

ORIGINAL RESEARCH

Open access

Explainable Machine Learning for Early Prediction of Unscheduled Hospital Readmissions: Integrating Global and Local Interpretability

Nguyen Thanh Huy^{1*}, Pham Quang Minh¹, Le Thi Bich², Tran Van Nam³

Abstract

At present, the majority of healthcare spending is driven by chronic patients who experience multiple readmissions due to their underlying conditions. Implementing individualized prevention frameworks could optimize the clinical management of this demographic. This study aimed to generate localized predictive models utilizing interpretable machine learning approaches to provide early identification of individual unscheduled hospital readmissions. To achieve this, a retrospective case-control investigation was performed, utilizing patient readmission data spanning 2018–2019. Following the cleaning and curation of the initial data repository ($n = 76,210$), the final sample size comprised $n = 29,026$ participants. A machine learning analysis evaluated multiple algorithms, employing unscheduled hospital readmission as the dependent variable, alongside local model-agnostic interpretability techniques.

The results demonstrated a 13% rate of unscheduled hospital readmissions. Statistically significant variations were identified regarding patient age and length of hospital stay ($p < 0.001$ in both instances). A logistic regression framework highlighted chronic therapy (odds ratio: 3.75), a history of diabetes mellitus (odds ratio: 1.14), and length of stay (odds ratio: 1.02) as statistically meaningful predictors. Machine learning algorithms demonstrated superior performance concerning sensitivity and complementary evaluation metrics. Within this analytical framework, length of stay and patient age emerged as the primary determinants for forecasting unscheduled readmissions. Notably, other clinical indicators, such as documented allergies and histories of adverse drug reactions, also demonstrated predictive relevance. Individualized prognostic models exhibited high sensitivity. In conclusion, this investigation identified critical determinants governing unscheduled hospital readmissions, highlighting the prominence of age and hospitalization duration. A personalized risk assessment framework capable of predicting readmissions with high accuracy was established. Future investigations should incorporate a broader array of clinical parameters to further optimize the predictive capacity of this model.

Keywords Machine learning, Hospital readmission, Chronic patient, Prediction, Shap-values

*Correspondence:

Nguyen Thanh Huy

huy.nguyen@vnua.edu.vn

¹ Department of Hospital Readmission Prediction and XAI, Faculty of Medicine, Vietnam National University of Agriculture, Hanoi, Vietnam

² Department of Global and Local Interpretability, Faculty of Engineering, Can Tho University of Medicine and Pharmacy, Can Tho, Vietnam

³ Department of Early Prediction and Clinical AI, Faculty of Medicine, Hue University, Hue, Vietnam

Introduction

Patients presenting with elevated comorbidity profiles frequently encounter a higher incidence of therapeutic complications, extended hospitalizations, increased surgical liabilities, predictable readmission patterns, and substantial utilization of healthcare resources and clinical services [1]. Consequently, forecasting these types of clinical complications could substantially optimize the delivery of healthcare services. Currently, the vast quantity of data collected during inpatient admissions offers an opportunity to address these challenges. Unfortunately, clinical departments—most notably emergency and surgical units—frequently experience overcrowding, which degrades care quality due to excessive clinical demand. To optimize resource management within these units, establishing a system to predict which patients will require repeated utilization of these services would be highly advantageous.

To date, numerous diagnostic instruments have been engineered to evaluate the probability of hospital readmission, among which the Charlson Comorbidity Index [2, 3] remains the most comprehensively evaluated. In accordance with the methodology outlined by the National Health System for the evaluation of inpatient care, an unexpected hospital readmission (UHR) is defined as an unplanned admission via the emergency department following a prior discharge from the same facility [4].

Readmissions are shaped by patient comorbidities and the specific morbidities managed during hospitalization. These events can serve as indicators of clinical stability during the progression of a disease, where readmission might be triggered by post-discharge complications that reflect substandard patient follow-up. Conversely, readmissions may also relate to immediate clinical stability, wherein premature hospital discharge precipitates a subsequent return to the facility [5]. Consequently, the Minimum Basic Data Set Register categorizes readmissions into two distinct intervals: those occurring within seven days of discharge, and those taking place between the second week and the 30th post-discharge day, which represents the focal point of this investigation. Given that readmissions within the first week are often tied to deficient acute clinical care, the subsequent period involves a more complex array of contributing factors.

In recent years, traditional data analysis has faced major obstacles stemming from inadequate statistical power and a shortage of advanced computational methods, which has constrained the thorough evaluation of extensive datasets and precise predictive modeling [6]. However, the emergence of modern computing technologies facilitates the implementation of diverse algorithms and machine learning methodologies supported by Artificial Intelligence, enabling more efficient and scalable processing of complex information [7]. This computational progression not only increases the analytical depth but also underpins the generation of more precise predictive frameworks through advanced machine learning workflows. The medical sector has long endeavored to integrate machine learning workflows into data management and clinical decision-making protocols [8]. This domain is undergoing rapid evolution due to machine learning, which facilitates the prediction, management, and diagnosis of pathological conditions with increased accuracy and dependability [9, 10].

On the basis of these considerations, the primary objective of this investigation was to evaluate the efficacy and performance of multiple machine learning algorithms in forecasting unexpected hospital readmissions within a University Clinic Hospital. Furthermore, this study sought to delineate the clinical profiles of patient populations at elevated risk for emergency department readmissions, thereby establishing individualized risk parameters for personalized prognostic models.

Materials and Methods

Design

This investigation employed a retrospective, case-control architecture using patient readmission records from the 2018–2019 period. Initially, a specialized data extraction template was engineered based on prior literature concerning unscheduled hospital readmissions (unexpected hospital readmissions). Following administrative authorization from the facility's Clinical Committee (granted on April 30, 2019) and institutional leadership, data anonymization was executed to capture the necessary study variables while strictly omitting personal identifiers. A preliminary dataset was subsequently

delivered for processing. Data cleaning, curation, and statistical protocols are detailed in the subsequent sections. All compiled information was managed under rigorous confidentiality safeguards, complying with the Spanish Organic Law on Data Protection (Organic Law 15/1999) alongside concurrent statutory regulations.

Data collection

The target cohort for this investigation comprised all individuals who received care at the emergency department of the University Clinic Hospital of Murcia (Spain) and were subsequently hospitalized through this unit, categorized as either a readmission or a non-readmission event. Analytical data spanned all emergency department registries from 2018 and 2019. Records from 2020–2022 were intentionally excluded to prevent confounding biases introduced by the COVID-19 pandemic, which could skew contemporary baseline realities.

From the gathered data repository, all registries originating from maternal and pediatric facilities—including obstetrics, gynecology, pediatrics, neonatology, and school-age services—were removed. Patients younger than 14 years of age were excluded from the analysis. Similarly, adolescents under 18 years of age who routinely received care within specialized clinical units, such as pediatric oncology and child psychiatry, were likewise omitted.

The dataset compiled sociodemographic variables, including biological sex, age, and ZIP code, alongside administrative metrics regarding admitting and discharging medical specialties. Pathological conditions were coded using two diagnostic frameworks: the International Classification of Diseases (ICD-10) and the Clinical Classification Software (CCS). Multiple clinical history variables were also harvested. Due to pronounced variance across medical records, reliable historical data could only be systematically extracted for hypertension, diabetes mellitus, dyslipidemia, and allergies. Alcohol and tobacco habits were documented, as were active chronic treatments and documented histories of adverse drug reactions. Finally, diagnostic imaging orders (including radiographs, ultrasounds, and computed tomography scans) were incorporated into the analytical file.

The initial raw database encompassed 76,210 patients admitted during the specified study timeframe. Two records were omitted due to unrecorded biological sex. Missing documentation regarding the admitting service necessitated the removal of 883 patients, while an absent discharge service record led to the exclusion of 424 individuals. Missing demographic information (age or ZIP code) accounted for the removal of an additional 51 subjects. Because this study specifically targeted individuals hospitalized via emergency services, selecting this subgroup yielded a modified dataset of $n = 50,729$ patients. The final data cleaning steps involved removing patients under 18 years of age ($n = 8,176$) and those associated with maternal specialties ($n = 13,527$). Ultimately, the curated analytical dataset tracking the population in this study comprised 29,026 individuals.

