

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<u>http://bmjopen.bmj.com</u>).

If you have any questions on BMJ Open's open peer review process please email <u>info.bmjopen@bmj.com</u>

BMJ Open

Identifying the most influential factors in lung transplant patients using a multivariate prediction model

Journal:	BMJ Open
Manuscript ID	bmjopen-2024-089796
Article Type:	Original research
Date Submitted by the Author:	09-Jun-2024
Complete List of Authors:	Gholamzadeh, Marsa; Tehran University of Medical Sciences, Safdari, Reza; Tehran University of Medical Sciences, Asadi Gharabaghi, Mehrnaz; Tehran University of Medical Sciences, Abtahi, Hamidreza; Tehran University of Medical Sciences, Department of Pulmonary Medicine; Tehran University of Medical Sciences, Pulmonary and Critical Care Department
Keywords:	Machine Learning, Pulmonary Disease < Lung Diseases, Transplant medicine < INTERNAL MEDICINE





I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our <u>licence</u>.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which <u>Creative Commons</u> licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

terez oni

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies



2		
3	1	Identifying the most influential factors in lung transplant patients using a
4	Ŧ	ruchting the most innuclitial factors in fung transplant patients using a
5	•	multivariate mudiction model
7	2	multivariate prediction model
8		
9	3	Marsa Gholamzadeh ¹ Reza Safdari ² Mehrnaz Asadi Gharahaghi ³ Hamidreza Abtahi ^{3,4*}
10	5	
11		
12 12	4	1. Ph.D. in Medical Informatics, Health Information Management Department, School of Allied Medical Sciences,
14	5	 Professor of Department of Health Information Management, School of Allied Medical Sciences, Tehran University
15	7	2. Frotessor of Department of Hearth Information Management, School of Africa Medical Sciences, Tehran University
16	2 2	3 Department of Pulmonary Medicine, Faculty of Medicine, Tehran University of Medical Sciences, Tehran, Iran
17	9	4 Associate Professor, Pulmonary and Critical Care Department, Thoracic Research Center, Imam Khomeini Hospital
18	10	Complex. Tehran University of Medical Sciences. Tehran, Iran.
19	11	
20	12	
22		
23	13	*Corresponding author: Hamidreza Abtahi
24		
25	14	E-mail address: hrabtahi2020research@gmail.com
20 27		
28	15	Tel: +9821-66192646
29	16	Postal address: Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical
30	10	Tostal address. Thoracle Research Center, infant Rhomenn Hospital Complex, Tentan Oniversity of Wedlear
31	17	Sciences, Qarib Ave, Keshavarz Blv, Tehran, Iran.
32 22	4.0	
34	18	
35		
36	19	ORCID ID:
37	20	
38	20	Marsa Gholamzaden, <u>https://orcid.org/0000-0001-6781-9342;</u>
39 40	21	Hamidreza Abtahi, https://orcid.org/0000-0002-1111-0497
41		numurezu notum, <u>meps./oreid.org/0000/0002/1111/010/</u> ,
42	22	Reza Safdari, https://orcid.org/0000-0002-4982-337X
43	• •	
44	23	Mehrnaz Asadi Gharabaghi: https://orcid.org/0000-0003-0852-1532
45 46	24	
47	27	
48	25	
49	25	
50	26	Word count: 3728 words
51	27	
52 53	27	
54	28	
55	20	
56	29	
57		
58 50		1
60 60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1

59

2		
3 4	30	Abstract
5 6	31	Objectives: In lung transplantation, a priority is assigned to each candidate on the waiting list. Our
7 8	32	primary objective was to identify the key factors that influence the allocation of priorities in lung
9 10	33	transplantation using machine learning (ML) techniques to enhance the process of prioritizing
11 12 13	34	patients.
14 15	35	Design: Developing a prediction model.
16 17	36	Setting and participants: Our data was retrieved from the UNOS open-source database of
18 19	37	transplant patients between 2005 and 2023.
20 21 22	38	Interventions: After the preprocessing process, a feature engineering technique was employed to
23 24	39	select the most relevant features. Then, six ML models with an optimized hyper-parameter
25 26	40	including Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector
27 28 20	41	Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model, and a deep
30 31	42	learning model (DL) were developed under trained data.
32 33	43	Primary and secondary outcome measures: The performance of each model was evaluated
34 35 26	44	using R-squared (R ²) and other error rate indexes. Next, the Shapley Additive Explanations
30 37 38	45	(SHAP) technique was utilized to identify the most important features in the prediction.
39 40	46	Results: The raw dataset contains 196,270 records with 545 features. After preprocessing, 32,966
41 42	47	records with 15 features remain. Among various models, the RF model achieved a high R2 score.
43 44 45	48	Additionally, the RF model exhibited the lowest error values indicating its superior precision
46 47	49	compared to other regression models SHAP technique in conjunction with the RF model revealed
48 49	50	the 11 most important features for priority allocation. Subsequently, we developed a web-based
50 51 52	51	decision support tool using Python and the Streamlit framework based on the best-fine-tuned
53 54	52	model.
55 56		
57 58		2

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

Conclusion: The deployment of the ML model has the potential to act as an automated tool to aid .n. .n. alocation score, Machin physicians in assessing the priority of lung transplants and identifying significant factors that play a role in patient survival. Keywords: Lung transplantation, allocation score, Machine learning, Prediction. For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

58	S	trengths and limitations of this study:
59	•	Despite the potential benefits of using ML algorithms in medical sciences, there is a scarcity of
60		studies examining the use of such algorithms in lung transplantation and organ allocation.
61	•	The use of various preprocessing and data cleaning techniques in our survey increased the
62		robustness and performance of the model.
63	•	Understanding the factors influencing the determination of lung transplant priority could
64		support clinicians in designing treatment plans and thus improving the quality of life of patients.
65	•	Deploying the developed ML model in the form of a decision support system increases its
66		applicability in clinical practice.
67		
68		

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

1-Introduction

Lung transplantation (LTx) is an advanced treatment option for patients suffering from end-stage lung disease. When no other treatment options are available and the patient is likely to die, lung transplant surgery is suggested as a well-established treatment option [1]. When a patient meets the inclusion criteria for transplantation, they are placed on a waiting list and assigned a priority. Various conditions may affect eligibility for lung transplantation and the patient's priority [2]. In some countries, a score is assigned to each patient on the waiting list to enhance the recipient selection process [3, 4]. Understanding the most influential factors in priority allocation for lung transplantation is beneficial for researchers worldwide, as it can improve post-transplant survival. Utilizing data mining methods and developing forecasting models in this field could aid clinicians in uncovering hidden patterns and relationships within patient data and allocation scores.

Machine learning (ML) methods have been developed across various fields of clinical medicine to assist clinicians in predicting and classifying diseases [5]. These methods are used to predict the length of stay in the Intensive Care Unit (ICU), diagnose septic infection[6], and extract disease patterns from big data [7, 8]. Nevertheless, there is a lack of studies on the development of predictive models and identification of important features using ML methods to predict lung transplantation priority [9, 10].

Thus, the primary objective of this study was to utilize ML techniques to identify the most influential factors that strongly impacted outcomes based on various developed ML methods to predict the priority using clinical and demographic data.

2- Methods

90 Throughout this section, the process of developing, comparing, and evaluating ML models is
91 shown schematically in Fig 1. Python programming language version 10 was used in this study
92 for developing and validating ML algorithms. For data preprocessing, Numpy and Pandas modules

were employed, while the sci-kit learn library was utilized for developing supervised classifier algorithms. 2-1-Dataset description and data retrieval The data for this study were obtained from the United Network for Organ Sharing (UNOS) online database [11]. Upon receiving written permission from UNOS, we accessed the recorded data pertaining to lung transplantation for our research. Our study included patients over 18 years old with end-stage lung disease who underwent lung transplants between 2005 and 2022. We performed a waiting list analysis using all available data entries from the United Network for Organ Sharing (UNOS) database for our study. The priority of candidates on the waiting list was considered as the outcome, while the clinical and demographic characteristics of patients were considered as features or predictors. **2-2-Pre-processing process** Data pre-processing is a crucial step in ML techniques, especially when dealing with raw data from clinical databases or medical records that often contain missing or unclear information. To ensure the development of more accurate models based on appropriate data, we followed a series of data pre-processing steps. The following steps were employed in this phase as pre-processing techniques. 1- Checking the duplicated values and records to remove the duplicates

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

and data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text

- 111 2- De-identify records and remove irreverent features
- 112 3- Convert nominal and categorical features to numerical values
- ⁹ 113 4- Identify missing data and missing values imputation
- 57 114 5- Outlier detection
 - 115 6- Feature engineering and feature selection

116 7- Data transformation and normalization

117 Duplicate checking and removal of irreverent features: After duplicate checking, we consulted
118 UNOS guidelines and experts to review all features and their definitions. Under their supervision,
119 we removed identification variables and irrelevant features, such as ID columns, hospital center
120 identification codes, and country of residence, to de-identify patients.

Following this, we converted the post-transplant survival days variable to years and excluded
patients with a survival rate of less than two years. Next, we filtered out patients below 18 years
of age and excluded any data before 2005. Additionally, records related to heart transplantation
were removed from the dataset.

After removing irrelevant features in the first data-cleaning phase, we utilized the discretized
operator to convert nominal values to numerical data. Categorical data were encoded using the
LabelEncoder class too.

Missing data management: To address missing data in our dataset, we conducted missing data imputation across the entire dataset. Initially, we assessed the specified columns or attributes to determine the extent of missing and unique data in each column. During this analysis, we discovered that the ICU column was empty and decided to delete it due to its lack of meaningful information. To impute the missing data, a threshold of 60% was set for feature removal. As a result, any column with more than 60% missing data was removed. For the missing data imputation, we adopted a strategy of replacing missing data in numerical features with the mean value of each respective feature. This approach allows us to retain the integrity of the dataset while minimizing the impact of missing data on our analysis. By performing these comprehensive steps of missing data imputation, we ensure the dataset is optimized for further analysis and modeling, enabling us to draw more accurate conclusions and insights.

Page 9 of 29

BMJ Open

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

Outliers handling: To address outliers in the dataset, we first create distribution plots to visualize
the data. Next, we apply the IQR method and use Box plots to identify outliers. Finally, we remove
these outliers to prepare the data for further processing.

Feature engineering and feature selection: Since a high-dimensional dataset was utilized in this study, feature engineering should be employed to reduce the dimensionality of the features and enhance model performance. Feature selection is the process of identifying relevant features while eliminating irrelevant and redundant ones, aiming to derive a subset of features that effectively describe the problem with minimal loss of efficiency[12].

As the first step in this phase, correlation analysis was conducted to assess the relationships between features and target variables. This analysis helps identify highly correlated features that can aid in feature selection and model development. Subsequently, a combination of filtering and embedded techniques, including variance threshold and XGBoost methods were utilized to select the most pertinent features for the model and enhance its performance. After carefully selecting the features, the most effective and appropriate features remained as predictors for modeling.

Data transformation and normalization: In the end, data normalization was carried out to optimize the features for modeling purposes.

0 155 2-4- Model development and tuning

The objective of this study was to develop a prediction model for a continuous numerical variable (priority score) using regression techniques to identify the most effective factors in selecting the most appropriate candidate for LTx. In this study, regression models were selected to examine the connection between input variables and output numerical values, as the target variable (outcome) is a continuous numerical value.

Page 10 of 29

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

> During the model development process, the dataset was divided into training and testing data sets in an 80:20 ratio. To determine the most influential factors and identify the best model, the performance of six regression-based models was evaluated: Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model (MLP—a class of feedforward artificial neural network), and a deep learning model (DL). The selection of these models was done based on the type of target variable and the study objectives.

A hyperparameter tuning optimization technique was employed in this phase to improve model performance by optimizing the training process by determining the best hyperparameters for each model. This technique was used to prevent models that underfit or overfit the data [13]. After tuning parameters in each model, the models were trained with updated best hyperparameters, and all metrics were calculated again to achieve the best performance. We employed the random search method, a hyperparameter tuning technique where hyperparameters are randomly chosen from a predefined set to train a model.

2-4-1- Multiple linear regression (MLR)

Multiple linear regression (MLR) is a statistical technique used to estimate the relationship between a dependent variable and one or more independent variables. It is an extension of linear regression, which requires more than one predictor variable to forecast the response variable[14]. MLR is a significant regression algorithm that models the linear association between a dependent continuous variable and multiple independent variables [15]. Hence, we have chosen this model to predict the continuous variable (priority score) based on several independent variables. The equation for multiple linear regression is demonstrated below[15]:

 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 (1)$

I	
2	
3	
4	
5	
6	
-	
/	
8	
9	
10	
11	
12	
12	
13	
14	
15	
16	
17	
10	
10	
19	
20	
21	
22	
23	
23	
24	
25	
26	
27	
28	
20	
20	
30	
31	
32	
33	
34	
25	
22	
30	
37	
38	
39	
40	
л <u>т</u>	
דד ⊿ר	
42	
43	
44	
45	
46	
47	
10	
40	
49	
50	
51	
52	
52	
55	
54	
55	
56	
57	
58	
50	

60

where y represents the priority; x_i is the considered variables; β_0 is the intercept; and β^i is the regression coefficients.

186 2-4-2- Random Forest Regressor (RF)

The random forest (RF) regression algorithm is a kind of ML approach that employs a group of decision trees, which are trained on a subset of the data, to make predictions. This technique is designed to stabilize the algorithm and decrease variance by using multiple trees. The RF regression algorithm is widely recognized as a popular model in developing regression models because of its strong performance with large datasets and diverse data types [16, 17].

192 2-4-3- Support Vector Machines Regressor (SVM)

193 Support vector machine regression (SVM) is a versatile regression function that can be used to 194 solve both classification and regression problems. SVM is a supervised learning algorithm that fits 195 a regression to the training data by reducing the distance between the sampled points and the fitted 196 hyperplane[18, 19]. One advantage of SVM is that it is a sparse algorithm, meaning that it only 197 needs information from a limited number of data points[20].

198 2-4-4- XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is another ML library that is available for free and offers a powerful and efficient implementation of the gradient boosting algorithm[21]. Gradient boosting is a technique that involves creating an ensemble of tree-based models and then combining them to create a more accurate overall model than any of the individual models in the sequence[22]. XGBoost is a popular choice for those who require an effective and optimized implementation of gradient boosting[23].

205 2-4-5- Multilayer Perceptron Model (MLP)

The Multilayer Perceptron (MLP) is considered one of the top regression models in the field of artificial neural networks. It is equipped with the capability to learn from training data using a

variety of training algorithms and rules. This feature allows the MLP to acquire numerous
advantages, including increased capacity. As a result, the MLP operates as a self-regulating model
that utilizes specific learning algorithms to enhance its performance when encountering new inputs
[24, 25].

212 2-4-6- Deep Learning Model (DL)

A deep learning model can be used for regression problems by learning a mapping from input features to the target output. Deep learning is an adaptable model proficient at effectively managing intricate data relationships. It proves especially beneficial when working with extensive datasets where traditional regression methods might not uncover intricate patterns. Nonetheless, to prevent overfitting and attain peak performance, these models necessitate meticulous calibration and validation[26-28]. Occasionally, due to the complex nature of implementing these models, simpler regression models may outperform them.

2-5- Performance evaluation

Typically, regression models are evaluated based on a function that measures the difference between the predicted and actual numerical value of the target variable, such as the priority score[29]. In this study, three popular evaluation metrics were used, including mean absolute error (MAE), root mean square error (RMSE), and R-squared (R²) score to assess the performance of the developed models [30].

(1)
$$MAE = \frac{\sum_{i=1}^{n} |R^*(i) - R(i)|}{n}$$

(2)
$$RMSE = \left\{ \frac{\sum_{i=1}^{n} (R^{*}(i) - R(i))^{2}}{n} \right\}^{1/2}$$

(3) $R^{2} = 1 - \frac{\sum_{i=1}^{n} (R^{*}(i) - R(i))^{2}}{\sum_{i=1}^{n} (R^{*}(i) - m(i))^{2}}$

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

In these formulas; variable n refers to the number of samples; $R^*(i)$ denotes the retrieved value predicted by the model; R(i) denotes the analyzed value; and m(i) denotes the average analyzed value.

To validate the developed machine learning (ML) models and reduce bias, we employed k-fold cross-validation. This technique overcomes the limitations of a simple train/test split by dividing the available data into multiple folds or subsets. By averaging the results across these folds, we achieve a more robust estimate of the model's performance compared to a simple train/test split.

2-6- Feature Importance

In the realm of machine learning models, a technique employed to elucidate the impact of each feature on the model is the SHAP (Shapley Additive Explanations) method. The SHAP method aims to enhance the transparency and interpretability of machine learning models by drawing on cooperative game theory [31]. For instance, linear models utilize their coefficients to gauge the significance of each feature. However, these coefficients are influenced by the scale of the variable itself, potentially resulting in misinterpretations [32]. The same can be found in tree-based models for feature ranking. This is precisely why SHAP (Shapley Additive Explanations) becomes valuable for model interpretation [33]. The absolute value of SHAP provides insight into how significantly an individual feature influences the prediction [34]. Once we identify the optimal model for priority prediction, we'll leverage the SHAP technique to assign weights to the most critical features. These features will then be ranked based on their importance and impact on the final priority score.

- 249 Patient and public involvement
- 250 None
- 4 251 **3-Results**

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

3-1- Dataset description

The raw dataset comprises information on 196,270 patients who underwent lung transplantation, as well as data related to lung donors. This comprehensive dataset includes 545 features, encompassing demographic and clinical details about the organ recipients, biomarkers, laboratory test results, and characteristics of the donated organs. Additionally, it provides insights into various patient outcomes, such as post-transplant survival rates, occurrences of acute organ rejection, priority levels, duration of intensive care unit stay, post-transplant infections, and instances of retransplantation.

To preprocess the dataset, we converted the transplantation date data type to a string format and extracted the year column by parsing the month, hour, and year components. We removed all data prior to 2005 due to the absence of a prioritization system during that period. Furthermore, to focus exclusively on adult transplants, we excluded information related to pediatric transplants for children and adolescents under 18 years of age. As a result, our initial dataset comprised 183,086 records.

Subsequently, we filtered the dataset to include only post-transplant survival records exceeding
 one year. After that, we eliminated any records with missing priority scores. Ultimately, our final
 dataset consisted of 45,966 records for subsequent analysis.

3-2- Exploratory data analysis

Following imputation of missing independent variable data and preprocessing steps, the overall patient population consisted of 66.88% men and 33.20% women, with a median age of 54.27 ± 14.24 years. Our target variable is the priority (or allocation) score, which represents a continuous numerical value. In Table A-1 in Appendix, we present descriptive analyses and the frequency distribution of various demographic and clinical variables within the dataset. Page 15 of 29

BMJ Open

Furthermore, we employed a data visualization method to enhance our comprehension of the data and dataset. This approach aids in verifying the integrity of the data and detecting any apparent inaccuracies. Incorporating data visualization is essential for all data science projects across various fields [35].

3-3- Data cleaning and preprocessing

In the method section, we present detailed information regarding preprocessing procedures applied to the whole dataset. After de-identifying the dataset, it was limited to 40,024 records and 445 features. During the initial data cleaning phase, we removed irrelevant features, reducing the total to 322. Subsequently, we performed missing values imputation, resulting in 215 features available for further analysis. Next, in the correlation analysis phase, our dataset contains over 165 features post-pre-processing.

Due to the dataset's high dimensionality, we applied pre-processing techniques to select only the most important features based on their importance scores. As a result, we narrowed down the dataset to 65 features in the first phase of feature engineering. After further applying feature engineering and selection techniques, our final dataset consisted of 32,966 records, containing 15 features.

3-4- Development and evaluation of regression models

The prediction models were developed by training several selected features obtained during the feature engineering phase. We used 80% of the dataset to train the algorithms and the rest 20% to test and validate their efficacy (80:20) and all six regression algorithms were trained based on trained data.

296 To address the bias of training using simple data splitting the average score, K-fold cross-297 validation was done and K was considered as 10 folds. The results showed that average scores of

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

10-fold cross-validation in six ML algorithms are the same as the simple splitting data process.
Subsequently, the hyperparameters were fine-tuned using a hyperparameter tuning technique to
enhance the performance of the developed models.

The optimized and selected hyperparameters are documented in Table 1. Next, the MAE, RMSE, and R^2 values for each optimized model were calculated. Finally, all the optimized ML models were compared based on their R^2 scores and other relevant metrics. The evaluation results for the regression models in terms of error rates are represented in Table 2.

In our quest to identify the best regression model, we focused on minimizing the error in terms of the R^2 score. As a result, the RF regression model emerged as the top performer among the developed prediction models. We made this determination based on a comprehensive evaluation of various metrics, utilizing the best features.

3-5- Most important features to select the most appropriate candidate

Upon selecting the best model, we proceeded to identify and weigh the most influential features using the SHAP library within the final model. Initially, a prediction model based on the chosen regression model was created. Subsequently, the importance of each feature was determined by analyzing the set of trees generated by the model using the SHAP technique. The SHAP library assigns a score to each feature based on its impact on the prediction model. The ranking of the variables used in the ultimate model is visually represented in Fig 2, which is a widely recognized and popular chart produced by SHAP.

Ultimately, the researchers pinpointed the 11 most effective features, each receiving the highest score in candidate prioritization. These features, along with their explanations, are detailed in Table 3. Notably, it showed that factors such as a patient's oxygen consumption and diagnosis played a significant role in prioritizing the waiting list. Additionally, the patient's waiting time on

BMJ Open

the transplant list emerged as another influential factor. Subsequently, we developed a web-based
 decision support tool using Python and the Streamlit framework based on the best-fine-tuned
 model.

⁰ 324 **4- Discussion**

The study aimed to explore the feasibility of utilizing machine learning (ML) methods to predict priority levels for patients on the waiting list for lung transplants and to pinpoint the critical factors influencing priority allocation. Despite the potential advantages of employing ML algorithms in organ allocation [36], there is a lack of research on their application specifically in lung transplantation. This investigation led to the development of a decision support tool for estimating transplantation priorities.

Currently, the decision-making process for prioritizing individuals on organ transplant waiting lists is predominantly reliant on physicians' subjective judgments, often following "first-come, first-served" or "longer waiting time" principles rather than utilizing sophisticated mathematical models [37, 38]. Researchers recommend that authorities explore more equitable and innovative solutions for allocating donor organs to patients on waiting lists. As a result, researchers in the field of transplantation have concentrated on developing advanced models to forecast priority rankings and outcomes for recipients based on pre-transplantation factors [39, 40]. Similarly, we employed ML models to investigate more appropriate factors in assigning organs to recipients.

Prior studies on organ allocation have focused only on classification models to predict the risk of mortality following transplantation [39, 41]. However, these approaches have not been highly effective in improving the prioritization of patients on lung transplant waiting lists[42, 43]. In contrast, our developed model takes into account various factors such as disease type, oxygen saturation, demographics, clinical tests, and functional status.

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

In the context of machine learning (ML), the effectiveness of methods depends not only on their design and techniques but also on the quality and suitability of the data they operate on. To overcome the limitations of prior research, which often relied on a single ML technique and small sample sizes, our study takes a different approach. We incorporate multiple ML techniques to enhance the accuracy of our results, leveraging a large dataset sourced from the United Network for Organ Sharing (UNOS) database.

Our algorithm yields slightly superior results. To enhance the robustness of our model, we employed various data preprocessing techniques and feature engineering methods. These approaches allowed us to identify the most relevant and informative features in the data while discarding redundant or noisy ones [44, 45]. Data preprocessing plays a crucial role in improving data quality and enhancing the accuracy of knowledge extraction [46]. Additionally, by reducing data complexity and dimensions, our models became better equipped to capture underlying patterns and relationships, resulting in improved predictive performance [10, 45].

