Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable SHAP space fall risk profiling from electronic health records for tailored inpatient prevention

Matthias Schulte-Althoff
Charité - University Medicine Berlin
Peter Krappen
Freie Universität Berlin
Felix Biessmann
Berliner Hochschule für Technik
Sebastian Jäger
Berliner Hochschule für Technik
Rahel Gubser
Charité - University Medicine Berlin
Armin Hauß
Charité - University Medicine Berlin
Felix Balzer
Charité - University Medicine Berlin
Tim Kilgus
Freie Universität Berlin
Daniel Fürstenau
daniel.fuerstenau@fu-berlin.de

Charité - University Medicine Berlin

Article

Keywords:

Posted Date: April 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9312593/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Falls are common, costly, and heterogeneous, which limits the effectiveness of one-size-fits-all prevention strategies. In a retrospective observational study using electronic health record data from a large German university hospital between 2016 and 2022, we trained an ensemble model to predict inpatient falls. We then derived patient-level SHapley Additive Explanations and clustered alerted patients in explanation space to generate interpretable risk profiles. On a held-out test set, the final model achieved an area under the receiver operating characteristic curve of 0.95 and an area under the precision-recall curve of 0.36, outperforming a baseline that used only guideline features. Explanation-space clustering identified nine clinically interpretable profiles with distinct dominant risk drivers and stronger separation of observed fall incidence than feature-space comparators. These profiles link model alerts to tailored, guideline-consistent prevention strategies. External validation and prospective implementation studies are needed before routine clinical deployment.

1. Introduction

Falls are a major patient-safety concern. Globally, they account for around 684,000 deaths each year and approximately 37.3 million events requiring medical attention, while 0.85% to 1.5% of total healthcare costs are attributable to falls [1,2]. In hospitals, falls are associated with injury, prolonged length of stay, and increased care burden. Although up to 42% of falls may be preventable through nonpharmacological interventions [3], effective prevention remains difficult because fall risk is multifactorial and heterogeneous across patients. With ageing populations, hospitals are increasingly caring for older, frailer inpatients who have high baseline fall risk and require resource-intensive prevention. At the same time, nursing teams must operationalize multifactorial recommendations under staffing and time constraints, making scalable risk stratification and actionable decision support a health-systems priority.

In routine care, fall risk is commonly assessed using manual checklist-based instruments such as the Morse Fall Scale, STRATIFY, and Hendrich II, often complemented by functional assessments such as Timed Up and Go or Tinetti [4–8]. These tools capture important person-, medication-, and mobility-related risk factors, but systematic reviews have shown limited predictive validity for distinguishing high- from low-risk patients and a tendency toward high sensitivity at the expense of specificity [9–13]. As a result, many inpatients may be flagged as high risk, which can dilute clinical attention and hinder targeted prevention. At the same time, the evidence base on fall risk has broadened to include factors such as diabetes, frailty, pain, and malnutrition [14–17], reinforcing the need for more nuanced and individualized assessment.

Recent guidelines have therefore shifted emphasis from generic screening toward personalized multifactorial prevention. The World Falls Guidelines [18] recommend that patients identified as high risk should receive a comprehensive multifactorial assessment and tailored interventions aligned with their specific combination of risk factors. Complementary national guidance, such as the German Expert Standard on Falls Prevention in Nursing, similarly emphasizes institution-specific assessment and prevention planning [19]. Operationalizing this recommendation in routine inpatient workflows remains difficult, as clinicians must make sense of interacting risks spanning mobility, cognition, prior falls, medications, comorbidity, and care environment, often under substantial time and staffing constraints [18,20].

Aiming towards a multifactorial, risk-specific assessment, recent machine-learning studies have used diverse data sources and settings for fall-risk modeling, including acute inpatient EHRs [21–24], long-term care data [25], disease-specific cohorts [26], sensor-based trajectories [27], and geriatric assessments [28]. For example, Ossai et al. [23] proposed real-time risk profiling in acute care, whereas Chu et al. [24], Thapa et al. [25], and Huang et al. [26] illustrate how multi-domain or subgroup-specific modeling can enrich fall-risk characterization. Complementary work has used sensor trajectories [27] or explainability on geriatric assessment features [28]. Together, these studies show that combining information across clinical domains can improve risk characterization beyond single-score tools, while increasingly digital clinical records enable such modeling at scale [29]. At the same time, clinically useful deployment requires more than predictive accuracy alone: model outputs must be translated into actionable representations that remain sensitive to patient-specific mechanisms of risk [30,31].

While most existing work focuses primarily on predictive performance, some approaches cluster patients in raw or latent feature space, which may reveal structure that is not necessarily tied to fall outcomes or intervention relevance [32–34]. Others combine clustering with explainability or outcome guidance to derive more clinically meaningful subgroup structure [35,36], while still often stopping short of producing reusable subgroup summaries for practice [28,33,35,43]. This gap matters because clinician-facing explanation formats can affect trust and decision behaviour [37], and because patients at similar predicted risk may still differ substantially in the mechanisms driving that risk.

This challenge is consistent with broader work on heterogeneous treatment effects: both risk drivers and preventive needs can vary substantially across patients, but estimating highly granular individual-level effects from observational electronic health record data is difficult in practice [38–42]. A pragmatic alternative is to identify subgroups of alerted patients who share similar model-attributed risk patterns and then link those patterns to guideline-consistent intervention domains. Such an approach can bridge prediction and intervention while preserving interpretability for clinical users.

In this retrospective observational study, we evaluate whether explainable electronic health record-based fall prediction can be translated into clinically meaningful subgroup-level risk profiles for inpatient fall prevention. We train machine-learning models on large-scale hospital electronic health record data, derive patient-level SHapley Additive exPlanations (SHAP), cluster alerted patients in explanation space, and assess whether these clusters separate observed fall incidence more strongly than feature-space comparators. Unlike clustering patient records independently of the prediction task, clustering local explanations identifies risk profiles linked to the outcome of interest. Moreover, unlike local explanations considered in isolation, clustering reveals recurrent structure across individual patient profiles. Accordingly, our aim is not only to identify patients at elevated risk, but also to provide interpretable profiles that support subgroup-tailored and guideline-consistent preventive action in routine care. The overall modeling idea is shown in Fig. 1.

2. Results

2.1 Predictive modeling

Table 1 presents selected evaluation metrics for representative models. A weighted ensemble model L3 achieved the highest AUPRC (0.357) on the test set, an AUROC of ~0.946, and among the highest recall, while maintaining good precision. It clearly outperformed the hospital's existing fall risk score baseline odds ratios (Supplementary Tables 1–3). The precision-recall performance was far above the no-skill baseline (prevalence ~1.2% so baseline AUPRC ~0.012).

By contrast, the simplified model that only used guideline-specified risk factors (26 features corresponding to factors like history of fractures, mobility impairment, cognitive impairment, depression, etc.) performed much worse (AUPRC ~0.05, AUROC ~0.82). This highlights that additional EHR variables carry important predictive signal that the traditional risk factors alone do not capture. For example, our full model found that certain procedure categories (like hemodialysis, or physical therapy orders) were among the top predictors of falls, whereas a guideline-based model focusing on diagnoses might disregard those entirely.

2.2 Threshold analysis for operational deployment

We performed a post-hoc threshold analysis to quantify the trade-off between sensitivity, precision, and operational workload across probability cut-offs (Fig. 2). At the default threshold of 0.50, the weighted ensemble achieved very high specificity (~99.96%) and precision (~81%), but sensitivity was low (~16%), indicating that only a small fraction of fallers would be identified under this conservative setting. This threshold might be too conservative for clinical use, as it would miss roughly 84% of true fall incidents.

If we lower the threshold to capture ~55–60% of fallers, sensitivity improves to ~57–67%, with precision dropping to ~16–21%, i.e. a threshold of 0.08 identified substantially more fallers (sensitivity ~55%) while yielding a precision of ~18%, corresponding to a number needed to evaluate of approximately 5.3 patients per true positive. From an implementation standpoint, the final model's Negative Predictive Value was extremely high (>99%) due to the rarity of falls. On the other hand, the Positive Predictive Value in the chosen operating range is more modest (on the order of 20%), but importantly, substantially higher than the ~1% pre-test probability of falls and higher than the precision of simpler screening tools. At stricter thresholds, alerts became more precise but far fewer fallers were captured; at more permissive thresholds, additional fallers would be identified at the cost of increased intervention burden.

