



OPEN Predicting high-need high-cost pediatric hospitalized patients in China based on machine learning methods

Peng Zhang^{1,5}, Bifan Zhu^{2,5}, Xing Chen³ & Linan Wang^{2,4}✉

Rapidly increasing healthcare spending globally is significantly driven by high-need, high-cost (HNHC) patients, who account for the top 5% of annual healthcare costs but over half of total expenditures. The programs targeting existing HNHC patients have shown limited long-term impact, and research predicting HNHC pediatric patients in China is limited. There is an urgent need to establish a specific, valid, and reliable prediction model using machine-learning-based methods to identify potential HNHC pediatric patients and implement proactive interventions before high costs arise. This study used a 7-year retrospective cohort dataset from two administrative databases in Shanghai, covering pediatric patients under 18 years. The machine-learning-based models were developed to predict HNHC status using logistic regression, k-nearest neighbors (KNN), random forest (RF), multi-layer perceptron (MLP), and Naive Bayes. This study divided the data from 2021–2022 into 70:30 as a training set and a test set, with the internal class balancing approach of the Synthetic Minority Over-sampling Technique (SMOTE). A grid search strategy was employed with k-fold cross-validation to optimize hyperparameters. Model performance was assessed by 5 metrics: Receiver Operating Characteristic-Area Under Curve (ROC-AUC), accuracy, sensitivity, specificity, and F1 score. The external validation from 2022–2023 data and the internal validation using different train-test ratios (80:20 and 90:10) were used to assess the robustness of the trained models. Among the 91,882 hospitalized children included in 2021, significant differences were found in socioeconomic, disease, healthcare service utilization, previous healthcare expenditure, and hospital characteristics between the HNHC and non-HNHC groups. The hospitalization costs for HNHC pediatric patients accounted for over 35% of total spending. The MLP model demonstrated the highest predictive performance (ROC-AUC: 0.872), followed by RF (0.869), KNN (0.836), and naive Bayes (0.828). The most important predictive factors included length of stay, number of hospitalizations, previous HNHC status, age, and presence of Top 20 HNHC diseases. MLP showed robustness as the most efficient model in external validation (ROC-AUC: 0.843) and internal validation using different train-test ratios (ROC-AUC: 0.826 in 80:20 ratio; 0.807 in 90:10 ratio). Machine learning models, particularly MLP, effectively predict HNHC pediatric patients, providing a basis for early identification of HNHC and proactive healthcare interventions into clinical practice. This approach can also assist policymakers and payers in optimizing healthcare resource allocation, controlling healthcare costs, and improving patient outcomes.

Rapidly increasing healthcare spending has become a significant challenge globally¹, primarily driven by high-need, high-cost (HNHC) patients, who are typically defined as those accounting for the top 5% of annual healthcare expenditures². Research has shown that the top 5% of HNHC patients accounted for over 50% of the total healthcare expenditures in the United Kingdom (54%, 2022)³, the United States (51.2%, 2021)⁴, Canada (60%), Spain (56%), and Germany (50%, 2019)⁵. The substantial financial burden imposed by HNHC patients underscores the urgent need for targeted interventions to optimize healthcare resource allocation and reduce costs⁶. Furthermore, studies demonstrate that HNHC patients and their families face a significant disease-

¹School of Humanities, Shanghai Institute of Technology, Haiquan Road 100, Fengxian District, Shanghai 201418, China. ²Shanghai Health Development Research Center, Room 804, No 1477, West Beijing Road, Jing'an District, Shanghai 20040, China. ³Fudan University, Dongan Street, 130, Xuhui District, Shanghai 200032, China. ⁴Shanghai Key Laboratory of Maternal Fetal Medicine, Shanghai Institute of Maternal-Fetal Medicine and Gynecologic Oncology, Shanghai First Maternity and Infant Hospital, School of Medicine, Tongji University, Shanghai 200092, China. ⁵Peng Zhang, Bifan Zhu authors are joint first author. ✉email: wanglinan@51mch.com

related economic burden while experiencing poorer health outcomes, suggesting inefficiencies in healthcare service delivery and resource allocation^{6–8}.

To address these challenges, many countries have prioritized HNHC patients in healthcare reform efforts by exploring various service delivery models, such as Adapting Stepped Care, Intensive Primary/Outpatient Care Programs, Interdisciplinary Transitional Care Programs, etc.^{9–15}. However, evaluations of these initiatives have yielded mixed results. While some programs have shown limited effectiveness in reducing emergency and inpatient services for specific patient subgroups, their long-term impact on cost control remains limited^{9,10,12,13,15}. Previous research emphasizes the limited effectiveness of targeting existing HNHC patients, highlighting the necessity for healthcare systems to proactively predict and identify potential high-risk HNHC patients before high needs and costs arise, enabling preventive interventions and personalized services to control healthcare expenses and improve patient outcomes^{10,12,16,17}.

Therefore, a prediction model for future HNHC patients has gained significant attention from policymakers and payers, as targeted interventions for this group are expected to be more effective. International studies have predominantly focused on predicting HNHC patients in adult populations, utilizing data from countries such as the United States, South Korea, and Japan, with no known prediction models for HNHC children^{18–20}. Previous studies have commonly utilized healthcare insurance claims or hospitalization data, employing machine-learning-based methods such as random forest^{18–25}, artificial neural network^{19,20,22,26,27}, k-nearest neighbors²⁴, naive Bayes^{21,28,29}, and XGBoost³⁰ to construct models, offering methodological insights for this study. In recent years, while studies have further subdivided the HNHC population and developed predictive models for sustained HNHC patients^{21,26}, there remains a critical gap in predictive models specifically tailored for pediatric HNHC patients globally.

In China, research has concentrated on single diseases such as heart disease and chronic obstructive pulmonary disease, developing models to predict HNHC patients with these specific conditions, and these studies lack comprehensive analyses across diverse disease categories^{22,28,31}. There is also a paucity of evidence on comprehensive prediction models for pediatric HNHC patients in China. However, evidence is limited as to HNHC patients with all disease categories in China. Significant differences in characteristics, prevalent diseases, healthcare utilization, and expenditures between children and adults further underscore the necessity of developing a specific, valid, and reliable machine-learning-based model tailored to pediatric HNHC patients³².

This study aims to develop a machine-learning-based prediction model for HNHC pediatric patients using a 7-year retrospective cohort dataset from China. The proposed model is intended to support decision-makers and payers in formulating effective strategies for early interventions, thereby mitigating the rapid escalation of healthcare costs.

Methods

Data resource

This study conducted a retrospective cohort study from January 1, 2017, to December 31, 2023, with data from two sources in Shanghai, China. One database originates from the administrative data of the Urban and Rural Resident Basic Medical Insurance (URBMI), a fundamental component of China's social medical insurance system providing coverage for children. Another database derives from the administrative records of the Children's Hospitalization Mutual Fund (CHMF), a non-profit, public welfare medical security mutual aid fund in Shanghai. Established in 1996, URBMI and CHMF cover almost all children in Shanghai, aiming to share the risk of hospitalization and major outpatient medical expenses for children.