Our analysis reveals that employing the RF regressor model, which incorporates 15 features from the most significant donor and recipient variables available prior to transplantation, represents an effective approach for assigning an allocation score to each candidate on the waiting list. This outperforms other regression models. RF was specifically chosen due to its favorable prediction performance in previous research [47]. A deployment model in the form of an AI-based decision support tool could assist clinicians in utilizing the survey results within the context of their decision-making process and point-of-care scenarios

ML-based models rely on intricate mathematical structures and multi-dimensional datasets, often yielding complex patterns and relationships that can be challenging for humans to grasp. To address this complexity and limitation, researchers have turned to SHAP (Shapley Additive

explanations) summary analysis. This technique identifies the top 11 influential features within the
final model. By doing so, it sheds light on which parameters should take precedence when selecting
the most suitable recipient with the highest priority—a factor that has not received extensive
exploration in prior studies.

While the suggested model demonstrated satisfactory performance, it does possess evident limitations. Despite the dataset under consideration being of a substantial size, it was obtained from a freely accessible dataset, not the Iranian transplantation data. in future studies, aligning with the structure of the UNOS database will allow for the collection of patient information tailored to researchers' requirements. Leveraging a local dataset can enhance its practical utility in pointof-care.

5-Conclusion

During this study, we succeeded in developing a priority prediction model based on the huge data of the UNOS database using ML models with the least error. Our research is among the pioneering studies that employ the SHAP method to enhance the comprehensibility of the proposed model intended for clinicians. Additionally, the automated auxiliary model that we created can assist clinicians in acquiring a better understanding of the transplant priority estimation and the crucial factors that influence patient survival. BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

6-Declaration

385 Ethics approval and consent to participate

The research was approved by the Tehran University of Medical Sciences Ethics Committee (IR.TUMS.IKHC.REC.1401.143). All methods were performed based on the relevant guidelines and regulations. Consent for participation was deemed unnecessary according to an Institutional Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee.

390

Consent for publication

5 6	391	Consent for publication was deemed unnecessary according to an Institutional Review Board
7 8	392	(IRB) of the Tehran University of Medical Sciences Ethics Committee.
9 10 11	393	Declaration of Competing Interest
12 13	394	The authors declare that they have no conflict of interest.
14 15	395	Availability of data and materials
16 17 18	396	The data used in this article can be obtained from the United Network for Organ Sharing (UNOS)
19 20	397	database by visiting www.unos.org/data. However, there are limitations on accessing this data, as
21 22	398	it was used under a license for the current study and is not accessible to the general public. The
23 24 25	399	interpretation and reporting of this data are the responsibility of the authors and in no way should
25 26 27	400	be seen as an official policy of or interpretation by the OPTN or the United States government.
28 29	401	Funding
30 31	402	This research was funded by the Thoracic Research Center through, Tehran University Medical
32 33 34	403	Sciences by Grant No (59042). The funding body played no role in the design of the study and
35 36	404	collection, analysis, interpretation of data, and in writing the manuscript.
37 38	405	Authors' contributions
39 40 41	406	Conception and design of the study: Hamidreza Abtahi, Marsa Gholamzadeh, Reza Safdari,
42 43	407	Mehrnaz Asadi Gharabaghi;
44 45	408	Data acquisition: Marsa Gholamzadeh, Hamidreza Abtahi;
46 47 48	409	Interpretation and/or analysis of data: Marsa Gholamzadeh, Hamidreza Abtahi, Reza Safdari,
49 50	410	Mehrnaz Asadi Gharabaghi;
51 52	411	Drafting the manuscript: Marsa Gholamzadeh, Hamidreza Abtahi;
53 54	412	Revising the manuscript critically for important intellectual content: All authors;
55 56 57		
58		19
60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

2		
3	413	Approval of the version of the manuscript to be published: All authors;
4		
5 6	414	Acknowledgments
7		
8	415	The data reported here have been supplied by the United Network for Organ Sharing
9		
10	416	(UNOS/OPTN) as the contractor for the Organ Procurement and Transplantation Network. We
11		
12 13	417	express our gratitude to the UNOS organization for allowing access to the data. We would like to
14		
15	418	extend our sincere thanks to the Thoracic Research Center of the Tehran University of Medical
16		
17	419	Sciences (TUMS) for their support and cooperation during this research.
18		
20	420	References
21	121	1 van der Mark SC, Heek PAS, Hellemens ME: Developments in lung transplantation over the past
22	421	1. Van der Mark SC, Hoek RAS, Hellemons ME. Developments in lung transplantation over the past
23	422	decade. European Respiratory Review 2020, 29 (157):190132.
24	423	2. Verleden GM, Dupont L, Yserbyt J, Schaevers V, Raemdonck DV, Neyrinck A, Vos R: Recipient
25	424	selection process and listing for lung transplantation. Journal of Thoracic Disease 2017,
26	425	9 (9):3372-3384.
27	426	3. Smits JM, Nossent G, Evrard P, Lang G, Knoop C, Kwakkel-van Erp JM, Langer F, Schramm R, van
28	427	de Graaf E, Vos R et al: Lung allocation score: the Eurotransplant model versus the revised US
29	428	model – a cross-sectional study. Transplant International 2018. 31(8):930-937.
30	429	4 Lancaster TS Miller IR Enstein DI DuPont NC Sweet SC Eghtesady P: Improved waitlist and
31	/20	transplant outcomes for nediatric lung transplantation after implementation of the lung
32	430	ellesetion score ///eart //ung Transplant 2017 26/EVE20 E29
33	431	anocation score. J Heart Lung Transplant 2017, 30 (5):520-528.
34	432	5. Satdari R, Rezayi S, Saeedi S, Tannapour M, Gnolamzaden M: Using data mining techniques to
35	433	fight and control epidemics: A scoping review. <i>Health and Technology</i> 2021, 11 (4):759-771.
36	434	6. Gholamzadeh M, Abtahi H, Safdari R: Comparison of different machine learning algorithms to
37	435	classify patients suspected of having sepsis infection in the intensive care unit. Informatics in
38	436	Medicine Unlocked 2023, 38 :101236.
39	437	7. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, McEvoy D, Stylianopoulos T, Munn
40	438	LL. Dutta S et al: Comparing machine learning algorithms for predicting ICU admission and
41	439	mortality in COVID-19 NPI Digit Med 2021 4(1):87
42	110	8 Safdari B. Deghatinour A. Gholamzadeh M. Maghooli K: Annlying data mining techniques to
43	440	dessify notions with suspected bonatitic C virus infection Intelligent Medicine 2022 2(4):102
44	441	classify patients with suspected nepatitis c virus infection. Intelligent Medicine 2022, 2(4).195-
45	442	
46	443	9. Gholamzadeh M, Abtahi H, Safdari R: Machine learning-based techniques to improve lung
47	444	transplantation outcomes and complications: a systematic review. BMC Medical Research
48	445	Methodology 2022, 22 (1):331.
49	446	10. Miller PE, Pawar S, Vaccaro B, McCullough M, Rao P, Ghosh R, Warier P, Desai NR, Ahmad T:
50	447	Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics:
51	448	Insights From the UNOS Database. Journal of Cardiac Failure 2019. 25(6):479-483.
52	449	11. LeClaire IM, Smith NJ, Chandratre S, Rein J, Kamalia MA, Kohmoto T, Joyce JD, Joyce DJ: Solid
53	150	organ donor-recipient race-matching: analysis of the United Network for Organ Sharing
54	450	database Transplint 2021 21 (1):640-647
55 56	TCF	autabase. Hanspinit 2021, 54(4).040-047.
50 57		
52		20
50		20
60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

2			
3	452	12.	Theng D, Bhoyar KK: Feature selection techniques for machine learning: a survey of more than
4	453		two decades of research. Knowledge and Information Systems 2024, 66(3):1575-1637.
5	454	13.	Li D, Liu Z, Armaghani DJ, Xiao P, Zhou J: Novel ensemble intelligence methodologies for
0	455		rockburst assessment in complex and variable environments. Sci Rep 2022, 12(1):1844.
/ 8	456	14.	Uyanık GK, Güler N: A Study on Multiple Linear Regression Analysis. Procedia - Social and
9	457		Behavioral Sciences 2013, 106 :234-240.
10	458	15.	Kavri M. Kavri I. Gencoglu MT: The performance comparison of Multiple Linear Regression.
11	459		Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data. In:
12	460		2017 14th International Conference on Engineering of Modern Electric Systems (EMES): 1-2 June
13	461		2017 2017 [,] 2017 [,] 2017 [,] 1-4
14	162	16	Dai B. Chen BC. Zhu SZ. Zhang WW: Using Random Forest Algorithm for Breast Cancer Diagnosis
15	462	10.	In: 2018 International Symposium on Computer Consumer and Control (IS2C): 6-8 Dec. 2018 2018:
16	403		
17	404	17	2010.445-452.
18	405	17.	Sinici PF, Ganesi S, Liu P. A comparison of random forest regression and multiple inear
19	400	10	regression for prediction in neuroscience. Journal of Neuroscience Methods 2013, 220(1):85-91.
20	467	18.	Yu W, Liu T, Valdez R, Gwinn M, Knoury WJ: Application of support vector machine modeling for
21	468		prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics
22	469		and Decision Making 2010, 10 (1):16.
24	470	19.	Sarker IH: Machine Learning: Algorithms, Real-World Applications and Research Directions. SN
25	471		<i>Computer Science</i> 2021, 2 (3):160.
26	472	20.	Huang H, Wei X, Zhou Y: An overview on twin support vector regression. Neurocomputing 2022,
27	473		490 :80-92.
28	474	21.	Bentéjac C, Csörgő A, Martínez-Muñoz G: A comparative analysis of gradient boosting
29	475		algorithms. Artificial Intelligence Review 2021, 54(3):1937-1967.
30	476	22.	Li S, Zhang X: Research on orthopedic auxiliary classification and prediction model based on
31	477		XGBoost algorithm. Neural Computing and Applications 2020, 32(7):1971-1979.
32	478	23.	Liu J, Wu J, Liu S, Li M, Hu K, Li K: Predicting mortality of patients with acute kidney injury in the
33 24	479		ICU using XGBoost model. PLOS ONE 2021, 16(2):e0246306.
25	480	24.	Sananmuang T, Mankong K, Chokeshaiusaha K: Multilayer perceptron and support vector
36	481		regression models for feline parturition date prediction. <i>Heliyon</i> 2024, 10 (6):e27992.
37	482	25.	Abiodun OI, Jantan A, Omolara AE, Dada KV, Umar AM, Linus OU, Arshad H, Kazaure AA, Gana U,
38	483		Kiru MU: Comprehensive Review of Artificial Neural Network Applications to Pattern
39	484		Recognition. IEEE Access 2019. 7:158820-158846.
40	485	26.	Sarker IH: Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications
41	486		and Research Directions. SN Computer Science 2021. 2 (6):420.
42	487	27	Ahmed SE, Alam MSB, Hassan M, Rozbu MR, Ishtiak T, Rafa N, Mofijur M, Shawkat Ali ABM
43	488	27.	Gandomi AH: Deen learning modelling techniques: current progress applications advantages
44	180		and challenges Artificial Intelligence Review 2023 56(11):13521-13617
45	100	20	LeCup V. Bengio V. Hinton G: Deen learning. Nature 2015. 50(11):15521 15017.
46	490	20.	Pamach A Pamamoorthy S Pubari SM: Forocasting Spread of COVID 19 Using Pogression
4/	491	29.	Algorithm In: Soft Computing for Droblem Soluting 2021/(2021: Singapore: Springer Singapore:
40 70	492		Algorium. In: Soft Computing for Problem Solving: 2021/ 2021; Singupore: Springer Singapore;
50	493	20	
51	494	30.	Karunasingha DSK: Root mean square error or mean absolute error? Use their ratio as well.
52	495		Information Sciences 2022, 585 :609-629.
53	496	31.	Ekanayake IU, Meddage DPP, Rathnayake U: A novel approach to explain the black-box nature
54	497		of machine learning in compressive strength predictions of concrete using Shapley additive
55	498		explanations (SHAP). Case Studies in Construction Materials 2022, 16:e01059.
56			
57			
58			21

59 60

1			
2			
4	499	32.	Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee
5	500		SI: From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach
6	501	~~	Intell 2020, 2 (1):56-67.
7	502	33.	Rodriguez-Pérez R, Bajorath J: Interpretation of machine learning models using shapley values:
8	503		application to compound potency and multi-target activity predictions. Journal of Computer-
9	504		Aided Molecular Design 2020, 34 (10):1013-1026.
10	505	34.	Kim Y, Kim Y: Explainable heat-related mortality with random forest and SHapley Additive
11	506		exPlanations (SHAP) models. Sustainable Cities and Society 2022, 79:103677.
12	507	35.	Patel Darshan R, Reddy PVB: The Importance of Data Visualization in Exploratory Data Analysis.
14	508		Journal of Advanced Zoology 2023, 44 (S6):923-929.
15	509	36.	Peloso A, Moeckli B, Delaune V, Oldani G, Andres A, Compagnon P: Artificial Intelligence: Present
16	510		and Future Potential for Solid Organ Transplantation. Transpl Int 2022, 35:10640.
17	511	37.	Bunnik EM: Ethics of allocation of donor organs. Curr Opin Organ Transplant 2023, 28(3):192-
18	512		
19	513	38.	Madwar S: United States officials propose further retreat from first-come, first-served organ
20	514	20	donation. <i>Cmaj</i> 2011, 183 (10):E639-640.
21	515	39.	Lau L, Kankanige Y, Rubinstein B, Jones R, Christophi C, Muralidharan V, Bailey J: Machine-
23	516		Learning Algorithms Predict Graft Failure After Liver Transplantation. Transplantation 2017,
24	517	10	101(4):e125-e132.
25	518	40.	Gotileb N, Azhie A, Sharma D, Spann A, Suo N-J, Tran J, Orchanian-Cheff A, Wang B, Goldenberg
26	519		A, Chasse M et al. The promise of machine learning applications in solid organ transplantation.
27	520		npj Digital Medicine 2022, 5(1):89.
28	521	41.	Jawitz OK, Raman V, Becerra D, Klapper J, Hartwig MG: Factors associated with short- versus
29 30	522	40	Iong-term survival after lung transplant. J Inorac Cardiovasc Surg 2022, 163(3):853-860.8852.
30	523	42.	Branmbhatt JN, Hee Wal I, Goss CH, Lease ED, Merio CA, Kaphadak SG, Ramos KJ: The lung
32	524		allocation score and other available models lack predictive accuracy for post-lung transplant
33	525	40	survival. J Heart Lung Transplant 2022, 41(8):1063-1074.
34	520	43.	Madela Faile to Improve Discrimination Deformance Chest 2022 162(1):152-162
35	527	4.4	Niddels Fails to improve Discrimination Performance. Cilest 2023, 163(1):152-163.
36	528	44.	Methods for Machine Learning Pased Disease Pick Prediction Frontiers in Riginformatics 2022
3/	529		Nethous for Machine Learning-based Disease Risk Prediction. Frontiers in Bioinjornatics 2022,
20 20	550	45	2. Sour V Into I Larrañago D: A review of feature coloction techniques in high-fermatics
40	531	45.	Saleys Y, Inza I, Larranaga P: A review of realure selection techniques in bioinformatics.
41	552	16	Biolinjorniulus 2007, 23(19).2307-2317.
42	555	40.	and prospects. <i>Big</i> Data Anglytics 2016 1(1):0
43	534	47	Ocka T. Johna H. Nakamata K. Vada V. Vakamichi H. Vamagata 7: Pandam forest approach for
44	555	47.	determining risk prediction and predictive factors of type 2 diabetes: large scale health check
45	550		undetering fisk prediction and predictive factors of type 2 diabetes. Taige-scale field in theck-
46	337		up uata in Japan . Bivis Nathtion, Prevention & amp, Health 2021, 4 (1).140-148.
47 48	538		
49			
50	539		
51			
52			
53			
54			
55			
50 57			
58			22
59			22
60			For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Tables

541 T 1 2 3 4 5	Table 1- The best selec Algorithm Multiple linear regression Random Forest Regressor	ted hyperparameters Hyperparameters positive= False, n_jobs= 2, fit_intercept= True, copy_X= True
1 2 3 4 5	Algorithm Multiple linear regression Random Forest Regressor	Hyperparameters positive= False, n_jobs= 2, fit_intercept= True, copy_X= True
1 2 3 4 5	Multiple linear regression Random Forest Regressor	positive= False, n_jobs= 2, fit_intercept= True, copy_X= True
2 3 4 5	Random Forest Regressor	
3 4 5		n_estimators= 90, min_samples_split= 2, min_samples_leaf= 1, max_samples 10000,
3 4 5		max_features: sqrt, max_depth=10
4	SVM Regressor	C =9.11158, loss='epsilon_insensitive', max_iter=5000
5	XGBoost Regressor	subsample=1, min_child_weight= 5, max_depth= 6, learning_rate=0.1, colsample_bytree=0.75
	MLP	solver= 'sgd', Learning_rate= 'adaptive', hidden_layer_sizes: (20,), alpha: 0.001, activation: logistic
6	DL	Optimizer= 'sgd', batch_size= 16, activation= 'relu'
542		
5/12		
545		
		23

544 Table 2- The evaluation metrics of developed models and comparison of the model performance

	Model	R ²	MSE	MAE	RMSE
1	Random Forest Regressor	95.168	12.548	2.056	3.542
2	XGBoost Regressor	83.012	58.326	4.487	7.637
3	Deep Learning algorithm	68.736	80.096	42.096	45.05
4	MLP Regressor	66.003	88.97	5.681	9.432
5	Linear Regression	52.259	123.989	6.984	11.131
6	Support Vector Machines	48.590	133.591	6.570	11.555
5					
6					
7					
/					

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES)

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

2	
2	
3	
4	
5	
6	
7	
/	
8	
9	
10	
10	
11	
12	
13	
14	
1	
15	
16	
17	
18	
10	
17	
20	
21	
22	
22	
22	
24	
25	
26	
27	
27	
28	
29	
30	
31	
27	
32	
33	
34	
35	
26	
50	
37	
38	
39	
10	
40	
41	
42	
43	
11	
44	
45	
46	
47	
<u>4</u> 8	
40	
49	
50	
51	
52	
52 52	
53	
54	
55	
56	
50	
5/	
58	

59

60

Table 3- The top 11 features identified by the SHAP method based on the prediction model

#	Feature	Description
1	INIT_02	The amount of oxygen needed when the transplant candidate is on the waiting list
2	GROUPING	Lung transplant candidate diagnosis group
3	DAYSWAIT_CHRON	The amount of waiting time of patients on the waiting list - up-to- date waiting time
4	MED_COND_TRR	The status of the patient's lungs at the time of the last clinical evaluation
5	HEMO_SYS_TRR	The latest status of Hemodynamics Pcw (Sys) MM/Hg
6	END_O2	O2 Requirement at rest
7	VENTILATOR_TCR	The patient's status in terms of the need for a ventilator
8	LIFE_SUP_TCR	The amount of social and financial support
9	CIG_Use	History of cigarette use
10	Vent_Support_TRR	Episode of ventilatory support
11	Transfusion	Events occurring between listing and transplant

549

1 2		
3 4	551	Figure legends
5	552	Fig 1. Schematic diagram of the proposed method
7 8	553	Fig 2-(a) SHAP summary plot of the top 11 features for predicting lung allocation score using random forest
9 10	554	regressor and (b) SHAP values to explain the predicted probabilities
11 12	555	
13 14	556	
15 16		
17 18		
19 20		
21 22		
23 24		
25 26		
20		
28 29		
30 31		
32 33		
34 35		
36 37		
38 39		
40 41		
42 43		
44 45		
46 47		
48		
49 50		
51 52		
53 54		
55 56		
57 58		26
59 60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml





Table A-1- Dataset description

	Variable	Range	Mean (SD)	SE	95% Cont
	A	19.59	54.27 (17.20)	0.005	20 2077
	Age	18-58	34.27 (17.30)	0.093	25.25977
Lung	BMI	14.99/- 44.//	25.3 (3.83)	0.021	25.2586
recipients	FEVI value	5-120	39.484 (17.30)	0.095	39.2977
1	Initial creatinine	0.1-24	0.841 (0.407)	0.002	0.8369
	Total Albumin serum	0.5-24	3.8787 (0.406)	0.002	3.8743
Summary sta	atistics of selected ca	tegorical predicto	ors (N=)		
	Variable		n	Percentage (%)	
	Conden	Male	18085	54.86	
	Gender	Female	14881	45.14	
		Α	1513	38.31	
	АВО	В	4529	11.32	
r		AB	1513	3.78	
Lung		0	18648	46.59	
recipients	History of Malignancy	Positive	363	1.10	
		Negative	30061	91.19	
		Unknown	2542	7.71	
	History of previous	Having	808	2.45	
	transplantation	Not Having	32158	97.55	
		Male	14154	42.94	
	GENDER	Female	14154	42.94	
		Unknown	9704	29.44	
		A	279	0.85	
		В	2544	7.72	
Donor	АВО	AB	11832	35.89	
		0	13038	39.55	
		Unknown	279	0.85	
		Positive	9756	29.59	
	History of	Negative	21300	64 61	
	Malignancy	Lulas	1010	5.70	

BMJ Open

Identifying the most influential factors in lung transplant patients using a multivariate prediction model: An analysis of UNOS datasets

Journal:	BMJ Open
Manuscript ID	bmjopen-2024-089796.R1
Article Type:	Original research
Date Submitted by the Author:	25-Mar-2025
Complete List of Authors:	Gholamzadeh, Marsa; Tehran University of Medical Sciences, Safdari, Reza; Tehran University of Medical Sciences, Asadi Gharabaghi, Mehrnaz; Tehran University of Medical Sciences, Abtahi, Hamidreza; Tehran University of Medical Sciences, Pulmonary and Critical Care Department; Tehran University of Medical Sciences, Thoracic Research Center
Primary Subject Heading :	Health informatics
Secondary Subject Heading:	Health informatics, Respiratory medicine, Health services research, Intensive care
Keywords:	Machine Learning, Pulmonary Disease < Lung Diseases, Transplant medicine < INTERNAL MEDICINE





I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our <u>licence</u>.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which <u>Creative Commons</u> licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

terez oni

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de I Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

1		
2		
3 4	1	Identifying the most influential factors in lung transplant patients using a
5		
6	2	multivariate prediction model: An analysis of UNOS datasets
7		
o 9	n	Marca Chalamandahl Dara Safdaril Mahmar Agadi Charahashi? Hamidrara Aktabi3*
10	5	Marsa Gholamzaden', Keza Saldan', Meninaz AsadiGharabagni', Hanndreza Abtani',
11		
12	4	1. Health Information Management and Medical Informatics Department, School of Allied Medical Sciences, Tehran
13 1/1	5	University of Medical Sciences, Tehran, Iran.
14	6	2. Department of Pulmonary Medicine, Faculty of Medicine, Tehran University of Medical Sciences, Tehran, Iran.
16	7	3. Pulmonary and Critical Care Department, Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran
17	8	University of Medical Sciences, Tehran, Iran.
18	9	
19	10	
20	11	*Corresponding author: Hamidreza Abtahi
21		Corresponding autory maintained and and
22	12	E-mail address: hrabtahi2020research@gmail.com
24		
25	13	Tel: +9821-66192646
26	10	
27	14	Postal address: Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical
28		
29	15	Sciences, Qarib Ave, Keshavarz Blv, Tehran, Iran.
31	16	
32	10	
33	17	
34	17	ORCID ID:
35	18	Marsa Gholamzadeh, https://orcid.org/0000-0001-6781-9342
36	10	
37 38	19	Hamidreza Abtahi, https://orcid.org/0000-0002-1111-0497;
39		
40	20	Reza Safdari, https://orcid.org/0000-0002-4982-337X
41	24	
42	21	Mehrnaz Asadi Gharabaghi: https://orcid.org/0000-0003-0852-1532
43	22	
44 45		
45	7 2	
47	25	
48	24	
49		
50	25	
51 52	26	
5∠ 53	-	
55 54	27	
55	28	Abstract
56	20	
57		
58		1
59 60		For peer review only - http://bmiopen.bmi.com/site/about/quidelines.xhtml

Page 3 of 41

BMJ Open

Objectives: In lung transplantation, a priority is assigned to each candidate on the waiting list. Our
 primary objective was to identify the key factors that influence the allocation of priorities in lung
 transplantation using machine learning (ML) techniques to enhance the process of prioritizing
 patients.