For inpatient falls, the relative costs of false negatives and false positives are asymmetric. In inpatient falls, a false negative—failing to identify a patient who then falls—can lead to injury, extended hospitalization, and associated costs. False positives—flagging someone who wouldn't have fallen—may lead to precautionary interventions like closer monitoring, bed alarms, assistance with toileting, etc. These interventions are generally low-risk and low-cost relative to the consequences of a fall. Therefore, many hospitals choose a threshold that errs on the side of sensitivity, accepting more false-positive alerts to prevent more falls. In our context, moving from 16% sensitivity to ~67% sensitivity would require about a six-fold increase in interventions per actual fall prevented, with precision dropping from 81% to 16%. Whether this is acceptable depends on resource availability. In the remainder of this paper, threshold adjustments used in our capacity analyses were not applied, e.g. when computing SHAP explanations.

2.3 Risk profiles by patient cluster

For the L3 model chosen, we examined the aggregated local feature importances using SHAP (Supplementary Fig. 1). Based on the calculated values we did a SHAP-based K-means clustering as described in the Methods section. We identified nine distinct risk clusters among the patients the model flagged as high risk. The right panel of Fig. 3 shows an inter-cluster distance map: cluster centroids were computed in SHAP space; for visualization only, centroids were embedded into a 2D space via MDS on the SHAP matrix. All inter-centroid distances reported refer to SHAP L2-space. Clusters 2, 5, and 6 lie adjacent in SHAP-value space, indicating partially overlapping risk profiles driven by complex treatment courses; by contrast, Cluster 3 is well separated. Cluster sizes ranged from ~177 patients (Cluster 3, ~6% of true positives) to ~685 patients (Cluster 1, ~23% of true positives). The clustering approach reveals various risk pathways: complex care needs (Clusters 5, 6), rehabilitation-focused care (Clusters 4, 9), and high-monitoring medical interventions (Clusters 7, 8), with Clusters 1 and 2 representing general care-dependent populations and Cluster 3 standing alone in its mental health focus. Table 2 summarizes the main characteristics of the clusters. For each cluster, interventions suggested by the World Fall Guidelines were added where appropriate. As Cluster 1 and 2 represent general care-dependent populations, general risk-preventing measures were listed. For the rest, tailored measurements were added. The Supplementary Note provides a more detailed cluster description.

With regard to the cluster assignment of new patient, as described in the SHAP-based patient clustering subsection, our empirical analysis suggests setting the Primary Distance Threshold τ_1 to the 90th percentile of the intra-cluster distances from our training data, and the predefined ratio threshold τ_2 to 1.5. These parameters were set based on the empirical distance distribution in the cluster-discovery (held-out test) cohort and achieved successful cluster assignment for 93% of patients in the held-out test set. For unassigned patients that do not meet the assignment criteria, the top three candidate clusters can be displayed by the system and considered based on clinical judgment.

2.4 Outcome separation benchmarking of clustering approaches

Across clustering methods, SHAP-space K-means showed the largest separation in observed fall incidence across clusters, according to our prespecified summary statistics—incidence range (max–min), incidence ratio (max/min), and IQR across clusters (Table 3). Specifically, incidence varied widely across its risk profiles (range 0.424, ratio 5.12, IQR 0.155), indicating both substantially higher- and lower-risk subgroups within the alerted population. In contrast, feature-space baselines—raw K-means/GMM/PCA-K-means/Ward/Spectral—yielded markedly narrower risk gradients (range 0.089–0.207, IQR 0.016–0.050), indicating more weakly differentiated outcome strata. Moreover, SHAP-space clusters showed the best alignment between clusters' mean predicted risk and observed incidence (Spearman 0.943, cluster-level calibration MAE 0.014), whereas alignment was substantially weaker for most feature-space baselines (Spearman 0.257–0.714, MAE 0.019–0.023). Taken together, these findings demonstrate that clustering in explanation space not only yields

interpretable risk profiles but also produces more outcome-relevant stratification among patients who would be classified as high risk.

3. Discussion

Using a large-scale EHR from 2016–2022 with 932,102 patients, our weighted ensemble model achieved AUROC ≈0.95 and AUPRC ≈0.36 on a held-out test set, substantially outperforming a guideline-features-only model (AUROC ≈0.82; AUPRC ≈0.05). Outcome-guided clustering of SHAP values on the test set yielded nine interpretable risk profiles. Benchmarking against feature-space and score-based comparators showed that SHAP-space clustering yields the largest between-cluster differences in observed fall incidence among alerted patients, while also maintaining strong alignment between cluster-level predicted risk and observed outcomes. A simple SHAP-space distance-mapping assigned 93% of test patients confidently to one of these clusters. Threshold analysis indicated feasible operating points. e.g., threshold ~0.08 and a sensitivity of ~0.55, PPV ~0.20, NPV >99%) that would allow units to tune alerts to resources. Together, the pipeline supports guideline-concordant, multifactorial, and subgroup-tailored prevention in routine care.

The resulting stacked ensemble model is competitive with or exceeds prior falls prediction models in the literature. For instance, one study by Hsu et al. used an XGBoost classifier on electronic health record (EHR) data to predict falls during hospital admission and found an Area Under the Curve (AUC) of 0.700 (95% CI: 0.614, 0.777) in a cohort of 639 adult patients (297 fall patients and 342 controls) [21]. Mishra et al. developed a ML model using gait and geriatric assessment data from 92 geriatric patients to predict 6-month fall risk and achieved an AUC of 0.80 (95% CI: 0.76, 0.85) [28]. Ye et al. used statewide health records from Maine (n = 265,225) to develop an XGBoost model predicting one-year fall risk, achieving a C-statistic (equivalent to the ROC-AUC) of 0.851 in the training set and 0.803 in the validation set [47]. Thapa et al. [25] achieved AUROC ~0.84 in long-term care settings. Complementing these “fall occurrence” models, Wang et al. developed an integrative multi-view ensemble model to predict injurious/severe inpatient falls among fall incidents, achieving AUC 0.808 (95% CI: 0.740–0.876) on a prospective test set, and AUC 0.86 in a high-risk subset—illustrating how machine learning can augment bedside screening by prioritizing patients most likely to experience harmful outcomes [22].

Fall prevention has a low tolerance for false negatives, i.e. missed fall-risk patients, which is why existing tools cast a wide net. Our threshold analysis explicitly allows tuning this balance. For example, if a unit is overwhelmed with alerts, they can raise the threshold modestly. Furthermore, by examining trade-offs, we inform discussions with hospital leadership. This kind of data-driven planning is not possible with legacy checklist scores (e.g., Morse, STRATIFY, Hendrich II) that typically yield point totals/categories rather than calibrated probabilities and do not support unit-specific probability cut-offs [4–6,9]. As a result, units cannot explicitly tune alert thresholds to local capacity, e.g., how many alerts must be evaluated to prevent one fall. Probabilistic ML models can provide such risk estimates, but to be actionable they must also expose why an alert fired and what factors to target. Accordingly, explainable ML methods such as SHAP can help translate model outputs into clinician-interpretable drivers of risk.

The black-box nature of ML models often complicates their interpretation in clinical settings. Therefore, the role of XAI methods, such as SHAP, is critical to build trust among medical professionals [48]. Detailed

exploration of different patient risk profiles provides invaluable insights for individualized and subgroup-specific interventions [21,28]. The emphasis on personalized fall risk assessment is consistent with several scientific findings [49–51] and the World Falls Guidelines assertion of the efficacy of tailored interventions [18]. Instead of a black-box risk score, the SHAP value approach allows for listing patient-specific factors. We found that some factors are consistently important, such as age and fall history, while others vary widely between individuals, e.g., mental health with younger female patients affected. This highlights the need for subgroup-specific strategies. Additionally, the model trained using only guideline factors performed poorly. Although the SHAP plot showed that well-known fall risk factors were among the most important, one may conclude that many factors should be considered when evaluating heterogeneous risk profiles, such as those associated with fall events.

A major impetus for this work was to operationalize the recommendations of the World Falls Guidelines in a hospital setting. Montero-Odasso et al. [18] call for multifactorial falls risk assessments and personalized interventions for those at high risk. In current practice, when a patient is labeled high-risk, generic precautions are applied, but personalized assessments are often not done due to time constraints. In intended use, the proposed model is a decision-support component that (i) flags elevated fall risk and (ii) assigns patients to an interpretable risk profile to guide selection of guideline-consistent prevention measures. Thus, a first-pass multifactorial assessment can prospectively be performed automatically: dozens of factors are being analyzed, and it is indicated which ones are driving a patient's risk via SHAP analysis. Then, by categorizing the patient's risk profile via clustering, clinicians can match the patient to an "intervention template" aligned with guideline-recommended domains.

