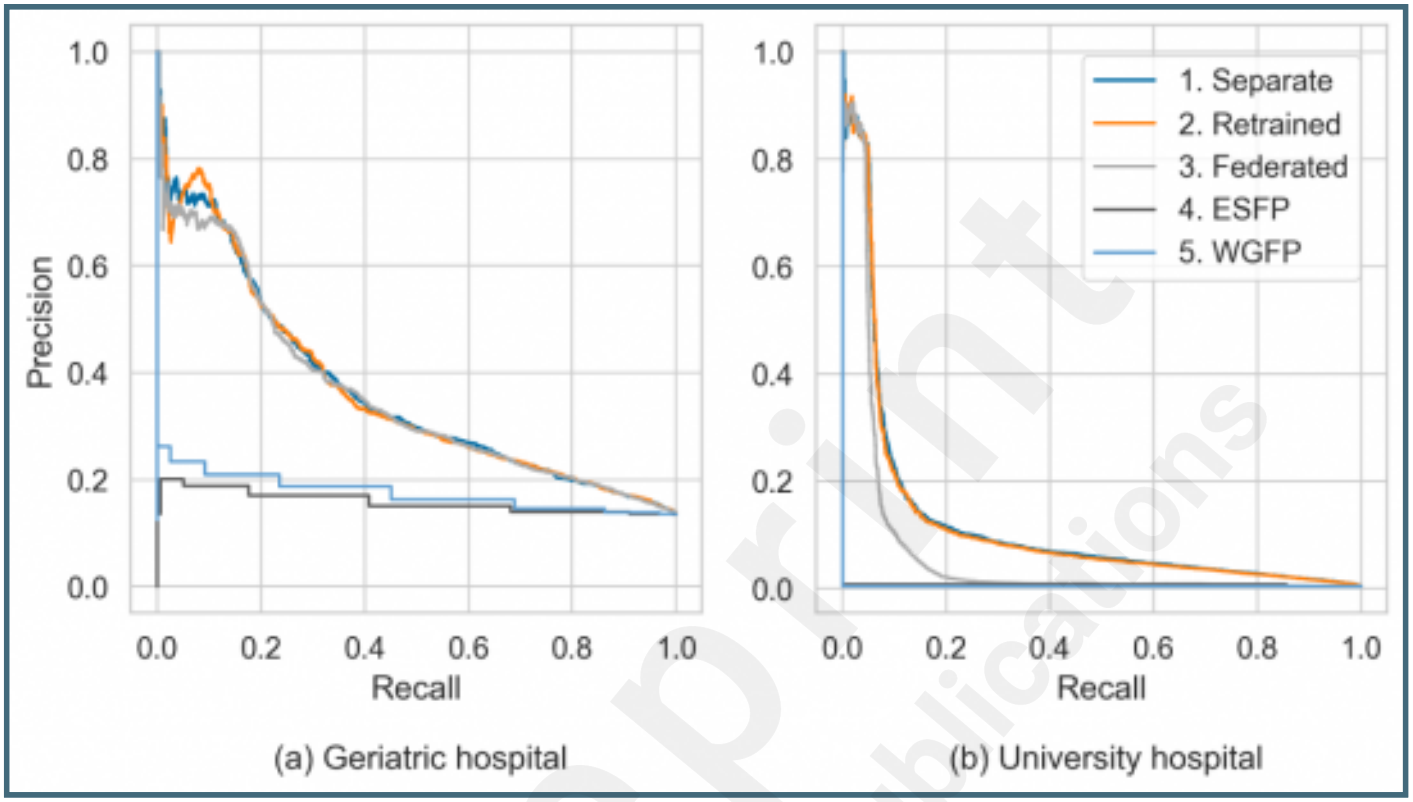
collaborations without sharing patient data. Sci Rep 2020 Jul 28;10(1):12598. doi: 10.1038/s41598-020-69250-1
40. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, Somani S, Paranjpe I, De Freitas JK, Wanyan T, Johnson KW, Bicak M, Klang E, Kwon YJ, Costa A, Zhao S, Miotto R, Charney AW, Böttinger E, Fayad ZA, Nadkarni GN, Wang F, Glicksberg BS. Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach. JMIR Med Inform 2021 Jan 27;9(1):e24207. doi: 10.2196/24207
41. Lee H, Chai YJ, Joo H, Lee K, Hwang JY, Kim S-M, Kim K, Nam I-C, Choi JY, Yu HW, Lee M-C, Masuoka H, Miyauchi A, Lee KE, Kim S, Kong H-J. Federated Learning for Thyroid Ultrasound Image Analysis to Protect Personal Information: Validation Study in a Real Health Care Environment. JMIR Med Inform 2021 May 18;9(5):e25869. doi: 10.2196/25869
42. Sarma KV, Harmon S, Sanford T, Roth HR, Xu Z, Tetreault J, Xu D, Flores MG, Raman AG, Kulkarni R, Wood BJ, Choyke PL, Priester AM, Marks LS, Raman SS, Enzmann D, Turkbey B, Speier W, Arnold CW. Federated learning improves site performance in multicenter deep learning without data sharing. Journal of the American Medical Informatics Association 2021 Jun 12;28(6):1259–1264. doi: 10.1093/jamia/ocaa341
43. Jones EW. Patient classification for long-term care : user's manual. United States Department of Health, Education, and Welfare; 1973.