The data used in this study is divided into three parts. The first part comprises 2017–2023 seven-year data comprehensively describing the most common 20 diseases among pediatric HNHC patients, based on the International Classification of Diseases Tenth Revision (ICD-10) and grouping ICD at the three-digit level. The second part consists of training and internal validation datasets from 2021 to 2022 for machine learning. The third part is an external validation dataset used for robustness analysis, employing 2022–2023 data to assess the robustness of the trained models.

Study population

The study included all children aged 18 and under hospitalized from January 1, 2017, to December 31, 2023, excluding patients outside the specified age range, records with missing cost data, or incomplete ICD-10 codes. Furthermore, measures were taken to identify and manage cost outliers through trimming procedures to uphold data integrity. The ethics of this study were approved by the Shanghai Health Development Research Center (approval number 2024001) and all the methods were performed in accordance with the Declaration of Helsinki.

Candidate predictors

The main predictive factors inputted into the model in this study consist of four categories, totaling 12 variables. The first category includes patient characteristics such as age categorized into five groups, gender, household registration status, health insurance type, and whether they belong to the Top 20 HNHC Diseases. The second category involves characteristics related to healthcare service utilization, including whether surgery was performed during hospitalization, length of hospital stays, and annual number of hospitalizations. The third category pertains to healthcare expenditure characteristics and previous research indicates a strong correlation between HNHC status in the previous year and the subsequent year²⁰, thus annual expenditure is computed to determine if patient costs rank within the top 5% as previous HNHC Status. The fourth category encompasses hospital characteristics such as hospital level, hospital type, and hospital region.

Outcomes

The outcome measured in this study was becoming an HNHC patient in the subsequent year, defined as individuals who rank within the top 5% of annual hospitalization costs. For the training dataset spanning 2021–2022, the predictive model assessed whether patients would become HNHC in 2022. For the external validation dataset spanning 2022–2023, the predictive model assessed whether patients would become HNHC in 2023.

Statistical analysis

Machine-learning-based models

This study employed five machine-learning-based models to predict HNHC pediatric patients: (1) logistic regression (used as the reference model)^{30,33}: A statistical model that predicts the probability of a binary outcome $P(y=1 | X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\dots+\beta_n X_n)}}$ based on linear relationships between predictor variables and the logarithm of odds, where X_1, X_2, \dots, X_n are predictor variables, and $\beta_1, \beta_2, \dots, \beta_n$ are model coefficients; (2) k-nearest neighbors' regression (KNN)³⁴: A non-parametric method that predicts the class of a query point by finding the majority class among its k nearest neighbors. The Euclidean distance is commonly used to determine nearest neighbors, defined as: $d(X_i, X_j) = \sqrt{\sum_{m=1}^n (X_{im} - X_{jm})^2}$, where X_i and X_j represent two points in the feature space; (3) random forest (RF)³⁵: An ensemble method that constructs multiple decision trees during training and outputs the mode (classification) or average prediction (regression) of the individual trees. The prediction function for a random forest can be written as: $\hat{f}(X) = \frac{1}{T} \sum_{t=1}^T f_t(X)$, where T is the number of trees and $f_t(X)$ is the prediction of the t -th tree; (4) Multi-layer Perceptron (MLP)³⁶: A type of artificial neural network composed of multiple layers of nodes and uses backpropagation for training, where each layer is fully connected to the next. The output of a single neuron is computed as: $z^{(l)} = \sigma(W^{(l)} x^{(l-1)} + b^{(l)})$, where $W^{(l)}$ is the weight matrix, $b^{(l)}$ is the bias, $\sigma(\cdot)$ is the activation function, and $x^{(l-1)}$ is the input from the previous layer; and (5) Naive Bayes³⁷: A probabilistic classifier, which was based on Bayes' theorem with the "naive" assumption of independence between features, making it efficient and effective for large datasets with categorical features. The posterior probability is calculated as: $P(C|X) = \frac{P(X|C)P(C)}{P(X)}$, where C is the class label, X is the feature set, and $P(X|C)$ assumes independence among features for computational efficiency.

Model development

This study used 70% of the data from 2021–2022 as a training set, employing the predictive factors of 2021 patients as model input parameters to construct models for predicting HNHC. Based on the initial model development, the remaining 30% of the 2021–2022 data served as a validation set to evaluate model performance. This process refined the training outcomes of the predictive model based on the actual HNHC cases in 2022, facilitating the selection of the optimal model. To address the issue of class imbalance between HNHC and non-HNHC groups within the dataset, internal class balancing was implemented to avoid the biased model performance particularly towards the majority class of HNHC³⁸. The approach of the Synthetic Minority Over-sampling Technique (SMOTE) ensures that the classifier is exposed to a more equitable distribution of samples, enhancing its ability to generalize across both majority and minority classes³⁹.

A grid search strategy was employed with k-fold cross-validation to optimize hyperparameters for the RF and KNN models^{25,40}. The MLP model, a machine-learning algorithm, utilizes multiple layers to effectively model nonlinear relationships between predictors and outcomes^{38,41}. The final models were also developed by manual tuning of the hyperparameters (i.e. the number of layers, etc.). In k-fold cross-validation, the dataset is divided into 'k' equal-sized subsets or folds to train the models 'k' times, each time using 'k-1' folds for training and the remaining fold for testing, which ensures that each data point is used for both training and testing, providing a more robust estimate of model performance⁴⁰. After primarily evaluating the performance across different values of $k=5$, $k=10$, and $k=20$, $k=10$ was selected for often providing a good balance between bias and variance.

Model assessment

The study assessed each model's predictive performance using the following 5 metrics. (1) ROC-AUC curve: A graphical plot that illustrates the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) across various threshold settings, providing a comprehensive assessment of a model's discriminatory ability. AUC represents the area under the receiver operating characteristic curve; (2) accuracy: The proportion of correct predictions (both true positives and true negatives) made by the model among all predictions, reflecting overall correctness, as shown in Eq. (1); (3) sensitivity: Known as the true positive rate (TP), sensitivity measures the proportion of actual positives (HNHC patients) correctly identified by the model, indicating its ability to detect true value of the positive cases and avoid false negatives (FN), as shown in Eq. (2); (4) specificity: The true negative rate (TN) measures the proportion of actual negatives correctly identified by the model among all negative instances, highlighting its ability to avoid false positives (FP), as shown in Eq. (3); (5) F1 score: The harmonic mean of precision and recall, F1 score balances both metrics and provides a single metric that combines precision (ability of the model to not label a negative sample as positive) and recall (ability of the model to find all positive samples), which is useful when there is an uneven class distribution, as shown in Eq. (4). The calculation formulas for the above metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

$$F1\text{score} = 2 \times \frac{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \quad (4)$$

Considering that the cohort of HNHC patients in this study comprised 91,882 children, which is significantly larger than the number of parameters used for prediction (12 parameters), the risk of overfitting was deemed low. The event per variable (EPV) for all models exceeded 80, which is well above the threshold of $EPV < 20$ typically associated with overfitting risks⁴².