This study has several limitations. It was retrospective and conducted at a single hospital, so documentation practices, case mix, and model performance may differ across institutions. We did not include all potentially relevant data streams, such as medication administration records or real-time physiological measures, which may further improve prediction. In addition, the risk profiles were derived by clustering SHAP patterns on a held-out cohort after model training. Although the prediction model can generate risk scores and patient-level explanations immediately, prospective profile assignment requires either an initial profile-discovery cohort or a burn-in period to establish stable centroids and assignment thresholds. Finally, intervention relevance was evaluated theoretically through guideline mapping rather than through a prospective cluster-guided implementation study.

Despite these limitations, the findings suggest that explainable EHR-based fall-risk profiling can translate prediction into clinically interpretable and potentially actionable prevention pathways. By linking alerts to dominant risk drivers and reusable subgroup-level intervention templates, the approach supports a more targeted form of multifactorial fall prevention than broad checklist-based screening alone. External validation and prospective implementation studies should now test transportability, workflow effects, and whether profile-guided prevention reduces falls or improves resource allocation in routine care.

4. Methods

4.1 Data sources and coding

We conducted a retrospective observational study using pseudonymized EHR data from a large German university hospital. Inpatient admissions from 2016 to 2022 were considered. To enrich the data for fall events, we included all patients who experienced one or more falls during the period (as documented in incident reports), as well as a sample of patients who did not fall; inclusion among all patients was limited to adults (age ≥18) and to ensure a broad representation across hospital services, non-fallers were drawn from the 20 departments with the highest fall counts (e.g., oncology, geriatrics, neurology, neurosurgery, internal medicine, psychiatry; excluding intensive care units)[1]. The final dataset comprised 932,102 unique patients, of whom 10,443 (1.12%) had at least one recorded fall during their hospital stay. This resulted in a highly imbalanced dataset (Table 4). We address class imbalance through appropriate choice of evaluation metrics, such as AUPRC and F1. For each patient, we extracted demographics (age, sex), diagnoses, procedures, nursing care data[2], which encompasses, among others, mobility, gait, medication, and fall risk assessment scores documented by nursing staff at admission and every 7 days thereafter (the hospital uses a three-level scale: “0” = no risk, “+” = some risk, “++” = high risk). Sex was extracted from the hospital administrative record as a binary recorded variable (female/male). No gender identity variable was available in the source systems.

Diagnosis data included ~12.18 million entries coded in ICD-10-GM (the German modification of the International Classification of Diseases, 10th Revision). Procedure data included ~5.72 million entries coded in OPS (Operationen- und Prozedurenschlüssel, the German procedure coding system). Both coding systems are hierarchical, allowing aggregation into broader categories when needed. Based on clinical consultation, we applied a data cleaning step: diagnoses marked with low certainty (about 4.1% of diagnosis entries) were removed to avoid spurious information. Missing age data for a small fraction of non-fallers (<0.05%) were imputed with the median age (57 years), with no age data missing for fallers. Nursing care data was absent in 76% of all patients, primarily because routine nursing documentation was not initiated for patients who were screened at admission as sufficiently independent/fit and therefore did not require a formal nursing care assessment within the nursing care module. We also derived from hospital records whether each patient had any documented falls prior to the first inpatient fall (for fallers) or during their hospital stays (for non-fallers), and we recorded the initial nursing risk assessment score for falls at admission. These serve as baseline comparisons for our model’s predictive performance. To prevent data leakage, only data available before a fall incident or before discharge for non-fallers was used via the clinic’s timestamps.

Fallers were significantly older (Cohen's d = 0.956) and had more complex medical histories. In our data, the five most common diagnosis categories among fallers were (1) family history of health risk factors, (2) ischemic heart disease, (3) metabolic disorders (e.g., diabetes), (4) hypertension, and (5) musculoskeletal disorders (see Supplementary Tables 4–5). Non-fallers shared many of the top diagnoses (reflecting overall hospital case mix), but fallers had higher rates of cardiovascular and metabolic issues, likely due to their higher age. On the procedure side, fallers most frequently underwent imaging, circulatory system procedures, e.g., catheterizations, and oncologic treatments, such as radiation or chemotherapy. Two procedure categories notably more prevalent in fallers were *computed tomography* (72% vs. 16%) and *patient monitoring* (56% vs. 12%).

2.2 Predictive modeling

We formulated fall prediction as a binary classification problem: fall vs. no-fall during the hospital stay. Features were constructed from the patient's data available up to the time of fall for fallers, or up to the end of the hospital stay for non-fallers (see Fig. 4). In practice, for patients who fell, we included all diagnoses and procedures recorded before their first fall event to avoid labeling consequences of the fall as causes. Guideline-specified fall risk factors included: fall during current stay; walking impairment; balance/gait impairment; use of a walking aid; gait assessment; fall within the last 12 months; number of falls in the last 12 months; transfer impairment; transfer assistance; cognitive impairment; disorientation (time/place/own person); agitation; confusion; bed-mobility impairment; bed-mobility assessment; skin condition relevant to mobility; presence of medical items (e.g., infusion, catheter); impairment of excretions; urinary/faecal incontinence; nocturia; use of psychotropic or sedative medications; decubitus; and fall during transfer. For patients who did not fall, we included data from their entire hospitalization. Each patient thus has a feature vector summarizing their demographics, diagnoses, procedures, and nursing care data prior to fall or prior to discharge for non-fallers. Categorical features (ICD-10-GM and OPS) were one-hot encoded using codes on the group level. A higher number of digits indicates a more specific diagnosis, e.g., "J45" indicates asthma, while "J45.0" indicates predominantly allergic asthma. Like the ICD, OPS codes have a hierarchical structure, e.g., "8-98" indicates various complex treatments, while "8-981" indicates a neurological complex treatment. The first two characters of the codes were used for grouping.

We leveraged an AutoML library (AutoGluon) to explore a suite of models—logistic regression, multilayer perceptron (neural network), random forest, gradient-boosted trees (XGBoost and LightGBM), and stacked ensembles. We performed a patient-wise random split into a training set used for cross-validation and model selection (80%; n=745,682) and a final held-out test set for all reported performance and threshold analyses (20%; n=186,420). Model selection and hyperparameter tuning were performed using 5-fold cross-validation on the training set, with the area under the precision–recall curve (AUPRC) as the primary metric for optimization. AUPRC is well-suited for imbalanced data because it focuses on performance in the minority class and is less dominated by the abundance of negative cases. It also evaluates model performance across all possible prediction thresholds, aligning with our interest in how different threshold choices affect precision and recall. For comparability, we additionally report macro-F1 and operating-point metrics (sensitivity/recall, specificity, precision/PPV) at selected probability thresholds.

To address severe class imbalance, we optimized model selection using AUPRC and evaluated all final results on the held-out test set with the natural outcome distribution. We avoided using oversampling techniques, such as SMOTE, during training to keep the training data as representative as possible. To improve the understanding of the model predictions and to investigate the underlying fall risk factors in detail, we used SHAP, a game-theoretic technique that can interpret model outcomes [43]. The resulting values, known as SHAP values, indicate the mean deviation from the mean prediction when a conditioning on a single feature, thereby revealing the impact of individual features on model outcomes. By relating the magnitude of features to their relevance in the model predictions, SHAP values can provide insight into different risk factors and latent interactions at the instance level [44]. For the implementation, we used FastSHAP [45] which has become a popular, more efficient variant of the original implementation.

2.3 Threshold analysis for model deployment

The envisioned use of our model is as a clinical decision support tool that runs in parallel with standard care processes. During patient admission and evaluation, the model would generate a fall risk prediction (with a probability score). If the predicted risk is above a chosen threshold (see Threshold Analysis below), the patient is flagged as high-risk for falls. At that point, the system provides two levels of explanation: (1) an individual SHAP explanation highlighting the patient's top risk-increasing and risk-decreasing factors, and (2) a cluster assignment suggestion indicating if the patient matches one of the pre-identified risk clusters. Fig. 1 illustrates this integration conceptually. In practice, a clinician reviewing the alert would see something like: "High fall risk predicted. Key factors: age 85 (+), recent surgery (+), use of sedative medications (+), independent mobility (-). Patient fits Cluster 2— 'Frail patients with rehabilitation needs'." The clinician can then confirm the relevance of this profile and initiate a tailored care plan accordingly. Each cluster is linked to a standardized set of recommended interventions.