Outcome variable

The primary target endpoint was unexpected hospital readmission occurring between the second week and the 30th post-discharge day, which was calculated utilizing patient registries from 2018–2019. The extracted database was restricted exclusively to individuals whose index admission originated directly from emergency services.

Statistical analysis

Descriptive statistics were computed for the sociodemographic attributes and all other potential predictive characteristics. Chi-squared tests of independence were executed to evaluate the associations between unexpected hospital readmission and biological sex, as well as other categorical variables. An unpaired t-test was conducted to determine potential differences in patient age and length of stay based on the occurrence or absence of an UHR. A multivariable logistic regression model was constructed to investigate variables independently associated with UHR.

Given the extensive number of independent variables and to mitigate the risks of multicollinearity, this modeling workflow was executed in three distinct stages. The initial model evaluated the association between UHR and the respective

admission and discharge medical specialties. The second model examined the predictive impact of the patient's primary disease classification. The third model analyzed the remaining clinical history parameters. Ultimately, a consolidated final model was derived via backward stepwise elimination, incorporating the variables that demonstrated significant predictive value in the preliminary models. Adjusted odds ratios alongside their corresponding 95% confidence intervals were reported.

All statistical procedures were carried out utilizing IBM SPSS software, version 27.0 (IBM SPSS software, Armonk, NY, USA). Graphical representations were generated using GraphPad Prism 9.0 (Graphpad Software, Boston, MA, USA). Spatial probability maps were constructed via Qgis 3.22 (<https://qgis.org/>, accessed on 7 July 2024).

Machine learning procedure

Data cleaning and pre-processing

The data pre-processing workflow is illustrated in **Figure 1**. Initially, the database was structured to isolate the target variable (UHR) from the array of input features. This technical phase was executed using the Python scikit-learn library and RStudio (<https://posit.co/download/rstudio-desktop/>, accessed on 7 July 2024). For nominal features lacking an intrinsic ordering, such as the Admission Service and Major Diagnostic Categories, we applied OneHotEncoder to transform them into sets of binary dummy variables. Ordinal attributes, such as emergency triage levels, were transformed via LabelEncoder to preserve their sequential progression, which could yield valuable patterns for the predictive architecture. Continuous numerical parameters were standardized using StandardScaler. This standardization process guaranteed that every feature contributed uniformly to the model's objective function, thereby optimizing general performance.

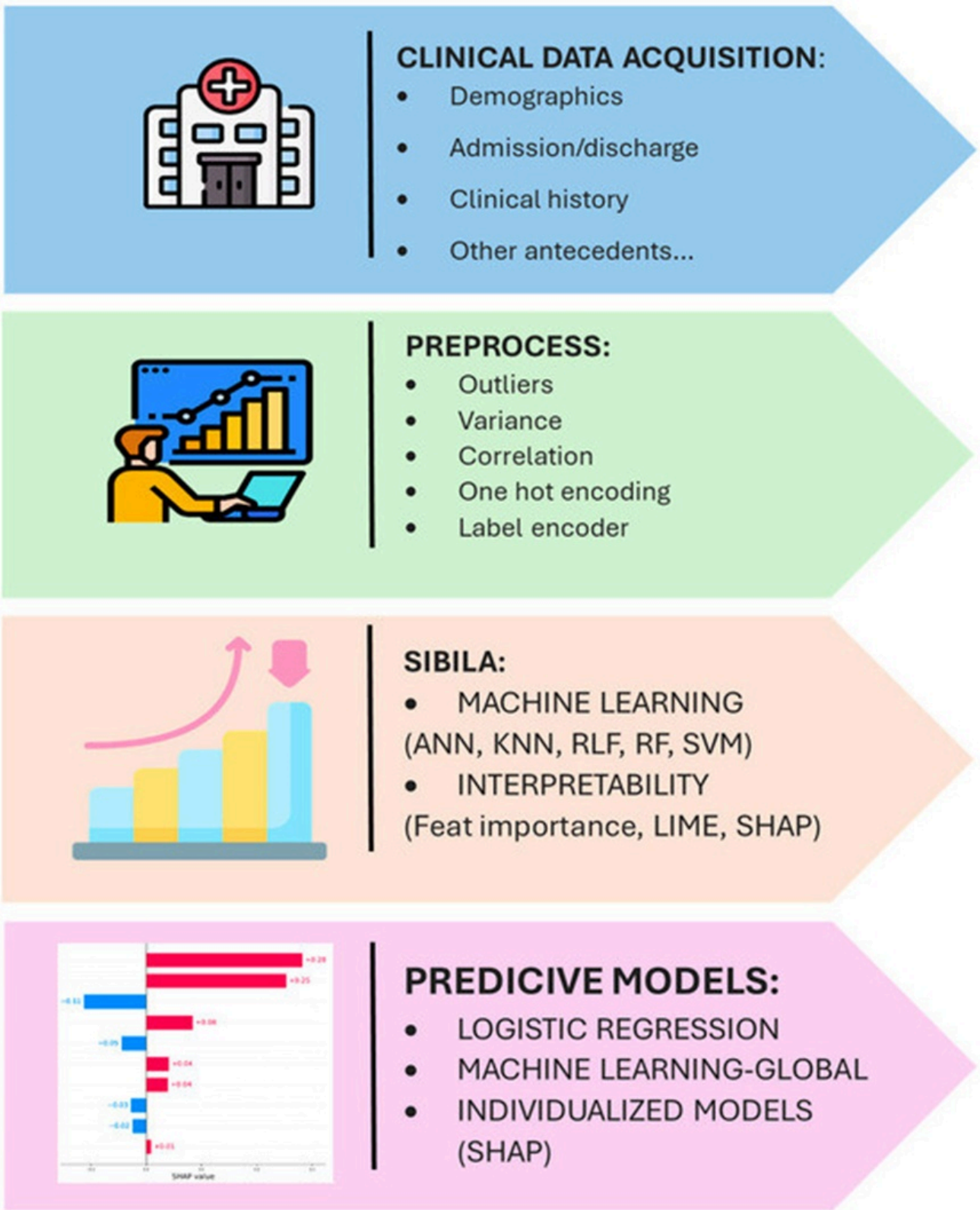


Figure 1. Overview of the study design.

The dataset, comprising 29,026 records following curation, was partitioned into training (80%) and validation (20%) subsets utilizing the .model_selection module. This split strategy, which allocated 23,205 entries for model training and

5,821 entries for validation, was implemented to strike an optimal balance between robust model training and reliable validation verification while optimizing computational resource usage.

Patient data anonymization constituted a fundamental phase of our preparation, executed to safeguard patient confidentiality while ensuring dataset consistency. The de-identification strategies included the absolute deletion of direct personal identifiers and the binning or aggregation of potentially traceable demographic features.

Machine learning analysis

For the quantitative evaluation of the dataset, this investigation deployed the SIBILA computational framework, an analytical tool engineered to utilize High-Performance Computing infrastructures for the development and assessment of diverse advanced machine learning models (Available at <https://github.com/bio-hpc/sibila>, accessed on 5 July 2024). This workflow enables the execution of Artificial Neural Networks, Random Forests, and eXtreme Gradient Boosting (XGBoost) models. These specific architectures were chosen due to their demonstrated efficacy in modeling the complex, high-dimensional datasets characteristic of medical research.

A key attribute of the SIBILA platform is its embedded integration of model interpretability frameworks, which include feature importance metrics, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP). These interpretability approaches are crucial for opening the "black box" mechanics of the utilized algorithms. They are essential for clarifying both the global feature hierarchy and the local case-specific feature contributions within the predictive architectures, ensuring a comprehensive understanding of the factors driving hospital readmission trends.

Additionally, SIBILA automatically generates detailed performance evaluations for each algorithm, placing a primary focus on the Area Under the Receiver Operating Characteristic (AUC-ROC) curve. Supplementary metrics, such as sensitivity, overall accuracy, and related indicators, were also systematically verified. The interpretability reports provided by SIBILA serve as empirical evidence to guide potential clinical implementation.

Results and Discussion

Preliminary analysis

Following the rigorous curation of the source repository, the final analytical cohort for this study comprised 29,026 individuals. An initial evaluation of unexpected hospital readmissions (UHR) across 2018 and 2019 demonstrated an overall incidence rate of 14.3% ($n = 4145$) within the observed population. The mean age of readmitted individuals was statistically significantly higher than that of patients who did not experience a readmission (64.1 versus 62.9 years, $p < 0.001$) (**Figure 2a**). Although the rate of UHR generally scaled upward with advancing age, it is worth noting that individuals aged 80 years or older displayed a lower incidence of UHR compared to the 70–79 age cohort (**Figure 2b**). Male patients exhibited a higher trend toward unexpected readmissions than female patients, though this specific demographic variation did not achieve statistical significance ($p = 0.139$) (**Figure 2c**). Furthermore, the duration of index hospitalizations was significantly longer in the cohort that experienced an UHR.

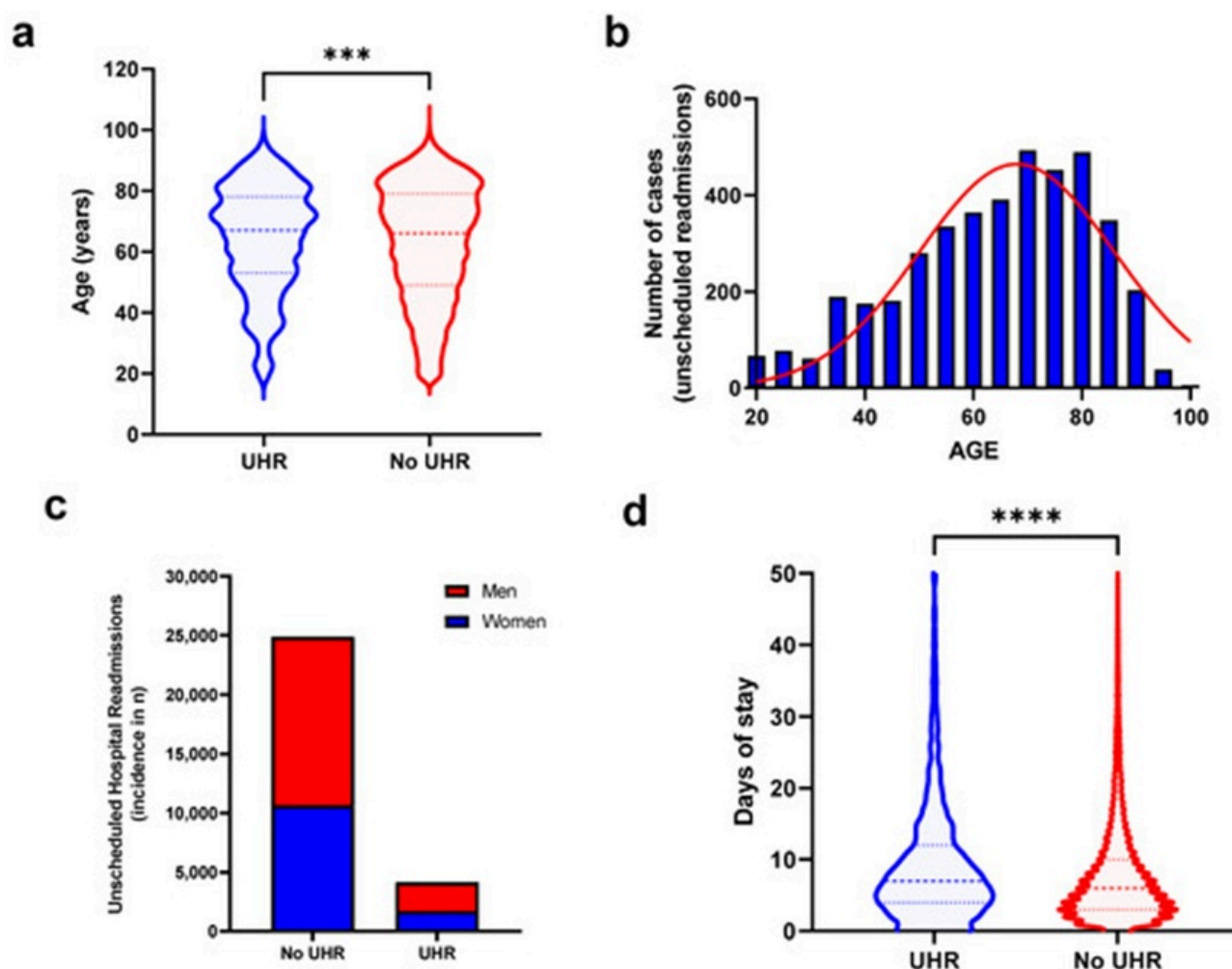


Figure 2. Characteristics of the population to the presence or not of unscheduled hospital readmission. (a) shows the distribution of ages depending on whether they were readmitted or not. The distribution of readmissions by age is shown in (b). The association with sex is also shown in (c). The distribution of patients' days of stay based on readmission is also shown in (d). Those patients that were readmitted were older (t-test $p < 0.001$) and stayed longer (t-test $p < 0.0001$).

The clinical department accounting for the highest volume of UHR cases was internal medicine. An identical trend was observed when evaluating the specific discharging services. Conversely, several departments—including pulmonology, maxillofacial surgery, and rheumatology—demonstrated UHR rates below 5%. According to the ICD-10 diagnostic framework, factors influencing health status and contact with health services (ICD section XXI) and diseases of the blood and blood-forming organs (ICD section III) were associated with the highest UHR incidence rates. Under the CCS classification, patients hospitalized for endocrine metabolic disorders experienced the highest frequency of UHR.

Wednesdays and Sundays represented the weekdays with the most prominent UHR rates. The baseline incidence of UHR remained relatively stable across the months, though slight increases were discernible during the summer season and throughout December. Hospitalizations initiated between 07:00 and 10:00 hours were linked to higher readmission rates, whereas the discharge hour associated with the peak UHR incidence occurred at 05:00 hours. As anticipated, individuals presenting with more severe medical histories were readmitted in greater proportions. Alcohol and tobacco consumption patterns also correlated with unexpected readmission outcomes. Interestingly, former consumers of alcohol and tobacco demonstrated higher UHR rates than active consumers or individuals who had never consumed these substances ($p < 0.001$ across all comparisons).

Lastly, we analyzed the spatial influence of patient residence locations on the likelihood of experiencing an UHR. Remarkably, we identified several residential zones where UHR probabilities exceeded 0.8, whereas other geographical sectors exhibited probabilities below 0.2, highlighting pronounced healthcare disparities between different neighborhoods within the same municipality (Figure 3).

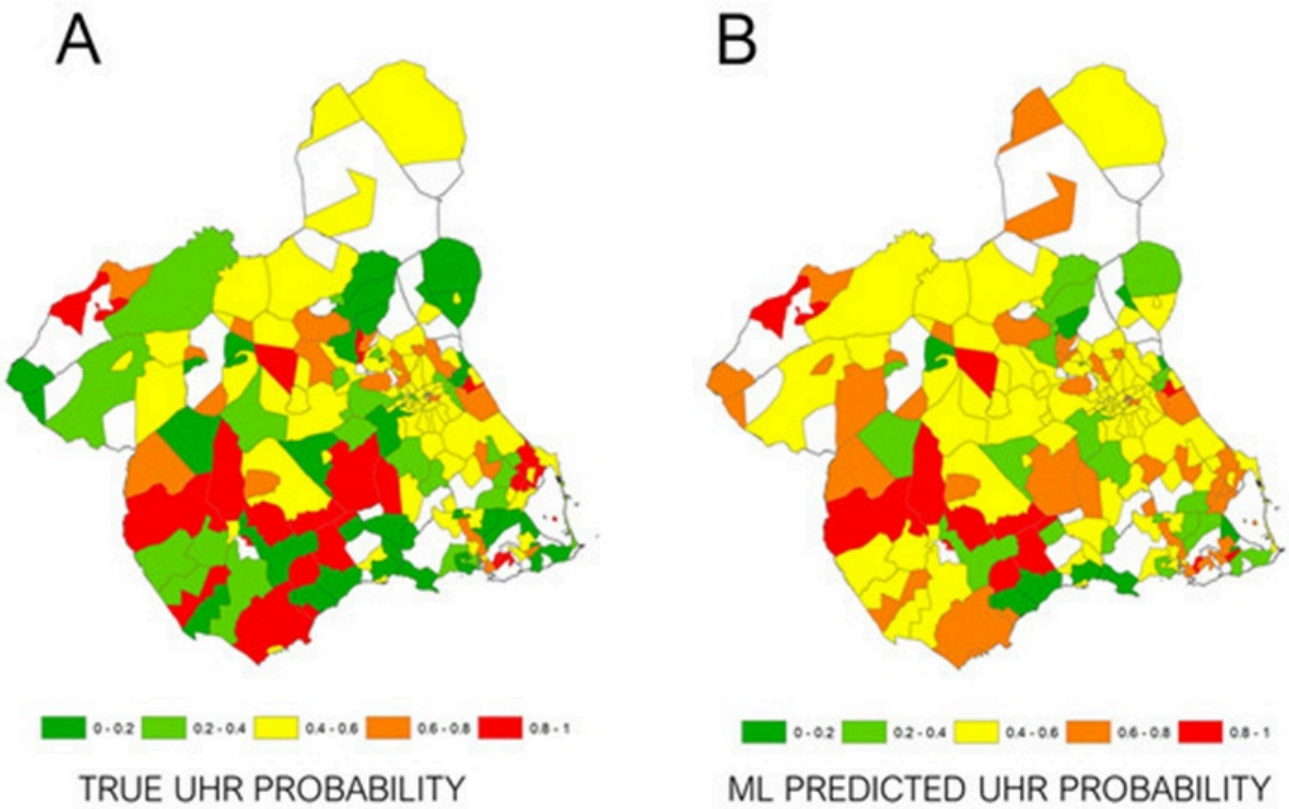


Figure 3. Observed (A) and machine learning (B) unscheduled hospital readmission (UHR) probability attending to the ZIP code. Colour scale represents probability of unscheduled hospital readmission, with green colour showing the lowest probability and red colours showing the highest.

Prediction models of unscheduled hospital readmission

To forecast unexpected clinical returns, a multivariable logistic regression framework was engineered utilizing unexpected hospital readmission as the dependent variable. Initially, a bivariate analysis was executed to filter out features lacking statistical significance in relation to UHR. Consequently, three distinct logistic frameworks were constructed to better characterize the specific predictors integrated into the forecasting architecture. To compile the final model, a stepwise multivariable logistic regression was executed, restricting entry exclusively to variables that displayed a statistically significant impact within the preceding models.

The final configuration (Table 1) demonstrated that active chronic therapy represented the attribute with the highest odds ratio (OR), establishing that patients undergoing continuous chronic treatment experienced a 3.75-fold increased risk of UHR. Individuals initially admitted to digestive surgery and hematology services also exhibited elevated liabilities for UHR (OR: 2.99 and 2.27, respectively). Hospitalization duration was likewise retained in the model, indicating that each additional day a patient remains hospitalized scales their probability of readmission upward by 2%.

Table 1. Multivariable logistic regression model to predict the unscheduled hospital readmission.

	Upper CI95%	Lower CI95%	Odd Ratio	p.	S.E. (B)	B
Chronic therapy	9.699	1.451	3.752	0.006	0.485	1.322
Diabetes mellitus	1.221	1.055	1.135	0.001	0.037	0.127
Days of stay	1.029	1.012	1.020	0.000	0.004	0.020
CCS-16	0.658	0.082	0.233	0.006	0.530	-1.457
CCS-17	0.801	0.188	0.388	0.010	0.369	-0.946
Adm:Cardiology	0.793	0.312	0.498	0.003	0.238	-0.697
Adm:Digestive Surgery	6.500	1.376	2.991	0.006	0.396	1.096
Adm:Haematology	5.083	1.011	2.267	0.047	0.412	0.819
Adm:Oncology	2.322	1.073	1.578	0.021	0.197	0.456
Adm:Internal Medicine	0.991	0.603	0.773	0.042	0.127	-0.257
Dis:Digestive Surgery	0.963	0.220	0.460	0.039	0.377	-0.777
Dis:Pneumology	0.706	0.144	0.319	0.005	0.406	-1.143

B: standard coefficient. S.E.: standard error. CI: confidence interval. Adm: admission service. Dis: discharge service. CCS: clinical classification software diagnosis: 16: injuries and poisoning. 17: symptoms; signs, and ill-defined conditions and factors influencing health status.

Regarding classification performance, the statistical data indicated that the regression framework correctly identified 80.2% of verified UHR cases. The analytical performance parameters demonstrated strong sensitivity (0.81), whereas the corresponding specificity remained constrained (0.42). This performance limitation prompted the adoption of a machine learning workflow to combine predictive attributes into an enhanced prognostic framework for UHR.

The empirical findings demonstrated that the Random Forest (RF) model exhibited the strongest predictive capacity among the candidate architectures, achieving an Area Under the Curve of 0.986. The model attained an overall accuracy of 94% alongside a specificity of 91%. Consequently, the Random Forest framework was chosen as the definitive predictive system.

Predictive attributes were ordered by their relative contribution to UHR forecasting based on the SHAP values derived from the model's outputs. Length of stay emerged as the primary determinant, mirroring its inclusion in the logistic framework and substantiating that prolonged hospitalizations correlate with an increased likelihood of UHR. The subsequent attribute in the feature hierarchy was patient age; however, in this context, the direction of its association was less uniform compared to length of stay.

Within the machine learning model, a history of allergies displayed a monotonic relationship with UHR risk. Notably, documented histories of adverse drug reactions and specific patient ZIP codes also represented crucial components for predicting UHR. The global impact and directional trends of the remaining features are illustrated in [Figure 4](#).

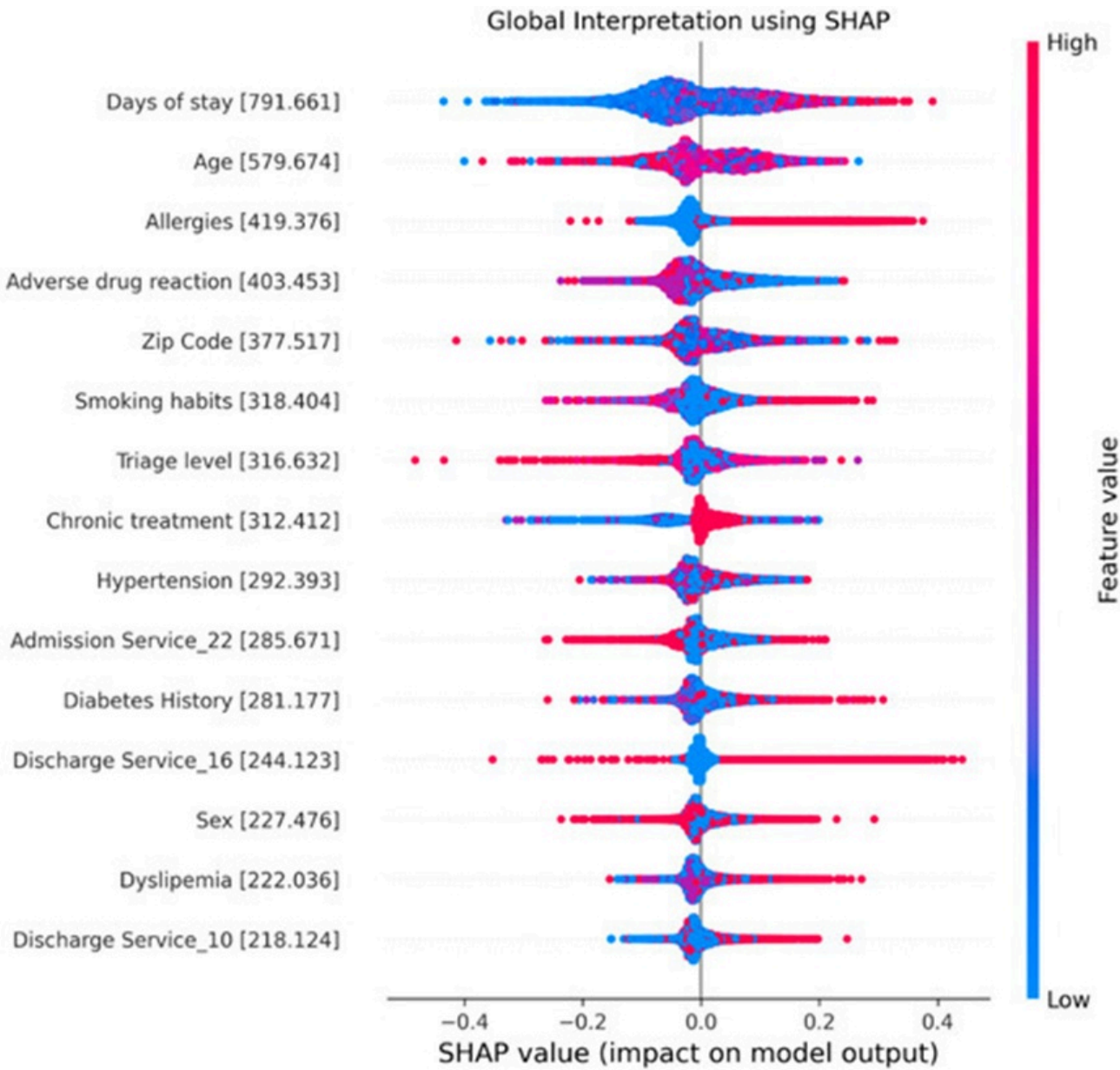


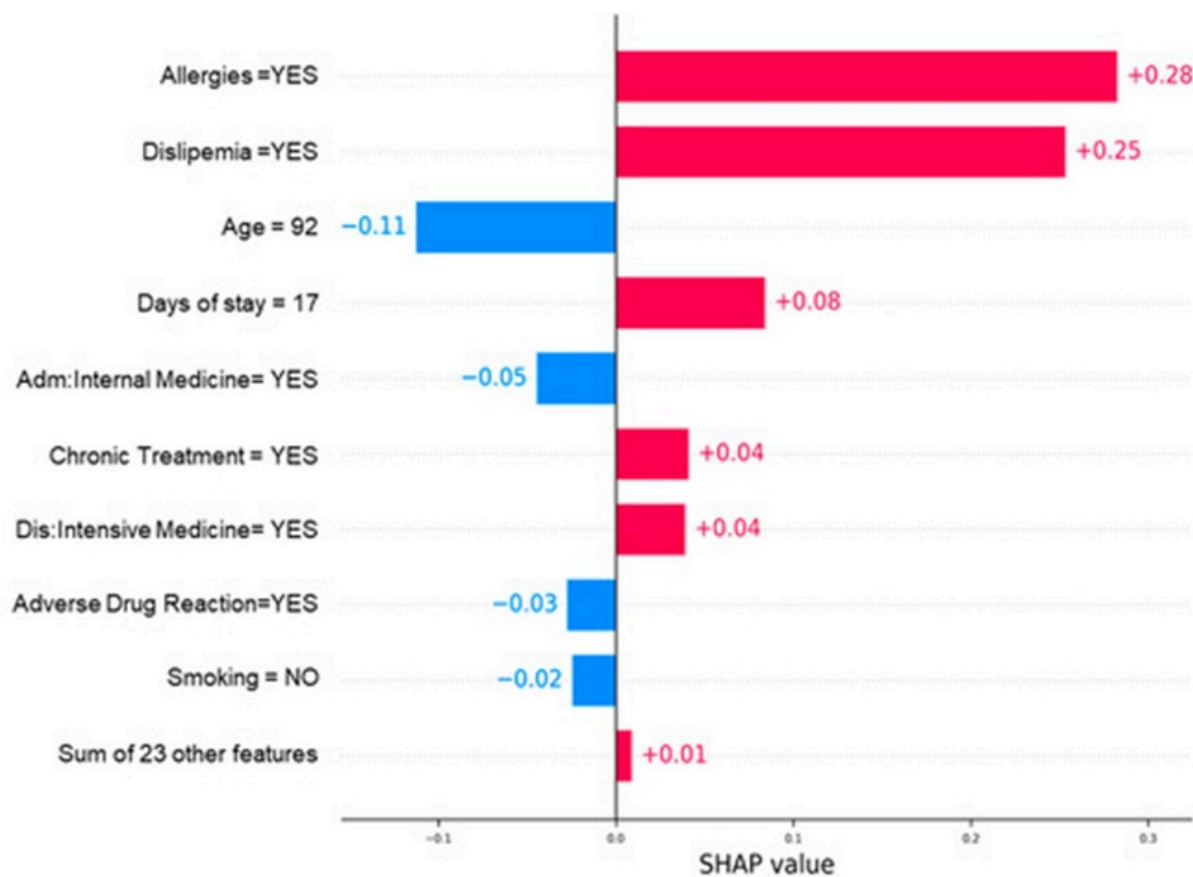
Figure 4. The impact of the input features on unscheduled hospital readmissions predictions. In the figure, each dot represents the effect of a feature on the prediction for one patient. The redder the colour of the dots is, the higher the value of the features is, and the bluer the colour of the dots is, the lower the value of the features is. Dots to the left x-axis represent patients with values of the features decreasing unexpected hospital readmission prediction, and dots to the right x-axis represent patients with values of the features increasing unexpected hospital readmission prediction.

Leveraging the superior precision and sensitivity profiles of the Random Forest configuration, we introduced an individualized risk factor profiling framework to clarify specific UHR predictions at the single-patient level. The clinical application of this personalized risk profiling tool is demonstrated for a patient who experienced an UHR (Figure 5a) alongside an individual who did not experience a readmission (Figure 5b). In both instances, the computed probability reached 0.999.

In the case-specific prediction of UHR, a documented history of allergies and dyslipidemia represented the most powerful features, driving an upward shift in the computed probability of readmission. Length of stay and active chronic regimens were similarly associated with heightened UHR risk for this specific individual. Conversely, for this same patient, an

advanced age of 92 years exerted an inverse, protective effect on the calculated readmission probability, which underscores the clinical importance of generating localized, patient-level interpretable explanations. In the alternative scenario (**Figure 5b**), the absence of active chronic regimens and the lack of a hypertension history served as the primary determinants driving a lower risk calculation for UHR.

a



b

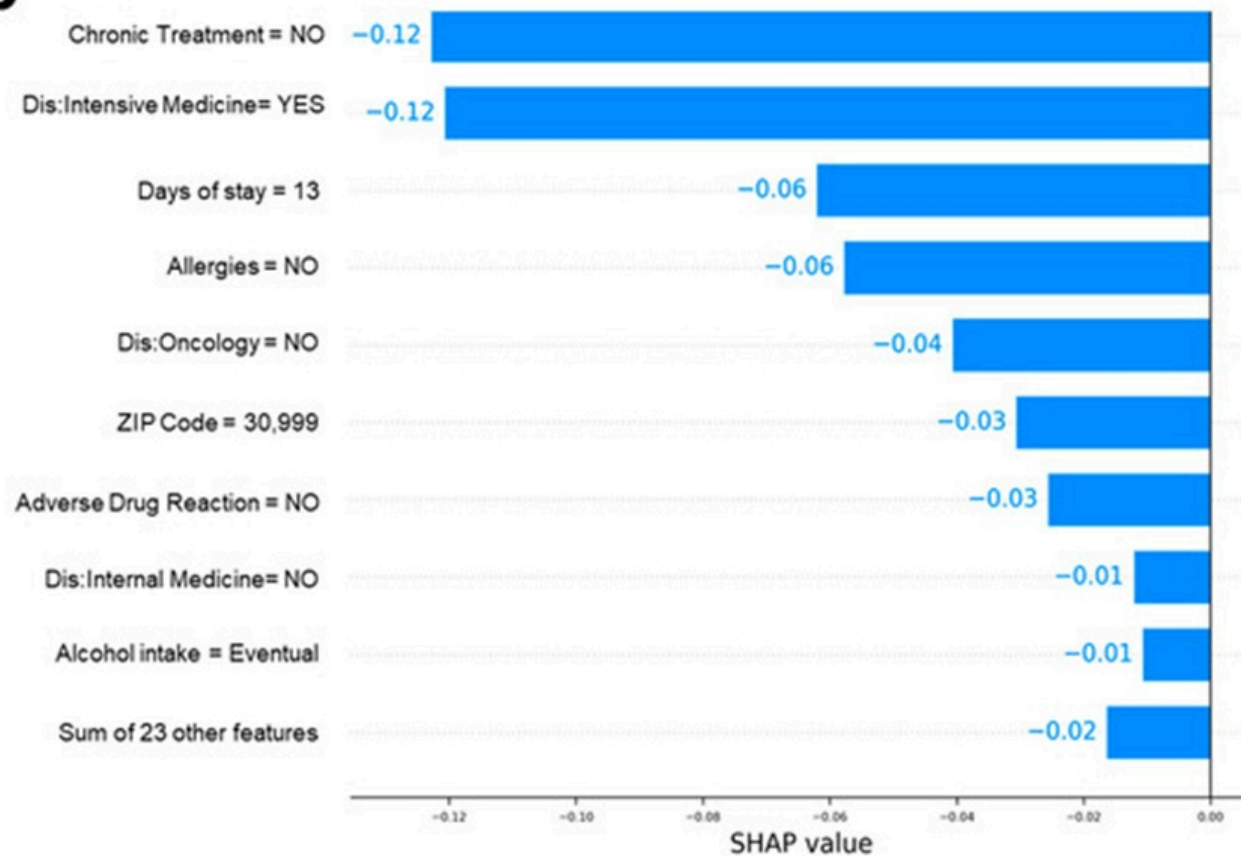


Figure 5. Examples of individualized unscheduled hospital readmissions (UHR) prediction. (a) is an example of personalized risk factor analysis for a patient in the test set identified as UHR. (b) is an example of personalized risk factor analysis for an individual identified as no-UHR.

The deployment of machine learning (ML) architectures within healthcare infrastructures has generated substantial clinical interest in recent years, presenting viable pathways for advancing patient outcomes and maximizing resource allocation efficiency [8]. In this study, we sought to address the persistent challenge of hospital readmissions through the design and cross-validation of an individualized, interpretable ML framework. This predictive structure was engineered to calculate the likelihood of an unexpected return within a 30-day post-discharge window. This prognostic model is intended to guide targeted clinical strategies to mitigate readmission rates.

An initial assessment of the target metrics revealed a 14.3% baseline UHR incidence across the examined sample. This frequency aligns with observations reported in prior literature evaluating cohorts older than 65 years, which document readmission rates ranging between 10.7% and 24.8% [11–14]. Large-scale epidemiological investigations, such as the seminal study by Jencks *et al.*, demonstrated that among 2,961,460 discharged beneficiaries, 19.6% experienced a readmission within the initial 30 days post-discharge, with the proportion rising to 34% by the 90th post-discharge day [15].

Individuals falling within the 70–79 age interval exhibited a higher baseline incidence of unexpected hospital readmissions relative to the rest of the study cohort. Advanced age is widely categorized as a prominent risk factor for early hospital readmissions, with readmission risks escalating alongside chronological age, particularly among populations over 65 years. Multiple large-scale epidemiological investigations involving adult cohorts and individuals ≥ 65 years have validated this positive correlation [16–18].

In contrast, alternative studies have failed to identify a significant correlation [19, 20]. Furthermore, even when incorporating age parameters into predictive readmission models, the majority of investigators have not observed an association substantial enough to justify its retention in final risk frameworks [21]. Interestingly, in our study, individuals exceeding 80 years of age demonstrated a lower probability of readmission. This divergence may stem from elevated mortality rates or more intensive specialized institutional care delivered to this specific demographic [22].

It remains critical to recognize that the elderly population displays marked heterogeneity; for example, the presence or absence of clinical frailty can fundamentally alter patient prognoses. As noted by Liperoti *et al.*, the systematic assessment of frailty among older individuals is vital for designing personalized healthcare pathways [23]. Within our study cohort, matching broader trends in published literature, male patients experienced a higher frequency of hospitalizations [24, 25]. However, biological sex has been reported as non-influential in alternative investigations [17, 26].

The medical specialty accounting for both the highest volume of total cases and the highest absolute number of UHR events was internal medicine. While limited published data exist concerning this specific department, an earlier retrospective cohort analysis by Fabbian *et al.* documented a lower readmission rate of approximately 5% [27]. Nevertheless, the mean age characterizing that historical sample was lower than the average age observed in our current investigation.

Concurrently, individuals presenting with more complex medical histories were readmitted at a higher rate. This clinical pattern has been widely documented, particularly among cohorts managing chronic heart failure, chronic obstructive pulmonary disease (COPD), and diabetes mellitus [28]. Tobacco and alcohol consumption habits were similarly associated with UHR outcomes. Tobacco exposure remains the leading source of preventable mortality globally and figures prominently in the pathogenesis and exacerbation of COPD. Indeed, comparative assessments of readmission metrics between European and United States infrastructures highlight COPD as a leading driver of hospital readmission rates [29].

The baseline frequency of unexpected hospital readmissions remained relatively consistent across the calendar months, though noticeable spikes in UHR incidence were recorded during the summer period and throughout December. Temporally, UHR incidence peaked during morning admission hours. To the best of our knowledge, no prior investigations have systematically evaluated the relationship between the exact hour of admission or discharge and the resulting UHR probability; hence, these findings establish a novel avenue of inquiry for future research.

Given that the central objective of this research was to construct an effective predictive framework for UHR, an initial multivariate logistic regression model was estimated. This model indicated that individuals subjected to ongoing chronic therapy experienced a 3.75-fold higher risk of UHR. This finding aligns with evidence from Morandi *et al.*, who observed that the administration of ≥ 7 medications escalated the probability of early readmission among chronic patients within a post-acute rehabilitation setting [30].

In addition, the duration of hospitalization was integrated into the predictive model, demonstrating that each supplementary day spent in the hospital increases a patient's risk of readmission by 2%. This continuous variable attracted substantial interest in foundational UHR literature, as investigators initially hypothesized that high-risk individuals subjected to an inappropriately "short" hospitalization would display elevated readmission rates driven by suboptimal care quality [31]. Nevertheless, the majority of subsequent empirical investigations have reported the inverse relationship, showing an escalated risk of readmission associated with prolonged stays [32].

It remains plausible that extended hospitalizations are concentrated among individuals suffering from severe pathologies that require specific medical specialties inherently tied to a higher frequency of emergency readmissions. For instance, patients discharged from digestive surgery units exhibited a three-fold greater likelihood of experiencing a UHR [33]. Conversely, individuals admitted to cardiology services demonstrated a 50% lower probability of UHR, which may reflect superior clinical management within this specialty or a lower baseline susceptibility to complications related to the underlying cardiac conditions [34].

To address the constrained performance of the logistic framework across several validation metrics, a machine learning workflow was developed. The empirical findings demonstrated that the Random Forest architecture yielded the highest predictive capacity among the candidate models evaluated. Within this algorithmic workflow, the length of stay was identified as the most prominent predictor. This outcome is highly noteworthy because it challenges the conventional assumption that brief hospitalizations reflect incomplete care or diagnostic errors; in this cohort, extended hospitalizations were not associated with improved patient outcomes [35]. Furthermore, individuals with a documented history of allergies also exhibited an increased probability of experiencing an unscheduled readmission [36].

A major advantage of machine learning-driven forecasting over classical statistical models is its capacity to identify and leverage implicit features that may initially seem clinically trivial, thereby enhancing overall predictive accuracy. The machine learning architecture engineered in this study identified several subtle features as vital for risk calculation, with the patient's ZIP code serving as a primary example. Traditional geospatial investigations, which are typically executed at a broad national scale, have been useful for locating regions with elevated disease incidence or prevalence.

However, our dataset permits more granular, localized analyses that isolate specific intra-urban zones characterized by a high probability of UHR. This localized strategy offers unprecedented operational insights, exposing specific geographic neighborhoods where post-discharge continuity or primary healthcare services may be insufficient, thereby driving readmission trends. Such outcomes highlight the capability of machine learning techniques not only to detect subtle, high-impact data patterns but also to inform the strategic optimization of healthcare resources aimed at reducing unscheduled readmission rates [37].

Nevertheless, while these collective parameters are highly informative and essential for predicting UHR, establishing individualized risk projections represents a more critical advancement. Until recently, generating personalized, patient-level predictive models was highly constrained by prolonged processing times and intensive computational resource

demands. Today, it is possible to isolate the most influential risk determinants unique to each individual. For example, in the case study presented previously, dyslipidemia was identified as the most dominant feature associated with UHR in that specific patient, despite the fact that neither the global machine learning framework nor the logistic regression model isolated this variable as a major predictor.

In this investigation, we have developed an individualized machine learning model designed to optimize patient care by tailoring clinical follow-up strategies to individual needs at the point of hospital admission. This framework differs from traditional "black-box" models by providing a clear explanation of the features driving its case-specific predictions, thereby supporting more precise and personalized clinical workflows. Deploying this innovative approach is key to reducing unplanned hospital readmissions, which represent a persistent challenge in clinical practice and discharge planning that imposes heavy financial and operational burdens on healthcare systems.

For instance, within the United States, annual expenditures linked to hospital readmissions reach USD 41.3 billion, positioning them among the most expensive healthcare events nationwide [38]. In our specific region, although the healthcare system is highly complex due to decentralized management, estimates suggest that hospital readmissions generate approximately EUR 500 million in annual costs [39]. Beyond financial metrics, UHR events severely diminish the quality of life for patients and their care circles. By integrating interpretable, individualized models, we can simultaneously advance patient care and foster a more sustainable healthcare infrastructure by mitigating the adverse effects of UHR.

Several limitations must be acknowledged in interpreting these findings. First, identifying a correlation between specific risk factors and rehospitalization does not establish definitive causal relationships; thus, we cannot conclude that the variables defining our predictive models are the direct causes of readmission. Second, retrospective data gathering may be susceptible to recall bias during clinical documentation. An additional limitation stems from the geographical origin of the source data. The data utilized in this study are representative of a specific regional population, meaning the baseline performance of this machine learning model must be validated in diverse external geographical regions prior to widespread implementation.

Although our final curated dataset included an extensive patient sample—substantially exceeding the sample sizes of many previous works focused on readmission modeling [37, 40, 41]—evaluating equivalent datasets from disparate healthcare jurisdictions remains essential to confirm the generalizability of the current model. Finally, the presence of unmeasured latent features that could influence readmission outcomes—such as exact longitudinal values from biochemical panels or specialized clinical assays—might further refine the machine learning model. Furthermore, incorporating functional prognostic indicators at the time of discharge, such as cognitive performance scores, would be highly valuable; this layer is of particular relevance to geriatric cohorts and specific medical specialties, as detailed by Fusco et al. [42].

Nevertheless, it is important to emphasize that a core objective of this study was to calculate readmission probabilities utilizing readily available, basic clinical variables without adding operational or financial costs to the facility. Our optimized model achieved an AUC of 99%, demonstrating excellent reliability and predictive performance using standard administrative and clinical records.

Conclusion

Our investigation utilized descriptive statistics to clarify the demographic profiles and baseline risk probabilities associated with emergency hospital admissions. Notably, the data revealed that the majority of readmission events occurred in patients under 80 years of age, with the patient's place of residence and biological sex acting as key contributing factors. To address the challenge of forecasting UHR, we executed a multivariate analysis targeting an array of demographic and clinical characteristics.

The duration of the initial hospital stay emerged as the primary predictor of readmission, confirming a direct relationship between extended hospitalizations and heightened readmission risks. Other significant risk determinants included a history of diabetes mellitus, documented allergies, and previous adverse drug reactions. The core achievement of this research is the construction of an individualized risk model capable of delivering personalized readmission risk assessments. While the current framework exhibits high accuracy and predictive reliability, its performance could be further enhanced by integrating additional continuous clinical variables, such as specific biochemical or laboratory parameters. This expansion would enable a more comprehensive clinical analysis and maximize the model's utility in active clinical practice.

Acknowledgements

The authors wish to express our gratitude to the Bioinformatics Service of the Hospital Virgen de la Arrixaca. We also wish to thank the Department of Biostatistics of the University of Murcia. Finally, we wish to thank Ascensión López (RN) for her assistance with data handling.

Conflict of interest

None

Financial support

This work has been funded by grants from the European Project Horizon 2020 SC1-BHC-02-2019 [REVERT, ID:848098]. Supercomputing resources in this work were supported by the Plataforma Andaluza de Bioinformática at the University of Málaga, the supercomputing infrastructure of the NLHPC (ECM-02, Powered@NLHPC), and the Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT is part of CIEMAT and the Government of Spain.

Ethics statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Hospital Universitario “Virgen de la Arrixaca” (CSV verification code: CARM-665aef0c-6cb0-44ff-7120-0050569b6280; date of approval 30 April 2019).

Received: 26 Jul 2023

Revised: 18 Sep 2023

Accepted: 20 Oct 2023

Published online: 15 March 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Guntaka SM, Tarazi JM, Chen Z, Vakharia R, Mont MA, Roche MW. Higher patient complexities are associated with increased length of stay, complications, and readmissions after total hip arthroplasty. *Surg Technol Int*. 2021;38:422-6.
<https://doi.org/10.52198/21.STI.38.OS1412>.

Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373-83.
[https://doi.org/10.1016/0021-9681\(87\)90171-8](https://doi.org/10.1016/0021-9681(87)90171-8).

de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity: a critical review of available methods. *J Clin Epidemiol*. 2003;56(3):221-9.
[https://doi.org/10.1016/S0895-4356\(02\)00585-1](https://doi.org/10.1016/S0895-4356(02)00585-1).

Ministerio de Sanidad. Registro de Altas de los Hospitales del Sistema Nacional de Salud (CMBD) [Internet]. Madrid: Ministerio de Sanidad; 2024 [cited 2024 Jun 28]. Available from: <https://www.sanidad.gob.es/estadEstudios/estadisticas/cmbdhome.htm>

Pohjanpää M, Ojala R, Luukkaala T, Gissler M, Tammela O. Association of early discharge with increased likelihood of hospital readmission in first four weeks for vaginally delivered neonates. *Acta Paediatr*. 2022;111(6):1144-56.
<https://doi.org/10.1111/apa.16290>.

Xue HT, Stanley-Baker M, Kong AWK, Li HL, Goh WWB. Data considerations for predictive modeling applied to the discovery of bioactive natural products. *Drug Discov Today*. 2022;27(8):2235-43.

Sutter T, Roth JA, Chin-Cheong K, Hug BL, Vogt JE. A comparison of general and disease-specific machine learning models for the prediction of unplanned hospital readmissions. *J Am Med Inform Assoc*. 2021;28(4):868-73.

Habebhh H, Gohel S. Machine learning in healthcare. *Curr Genom*. 2021;22(4):291-300.

Zhou Y, Han W, Yao X, Xue JJ, Li Z, Li Y. Developing a machine learning model for detecting depression, anxiety, and apathy in older adults with mild cognitive impairment using speech and facial expressions: a cross-sectional observational study. *Int J Nurs Stud*. 2023;146:104562.

Sammur SJ, Crispin-Ortuzar M, Chin SF, Provenzano E, Bardwell HA, Ma W, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022;601(7894):623-9.
<https://doi.org/10.1038/s41586-021-04278-5>.

Robinson S, Howie-Esquivel J, Vlahov D. Readmission risk factors after hospital discharge among the elderly. *Popul Health Manag*. 2012;15(6):338-51.
<https://doi.org/10.1089/pop.2011.0095>.

Lanièce I, Couturier P, Dramé M, Gavazzi G, Lehman S, Jolly D, et al. Incidence and main factors associated with early unplanned hospital readmission among French medical inpatients aged 75 and over admitted through emergency units. *Age Ageing*. 2008;37(4):416-22.

Donzé J, Lipsitz S, Bates DW, Schnipper JL. Causes and patterns of readmissions in patients with common comorbidities: retrospective cohort study. *BMJ*. 2013;347:f7171.
<https://doi.org/10.1136/bmj.f7171>.

Cornette P, D'Hoore W, Malhomme B, Van Pee D, Meert P, Swine C. Differential risk factors for early and later hospital readmission of older patients. *Aging Clin Exp Res*. 2005;17(4):322-8.

Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. *N Engl J Med*. 2009;360(14):1418-28.
<https://doi.org/10.1056/NEJMsa0803563>.

Billings J, Blunt I, Steventon A, Georgiou T, Lewis G, Bardsley M. Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). *BMJ Open*. 2012;2(4):e001667.
<https://doi.org/10.1136/bmjopen-2012-001667>.

Allaudeen N, Vidyarthi A, Maselli J, Auerbach A. Redefining readmission risk factors for general medicine patients. *J Hosp Med*. 2011;6(2):54-60.
<https://doi.org/10.1002/jhm.805>.

Coderch J, Sánchez-Pérez I, Ibern P, Carreras M, Pérez-Berrueto X, Inoriza JM. Predicción del riesgo individual de alto coste sanitario para la identificación de pacientes crónicos complejos. *Gac Sanit*. 2014;28(4):292-300.

Fleming LM, Gavin M, Piatkowski G, Chang JD, Mukamal KJ. Derivation and validation of a 30-day heart failure readmission model. *Am J Cardiol*. 2014;114(9):1379-82.
<https://doi.org/10.1016/j.amjcard.2014.07.071>.

Dharmarajan K, Hsieh AF, Lin Z, Bueno H, Ross JS, Horwitz LI, et al. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *JAMA*. 2013;309(4):355-63.
<https://doi.org/10.1001/jama.2012.216476>.

Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. *JAMA*. 2011;306(15):1688-98.
<https://doi.org/10.1001/jama.2011.1515>.

Glans M, Kragh Ekstam A, Jakobsson U, Bondesson Å, Midlöv P. Risk factors for hospital readmission in older adults within 30 days of discharge—a comparative retrospective study. *BMC Geriatr*. 2020;20:467.
<https://doi.org/10.1186/s12877-020-01867-3>.

Liperoti R, Vetrano DL, Palmer K, Targowski T, Cipriani MC, Lo Monaco MR, et al. Association between frailty and ischemic heart disease: a systematic review and meta-analysis. *BMC Geriatr*. 2021;21:357.

Silverstein MD, Qin H, Mercer SQ, Fong J, Haydar Z. Risk factors for 30-day hospital readmission in patients 65 years of age. *Proc (Bayl Univ Med Cent)*. 2008;21(4):363-72.

Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med Care*. 2010;48(11):981-8.

Bisiani MA, Jurgens CY. Do collaborative case management models decrease hospital readmission rates among high-risk patients? *Prof Case Manag*. 2015;20(4):188-96.

Fabbian F, Boccafogli A, De Giorgi A, Pala M, Salmi R, Melandri R, et al. The crucial factor of hospital readmissions: a retrospective cohort study of patients evaluated in the emergency department and admitted to the department of medicine of a general hospital in Italy. *Eur J Med Res*. 2015;20:6.

Kroeze ED, de Groot AJ, Smorenburg SM, Mac Neil Vroomen JL, van Vught AJAH, Buurman BM. A case vignette study to refine the target group of an intermediate care model: the Acute Geriatric Community Hospital. *Eur Geriatr Med*. 2024;15(4):977-89.
<https://doi.org/10.1007/s41999-024-00947-6>.

Westert GP, Lagoe RJ, Keskimäki I, Leyland A, Murphy M. An international study of hospital readmissions and related utilization in Europe and the USA. *Health Policy*. 2002;61(3):269-78.

Morandi A, Bellelli G, Vasilevskis EE, Turco R, Guerini F, Torpilliesi T, et al. Predictors of rehospitalization among elderly patients admitted to a rehabilitation hospital: the role of polypharmacy, functional status and length of stay. *J Am Med Dir Assoc*. 2013;14(10):761-7.
<https://doi.org/10.1016/j.jamda.2013.03.013>.

Grovu R, Huo Y, Nguyen A, Mourad O, Pan Z, El-Gharib K, et al. Machine learning: predicting hospital length of stay in patients admitted for lupus flares. *Lupus*. 2023;32(12):1418-29.
<https://doi.org/10.1177/09612033231206830>.

Zanocchi M, Maero B, Martinelli E, Cerrato F, Corsinovi L, Gonella M, et al. Early re-hospitalization of elderly people discharged from a geriatric ward. *Aging Clin Exp Res*. 2006;18(1):63-9.

Im KM, Kim EY. Identification of ICU patients with high nutritional risk after abdominal surgery using modified NUTRIC score and the association of energy adequacy with 90-day mortality. *Nutrients*. 2022;14(5):946.
<https://doi.org/10.3390/nu14050946>.

Albinali HHA, Singh R, Al Arabi A, Al Qahtani A, Asaad N, Al Suwaidi J. Predictors of 30-day re-admission in cardiac patients at Heart Hospital, Qatar. *Heart Views*. 2023;24(3):125-35.
https://doi.org/10.4103/heartviews.heartviews_91_22.

Mechelli A, Lin A, Wood S, McGorry P, Amminger P, Tognin S, et al. Using clinical information to make individualized prognostic predictions in people at ultra high risk for psychosis. *Schizophr Res*. 2017;184:32-8.
<https://doi.org/10.1016/j.schres.2016.11.047>.

Ortiz-Barrios M, Altamar-Maldonado Z, Martínez-Solano C, Petrillo A, De Felice F, Jiménez-Delgado G, et al. Predicting 15-day unplanned readmissions in hospitalization departments: an application of logistic regression. *Ingeniare Rev Chil Ing*. 2021;29(2):378-98.
<https://doi.org/10.4067/S0718-33052021000200378>.

Conilione P, Jessup R, Gust A. Novel machine learning model for predicting multiple unplanned hospitalisations. *BMJ Health Care Inform*. 2023;30(1):e100682.
<https://doi.org/10.1136/bmjhci-2022-100682>.

Yhdego HH, Nayebeazar A, Amrollahi F, Boussina A, Shashikumar S, Wardi G, et al. Prediction of unplanned hospital readmission using clinical and longitudinal wearable sensor features. *medRxiv*. 2023.
<https://doi.org/10.1101/2023.11.20.23298736>.

The Lancet Public Health. COVID-19 in Spain: a predictable storm? *Lancet Public Health*. 2020;5(11):e568.
[https://doi.org/10.1016/S2468-2667\(20\)30239-5](https://doi.org/10.1016/S2468-2667(20)30239-5).

Lo YT, Liao JC, Chen MH, Chang CM, Li CT. Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms. *BMC Med Inform Decis Mak*. 2021;21:288.
<https://doi.org/10.1186/s12911-021-01639-y>.

Yu MY, Son YJ. Machine learning-based 30-day readmission prediction models for patients with heart failure: a systematic review. *Eur J Cardiovasc Nurs*. 2024;23(5):e1-e10.

Fusco A, Galluccio C, Castelli L, Pazzaglia C, Pastorino R, Pires Marafon D, et al. Severe acquired brain injury: prognostic factors of discharge outcome in older adults. *Brain Sci*. 2022;12(9):1232.
<https://doi.org/10.3390/brainsci12091232>.