Design: Developing a prediction model.

Setting and participants: Our data was retrieved from the UNOS open-source database of
transplant patients between 2005 and 2023.

Interventions: After the preprocessing process, a feature engineering technique was employed to
select the most relevant features. Then, six ML models with an optimized hyper-parameter
including Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector
Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model, and a deep
learning model (DL) were developed based on UNOS dataset.

41 Primary and secondary outcome measures: The performance of each model was evaluated
42 using R-squared (R²) and other error rate metrics. Next, the Shapley Additive Explanations
43 (SHAP) technique was utilized to identify the most important features in the prediction.

Results: The raw dataset contains 196,270 records with 545 features in all organs. After preprocessing, 32,966 records with 15 features remain. Among various models, the RF model achieved a high R2 score. Additionally, the RF model exhibited the lowest error values indicating its superior precision compared to other regression models SHAP technique in conjunction with the RF model revealed the 11 most important features for priority allocation. Subsequently, we developed a web-based decision support tool using Python and the Streamlit framework based on the best-fine-tuned model.
BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

<text><text><text> **Conclusion**: The deployment of the ML model has the potential to act as an automated tool to aid physicians in assessing the priority of lung transplants and identifying significant factors that play a role in patient survival. Keywords: Lung transplantation, allocation score, Machine learning, Prediction. For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1 2			
3 4	56	S	trengths and limitations of this study:
5 6	57	•	To ensure transparency and interpretability in our machine learning models, we employed
/ 8 9	58		Explainable Artificial Intelligence (XAI) techniques, specifically the SHAP (Shapley Additive
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29	59		Explanations) method.
12 13	60	•	The use of various preprocessing and data cleaning techniques in our survey increased the
14 15 16	61		robustness and performance of the model.
17 18	62	•	Understanding the factors influencing the determination of lung transplant priority could
19 20	63		support clinicians in designing treatment plans and thus improving the quality of life of patients.
21 22 22	64	•	Deploying the developed ML model in the form of a decision support system increases its
23 24 25	65		applicability in clinical practice.
26 27	66		
28 29 30	67		
31 32			
33 34			
35 36 37			
38 39			
40 41 42			
43 44			
45 46			
47 48 40			
50 51			
52 53			
54 55 56			
57 58			4
59 60			For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Lung transplantation (LTx) is an advanced treatment option for patients suffering from end-stage lung disease. When no other treatment options are available and the patient is likely to die, lung transplant surgery is suggested as a well-established treatment option [1]. When a patient meets the inclusion criteria for transplantation, they are placed on a waiting list and assigned a priority. Various conditions may affect eligibility for lung transplantation and the patient's priority [2]. In some countries, a score is assigned to each patient on the waiting list to enhance the recipient selection process [3, 4]. Understanding the most influential factors in priority allocation for lung transplantation is beneficial for researchers worldwide, as it can improve post-transplant survival. Utilizing data mining methods and developing forecasting models in this field could aid clinicians in uncovering hidden patterns and relationships within patient data and allocation scores.

Machine learning (ML) methods have been developed across various fields of clinical medicine to assist clinicians in predicting and classifying diseases [5]. These methods are used to predict the length of stay in the Intensive Care Unit (ICU), diagnose septic infection[6], and extract disease patterns from big data [7, 8]. Nevertheless, there is a lack of studies on the development of predictive models and identification of important features using ML methods to predict lung transplantation priority [9, 10]. Thus, the primary objective of this study was to utilize ML techniques to identify the most influential factors that strongly impacted outcomes based on various developed ML methods to predict the priority using clinical and demographic data.

2- Methods

Throughout this section, the process of developing, comparing, and evaluating ML models is shown schematically in Fig 1. Python programming language version 10 was used in this study for developing and validating ML algorithms. For data preprocessing, Numpy and Pandas modules

BMJ Open

ware employed, while the sei kit learn library was utilized for developing supervised elessifier	
elections	
algorithms.	
2-1-Dataset description and data retrieval	
The data for this study were obtained from the United Network for Organ Sharing (UNOS) online	Prot
database [11]. Upon receiving written permission from UNOS, we accessed the recorded data	ected I
pertaining to lung transplantation for our research. Our study included patients over 18 years old	by сор
with end-stage lung disease who underwent lung transplants between 2005 and 2022. We	yright,
performed a waiting list analysis using all available data entries from the United Network for	includ
Organ Sharing (UNOS) database for our study.	ling for
The priority of candidates on the waiting list was considered as the outcome, while the clinical and	Enst uses
demographic characteristics of patients were considered as features or predictors.	related
2-2-Pre-processing process	to tex
Data pre-processing is a crucial step in ML techniques, especially when dealing with raw data from	perieu t and d
clinical databases or medical records that often contain missing or unclear information. To ensure	r (ABE lata mi
the development of more accurate models based on appropriate data, we followed a series of data	S). ning, A
pre-processing steps. The following steps were employed in this phase as pre-processing	Al train
techniques.	ing, a
1- Checking the duplicated values and records to remove the duplicates	nd sim
2- De-identify records and remove irreverent features	ilar tec
3- Convert nominal and categorical features to numerical values	;hnolo
4- Identify missing data and missing values imputation	gies.
5- Outlier detection	
6- Feature engineering and feature selection	
6	
For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

BN

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

114 7- Data transformation and normalization

2-2-1-Duplicate checking and removal of irreverent features

After duplicate checking, we consulted UNOS guidelines and experts to review all features and their definitions. Under their supervision, we removed identification variables and irrelevant features, such as ID columns, hospital center identification codes, and country of residence, to deidentify patients.

Following this, we converted the post-transplant survival days variable to years and excluded
patients with a survival rate of less than two years. Next, we filtered out patients below 18 years
of age and excluded any data before 2005. Additionally, records related to heart transplantation
were removed from the dataset.

After removing irrelevant features in the first data-cleaning phase, we utilized the discretized operator to convert nominal values to numerical data. Categorical data were encoded using the LabelEncoder class too.

2-2-2- Missing data management:

To address missing data in our dataset, we conducted missing data imputation across the entire dataset. Initially, we assessed the specified columns or attributes to determine the extent of missing and unique data in each column. During this analysis, we discovered that the ICU column was empty and decided to delete it due to its lack of meaningful information. To impute the missing data, a threshold of 80% was set for feature removal with expert consultation. As a result, any column with more than 80% missing data was removed. If the missingness is due to inconsistent reporting rather than clinical irrelevance, dropping the column could exclude critical information about high-risk patients. In this case, domain experts might recommend retaining the column and using advanced imputation techniques or creating a binary indicator for missingness.

Page 9 of 41

BMJ Open

Along with feature removal, the sensitivity analysis was conducted on dataset which revealed that
the inclusion of omitted features had a detrimental effect on the performance of the Random Forest
Regressor model. Specifically, these features led to a decrease in the R² score from 0.95 to 0.68
and an increase in RMSE from 3.5 to 4.8.

For the missing data imputation, we adopted a strategy of replacing missing data in numerical features with the mean value of each respective feature. This approach allows us to retain the integrity of the dataset while minimizing the impact of missing data on our analysis. By performing these comprehensive steps of missing data imputation, we ensure the dataset is optimized for further analysis and modeling, enabling us to draw more accurate conclusions and insights.

2-2-3- Outliers handling

Outliers can significantly impact the performance and interpretability of machine learning models. Therefore, it is essential to investigate their causes before deciding whether to exclude, transform, or retain them. This exploration ensures that the preprocessing steps are justified and scientifically sound. To address outliers in the dataset, we first create distribution plots to visualize the data. Next, we apply the IQR method and use Box plots to identify outliers. Finally, we remove these outliers to prepare the data for further processing. BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

To address outliers in the dataset, a comprehensive approach was employed that included visual and statistical analysis to identify and understand the nature of the outliers. The distribution plots and boxplots were created to visualize outliers and applied the IQR method to quantify their extent. Additionally, statistical analysis was conducted to assess the impact of outliers on the dataset and performed sensitivity analysis to evaluate their influence on model performance. Throughout this process, we ensured transparency and justification by documenting all outliers and providing context-specific reasons for their exclusion, such as data entry errors or clinically irrelevant

160 extreme values [12, 13]. This rigorous approach ensured that the removal of outliers was161 methodologically sound and did not compromise the integrity of our analysis.

2-2-4-Feature engineering and feature selection

Given the high-dimensional nature of the dataset used in this study, feature engineering and selection were critical steps to reduce dimensionality, eliminate irrelevant or redundant features, and enhance model performance. Feature selection aims to identify a subset of features that effectively describe the problem with minimal loss of information and computational efficiency [14]. All phases of feature engineering and selection were conducted under the supervision of clinical experts to ensure the relevance and validity of the selected features.

Step one, correlation analysis: As the first step, correlation analysis was performed to assess the relationships between features and the target variable, as well as inter-feature correlations. This analysis helped identify highly correlated features that could introduce multicollinearity and redundancy into the model using heatmap graph. Features with a correlation coefficient above a predefined threshold were flagged for further evaluation.

Step two, variance threshold filtering: To eliminate low-variance features that contribute little
 to the model's predictive power, a variance threshold was applied. Features with variance below a
 specified threshold (e.g., 0.01) were removed, as they were deemed to have minimal impact on the
 target variable.

Step three, embedded feature selection with XGBoost: Following the initial filtering, an
 embedded feature selection technique was employed using the XGBoost algorithm. XGBoost
 provides intrinsic feature importance scores based on metrics such as gain, cover, and frequency.
 Features with very low importance scores (negative scores indicating no correlation with target
 value) were excluded from the final feature set.

BMJ Open

Step four, expert review and validation: All selected features were reviewed and validated by subject matter experts to ensure their clinical, practical, and scientific relevance. This step was critical to avoid eliminating features that, although statistically significant, may not be clinically significant. For example, some features were retained for model customization based on expert consultation, despite having modest statistical significance.

Through a comprehensive and expert-guided feature selection process, we identified a subset of features that were statistically significant, domain-relevant, and impactful for model performance. This rigorous approach ensured the final model was both robust and clinically meaningful, with the selected features deemed critical for predicting the target variable.

24 192 2-2-5-Data transformation and normalization: In the end, data normalization was carried out to
 25 193 optimize the features for modeling purposes.

²⁸ 29 194 **2-3- Splitting data and validation technique**

During the model development process, the dataset was divided into training and testing data in an 80:20 ratio where 80% of the data was used for training the models and the remaining 20% was reserved for testing and validation. The training dataset is used to train the model, allowing it to learn patterns and relationships within the data based on the available data. The training dataset typically contains the bulk of the available data. In contrast, the testing dataset is intended solely to evaluate the model's performance on unseen data, ensuring an unbiased assessment of its generalizability. This dataset is kept separate from the training process to provide a true measure of how the model performs in real-world scenarios. Both datasets are often split randomly, with common ratios such as 80:20 or 70:30, depending on the size and nature of the data.

This split ensured that the models were evaluated on unseen data to assess their generalization capability. To mitigate potential bias introduced by simple data splitting, cross-validation

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

technique was employed. Through this technique, the dataset was divided into k folds, and each model was trained and validated k times, with each fold serving as the validation set once. The average performance metrics across all folds were calculated to ensure robust evaluation. The results of the k-fold cross-validation were consistent with those obtained from the simple 80:20 split, confirming the reliability of the initial approach.

2-4- Model development and tuning

The objective of this study was to develop a prediction model for a continuous numerical variable (priority score) using regression techniques to identify the most effective factors in selecting the most appropriate candidate for LTx. In this study, regression models were selected to examine the connection between input variables and output numerical values, as the target variable (outcome) is a continuous numerical value.

To determine the most influential factors and identify the best model, the performance of six regression-based models was evaluated: Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model (MLP—a class of feedforward artificial neural network), and a deep learning model (DL). The selection of these models was done based on the type of target variable and the study objectives.

A hyperparameter tuning optimization technique was employed in this phase to improve model performance by optimizing the training process by determining the best hyperparameters for each model. This technique was used to prevent models that underfit or overfit the data [15]. After tuning parameters in each model, the models were trained with updated best hyperparameters, and all metrics were calculated again to achieve the best performance. We employed the random search

BMJ Open

2		
3 4	228	method, a hyperparameter tuning technique where hyperparameters are randomly chosen from a
5 6 7	229	predefined set to train a model.
7 8 9	230	2-4-1- Multiple linear regression (MLR)
10 11	231	Multiple linear regression (MLR) is a statistical technique used to estimate the relationship
12 13	232	between a dependent variable and one or more independent variables. It is an extension of linear
14 15	233	regression, which requires more than one predictor variable to forecast the response variable[16].
16 17 19	234	MLR is a significant regression algorithm that models the linear association between a dependent
19 20	235	continuous variable and multiple independent variables[17]. Hence, we have chosen this model to
21 22	236	predict the continuous variable (priority score) based on several independent variables. The
23 24	237	equation for multiple linear regression is demonstrated below[17]:
25 26 27	238	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 (1)$
27 28 29	239	where y represents the priority; x_i is the considered variables; β_0 is the intercept; and β^i is the
30 31	240	regression coefficients.
32 33	241	2-4-2- Random Forest Regressor (RF)
34 35 26	242	The random forest (RF) Regressor algorithm is a kind of ML approach that employs a group
30 37 38	243	of decision trees, which are trained on a subset of the data, to make predictions. This technique is
39 40	244	designed to stabilize the algorithm and decrease variance by using multiple trees. The RF regressor
41 42	245	algorithm is widely recognized as a popular model in developing regression models because of its
43 44 45	246	strong performance with large datasets and diverse data types [18, 19].
46 47	247	2-4-3- Support Vector Machines Regressor (SVM)
48 49	248	Support vector machine regression (SVM) is a versatile regression function that can be used to
50 51 52	249	solve both classification and regression problems. SVM is a supervised learning algorithm that fits
52 53 54 55	250	a regression to the training data by reducing the distance between the sampled points and the fitted
50 57 58		12

hyperplane[20, 21]. One advantage of SVM is that it is a sparse algorithm, meaning that it onlyneeds information from a limited number of data points[22].

253 2-4-4- XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is another ML library that is available for free and offers a
powerful and efficient implementation of the gradient boosting algorithm[23]. Gradient boosting
is a technique that involves creating an ensemble of tree-based models and then combining them
to create a more accurate overall model than any of the individual models in the sequence[24].
XGBoost is a popular choice for those who require an effective and optimized implementation of
gradient boosting[25].

2-4-5- Multilayer Perceptron Model (MLP)

The Multilayer Perceptron (MLP) is considered one of the top regression models in the field of artificial neural networks. It is equipped with the capability to learn from training data using a variety of training algorithms and rules. This feature allows the MLP to acquire numerous advantages, including increased capacity. As a result, the MLP operates as a self-regulating model that utilizes specific learning algorithms to enhance its performance when encountering new inputs [26, 27].

40 267 **2-4-6- Deep Learning Model (DL)**

A deep learning model can be used for regression problems by learning a mapping from input features to the target output. Deep learning is an adaptable model proficient at effectively managing intricate data relationships. It proves especially beneficial when working with extensive datasets where traditional regression methods might not uncover intricate patterns. Nonetheless, to prevent overfitting and attain peak performance, these models necessitate meticulous calibration and validation[28-30]. Occasionally, due to the complex nature of implementing these models, simpler regression models may outperform them.

2-5- Performance evaluation

Typically, regression models are evaluated based on a function that measures the difference between the predicted and actual numerical value of the target variable, such as the priority score[31]. In this study, three popular evaluation metrics were used, including mean absolute error (MAE), root mean square error (RMSE), and R-squared (R²) score to assess the performance of the developed models [32].

281
(1)
$$MAE = \frac{\sum_{i=1}^{n} |R^{*}(i) - R(i)|}{n}$$

282
(2) $RMSE = \left\{\frac{\sum_{i=1}^{n} (R^{*}(i) - R(i))^{2}}{n}\right\}^{1/2}$
283
(3) $R^{2} = 1 - \frac{\sum_{i=1}^{n} (R^{*}(i) - R(i))^{2}}{\sum_{i=1}^{n} (R^{*}(i) - m(i))^{2}}$

In these formulas; variable n refers to the number of samples; $R^*(i)$ denotes the retrieved value predicted by the model; R(i) denotes the analyzed value; and m(i) denotes the average analyzed value. BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

To validate the developed machine learning (ML) models and reduce bias, we employed k-fold cross-validation. This technique overcomes the limitations of a simple train/test split by dividing the available data into multiple folds or subsets. By averaging the results across these folds, we achieve a more robust estimate of the model's performance compared to a simple train/test split.

2-6- Feature Importance

To implement Explainable AI (XAI), a technique employed to elucidate the impact of each feature on the model is the SHAP (Shapley Additive Explanations) method. The SHAP method aims to enhance the transparency and interpretability of machine learning models by drawing on cooperative game theory [33]. For instance, linear models utilize their coefficients to gauge the significance of each feature. However, these coefficients are influenced by the scale of the variable itself, potentially resulting in misinterpretations [34]. The same can be found in tree-based models

Page 16 of 41

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) .

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

BMJ Open

> for feature ranking. This is precisely why SHAP (Shapley Additive Explanations) becomes valuable for model interpretation [35]. The absolute value of SHAP provides insight into how significantly an individual feature influences the prediction [36]. Once we identify the optimal model for priority prediction, we'll leverage the SHAP technique to assign weights to the most critical features. These features will then be ranked based on their importance and impact on the final priority score.

304 Patient and public involvement

Patients and/or the public were not involved in the design, or conduct, or reporting, ordissemination plans of this research.

3-Results

3-1- Dataset description

The raw dataset including all organs comprises information on 196,270 patients who underwent lung transplantation, as well as data related to lung donors. This comprehensive dataset includes 545 features, encompassing demographic and clinical details about the organ recipients, biomarkers, laboratory test results, and characteristics of the donated organs (Table A-2 in Appendix A). Additionally, it provides insights into various patient outcomes, such as posttransplant survival rates, occurrences of acute organ rejection, priority levels, duration of intensive care unit stay, post-transplant infections, and instances of re-transplantation.

To preprocess the dataset, we converted the transplantation date data type to a string format and extracted the year column by parsing the month, hour, and year components. We removed all data prior to 2005 due to the absence of a prioritization system during that period. Furthermore, to focus exclusively on adult transplants, we excluded information related to pediatric transplants for

children and adolescents under 18 years of age. As a result, our initial dataset comprised 183,086records.

Subsequently, we filtered the dataset to include only post-transplant survival records exceeding
one year. After that, we eliminated any records with missing priority scores. Ultimately, our final
dataset consisted of 45,966 records for subsequent analysis.

3-2- Exploratory data analysis

Following imputation of missing independent variable data and preprocessing steps, the overall patient population consisted of 66.88% men and 33.20% women, with a median age of 54.27±14.24 years. Our target variable is the priority (or allocation) score, which represents a continuous numerical value. In Table A-1 in Appendix A, we present descriptive analysis and the frequency distribution of various demographic and clinical variables within the dataset.

Furthermore, we employed a data visualization method to enhance our comprehension of the data and dataset. This approach aids in verifying the integrity of the data and detecting any apparent inaccuracies. Incorporating data visualization is essential for all data science projects across various fields [37].

3-3- Data cleaning and preprocessing

In the method section, we present detailed information regarding preprocessing procedures applied to the whole dataset. After de-identifying the dataset, it was limited to 40,024 records and 445 features. During the initial data cleaning phase, we removed irrelevant features, reducing the total to 322. Subsequently, we performed missing values imputation, resulting in 215 features available for further analysis. Next, in the correlation analysis phase, our dataset contains over 165 features post-pre-processing.

Due to the dataset's high dimensionality, we applied pre-processing techniques to select only the most important features based on their importance scores. As a result, we narrowed down the dataset to 65 features in the first phase of feature engineering. After further applying feature engineering and selection techniques, our final dataset consisted of 32,966 records, containing 15 features.

3-4- Development and evaluation of regression models

The prediction models were developed by training several selected features obtained during the feature engineering phase. We used 80% of the dataset to train the algorithms and the rest 20% to test and validate their efficacy (80:20) and all six regression algorithms were trained based on trained data.

To address the bias of training using simple data splitting the average score, K-fold cross-validation was done and K was considered as 10 folds. The results showed that average scores of 10-fold cross-validation in six ML algorithms are the same as the simple splitting data process. Subsequently, the hyperparameters were fine-tuned using a hyperparameter tuning technique to enhance the performance of the developed models.

The optimized and selected hyperparameters are documented in Table 1. Next, the MAE, RMSE, and R² values for each optimized model were calculated and represented in Table 2. Finally, all the optimized ML models were compared based on their R² scores in combination with other relevant metrics. In this task, the RF Regressor emerges as the most robust model, demonstrating superior performance with an impressive 95.168% R² value, which indicates it explains nearly 96% of the variance in the data, significantly outperforming other techniques. The Adjusted R^2 metric, which penalizes unnecessary model complexity, closely mirrors the standard R^2 here (95.163%). The model's exceptional performance is evidenced by its lowest Mean Squared Error

(12.548), Mean Absolute Error (2.056), and Root Mean Square Error (3.542), suggesting highly accurate and precise predictions. The superior performance of RF model can be attributed to its ability to handle complex, non-linear relationships in medical data through ensemble learning, where multiple decision trees are combined to create a more flexible and generalized predictive model. In contrast, traditional linear methods like Linear Regression and Support Vector Machines struggled, achieving R^2 values below 53%, which suggests the priority prediction requires sophisticated, non-linear modeling approaches that can capture intricate patterns in medical datasets. The progression from linear to ensemble and advanced machine learning techniques clearly demonstrates the importance of selecting appropriate algorithms for complex predictive challenges in healthcare.

As a result, the RF Regressor model emerged as the top performer among the developed prediction
models. We made this determination based on a comprehensive evaluation of various metrics,
utilizing the best features.

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

3-5- Most important features to select the most appropriate candidate

Upon selecting the best model, we proceeded to identify and weigh the most influential features using the SHAP library within the final model. Initially, a prediction model based on the chosen regression model was created. Subsequently, the importance of each feature was determined by analyzing the set of trees generated by the model using the SHAP technique. The SHAP library assigns a score to each feature based on its impact on the prediction model. The ranking of the variables used in the ultimate model is visually represented in Fig 2, which is a widely recognized and popular chart produced by SHAP.

386 Ultimately, the researchers pinpointed the 11 most effective features, each receiving the highest387 score in candidate prioritization.

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

These features, along with their explanations, are detailed in Table 3. Notably, it showed that factors such as a patient's oxygen consumption and diagnosis played a significant role in prioritizing the waiting list. Additionally, the patient's waiting time on the transplant list emerged as another influential factor. Subsequently, we developed a web-based decision support tool using Python and the Streamlit framework based on the best-fine-tuned model (Fig 3).