A practical aspect is choosing the probability threshold at which to classify a patient as "high risk." In training, we optimized for overall AUPRC, considering all possible thresholds simultaneously. Depending on hospital priorities, one might prefer a sensitive or a more specific model. For deployment, the model outputs a fall-risk probability score per admission that can be converted into an alert via a probability threshold. We therefore conducted a post-hoc threshold analysis on the held-out test set using candidate thresholds and reporting sensitivity, specificity, precision (PPV), and the number needed to evaluate (NNE = 1/PPV). The PPV can characterize the workload–benefit trade-off for clinic administrators and nursing staff.

2.4 SHAP-Based patient clustering

Our next step was to identify clusters of high-risk patients with similar risk-factor patterns in their local explanations, rather than clustering raw features. In this representation, each patient is described by a vector of feature-attribution values, where larger positive values indicate stronger contributions toward the model predicting a fall. Clustering in SHAP space groups patients by the model-attributed drivers of risk (e.g., mobility, cognition, medications), rather than by feature similarity alone, yields subgroup-level "why" profiles that are globally interpretable, align directly with guideline-based multifactorial assessment, and can be translated into tailored intervention bundles for each cluster. To ensure an unbiased summary, we clustered only the held-out test-set predictions and their SHAP vectors, thus preventing information leakage from training/validation into subgroup discovery. We constructed a vector of SHAP values for each patient using the final model: $Sp = (s_1, s_2, \ldots, s_m)$, where m is the number of features. Here s_i is the SHAP value for feature i, representing how much that feature pushed the model's prediction for patient p above or below the baseline fall probability.

Patients who have similar SHAP value profiles can be considered to have similar risk factor contributions, even if their raw feature values differ. For example, two patients may differ in absolute age or diagnosis codes, yet both may be 'high-risk' primarily due to prior falls and sedating medication exposure, leading to similar SHAP patterns. We applied K-means clustering to the matrix of SHAP value vectors for patients who were predicted by the model to be at risk of falling. To determine an appropriate number of clusters k, we used the elbow method on the clustering distortion—the within-cluster sum of squared distances. As shown in Fig. 3, the elbow plot suggested an optimal clustering around $k = 9$ clusters for our data. Increasing beyond

9 yielded diminishing returns (clusters became too small or too similar), while fewer than 9 clusters would merge distinct risk profiles and reduce granularity. We prioritized a solution that maximized within-cluster similarity of SHAP profiles, because our goal is to define groups for which very specific, tailored interventions can be designed. After clustering, we characterized each cluster both externally by patient attributes and in terms of their respective risk factors.

We evaluated whether explanation-space clustering yields risk profiles that are (i) coherent in terms of risk drivers and (ii) clinically meaningful in terms of fall incidence separation among patients who would trigger an alert. We benchmarked SHAP-space K-means against representative baselines operating on (a) raw feature space (K-means, Gaussian mixture model, Ward agglomerative, spectral clustering), (b) reduced feature representations (PCA(95%) + K-means), and (c) score-only stratification (predicted-risk quantile bins).

To quantify clinical stratification among alerted patients, we defined a held-out high-risk evaluation cohort as test-set admissions that would trigger an operational alert at a prespecified probability threshold ($p \geq 0.08$). For each method, we computed cluster-level observed fall incidence with 95% Wilson confidence intervals and summarized outcome separation by (i) incidence range (max–min), (ii) incidence ratio (max/min), and (iii) IQR of incidence across clusters. We refer to a method as having stronger outcome separation when it yields larger values on these stratification summaries, reflecting greater dispersion of observed incidence across clusters in the alerted cohort. We additionally assessed risk–outcome alignment using Spearman correlation between cluster mean predicted risk and observed incidence, and cluster-level calibration error (MAE). For method-level summaries, clusters contributing <1% of the evaluation cohort were collapsed.

In a prospective deployment, new inpatients (not seen during training) predicted as high risk must be assignable to one of the discovered risk clusters to surface the most relevant, subgroup-specific risk drivers. This enables tailored interventions aligned with guideline-recommended multifactorial care. We propose a simple distance-based method to assign new patients to clusters in the SHAP space.

To support prospective use, we map newly alerted patients to the discovered SHAP-based risk profiles by nearest-centroid assignment in SHAP space via k-means prediction. To avoid unstable labels, we apply two confidence checks: (i) the patient must lie within a distance threshold defined by the empirical distribution of within-cluster distances (e.g., 90th percentile), and (ii) the closest cluster must be sufficiently closer than the second-closest. Patients failing either check are marked unassigned.

To operationalize phenotype assignment for newly alerted patients, we use nearest-centroid assignment in SHAP space (k-means prediction) combined with a reject option for low-confidence matches. For a newly alerted patient with SHAP vector S_p and cluster centroids μ_k, we compute Euclidean distances $d(S_p, \mu_k) = \|S_p - \mu_k\|_2$. We assign the patient to the nearest centroid $k^* = \text{argmin}_k d(p,k)$ only if two criteria are met: (i) *Primary Distance Threshold* (**absolute proximity)** $d(p,k^*) \leq \tau_1$, where τ_1 is set to the 90th percentile of within-cluster distances estimated on the risk-profile-cohort; and (ii) *Ratio Threshold (***relative uniqueness)** $d(p,k^*)/d(p,k^{(2)}) \leq \tau^2$, where $k^{(2)}$ is the second-closest centroid and τ_2 is a prespecified ratio threshold (1.5 in our experiments). Patients failing either criterion are marked “unassigned”, and the top candidate risk profiles may be displayed to support clinical judgement. The overall modeling pipeline is sketched in Fig. 5.

2.5 Ethics approval and consent to participate

This retrospective study was approved by the Charité Ethics Commission (Approval No. EA2/184/21). Original approval dated 20 August 2021; amendment approved on 2 December 2024 (amendment submitted on 20 November 2024). For this retrospective study, informed consent was waived by the ethics board. All procedures were performed in accordance with the Declaration of Helsinki and its later amendments (or comparable ethical standards). For retrospective analyses of routinely collected data, informed consent was waived by the ethics committee where applicable.

Declarations

Ethics approval and consent to participate

This retrospective study was approved by the Charité Ethics Commission (Approval No. EA2/184/21). Original approval dated 20 August 2021; amendment approved on 2 December 2024 (amendment submitted on 20 November 2024). For this retrospective study, informed consent was waived by the ethics board. All procedures were performed in accordance with the Declaration of Helsinki and its later amendments (or comparable ethical standards). For retrospective analyses of routinely collected data, informed consent was waived by the ethics committee where applicable.

Competing interests

The authors declare that they have no competing interests.

Author Contribution

Matthias Schulte-Althoff: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, writing—original draft.Peter Krappen: Conceptualization, data curation, formal analysis, methodology, writing—original draft, writing—review and editing.Felix Biessmann: Conceptualization, supervision, project administration, writing—review and editing.Sebastian Jaeger: Supervision, writing—review and editing.Rahel Gubser: Conceptualization, writing—review and editing.Armin Hauss: Investigation, validation, resources.Felix Balzer: Supervision, writing—review and editing.Tim Kilgus: Formal analysis, writing—review and editing.Daniel Fuerstenau: Conceptualization, funding acquisition, data curation, methodology, project administration, writing—review and editing.

Acknowledgements

This work was part of the KIP-SDM project funded by the Federal Ministry for Research, Technology and Space (BMFTR), Grant No. 16SV8854.

Data Availability

The patient-level electronic health record data analysed in this study are not publicly available due to privacy, ethical, and legal restrictions. De-identified data may be made available from the corresponding author on reasonable request and subject to institutional approval, applicable data-protection regulations, and a data-use agreement that prohibits re-identification and onward sharing.

Code availability

Custom code central to the reported findings will be made available to editors and referees upon request during peer review and to readers upon publication. At submission, code is available from the corresponding author upon reasonable request.