Robustness analysis

The external validation dataset 2022–2023 was used to assess the robustness of the trained models. These prediction models were validated using predictors in the 2022 data and the outcome in the 2023 data. For internal validation, changing the 70:30 train-test ratio in the main models, additional experiments were conducted using different train-test ratios, namely 80:20 and 90:10²⁹, to assess the robustness of the models. The same 5 metrics were used to assess each model's predictive performance.

In the robustness analysis section, additional metrics were also incorporated across various train-test ratios to provide a more comprehensive evaluation of model performance, including: (1) the Mean Squared Error (MSE) quantifies the average squared difference between predicted and actual values, emphasizing larger errors, while the Root Mean Squared Error (RMSE) is the square root of MSE, providing an error metric in the same units as the target variable, facilitating more intuitive interpretation of model performance; (2) the Positive Likelihood Ratio (PLR) and Negative Likelihood Ratio (NLR) are statistical measures used to evaluate the diagnostic performance of a test, providing an assessment of how well a test can distinguish between the presence and absence of a condition; (3) the Diagnostic Odds Ratio (DOR) is a single indicator of test performance, defined as the ratio of the odds of a positive test result in individuals with the condition to the odds of a positive result in those without the condition; a higher DOR signifies better discriminatory power of the model; and (4) the Matthews Correlation Coefficient (MCC) is a measure of the quality of binary classifications, taking into account true and false positives and negatives; it returns a value between -1 and 1, where 1 indicates perfect predictions, 0 indicates random predictions, and -1 indicates total disagreement between prediction and observation.

These models were developed and optimized with Sci-kit Learn in Python version 3.0.

Results

Characteristics of the study sample

The study included a total of 91,882 hospitalized children in 2021. These children were categorized into HNHC ($n = 4,595$) and non-HNHC ($n = 87,287$) groups based on whether their annual medical costs ranked in the top 5%, and their predictor variables were described and compared accordingly (Table 1).

Compared to non-HNHC children, HNHC children were more likely to be under 1 year old or over 13 years old, female, Shanghai residents, and dual-insured (URBMI + CHMF). Analysis of diseases among HNHC children from 2017 to 2023 indicated that The top 20 diseases identified in the study can be categorized into five groups: (1) common neonatal diseases (e.g., preterm birth, neonatal respiratory distress syndrome), (2) traumas and limb deformities (e.g., fractures, dislocations, sprains), (3) respiratory diseases (e.g., pneumonia), (4) malignant neoplasms, and (5) cardiovascular, cerebrovascular, and rare diseases (e.g., congenital heart malformations, hemophilia, Crohn's disease) (Table 2). The results revealed that 49.14% had one of the top 20 HNHC diseases, a significantly higher rate than that of non-HNHC children. Regarding healthcare service utilization, 91.64% of HNHC children underwent surgery (74.94% of non-HNHC), with a longer average hospital stay (34.15 days) and more frequent annual hospitalizations (2.86 times). In terms of hospital characteristics, HNHC children were more likely to be hospitalized in tertiary hospitals located in the city center, either in general or specialized hospitals.

Between 2021 and 2023, the total hospitalization spending and HNHC patient spending experienced slight fluctuations. In 2021, the median annual hospitalization spending among pediatric patients was 190.75 million dollars and 67.50 million dollars was spent on 5% HNHC patients. In three years, the proportion of total spending attributed to HNHC patients remained stable, ranging between 35 and 36% (Fig. 1).

Prediction of HNHC patients

The prediction abilities of the reference model and four machine-learning-based prediction models for HNHC patients are presented in Table 3, including ROC-AUC, accuracy, sensitivity, specificity, and F1-score for each model. There were no predictors with high variance inflation factors ($VIF > 10$) among the parameters included in the reference model, indicating that collinearity is not a concern for the prediction models. Based on the metrics in the majority of previous studies^{19,20,22,26,28,31,43}, this study utilized the ROC curve as the criterion for selecting the optimal model, as demonstrated in Table 3 and Fig. 2.

Four machine-learning-based models had a higher discriminative ability than the logistic regression model, which was the reference model ($ROC\text{-}AUC = 0.827$, $F1\text{-}score = 0.127$). The MPL model (0.872) had the highest ROC-AUC of 0.869, followed by the random forest model (0.869), the KNN model (0.836), and the Naive Bayes model (0.828). The Naive Bayes model had the highest sensitivity (0.567) and the specificity of the random

Predictors	HNHC		Non-HNHC	
	n (%)		n (%)	
Patient characteristics				
Age***				
< = 30d	732	15.93	12,443	14.26
> 30d to < 1y	192	4.18	1,864	2.14
1-4y	989	21.52	32,363	37.08
5-12y	974	21.20	28,001	32.08
13-18y	1,708	37.17	12,616	14.45
Sex**				
Girls	1,927	41.94	35,051	40.16
Boys	2668	58.06	52,236	59.84
Household register***				
Shanghai	3,641	79.24	67,151	76.93
Others	954	20.76	20,136	23.07
Health insurance type***				
URBMI + CHMF	3,687	80.24	65,980	75.59
URBMI	446	9.71	8,757	10.03
CHMF	462	10.05	12,550	14.38
Top 20 HNHC Disease***				
No	2,337	50.86	69,143	79.21
Yes	2,258	49.14	18,144	20.79
Healthcare service utilization characteristics				
Operation***				
No	384	8.36	21,874	25.06
Yes	4,211	91.64	65,413	74.94
Length of stay, mean (std)***	34.15	46.76	4.81	4.99
Number of hospitalizations, mean (std)***	2.86	3.77	1.12	0.51
Healthcare expenditure characteristics				
Previous HNHC Status***				
No	0	0.00	87,287	100.00
Yes	4,595	100.00	0	0.00
Hospital characteristics				
Hospital level***				
Primary	3	0.07	28	0.03
Secondary	191	4.16	14,844	17.01
Tertiary	4,401	95.78	72,415	82.96
Hospital type***				
General hospital	2,246	48.88	41,831	47.92
Specialized hospital	1,905	41.46	33,141	37.97
Others	444	9.66	12,315	14.11
Hospital region***				
Suburb	1,739	37.85	41,315	47.33
City center	2,856	62.15	45,972	52.67

Table 1. Predictor variables of the study populations in 2021 (n = 91,882). The symbols in the table represent significant differences between the two groups with * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

forest model, the MPL model and the KNN model was 1.000. Considering the low risk of overfitting, the event per variable (EPV) for prediction models was posited and all EPVs for all models were over 80 (EPV < 20 is indicative of potential overfitting), which further supported the robustness and generalizability of these models.

The ROC curve results for the prediction models are shown in Fig. 2, where the x-axis represents sensitivity and the y-axis represents specificity, and the corresponding values of the area under the ROC curve for each model are presented in Table 3. Compared to the red line representing the logistic regression model, the other machine learning models demonstrate higher predictive power. The Multi-layer Perceptron model exhibits the highest predictive performance, as indicated by the green line with the largest area under the ROC curve.