4- Discussion

The study aimed to explore the feasibility of utilizing machine learning (ML) methods to predict priority levels for patients on the waiting list for lung transplants and to pinpoint the critical factors influencing priority allocation. Despite the potential advantages of employing ML algorithms in organ allocation [38], there is a lack of research on their application specifically in lung transplantation. This investigation led to the development of a decision support tool for estimating transplantation priorities.

Currently, the decision-making process for prioritizing individuals on organ transplant waiting lists is predominantly reliant on physicians' subjective judgments, often following "first-come, first-served" or "longer waiting time" principles rather than utilizing sophisticated mathematical models [39, 40]. Researchers recommend that authorities explore more equitable and innovative solutions for allocating donor organs to patients on waiting lists. As a result, researchers in the field of transplantation have concentrated on developing advanced models to forecast priority rankings and outcomes for recipients based on pre-transplantation factors [41, 42]. Similarly, we employed ML models to investigate more appropriate factors in assigning organs to recipients.

Prior studies on organ allocation have focused only on classification models to predict the risk of
mortality following transplantation [41, 43]. However, these approaches have not been highly
effective in improving the prioritization of patients on lung transplant waiting lists[44, 45]. In

Page 21 of 41

BMJ Open

411 contrast, our developed model takes into account various factors such as disease type, oxygen412 saturation, demographics, clinical tests, and functional status.

In the context of machine learning (ML), the effectiveness of methods depends not only on their design and techniques but also on the quality and suitability of the data they operate on. To overcome the limitations of prior research, which often relied on a single ML technique and small sample sizes, our study takes a different approach. We incorporate multiple ML techniques to enhance the accuracy of our results, leveraging a large dataset sourced from the United Network for Organ Sharing (UNOS) database.

Our algorithm yields slightly superior results. To enhance the robustness of our model, we employed various data preprocessing techniques and feature engineering methods. These approaches allowed us to identify the most relevant and informative features in the data while discarding redundant or noisy ones [46, 47]. Data preprocessing plays a crucial role in improving data quality and enhancing the accuracy of knowledge extraction [48]. Additionally, by reducing data complexity and dimensions, our models became better equipped to capture underlying patterns and relationships, resulting in improved predictive performance [10, 47]. BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Our analysis reveals that employing the RF regressor model, which incorporates 15 features from the most significant donor and recipient variables available prior to transplantation, represents an effective approach for assigning an allocation score to each candidate on the waiting list. This outperforms other regression models. RF was specifically chosen due to its favorable prediction performance in previous research [49]. Implementing the developed model as an AI-based decision support tool could assist physicians in integrating clinical insights into their decision-making processes and point-of-care scenarios, thereby enhancing the practical utility of the data.

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

ML-based models rely on intricate mathematical structures and multi-dimensional datasets, often yielding complex patterns and relationships that can be challenging for humans to grasp. To address this complexity and limitation, researchers have turned to SHAP (Shapley Additive explanations) summary analysis. This technique identifies the top influential features within the final model. By doing so, it sheds light on which parameters should take precedence when selecting the most suitable recipient with the highest priority—a factor that has not received extensive exploration in prior studies. On the other hand, as the research community increasingly shifts toward explainable AI (XAI) methods [50, 51], the adoption of this approach represents a significant step forward. By employing XAI techniques, the performance of developed models can be interpreted and explained more transparently, fostering greater trust and understanding in their outcomes.

Our study possesses some limitations. Despite the dataset under consideration being of a substantial size, it was obtained from a freely accessible dataset, aligning with the structure of the UNOS database will allow for the collection of patient information tailored to researchers' requirements. While our study demonstrates the effectiveness of the random forest (RF) model in predicting outcomes for lung transplant patients using the UNOS dataset, it is important to note the lack of external validation as a limitation. The model was developed and validated only on the UNOS dataset, which, although comprehensive, may contain biases related to specific populations and practices in the United States. However, we plan to focus on collaborating with international transplant registries or multicenter studies to validate the performance of the model in different populations and healthcare settings. This will enhance the validity of the model and its potential for widespread clinical adoption. External validation on independent datasets from different geographic regions or healthcare systems is essential to ensure the generalizability and robustness

Page 23 of 41

BMJ Open

of our findings. As part of future work, we are developing an intelligent lung transplant patient
information system at our center. Building on previous efforts to apply AI-based techniques in
solid organ transplantation [52-55], this system aims to integrate the current model with patients'
medical records while leveraging additional AI-based models to enhance its performance.

5-Conclusion

461 During this study, we succeeded in developing a priority prediction model based on the huge data 462 of the UNOS database using ML models with the least error. Our research is among the pioneering 463 studies that employ the SHAP method as an XAI technique to enhance the comprehensibility of 464 the proposed model intended for clinicians. Additionally, the automated auxiliary model that we 465 created can assist clinicians in acquiring a better understanding of the transplant priority estimation 466 and the crucial factors that influence patient survival. BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) .

and data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text

6-Declaration

468 Ethics approval and consent to participate

469 The research was approved by the Tehran University of Medical Sciences Ethics Committee
470 (IR.TUMS.IKHC.REC.1401.143). All methods were performed based on the relevant guidelines
471 and regulations. Consent for participation was deemed unnecessary according to an Institutional
472 Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee.

- 473 Consent for publication
- 474 Consent for publication was deemed unnecessary according to an Institutional Review Board
- 475 (IRB) of the Tehran University of Medical Sciences Ethics Committee.
- ⁹ 476 **Declaration of Competing Interest**
- $\frac{1}{2}$ 477 The authors declare that they have no conflict of interest.
 - 478 Availability of data and materials

Page 24 of 41

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES)

and data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text

BMJ Open

> The data used in this article can be obtained from the United Network for Organ Sharing (UNOS) database by visiting www.unos.org/data. However, there are limitations on accessing this data, as it was used under a license for the current study and is not accessible to the general public. The interpretation and reporting of this data are the responsibility of the authors and in no way should be seen as an official policy of or interpretation by the OPTN or the United States government.

484 Funding

This research was funded by the Thoracic Research Center through, Tehran University Medical
Sciences by Grant No (59042). The funding body played no role in the design of the study and
collection, analysis, interpretation of data, and in writing the manuscript.

488 Authors' contributions

H.A., M.G., R.S., and M.A.G. contributed to the conception and design of the study. M.G. and
H.A. acquired the data. M.G., H.A., R.S., and M.A.G. were involved in data interpretation and
analysis. M.G. and H.A. drafted the manuscript. All authors critically revised the manuscript for
important intellectual content and approved the final version to be published. H.A. is the
guarantor.;

494 Acknowledgments

The data reported here have been supplied by the United Network for Organ Sharing (UNOS/OPTN) as the contractor for the Organ Procurement and Transplantation Network. We express our gratitude to the UNOS organization for allowing access to the data. We would like to extend our sincere thanks to the Thoracic Research Center of the Tehran University of Medical Sciences (TUMS) for their support and cooperation during this research.

References

5011.Van der Mark SC, Hoek RAS, Hellemons ME: Developments in lung transplantation over the502past decade. European Respiratory Review 2020, 29(157):190132.

BMJ Open

1			
2			
3	503	2.	Verleden GM, Dupont L, Yserbyt J, Schaevers V, Raemdonck DV, Neyrinck A, Vos R: Recipient
4	504		selection process and listing for lung transplantation. Journal of Thoracic Disease 2017,
5	505		9 (9):3372-3384.
6	506	3.	Smits JM, Nossent G, Evrard P, Lang G, Knoop C, Kwakkel-van Erp JM, Langer F, Schramm R,
/	507		van de Graaf E, Vos R et al: Lung allocation score: the Eurotransplant model versus the revised
8	508		US model – a cross-sectional study . Transplant International 2018, 31 (8):930-937.
9	509	4.	Lancaster TS, Miller JR, Epstein DJ, DuPont NC, Sweet SC, Eghtesady P: Improved waitlist and
10	510		transplant outcomes for pediatric lung transplantation after implementation of the lung
11	511		allocation score J Heart Lung Transplant 2017 36(5):520-528
12	512	5	Safdari R Rezavi S Saeedi S Tanhanour M Gholamzadeh M: Using data mining techniques to
13	513	0.	fight and control enidemics: A sconing review <i>Health and Technology</i> 2021 11(4):759-771
14	514	6	Gholamzadeh M Abtahi H Safdari R [.] Comparison of different machine learning algorithms
15	515	0.	to classify nationts suspected of having sensis infaction in the intensive care unit Informatics
10	515		in Medicing Unlocked 2022 39:101226
12	510	7	Subudhi S. Varma A. Patal A.B. Hardin CC. Khandakar MI. Lea H. MaEvov D. Stulianonoulos T.
10	517 E10	1.	Munn LL Dutte S at al: Comparing machine learning elegrithms for predicting ICU
20	510		admission and mostality in COVID 10 NDLD is Mad 2021 4(1):97
20	519	0	Sofderi D. Daghetingur A. Chalamzadah M. Maghaali V. Annlying data mining tachniques to
22	520	0.	satuali K, Degnatipoul A, Gholanizaden M, Magnoon K. Applying data mining techniques to
23	521		classify patients with suspected nepatitis C virus infection. Intelligent Medicine 2022, 2(4):193-
24	522	0	
25	523	9.	Gnolamzaden M, Abtani H, Satdari K. Machine learning-based techniques to improve lung
26	524		transplantation outcomes and complications: a systematic review. BMC Medical Research
27	525	10	Methodology 2022, 22(1):331.
28	526	10.	Miller PE, Pawar S, Vaccaro B, McCullough M, Rao P, Ghosh R, Warier P, Desai NR, Ahmad T:
29	527		Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset
30	528		Characteristics: Insights From the UNOS Database. Journal of Cardiac Failure 2019,
31	529		25 (6):479-483.
32	530	11.	LeClaire JM, Smith NJ, Chandratre S, Rein L, Kamalia MA, Kohmoto T, Joyce LD, Joyce DL:
33	531		Solid organ donor-recipient race-matching: analysis of the United Network for Organ
34	532		Sharing database. Transpl Int 2021, 34(4):640-647.
35	533	12.	Mazarei A, Sousa R, Mendes-Moreira J, Molchanov S, Ferreira HM: Online boxplot derived
36	534		outlier detection. International Journal of Data Science and Analytics 2025, 19(1):83-97.
37	535	13.	Kalaivani B, Ranichitra A: Unveiling the Impact of Outliers: An Improved Feature
38	536		Engineering Technique for Heart Disease Prediction. In: 2024; Singapore: Springer Nature
39	537		Singapore; 2024: 469-478.
40	538	14.	Theng D, Bhoyar KK: Feature selection techniques for machine learning: a survey of more
41	539		than two decades of research. Knowledge and Information Systems 2024, 66(3):1575-1637.
42	540	15.	Li D, Liu Z, Armaghani DJ, Xiao P, Zhou J: Novel ensemble intelligence methodologies for
43	541		rockburst assessment in complex and variable environments. Sci Rep 2022, 12(1):1844.
44	542	16.	Uyanık GK, Güler N: A Study on Multiple Linear Regression Analysis. Procedia - Social and
45	543		Behavioral Sciences 2013, 106:234-240.
46	544	17.	Kayri M, Kayri I, Gencoglu MT: The performance comparison of Multiple Linear Regression,
4/	545		Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data.
48	546		In: 2017 14th International Conference on Engineering of Modern Electric Systems (EMES): 1-2
49 50	547		June 2017 2017: 2017: 1-4.
50	548	18.	Dai B. Chen RC. Zhu SZ. Zhang WW: Using Random Forest Algorithm for Breast Cancer
52	549		Diagnosis . In: 2018 International Symposium on Computer. Consumer and Control (IS3C): 6-8
52 52	550		Dec. 2018 2018: 2018: 449-452.
55	551	19	Smith PF. Ganesh S. Liu P: A comparison of random forest regression and multiple linear
55	552	17.	regression for prediction in neuroscience Journal of Neuroscience Methods 2013 220(1):85-91
56	332		$-\mathbf{G} = \mathbf{G} =$
57			
58			24
59			_ ·
60			For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 26 of 41

BMJ Open

3	553	20	Yu W Liu T Valdez R Gwinn M Khoury MJ [.] Application of support vector machine modeling
4	554	-0.	for prediction of common diseases: the case of diabetes and pre-diabetes. <i>BMC Medical</i>
5	555		Informatics and Decision Making 2010 10(1):16
6	556	21	Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research
7	557	21.	Directions SN Computer Science 2021 2(3):160
8	558	22	Huang H Wei X Zhou Y. An overview on twin support vector regression Neurocomputing
9	559		2022 490 ·80-92
10	560	23	Bentéjac C Csörgő A Martínez-Muñoz G [.] A comparative analysis of gradient boosting
11	561	23.	algorithms Artificial Intelligence Review 2021 54(3):1937-1967
12	562	24	Li S. Zhang X: Research on orthopedic auxiliary classification and prediction model based on
13	563	27.	XGBoost algorithm Neural Computing and Applications 2020 32 (7):1971-1979
14 15	564	25	Lin I Wu I Lin S Li M Hu K Li K: Predicting mortality of nations with acute kidney injury
15	565	20.	in the ICU using XGBoost model PLOS ONE 2021 16(2):e0246306
10	566	26	Sananmuang T Mankong K Chokeshajusaha K: Multilayer percentron and support vector
18	567	20.	regression models for feline nerturition date prediction Halivon 2024 10(6):e27002
19	568	27	Abiodun OL Jantan A. Omolara AF. Dada KV. Umar AM. Linus OL. Arshad H. Kazaure AA
20	560	21.	Gana II Kiru MII: Comprehensive Review of Artificial Neural Network Applications to
21	570		Pattern Decognition IEEE Access 2010 7:158820 158846
22	570	20	Sarker III: Doon Learning: A Comprehensive Overview on Techniques Tevenomy
23	571	20.	Applications and Besearch Directions SN Computer Science 2021 2(6):420
24	572	20	Abmed SE Alam MSB Hassan M Dozby MD Ishtiak T Dafa N Mofijur M Shawkat Ali ABM
25	575	29.	Condomi AH: Doon loorning modelling techniques: current progress emploations
26	574		oducini Ari. Deep learning moderning techniques: current progress, applications, advantages and shallonges Artificial Intelligence Provide 2022 56(11):12521 12617
27	575	20	LaCun V. Dangia V. Hinton G: Doon loorning. Nature 2015, 50(11):15521-15017.
28	570	50. 21	Demosh A Domemosthy S Duberi SM: Forecosting Spread of COVID 10 Using Degression
29	5//	51.	Algorithm In: Soft Computing for Duchlam Scheing, 2021// 2021, Singgroup: Springer Singenere:
30	578		Algorithm. In. Soft Computing for Problem Solving: 2021/ 2021, Singapore. Springer Singapore,
31	5/9	22	2021. 401-407.
32	580	32.	Lufarmantian Sciences 2022 595:(00.620
33	581	22	Information Sciences 2022, 585:009-029.
34	582	33.	Ekanayake IU, Meddage DPP, Ratnnayake U: A novel approach to explain the black-box hature
35	583		of machine learning in compressive strength predictions of concrete using Shapley additive
36	584	2.4	explanations (SHAP). Case studies in Construction Materials 2022, 16:e01059.
3/	585	34.	Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz K, Himmelfard J, Bansal
30	586		N, Lee SI: From Local Explanations to Global Understanding with Explainable AI for Trees.
29 40	587	25	Nat Mach Intell 2020, 2(1):50-67.
40 //1	588	35.	Rodriguez-Perez R, Bajorath J: Interpretation of machine learning models using shapley
42	589		values: application to compound potency and multi-target activity predictions. Journal of
43	590	26	Computer-Aided Molecular Design 2020, 34 (10):1013-1026.
44	591	36.	Kim Y, Kim Y: Explainable neat-related mortality with random forest and Shapley Additive
45	592	27	exPlanations (SHAP) models. Sustainable Cities and Society 2022, 79:103677.
46	593	37.	Patel Darshan R, Reddy PVB: The Importance of Data Visualization in Exploratory Data
47	594	20	Analysis. Journal of Advanced Zoology 2023, 44(86):923-929.
48	595	38.	Peloso A, Moeckli B, Delaune V, Oldani G, Andres A, Compagnon P: Artificial Intelligence:
49	596	•	Present and Future Potential for Solid Organ Transplantation. Transpl Int 2022, 35:10640.
50	597	39.	Bunnik EM: Ethics of allocation of donor organs. Curr Opin Organ Transplant 2023, 28(3):192-
51	598	40	
52	599	40.	Madwar S: United States officials propose further retreat from first-come, first-served organ
53	600	4.1	donation. <i>Cmaj</i> 2011, 183 (10):E639-640.
54	601	41.	Lau L, Kankanige Y, Rubinstein B, Jones R, Christophi C, Muralidharan V, Bailey J: Machine-
55	602		Learning Algorithms Predict Graft Failure After Liver Transplantation. Transplantation
56	603		2017, 101 (4):e125-e132.
57			
58			25

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

604 605	42.	Gotlieb N, Azhie A, Sharma D, Spann A, Suo N-J, Tran J, Orchanian-Cheff A, Wang B, Goldenberg A, Chassé M <i>et al</i> : The promise of machine learning applications in solid organ
606		transplantation. <i>npi Digital Medicine</i> 2022, 5 (1):89.
607	43.	Jawitz OK, Raman V, Becerra D, Klapper J, Hartwig MG: Factors associated with short- versus
608		long-term survival after lung transplant. J Thorac Cardiovasc Surg 2022, 163(3):853-860.e852.
609	44.	Brahmbhatt JM, Hee Wai T, Goss CH, Lease ED, Merlo CA, Kapnadak SG, Ramos KJ: The lung
610		allocation score and other available models lack predictive accuracy for post-lung transplant
611		survival. J Heart Lung Transplant 2022, 41(8):1063-1074.
612	45.	Dalton JE, Lehr CJ, Gunsalus PR, Mourany L, Valapour M: Refining the Lung Allocation Score
613	10	Models Fails to Improve Discrimination Performance. Chest 2023, 163(1):152-163.
614 615	46.	Pudjinartono N, Fadason I, Kempa-Lienr AW, O'Sullivan JM: A Review of Feature Selection Matheds for Machine Learning Pased Disease Pick Prediction Evolutions in Right Provide Selection
616		2022 2
617	47	Saevs Y Inza I Larrañaga P [.] A review of feature selection techniques in bioinformatics
618	17.	Bioinformatics 2007. 23(19):2507-2517.
619	48.	García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F: Big data preprocessing: methods
620		and prospects. Big Data Analytics 2016, 1(1):9.
621	49.	Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z: Random forest approach
622		for determining risk prediction and predictive factors of type 2 diabetes: large-scale health
623	- 0	check-up data in Japan. BMJ Nutrition, Prevention & amp; Health 2021, 4(1):140-148.
624	50.	S Band S, Yarahmadi A, Hsu C-C, Biyari M, Sookhak M, Ameri R, Dehzangi I, Chronopoulos AT,
625		Liang H-W: Application of explainable artificial intelligence in medical health: A systematic
620	51	Sadaghi Z Alizadehsani R Cifei MA Kausar S Rehman R Mahanta P Bora PK Almasri A
628	51.	Alkhawaldeh RS Hussain S <i>et al</i> : A review of Explainable Artificial Intelligence in healthcare
629		Computers and Electrical Engineering 2024 118 :109370
630	52.	Abtahi H, Shahmoradi L, Amini S, Gholamzadeh M: Design and evaluation of a Mobile-Based
631		decision support system to enhance lung transplant candidate assessment and management:
632		knowledge translation integrated with clinical workflow. BMC Medical Informatics and
633		Decision Making 2023, 23 (1):145.
634	53.	Gholamzadeh M, Safdari R, Amini S, Abtahi H: Feasibility study and determination of
635		prerequisites of telecare programme to enhance patient management in lung transplantation:
636		a qualitative study from the perspective of Iranian healthcare providers. BMJ Open 2023,
637	51	13(6):eU/33/U. Abtabi II. Safdari D. Chalamzadah M: Bragmatia solutions to anhance salf management skills
620	34.	in solid organ transplant patients: systematic review and thematic analysis <i>BMC Primary</i>
640		Care 2022 23(1):166
641	55.	Gholamzadeh M. Abtahi H. Safdari R: Telemedicine in lung transplant to improve patient-
642		centered care: A systematic review. International Journal of Medical Informatics 2022,
643		167 :104861.
611		
044		
645		
		26
		20
		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Tables

646	Tables	
647 [′]	Table 1- The best selec	ted hyperparameters
	Algorithm	Hyperparameters
1	Multiple linear regression	positive= False, n_jobs= 2, fit_intercept= True, copy_X= True
2	Random Forest Regressor	n_estimators= 90, min_samples_split= 2, min_samples_leaf= 1, max_samples 10000,
		max_features: sqrt, max_depth=10
3	SVM Regressor	C =9.11158, loss='epsilon_insensitive', max_iter=5000
4	XGBoost Regressor	subsample=1, min_child_weight= 5, max_depth= 6, learning_rate=0.1, colsample_bytree=0.75
5	MLP	solver= 'sgd', Learning_rate= 'adaptive', hidden_layer_sizes: (20,), alpha: 0.001, activation: logistic
6	DL	Optimizer= 'sgd', batch_size= 16, activation= 'relu'
648		
610		
049		
		27

	Model	R ² (%)	Adjusted R ² (%)	MSE	MAE	RMSE
1	Random Forest Regressor	95.168	95.163	12.548	2.056	3.542
2	XGBoost Regressor	82.88	82.87	58.326	4.487	7.637
3	Deep Learning algorithm	68.736	68.23	80.096	42.096	45.05
4	MLP Regressor	66.003	65.98	88.97	5.681	9.432
5	Linear Regression	52.259	52.23	123.989	6.984	11.131
6	Support Vector Machines	48.590	48.55	133.591	6.570	11.555
651						
652						
002						
653						

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.	BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de I
---	---

	Feature	Description
1	INIT_O2	The amount of oxygen needed when the transplant candidate is the waiting list
2	GROUPING	Lung transplant candidate diagnosis group
3	DAYSWAIT_CHRON	The amount of waiting time of patients on the waiting list - up date waiting time
4	MED_COND_TRR	The status of the patient's lungs at the time of the last clinical evaluation
5	HEMO_SYS_TRR	The latest status of Hemodynamics Pcw (Sys) MM/Hg
6	END O2	O2 Requirement at rest
7	VENTILATOR_TCR	The patient's status in terms of the need for a ventilator
8	LIFE_SUP_TCR	The amount of social and financial support
9	CIG Use	History of cigarette use
10	Vent Support TRR	Episode of ventilatory support
11	Transfusion	Events occurring between listing and transplant

BMJ Open

Figure legends

Fig 1. Schematic diagram illustrating the proposed methodology for developing machine learning models. The process includes data preprocessing, feature engineering, model training, and evaluation, followed by a systematic comparison of multiple models using performance metrics to identify and select the optimal model for deployment.

Fig 2-(a) SHAP summary plot of the top 11 features for predicting lung allocation score using random forest regressor and (b) SHAP values to explain the predicted probabilities

Fig 3- Interactive web-based interface for the machine learning model, developed using the Streamlit framework. The tool allows users to input data, visualize predictions, and explore model performance metrics in real-time, providing an accessible platform for researchers and practitioners to interact with the

developed algorithm

 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml





Priority Score Calculator

Enter Patient Data

Initial Oxygen Level		(?)	Ventilator Usage Score		(?)
50.00	-	+	5	-	+
END_02			Life Support Score		
50.00	-	+	5	-	+
GROUPING		0	Cigarette Use level		
5	-	+	5	-	+
Days Waiting for Chronic Care			Ventilator Support		
10	-	+	5	-	+
Medical Condition Score			Transfusion Score		(?)
5	-	+	5	-	+
Hemodynamic System Score					
5	-	+			

📊 Priority Score

The calculated priority score is: 15.00

Notes:

- The Priority Score is prediced based on the input values.
- Each input field has a specific weight assigned to it.