References

1. Chen, Y. et al. Global, regional, and national burden of falls among older adults: findings from the Global Burden of Disease Study 2021 and projections to 2040. npj Aging 11, 85 (2025).
2. Heinrich, S., Rapp, K., Rissmann, U., Becker, C. & König, H. H. Cost of falls in old age: a systematic review. Osteoporos. Int. 21, 891-902 (2010).
3. Hshieh, T. T., Yang, T., Gartaganis, S. L., Yue, J. & Inouye, S. K. Hospital Elder Life Program: systematic review and meta-analysis of effectiveness. Am. J. Geriatr. Psychiatry 26, 1015-1033 (2018).
4. Morse, J. M., Morse, R. M. & Tylko, S. J. Development of a scale to identify the fall-prone patient. Can. J. Aging 8, 366-377 (1989).
5. Oliver, D., Britton, M., Seed, P., Martin, F. C. & Hopper, A. H. Development and evaluation of evidence-based risk assessment tool (STRATIFY) to predict which elderly inpatients will fall: case-control and cohort studies. BMJ 315, 1049-1053 (1997).
6. Gray-Miceli, D. & Greenberg, S. Fall risk assessment for older adults: the Hendrich II fall risk model. Am. J. Nurs. 107, 50-58 (2007).
7. Podsiadlo, D. & Richardson, S. The Timed Up and Go: a test of basic functional mobility for frail elderly persons. J. Am. Geriatr. Soc. 39, 142-148 (1991).
8. Tinetti, M. E. Performance-oriented assessment of mobility problems in elderly patients. J. Am. Geriatr. Soc. 34, 119-126 (1986).
9. Park, S. H. Tools for assessing fall risk in the elderly: a systematic review and meta-analysis. Aging Clin. Exp. Res. 30, 1-16 (2018).
10. Ivziku, D., Matarese, M. & Pedone, C. Predictive validity of the Hendrich Fall Risk Model II in an acute geriatric unit. Int. J. Nurs. Stud. 48, 468-474 (2011).
11. Cho, E. H. et al. Comparison of the predictive validity of three fall risk assessment tools and analysis of fall-risk factors at a tertiary teaching hospital. J. Clin. Nurs. 29, 3482-3493 (2020).
12. Matarese, M., Ivziku, D., Bartolozzi, F., Piredda, M. & De Marinis, M. G. Systematic review of fall risk screening tools for older patients in acute hospitals. J. Adv. Nurs. 71, 1198-1209 (2015).
13. Turner, K. et al. Alarm with care: a de-implementation strategy to reduce fall prevention alarm use in US hospitals: a study protocol for a hybrid 2 effectiveness-implementation trial. Implement. Sci. 18, 70

(2023).
14. Yang, Y., Hu, X., Zhang, Q. & Zou, R. Diabetes mellitus and risk of falls in older adults: a systematic review and meta-analysis. Age Ageing 45, 761-767 (2016).
15. Kojima, G. Frailty as a predictor of future falls among community-dwelling older people: a systematic review and meta-analysis. J. Am. Med. Dir. Assoc. 16, 1027-1033 (2015).
16. Lan, X., Li, H., Wang, Z. & Chen, Y. Frailty as a predictor of future falls in hospitalized patients: a systematic review and meta-analysis. Geriatr. Nurs. 41, 69-74 (2020).
17. Trevisan, C. et al. Nutritional status, body mass index, and the risk of falls in community-dwelling older adults: a systematic review and meta-analysis. J. Am. Med. Dir. Assoc. 20, 569-582.e7 (2019).
18. Montero-Odasso, M. et al. World guidelines for falls prevention and management for older adults: a global initiative. Age Ageing 51, afac205 (2022).
19. Deutsches Netzwerk für Qualitätsentwicklung in der Pflege. Expertenstandard Sturzprophylaxe in der Pflege: 2. Aktualisierung 2022 (DNQP, 2022).
20. Morris, M. E. et al. Interventions to reduce falls in hospitals: a systematic review and meta-analysis. Age Ageing 51, afac077 (2022).
21. Hsu, Y.-C. et al. Prediction of fall events during admission using eXtreme Gradient Boosting: a comparative validation study. Sci. Rep. 10, 16777 (2020).
22. Wang, L. et al. Preventing inpatient falls with injuries using integrative machine learning prediction: a cohort study. npj Digit. Med. 2, 127 (2019).
23. Ossai, C. I., O'Connor, L. & Wickramasinghe, N. Real-time inpatients risk profiling in acute care: a comparative study of falls and pressure injuries vulnerabilities. In Proceedings of the 33rd Bled eConference (2020).
24. Chu, W.-M. et al. 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. 9, 937216 (2022).
25. Thapa, R. et al. Predicting falls in long-term care facilities: machine learning study. JMIR Aging 5, e35373 (2022).
26. Huang, X. et al. A multi-view classification framework for falls prediction: multiple-domain assessments in Parkinson's disease. In Proceedings of the 54th Hawaii International Conference on System Sciences 1-9 (2021).
27. Yu, S., Chai, Y., Samtani, S., Liu, H. & Chen, H. Motion sensor-based fall prevention for senior care: a Hidden Markov Model with generative adversarial network approach. Inf. Syst. Res. 35, 1-15 (2024).
28. Mishra, A. K. et al. Explainable fall risk prediction in older adults using gait and geriatric assessments. Front. Digit. Health 4, 869812 (2022).
29. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. Big data in healthcare: using analytics to identify and manage high-risk and high-cost patients. Health Aff. 33, 1123-1131 (2014).
30. Garriga, R. et al. Machine learning model to predict mental health crises from electronic health records. Nat. Med. 28, 1240-1248 (2022).

31. Shmueli, G. & Koppius, O. R. Predictive analytics in information systems research. MIS Q. 35, 553-572 (2011).
32. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci. Rep. 6, 26094 (2016).
33. Loftus, T. J. et al. Phenotype clustering in health care: a narrative review for clinicians. Front. Artif. Intell. 5, 842306 (2022).
34. Lee, C., Rashbass, J. & van der Schaar, M. Outcome-oriented deep temporal phenotyping of disease progression. IEEE Trans. Biomed. Eng. 68, 2423-2434 (2021).
35. Li, Y. et al. Towards interpretability and personalization: a predictive framework for clinical time-series analysis. In 2021 IEEE International Conference on Data Mining 340-349 (IEEE, 2021).
36. Carr, O. et al. Longitudinal patient stratification of electronic health records with flexible adjustment for clinical outcomes. In Proceedings of Machine Learning for Health 220-238 (PMLR, 2021).
37. Hur, S., Lee, Y. & Park, J. Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour. npj Digit. Med. 8, 578 (2025).
38. Mithas, S., Chen, Y., Lim, Y. & Silveira, A. D. O. On causality and plausibility of treatment effects in operations management research. Prod. Oper. Manag. 31, 4558-4571 (2022).
39. Kent, D. M., Steyerberg, E. & van Klaveren, D. Personalized evidence-based medicine: predictive approaches to heterogeneous treatment effects. BMJ 363, k4245 (2018).
40. Dahabreh, I. J., Hayward, R. & Kent, D. M. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. Int. J. Epidemiol. 45, 2184-2193 (2016).
41. Holland, P. W. Statistics and causal inference. J. Am. Stat. Assoc. 81, 945-960 (1986).
42. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66, 688-701 (1974).
43. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 4765-4774 (2017).
44. Chen, H., Covert, I. C., Lundberg, S. M. & Lee, S.-I. Algorithms to estimate Shapley value feature attributions. Nat. Mach. Intell. 5, 590-601 (2023).
45. Jethani, N., Sudarshan, M., Covert, I., Lee, S.-I. & Ranganath, R. FastSHAP: real-time Shapley value estimation. In International Conference on Learning Representations (2022).
46. Kruskal, J. B. & Wish, M. Multidimensional Scaling (Sage, 1978).
47. Ye, C. et al. Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. Int. J. Med. Inform. 137, 104105 (2020).
48. Zhang, Y., Weng, Y. & Lund, J. Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics 12, 237 (2022).
49. Radecki, B., Reynolds, S. & Kara, A. Inpatient fall prevention from the patient's perspective: a qualitative study. Appl. Nurs. Res. 43, 114-119 (2018).
50. Spano-Szekely, L. et al. Individualized fall prevention program in an acute care setting: an evidence-based practice improvement. J. Nurs. Care Qual. 34, 127-132 (2019).

51. LeLaurin, J. H. & Shorr, R. I. Preventing falls in hospitalized patients: state of the science. Clin. Geriatr. Med. 35, 273-283 (2019).

Footnotes

1. Our evaluation of the university hospital's fall risk model demonstrated varying levels of accuracy, precision, and recall, as detailed in Supplementary Tables 1–3. Based on logistic regression, patients labeled "+" (increased fall risk) are over four times (OR 4.17) more likely to fall, while those with "++" (strongly increased fall risk) have over six times (OR 6.31) the risk compared to those marked "0".
2. In some cases, diagnoses may not have been coded until the end of the hospital stay. In these cases, the diagnosis date is too late and is sometimes disregarded for machine learning. However, procedures necessary for billing purposes were subject to more precise dating.

Tables

Table 1 | Model performance metrics on the test set. AUPRC, AUROC and F1 (macro) metrics for various prediction models to predict falls based on patient history. The XGBoost: Fall Guidelines model has been trained only on features that are considered in the World Falls Guidelines and a common German fall standard [18,19].