44. Operationen- und Prozedurenschlüssel Version 2024. 2024. Available from: https://klassifikationen.bfarm.de/ops/kode-suche/htmlops2024/index.htm [accessed Jun 10, 2024]
45. Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme. 2024. Available from: https://klassifikationen.bfarm.de/icd-10-gm/kode-suche/htmlgm2024/index.htm
46. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res 2016 Dec 16;18(12):e323. doi: 10.2196/jmir.5870
47. Kuratorium Deutsche Altershilfe. 2024. Available from: https://kda.de/
48. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2012 Nov;12:2825–2830. doi: 10.5555/1953048.2078195
49. XGBoost Documentation. Available from: https://xgboost.readthedocs.io/en/stable/ [accessed Jun 10, 2024]
50. Flower: A Friendly Federated Learning Framework. Available from: https://flower.ai/ [accessed Jun 10, 2024]
51. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012 Mar 1;13:281–305. doi: 10.5555/2188385.2188395
52. Rehfeld S, Schulte-Althoff M, Schreiber F, Fürstenau D, Näher A-F, Hauss A, Köhler C, Balzer F. The Prediction of Fall Circumstances Among Patients in Clinical Care – A Retrospective Observational Study. Challenges of Trustable AI and Added-Value on Health IOS Press; 2022. p. 575–576. doi: 10.3233/SHTI220530
53. Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning: Limitations and Opportunities. MIT Press; 2023. Available from: fairmlbook.org
54. Verma S, Rubin J. Fairness definitions explained. Proceedings of the International Workshop on Software Fairness Gothenburg Sweden: ACM; 2018. p. 1–7. doi: 10.1145/3194770.3194776
55. GitHub - shap: A game theoretic approach to explain the output of any machine learning model. Available from: https://github.com/shap/shap [accessed Jun 10, 2024]
56. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, Ourselin S, Sheller M, Summers RM, Trask A, Xu D, Baust M,

Cardoso MJ. The future of digital health with federated learning. npj Digit Med 2020;3(1):1–7. doi: 10.1038/s41746-020-00323-1
57. Prayitno, Shyu C-R, Putra KT, Chen H-C, Tsai Y-Y, Hossain KSMT, Jiang W, Shae Z-Y. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. Applied Sciences Multidisciplinary Digital Publishing Institute; 2021 Jan;11(23):11191. doi: 10.3390/app112311191
58. Hsieh K, Phanishayee A, Mutlu O, Gibbons PB. The Non-IID Data Quagmire of Decentralized Machine Learning. ICML 2020;4387–4398. doi: 10.5555/3524938.3525346
59. Dhade P, Shirke P. Federated Learning for Healthcare: A Comprehensive Review. RAiSE-2023 MDPI; 2024. p. 230. doi: 10.3390/engproc2023059230

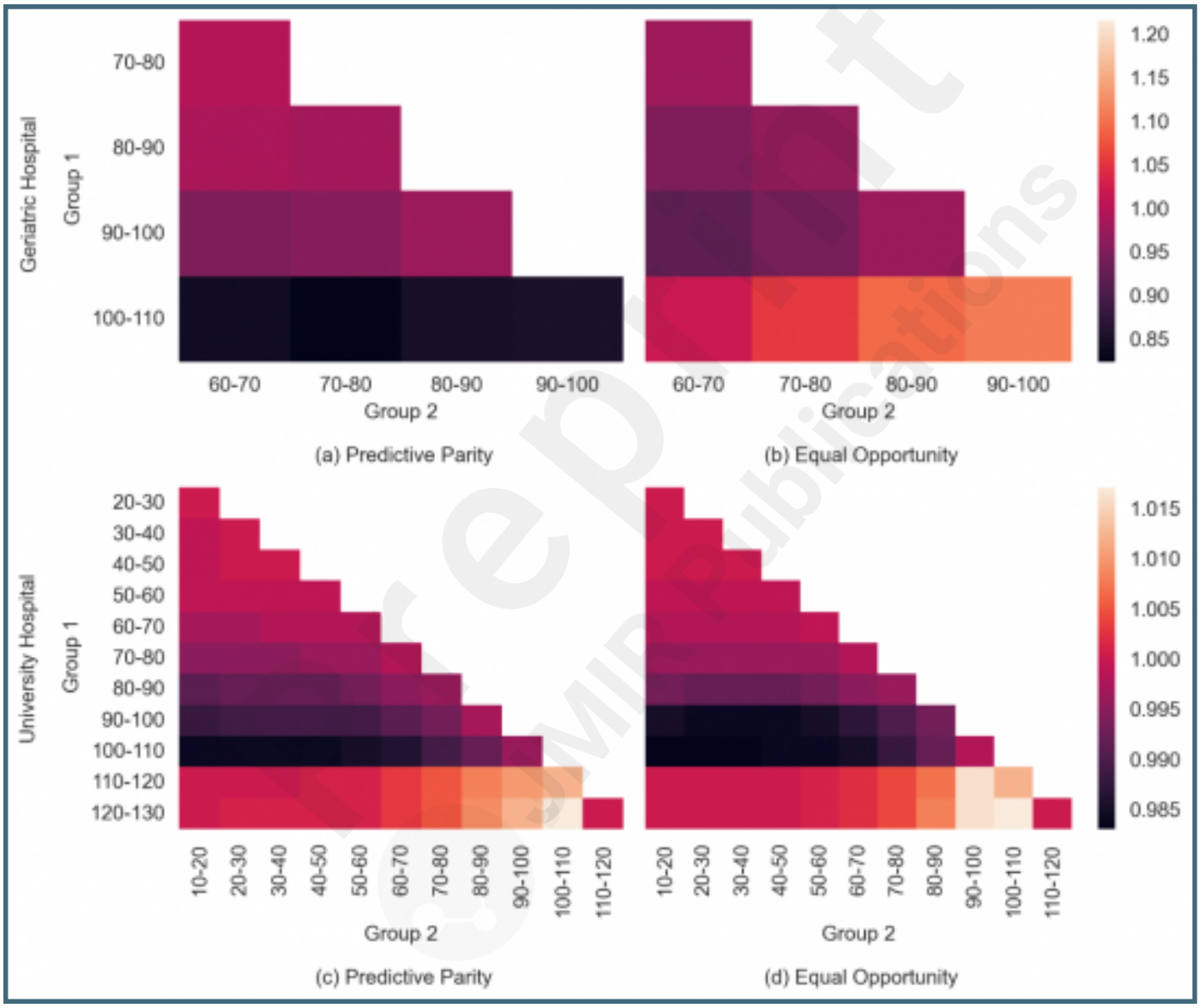
Supplementary Files

Figures

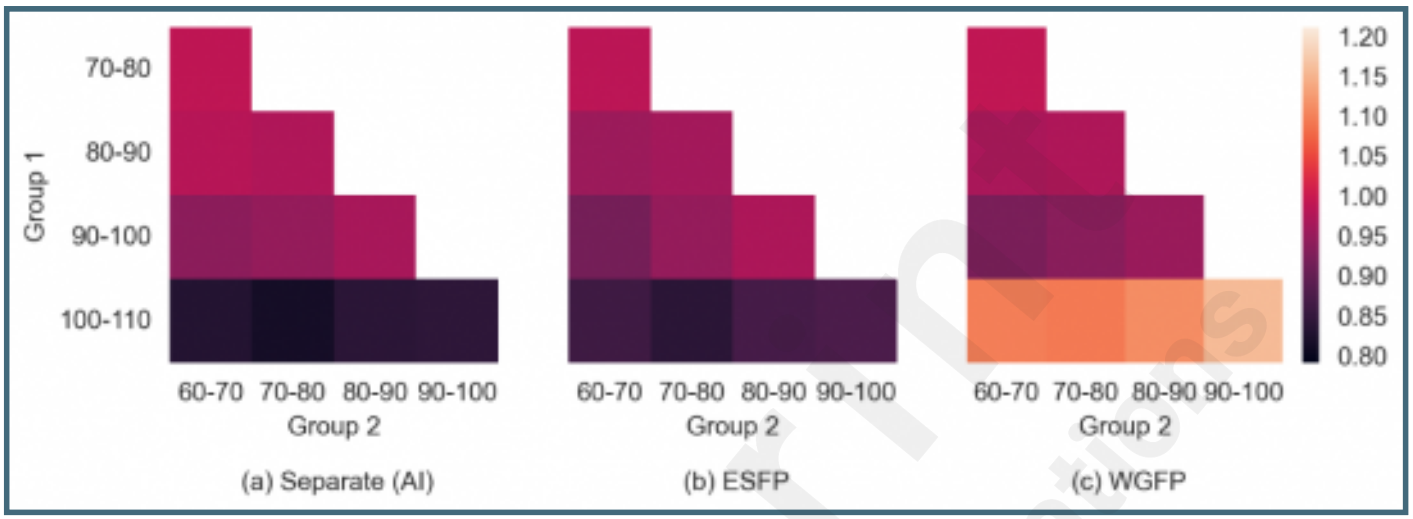
The AI based models (Separate, Retrained and Federated) substantially outperform the baseline rule-based models (ESFP, WGFP). FL (gray curves) struggles to effectively integrate knowledge from the two hospitals with considerable demographic shifts, leading to reduced performance for the university hospital.



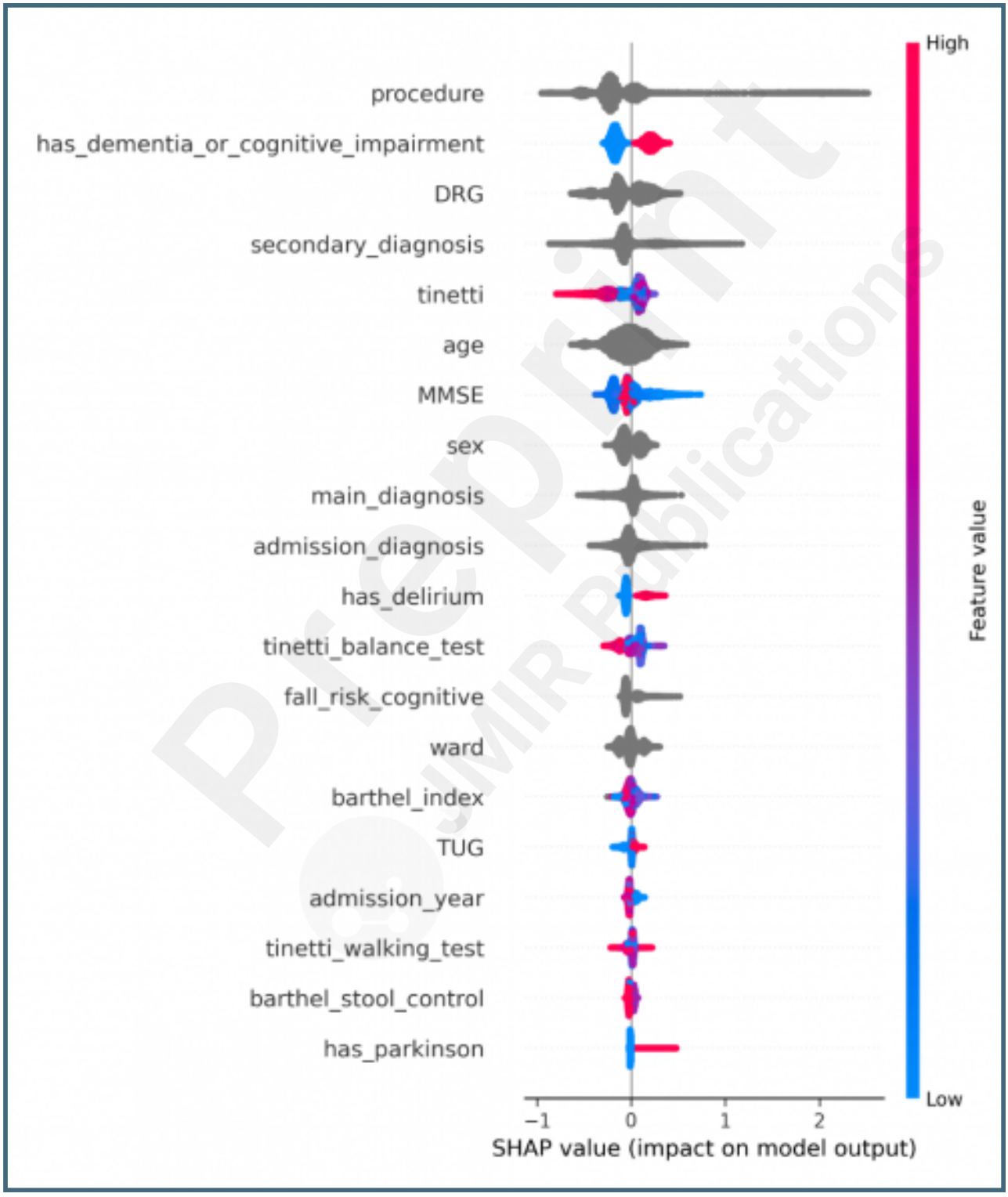
Fairness analysis indicates infringements of AI fall risk assessment models for some age groups. Models demonstrate lower precision for patients aged 100-110, indicating overestimation of fall risk, particularly in the geriatric hospital (a, c). However, this fairness disparity is less pronounced in the university hospital, where the precision score is closer to parity. In contrast, the geriatric hospital model shows reduced recall for patients aged 90-100 (b), failing to accurately identify those patients who fall, in comparison to patients aged 100-110. In the university hospital model (d), the recall score for patients aged 90-100 and 100-110 is also lower, though still near parity. The ratios represent the score of Group 1 over the score of Group 2. Darker shades indicate that Group 1 has a lower score than Group 2, a lighter shade indicates that Group 1 has a higher score than Group 2, a score of 1 indicates that both groups have equal scores.



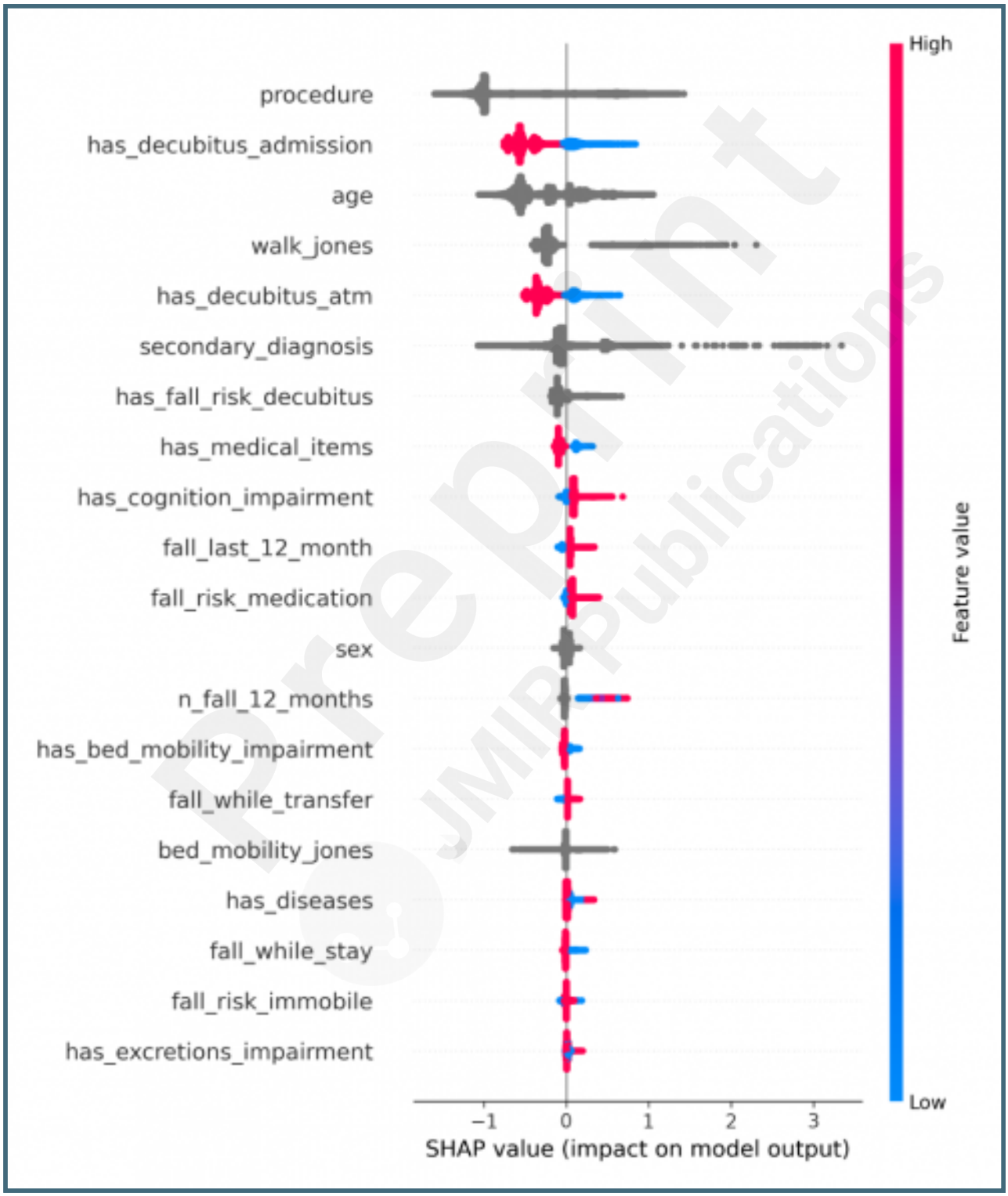
Predictive parity comparison—geriatric hospital. The AI model and the ESFP demonstrate lower precision for patients aged 100-110, indicating overestimation of fall risk, particularly the AI model (a). However, this fairness disparity is lower for the ESFP (b). In contrast, the WGFP model shows higher precision for patients aged 100-110 (c), having more false positives for the rest of the groups. The ratios represent the score of Group 1 over the score of Group 2. Darker shades indicate that Group 1 has a lower score than Group 2, a lighter shade indicates that Group 1 has a higher score than Group 2, a score of 1 indicates equal scores across both groups.



Geriatric hospital—each point along the x-axis represents a single patient and indicates its SHAP value for that feature. Red color means that the feature has higher value, and a blue color means that the feature has lower value. Categorical features are shown as gray points. The most important features include procedure, has_dementia_or_cognitive_impairment, the billable diagnosis code (DRG), secondary_diagnosis, age, the Tinetti score, and MMSE. Individuals with dementia or cognitive impairment face a higher fall risk (red points along the X axis). Patients with higher Tinetti scores are mostly at lower risk of falling.



University hospital—each point along the x-axis represents a single patient and indicates its SHAP value for that feature. Red color means that the feature has higher value, and a blue color means that the feature has lower value. Categorical features are shown as gray points. The most influential features are procedure, has_decubitus_admission, age, the jones score, has_decubitus_atm and the secondary_diagnosis. Individuals who have decubitus on admission or at the moment of the assessment are at lower risk of falling (red points along the X axis).



Multimedia Appendixes

Supplementary Material.
URL: http://asset.jmir.pub/assets/04cfeebd55bc111a72be0eccb2e94be6.docx