Fig 3- Interactive web-based interface for the machine learning model, developed using the Streamlit framework. The tool allows users to input data, visualize predictions, and explore model performance metrics in real-time, providing an accessible platform for researchers and practitioners to interact with the developed algorithm

122x136mm (144 x 144 DPI)

Summary sta	atistics of the selected	d continuous preo	dictors (N=)		
	Variable	Range	Mean (SD)	SE	95% Co
	Age	18-58	54.27 (17.30)	0.095	39.297
Lung	BMI	14.997- 44.77	25.3 (3.83)	0.021	25.2586
Lung	FEV1 value	5-120	39.484 (17.30)	0.095	39.297
recipients	Initial creatinine	0.1-24	0.841 (0.407)	0.002	0.8369
	Total Albumin serum	0.5-24	3.8787 (0.406)	0.002	3.8743
Summary sta	atistics of selected ca	tegorical predicto	ors (N=)		
	Variable		n	Percentage (%)	
	Condon	Male	18085	54.86	
	Gender	Female	14881	45.14	
		Α	1513	38.31	
		В	4529	11.32	
Lung	ABU	AB	1513	3.78	
Lung		0	18648	46.59	
recipients	History of	Positive	363	1.10	
		Negative	30061	91.19	
	Manghancy	Unknown	2542	7.71	
	History of previous	Having	808	2.45	
	transplantation	Not Having	32158	97.55	
		Male	14154	42.94	
	GENDER	Female	14154	42.94	
		Unknown	9704	29.44	
		А	279	0.85	
		В	2544	7.72	
Donor	ABO	AB	11832	35.89	
		0	13038	39.55	
		Unknown	279	0.85	
	History of	Positive	9756	29.59	
	Malignanev	Negative	21300	64.61	
	mangnancy	Unknown	1910	5.79	

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.	Enseignement Superieur (ABES) .	MJ Open: first published as 10.1136/bmiopen-2024-089796 on 16 May 2025. Downloaded from http://bmiopen.bmi.com/ on June 7. 2025 at Agence Bibliographique de
--	---------------------------------	--

 Table A-2- Baseline Dataset description from UNOS database

Variables	count	mean	std	min	25%	50%	75%	max
variables	32966	1 9387	1 3160	0	0	3	3	3
GROUPING	52900	25	87	Ū		5	5	5
GENDER	32966	0.5485 96	0.4976 4	0	0	1	1	1
АВО	32966	1.6676 58	1.3853 79	0	0	2	3	3
WGT_KG_TCR	32966	72.416 14	17.805 07	3.1	59.422	72.575	84.822	212
HGT_CM_TCR	32966	168.29 24	12.245 25	5	160.02	168.61 97	175.5	210.82
FUNC_STAT_TCR	32966	2063.0 03	278.18 43	1	2040	2060	2070	4100
DIAB	32966	6.7644 54	72.834 96	0	1	1	1	998
MALIG_TCR	32966	0.1652 31	0.5405 16	0	0	0	0	2
TOT_SERUM_ALBUM	32966	3.8786 76	0.4068 92	0.5	3.8723 02	3.8723 02	3.8723 02	9.8
RESIST_INF	32966	0.1060 49	0.3936	0	0	0	0	2
HEMO_SYS_TCR	32966	42.636 04	16.843 37	0	32	39	47	180
HEMO_PA_DIA_TCR	32966	17.793 79	8.9353 08	0	12	17	21	110
HEMO_PA_MN_TCR	32966	27.511 46	10.971 3	0	21	26	31	110
HEMO_PCW_TCR	32966	10.695 36	5.3746 86	0	7	10	14	50
HEMO_CO_TCR	32966	5.2616 08	1.4071 82	0.2	4.4	5	5.96	15
CIG_USE	32966	0.5656	0.4956 84	0	0	1	1	1
TCR_DUR_ABSTAIN	32966	57.322 61	166.72 62	1	7	7	56	998
LAST_INACT_REASON	32966	5.2636 35	1.5007 13	1	5	5	5	16
INIT_STAT	32966	7043.2 71	178.32 2	7010	7010	7010	7010	7999
INIT_02	32966	4.0979 41	4.1383 58	0	2	3	4.1564 1	35
END_O2	32966	5.4139 94	5.2037 36	0	2.5	4	6	26.3
INIT_CREAT	32966	0.8412 88	0.4078 41	0.1	0.68	0.8	0.98	24
END_CREAT	32966	0.8471 1	0.4216 2	0.08	0.66	0.8	0.99	25
CALC_LAS_LISTDATE	32966	43.041 33	16.167 66	0	33.676 89	37.545 7	45.019 99	96.224 91
DAYSWAIT_CHRON	32966	213.51 44	367.05 53	0	20	73	236	5120

Variables	count	mean	std	min	25%	50%	75%	max
INIT_AGE	32966	54.273 98	14.244 29	0	48	58	64	81
LIFE_SUP_TCR	32966	0.0705 27	0.2560 37	0	0	0	0	1
VENTILATOR_TCR	32966	0.0393	0.1944 14	0	0	0	0	1
INIT_LLU_FLG	32966	0.3566	0.4790 24	0	0	0	1	1
INIT_RLU_FLG	32966	0.3546 38	0.4784 11	0	0	0	1	1
INIT_BLU_FLG	32966	0.8302 49	0.3754 19	0	1	1	1	1
END_LLU_FLG	32966	0.3630 71	0.4808 92	0	0	0	1	1
END_RLU_FLG	32966	0.3613 42	0.4803 97	0	0	0	1	1
END_BLU_FLG	32966	0.8421	0.3646 43	0	1	1	1	1
DR51	32966	19.820 75	38.825 21	0	0	0	0	99
DR51_2	32966	2.9355	16.525 55	0	0	0	0	99
DR52	32966	20.401 29	39.140 27	0	0	0	0	99
DR52_2	32966	2.8566	16.278 53	0	0	0	0	99
DR53	32966	20.145	39.013 53	0	0	0	0	99
DR53_2	32966	2.9363	16.518 35	0	0	0	0	99
DQ1	32966	2.2774 98	20.304 64	0	0	0	1	609
DQ2	32966	3.0090	29.151 05	0	0	0	0	609
MED_COND_TRR	32966	1.8775 71	1.3539 62	0	0	3	3	3
CREAT_TRR	32966	0.8561	0.3541	0.1	0.71	0.8643	0.9	25
DIAL_AFTER_LIST	32966	0.3119	0.4688	0	0	0	1	2
FEV1_TRR	32966	39.484 49	17.302 73	5	26	39.941 8	45	120
HEMO_CO_TRR	32966	5.3365	1.1499	1	4.86	5.3473	5.53	15
HEMO_PA_DIA_TRR	32966	17.906	7.3117	0	14	17.951	19	110
HEMO_PA_MN_TRR	32966	27.369	8.6623 76	0	23	27.433	28	110
HEMO_PCW_TRR	32966	10.680	4.3508	0	9	10.687	12	50
HEMO SYS TRR	32966	42.506	13.378	0	35	42.667	42.667	180

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Variables	count	mean	std	min	25%	50%	75%	max
INFECT_IV_DRUG_TRR	32966	0.4634 47	0.6279 66	0	0	0	1	2
INOTROP_VASO_CO_TRR	32966	0.6707 52	0.5336 64	0	0	1	1	2
INOTROP_VASO_DIA_TRR	32966	0.6995	0.5285	0	0	1	1	2
INOTROP_VASO_MN_TRR	32966	0.6828	0.5329	0	0	1	1	2
INOTROP_VASO_PCW_TRR	32966	0.6769	0.5297	0	0	1	1	2
INOTROP_VASO_SYS_TRR	32966	0.7045	0.5271	0	0	1	1	2
OTH_LIFE_SUP_TRR	32966	0.0138	0.1166	0	0	0	0	1
PCO2_TRR	32966	47.768	10.660 06	10	42.1	47.648	48	120
PRIOR_LUNG_SURG_TRR	32966	0.3923	0.5680	0	0	0	1	2
STEROID	32966	0.9488	0.8256	0	0	1	2	2
TBILI	32966	0.5944	0.6892	0.1	0.4	0.6	0.6113	36
TRANSFUSIONS	32966	0.3857	0.5528	0	0	0	1	2
VENT_SUPPORT_TRR	32966	0.4269	0.6005	0	0	0	1	2
VENTILATOR_TRR	32966	0.0396	0.1950	0	0	0	0	1
INHALED_NO_TRR	32966	0.0054	0.0736	0	0	0	0	1
PRIOR_CARD_SURG_TYPE_O STXT_TRR	32966	94.764 45	3.8366	0	95	95	95	95
PROSTACYCLIN_TRR	32966	0.0041	0.0643	0	0	0	0	1
TRACHEOSTOMY_TRR	32966	0.3548	0.5263	0	0	0	1	2
ECMO_72HOURS	32966	0.7065	0.4923	0	0	1	1	2
INHALEDNO_72HOURS	32966	0.7239	0.5000	0	0	1	1	2
INTUBATED_72HOURS	32966	0.8559	0.5551 87	0	1	1	1	2
HBV_CORE	32966	1.0421	1.3775	0	0	0	3	3
HBV_SUR_ANTIGEN	32966	0.9615	1.3870 78	0	0	0	3	3
HBV_SURF_TOTAL	32966	2.2479	1.2260	0	2	3	3	3
CMV_STATUS	32966	1.7322	1.1952 81	0	0	2	3	3
HIV_SEROSTATUS	32966	0.9831	1.3915	0	0	0	3	3

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Variables	count	mean	std	min	25%	50%	75%	max
HCV_SEROSTATUS	32966	0.9818	1.3788	0	0	0	3	3
EBV_SEROSTATUS	32966	2.1486 08	0.7635 92	0	2	2	3	3
CRSMATCH_DONE	32966	1.5881 51	0.5852 56	0	1	2	2	2
CPRA	32966	747.29 97	431.05 68	0	86	999	999	999
PREV_TX	32966	0.0245	0.1546 29	0	0	0	0	1
PREV_TX_ANY	32966	0.0257 84	0.1584 93	0	0	0	0	1
DA1	32966	9.9479 16	71.844 42	1	2	3	10	6802
DA2	32966	45.965 09	187.60 4	0	24	33	48	6802
DB1	32966	43.607 44	179.94 94	7	13	44	46	550
DB2	32966	71.462	234.55	0	44	60	76	820
DDR1	32966	10.588	29.918 02	1	4	11	11	150
DDR2	32966	23.310	55.512	0	13	15	24	160
RA1	32966	20.054	130.37	0	2	3	23	680
RA2	32966	80.445 43	330.79 92	0	24	68	92	680
RB1	32966	80.574 83	311.74	0	8	44	92	570
RB2	32966	142.11 97	480.61	0	44	61	162	820
RDR1	32966	21.514	82.279 82	0	4	13	24	160
RDR2	32966	46.769	139.00	0	13	17	53	160
AMIS	32966	35.698	46.580	0	1	2	99	99
DRMIS	32966	35.776	46.563	0	1	2	99	99
HLAMIS	32966	37.797	45.080	0	4	5	99	99
MALIG_TRR	32966	0.6257	0.4866	0	0	1	1	2
CMV_IGG	32966	2.3041	1.0777	0	2	3	3	3
CMV_IGM	32966	2.0563	1.2855	0	1	3	3	3
HIST_COCAINE_DON	32966	0.5201	0.6779	0	0	0	1	2
AGE DON	32966	34.465	11.732	6	26	34	40	76
Page 40 c	of 41							
-----------	-------							
-----------	-------							

Variables	count	mean	std	min	25%	50%	75%	max
HBV_CORE_DON	32966	1.9179 15	1.3698 88	0	1	1	4	4
HBV_SUR_ANTIGEN_DON	32966	1.8853 06	1.3676 88	0	1	1	4	4
ABO_DON	32966	2.8215 43	1.3835 85	0	3	3	4	4
ALCOHOL_HEAVY_DON	32966	0.5142 27	0.6755 2	0	0	0	1	2
GENDER_DON	32966	1.0180 79	0.7552 08	0	0	1	2	2
HEP_C_ANTI_DON	32966	1.9044 47	1.3691 03	0	1	1	4	4
ANTIHYPE_DON	32966	0.7437 66	0.7968 65	0	0	1	1	2
BUN_DON	32966	19.683 7	13.757 26	0.4	12	20.214 87	20.214 87	245
CREAT_DON	32966	1.4041 16	1.2691 58	0.07	0.82	1.3	1.4200 9	37
PT_DIURETICS_DON	32966	1.2838 68	0.7875 79	0	1	1	2	2
PT_STEROIDS_DON	32966	1.3480 25	0.7613 86	0	1	2	2	2
PT_T3_DON	32966	0.3065 28	0.4690 17	0	0	0	1	2
PT_T4_DON	32966	1.2296 61	0.8051 9	0	1	1	2	2
PT_OTH2_OSTXT_DON	32966	5972.5 15	2591.6 37	0	4203.2 5	7694	7694	9269
PULM_INF_DON	32966	1.2977 92	1.1749 39	0	0	1	3	3
SGOT_DON	32966	96.904 49	306.02 49	0.3	31	68	97.527 13	20000
SGPT_DON	32966	97.767 02	365.39 33	0.4	25	56	97	44117
TBILI_DON	32966	0.9881	1.0570 65	0	0.6	0.9877 6	0.9877 6	59
URINE_INF_DON	32966	0.9634 78	1.3424 44	0	0	0	3	3
VASODIL_DON	32966	0.5357 64	0.6966 22	0	0	0	1	2
VDRL_DON	32966	2.1988 11	1.8183 14	0	1	1	5	5
CLIN_INFECT_DON	32966	1.2380 03	0.7994	0	1	1	2	2
HIST_CIG_DON	32966	0.4278	0.6079	0	0	0	1	2
HIST_HYPERTENS_DON	32966	0.6235 82	0.7483 24	0	0	0	1	2
HIST_CANCER_DON	32966	0.3236 97	0.4948 65	0	0	0	1	2
DIABETES_DON	32966	0.3953 47	0.5802 36	0	0	0	1	2

Variables	count	mean	std	min	25%	50%	75%	max
HIST_OTH_DRUG_DON	32966	0.8759 02	0.8238 69	0	0	1	2	2
CMV_DON	32966	2.7614 21	1.1451 94	0	1	3	4	4
DDAVP_DON	32966	0.5305 47	0.6931	0	0	0	1	2
HEPARIN_DON	32966	1.6638 66	0.5103	0	1	2	2	2
ARGININE_DON	32966	1.2218 65	0.8072 42	0	1	1	2	2
WGT_KG_DON_CALC	32966	77.364 88	14.876 9	23.5	70	77.353 48	81.6	189
BMI_DON_CALC	32966	26.190 33	4.5802 42	10.588 66	23.667 83	26.221 13	27.173 1	66.035 64
HBV_NAT_DON	32966	2.0281 81	1.4016 18	0	0	3	3	3
ABO_MAT	32966	1.6387 79	0.9052 58	1	1	1	3	3
DIAL_PRIOR_TX	32966	0.6500 64	0.4834 63	0	0	1	1	2
ISCHTIME	32966	5.2922 57	1.5282 11	0.042	4.5664 06	5.3519 05	5.6328 13	25
O2_REQ_CALC	32966	5.3750 88	4.2890 68	0	3	5.4579 95	5.4579 95	26.3
PRIOR_CARD_SURG_TRR	32966	0.3284	0.4908 72	0	0	0	1	2
MALIG	32966	0.4118	0.5984	0	0	0	1	2
HGT_CM_CALC	32966	169.99 18	8.2559 88	122	165.1	170.03 08	175	210.82
BMI_CALC	32966	25.299 95	3.8325 76	14.997 85	23.382 51	25.392 36	27.331 17	44.777 87
DISTANCE	32966	211.25 96	210.94 01	0	67	215	221	4137
VENT_SUPPORT_AFTER_LIS T	32966	0.4269	0.6005	0	0	0	1	2
PROTEIN_URINE	32966	0.8621	0.8219 72	0	0	1	2	2
CARDARREST_NEURO	32966	0.3973 79	0.5680 28	0	0	0	1	2
PO2	32966	380.25 4	121.57 52	3.2	368	382.31 52	457	754
HIST_MI	32966	0.3272	0.4933	0	0	0	1	2
LV_EJECT	32966	58.086 24	9.4403 97	1	58	58.118 68	61	99
CORONARY_ANGIO	32966	2.0919 74	1.3279 38	1	1	1	4	4
BIOPSY_DGN	32966	3.0819	1.9978 58	1	1	5	5	5
HBSAB_DON	32966	3.8937	1.4628	0	3	3	6	6

Variables	count	mean	std	min	25%	50%	75%	max
EBV_IGG_CAD_DON	32966	4.8062 85	1.5472 43	0	4	4	7	7
EBV_IGM_CAD_DON	32966	2.7079 72	2.2522 66	0	1	1	6	6
CDC_RISK_HIV_DON	32966	0.5345 81	0.6988 12	0	0	0	1	2
INOTROP_SUPPORT_DON	32966	0.9822 54	0.8363 9	0	0	1	2	2
TRANSFUS_TERM_DON	32966	294.91 72	454.81 1	0	0	1	998	998
PO2_FIO2_DON	32966	86.126 1	21.280 58	1	86	100	100	100
PCO2_DON	32966	37.029 91	5.5849 25	10	34.7	37	39	110
BRONCHO_LT_DON	32966	321.94 15	464.84	1	2	2	998	998
BRONCHO_RT_DON	32966	331.79	468.46	1	2	2	998	998
CHEST_XRAY_DON	32966	300.36	455.52 74	1	2	5	999	999
PH_DON	32966	7.4153	0.0835	5	7.4	7.4155	7.44	8
HEMATOCRIT_DON	32966	29.215 83	4.4611	2.5	27	29.149 96	30.6	71

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

Identifying the most influential factors in lung transplant patients using a multivariate prediction model: An analysis of UNOS datasets

Journal:	BMJ Open
Manuscript ID	bmjopen-2024-089796.R2
Article Type:	Original research
Date Submitted by the Author:	06-Apr-2025
Complete List of Authors:	Gholamzadeh, Marsa; Tehran University of Medical Sciences, Safdari, Reza; Tehran University of Medical Sciences, Asadi Gharabaghi, Mehrnaz; Tehran University of Medical Sciences, Abtahi, Hamidreza; Tehran University of Medical Sciences, Pulmonary and Critical Care Department; Tehran University of Medical Sciences, Thoracic Research Center
Primary Subject Heading :	Health informatics
Secondary Subject Heading:	Health informatics, Respiratory medicine, Health services research, Intensive care
Keywords:	Machine Learning, Pulmonary Disease < Lung Diseases, Transplant medicine < INTERNAL MEDICINE





I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our <u>licence</u>.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which <u>Creative Commons</u> licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

terez oni

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Identifying the most influential factors in lung transplant patients using a multivariate prediction model: An analysis of UNOS datasets Marsa Gholamzadeh¹, Reza Safdari¹, Mehrnaz AsadiGharabaghi², Hamidreza Abtahi^{3,*} 1. Health Information Management and Medical Informatics Department, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran. 2. Department of Pulmonary Medicine, Faculty of Medicine, Tehran University of Medical Sciences, Tehran, Iran. 3. Pulmonary and Critical Care Department, Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Tehran, Iran. *Corresponding author: Hamidreza Abtahi E-mail address: hrabtahi2020research@gmail.com Tel: +9821-66192646 Postal address: Thoracic Research Center, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Qarib Ave, Keshavarz Blv, Tehran, Iran. 7.64 **ORCID ID:** Marsa Gholamzadeh, https://orcid.org/0000-0001-6781-9342; Hamidreza Abtahi, https://orcid.org/0000-0002-1111-0497; Reza Safdari, https://orcid.org/0000-0002-4982-337X Mehrnaz Asadi Gharabaghi: https://orcid.org/0000-0003-0852-1532 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

59

2		
3 4	29	Abstract
5 6	30	Objectives : In lung transplantation, a priority is assigned to each candidate on the waiting list. Our
7 8	31	primary objective was to identify the key factors that influence the allocation of priorities in lung
9 10	32	transplantation using machine learning (ML) techniques to enhance the process of prioritizing
11 12 13	33	patients.
14 15	34	Design: Developing a prediction model.
16 17	35	Setting and participants: Our data was retrieved from the UNOS open-source database of
18 19	36	transplant patients between 2005 and 2023.
20 21 22	37	Interventions: After the preprocessing process, a feature engineering technique was employed to
23 24	38	select the most relevant features. Then, six ML models with an optimized hyper-parameter
25 26	39	including Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector
27 28 20	40	Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model, and a deep
30 31	41	learning model (DL) were developed based on UNOS dataset.
32 33	42	Primary and secondary outcome measures: The performance of each model was evaluated
34 35 26	43	using R-squared (R ²) and other error rate metrics. Next, the Shapley Additive Explanations
36 37 38	44	(SHAP) technique was utilized to identify the most important features in the prediction.
39 40	45	Results: The raw dataset contains 196,270 records with 545 features in all organs. After
41 42	46	preprocessing, 32,966 records with 15 features remain. Among various models, the RF model
43 44 45	47	achieved a high R2 score. Additionally, the RF model exhibited the lowest error values indicating
46 47	48	its superior precision compared to other regression models SHAP technique in conjunction with
48 49	49	the RF model revealed the 11 most important features for priority allocation. Subsequently, we
50 51 52	50	developed a web-based decision support tool using Python and the Streamlit framework based on
52 53 54 55	51	the best-fine-tuned model.
56 57 58		2

<text> **Conclusion**: The deployment of the ML model has the potential to act as an automated tool to aid physicians in assessing the priority of lung transplants and identifying significant factors that play a role in patient survival. Keywords: Lung transplantation, allocation score, Machine learning, Prediction. For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1 2			
2 3 4	57	S	trengths and limitations of this study:
5 6	58	•	The use of various preprocessing and data cleaning techniques in our survey increased the
7 8 9	59		robustness and performance of the model.
10 11	60	•	To ensure transparency and interpretability in our machine learning models, we employed
12 13	61		Explainable Artificial Intelligence (XAI) techniques, specifically the SHAP (Shapley Additive
14 15 16	62		Explanations) method.
17 18	63	•	Understanding the factors influencing the determination of lung transplant priority could
19 20 21	64		support clinicians in designing treatment plans and thus improving the quality of life of patients.
22 23	65	•	Deploying the developed ML model in the form of a decision support system increases its
24 25	66		applicability in clinical practice.
26 27 28	67	•	The model was validated on the UNOS dataset but requires external validation across diverse
29 30	68		populations and healthcare systems to ensure generalizability and clinical applicability.
31 32	69		
33 34 25	70		
36 27			
37 38 39			
40 41			
42 43			
44 45			
46			
47 48			
49			
50 51			
52			
53			
54 55			
56			
57			
58 50			4
60			For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

1-Introduction

Lung transplantation (LTx) is an advanced treatment option for patients suffering from end-stage lung disease. When no other treatment options are available and the patient is likely to die, lung transplant surgery is suggested as a well-established treatment option [1]. When a patient meets the inclusion criteria for transplantation, they are placed on a waiting list and assigned a priority. Various conditions may affect eligibility for lung transplantation and the patient's priority [2]. In some countries, a score is assigned to each patient on the waiting list to enhance the recipient selection process [3, 4]. Understanding the most influential factors in priority allocation for lung transplantation is beneficial for researchers worldwide, as it can improve post-transplant survival. Utilizing data mining methods and developing forecasting models in this field could aid clinicians in uncovering hidden patterns and relationships within patient data and allocation scores. Machine learning (ML) methods have been developed across various fields of clinical medicine to assist clinicians in predicting and classifying diseases [5]. These methods are used to predict the

84 length of stay in the Intensive Care Unit (ICU), diagnose septic infection[6], and extract disease 85 patterns from big data [7, 8]. Nevertheless, there is a lack of studies on the development of 86 predictive models and identification of important features using ML methods to predict lung 87 transplantation priority [9, 10]. Thus, the primary objective of this study was to utilize ML 88 techniques to identify the most influential factors that strongly impacted outcomes based on 89 various developed ML methods to predict the priority using clinical and demographic data.