ML Metric	Logistic Regression	Neural Net for Classification	RandomForest Gini BAG L2	Weighted Ensemble_L3	LightGBMXT BAG L2	XGBoost: Fall Guidelines
F1 (macro)	0.5053	0.6084	**0.6345**	0.6313	0.6318	0.5057
AUPRC	0.1079	0.1639	0.3343	**0.3568**	0.3449	0.0486
AUROC	0.9044	0.8802	0.9373	0.9463	**0.9472**	0.8225

Table 2 | Cluster descriptions, including mean principal diagnoses and procedures based on SHAP values. Relative cluster size is calculated as the number of patients assigned to the cluster divided by the number of all patients correctly classified as fallers (true positives). The final column suggests interventions targeted to the cluster’s risk profile based upon the World Falls Guidelines by Montero-Odasso et al. [18].

Cluster	*Size*	*Age*	*Sex (m)*	*Salient features based on SHAP values*	*Guideline aligned intervention domains [18]*
C1: *General Care*	685 (.23)	79.0	0.52	9-98 to 9-99: Supplementary measures and information (2.63) Falls in the last 12 months (1.61) I10 to I15: Hypertension (0.50) Z80 to Z99: Health risks based on family or self-history (0.36)	*Multidomain intervention based on general patient education and awareness (e.g., fear or concerns of falling, fall history etc. yoga, Pilates—p.23 / p.24) Consider individual patient preferences and values. Measures of physicality; Considerations on activities in daily life (ADL); consideration on environmental context (e.g., getting familiar with surroundings; adequate lightning; low bed); general health state of patient (hearing aid; walking impairment, Vitamin D supplements—p. 23 etc.);Medication review* (especially psychoactives, antihypertensives); *Continence care* (to avoid urgent unsafe toileting); *Cognitive support* and general mental state (e.g., depression, morbidity).
C2: *General Complex Care*	297 (.10)	61.2	0.54	Falls in the last 12 months (2.43) 8-97 to 8-98: Complex care (2.15) 8-80 to 8-85: Circulatory measures (0.59) 3-20 to 3-26: Computed tomography CT (0.48)	
C3: *Behavioral abnormality*	177 (.06)	64.2	0.31	9-60 to 9-64: Mental, psychosomatic and behavioral disorders (2.99) 9-98 to 9-99: Other complementary measures (1.06) Z00 to Z13: Seeking health care for examination and clarification (0.55)	*Nutritional assessment* (adequate vitamin D intake; serum 25 OH vitamin D levels), *assessment of substance abuse and excessive alcohol intake*—p. 18; *Physical exercise* (in community settings; Pilates, Yoga) p.23.; *Behavioural therapy* (for mild cognitive impairment/ dementia: group exercise. cognitive behavioural therapy and Occupational therapy)—p.12

				I10 to I15: Hypertension (0.32)	
C4: ***Frail & Rehab-Focused***	307 (.10)	87.3	0.46	Falls in the last 12 months (2.18) 8-55 to 8-60: Early rehabilitation and physical therapy (1.93) Walking and transfer impairments (1.84) I10 to I15: Hypertension (0.50)	*Balance and strength training (for stroke: balance, strength, walking) p.12; Multimodal patient education (e.g., face-to-face, handout, video) p.24. Physical training (for balance to people with complex phase PD) p.13. Medication optimization (to maximize motor function and minimize side effects—dyskinesia and hypotension) p. 25. Nutrition & vitamin D* and serum 25 OH supplementation (address malnutrition, bone health).—p.18; *Backward Chaining: Re-learning the skill to rise from the floor through practice of each of the specific movements required where the last step in the chain is taught first.—p.19; Fracture Management incl. Osteoporosis treatment—p.18*
C5: ***Complex Treatments*** **(ICU/Neuro)**	293 (.10)	77.8	0.57	8-97 to 8-98: Complex treatment (2.21) Falls in the last 12 months (1.50) I10 to I15: Hypertension (0.54) Z80 to Z99: Health risks based on family or self-history (0.39)	*Considerations on activities in daily life (ADL);"cognitive stimulation, daily orientation, early mobilisation, vision and hearing, fluid management, constipation management, feeding assistance, sleep and family involvement."—p.16; Physical Measures (progressive resistance strength training e.g., TaiChi); Medication optimization (avoid weak opioids, "as the adverse events risk may outweigh the benefit. For neuropathic pain, first line treatment includes serotonin norepinephrine reuptake inhibitors, gabapentinoids and Transdermallidocaineorcapsaicin"—p. 22); Modification of orthostatic blood pressure—p.21; Nutrition & vitamin D* and serum 25 OH supplementation (address malnutrition, bone health)—p.18; *involvement* and training of informal (unpaid) and/or formal (paid) caregivers.—p.19
C6: ***Complex Treatments + Metabolic***	383 (.13)	81.8	0.64	8-97 to 8-98: Complex treatment (2.99) Walking and transfer impairments (1.52) I10 to I15: Hypertension (0.60) E70 to E90: Metabolic	

				disorders (0.43)	
C7: ***Radiation and Chemo-therapy & Circulatory***	268 (.09)	64.9	0.50	8-80 to 8-85: Measures for blood circulation (1.40) 8-52 to 8-54: Radiation therapy and chemotherapy (0.85) 6-00 to 6-00: Application of medications (0.72) Z00 to Z13: Seeking health care for examination and clarification (0.71)	*Medication optimization (avoid weak opioids, "as the adverse events risk may outweigh the benefit. For neuropathic pain, first line treatment includes serotonin norepinephrine reuptake inhibitors, gabapentinoids and Transdermallidocaineorcapsaicin"—p. 22); Pain Management (defining cause, type (nociceptive, neuropathic) and intensity with pain rating scale for elderly patients)—p.17. Non-pharmacological approaches (physiotherapy and cognitive behavioural therapy)—p.22*
C8: ***Patient Monitoring (Cardiac/Neuro)***	312 (.10)	84.5	0.60	Falls in the last 12 months (1.93) 8-92 to 8-93: Patient monitoring (0.70) I10 to I15: Hypertension (0.43) Z80 to Z99: Health risks based on family or self-history (0.31)	*Balance and strength training (for stroke: balance, strength, walking) p.12; Multimodal patient education (e.g., face-to-face, handout, video) p.24.; Hypotension* (management of orthostatic hypotension incl. modification of possible culprit medications, rehydration, compression garments and medications (e.g., fludrocortisone and midodrine)—p.12; p.21; *ADL* (general physical activity; promotion healthy lifestyle)—p.19; *Medication Review "*for bradycardic disorders (sinus node disease, atrioventricular conduction disorders, vasovagal syndrome and carotid sinus syndrome) and *tachyarrhythmias* (atrial fibrillation, supraventricular and ventricular tachycardia) include modification of culprit medications, specific anti-arrhythmic medication and, in some cases, implantable devices (such as pacemakers and implantable cardioverter-defibrillators); *Syncope Management* (as per local syncope guidelines)."—p.21
C9: ***Rehab + Speech/Hearing***	294 (.10)	74.7	0.64	8-55 to 8-60: Early rehabilitation	*Positioning "Particle repositioning (Epley) manoeuvres are an effective treatment for benign paroxysmal position and vertigo*

and physical therapy (1.70) Walking and transfer impairments (1.39) 9-31 to 9-32: Phoniatric and pediatric audiology therapy (0.88) 8-97 to 8-98: Complex care (0.61)	*and may reduce falls rates, but evidence is limited. Vestibular rehabilitation therapy improves postural and gait stability in cases of vestibular dysfunction although the optimal approach is still unclear as is the effect on falls rates"—p.22*

Table 3 | Outcome separation benchmarking of clustering approaches in a high-risk evaluation cohort (p≥0.08).

Method	$k_{effective}$	**Observed incidence range across clusters (*max−min*)**	**Observed incidence ratio (*max/min*)**	**IQR of observed incidence**	**Spearman Correlation (*risk_cluster*, *obs_rate_cluster*)**	**Cluster calibration MAE**
SHAP-space K-means	6	**0.4239**	**5.1183**	**0.1550**	**0.9429**	**0.0141**
Raw K-means	8	0.2050	2.5983	0.0175	0.3810	0.0231
Raw GMM	8	0.2069	2.6179	0.0163	0.3333	0.0233
PCA(95%)K-means	6	0.2045	2.6170	0.0263	0.2571	0.0229
Raw Ward	6	0.0888	1.9060	0.0496	0.7143	0.0188
Raw Spectral	8	0.1462	2.6187	0.0364	0.3095	0.0230
Risk-score bins	2*	0.6705	6.2659	0.3353	NA*	0.0182

Note: Metrics summarize the distribution of observed fall incidence across clusters for each method. Clusters contributing <1% of the evaluation cohort were collapsed for reporting; thus $k_{effective}$ may be < 9. Full cluster-level outcomes are provided in Supplementary Table 6. Lower is better for Calibration MAE; higher is better for Spearman correlation between cluster mean predicted risk and cluster observed fall incidence.
*Collapsed to 2 bins due to tied/low-variance predicted probabilities; to be treated as a stratification comparator rather than a risk-profile discovery method

Table 4 | Patient attributes and faller vs. non-faller comparison from 2016 to 2022. Values are presented as mean ± standard deviation or count (percentage).