The Losses or Error Graph for the prediction models are shown in Fig. 3, where the y-axis represents different error rates for the corresponding models. As shown in the Losses or Error Graph, the highest error rate was observed for the Naive Bayes model, while the MLP maintained the lowest error rate—lower than those of RF,

Disease	ICD10	HNHC patients		Other patients	
		n	%	n	%
Preterm birth	O60	2,950	8.57	11,998	1.84
Congenital malformation of cardiac septa	Q21	2,373	6.90	655	0.10
Fracture of lower leg	S82	1,217	3.54	2,045	0.31
Pneumonia (unspecified organism)	J18	1,198	3.48	144,103	22.04
Dislocation, sprain, and strain of knee	S83	1,194	3.47	1,254	0.19
Neonatal respiratory distress syndrome	P22	1,191	3.46	4,878	0.75
Fracture of forearm	S52	962	2.80	4,170	0.64
Fracture of shoulder and upper arm	S42	923	2.68	5,483	0.84
Malignant neoplasm	C95	772	2.24	702	0.11
Epilepsy	G40	504	1.46	8,916	1.36
Disorders related to short gestation and low birth weight	P07	486	1.41	927	0.14
Congenital malformation of aorta	Q25	439	1.28	433	0.07
Acquired deformities of musculoskeletal system/connective tissue	M95	431	1.25	401	0.06
Hemophilia (Factor VIII deficiency)	D66	348	1.01	173	0.03
Other disorders of brain	G93	338	0.98	324	0.05
Pneumothorax	J93	335	0.97	1,039	0.16
Paroxysmal tachycardia	I47	333	0.97	310	0.05
Other congenital malformations of heart	Q24	330	0.96	247	0.04
Crohn's disease	K50	328	0.95	579	0.09
Acquired deformities of fingers and toes	M21	324	0.94	354	0.05

Table 2. The top 20 most prevalent diseases among HNHC patients during 2017–2023.

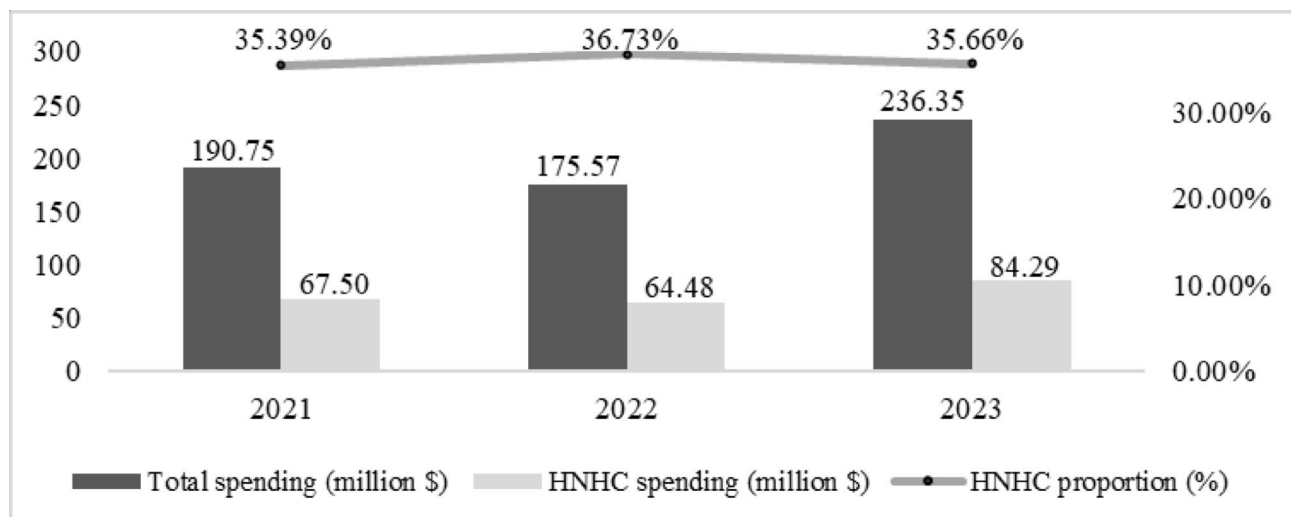


Fig. 1. Distribution of annual hospitalization spending among children 2021–2023.

Outcome	ROC-AUC	accuracy	sensitivity	specificity	F1-score
Logistic regression	0.827	0.992	0.074	0.999	0.127
Naive bayes	0.828	0.950	0.567	0.953	0.150
KNN	0.836	0.992	0.012	1.000	0.159
RF	0.869	0.992	0.098	0.999	0.167
MLP	0.872	0.992	0.042	1.000	0.138

Table 3. Prediction ability of models for HNHC patients (Train-test Ratio of 70:30).

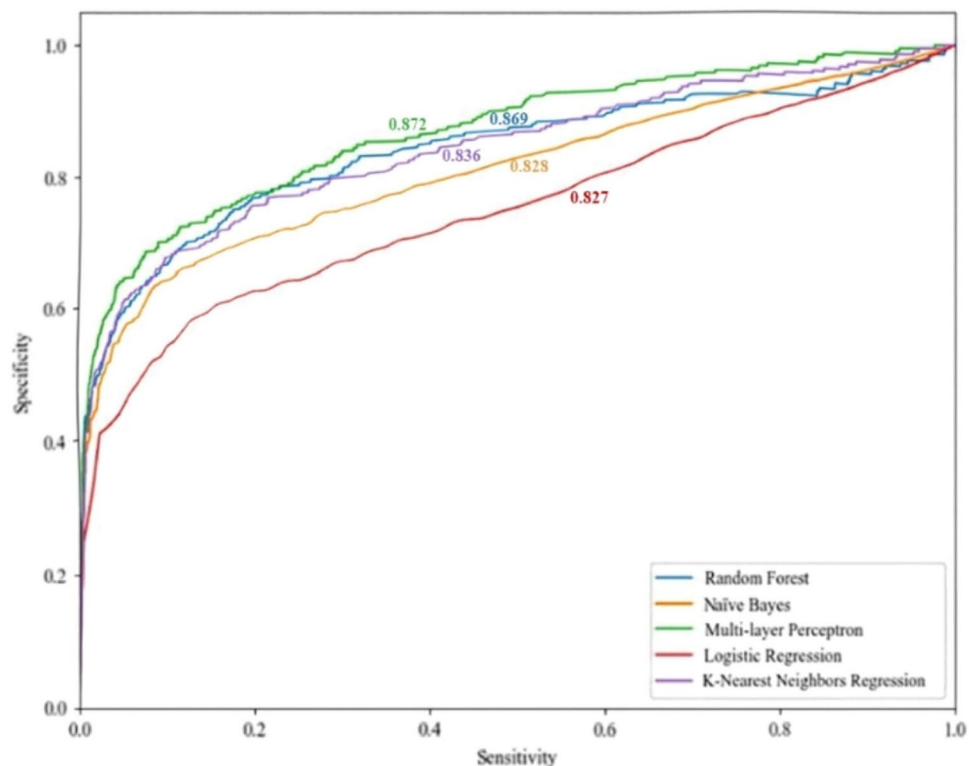


Fig. 2. Receiver-operating characteristics (ROC) curves.

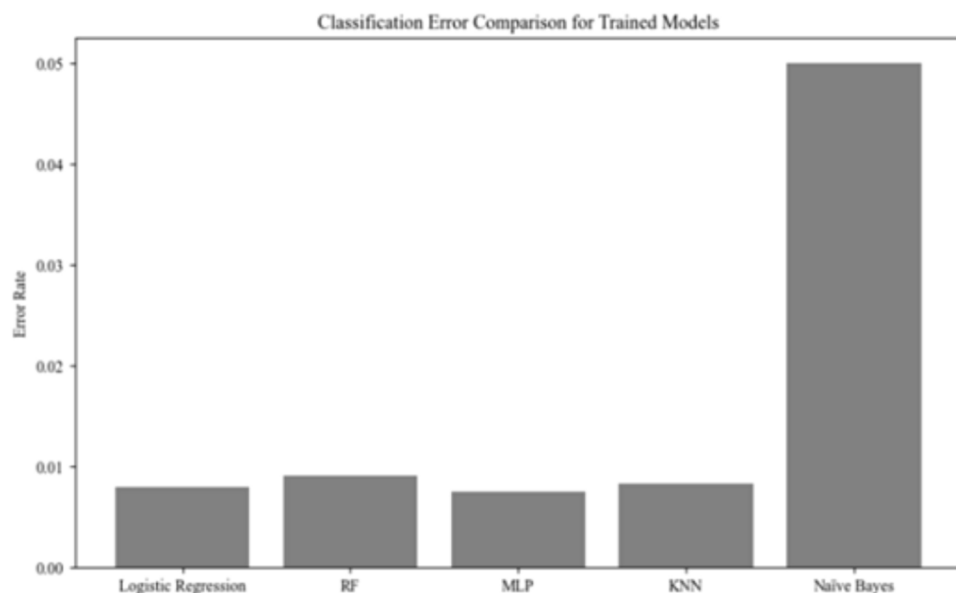


Fig. 3. The losses or error graph among different prediction models.

Logistic Regression, and KNN. The MLP model achieved high predictive performance while maintaining the lowest error rate during training, indicating superior model performance.

Predictors importance

Based on the results of the MLP model, the SHAP (SHapley Additive exPlanations) plots were utilized to visually elucidate machine learning model predictions by quantifying the contribution of each feature in Fig. 4. In Fig. 4, the impact of 12 indicators on model predictions was identified to aggregate feature importance across all variables in the summary SHAP plot by assigning a SHAP value to each feature relative to a baseline prediction

(often the dataset's mean), and the magnitude of SHAP values indicating the extent to which features influence the model's predicted output. In the MLP model, the factors that most significantly influence the predictive results are length of stay, number of hospitalizations, previous HNHC status, age, and presence of Top 20 HNHC diseases.

Robustness analyses

The results for robustness analyses are presented in Table 4. Using these five models, the study predicted HNHC for 2023 based on relevant characteristics of a total of 77,774 hospitalized children in 2022 to assess the discriminative ability of the models based on actual outcomes, and additional experiments using different train-test ratios, namely 80:20, and 90:10 were conducted to assess the robustness of the models.

In the external validation dataset, the predictive performance of all models showed a decrease compared to the main results. However, the MLP model remained the best-performing model (ROC-AUC of 0.843, accuracy of 0.986, sensitivity of 0.122, specificity of 0.999, F1-score of 0.203). Additionally, the MLP model's performance was significantly higher than that of the reference logistic regression model (ROC-AUC of 0.808). The external validation results indicate the robustness of the optimal model identified in the main analysis. The model trained using MLP demonstrates effective prediction of HNHC occurrence among hospitalized children.

After adjusting the train-test ratio of the internal validation data to 80:20 and 90:10, the predictive performance of all models demonstrated robustness and consistency compared to the main results. The MLP model continued to exhibit the best fit (ROC-AUC: 0.872 in 80:20 ratio; 0.883 in 90:10 ratio), significantly outperforming the logistic regression (ROC-AUC: 0.826 in 80:20 ratios; 0.807 in 90:10 ratios). There was also no substantial change in ranking for the prediction ability of other models in 80:20 and 90:10 ratios, which kept the same with the main result in 70:30 ratios (MLP > RF > KNN > Naive Bayes > logistic regression). The alteration of the train-test ratio for the internal validation data had minimal impact on the primary results.

In the robustness analysis, additional metrics including RMSE, MSE, PLR, NLR, DOR, and MCC were incorporated across different train-test ratios (Table 5). Results indicate that both RMSE and MSE remain within acceptable ranges; aside from the Naïve Bayes model, the performance of the other models is consistent across all ratios. Moreover, favorable PLR and DOR values are observed—particularly under the primary 70:30 split, where high PLR values are recorded and the MLP model maintains optimal predictive performance.

Discussion

Using retrospective cohort data of hospitalized children in Shanghai, China, this study found that hospitalization costs for HNHC pediatric patients accounted for over 35% of total spending, consistent with findings from studies in other countries^{3,44}. Machine learning models outperformed the logistic regression model in predicting HNHC patients for the following year, with the MLP model showing the highest predictive efficacy. The results of external validation from 2022–2023 data and the internal validation using different train-test ratios (80:20 and 90:10) corroborated the main findings, affirming MLP as the optimal model and demonstrating robustness.

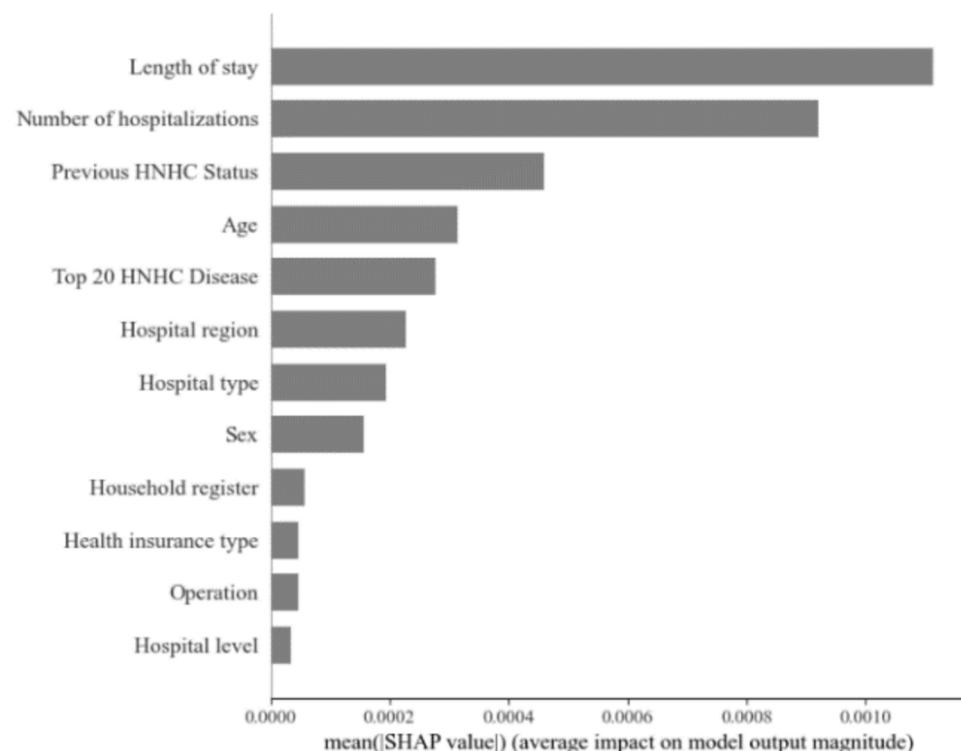


Fig. 4. Importance of each predictor in the MLP model.