2- Methods

91 Throughout this section, the process of developing, comparing, and evaluating ML models is
92 shown schematically in Fig 1. Python programming language version 10 was used in this study
93 for developing and validating ML algorithms. For data preprocessing, Numpy and Pandas modules

1 2	
- 3 4	9
5 6	9
7 8	9
9 10	9
11 12	9
13 14 15	٥
15 16 17	5
17 18 10	10
19 20 21	10
21 22 22	10
23 24 25	10
25 26 27	10
28 29	10
30 31	10
32 33	10
34 35	10
36 37	10
38 39	10
40 41	11
42 43	11
44 45	11
46 47	11
48 49 50	11
50 51 52	11
52 53 54	11
55 56	ΤT
57	
58 59	

60

were employed, while the sci-kit learn library was utilized for developing supervised classifier algorithms.

96 **2-1-Dataset description and data retrieval**

97 The data for this study were obtained from the United Network for Organ Sharing (UNOS) online 98 database [11]. Upon receiving written permission from UNOS, we accessed the recorded data 99 pertaining to lung transplantation for our research. Our study included patients over 18 years old 100 with end-stage lung disease who underwent lung transplants between 2005 and 2022. We 101 performed a waiting list analysis using all available data entries from the United Network for 102 Organ Sharing (UNOS) database for our study.

103 The priority of candidates on the waiting list was considered as the outcome, while the clinical and
 104 demographic characteristics of patients were considered as features or predictors.

105 **2-2-Pre-processing process**

Data pre-processing is a crucial step in ML techniques, especially when dealing with raw data from clinical databases or medical records that often contain missing or unclear information. To ensure the development of more accurate models based on appropriate data, we followed a series of data pre-processing steps. The following steps were employed in this phase as pre-processing techniques.

- 111 1- Checking the duplicated values and records to remove the duplicates
- ⁴⁵ 112 2- De-identify records and remove irreverent features
- 113 3- Convert nominal and categorical features to numerical values
- ¹⁹ 114 4- Identify missing data and missing values imputation
- 57 115 5- Outlier detection
 - 116 6- Feature engineering and feature selection

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

117 7- Data transformation and normalization

2-2-1-Duplicate checking and removal of irreverent features

After duplicate checking, we consulted UNOS guidelines and experts to review all features and their definitions. Under their supervision, we removed identification variables and irrelevant features, such as ID columns, hospital center identification codes, and country of residence, to deidentify patients.

Following this, we converted the post-transplant survival days variable to years and excluded
patients with a survival rate of less than two years. Next, we filtered out patients below 18 years
of age and excluded any data before 2005. Additionally, records related to heart transplantation
were removed from the dataset.

After removing irrelevant features in the first data-cleaning phase, we utilized the discretized operator to convert nominal values to numerical data. Categorical data were encoded using the LabelEncoder class too.

2-2-2- Missing data management:

To address missing data in our dataset, we conducted missing data imputation across the entire dataset. Initially, we assessed the specified columns or attributes to determine the extent of missing and unique data in each column. During this analysis, we discovered that the ICU column was empty and decided to delete it due to its lack of meaningful information. To impute the missing data, a threshold of 80% was set for feature removal with expert consultation. As a result, any column with more than 80% missing data was removed. If the missingness is due to inconsistent reporting rather than clinical irrelevance, dropping the column could exclude critical information about high-risk patients. In this case, domain experts might recommend retaining the column and using advanced imputation techniques or creating a binary indicator for missingness.

Page 9 of 42

BMJ Open

Along with feature removal, the sensitivity analysis was conducted on dataset which revealed that the inclusion of omitted features had a detrimental effect on the performance of the Random Forest Regressor model. Specifically, these features led to a decrease in the R² score from 0.95 to 0.68 and an increase in RMSE from 3.5 to 4.8. It means that the sensitivity analysis demonstrated that omitting features with missing data significantly degraded model performance (R² decreased from 0.95 to 0.68; RMSE increased from 3.5 to 4.8), suggesting these missing values were not missing at random (MNAR). The observed performance drop implies that the missingness may depend on unobserved factors systematically related to the outcome, warranting MNAR-specific methods for robust inference.

For the missing data imputation, we adopted a strategy of replacing missing data in numerical features with the mean value of each respective feature. This approach allows us to retain the integrity of the dataset while minimizing the impact of missing data on our analysis. By performing these comprehensive steps of missing data imputation, we ensure the dataset is optimized for further analysis and modeling, enabling us to draw more accurate conclusions and insights.

2-2-3- Outliers handling

Outliers can significantly impact the performance and interpretability of machine learning models. Therefore, it is essential to investigate their causes before deciding whether to exclude, transform, or retain them. This exploration ensures that the preprocessing steps are justified and scientifically sound. To address outliers in the dataset, we first create distribution plots to visualize the data. Next, we apply the IQR method and use Box plots to identify outliers. Finally, we remove these outliers to prepare the data for further processing.

161 To address outliers in the dataset, a comprehensive approach was employed that included visual162 and statistical analysis to identify and understand the nature of the outliers. The distribution plots

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

and boxplots were created to visualize outliers and applied the IQR method to quantify their extent.
Additionally, statistical analysis was conducted to assess the impact of outliers on the dataset and
performed sensitivity analysis to evaluate their influence on model performance. Throughout this
process, we ensured transparency and justification by documenting all outliers and providing
context-specific reasons for their exclusion, such as data entry errors or clinically irrelevant
extreme values [12, 13]. This rigorous approach ensured that the removal of outliers was
methodologically sound and did not compromise the integrity of our analysis.

2-2-4-Feature engineering and feature selection

Given the high-dimensional nature of the dataset used in this study, feature engineering and selection were critical steps to reduce dimensionality, eliminate irrelevant or redundant features, and enhance model performance. Feature selection aims to identify a subset of features that effectively describe the problem with minimal loss of information and computational efficiency [14]. All phases of feature engineering and selection were conducted under the supervision of clinical experts to ensure the relevance and validity of the selected features.

Step one, correlation analysis: As the first step, correlation analysis was performed to assess the relationships between features and the target variable, as well as inter-feature correlations. This analysis helped identify highly correlated features that could introduce multicollinearity and redundancy into the model using heatmap graph. Features with a correlation coefficient above a predefined threshold were flagged for further evaluation.

Step two, variance threshold filtering: To eliminate low-variance features that contribute little
 to the model's predictive power, a variance threshold was applied. Features with variance below a
 specified threshold (e.g., 0.01) were removed, as they were deemed to have minimal impact on the
 target variable.

Page 11 of 42

BMJ Open

Step three, embedded feature selection with XGBoost: Following the initial filtering, an embedded feature selection technique was employed using the XGBoost algorithm. XGBoost provides intrinsic feature importance scores based on metrics such as gain, cover, and frequency. Features with very low importance scores (negative scores indicating no correlation with target value) were excluded from the final feature set. Step four, expert review and validation: All selected features were reviewed and validated by subject matter experts to ensure their clinical, practical, and scientific relevance. This step was critical to avoid eliminating features that, although statistically significant, may not be clinically significant. For example, some features were retained for model customization based on expert consultation, despite having modest statistical significance. Through a comprehensive and expert-guided feature selection process, we identified a subset of features that were statistically significant, domain-relevant, and impactful for model performance. This rigorous approach ensured the final model was both robust and clinically meaningful, with the selected features deemed critical for predicting the target variable. **2-2-5-Data transformation and normalization:** In the end, data normalization was carried out to optimize the features for modeling purposes. 2-3- Splitting data and validation technique During the model development process, the dataset was divided into training and testing data in an 80:20 ratio where 80% of the data was used for training the models and the remaining 20% was reserved for testing and validation. The training dataset is used to train the model, allowing it to learn patterns and relationships within the data based on the available data. The training dataset typically contains the bulk of the available data. In contrast, the testing dataset is intended solely to evaluate the model's performance on unseen data, ensuring an unbiased assessment of its

Page 12 of 42

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) .

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

BMJ Open

generalizability. This dataset is kept separate from the training process to provide a true measure
of how the model performs in real-world scenarios. Both datasets are often split randomly, with
common ratios such as 80:20 or 70:30, depending on the size and nature of the data.

This split ensured that the models were evaluated on unseen data to assess their generalization capability. To mitigate potential bias introduced by simple data splitting, cross-validation technique was employed. Through this technique, the dataset was divided into k folds, and each model was trained and validated k times, with each fold serving as the validation set once. The average performance metrics across all folds were calculated to ensure robust evaluation. The results of the k-fold cross-validation were consistent with those obtained from the simple 80:20 split, confirming the reliability of the initial approach.

2-4- Model development and tuning

The objective of this study was to develop a prediction model for a continuous numerical variable (priority score) using regression techniques to identify the most effective factors in selecting the most appropriate candidate for LTx. In this study, regression models were selected to examine the connection between input variables and output numerical values, as the target variable (outcome) is a continuous numerical value.

To determine the most influential factors and identify the best model, the performance of six regression-based models was evaluated: Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Machines (SVM) Regressor, XGBoost Regressor, a multilayer perceptron model (MLP—a class of feedforward artificial neural network), and a deep learning model (DL). The selection of these models was done based on the type of target variable and the study objectives.

A hyperparameter tuning optimization technique was employed in this phase to improve model performance by optimizing the training process by determining the best hyperparameters for each model. This technique was used to prevent models that underfit or overfit the data [15]. After tuning parameters in each model, the models were trained with updated best hyperparameters, and all metrics were calculated again to achieve the best performance. We employed the random search method, a hyperparameter tuning technique where hyperparameters are randomly chosen from a predefined set to train a model.

2-4-1- Multiple linear regression (MLR)

Multiple linear regression (MLR) is a statistical technique used to estimate the relationship between a dependent variable and one or more independent variables. It is an extension of linear regression, which requires more than one predictor variable to forecast the response variable[16]. MLR is a significant regression algorithm that models the linear association between a dependent continuous variable and multiple independent variables [17]. Hence, we have chosen this model to predict the continuous variable (priority score) based on several independent variables. The equation for multiple linear regression is demonstrated below[17]:

 $v = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 (1)$

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

where y represents the priority; x_i is the considered variables; β_0 is the intercept; and β^i is the regression coefficients.

2-4-2- Random Forest Regressor (RF)

The random forest (RF) Regressor algorithm is a kind of ML approach that employs a group of decision trees, which are trained on a subset of the data, to make predictions. This technique is designed to stabilize the algorithm and decrease variance by using multiple trees. The RF regressor algorithm is widely recognized as a popular model in developing regression models because of its strong performance with large datasets and diverse data types [18, 19].

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

255 2-4-3- Support Vector Machines Regressor (SVM)

Support vector machine regression (SVM) is a versatile regression function that can be used to solve both classification and regression problems. SVM is a supervised learning algorithm that fits a regression to the training data by reducing the distance between the sampled points and the fitted hyperplane[20, 21]. One advantage of SVM is that it is a sparse algorithm, meaning that it only needs information from a limited number of data points[22].

261 2-4-4- XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is another ML library that is available for free and offers a powerful and efficient implementation of the gradient boosting algorithm[23]. Gradient boosting is a technique that involves creating an ensemble of tree-based models and then combining them to create a more accurate overall model than any of the individual models in the sequence[24]. XGBoost is a popular choice for those who require an effective and optimized implementation of gradient boosting[25].

268 2-4-5- Multilayer Perceptron Model (MLP)

The Multilayer Perceptron (MLP) is considered one of the top regression models in the field of artificial neural networks. It is equipped with the capability to learn from training data using a variety of training algorithms and rules. This feature allows the MLP to acquire numerous advantages, including increased capacity. As a result, the MLP operates as a self-regulating model that utilizes specific learning algorithms to enhance its performance when encountering new inputs [26, 27].

2-4-6- Deep Learning Model (DL)

A deep learning model can be used for regression problems by learning a mapping from input
features to the target output. Deep learning is an adaptable model proficient at effectively managing
intricate data relationships. It proves especially beneficial when working with extensive datasets

Page 15 of 42

BMJ Open

where traditional regression methods might not uncover intricate patterns. Nonetheless, to prevent
overfitting and attain peak performance, these models necessitate meticulous calibration and
validation[28-30]. Occasionally, due to the complex nature of implementing these models, simpler
regression models may outperform them.

2-5- Performance evaluation

Typically, regression models are evaluated based on a function that measures the difference between the predicted and actual numerical value of the target variable, such as the priority score[31]. In this study, three popular evaluation metrics were used, including mean absolute error (MAE), root mean square error (RMSE), and R-squared (R²) score to assess the performance of the developed models [32].

(1)
$$MAE = \frac{\sum_{i=1}^{n} |R^*(i) - R(i)|}{n}$$

(2) $RMSE = \left\{ \frac{\sum_{i=1}^{n} (R^*(i) - R(i))^2}{n} \right\}^{1/2}$
(3) $R^2 = 1 - \frac{\sum_{i=1}^{n} (R^*(i) - R(i))^2}{n}$

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) .

and data mining, Al training, and similar technologies

Protected by copyright, including for uses related to text

³ 251 $\sum_{i=1}^{n} (R^*(i) - m(i))^2$ ⁴ 292 In these formulas; variable n refers to the number of samples; $R^*(i)$ denotes the retrieved value ⁶ predicted by the model; R(i) denotes the analyzed value; and m(i) denotes the average analyzed ⁹ 294 value.

To validate the developed machine learning (ML) models and reduce bias, we employed k-fold cross-validation. This technique overcomes the limitations of a simple train/test split by dividing the available data into multiple folds or subsets. By averaging the results across these folds, we achieve a more robust estimate of the model's performance compared to a simple train/test split.

2-6- Feature Importance

To implement Explainable AI (XAI), a technique employed to elucidate the impact of each feature
on the model is the SHAP (Shapley Additive Explanations) method. The SHAP method aims to

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) .

Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

enhance the transparency and interpretability of machine learning models by drawing on cooperative game theory [33]. For instance, linear models utilize their coefficients to gauge the significance of each feature. However, these coefficients are influenced by the scale of the variable itself, potentially resulting in misinterpretations [34]. The same can be found in tree-based models for feature ranking. This is precisely why SHAP (Shapley Additive Explanations) becomes valuable for model interpretation [35]. The absolute value of SHAP provides insight into how significantly an individual feature influences the prediction [36]. Once we identify the optimal model for priority prediction, we'll leverage the SHAP technique to assign weights to the most critical features. These features will then be ranked based on their importance and impact on the final priority score.

312 Patient and public involvement

Patients and/or the public were not involved in the design, or conduct, or reporting, ordissemination plans of this research.

3-Results

3-1- Dataset description

The raw dataset including all organs comprises information on 196,270 patients who underwent lung transplantation, as well as data related to lung donors. This comprehensive dataset includes 545 features, encompassing demographic and clinical details about the organ recipients, biomarkers, laboratory test results, and characteristics of the donated organs (Table A-2 in Appendix A). Additionally, it provides insights into various patient outcomes, such as posttransplant survival rates, occurrences of acute organ rejection, priority levels, duration of intensive care unit stay, post-transplant infections, and instances of re-transplantation. Page 17 of 42

BMJ Open

To preprocess the dataset, we converted the transplantation date data type to a string format and extracted the year column by parsing the month, hour, and year components. We removed all data prior to 2005 due to the absence of a prioritization system during that period. Furthermore, to focus exclusively on adult transplants, we excluded information related to pediatric transplants for children and adolescents under 18 years of age. As a result, our initial dataset comprised 183,086 records.

Subsequently, we filtered the dataset to include only post-transplant survival records exceeding
one year. After that, we eliminated any records with missing priority scores. Ultimately, our final
dataset consisted of 45,966 records for subsequent analysis.

3-2- Exploratory data analysis

Following imputation of missing independent variable data and preprocessing steps, the overall patient population consisted of 66.88% men and 33.20% women, with a median age of 54.27±14.24 years. Our target variable is the priority (or allocation) score, which represents a continuous numerical value. In Table A-1 in Appendix A, we present descriptive analysis and the frequency distribution of various demographic and clinical variables within the dataset. BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Furthermore, we employed a data visualization method to enhance our comprehension of the data and dataset. This approach aids in verifying the integrity of the data and detecting any apparent inaccuracies. Incorporating data visualization is essential for all data science projects across various fields [37].

7 343 **3-3- Data cleaning and preprocessing**

In the method section, we present detailed information regarding preprocessing procedures applied to the whole dataset. After de-identifying the dataset, it was limited to 40,024 records and 445 features. During the initial data cleaning phase, we removed irrelevant features, reducing the total

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

to 322. Subsequently, we performed missing values imputation, resulting in 215 features available
for further analysis. Next, in the correlation analysis phase, our dataset contains over 165 features
post-pre-processing.

Due to the dataset's high dimensionality, we applied pre-processing techniques to select only the most important features based on their importance scores. As a result, we narrowed down the dataset to 65 features in the first phase of feature engineering. After further applying feature engineering and selection techniques, our final dataset consisted of 32,966 records, containing 15 features.

3-4- Development and evaluation of regression models

The prediction models were developed by training several selected features obtained during the feature engineering phase. We used 80% of the dataset to train the algorithms and the rest 20% to test and validate their efficacy (80:20) and all six regression algorithms were trained based on trained data.

To address the bias of training using simple data splitting the average score, K-fold crossvalidation was done and K was considered as 10 folds. The results showed that average scores of 10-fold cross-validation in six ML algorithms are the same as the simple splitting data process. Subsequently, the hyperparameters were fine-tuned using a hyperparameter tuning technique to enhance the performance of the developed models.

The optimized and selected hyperparameters are documented in Table 1. Next, the MAE, RMSE, and R^2 values for each optimized model were calculated and represented in Table 2. Finally, all the optimized ML models were compared based on their R^2 scores in combination with other relevant metrics. In this task, the RF Regressor emerges as the most robust model, demonstrating superior performance with an impressive 95.168% R^2 value, which indicates it explains nearly Page 19 of 42

BMJ Open

96% of the variance in the data, significantly outperforming other techniques. The Adjusted R^2 metric, which penalizes unnecessary model complexity, closely mirrors the standard R² here (95.163%). The model's exceptional performance is evidenced by its lowest Mean Squared Error (12.548), Mean Absolute Error (2.056), and Root Mean Square Error (3.542), suggesting highly accurate and precise predictions. The superior performance of RF model can be attributed to its ability to handle complex, non-linear relationships in medical data through ensemble learning, where multiple decision trees are combined to create a more flexible and generalized predictive model. In contrast, traditional linear methods like Linear Regression and Support Vector Machines struggled, achieving R² values below 53%, which suggests the priority prediction requires sophisticated, non-linear modeling approaches that can capture intricate patterns in medical datasets. The progression from linear to ensemble and advanced machine learning techniques clearly demonstrates the importance of selecting appropriate algorithms for complex predictive challenges in healthcare.

As a result, the RF Regressor model emerged as the top performer among the developed prediction models. We made this determination based on a comprehensive evaluation of various metrics, utilizing the best features. BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

⁰ 386 **3-5- Most important features to select the most appropriate candidate**

Upon selecting the best model, we proceeded to identify and weigh the most influential features using the SHAP library within the final model. Initially, a prediction model based on the chosen regression model was created. Subsequently, the importance of each feature was determined by analyzing the set of trees generated by the model using the SHAP technique. The SHAP library assigns a score to each feature based on its impact on the prediction model. The ranking of the

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

variables used in the ultimate model is visually represented in Fig 2, which is a widely recognizedand popular chart produced by SHAP.

394 Ultimately, the researchers pinpointed the 11 most effective features, each receiving the highest395 score in candidate prioritization.

These features, along with their explanations, are detailed in Table 3. Notably, it showed that factors such as a patient's oxygen consumption and diagnosis played a significant role in prioritizing the waiting list. Additionally, the patient's waiting time on the transplant list emerged as another influential factor. Subsequently, we developed a web-based decision support tool using Python and the Streamlit framework based on the best-fine-tuned model (Fig 3).

4- Discussion

The study aimed to explore the feasibility of utilizing machine learning (ML) methods to predict priority levels for patients on the waiting list for lung transplants and to pinpoint the critical factors influencing priority allocation. Despite the potential advantages of employing ML algorithms in organ allocation [38], there is a lack of research on their application specifically in lung transplantation. This investigation led to the development of a decision support tool for estimating transplantation priorities.

408 Currently, the decision-making process for prioritizing individuals on organ transplant waiting lists 409 is predominantly reliant on physicians' subjective judgments, often following "first-come, first-410 served" or "longer waiting time" principles rather than utilizing sophisticated mathematical models 411 [39, 40]. Researchers recommend that authorities explore more equitable and innovative solutions 412 for allocating donor organs to patients on waiting lists. As a result, researchers in the field of 413 transplantation have concentrated on developing advanced models to forecast priority rankings

and outcomes for recipients based on pre-transplantation factors [41, 42]. Similarly, we employed
ML models to investigate more appropriate factors in assigning organs to recipients.

Prior studies on organ allocation have focused only on classification models to predict the risk of mortality following transplantation [41, 43]. However, these approaches have not been highly effective in improving the prioritization of patients on lung transplant waiting lists[44, 45]. In contrast, our developed model takes into account various factors such as disease type, oxygen saturation, demographics, clinical tests, and functional status.

In the context of machine learning (ML), the effectiveness of methods depends not only on their design and techniques but also on the quality and suitability of the data they operate on. To overcome the limitations of prior research, which often relied on a single ML technique and small sample sizes, our study takes a different approach. We incorporate multiple ML techniques to enhance the accuracy of our results, leveraging a large dataset sourced from the United Network for Organ Sharing (UNOS) database. BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Our algorithm yields slightly superior results. To enhance the robustness of our model, we employed various data preprocessing techniques and feature engineering methods. These approaches allowed us to identify the most relevant and informative features in the data while discarding redundant or noisy ones [46, 47]. Data preprocessing plays a crucial role in improving data quality and enhancing the accuracy of knowledge extraction [48]. Additionally, by reducing data complexity and dimensions, our models became better equipped to capture underlying patterns and relationships, resulting in improved predictive performance [10, 47].

Our analysis reveals that employing the RF regressor model, which incorporates 15 features from
the most significant donor and recipient variables available prior to transplantation, represents an
effective approach for assigning an allocation score to each candidate on the waiting list. This

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

outperforms other regression models. RF was specifically chosen due to its favorable prediction
performance in previous research [49]. Implementing the developed model as an AI-based decision
support tool could assist physicians in integrating clinical insights into their decision-making
processes and point-of-care scenarios, thereby enhancing the practical utility of the data.

ML-based models rely on intricate mathematical structures and multi-dimensional datasets, often yielding complex patterns and relationships that can be challenging for humans to grasp. To address this complexity and limitation, researchers have turned to SHAP (Shapley Additive explanations) summary analysis. This technique identifies the top influential features within the final model. By doing so, it sheds light on which parameters should take precedence when selecting the most suitable recipient with the highest priority—a factor that has not received extensive exploration in prior studies. On the other hand, as the research community increasingly shifts toward explainable AI (XAI) methods [50, 51], the adoption of this approach represents a significant step forward. By employing XAI techniques, the performance of developed models can be interpreted and explained more transparently, fostering greater trust and understanding in their outcomes.