Characteristic	All Patients (n = 932,102, 100%)	Non-Faller (n = 921,658, 98.8%)	Faller (n = 10,443, 1.12%)
Recorded sex male	435,324 (46.71%)	429,578 (46.62%)	5,746 (55.02%)
Recorded sex female	496,588 (53.29%)	491,891 (53.38%)	4,697 (44.98%)
Age (years)	57.44 ± 19.76	57.23 ± 19.70	76.03 ± 15.72
Diagnoses / patient	4.82 ± 6.73	4.67 ± 6.48	17.96 ± 12.46
Procedures / patient	1.96 ± 3.46	1.88 ± 3.34	8.90 ± 6.12

Figures

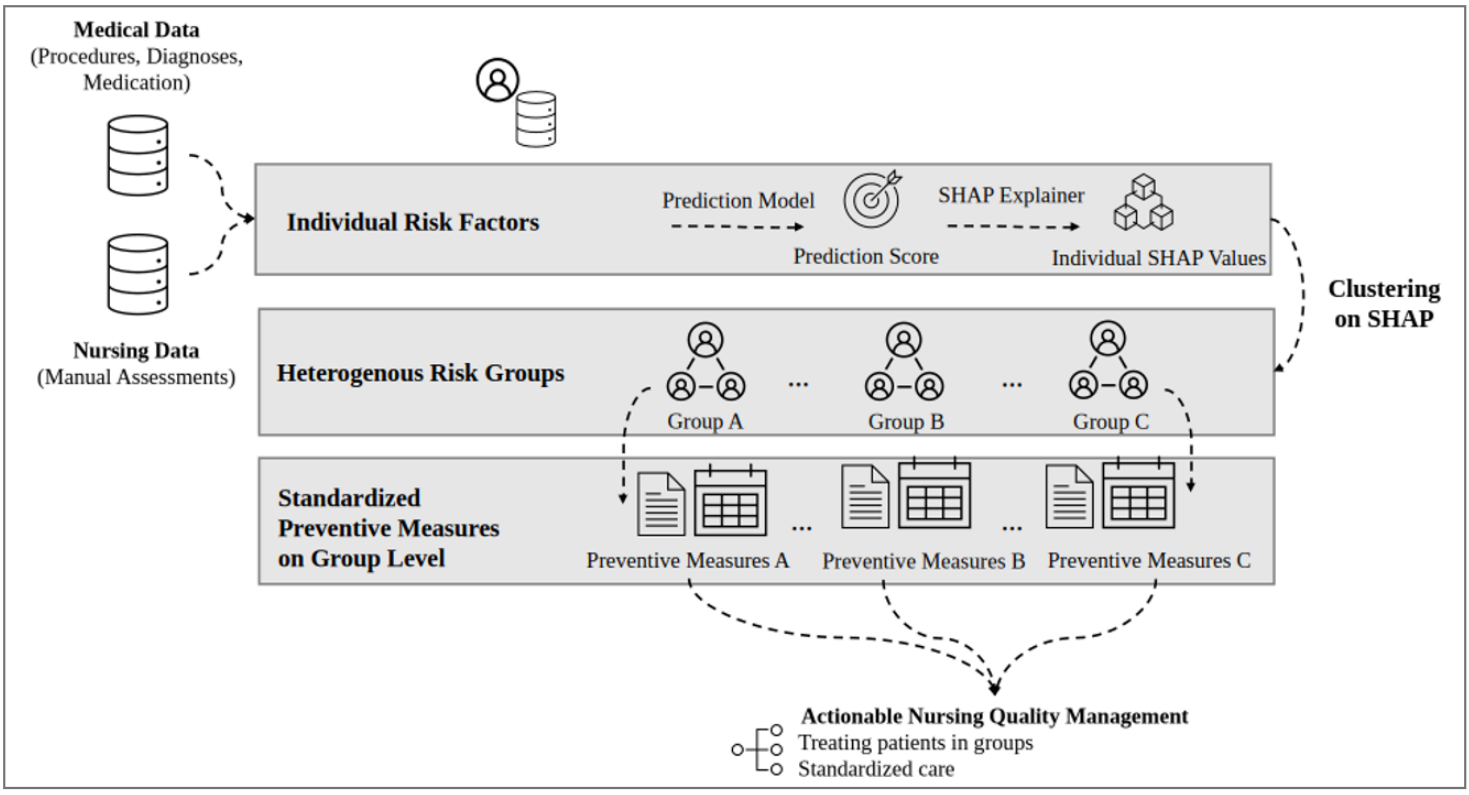


Figure 1

From individual risk factors to subgroup-tailored inpatient fall prevention. Individual electronic health record risk factors are transformed into patient-level fall-risk predictions and corresponding SHAP explanations. Clustering in explanation space groups high-risk patients with similar model-attributed risk patterns into

clinically interpretable subgroups. These subgroup profiles can then be mapped to standardized prevention pathways, enabling more targeted inpatient fall-prevention planning than risk scoring alone.

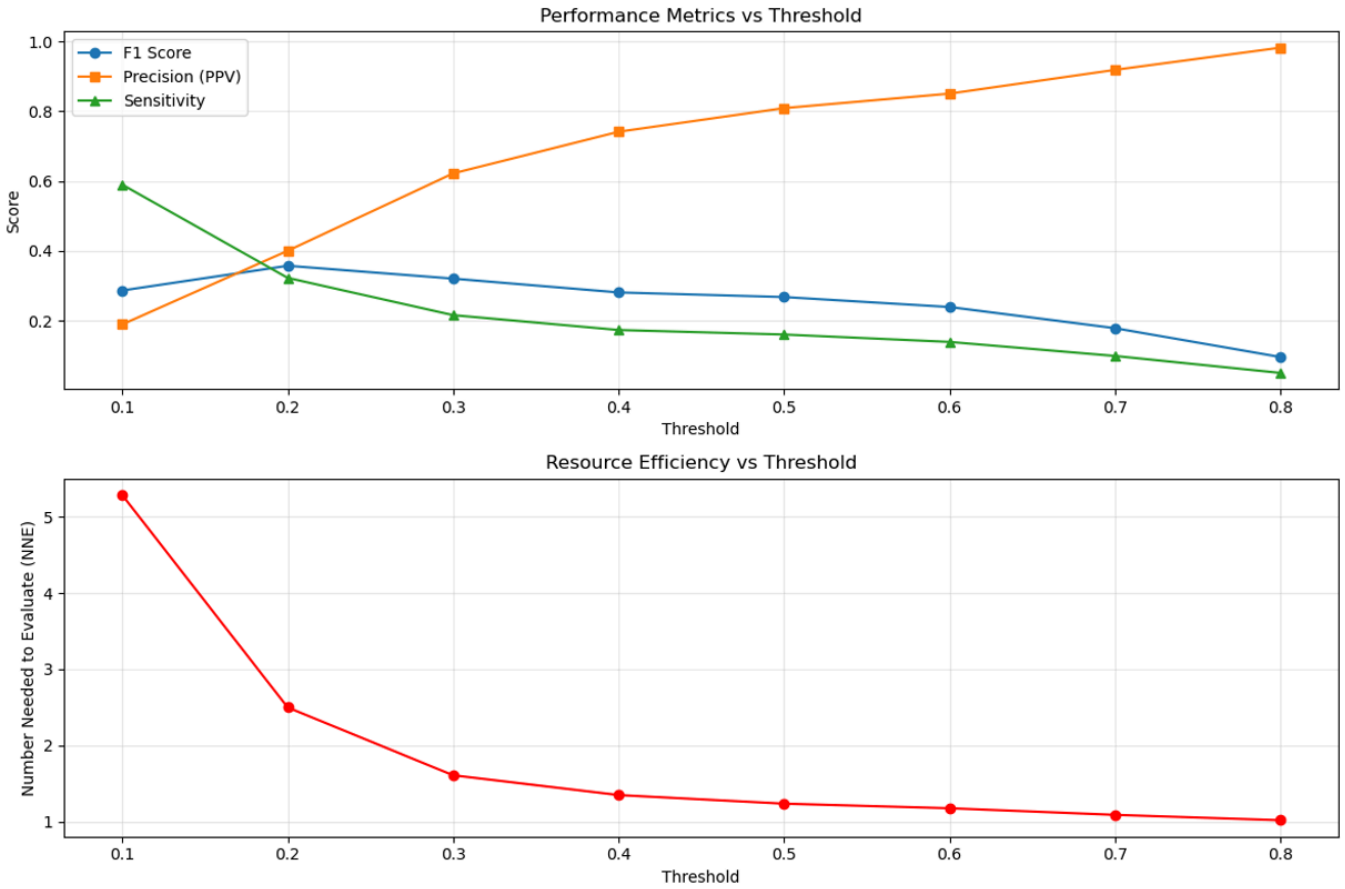


Figure 2

Threshold analysis for operational deployment. *The top figure shows performance metrics at different prediction thresholds. Higher thresholds (right side) yield fewer alerts with high precision but miss more falls (low sensitivity). Lowering the threshold (left side) increases sensitivity (catching more true fall cases) at the expense of more false positives (lower precision). The chosen operating point should balance the cost of missed falls vs. the effort of unnecessary interventions. The bottom figure shows the Number Needed to Evaluate NNE = 1/PPV—that is the number of patients who need to be evaluated to find one true positive.*

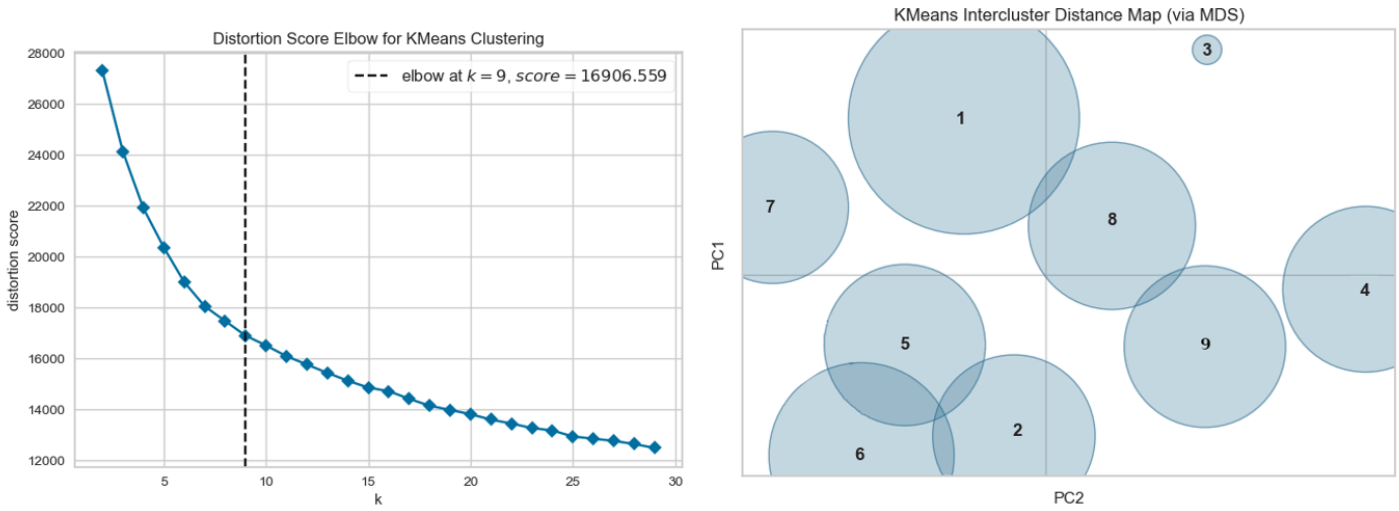


Figure 3

K-elbow plot to determine the number of clusters with an optimum of k=9 and (2) intercluster distance map with all features reduced to a two-dimensional vector space using multidimensional scaling (MDS). *MDS is a distance-preserving technique that places high-dimensional points in 2D while approximately maintaining their pairwise dissimilarities [46].*

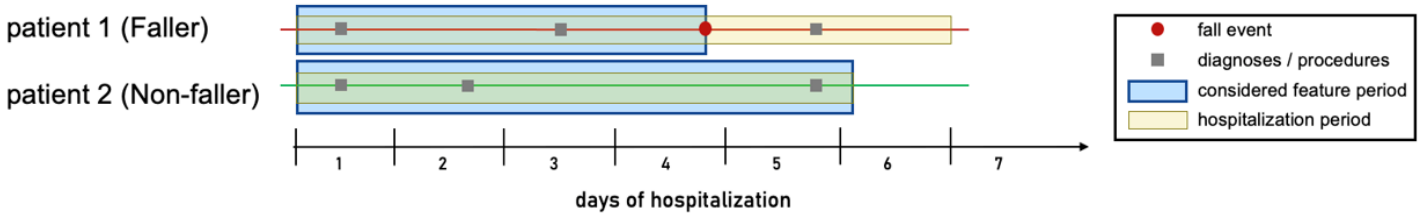


Figure 4

Construction of patient-specific feature windows during hospitalization. *For fallers, only diagnoses and procedures recorded before the first fall event were included in the predictor window to avoid leakage from post-event information. For non-fallers, features were accumulated until discharge. The schematic illustrates the different temporal cutoffs used to construct model inputs.*

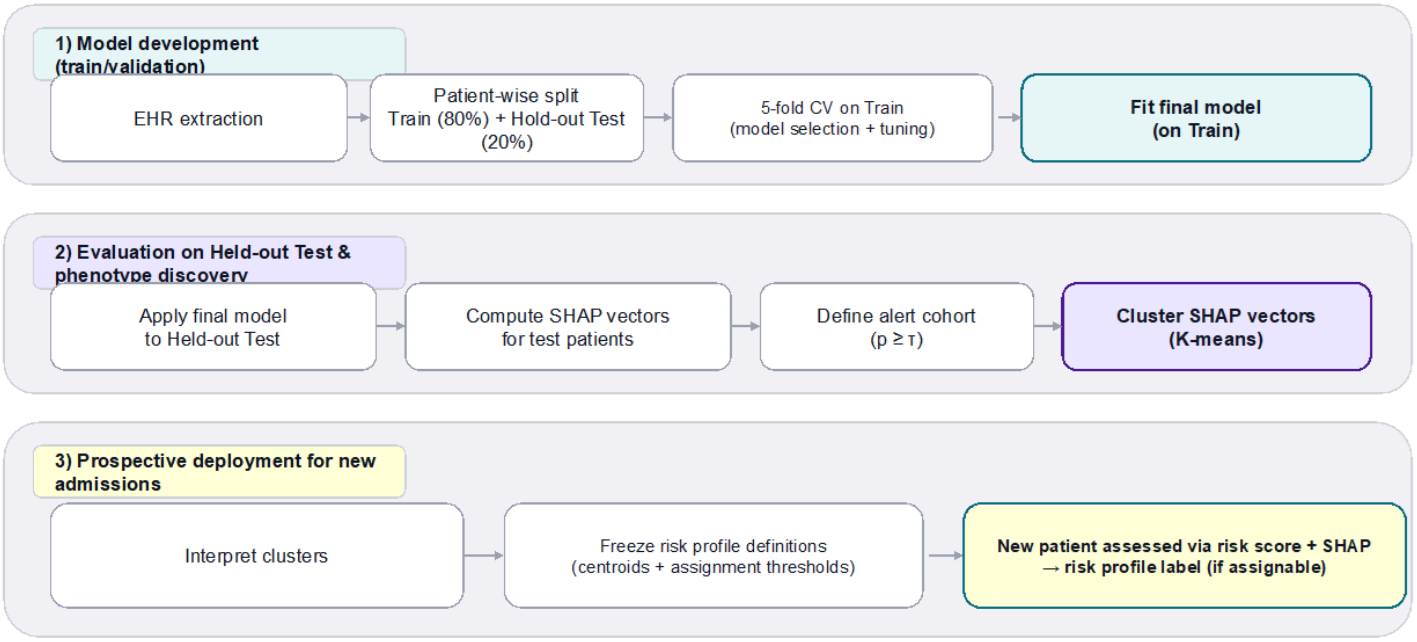


Figure 5

Workflow for model development, phenotype discovery, and prospective deployment. *The final prediction model is trained on the development cohort and applied to the held-out test cohort, where SHAP vectors are computed and clustered within the alert cohort ($p \geq \tau$) to define risk profiles. Cluster centroids and assignment thresholds are then frozen and used prospectively to map newly alerted patients to a risk profile when assignable.*

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationNPJHealthSys.docx