Outcome	ROC-AUC	accuracy	sensitivity	specificity	F1-score
External validation					
Logistic regression	0.808	0.986	0.053	1.000	0.097
KNN	0.686	0.985	0.090	0.998	0.150
RF	0.757	0.985	0.101	0.998	0.165
Naive bayes	0.813	0.950	0.531	0.956	0.234
MLP	0.843	0.986	0.122	0.999	0.203
Train-test ratio of 80:20					
Logistic regression	0.826	0.992	0.074	0.999	0.125
Naive bayes	0.831	0.950	0.567	0.953	0.161
KNN	0.838	0.992	0.012	0.999	0.159
RF	0.869	0.992	0.091	0.999	0.167
MLP	0.872	0.992	0.096	1.000	0.167
Train-test ratio of 90:10					
Logistic regression	0.807	0.991	0.082	0.999	0.140
Naive bayes	0.841	0.951	0.567	0.954	0.174
KNN	0.855	0.991	0.016	0.999	0.160
RF	0.873	0.992	0.094	0.999	0.170
MLP	0.883	0.990	0.306	0.996	0.356

Table 4. Robustness analyses of prediction ability for HNHC patients.

Outcome	ROC-AUC	RMSE	MSE	PLR	NLR	DOR	MCC
Train-test ratio of 70:30							
Logistic regression	0.827	0.09	0.01	95.24	0.91	104.23	0.20
Naive bayes	0.828	0.20	0.05	13.19	0.41	31.92	0.24
KNN	0.836	0.09	0.01	73.36	0.86	84.84	0.23
RF	0.869	0.09	0.01	77.19	0.87	88.86	0.23
MLP	0.872	0.09	0.02	101.65	0.79	129.34	0.32
Train-test ratio of 80:20							
Logistic regression	0.826	0.09	0.01	121.74	0.89	136.19	0.23
Naive bayes	0.831	0.20	0.04	14.10	0.37	38.19	0.25
KNN	0.838	0.09	0.01	70.77	0.87	81.39	0.22
RF	0.869	0.09	0.01	57.29	0.90	63.59	0.18
MLP	0.872	0.09	0.01	85.93	0.89	96.78	0.22
Train-test ratio of 90:10							
Logistic regression	0.807	0.09	0.01	109.37	0.88	124.15	0.24
Naive bayes	0.841	0.20	0.04	13.92	0.38	36.89	0.24
KNN	0.855	0.09	0.02	68.36	0.88	77.54	0.20
RF	0.873	0.09	0.01	75.95	0.87	87.48	0.22
MLP	0.883	0.09	0.01	78.63	0.85	91.97	0.24

Table 5. Robustness analyses of model performance metrics comparison across different train-test ratios. PLR, the positive likelihood ratio; NLR, negative likelihood ratio; DOR, diagnostic odds ratio; MCC, Matthews correlation coefficient; RMSE, the Root Mean Square Error; MSE, Mean Square Error.

Taken together, this study provides foundational evidence for the status of HNHC children in China and fills a gap in the international literature on predictive modeling for HNHC children.

The study found significant differences between the HNHC and non-HNHC groups in key predictive factors based on the results describing both groups, consistent with previous studies^{32,44}. HNHC children were more likely to be under 1 year old or over 13, female, and residents of Shanghai with dual insurance (URBMI + CHMF). They also showed higher rates of severe diseases and more frequent use of healthcare services, including surgeries, longer hospital stays, and more frequent annual hospitalizations. Additionally, HNHC children were more often hospitalized in tertiary hospitals in urban areas. At the same time, the top 20 most common diseases among HNHC children were identified, distinguishing the causes of HNHC in children from those in adults from a disease perspective^{18,20}. The top 20 diseases identified in the study can be categorized into five groups, including common neonatal diseases, traumas and limb deformities, respiratory diseases, malignant neoplasms, and

cardiovascular, cerebrovascular, and rare diseases. This categorization provides valuable insights for pediatric resource allocation in clinical settings and serves as a strong predictive factor for subsequent model analysis.

Numerous studies have been conducted to identify prospective HNHC adult patients using traditional methods like logistic regression models, which have been proven to have lower predictive power compared to machine-learning-based models^{18–22,43}. In predictive models for HNHC patients, other studies have identified the following best-performing methods and their predictive efficacy (often measured by AUC): XGBoost model (AUC: 0.790⁴³ and 0.801³¹), Smooth Bayesian network model (AUC: 0.840²⁸), neural networks (AUC: 0.840²⁰), RF model (AUC: 0.883¹⁹), and LightGBM (AUC: 0.900⁴³). For different populations, datasets, and model constructions, the optimal type of model may vary. However, the AUC metric remains a reliable measure for assessing the predictive efficacy of different models within the same study or for cross-study comparisons^{19,20,22,26,28,31,43}. Generally, an AUC greater than 0.8 indicates good predictive performance of the model⁴⁰, and in this study, the MLP model achieved an AUC of 0.872, which is similar to the predictive performance observed in previous adult HNHC models. Additionally, when selecting machine learning models for large-scale datasets, understanding the time complexity and computational efficiency is crucial, as the measures of how fast or slow an algorithm will perform for the input size⁴⁵. Models such as logistic regression, Naive Bayes, and KNN exhibit linear time complexity concerning the number of samples and features, making them computationally efficient for both training and prediction^{46,47}. In contrast, models like RF and MLP often offer enhanced predictive performance, which may require more computational resources due to their complexity and parameter tuning, with additional computational cost depending on the application's requirements⁴⁸.

In the analysis of predictor importance for prediction models, it was found that incorporating previous year healthcare service utilization (length of stay and number of hospitalizations) and HNHC status significantly enhances the predictive power of the model. Prior studies have shown strong correlations between healthcare costs and utilization from the previous year and those in the subsequent year^{8,20,49}, which supports the notion that adding these relevant factors substantially improves the predictive performance of the model. Based on a retrospective cohort study spanning 2017 to 2023, the top three-digit ICD-10 codes of HNHC patients were ranked to identify the highest prevalence of diseases within the top 20. This process enabled us to pinpoint the pediatric conditions posing the highest risk of HNHC, demonstrating substantial differences between pediatric and adult HNHC conditions^{18,19}. Thus, enabling early patient classification based on these findings as predictive factors inputted into the forecasting model, which was the top 4 important predictor, suggesting that clinical practice should focus on these potential population to optimize services and prevent the occurrence of high healthcare costs.

Moreover, few studies validate their constructed models using external datasets, and even prior studies have primarily conducted sensitivity analyses focused on defining HNHC populations, essentially using a single set of training data for analysis²⁰. This study innovatively employed external data for 77,774 children from 2022–2023 to validate the robustness of the models, which confirmed the MLP model as optimal, with sustained high efficacy (ROC-AUC: 0.843), and remained consistent with the main result. However, the external validation results demonstrated a slight decline in performance across all models compared to the main results. This decrease might be attributed to several factors, including differences in patient demographics, clinical practices, and data collection methods between the training and external datasets, as well as temporal variability in healthcare practices and patient characteristics⁵⁰. Also, a degree of overfitting to the internal data may have contributed to the reduced generalizability³⁸. To improve robustness across diverse datasets, future studies should incorporate more heterogeneous external data from multiple institutions or regions, employ domain adaptation techniques to better align the training and external data distributions, and explore ensemble methods to reduce variance²¹. Based on the MLP model in this study, these strategies provide a pathway for refining predictive models and enhancing their applicability in real-world settings to analyze HNHC cases among hospitalized children in China.

The contributions of this study are reflected in two key aspects. From an academic perspective, this is the first study, to the best of current knowledge, to conduct HNHC prediction among pediatric populations in China, comparing multiple machine learning methods that outperformed traditional statistical models. This study found significant differences in key predictive factors, such as disease types, between children and adults in HNHC predictions, suggesting the need for specific factors when predicting HNHC pediatric patients^{18–20}. The use of internal and external validation enhances the robustness of the results and provides a reference for future research to further improve predictive accuracy using methods such as MLP. The utilization of the machine learning model in HNHC also raises ethical concerns relating to data privacy and the potential for discrimination, and it is essential to address these issues by improving data transparency and adopting a feasible approach for data collection in the future.

From a policy and practical application perspective, this study provides a predictive model for the early identification of HNHC pediatric patients, which can help optimize and improve healthcare service models that have been attempted but yielded limited success^{9,10,12,13,15}. In the future, the integration of the trained MLP model into clinical workflows can enable real-time identification and risk stratification of HNHC patients, facilitating early intervention and resource optimization. The model could be involved in integration with the hospital's electronic medical record (EMR) system and the claims database, as a decision-support tool, and be optimized by the feedback from clinicians and continuous monitoring to update the model's real-world utility. By pinpointing the optimal intervention window and initiating services early, the model contributes to preventing high healthcare costs.

This study has several limitations. Firstly, due to barriers to accessing medical data in China, it's difficult to link outpatient and inpatient data. More social, physiological, and biochemical information could not be obtained, which is a common data limitation in Chinese research. While this study endeavored to collect all factors that maybe influencing the predictive performance of HNHC in theory and incorporated them into the

predictive models to provide robust results within the available data scope, future research should investigate more predictors in improving predictive models of HNHC pediatric patients. Secondly, because machine-learning-based models rely on training data, the generalizability of this study's results to predict HNHC children in other countries or generalize to other populations (eg, older adults), different settings, or other data sources should be discussed with caution. This study used health insurance claims data for all hospitalized children in Shanghai, a major city in China, to reduce sampling bias and leverage the advantages of big data as possible. Future studies could further validate these findings and incorporate additional data sources to refine the model. Finally, due to the impact of COVID-19 lockdown policies, there may be some fluctuations in the 2021–2022 data. To address this issue, additionally external validation was conducted using a dataset from 2022–2023, and the ranking of the efficacy of the models remained robust, indicating the feasibility of using the MLP model for prediction.

Conclusion

In summary, this study revealed that the top 5% of HNHC pediatric patients accounted for over 35% of total hospitalization costs. Using a large retrospective cohort of 91,882 children for training and 77,774 for external validation to develop the machine learning models, the MLP indicated the best predictive performance. Key predictors of HNHC pediatric patients were identified, including length of stay, number of hospitalizations, previous HNHC status, age, and presence of Top 20 HNHC diseases. These findings underscore the critical importance of the integration of predictive models into clinical practice and policymaking to realize early identification of HNHC pediatric populations for optimizing healthcare resource allocation, facilitating targeted interventions, and controlling healthcare costs. In future studies, real-world longitudinal data from multiple regions, additional predictive factors related to sociodemographic characteristics and health behaviors, and more ensemble methods of comprehensive machine learning algorithms should be employed to enhance the predictive performance and generalizability of the models.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 20 January 2025; Accepted: 21 April 2025

Published online: 08 May 2025

References

1. OECD. Health at a Glance 2023[R]. OECD Publishing, 2023.
2. Agency for Healthcare Research and Quality. The high concentration of U.S. health care expenditures[EB/OL]. [2024–12–23]. <http://www.ahrq.gov/research/ria19/pendria.htm>.
3. Punjabi, N. et al. Categorising high-cost high-need children and young people. *Arch. Dis. Child.* **107**(4), 346–350 (2022).
4. Adriana Hernandez-Viver, Mitchell, E. M. Concentration of healthcare expenditures and selected characteristics of persons with high expenses, United States civilian noninstitutionalized population, 2018–2021. Agency for Healthcare Research and Quality, 2024.
5. Tanke, M. A. C. et al. A challenge to all: A primer on inter-country differences of high-need, high-cost patients. *PLoS ONE* **14**(6), 123–129 (2019).
6. Papanicolas, I. et al. Differences in health outcomes for high-need high-cost patients across high-income countries. *Health Serv. Res.* **56**, 1347–1357 (2021).
7. Leininger, L. J., Tomaino, M. & Meara, E. Health-related quality of life in high-cost, high-need populations. *Am. J. Manag. Care* **29**(7), 362–371 (2023).
8. Pourat, N. et al. Differences in health care utilization of high-need and high-cost patients of federally funded health centers versus other primary care providers. *Med. Care* **62**(1), 52–59 (2024).
9. Bailey, J. E. et al. Effect of intensive interdisciplinary transitional care for high-need, high-cost patients on quality, outcomes, and costs: A quasi-experimental study. *J. Gen. Intern. Med.* **34**(9), 1815–1824 (2019).
10. Brenner, J. Lessons learned from the Camden Coalition's work with high needs, high complexity patients. *Popul. Health Manag.* **26**(4), 227–229 (2023).
11. Komaromy, M. et al. A novel intervention for high-need, high-cost medicaid patients: A study of ECHO care. *J. Gen. Intern. Med.* **35**(1), 21–27 (2020).
12. Pestka, D. L. et al. Delivering care to high-cost high-need patients: lessons learned in the development of a complex care primary care team. *J. Prim. Care Commun. Health* <https://doi.org/10.1177/21501327211023888> (2021).
13. Quinton, J. K. et al. Differential impact of a plan-led standardized complex care management intervention on subgroups of high-cost high-need Medicaid patients. *Popul. Health Manag.* **26**(2), 100–106 (2023).
14. Smeets, R. G. M. et al. Person-centred and efficient care delivery for high-need, high-cost patients: primary care professionals' experiences. *BMC Fam. Pract.* <https://doi.org/10.1186/s12875-020-01172-3> (2020).
15. Wong, M. S. et al. Outcomes that matter: high-needs patients' and primary care leaders' perspectives on an intensive primary care pilot. *J. Gen. Intern. Med.* **36**(11), 3366–3372 (2021).
16. Chang, E., Ali, R. & Berkman, N. D. Unpacking complex interventions that manage care for high-need, high-cost patients: A realist review. *BMJ Open* **12**(6), 29–38 (2022).
17. de Ruijter, U. W. et al. Prediction models for future high-need high-cost healthcare use: A systematic review. *J. Gen. Intern. Med.* **37**(7), 1763–1770 (2022).
18. Choi, Y. et al. Development and evaluation of machine learning-based high-cost prediction model using health check-up data by the national health insurance service of Korea. *Int. J. Environ. Res. Pub. Health* **19**(20), 13672 (2022).
19. Langenberger, B., Schulte, T. & Groene, O. The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data. *PLoS ONE* **18**(1), e0279540 (2023).
20. Osawa, I. et al. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. *NPJ Dig Med* <https://doi.org/10.1038/s41746-020-00354-8> (2020).
21. Howson, S. N. et al. Improving the prediction of persistent high health care utilizers: Retrospective analysis using ensemble methodology. *JMIR Med. Inform.* **10**(3), e33212 (2022).

22. Yang, P. et al. Early prediction of high-cost inpatients with ischemic heart disease using network analytics and machine learning. *Exp. Syst. Appl* **210**, 118541 (2022).
23. Sharma, A. et al. A comparative study of heterogeneous machine learning algorithms for arrhythmia classification using feature selection technique and multi-dimensional datasets. *Eng. Res. Expr.* **6**(3), 35209 (2024).
24. Fu R, Coyte, P. C. IEEE. A machine learning approach to identify high-cost elderly renal transplant recipients: 2019 In *4th IEEE/ACM International Conference on Connected Health - Applications, Systems and Engineering Technologies (CHASE)*: 20195–6 (2019).
25. Dhanka S, Maini S. Random forest for heart disease detection: A classification approach In//: IEEE 1–3 (2021).
26. Ng, S. H. X. et al. Characterising and predicting persistent high-cost utilisers in healthcare: a retrospective cohort study in Singapore. *BMJ Open* **10**(1), e031622 (2020).
27. Barbosa, F. M., & Ishii, R. P. A Neural Network Approach to High Cost Patients Detection: Computational science and its applications, ICCSA 2021, PT III[Z]. GERVASI O, MURGANTE B, MISRA S, et al. In *21st International Conference on Computational Science and Its Applications (ICCSA)*: 2021: 12951, 527–540.
28. Lin, S. et al. Smooth Bayesian network model for the prediction of future high-cost patients with COPD. *Int. J. Med. Informatics* **126**, 147–155 (2019).
29. Dhanka S, Maini S. Multiple Machine Learning Intelligent Approaches for the Heart Disease Diagnosis In//: IEEE 147 152 (2023).
30. Dhanka, S., Bhardwaj, V. K. & Maini, S. Comprehensive analysis of supervised algorithms for coronary artery heart disease detection. *Expert. Syst.* **40**(7), e13300 (2023).
31. Luo, L. et al. Using machine learning approaches to predict high-cost chronic obstructive pulmonary disease patients in China. *Health Inform. J.* **26**(3), 1577–1598 (2020).
32. Zhang, P., Zhu, B. & Wang, L. Assessing differences among persistent, episodic, and non- high-need high-cost hospitalized children in China after categorization by an unsupervised learning algorithm. *Arch. Pub. Health* <https://doi.org/10.1186/s13690-024-0144-2> (2024).
33. Sci-kit learn. logistic regression[EB/OL]. <https://www.scikitlearn.com.cn/0.21.3/3/>.
34. Sci-kit learn. k-nearest neighbors' regression[EB/OL]. <https://www.scikitlearn.com.cn/0.21.3/7/>.
35. Sci-kit learn. random forest[EB/OL]. <https://www.scikitlearn.com.cn/0.21.3/11/>.
36. Sci-kit learn. Multi-layer Perceptron[EB/OL]. <https://www.scikitlearn.com.cn/0.21.3/18/>.
37. Sci-kit learn. Naïve Bayes[EB/OL]. <https://www.scikitlearn.com.cn/0.21.3/10/>.
38. Efthimiou, O. et al. Developing clinical prediction models: A step-by-step guide. *BMJ* **386**, e078276 (2024).
39. Chawla, N. V. et al. SMOTE: Synthetic minority over-sampling technique. *J. artif. Intell. Res.* **16**, 321–357 (2002).
40. Sci-kit learn. cross-validation[EB/OL]. <https://www.scikitlearn.com.cn/0.21.3/31/>.
41. Cao C, Liu F, Tan H, et al. Deep learning and its applications in biomedicine. *Genom. Proteom. Bioinform.* 2018.
42. Peduzzi, P. et al. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **49**(12), 1373–1379 (1996).
43. Nguyen, N. H. H. et al. Simplified machine learning models can accurately identify high-need high-cost patients with inflammatory bowel disease. *Clin. Transl. Gastroenterol.* **13**(7), e00507 (2022).
44. Gill, P. J. et al. Identifying conditions with high prevalence, cost, and variation in cost in us children's hospitals. *JAMA Netw. Open* **4**(7), e2117816 (2021).
45. Kukreja, M., Johnston, S. A. & Stafford, P. Comparative study of classification algorithms for immunosignaturing data. *BMC Bioinform.* **13**(1), 139 (2012).
46. Zhang, S. et al. Efficient kNN classification with different numbers of nearest neighbors. *IEEE Trans. Neural Netw. Learn. Syst.* **29**(5), 1774–1785 (2017).
47. Tsangaratos, P. & Ilia, I. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. *CATENA* **145**, 164–179 (2016).
48. Chugh, R. S., et al. A comparative analysis of classifiers for image classification: In *IEEE* 248 253 2020.
49. Yang, C. et al. Machine learning approaches for predicting high cost high need patient expenditures in health care. *Biomed Eng. Online* <https://doi.org/10.1186/s12938-018-0568-3> (2018).
50. Zhang, Y. & Kaushal, R. Method for optimizing classification of patient data from large patient datasets for identifying and classifying high-need high-cost patients. (2024).

Acknowledgements

Peng Zhang would like to thank my friend Zemin Xia for providing valuable resources for data retrieval.

Author contributions

P.Z: Data curation, Methodology, Software, Writing-Original draft preparation, Writing-Editing. B.Z: Conceptualization, Investigation, Validation, Writing- Reviewing. X.C.: Methodology, Software, Visualization. L.W: Supervision, Writing- Reviewing.

Funding

Science and Technology Commission of Shanghai Municipality, No.23YF1440900, No. 23692109000

Declarations

Competing interests

The authors declare no competing interests.

Consent statement

Due to the retrospective nature of the study and the data we used came from anonymized and secondary databases, the Shanghai Health Development Research Center (approval number 2024001) waived the need of obtaining informed consent.

Human ethics

The ethics of this study were approved by the Shanghai Health Development Research Center (approval number 2024001). All methods were carried out in accordance with relevant guidelines and regulations (Declarations of Helsinki, 1996).

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-99546-z>.

Correspondence and requests for materials should be addressed to L.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025