Our study possesses some limitations. Despite the dataset under consideration being of a substantial size, it was obtained from a freely accessible dataset, aligning with the structure of the UNOS database will allow for the collection of patient information tailored to researchers' requirements. While our study demonstrates the effectiveness of the random forest (RF) model in predicting outcomes for lung transplant patients using the UNOS dataset, it is important to note the lack of external validation as a limitation. The model was developed and validated only on the UNOS dataset, which, although comprehensive, may contain biases related to specific populations and practices in the United States. However, we plan to focus on collaborating with international

Page 23 of 42

BMJ Open

transplant registries or multicenter studies to validate the performance of the model in different populations and healthcare settings. This will enhance the validity of the model and its potential for widespread clinical adoption. External validation on independent datasets from different geographic regions or healthcare systems is essential to ensure the generalizability and robustness of our findings. As part of future work, we are developing an intelligent lung transplant patient information system at our center. Building on previous efforts to apply AI-based techniques in solid organ transplantation [52-55], this system aims to integrate the current model with patients' medical records while leveraging additional AI-based models to enhance its performance.

5-Conclusion

469 During this study, we succeeded in developing a priority prediction model based on the huge data 470 of the UNOS database using ML models with the least error. Our research is among the pioneering 471 studies that employ the SHAP method as an XAI technique to enhance the comprehensibility of 472 the proposed model intended for clinicians. Additionally, the automated auxiliary model that we 473 created can assist clinicians in acquiring a better understanding of the transplant priority estimation 474 and the crucial factors that influence patient survival.

6-Declaration

0 476 Ethics approval and consent to participate

The research was approved by the Tehran University of Medical Sciences Ethics Committee
(IR.TUMS.IKHC.REC.1401.143). All methods were performed based on the relevant guidelines
and regulations. Consent for participation was deemed unnecessary according to an Institutional
Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee.

481 Consent for publication

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de Enseignement Superieur (ABES). Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies

BMJ Open

Consent for publication was deemed unnecessary according to an Institutional Review Board (IRB) of the Tehran University of Medical Sciences Ethics Committee. **Declaration of Competing Interest** The authors declare that they have no conflict of interest. Availability of data and materials The data used in this article can be obtained from the United Network for Organ Sharing (UNOS) database by visiting www.unos.org/data. However, there are limitations on accessing this data, as it was used under a license for the current study and is not accessible to the general public. The interpretation and reporting of this data are the responsibility of the authors and in no way should be seen as an official policy of or interpretation by the OPTN or the United States government. Funding This research was funded by the Thoracic Research Center through, Tehran University Medical Sciences by Grant No (59042). The funding body played no role in the design of the study and collection, analysis, interpretation of data, and in writing the manuscript. **Authors' contributions** H.A., M.G., R.S., and M.A.G. contributed to the conception and design of the study. M.G. and H.A. acquired the data. M.G., H.A., R.S., and M.A.G. were involved in data interpretation and analysis. M.G. and H.A. drafted the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version to be published. H.A. is the guarantor.; Acknowledgments The data reported here have been supplied by the United Network for Organ Sharing (UNOS/OPTN) as the contractor for the Organ Procurement and Transplantation Network. We

1			
2			
3 4	505	expres	ss our gratitude to the UNOS organization for allowing access to the data. We would like to
5			
6	506	extend	d our sincere thanks to the Thoracic Research Center of the Tehran University of Medical
7			
8	507	Scienc	ces (TUMS) for their support and cooperation during this research.
9			
10	508	Refer	ences
11			
12	509	1.	Van der Mark SC, Hoek RAS, Hellemons ME: Developments in lung transplantation over the
15 1/	510	•	past decade. European Respiratory Review 2020, 29 (157):190132.
14	511	2.	Verleden GM, Dupont L, Yserbyt J, Schaevers V, Raemdonck DV, Neyrinck A, Vos R: Recipient
16	512		selection process and listing for lung transplantation. Journal of Thoracic Disease 2017,
17	513	•	9(9):3372-3384.
18	514	3.	Smits JM, Nossent G, Evrard P, Lang G, Knoop C, Kwakkel-van Erp JM, Langer F, Schramm R,
19	515		van de Graaf E, Vos R <i>et al</i> : Lung allocation score: the Eurotransplant model versus the revised
20	516	4	US model – a cross-sectional study. <i>Transplant International</i> 2018, 31(8):930-937.
21	517	4.	Lancaster 1S, Miller JR, Epstein DJ, DuPont NC, Sweet SC, Eghtesady P: Improved waitlist and
22	518		transplant outcomes for pediatric lung transplantation after implementation of the lung
23	519	~	allocation score. J Heart Lung Transplant 2017, 36 (5):520-528.
24	520	5.	Satdari R, Rezayi S, Saeedi S, Tanhapour M, Gholamzadeh M: Using data mining techniques to
25	521	6	fight and control epidemics: A scoping review. Health and Technology 2021, 11(4):/59-7/1.
26	522	6.	Gholamzadeh M, Abtahi H, Safdari R: Comparison of different machine learning algorithms
27	523		to classify patients suspected of having sepsis infection in the intensive care unit. Informatics
28	524	7	in Medicine Unlocked 2023, 38 :101236.
29	525	1.	Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, McEvoy D, Stylianopoulos I,
30 21	526		Munn LL, Dutta S et al. Comparing machine learning algorithms for predicting ICU
27	527	0	admission and mortality in COVID-19. NPJ Digit Med 2021, 4(1):87.
32	528	8.	Sandari R, Degnatipour A, Gnolamzaden M, Magnooli K. Applying data mining techniques to
34	529		classify patients with suspected nepatitis C virus infection. Intelligent Medicine 2022, 2(4):193-
35	530	0	198. Chalamzadah M. Ahtahi H. Safdari D. Mashing laguning bagad taahniguag ta imputatio lung
36	551	9.	transplantation outcomes and complications: a systematic review <i>DMC</i> Madical Passarch
37	552		Methodology 2022, 22(1):221
38	555	10	Methodology 2022, 22(1).551. Miller DE Dawar S. Vaccaro P. McCullouch M. Bao D. Chach P. Wariar D. Dasai ND. Ahmad T:
39	554 525	10.	Predictive Abilities of Machine Learning Techniques May Be Limited by Detect
40	525		Characteristics: Insights From the UNOS Database Journal of Cardiac Eailure 2010
41	550		25 (6):470 483
42	538	11	LeClaire IM Smith NL Chandratre S Rein L Kamalia MA Kohmoto T Joyce ID Joyce DI
43	530	11.	Solid organ donor-recipient race-matching: analysis of the United Network for Organ
44	540		Sharing database Transpl Int 2021 34(A):640-647
45	5/1	12	Mazarej A Sousa R Mendes-Moreira I Molchanov S Ferreira HM: Online boxnlot derived
40 47	541	12.	outlier detection International Journal of Data Science and Analytics 2025 19(1):83-97
47 //8	5/2	13	Kalaivani B Ranichitra A: Unvailing the Impact of Outliers: An Improved Feature
40 49	543	15.	Figure Technique for Heart Disease Prediction In: 2024: Singapore: Springer Nature
50	545		Singanore: 2024: 469-478
51	546	14	Theng D Bhovar KK: Feature selection techniques for machine learning: a survey of more
52	547	17.	than two decades of research Knowledge and Information Systems 2024 66(3):1575_1637
53	5/18	15	Li D. Liu Z. Armaghani DI. Xiao P. Zhou I: Novel ensemble intelligence methodologies for
54	540	10.	rockhurst assessment in complex and variable environments. Sci Ron 2022 12(1):1844
55	545		2000000000000000000000000000000000000
56			
57			
58			24
59			

1			
2			
3	550	16	Uyanık GK Güler N [.] A Study on Multiple Linear Regression Analysis <i>Procedia - Social and</i>
4	551		Rehavioral Sciences 2013 106:234-240
5	552	17	Kavri M. Kavri I. Generadu MT: The performance comparison of Multiple Linear Regression
6	552	17.	Dandam Forest and Artificial Noural Network by using photovoltais and atmospheric data
7	555		Kanuoin Forest and Artificial Neural Network by using photovoltaic and atmospheric data.
8	554		III. 2017 14th International Conference on Engineering of Modern Electric Systems (EMES): 1-2
9	555	10	June 2017 2017; 2017: 1-4.
10	556	18.	Dai B, Chen RC, Zhu SZ, Zhang WW: Using Random Forest Algorithm for Breast Cancer
11	557		Diagnosis . In: 2018 International Symposium on Computer, Consumer and Control (IS3C): 6-8
12	558		<i>Dec. 2018 2018</i> ; 2018: 449-452.
13	559	19.	Smith PF, Ganesh S, Liu P: A comparison of random forest regression and multiple linear
14	560		regression for prediction in neuroscience . Journal of Neuroscience Methods 2013, 220 (1):85-91.
15	561	20.	Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ: Application of support vector machine modeling
16	562		for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical
17	563		Informatics and Decision Making 2010, 10(1):16.
18	564	21.	Sarker IH: Machine Learning: Algorithms, Real-World Applications and Research
19	565		Directions SN Computer Science 2021 2(3):160
20	566	22	Huang H Wei X Zhou Y. An overview on twin support vector regression Neurocomputing
21	567	<i>22</i> .	2022 490.80-02
22	569	22	2022, 470.00-72. Bortáina C. Csörgő A. Martínaz Muñaz G: A comparativa analysis of gradient boosting
23	500	23.	algorithms, Artificial Intelligence Device 2021 54(2):1027 1067
24	509	24	algorithms. Artificial Intelligence Review 2021, 54(5):1957-1907.
25	570	24.	Li S, Zhang X: Research on orthopedic auxiliary classification and prediction model based on
26	5/1		XGBoost algorithm . Neural Computing and Applications 2020, 32 (7):1971-1979.
27	572	25.	Liu J, Wu J, Liu S, Li M, Hu K, Li K. Predicting mortality of patients with acute kidney injury
28	573		in the ICU using XGBoost model. <i>PLOS ONE</i> 2021, 16(2):e0246306.
29	574	26.	Sananmuang T, Mankong K, Chokeshaiusaha K: Multilayer perceptron and support vector
30	575		regression models for feline parturition date prediction. <i>Heliyon</i> 2024, 10(6):e27992.
31	576	27.	Abiodun OI, Jantan A, Omolara AE, Dada KV, Umar AM, Linus OU, Arshad H, Kazaure AA,
32	577		Gana U, Kiru MU: Comprehensive Review of Artificial Neural Network Applications to
33	578		Pattern Recognition. IEEE Access 2019, 7:158820-158846.
34	579	28.	Sarker IH: Deep Learning: A Comprehensive Overview on Techniques, Taxonomy,
35	580		Applications and Research Directions. SN Computer Science 2021, 2(6):420.
36	581	29	Ahmed SF Alam MSB Hassan M Rozbu MR Ishtiak T Rafa N Mofiliur M Shawkat Ali ABM
37	582		Gandomi AH. Deen learning modelling techniques: current progress applications
38	583		advantages and challenges Artificial Intelligence Review 2023 56(11):13521-13617
39	58/	30	LeCun V Bengio V Hinton G: Deen learning Nature 2015, 521(7553):436-444
40	504 E0E	21	Domash A. Damamaarthy S. Duhari SM: Equations Spread of COVID 10 Using Degression
41	202	51.	Algorithm In Set Computing for Ducklow Solving 2021// 2021, Singer and Sminger Singer and
42	580		Algorithm. In. soli Computing for Problem Solving: 2021/ 2021; Singapore. Springer Singapore,
43	587	22	
44	588	32.	Karunasingha DSK: Root mean square error or mean absolute error? Use their ratio as well.
45	589		Information Sciences 2022, 585:609-629.
46	590	33.	Ekanayake IU, Meddage DPP, Rathnayake U: A novel approach to explain the black-box nature
40	591		of machine learning in compressive strength predictions of concrete using Shapley additive
47 48	592		explanations (SHAP). Case Studies in Construction Materials 2022, 16:e01059.
40	593	34.	Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal
50	594		N, Lee SI: From Local Explanations to Global Understanding with Explainable AI for Trees.
51	595		Nat Mach Intell 2020, 2(1):56-67.
52	596	35.	Rodríguez-Pérez R, Bajorath J: Interpretation of machine learning models using shapley
52	597	•	values: application to compound potency and multi-target activity predictions. Journal of
55	598		Computer-Aided Molecular Design 2020 34 (10):1013-1026
54	590	36	Kim V Kim V: Explainable heat-related mortality with random forest and SHanley Additive
55	600	50.	av Planations (SHAP) models Systemable Cities and Society 2022 70.102677
57	000		cal lanatons (SIIAI) mouchs. Sustainable Cities and Society 2022, 17.103077.
58			25
50			(.)

BMJ Open

2			
3	601	37.	Patel Darshan R. Reddy PVB: The Importance of Data Visualization in Exploratory Data
4	602		Analysis Journal of Advanced Zoology 2023 44(S6):923-929
5	603	38	Peloso A Moeckli B Delaune V Oldani G Andres A Compagnon P. Artificial Intelligence.
6	604	50.	Present and Future Potential for Solid Organ Transplantation Transpl Int 2022 35:10640
7	605	39	Bunnik FM [•] Ethics of allocation of donor organs <i>Curr Opin Organ Transplant</i> 2022, 33:10010.
8	606	57.	106
9	607	40	170. Madwar S: United States officials propose further retreat from first some first served organ
10	609	40.	denotion Creat 2011 192(10):E620.640
11	600	41	Lou I. Kankaniga V. Dubinstein D. Jones D. Christophi C. Muralidharan V. Dailay I: Machine
12	610	41.	Lau L, Kankanige I, Kuonisteni D, Jones K, Christophi C, Mutanunaran V, Baney J. Machine-
13	610		Learning Algorithms Predict Grait Failure Alter Liver Transplantation. Transplantation
14	611	40	2017, 101(4):e125-e132.
15	612	42.	Gouleo N, Aznie A, Snarma D, Spann A, Suo N-J, Iran J, Orchanian-Chell A, wang B,
16	613		Goldenberg A, Chasse M et al. The promise of machine learning applications in solid organ
1/	614	10	transplantation. npj Digital Medicine 2022, 5(1):89.
18	615	43.	Jawitz OK, Raman V, Becerra D, Klapper J, Hartwig MG: Factors associated with short- versus
19	616		long-term survival after lung transplant. J Thorac Cardiovasc Surg 2022, 163(3):853-860.e852.
20	617	44.	Brahmbhatt JM, Hee Wai T, Goss CH, Lease ED, Merlo CA, Kapnadak SG, Ramos KJ: The lung
21	618		allocation score and other available models lack predictive accuracy for post-lung transplant
22	619		survival. J Heart Lung Transplant 2022, 41(8):1063-1074.
25 24	620	45.	Dalton JE, Lehr CJ, Gunsalus PR, Mourany L, Valapour M: Refining the Lung Allocation Score
24	621		Models Fails to Improve Discrimination Performance. <i>Chest</i> 2023, 163 (1):152-163.
25	622	46.	Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM: A Review of Feature Selection
20	623		Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics
27	624		2022, 2 .
20	625	47.	Saeys Y, Inza I, Larrañaga P: A review of feature selection techniques in bioinformatics.
30	626		Bioinformatics 2007, 23(19):2507-2517.
31	627	48.	García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F: Big data preprocessing: methods
32	628		and prospects. Big Data Analytics 2016, 1(1):9.
33	629	49.	Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z: Random forest approach
34	630		for determining risk prediction and predictive factors of type 2 diabetes: large-scale health
35	631		check-up data in Japan. BMJ Nutrition, Prevention & amp; Health 2021, 4(1):140-148.
36	632	50.	S Band S, Yarahmadi A, Hsu C-C, Biyari M, Sookhak M, Ameri R, Dehzangi I, Chronopoulos AT,
37	633		Liang H-W: Application of explainable artificial intelligence in medical health: A systematic
38	634		review of interpretability methods. Informatics in Medicine Unlocked 2023, 40:101286.
39	635	51.	Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, Bora PK, Almasri A,
40	636		Alkhawaldeh RS, Hussain S <i>et al</i> : A review of Explainable Artificial Intelligence in healthcare.
41	637		Computers and Electrical Engineering 2024, 118 :109370.
42	638	52.	Abtahi H, Shahmoradi L, Amini S, Gholamzadeh M: Design and evaluation of a Mobile-Based
43	639		decision support system to enhance lung transplant candidate assessment and management:
44	640		knowledge translation integrated with clinical workflow. BMC Medical Informatics and
45	641		Decision Making 2023. 23(1):145.
46	642	53.	Gholamzadeh M. Safdari R. Amini S. Abtahi H: Feasibility study and determination of
47	643		prerequisites of telecare programme to enhance nationt management in lung transplantation:
48	644		a qualitative study from the perspective of Iranian healthcare providers <i>BMI Open</i> 2023
49	645		13(6):e073370
50	646	54	Abtahi H Safdari R Gholamzadeh M [•] Pragmatic solutions to enhance self-management skills
51 52	647		in solid organ transplant patients: systematic review and thematic analysis <i>BMC Primary</i>
52 53	648		Care 2022 23(1):166
55	649	55	Gholamzadeh M. Abtahi H. Safdari R [.] Telemedicine in lung transplant to improve patient-
55	650		centered care: A systematic review International Journal of Medical Informatics 2022
56	651		167 ·104861
57	55 I		
58			26
59			
60			For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	Algorithm	Hyperparameters
1	Multiple linear regression	positive= False, n_jobs= 2, fit_intercept= True, copy_X= True
2	Random Forest Regressor	n_estimators= 90, min_samples_split= 2, min_samples_leaf= 1, max_samples 10000,
		max_features: sqrt, max_depth=10
3	SVM Regressor	C =9.11158, loss='epsilon_insensitive', max_iter=5000
4	XGBoost Regressor	subsample=1, min_child_weight= 5, max_depth= 6, learning_rate=0.1, colsample_bytree=0.75
5	MLP	solver= 'sgd', Learning_rate= 'adaptive', hidden_layer_sizes: (20,), alpha: 0.001, activation: logistic
6	DL	Optimizer= 'sgd', batch_size= 16, activation= 'relu'
656		
657		
		28

1 2 3	с го Т
4	658 I
5 6	
7	
8	1
9 10	$\frac{2}{3}$
11 12	4
12	5
14 15	6
15 16	659
17	660
18 19	
20	661
21 22	
23	
24 25	
26	
27 28	
29	
30 31	
32	
33 34	
34 35	
36	
37 38	
39	
40 41	
42	
43 44	
45	
46 47	
48	
49 50	
50 51	
52	
53 54	
55	
56 57	

658 Table 2- The evaluation	metrics of develope	d models and com	narison of the	model performance
0.00 1 0.00 2^{-1} 1 fie evaluation	i metries of develope	a models and com	iparison or the	model performance

	Model	R ² (%)	Adjusted R ² (%)	MSE	MAE	RMSE
1	Random Forest Regressor	95.168	95.163	12.548	2.056	3.542
2	XGBoost Regressor	82.88	82.87	58.326	4.487	7.637
3	Deep Learning algorithm	68.736	68.23	80.096	42.096	45.05
4	MLP Regressor	66.003	65.98	88.97	5.681	9.432
5	Linear Regression	52.259	52.23	123.989	6.984	11.131
6	Support Vector Machines	48.590	48.55	133.591	6.570	11.555
59						
60						
/00						
61						

662			
663 T	able 3- Th	e top 11 features identified by the	e SHAP method based on the prediction model
	#	Feature	Description
	1	INIT_O2	The amount of oxygen needed when the transplant candidate is o the waiting list
_	2	GROUPING	Lung transplant candidate diagnosis group
	3	DAYSWAIT_CHRON	The amount of waiting time of patients on the waiting list - up-to date waiting time
	4	MED_COND_TRR	The status of the patient's lungs at the time of the last clinical evaluation
_	5	HEMO_SYS_TRR	The latest status of Hemodynamics Pcw (Sys) MM/Hg
	6	END_O2	O2 Requirement at rest
	7	VENTILATOR_TCR	The patient's status in terms of the need for a ventilator
	8	LIFE_SUP_TCR	The amount of social and financial support
	9	CIG_Use	History of cigarette use
_	10	Vent_Support_TRR	Episode of ventilatory support
_	11	Transfusion	Events occurring between listing and transplant
			30

BMJ Open: first published as 10.1136/bmjopen-2024-089796 on 16 May 2025. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Figure legends

Fig 1. Schematic diagram illustrating the proposed methodology for developing machine learning models. The process includes data preprocessing, feature engineering, model training, and evaluation, followed by a systematic comparison of multiple models using performance metrics to identify and select the optimal model for deployment.

Fig 2-(a) SHAP summary plot of the top 11 features for predicting lung allocation score using random forest regressor and (b) SHAP values to explain the predicted probabilities

Fig 3- Interactive web-based interface for the machine learning model, developed using the Streamlit framework. The tool allows users to input data, visualize predictions, and explore model performance metrics in real-time, providing an accessible platform for researchers and practitioners to interact with the developed algorithm




Priority Score Calculator

Enter Patient Data

Initial Oxygen Level		0	Ventilator Usage Score		3
50.00	-	+	5	-	+
END_02			Life Support Score		
50.00	-	+	5	-	+
GROUPING		(?)	Cigarette Use level		
5	-	+	5	~	+
Days Waiting for Chronic Care			Ventilator Support		
10	-	+	5	-	+
Medical Condition Score			Transfusion Score		0
5	-	+	5	-	+
Hemodynamic System Score					
5	-	+			

📊 Priority Score

The calculated priority score is: 15.00

📝 Notes:

- The Priority Score is prediced based on the input values.
- Each input field has a specific weight assigned to it.

Fig 3- Interactive web-based interface for the machine learning model, developed using the Streamlit framework. The tool allows users to input data, visualize predictions, and explore model performance metrics in real-time, providing an accessible platform for researchers and practitioners to interact with the developed algorithm

122x136mm (144 x 144 DPI)

Table A-1- Dataset description

	Variabla	Danca	Moor (SD)	SE	050/ Card
	variable	Kange	Mean (SD)	SE	95% Con
	Age	18-58	54.27 (17.30)	0.095	39.2977
Lung	BMI	14.997- 44.77	25.3 (3.83)	0.021	25.2586
e raciniants	FEV1 value	5-120	39.484 (17.30)	0.095	39.2977
recipients	Initial creatinine	0.1-24	0.841 (0.407)	0.002	0.8369
	Total Albumin serum	0.5-24	3.8787 (0.406)	0.002	3.8743
Summary sta	ntistics of selected ca	tegorical predicto	ors (N=)		
	Variable		n	Percentage (%)	
	Gurden	Male	18085	54.86	
	Gender	Female	14881	45.14	
		Α	1513	38.31	
		В	4529	11.32	
r	ABO	AB	1513	3.78	
Lung		0	18648	46.59	
recipients		Positive	363	1.10	
	History of	Negative	30061	91.19	
	Malignancy	Unknown	2542	7.71	
	History of previous	Having	808	2.45	
	transplantation	Not Having	32158	97.55	
		Male	14154	42.94	
	GENDER	Female	14154	42.94	
		Unknown	9704	29.44	
		А	279	0.85	
		В	2544	7.72	
Donor	ABO	AB	11832	35.89	
		0	13038	39.55	
		Unknown	279	0.85	
		Positive	9756	29.59	
	History of	Negative	21300	64.61	
	Malignancy	Unknown	1010	5 70	

1 2
2
4
5
6
7
ð Q
10
11
12
13
14
15
17
18
19
20
21
22
24
25
26
27
28 29
30
31
32
33
34 25
36
37
38
39
40 41
41 42
43
44
45
46
4/ /9
40 49
50
51
52
53
54 55
56
57
58
59
60

Table A-2- Baseline Dataset des	cription	from UP	NOS data	abase				
Variables	count	mean	std	min	25%	50%	75%	max
GROUPING	32966	1.9387 25	1.3160 87	0	0	3	3	3
GENDER	32966	0.5485 96	0.4976 4	0	0	1	1	1
ABO	32966	1.6676 58	1.3853 79	0	0	2	3	3
WGT_KG_TCR	32966	72.416 14	17.805 07	3.1	59.422	72.575	84.822	212
HGT_CM_TCR	32966	168.29 24	12.245 25	5	160.02	168.61 97	175.5	210.82
FUNC_STAT_TCR	32966	2063.0 03	278.18 43	1	2040	2060	2070	4100
DIAB	32966	6.7644 54	72.834 96	0	1	1	1	998
MALIG_TCR	32966	0.1652 31	0.5405 16	0	0	0	0	2
TOT_SERUM_ALBUM	32966	3.8786 76	0.4068 92	0.5	3.8723 02	3.8723 02	3.8723 02	9.8
RESIST_INF	32966	0.1060 49	0.3936	0	0	0	0	2
HEMO_SYS_TCR	32966	42.636 04	16.843 37	0	32	39	47	180
HEMO_PA_DIA_TCR	32966	17.793 79	8.9353 08	0	12	17	21	110
HEMO_PA_MN_TCR	32966	27.511 46	10.971 3	0	21	26	31	110
HEMO_PCW_TCR	32966	10.695 36	5.3746 86	0	7	10	14	50
HEMO_CO_TCR	32966	5.2616 08	1.4071 82	0.2	4.4	5	5.96	15
CIG_USE	32966	0.5656	0.4956 84	0	0	1	1	1
TCR_DUR_ABSTAIN	32966	57.322 61	166.72 62	1	7	7	56	998
LAST_INACT_REASON	32966	5.2636 35	1.5007 13	1	5	5	5	16
INIT_STAT	32966	7043.2 71	178.32 2	7010	7010	7010	7010	7999
INIT_02	32966	4.0979 41	4.1383 58	0	2	3	4.1564 1	35
END_O2	32966	5.4139 94	5.2037 36	0	2.5	4	6	26.3
INIT_CREAT	32966	0.8412 88	0.4078 41	0.1	0.68	0.8	0.98	24
END_CREAT	32966	0.8471 1	0.4216 2	0.08	0.66	0.8	0.99	25
CALC_LAS_LISTDATE	32966	43.041 33	16.167 66	0	33.676 89	37.545 7	45.019 99	96.224 91
DAYSWAIT_CHRON	32966	213.51 44	367.05 53	0	20	73	236	5120

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Variables	count	mean	std	min	25%	50%	75%	max
INIT_AGE	32966	54.273 98	14.244 29	0	48	58	64	81
LIFE_SUP_TCR	32966	0.0705 27	0.2560 37	0	0	0	0	1
VENTILATOR_TCR	32966	0.0393	0.1944 14	0	0	0	0	1
INIT_LLU_FLG	32966	0.3566	0.4790 24	0	0	0	1	1
INIT_RLU_FLG	32966	0.3546 38	0.4784 11	0	0	0	1	1
INIT_BLU_FLG	32966	0.8302 49	0.3754 19	0	1	1	1	1
END_LLU_FLG	32966	0.3630 71	0.4808 92	0	0	0	1	1
END_RLU_FLG	32966	0.3613	0.4803 97	0	0	0	1	1
END_BLU_FLG	32966	0.8421	0.3646 43	0	1	1	1	1
DR51	32966	19.820 75	38.825 21	0	0	0	0	99
DR51_2	32966	2.9355 09	16.525 55	0	0	0	0	99
DR52	32966	20.401 29	39.140 27	0	0	0	0	99
DR52_2	32966	2.8566 1	16.278 53	0	0	0	0	99
DR53	32966	20.145 18	39.013 53	0	0	0	0	99
DR53_2	32966	2.9363 59	16.518 35	0	0	0	0	99
DQ1	32966	2.2774 98	20.304 64	0	0	0	1	609
DQ2	32966	3.0090 7	29.151 05	0	0	0	0	609
MED_COND_TRR	32966	1.8775 71	1.3539 62	0	0	3	3	3
CREAT_TRR	32966	0.8561 59	0.3541 49	0.1	0.71	0.8643	0.9	25
DIAL_AFTER_LIST	32966	0.3119 58	0.4688 32	0	0	0	1	2
FEV1_TRR	32966	39.484 49	17.302 73	5	26	39.941 8	45	120
HEMO_CO_TRR	32966	5.3365 05	1.1499 13	1	4.86	5.3473 03	5.53	15
HEMO_PA_DIA_TRR	32966	17.906 18	7.3117	0	14	17.951 62	19	110
HEMO_PA_MN_TRR	32966	27.369 64	8.6623 76	0	23	27.433 37	28	110
HEMO_PCW_TRR	32966	10.680 73	4.3508 63	0	9	10.687 58	12	50
HEMO_SYS_TRR	32966	42.506 21	13.378 96	0	35	42.667 5	42.667 5	180

Variables	count	mean	std	min	25%	50%	75%	max
INFECT IV DRUG TRR	32966	0.4634	0.6279	0	0	0	1	2
	22055	47	66					
INOTROP_VASO_CO_TRR	32966	0.6707	0.5336	0	0	1	1	2
	32966	0.6995	0.5285	0	0	1	1	2
INOTROP_VASO_DIA_TRR	52700	39	4	0		1	1	2
NOTROR VACO MN TRR	32966	0.6828	0.5329	0	0	1	1	2
INUTROP_VASO_MIN_TRR		55	5					
INOTROP VASO PCW TRR	32966	0.6769	0.5297	0	0	1	1	2
	22066	4	01		0	1	1	
INOTROP_VASO_SYS_TRR	32966	0.7045	0.5271	0	0	1	1	2
	32966	0.0138	0.1166	0	0	0	0	1
OTH_LIFE_SUP_TRR	52700	02	71	0		Ū	Ū	1
	32966	47.768	10.660	10	42.1	47.648	48	120
PC02_IRR		1	06			14		
PRIOR LUNG SURG TRR	32966	0.3923	0.5680	0	0	0	1	2
Thiok_Long_song_ink		74	63					-
STEROID	32966	0.9488	0.8256	0	0	1	2	2
	32066	0 5044	43	0.1	0.4	0.6	0.6113	36
TBILI	32900	13	98	0.1	0.4	0.0	84	30
	32966	0.3857	0.5528	0	0	0	1	2
TRANSFUSIONS			85					
VENT SUPPORT TRR	32966	0.4269	0.6005	0	0	0	1	2
VENT_SUITORI_IRR		85	65					
VENTILATOR TRR	32966	0.0396	0.1950	0	0	0	0	1
_	22066	1/	0 0726	0	0	0	0	1
INHALED_NO_TRR	52900	0.0034	92	0	0	0	0	1
PRIOR CARD SURG TYPE O	32966	94 764	3 8366	0	95	95	95	95
STXT TRR	02,000	45	07		10	20	10	20
BDOSTACVCI IN TDD	32966	0.0041	0.0643	0	0	0	0	1
		56	32					
TRACHEOSTOMY TRR	32966	0.3548	0.5263	0	0	0	1	2
	22066	5	56	0		1	1	2
ECMO_72HOURS	32966	0.7065	0.4923	0	0	1	1	2
	32966	0 7239	0 5000	0	0	1	1	2
INHALEDNO_72HOURS	52900	28	37	Ű	Ů		1	-
INTUDATED 711010S	32966	0.8559	0.5551	0	1	1	1	2
INTUBATED_72HOURS		42	87					
HBV CORE	32966	1.0421	1.3775	0	0	0	3	3
	22055	04	25					
HBV_SUR_ANTIGEN	32966	0.9615	1.3870	0	0	0	3	3
	32966	2 2479	1 2260	0	2	3	3	3
HBV_SURF_TOTAL	52700	52	29			5	5	5
	32966	1.7322	1.1952	0	0	2	3	3
		7	81					
HIV SEROSTATUS	32966	0.9831	1.3915	0	0	0	3	3
		34	81					

Variables	count	mean	std	min	25%	50%	75%	max
HCV_SEROSTATUS	32966	0.9818	1.3788 63	0	0	0	3	3
EBV_SEROSTATUS	32966	2.1486 08	0.7635 92	0	2	2	3	3
CRSMATCH_DONE	32966	1.5881	0.5852	0	1	2	2	2
CPRA	32966	747.29 97	431.05	0	86	999	999	999
PREV_TX	32966	0.0245	0.1546	0	0	0	0	1
PREV_TX_ANY	32966	0.0257 84	0.1584 93	0	0	0	0	1
DA1	32966	9.9479	71.844	1	2	3	10	6802
DA2	32966	45.965	187.60	0	24	33	48	6802
DB1	32966	43.607	179.94 94	7	13	44	46	5501
DB2	32966	71.462	234.55	0	44	60	76	8201
DDR1	32966	10.588	29.918 02	1	4	11	11	1501
DDR2	32966	23.310	55.512	0	13	15	24	1602
RA1	32966	20.054	130.37	0	2	3	23	6801
RA2	32966	80.445 43	330.79 92	0	24	68	92	6802
RB1	32966	80.574 83	311.74	0	8	44	92	5703
RB2	32966	142.11 97	480.61	0	44	61	162	8201
RDR1	32966	21.514	82.279 82	0	4	13	24	1601
RDR2	32966	46.769	139.00	0	13	17	53	1602
AMIS	32966	35.698	46.580	0	1	2	99	99
DRMIS	32966	35.776	46.563	0	1	2	99	99
HLAMIS	32966	37.797 79	45.080 84	0	4	5	99	99
MALIG_TRR	32966	0.6257	0.4866	0	0	1	1	2
CMV_IGG	32966	2.3041	1.0777 62	0	2	3	3	3
CMV_IGM	32966	2.0563	1.2855	0	1	3	3	3
HIST_COCAINE_DON	32966	0.5201	0.6779	0	0	0	1	2
AGE_DON	32966	34.465	, 11.732 68	6	26	34	40	76

Variables	count	mean	std	min	25%	50%	75%	max
HBV_CORE_DON	32966	1.9179 15	1.3698 88	0	1	1	4	4
HBV_SUR_ANTIGEN_DON	32966	1.8853 06	1.3676 88	0	1	1	4	4
ABO_DON	32966	2.8215 43	1.3835 85	0	3	3	4	4
ALCOHOL_HEAVY_DON	32966	0.5142	0.6755 2	0	0	0	1	2
GENDER_DON	32966	1.0180 79	0.7552 08	0	0	1	2	2
HEP_C_ANTI_DON	32966	1.9044 47	1.3691 03	0	1	1	4	4
ANTIHYPE_DON	32966	0.7437 66	0.7968 65	0	0	1	1	2
BUN_DON	32966	19.683 7	13.757 26	0.4	12	20.214 87	20.214 87	245
CREAT_DON	32966	1.4041 16	1.2691 58	0.07	0.82	1.3	1.4200 9	37
PT_DIURETICS_DON	32966	1.2838 68	0.7875 79	0	1	1	2	2
PT_STEROIDS_DON	32966	1.3480 25	0.7613 86	0	1	2	2	2
PT_T3_DON	32966	0.3065	0.4690	0	0	0	1	2
PT_T4_DON	32966	1.2296	0.8051	0	1	1	2	2
PT_OTH2_OSTXT_DON	32966	5972.5	2591.6	0	4203.2	7694	7694	9269
PULM_INF_DON	32966	1.2977	1.1749	0	0	1	3	3
SGOT_DON	32966	96.904 49	306.02 49	0.3	31	68	97.527 13	20000
SGPT_DON	32966	97.767	365.39	0.4	25	56	97	44117
TBILI_DON	32966	0.9881	1.0570	0	0.6	0.9877	0.9877	59
URINE_INF_DON	32966	0.9634	1.3424 44	0	0	0	3	3
VASODIL_DON	32966	0.5357	0.6966	0	0	0	1	2
VDRL_DON	32966	2.1988	1.8183	0	1	1	5	5
CLIN_INFECT_DON	32966	1.2380	0.7994	0	1	1	2	2
HIST_CIG_DON	32966	0.4278	0.6079	0	0	0	1	2
HIST_HYPERTENS_DON	32966	0.6235	0.7483	0	0	0	1	2
HIST_CANCER_DON	32966	0.3236	0.4948	0	0	0	1	2
DIABETES DON	32966	0.3953	0.5802	0	0	0	1	2

Variables	count	mean	std	min	25%	50%	75%	max
HIST_OTH_DRUG_DON	32966	0.8759	0.8238	0	0	1	2	2
CMV_DON	32966	2.7614 21	1.1451 94	0	1	3	4	4
DDAVP_DON	32966	0.5305	0.6931	0	0	0	1	2
HEPARIN_DON	32966	1.6638	0.5103	0	1	2	2	2
ARGININE_DON	32966	1.2218	0.8072	0	1	1	2	2
WGT_KG_DON_CALC	32966	77.364	14.876	23.5	70	77.353 48	81.6	189
BMI_DON_CALC	32966	26.190 33	4.5802 42	10.588 66	23.667 83	26.221	27.173	66.035 64
HBV_NAT_DON	32966	2.0281 81	1.4016 18	0	0	3	3	3
ABO_MAT	32966	1.6387 79	0.9052 58	1	1	1	3	3
DIAL_PRIOR_TX	32966	0.6500 64	0.4834 63	0	0	1	1	2
ISCHTIME	32966	5.2922 57	1.5282 11	0.042	4.5664 06	5.3519 05	5.6328 13	25
O2_REQ_CALC	32966	5.3750 88	4.2890 68	0	3	5.4579 95	5.4579 95	26.3
PRIOR_CARD_SURG_TRR	32966	0.3284 29	0.4908 72	0	0	0	1	2
MALIG	32966	0.4118	0.5984 24	0	0	0	1	2
HGT_CM_CALC	32966	169.99 18	8.2559 88	122	165.1	170.03 08	175	210.82
BMI_CALC	32966	25.299 95	3.8325 76	14.997 85	23.382 51	25.392 36	27.331 17	44.777 87
DISTANCE	32966	211.25 96	210.94 01	0	67	215	221	4137
VENT_SUPPORT_AFTER_LIS T	32966	0.4269	0.6005	0	0	0	1	2
PROTEIN_URINE	32966	0.8621	0.8219	0	0	1	2	2
CARDARREST_NEURO	32966	0.3973 79	0.5680 28	0	0	0	1	2
PO2	32966	380.25 4	121.57 52	3.2	368	382.31 52	457	754
HIST_MI	32966	0.3272 46	0.4933 55	0	0	0	1	2
LV_EJECT	32966	58.086 24	9.4403 97	1	58	58.118 68	61	99
CORONARY_ANGIO	32966	2.0919 74	1.3279 38	1	1	1	4	4
BIOPSY_DGN	32966	3.0819	1.9978 58	1	1	5	5	5
HBSAB_DON	32966	3.8937	1.4628 7	0	3	3	6	6

EBV_IGG_CAD_DON 32966 4.8062 1.5472 0 4 4 7 7 EBV_IGM_CAD_DON 32966 2.7079 2.2522 0 1 1 6 6 CDC_RISK_HIV_DON 32966 0.6988 0 0 0 1 2 INOTROP_SUPPORT_DON 32966 0.9822 0.8363 0 0 1 2 2 TRANSFUS_TERM_DON 32966 294.91 454.81 0 0 1 2 2 PO2_FIO2_DON 32966 37.029 5.5849 10 34.7 37 39 110 PC02_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_LT_DON 32966 331.79 468.46 1 2 2 998 998 BRONCHO_RT_DON 32966 30.36 455.52 1 2 2 998 998 BRONCHO_RT_DON 32966 7.4153	EBV_IGG_CAD_DON 32966 4.8062 1.5472 0 4 4 7 7 EBV_IGM_CAD_DON 32966 2.7079 2.2522 0 1 1 6 6 CDC_RISK_HIV_DON 32966 0.5345 0.6988 0 0 0 1 2 INOTROP_SUPPORT_DON 32966 0.9822 0.8363 0 0 1 2 2 TRANSFUS_TERM_DON 32966 0.9822 0.8363 0 0 1 2 2 TRANSFUS_TERM_DON 32966 264.26 1.28363 0 0 1 2 2 PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PC02_DON 32966 37.02 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 331.79 468.46 1 2 2 998 998 BRONCHO_RT_DON 32966	Variables	count	mean	std	min	25%	50%	75%	max
EBV_IGM_CAD_DON 32966 2.7079 2.2522 0 1 1 6 6 CDC_RISK_HIV_DON 32966 0.5345 0.6988 0 0 0 1 2 INOTROP_SUPPORT_DON 32966 0.9822 0.8363 0 0 1 2 2 TRANSFUS_TERM_DON 32966 294.91 454.81 0 0 1 2 2 TRANSFUS_TERM_DON 32966 86.126 21.280 1 86 100 100 100 PO2_FIO2_DON 32966 37.029 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 301.37 468.46 1 2 2 998 998 CHEST_XRAY_DON 32966 30.36 455.52 1 2 5 999 999 999 PH_DON <td>EBV_IGM_CAD_DON 32966 2.7079 2.2522 0 1 1 6 6 CDC_RISK_HIV_DON 32966 0.5345 0.6988 0 0 0 1 2 INOTROP_SUPPORT_DON 32966 0.9822 0.8363 0 0 1 2 2 TRANSFUS_TERM_DON 32966 294.91 454.81 0 0 1 998 998 PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PC02_FIO2_DON 32966 37.029 5.849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 37.029 5.849 10 34.7 37 39 110 BRONCHO_RT_DON 32966 30.029 5.849 10 34.7 37 39 998 BRONCHO_RT_DON 32966 30.36 455.52 1 2 2 998 998 BRONCHO_RT_DON</td> <td>EBV_IGG_CAD_DON</td> <td>32966</td> <td>4.8062 85</td> <td>1.5472 43</td> <td>0</td> <td>4</td> <td>4</td> <td>7</td> <td>7</td>	EBV_IGM_CAD_DON 32966 2.7079 2.2522 0 1 1 6 6 CDC_RISK_HIV_DON 32966 0.5345 0.6988 0 0 0 1 2 INOTROP_SUPPORT_DON 32966 0.9822 0.8363 0 0 1 2 2 TRANSFUS_TERM_DON 32966 294.91 454.81 0 0 1 998 998 PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PC02_FIO2_DON 32966 37.029 5.849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 37.029 5.849 10 34.7 37 39 110 BRONCHO_RT_DON 32966 30.029 5.849 10 34.7 37 39 998 BRONCHO_RT_DON 32966 30.36 455.52 1 2 2 998 998 BRONCHO_RT_DON	EBV_IGG_CAD_DON	32966	4.8062 85	1.5472 43	0	4	4	7	7
CDC_RISK_HIV_DON 32966 0.5345 0.6988 0 0 0 1 2 INOTROP_SUPPORT_DON 32966 0.9822 0.8363 0 0 1 2 2 TRANSFUS_TERM_DON 32966 294.91 454.81 0 0 1 998 998 PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PC02_DON 32966 37.029 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 CHEST_XRAY_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71 HEMATOCRIT_DON 3	CDC_RISK_HIV_DON 32966 0.5345 0.6988 0 0 1 2 INOTROP_SUPPORT_DON 32966 0.9822 0.8363 0 0 1 2 2 TRANSFUS_TERM_DON 32966 294.91 454.81 0 0 1 998 998 PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PC02_DON 32966 37.029 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 31.79 468.46 1 2 2 998 998 BRONCHO_RT_DON 32966 303.6 455.52 1 2 5 999 999 BRONCHO_RT_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 PH_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71 PH_DON 32966 2	EBV_IGM_CAD_DON	32966	2.7079 72	2.2522 66	0	1	1	6	6
INOTROP_SUPPORT_DON 32966 0.9822 0.8363 0 0 1 2 2 FRANSFUS_TERM_DON 32966 294.91 454.81 0 0 1 998 998 PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PC02_DON 32966 37.029 5.5849 10 34.7 37 39 1110 BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 BRONCHO_RT_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	INOTROP_SUPPORT_DON 32966 0.9822 0.8363 0 0 1 2 2 TRANSFUS_TERM_DON 32966 294,91 454.81 0 0 1 998 998 PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PCO2_DON 32966 37.029 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 CHEST_XRAY_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 74153 0.0835 5 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	CDC_RISK_HIV_DON	32966	0.5345	0.6988	0	0	0	1	2
TRANSFUS_TERM_DON 32966 294,91 454,81 0 0 1 998 998 PO2_FIO2_DON 32966 86,126 21,280 1 86 100 100 100 PC02_DON 32966 37,029 5,5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 321,94 464,84 1 2 2 998 998 BRONCHO_RT_DON 32966 331,79 468,46 1 2 2 998 998 BRONCHO_RT_DON 32966 300,36 455,52 1 2 5 999 999 PH_DON 32966 7,4153 0,0835 5 7,4 7,4155 7,44 8 HEMATOCRIT_DON 32966 29,215 4,4611 2.5 27 29,149 30.6 71	TRANSFUS_TERM_DON 32966 294.91 454.81 0 0 1 998 998 PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PCO2_DON 32966 37.029 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 CHEST_XRAY_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71 96 30.6 71 83 27 25 27 29.149 30.6 71	INOTROP_SUPPORT_DON	32966	0.9822	0.8363	0	0	1	2	2
PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PC02_DON 32966 37.029 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 CHEST_XRAY_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	PO2_FIO2_DON 32966 86.126 21.280 1 86 100 100 100 PC02_DON 32966 37.029 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 CHEST_XRAY_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 7.4155 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	FRANSFUS_TERM_DON	32966	294.91 72	454.81	0	0	1	998	998
PCO2_DON 32966 37.029 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 BRONCHO_RT_DON 32966 300.36 455.52 1 2 5 999 999 CHEST_XRAY_DON 32966 15 5 7.4 7.4155 7.44 8 PH_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	PCO2_DON 32966 37.029 5.5849 10 34.7 37 39 110 BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 BRONCHO_RT_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	PO2_FIO2_DON	32966	86.126	21.280	1	86	100	100	100
BRONCHO_LT_DON 32966 321.94 468.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 CHEST_XRAY_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 7.4153 0.00835 5 7.4 7.4155 7.44 8 PH_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71 96 83 27 96 30.6 71	BRONCHO_LT_DON 32966 321.94 464.84 1 2 2 998 998 BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 BRONCHO_RT_DON 32966 301.79 468.46 1 2 2 998 998 CHEST_XRAY_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	PCO2_DON	32966	37.029	5.5849	10	34.7	37	39	110
Image: system Image: s	BRONCHO_RT_DON 32966 331.79 468.46 1 2 2 998 998 CHEST_XRAY_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	BRONCHO_LT_DON	32966	321.94	464.84	1	2	2	998	998
CHEST_XRAY_DON 32966 300.36 455.52 1 2 5 999 999 PH_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 PH_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	L IS S I <thi< th=""> I I <thi< th=""></thi<></thi<>	BRONCHO RT DON	32966	331.79	468.46	1	2	2	998	998
H_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	PH_DON 32966 7.4153 0.0835 5 7.4 7.4155 7.44 8 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	CHEST XRAY DON	32966	300.36	455.52	1	2	5	999	999
H_DOX 15 11 55 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	HE_DOIN 15 11 55 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71 HEMATOCRIT_DON 32966 29.215 4.4611 2.5 27 29.149 30.6 71	PH DON	32966	43 7.4153	74 0.0835	5	7.4	7.4155	7.44	8
		HEMATOCRIT DON	32966	15 29.215	11 4.4611	2.5	27	55 29.149	30.6	71

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml