Impact of Lung Transplant Patient Characteristics on Blood Transfusion Requirements and Clinical Outcomes

By

Leen Madani, Alanna Olteanu, Nishant Sarkar

Institute for Management and Innovation

University of Toronto Mississauga

BTC1877H Data Science in Health II

Dr. Nicholas Mitsakakis

December 4, 2023

1. Introduction

1.1 Literature Review

Lung transplantation for end-stage lung disease is a treatment method used in patients with a variety of pulmonary conditions, including cystic fibrosis, pulmonary fibrosis, and primary pulmonary arterial hypertension.[1] The procedure has advanced significantly since the first successful lung transplant in 1983. Despite these advances, the median long-term survival for lung transplant patients is 5.8 years, which is notably less than other transplant areas.[2] A contributor to this problem is that ~25% of lung transplant patients experience primary graft dysfunction, and a significant proportion experience other complications including infections and renal failure.[3]

A growing body of evidence suggests that a key player in these complications is intraoperative blood loss and the resulting need for blood transfusions.[4] Indeed, some studies have indicated that large-volume (generally >1 liter, or approximately 2.22 units of blood) intraoperative transfusion of red blood cells is independently associated with primary graft dysfunction.[5] While the link between transfusion and patient mortality is an evolving area of research, there is evidence that understanding the predictors of blood transfusion requirements in lung transplant patients can be crucial for clinical decision-making. Knowledge of these predictors can help clinicians anticipate and prepare for potential transfusion needs, thereby enabling more precise surgical planning and resource allocation. This foresight help minimize intraoperative blood transfusion, which in turn may reduce the incidence of primary graft dysfunction and other postoperative complications. However, the specific predictors associated with the need and amount of transfusions in lung transplant patients are not entirely understood.[6]

Some investigators have linked certain predictors to an increased need for blood transfusion (or increased levels of blood loss) in lung cancer patients. For example, bilateral lung transplants or redo lung transplants are generally more complex procedures that may require invasive methods including cardiopulmonary bypass (CPB), and thus are strong predictors of additional blood loss during surgery.[7] Overall, the use of more complex and invasive techniques is often associated with the need for additional transfusions, including CPB,[8] ex vivo lung perfusion (involving the external perfusion and assessment of donor lungs),[9] extracorporeal membrane oxygenation (ECMO),[10] The presence of particular pulmonary diseases is also often a strong predictor of blood loss. Cystic fibrosis patients are more likely to require extensive surgery and bilateral lung transplants, and as such generally require more blood.[11] Similarly, interstitial lung disease (ILD) patients have been shown to lose more blood during lung transplant surgery, likely a result of the increased time and difficulty of surgery for ILD cases.[12]

Insights can also be drawn from other disease areas: Predictors for transfusion or blood loss in other surgical procedures, particularly those closely related to lung transplant, may generalize to lung transplant as well. A good example is cardiac surgeries, which also occur in the thoracic area. Increased BMI has been found to be a predictor of increased bleeding in cardiac surgery,[13] and preoperative stroke and renal failure have been linked to a higher risk of requiring transfusion.[14] Beyond cardiac surgery, some inferences can be made based on somewhat-unrelated transplant areas. For example, age is a component of the Charleston Comorbidity Index, and higher CCI index scores have been linked to increased blood

loss in orthopedic surgery.[4] Similarly, males appear to experience more perioperative blood loss during orthopedic surgery; these may apply to lung transplant as well.[4]

Still other possible predictors are those that have not been directly associated with blood loss during types of transplant surgery, but have a plausible causal link to increased blood loss and thus the need for transfusion. An example is the presence of chronic obstructive pulmonary disease (COPD); a significant proportion of the population with COPD is also affected by anemia, and anemia is associated with increased bleeding and slower wound healing.[15] Another is the presence of liver diseases, given that the liver is an important source for coagulation factors which are often impeded during disease.[16] Finally, preoperative blood test results (including hemoglobin levels, platelet count, hematocrit, and international normalized ratio) can offer insights into a patient's coagulation status and overall blood health ahead of surgery. Lower hemoglobin levels or platelet counts, for instance, might predispose patients to increased intraoperative bleeding, thereby elevating the likelihood of transfusion requirements.[16]

In summary, the possible predictors for blood loss in lung cancer patients form a complex landscape, where multiple factors, ranging from specific pulmonary diseases and surgical techniques to broader health indicators and preoperative conditions, may influence the need for blood transfusions in lung transplantation. While previous studies provide a strong starting point, they also reveal significant gaps in the understanding of these predictors in the specific context of lung transplantation. This gap highlights the need for a more focused analysis, tailored specifically to lung transplant surgery.

1.2 Objective and Research Question

The objective of this analysis is to better understand the factors associated with patients requiring transfusions during lung transplant surgery, with the goal of identifying key predictors for transfusion needs that could be used to enhance pre-, intra- and post-operative decision making and potentially reduce the incidence of postoperative complications associated with transfusions. Furthermore, the analysis seeks to investigate the relationship between transfusions and poorer patient outcomes (particularly patient mortality). In this manner, ‘transfusion’ can be considered an intermediate variable between patient predictors and mortality, further enhancing the clinical relevance and utility of the findings. Previous studies have hypothesized a link between several patient predictors and blood loss/need for transfusion during surgery, and the link between blood loss during surgery and poorer outcomes. However, these investigations are often not comprehensive in their findings or more relevant to other surgical areas, and as such further investigation is warranted.

To this end, the main research questions guiding this analysis were:

1. What are the characteristics of patients that require transfusions, and what are the factors influencing the need and amount of transfusions?
2. In conjunction with the above, what is the impact of transfusion on patient outcomes, including mortality?

2. Methods

2.1 Dataset

This study leveraged a dataset containing physiological and treatment information from 193 lung transplant patients receiving either a single or bilateral lung transplant from 2018-01-05 to 2018-12-30. A total of 113 variables are associated with each patient, with the exception of those patients with missing data for some variables. The outcome variables of interest include the amount of transfusion each patient received during (denoted as intra-) and after their surgeries (over the course of 24, 48, and 72 hours). These are documented across four types of transfusion: red blood cells (RBC) or packed cells, fresh-frozen plasma (FFP), platelets (Plt), and cryoprecipitate (Cryo). All transfusion amounts are assumed to be in 'units.' Another outcome of interest is whether the patient had a Massive Transfusion, with RBC units > 10.

Demographic data associated with each patient includes their gender, age, and BMI, etc. Information about patients' respiratory diagnoses (including whether the patient had COPD, cystic fibrosis, interstitial lung disease, etc.) and other diagnoses (including coronary artery disease, hypertension, renal failure, etc.) is included, along with important treatment decisions made during surgery like the use of ex-vivo lung perfusion or pre/intraoperative extracorporeal membrane oxygenation (ECLS-ECMO) or cardiopulmonary bypass (ECLS-CPB). The dataset contains results of each patient's pre-, intra-, and post-immediate, and post-1 day operative blood work, including hemoglobin (Hb), fibrinogen, and creatinine levels along with other measures like prothrombin time (PT) and international normalized ratio (INR). Finally, the dataset includes variables associated with the patients' survival (eg. date of death and whether they were alive for 30, 90, and 365 days after surgery) and time spent in the hospital (eg. ICU admission date and length of stay).

2.2 Analysis Software

This analysis was conducted using the R programming language. The initial phase of exploratory data analysis was conducted using the 'funModeling', 'naniar', and 'ggplot2' packages to generate summary statistics and data/missing data visualizations. Data manipulation was aided by the 'dplyr' package. Data imputation was done using the 'mice' package. Several packages were used in the investigation of research questions 1 and 2: 'car', 'glmnet', and 'survival'.

2.3 Data Preprocessing

First, the missingness of the data was investigated; variables with missingness >30% were removed altogether. This step reduced the number of candidate variables in the dataset from 113 to 80. Next, all variables were recoded to better reflect what they were representing: most variables which would be better represented as factors (ex. Gender, disease statuses, the use of intraoperative ECLS, etc.) were originally encoded as characters, which was remedied. Intra_Albumin.5.mL was found to be encoded as a character when it should have been numeric; this was also remedied.

The dataset was then investigated to find any strange values that were not properly encoded as missing or negative values which were not possible, which were found in Intra_Albumin, DCD.vs.DBD, PostImmediate_PTT, and PostDay1_PTT - as there was no way of knowing what these variables should have been, they were replaced by missing values (NA). Furthermore, some time variables were represented twice in the dataset (including ICU.Admission.Date.Time, Duration.of.ICU.Stay..days., etc.); for variables with duplicates, the version of the variable with more missing values was removed due to redundancy.

Subsequently, the dataset was used to define some new composite variables for downstream analysis. The two 'transplant type' variables (which denoted if the patient was on their first or second transplant) were combined into one factored variable, 'Transplant_Type,' with levels 'FIRST' and 'SECOND'; the original 'transplant type' variables were then removed. Next, the three "Alive Status" variables (which denoted whether the patient was alive at 30 days, 90 days, or 1 year) were combined into one numeric variable, "Minimum_Alive_Days," which stored the highest number of days a patient was known to be alive. The original "Alive Status" variables were also removed. Finally, a new factored variable "Had.Transfusion" was defined, describing whether the patient had any transfusion after 72 hours (whether RBC, packed cells, FFP, Plt, or Cryo). This was done because some of the variables pre-combination had few patients who actually received the given transfusion, so a composite variable was created to enable investigations of transfusions beyond RBC.

2.4 Initial Variable Selection

The selection of an initial set of candidate variables for this study was guided by an extensive review of relevant literature (**Supplementary Table 1**). The aim was to identify variables that have a demonstrable impact on the likelihood of blood transfusions and the quantity of blood transfused. This literature review encompassed various studies and reports, focusing on both clinical and demographic factors that influence transfusion needs. Variables were chosen based on their recurrence in the literature as significant factors affecting transfusion requirements across surgical areas, or had a plausible causal link to increased transfusion requirements. The final set of variables included continuous factors such as Age, BMI, various preoperative and postoperative blood parameters (e.g., Hemoglobin, Hematocrit, Platelets, INR), intraoperative measures (e.g., Albumin administration, Crystalloid volume, Packed Cells), as well as categorical factors like patient gender, presence of chronic conditions (e.g., COPD, Cystic Fibrosis, Interstitial Lung Disease, Coronary Artery Disease, Hypertension, Renal Failure, Liver Disease), type of surgery or transplant received, and the use of extracorporeal support systems (e.g., ECLS/ECMO).

2.5 Data Imputation and Collinearity

In clinical datasets, missing data is a significant challenge that can introduce significant biases if not addressed appropriately. For this study, a stochastic imputation method was chosen over multiple imputation, primarily due to its compatibility and effectiveness in datasets intended for Lasso regression models. Stochastic imputation involves the use of probabilistic models to estimate missing values. This method is particularly advantageous in clinical datasets where missingness is often not random and may be correlated with patients' underlying health conditions or treatment modalities. Lasso regression models are sensitive to the method of imputation due to their inherent nature of feature selection and

regularization. Furthermore, given the bootstrapped Lasso regression approach used (see **Section 2.6**), if multiple imputation was used, the models would take several times longer to train, and would likely experience issues in pooling afterward. Also, only a small number of datapoints were missing overall, minimizing the possibility for the introduction of significant bias. Consequently, stochastic imputation was determined to be the best option. The imputed values appeared to closely align with the distribution of the original variables (**Figure 1**).

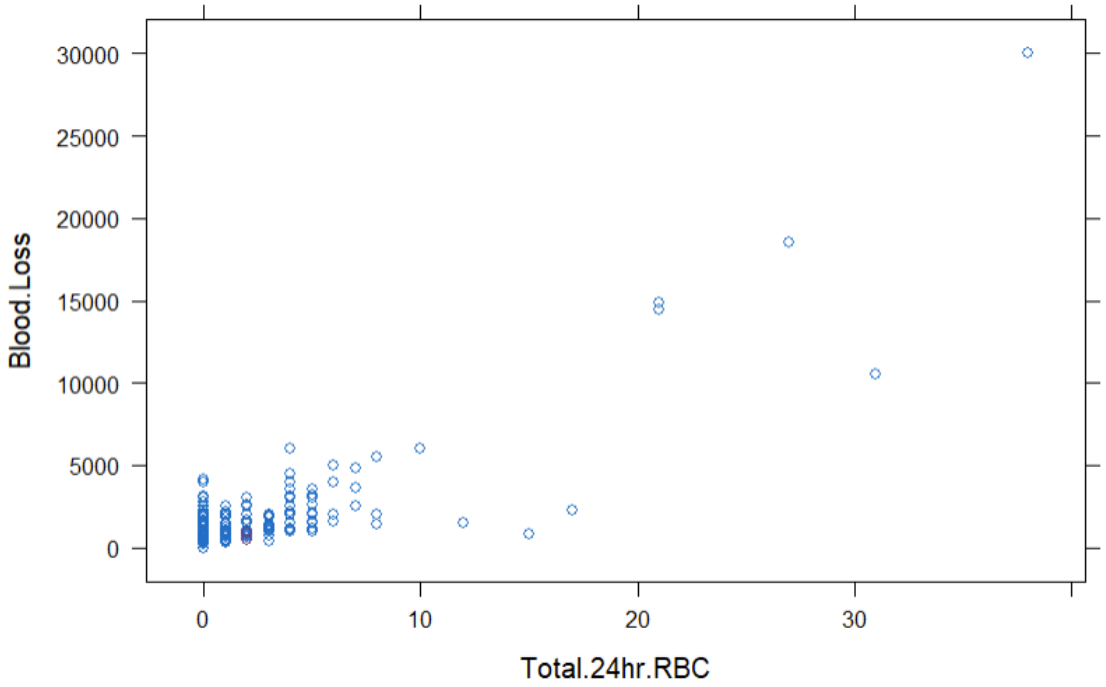


Figure 1: Relationship between Total.24hr.RBC and Blood.Loss with stochastic regression imputation. Created using the 'mice' package. Imputed points in red.

Following imputation, the dataset was subdivided into two data frames: Pre_df and Post_df. Pre_df contained candidate predictors for Research Question 1, consisting of measurements of patient status from before or during the surgery that would be useful for predicting the need for transfusion during surgery. Post_df contained variables useful for Research Question 2, consisting of outcome, survival, and post-surgical variables.

Next, to assess the multicollinearity among the predictors in Pre_df, we utilized Variance Inflation Factor (VIF) analysis. This was carried out using the 'car' package in R, leveraging a full model that included all candidate predictors with Total.24hr.RBC as outcome. This step was particularly important when using Lasso regression in subsequent steps, as Lasso tends to select one variable from a group of highly correlated variables, potentially overlooking others that are equally important. Variables associated with a VIF > 5 were removed if they were found to not be useful in downstream analysis.

2.6 Research Question 1.1 - Characteristics of Transfusion Patients (Exploratory Data Analysis)

The Exploratory Data Analysis (EDA) was conducted with the objective of uncovering patterns, anomalies, trends, and relationships within the data, particularly focusing on the transfusion of RBCs. This phase involved a variety of graphical and statistical techniques:

1. Histograms and Box Plots: These were utilized to examine the distribution of continuous variables such as age, BMI, and various blood parameters both preoperatively and postoperatively.
2. Scatter Plots: Scatter plots helped in exploring potential correlations between continuous variables such as preoperative hemoglobin levels (Pre_Hb) and the total amount of RBC transfused within 24 hours (Total.24hr.RBC).
3. Bar Charts: These were particularly useful for analyzing categorical data, such as the presence of specific chronic conditions (e.g., Liver Disease) and their relationship to the amount of RBC transfused.
4. Correlation Heatmaps: These were used to visualize the correlation matrix of continuous variables, offering insights into the interdependencies among various factors in the dataset.

Each of these methods was selected for its ability to effectively visualize and analyze the dataset's complexity, with a particular focus on identifying key factors influencing blood transfusion requirements in patients.

2.7 Research Question 1.2 - Factors Influencing Need and Amount of Transfusion

Given the selection of a set of candidate predictors from the literature and leveraging the insights from the EDA process to determine some characteristics of patients requiring transfusion, we advanced our analysis using bootstrapped Lasso regression. This approach was done with the goal of further narrowing down the candidate predictors in order to identify the most influential factors with respect to the need and amount of transfusions in these lung transplant patients. Lasso regression, involving L1 regularization, is particularly suited for situations with numerous potential predictors, as it integrates variable selection with model fitting. By imposing a penalty on the size of the coefficients, Lasso effectively zeroes out less influential variables, selecting only those that contribute significantly to the model (and thus are the most influential in determining the need and amount of transfusions, depending on the outcome variable). To tune each Lasso model, cross-validation was used to determine the most appropriate (minimum) lambda value associated with the least predictive error, avoiding overfitting.

However, running Lasso regression only once would not be sufficient: depending on how the data is divided during cross-validation, different sets of predictors can be selected as the “most influential,” affecting the consistency and stability of this approach. To address this challenge, we employed a bootstrapping technique. This involved generating 2000 resamples of the dataset, each with replacement, and applying Lasso regression to each resample. Through this repetitive process, we were able to observe the frequency of each candidate predictor’s selection across 2000 iterations of bootstrapped sampling and model training. This provided a measure of each predictor’s frequency of inclusion in the models associated with the minimum lambda, and thus their relative importance and stability, identifying those variables that were consistently influential across various samples. Predictors with a >80% frequency of inclusion were considered to be the most influential for their respective outcome. Previous applications of

this approach have used frequency 'cutoffs' of 50-90%;[17] 80% was chosen to ensure a balance between including predictors with a high likelihood of genuine association with the outcome and excluding those that might be less consistently relevant.

Notably, as 2000 trained models are involved, this bootstrapping approach could not result in a single, unified set of model weights that could be used to directly interpret the effect of each influential predictor. Consequently, we next fitted the most frequently selected predictors from the bootstrapped Lasso regression into linear or logistic regression models, depending on the nature of the outcome variable. The beta coefficients from these regression models provided insights into the relative magnitude and direction of each predictor's effect. We highlight that the interpretability of the p-values derived from these regression models is limited. The predictors were pre-selected in a data-driven approach based on their performance in the Lasso model, a process that inherently carries a form of selection bias, so the resulting p-values do not accurately reflect the uncertainty associated with the variable selection process. As a result, only the beta coefficients were analyzed.

This approach of bootstrapped Lasso regression, predictor selection, and fitting/analysis of standard regression models, was repeated thrice with three outcome variables. The first outcome was total red blood cell units received over 24 hours (Total.24hr.RBC, a continuous variable requiring Lasso regression and a linear regression model), to investigate the factors influencing the amount of transfusion. To investigate the factors influencing the need for transfusion, the second outcome was whether the patient had a massive transfusion (Massive.Transfusion), and the third was whether the patient had any form of transfusion (Had.Transfusion, a newly defined variableboth this and Massive.Transfusion were categorical variables requiring Lasso classifiers and a logistic regression model).

2.8 Research Question 2 - Impact of Transfusion on Patient Outcomes

The second objective of our study was to assess the impact of transfusion on patient outcomes, including mortality, in lung transplant recipients. We selected post-operative variables based on a literature review and our exploratory data analysis (EDA), which included examining a correlation matrix for variables with high correlations (greater than 0.8). Collinearity was evaluated by fitting a linear model with the predictor being the total RBCs administered within 24 hours and calculating the VIF. We excluded intraoperative measurements such as crystalloids and albumins because they are universally administered and therefore unlikely to influence patient outcomes.[18]

Our primary analysis was centered on the total number of RBCs given in 24 hours and the occurrence of massive transfusion. Linear regression models were applied to determine the impact of these predictors on the length of hospital stay. Logistic regression was utilized for binary outcomes, including 30-day, 90-day, and 12-month survival rates, death, and the necessity for reoperation due to bleeding within 24 hours post-surgery.

In addition, we conducted survival analysis using ICU admission and discharge dates to compute survival time, which was defined as the time from ICU admission to either death or discharge. We fitted a Cox proportional hazards regression model, using total RBCs transfused within 24 hours as a predictor. This variable was selected for its significant association with mortality in the simple regression model and was

evaluated alongside other patient characteristics. A complementary log-log (cloglog) plot was produced to verify the Cox hazard proportionality assumption.

Kaplan-Meier curves were generated to illustrate the effect of the total number of RBCs transfused within 24 hours on patient survival. We categorized the transfused RBCs into three groups: 0-10 units (Regular), 10-20 units (Massive), and over 20 units (Ultra-Massive).[19]

3. Results

3.1 Overview of Data, Collinearity Testing and Graphical Display of Patient Characteristics

The initial phase of our analysis focused on selecting pertinent variables from the lung transplant dataset, which contained a diverse range of patient demographics, preoperative characteristics, intraoperative details, and postoperative outcomes. Key variables identified for detailed analysis included demographic factors (Age, Gender), preoperative clinical measures (Pre-Hb, Pre-Hct, Pre_Platelets, Pre_INR), and intraoperative factors (Blood.Loss, Intra_Packed.Cells). Given the presence of missing values in our dataset, we employed the Stochastic Imputation method, resulting in a more comprehensive dataset. Post-imputation, key variables like Pre_Hb exhibited a notable change in distribution, emphasizing the importance of addressing missing data in clinical datasets.

VIF testing highlighted problematic multicollinearity (VIF > 5) between several key variables (**Supplementary Table 2**). These included Pre_Hb (VIF = 16.06), Pre_Hct (15.66), Intra_Packed.Cells (11.19), Blood.Loss (7.46), RBC.72hr.Total (10.86), and Plt.72hr.Total (8.94). Pre_Hb and Pre_Hct were found to be significantly correlated (which aligns with physiological expectations), so hematocrit was removed to limit this effect. Intra_Packed.Cells and Blood.Loss were removed as they would be directly related to the amount of RBCs received, and would not be useful predictors. Finally, RBC.72hr.Total, Plt.72hr.Total, and the the 72 hour total variables for FFP and Cryo were removed, as they were already represented by the composite variable Had.Transfusion that would be used for analysis (**Supplementary Table 3)**.

Our EDA commenced with the generation of histograms for all variables, including both continuous and categorical ones (**Figure 2, Figure 3**). This provided an immediate visual insight into the distribution and skewness of each variable. The creation of histograms for each variable facilitated a preliminary understanding of data distribution and potential outliers.

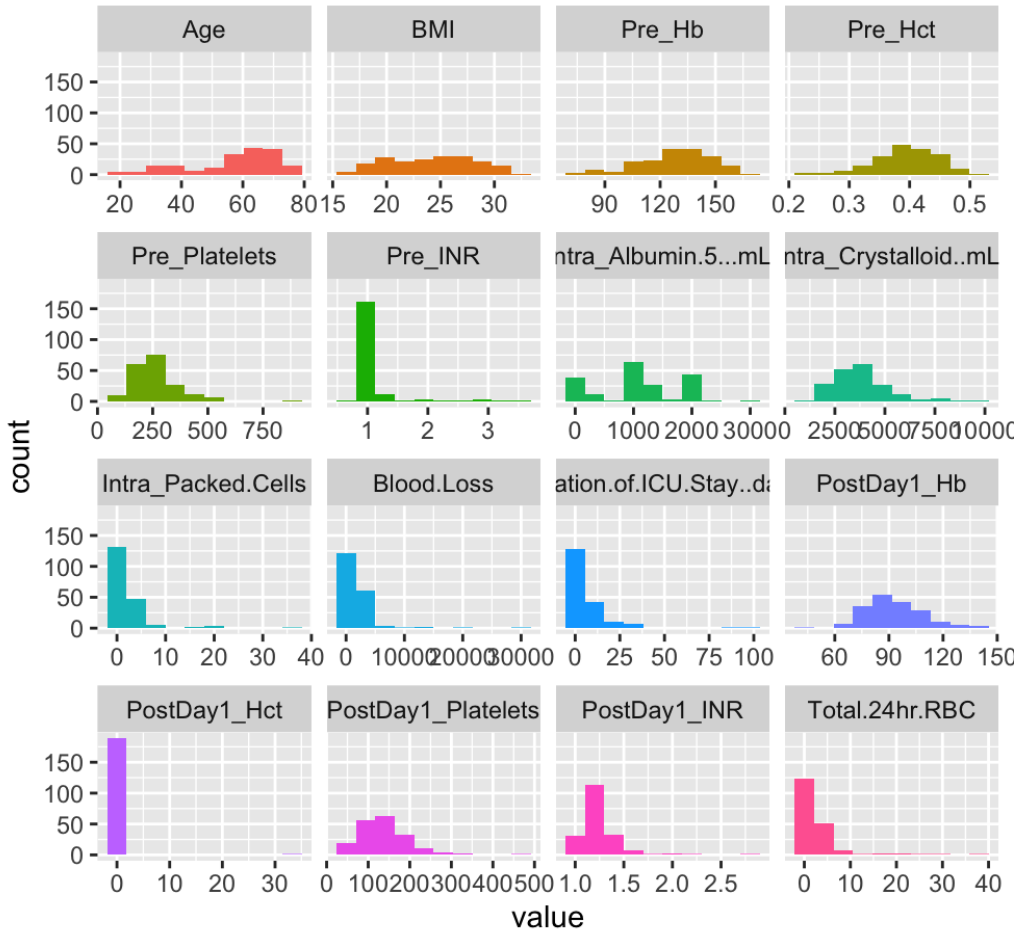


Figure 2: Histograms for All the Continuous Variables Used

We examined all the continuous variables in the dataset to gain insights into their distributions and potential relationships. As shown in **Figure 2**, we observed that the continuous variables exhibit various patterns. Notably, 'Age' displayed a mean of approximately 56.35 years (significant error ± 16.25), following an approximately normal distribution. Similarly, 'BMI' had a mean of approximately 24.1 (significant error ± 4.564) with a unimodal distribution. 'Pre_Hb' exhibited a mean of approximately 127.8 g/dL (significant error ± 22.02), also following a unimodal distribution. 'Pre_Hct' displayed a mean of approximately 0.3936 (significant error ± 0.06064) with a unimodal distribution as well. 'Pre_Platelets' had a mean of approximately 264.6 (significant error ± 107.4) and showed a unimodal distribution. 'Pre_INR' exhibited a mean of approximately 1.114 (significant error ± 0.3219), with a slight right skew. 'Intra_Albumin.5...mL.' showed a mean of approximately 1107 (significant error ± 812.5) and exhibited a skewed distribution. 'Intra_Crystalloid..mL.' had a mean of approximately 3696 (significant error ± 1539) and displayed a slightly right-skewed distribution. 'Intra_Packed.Cells' exhibited a mean of approximately 2.274 (significant error ± 3.227) and was right-skewed. 'Blood.Loss' displayed a mean of approximately 1905 (significant error ± 1958) and was right-skewed. 'Duration.of.ICU.Stay..days.' had a mean of approximately 7.284 (significant error ± 8.164) and showed a unimodal distribution. 'PostDay1_Hb' exhibited a mean of approximately 93.21 g/dL (significant error ± 18.5) and followed a unimodal distribution. 'PostDay1_Hct' displayed a mean of approximately 0.4549 (significant error ± 0.4017) and was unimodal. 'PostDay1_Platelets' had a mean of approximately 136.5 (significant error ± 64.81) and

exhibited a unimodal distribution. 'PostDay1_INR' exhibited a mean of approximately 1.264 (significant error ± 0.2113), with a unimodal distribution. This EDA provides a preliminary understanding of the data's characteristics, which will inform our subsequent analyses and model building.

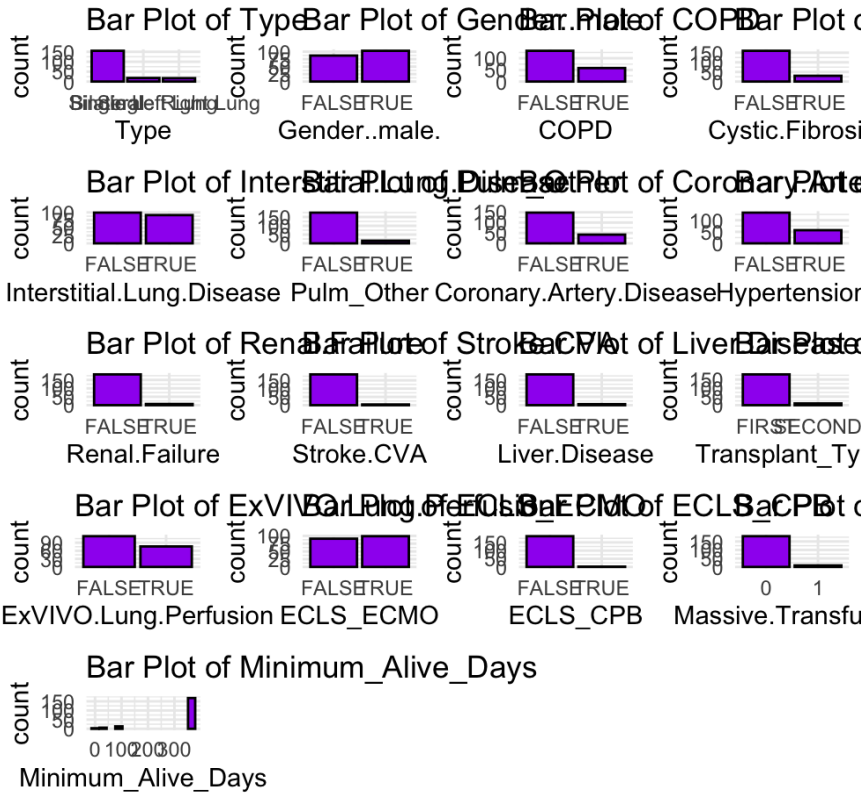


Figure 3: Barplots for All the Categorical Variables used

We further examined several categorical variables in the dataset to understand their distributions and characteristics, displayed in **Figure 3**. The categorical variables include 'Type,' 'Gender (male),' 'COPD (Chronic Obstructive Pulmonary Disease),' 'Cystic Fibrosis,' 'Interstitial Lung Disease,' 'Pulmonary Other,' 'Coronary Artery Disease,' 'Hypertension,' 'Renal Failure,' 'Stroke/CVA,' 'Liver Disease,' 'Transplant Type,' 'ExVIVO Lung Perfusion,' 'ECLS ECMO,' 'ECLS CPB,' 'Massive Transfusion,' and 'Minimum Alive Days.' For 'Type,' the majority of cases (81.6%) are 'Bilateral,' with 'Single Left Lung' and 'Single Right Lung' accounting for 9.5% and 8.9%, respectively. In terms of gender, 54.2% of patients are male ('TRUE'). Regarding health conditions, 'COPD' is present in 30.5% of cases, 'Cystic Fibrosis' in 15.8%, 'Interstitial Lung Disease' in 47.9%, 'Pulmonary Other' in 7.9%, 'Coronary Artery Disease' in 22.1%, 'Hypertension' in 30%, 'Renal Failure' in 3.7%, and 'Stroke/CVA' in 2.1%. Additionally, 'Liver Disease' is present in 3.2% of cases. For transplant types, 'FIRST' comprises 94.7% of cases, and 'SECOND' makes up 5.3%. 'ExVIVO Lung Perfusion' is 'TRUE' in 40% of cases. In terms of extracorporeal life support, 'ECLS ECMO' is 'TRUE' in 52.1%, while 'ECLS CPB' is 'TRUE' in only 1.1% of cases. 'Massive Transfusion' is required in 4.7% of cases. Lastly, 'Minimum Alive Days' varies, with 88.4% having a value of 365 days, and smaller proportions having values of 0, 30, or 90 days. This comprehensive examination of categorical variables in our dataset provides valuable insights into the distribution of these variables, which will inform subsequent analyses and modeling efforts.

Further graphical analyses played a crucial role in determining the key variables related to the need and amount of blood transfusion. By selecting certain important variables and plotting them against our dependent variable (Total.24hr.RBC), we were able to determine which characteristics were more likely to cause a patient to have blood transfusion. Correlation plots and heat maps further demonstrated which factors influence each other when determining the need for blood transfusion.

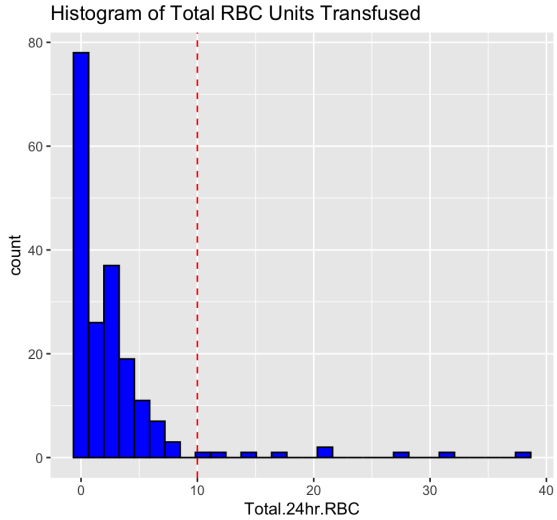


Figure 4: Histogram of Total RBC Units Transfused over 24 hours

Figure 4 displays a histogram Total.24hr.RBC. The distribution is right-skewed, with the majority of patients receiving fewer than 10 units of RBCs, as evidenced by the tall bars on the left. The number of patients decreases significantly as the number of RBC units increases, therefore indicating that a majority of patients required a smaller amount of RBC Units transfused within the 24 hr period. The red dashed line is at the 10 unit mark, representing the distinction between massive transfusion.

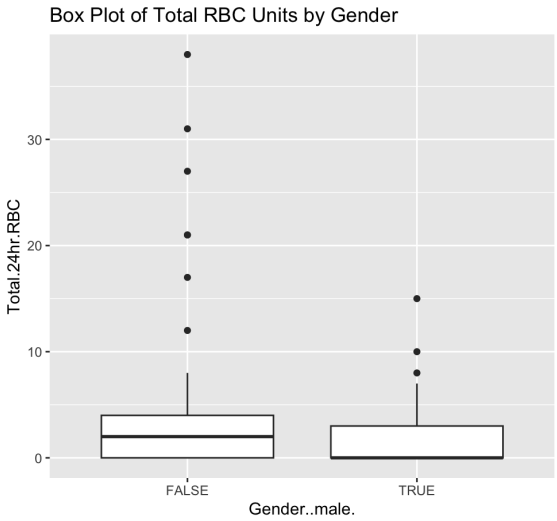


Figure 5: Box Plot of Total RBC Units by Gender

Figure 5 visualizes the distribution of total RBC units transfused within a 24-hour period, categorized by gender. The x-axis separates the data into two categories labeled "FALSE" and "TRUE," which likely correspond to female and male patients, respectively. For both categories, the central rectangle (the "box"

in the box plot) spans from the first quartile (Q1) to the third quartile (Q3) of the data, representing the middle 50% of the scores for each gender. The horizontal line within each box indicates the median value of RBC units transfused. The plot shows that the median transfusion volume is low for both genders, with a slightly higher median for males ("TRUE"). This is an early indication that males may require less transfusion, contradicting the literature, though further investigation is required. There are outliers for both genders, indicated by individual dots beyond the "whiskers" of the box plot, which extend to the smallest and largest values within 1.5 times the interquartile range (IQR) from the edges of the box. The presence of outliers suggests that some patients received significantly more RBC units than the typical range.

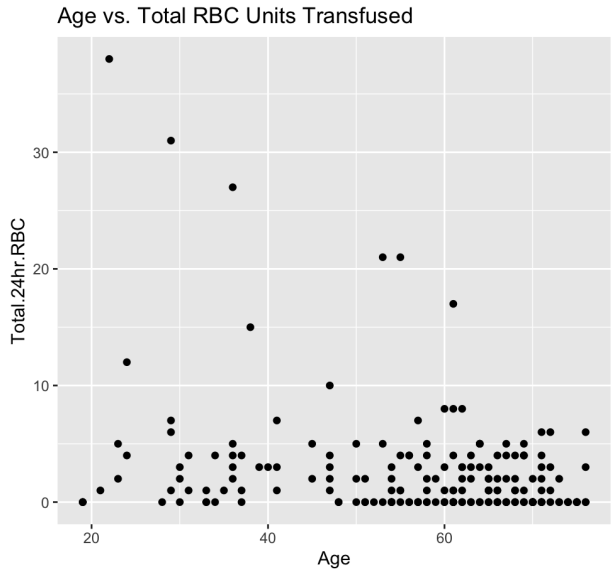


Figure 6: Scatter Plot of Blood Transfusion by Age

Figure 6 depicts the relationship between patient age and the total number of RBC units transfused within a 24-hour period. The x-axis represents the age of patients, and the y-axis represents the total RBC units transfused. From the distribution of points, there does not appear to be a clear linear relationship between age and the amount of RBC units transfused; the data points are widely dispersed. Most of the transfusion amounts are clustered at the lower end of the y-axis, indicating that regardless of age, the majority of patients received a smaller number of transfusions. There are, however, a few instances of higher transfusion volumes across various ages, which are depicted as points scattered above the main cluster. A concentration of data points near the zero line of RBC units suggests that many patients in the dataset did not receive transfusions or received very few units. There is a higher volume of dots near the higher values of the x-axis, indicating that the majority of patients are older.

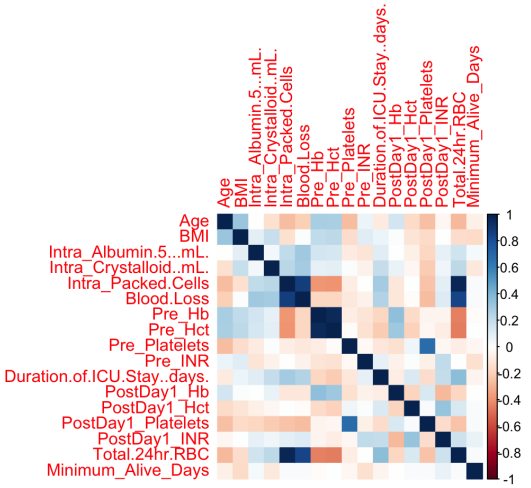


Figure 7: Correlation map demonstrating correlation between continuous variables

Figure 7 depicts a correlation matrix heatmap, which is a visual representation of the Pearson correlation coefficients between pairs of continuous variables in a dataset. In this heatmap, each square shows the correlation between the variables on the x-axis and the y-axis. The color scale on the right indicates the strength and direction of the correlation, where 1 is a perfect positive correlation, -1 is a perfect negative correlation, and 0 indicates no correlation. Several pairs of variables show strong correlations, as indicated by darker shades. For instance, there appears to be a strong positive correlation between 'Intra_Packed.Cells' and 'Blood.Loss', which is intuitive because more significant blood loss during surgery often requires more packed cell transfusions. Similarly, 'Pre_Hb' (preoperative hemoglobin) and 'Pre_Hct' (preoperative hematocrit) are also positively correlated, which is expected since hematocrit is the percentage of blood volume that is made up of red blood cells, to which hemoglobin is central. There is a noticeable negative correlation between Pre_Hb and Total.24hr.RBC, which could indicate that it is a predictor for amount of transfusion.

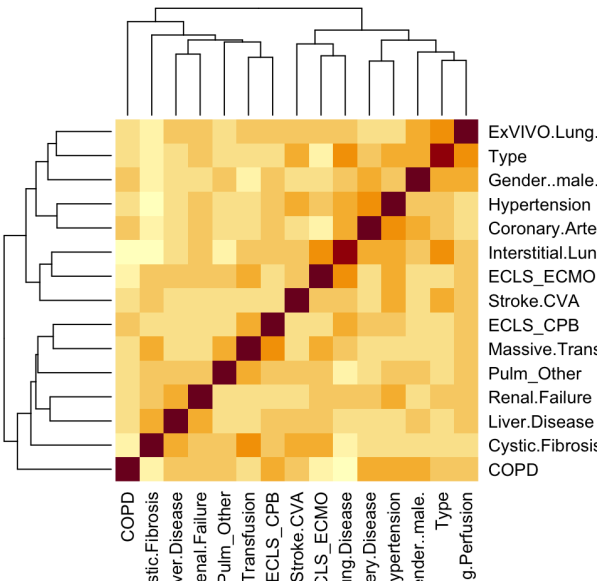


Figure 8: Heat map demonstrating correlation between the categorical variables

Figure 8 illustrates the associations between various categorical variables, reflecting the prevalence of comorbidities and characteristics within a patient cohort undergoing lung transplantation. Each cell's color intensity represents the strength of association between the row and column variables, with darker colors indicating a higher degree of association, which could be due to a higher co-occurrence within the patient data. Clusters identified by the dendrograms on the top and left sides demonstrate groupings of variables based on their associations. For example, closely grouped variables may often co-occur within the same patients (e.g., particular comorbidities like "Hypertension" and "Coronary.Artery.Disease" might frequently appear together). Variables such as "ECLS_ECMO" and "ECLS_CPB" are relatively closely associated, which makes clinical sense given that both relate to extracorporeal life support technologies. Similarly, "Cystic.Fibrosis" and "COPD" appear to have less association with other variables, which may suggest these conditions have distinct profiles in this dataset. Notably, there appears to be a correlation between the presence of cystic fibrosis and massive transfusion, and the use of CPB and massive transfusion, suggesting that patients with CF or who underwent CPB may be at higher risk of requiring a massive transfusion.

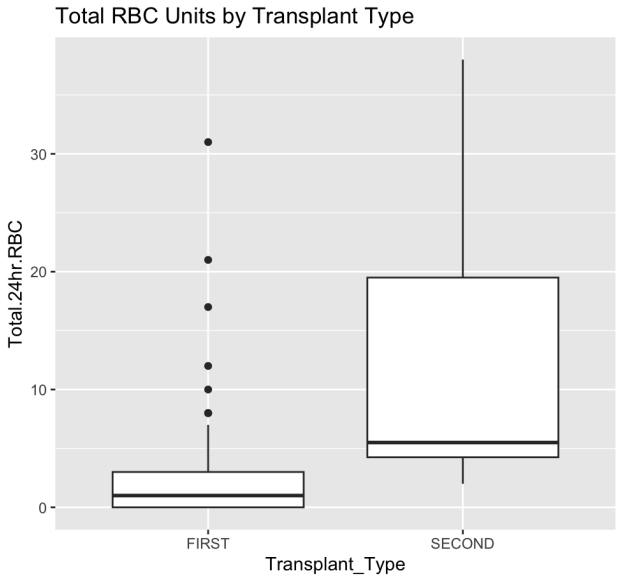


Figure 9: Box Plot of total RBC Units transfused by transplant type

Figure 9 compares the total RBC units transfused over 24 hours between two groups of lung transplant patients, categorized by the type of transplant received—labeled as "FIRST" and "SECOND" on the x-axis—in order to compare transfusions between patients undergoing their first lung transplant and those receiving a subsequent transplant. The minimum, 1st quartile, median, 3rd quartile, and maximum amount of RBCs transfused are visibly higher for patients undergoing a second transplant, aligning with the literature investigation.

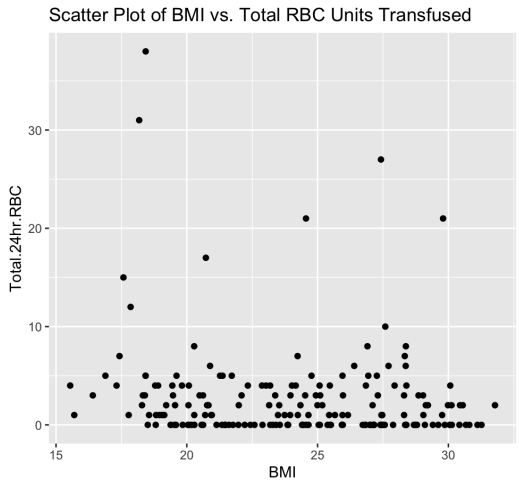


Figure 10: Scatter plot of BMI by Total RBC Units Transfused

Figure 10 illustrates the relationship between Body Mass Index (BMI) and the total number of red blood cell (RBC) units transfused within a 24-hour period. The plot indicates a wide distribution of BMI values, primarily ranging from around 15 to 30. The majority of data points are concentrated at the lower end of the y-axis, suggesting that most patients received only a small number of RBC units, regardless of their BMI. There are, however, several outliers with higher RBC transfusions across the BMI spectrum. There does not appear to be a clear linear trend or correlation evident from the scatter plot, as data points are spread without a discernible pattern that would indicate a strong relationship between BMI and RBC transfusion amounts. The plot shows that patients across a variety of BMI levels generally received low to moderate amounts of RBC units, with a few exceptions receiving higher amounts. This could imply that BMI alone is not a strong predictor of the volume of RBC units a patient might require post-transplant.

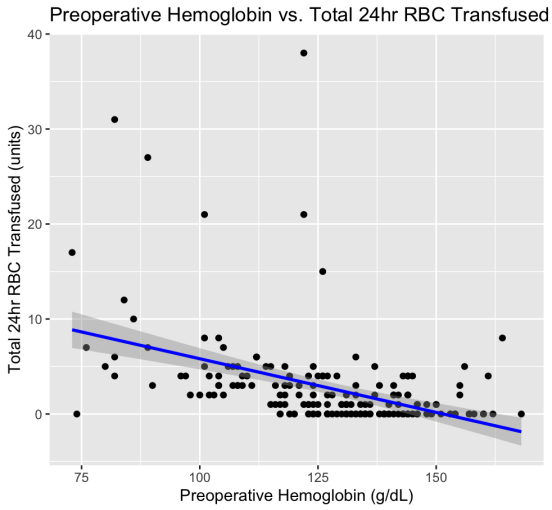


Figure 11: Pre_Hb levels vs. RBC Transfusion Levels

Figure 11 displays the relationship between preoperative hemoglobin levels (measured in grams per deciliter, g/dL) and the total number of red blood cell (RBC) units transfused within a 24-hour period. The x-axis represents preoperative hemoglobin levels, while the y-axis shows the total RBC units transfused. A trend line has been added to the scatter plot, indicating a negative correlation between the two variables. The trend line, along with the shaded area representing the confidence interval, suggests that as

preoperative hemoglobin levels increase, the total number of RBC units transfused tends to decrease. This is in line with clinical expectations, as patients with higher hemoglobin levels before surgery are generally less likely to require large amounts of blood transfusions. The concentration of data points near the lower end of the y-axis and their dispersion as hemoglobin levels decrease highlight that patients with lower preoperative hemoglobin are more variable in their transfusion requirements.

In conclusion, our exploratory data analysis (EDA) revealed several noteworthy insights into the characteristics of patients in the dataset. Firstly, it was observed that the majority of patients did not require a significant volume of RBC transfusions, suggesting that blood transfusion was not a common procedure among the patients. Interestingly, the analysis suggested a potential gender-related difference, as males appeared to require fewer transfusions compared to females. This observation contradicts existing literature and warrants further investigation to identify the significance of this relationship. Furthermore, our findings align with existing literature in the case of patients on their second transplant, who required a substantially higher volume of Total 24-hour RBC transfusions. This highlights the importance of considering previous transplant history when assessing transfusion needs. Additionally, a noticeable negative correlation was identified between the preoperative hemoglobin (Pre_Hb) levels and the volume of Total 24-hour RBC transfusions. This correlation underscores the significance of pre-operative hemoglobin levels in predicting transfusion requirements. Lastly, our analysis revealed correlations between cystic fibrosis and massive transfusion, as well as between CPB (Cardiopulmonary Bypass) and massive transfusion. These findings are consistent with existing literature, suggesting that patients with cystic fibrosis and those undergoing CPB may be at a higher risk of requiring massive transfusions. These patient characteristics associated with the need for transfusions were followed up with more rigorous statistical analysis in the following section.

3.2 Research Question 1.2 - Factors Influencing Need and Amount of Transfusion

To follow up on the insights from **Section 3.2** and assess which of the candidate predictors were most important in influencing the need and amount of transfusions for the lung transplant patients in this dataset, the Pre_df dataset (containing all candidate predictors for Research Question 1.2) was used to train Lasso regression models in a bootstrapped fashion, with a total of 2000 bootstraps (as described in **Section 2.7**). The predictors which were present in >80% of the models associated with the minimum lambda across the 2000 iterations were considered most important, and were used to fit a linear or logistic regression model. As previously stated, the resulting p-values cannot be normally interpreted due to the bootstrapped Lasso process for variable selection. However, the beta-coefficients offer some insights.

First, this process was applied with Total.24hr.RBC (a continuous variable; the total amount of red blood cell units transfused to the patient over 24 hours) as the outcome variable, to determine which factors influence the amount of transfusions in these patients. The exploratory data analysis indicated that Total.24hr.RBC likely has a linear relationship with some of these predictors, showing that Lasso regression is likely an appropriate approach (independence is also assumed; as the model will not be used for inference, normality and homoscedasticity of residuals were not assessed).

Nine predictors were retained in at least 80% of the trained models (**Figure 12**): Pre_Hb (hemoglobin level before surgery), Transplant_Type (whether it was the patients' first or second transplant),

ECLS_ECMO (the use of ECMO), Intra_Albumin.5.mL (the amount, in mL, of 5% albumin solution received during surgery), Gender.male (whether the patient was male), Intra_Crystalloid.mL (the amount, in mL, of crystalloid fluid received during surgery), ECLS_CPB (the use of CPB), Pre_Platelets (platelet count before surgery), and ExVIVO.Lung.Perfusion (whether ex vivo lung perfusion was used). The consistent selection of these predictors across a large number of bootstrapped Lasso models indicates their importance as predictors for the amount of transfusion in these lung transplant patients. These predictors were subsequently used to fit a linear regression model with Total.24hr.RBC as outcome (**Table 1**).

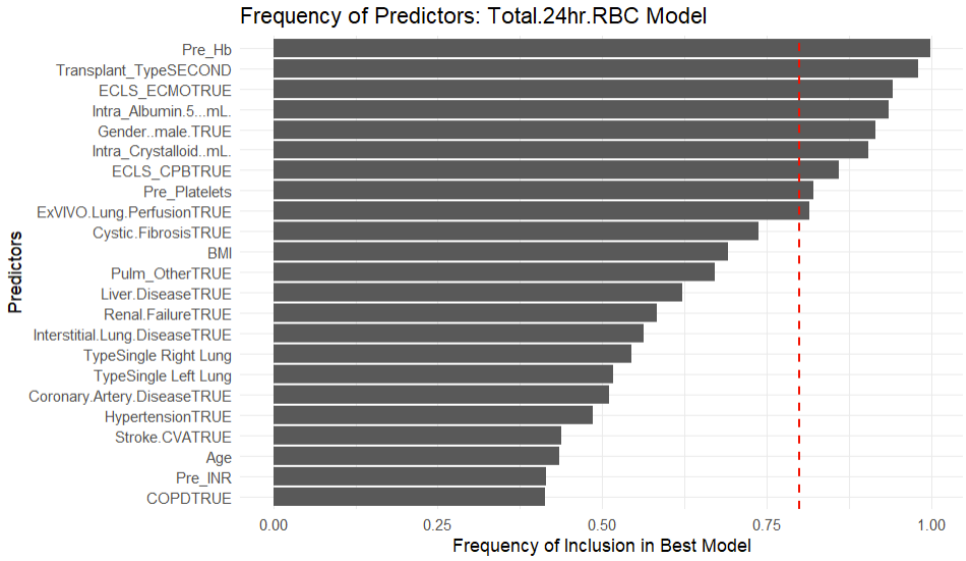


Figure 12: Frequency plot displaying the frequency of predictors' inclusion in the Total.24hr.RBC Lasso regression model associated with the optimal lambda, over 2000 bootstraps. Dashed red line denotes frequency cutoff for 'most important' predictors (80%).

Table 1: Beta coefficients, t-values, and p-values associated with each predictor in the linear regression model for Total.24hr.RBC.

Linear Regression Model: Total.24hr.RBC			
Predictor	**Coefficient**	**T-value**	**P-value**
β_0 (Intercept)	13.175	6.687	2.70e-10
Pre_Hb	-0.102	-7.156	1.97e-11
Transplant_Type (SECOND)	7.835	6.526	6.51e-10
ECLS_ECMO (TRUE)	0.991	1.691	0.093
Intra_Albumin.5.mL	0.001	3.452	0.0007
Gender.male (TRUE)	-1.345	-2.465	0.015
Intra_Crystalloid.mL	0.0004	2.127	0.035
ECLS_CPB (TRUE)	12.184	4.845	2.70e-6
Pre_Platelets	-0.004	-1.700	0.091
ExVIVO.Lung.Perfusion (TRUE)	1.054	2.003	0.047

In **Table 1**, the estimated intercept coefficient β_0 represents the expected units of RBC a patient would have transfused assuming that all predictors were set at zero or their reference level. For the remaining predictors, the coefficient represents the change in the expected value of RBC units transfused over 24 hours for a one-unit change in the corresponding predictor, while holding all other predictors constant. For the categorical predictors, the coefficient represents the change in expected transfused RBC units between the denoted level and the reference level (eg. between Gender.Male being TRUE and FALSE).

Notably, higher hemoglobin levels before surgery appear to be inversely related to the need for transfusion, as observed in **Section 3.1**. This suggests that patients with higher initial hemoglobin levels may require fewer RBC units transfused, potentially due to a greater existing reserve of red blood cells. Higher platelet count before surgery has a similar relationship to the outcome but with a smaller magnitude, likely for similar reasons. Notably, the model suggests that male patients require less RBC transfusion, aligning with **Section 3.1** and contradicting literature descriptions of male blood loss during orthopedic surgery, suggesting that gender may play different roles in blood loss depending on the surgery.

Patients on their second (or redo) lung transplant require sharply more RBC units; this also reflects **Section 3.1** and assertions in the literature that previous lung transplantation is a risk factor for increased blood loss in lung transplant surgery. The use of CPB, ex vivo lung perfusion, and to a much smaller extent ECMO, are also associated with increased RBC transfusion in these patients; this also reflects the previous investigation and the literature. Finally, the administration of intra-surgery albumin and crystalloid were also associated with increased RBC transfusion, reflecting previous hypotheses that albumin administration results in increased risk for blood transfusion and suggesting that the poorer outcomes experienced by patients with a higher crystalloid:RBC transfusion ratio may be a result of crystalloid requiring the patients to receive even more blood.

Next, the analytical workflow was applied with Massive.Transfusion (a categorical variable denoting whether the patient had a transfusion of >10 RBC units) as the outcome variable, to determine factors influencing the need for a *massive* transfusion. Three predictors were retained in at least 80% of the tuned classifiers: Pre_Hb, Transplant_Type, and Pre_Platelets (**Figure 13**). These predictors were used to fit a logistic regression model with Massive.Transfusion as the outcome (**Table 2**).

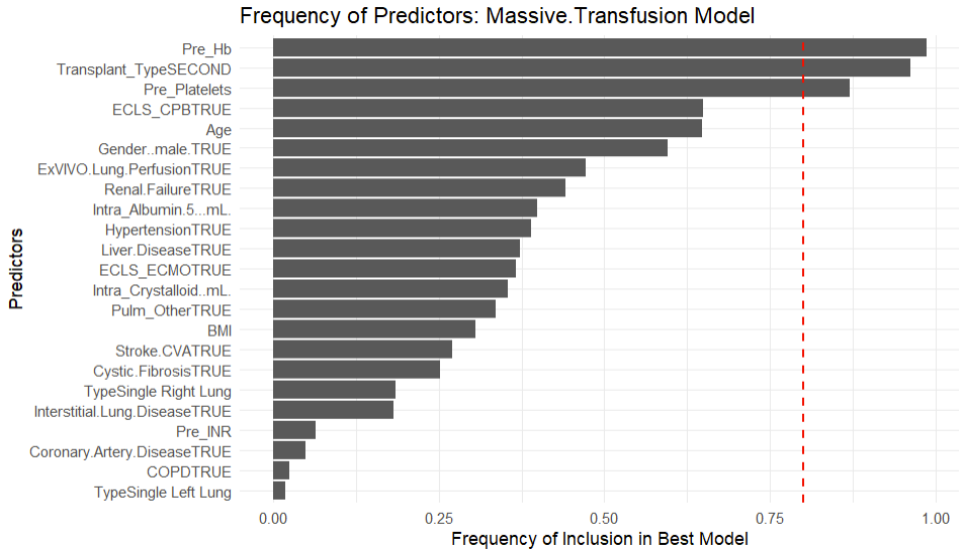


Figure 13: Frequency plot displaying the frequency of predictors' inclusion in the Massive.Transfusion Lasso classifier associated with the optimal lambda, over 2000 bootstraps. Dashed red line denotes frequency cutoff for 'most important' predictors (80%).

Table 2: Beta coefficients, t-values, and p-values associated with each predictor in the logistic regression model for Massive.Transfusion.

Logistic Regression Model: Massive.Transfusion			
Predictor	**Coefficient**	**T-value**	**P-value**
β_0 (Intercept)	7.307	2.857	0.004
Pre_Hb	-0.076	-3.341	0.0008
Transplant_Type (SECOND)	3.572	3.423	0.0006
Pre_Platelets	-0.011	-1.885	0.059

In **Table 2**, the estimated intercept β_0 represents the log of odds of the participant requiring massive transfusion (Massive.Transfusion = 1). For the continuous predictors (Pre_Hb and Pre_Platelets), the coefficient represents the change in log of odds of massive transfusion for a one-unit increase in the predictor, assuming the other predictors remain constant. For Transplant_Type, the coefficient represents the change in log of odds of massive transfusion for patients who are on their second (or redo) transplant compared to the reference level, patients who are on their first.

Interestingly, the bootstrapping process converged on a much smaller set of very influential predictors; this may indicate that only a few factors are particularly important in determining the likelihood of a massive transfusion. Similarly to the previous linear regression model, higher levels of Pre_Hb and Pre_Platelets are associated with a lower probability of the patient requiring massive transfusion, and patients on a redo transplant have a significantly higher probability of requiring massive transfusion. These predictors were represented in both models, further underscoring their utility as pre-surgery predictors of whether and how much transfusion a patient would require.

Finally, the workflow was applied with Had.Transfusion (a categorical variable denoting if a patient had any transfusion of RBCs, FFP, Plt, or Cryo) as the outcome variable, to determine the factors influencing need for *any* transfusion. Seventeen predictors were present in >80% of the tuned classifiers: Pre_Hb, ECLS_ECMO, Pulm_Other (whether the patient had other pulmonary diagnoses), Intra_Albumin, Intra_Crystalloid, Type (if the patient had a single transplant of either lung), Pre_Platelets, BMI, Gender.Male, Renal.Failure, Hypertension, ExVIVO.Lung.Perfusion, Coronary.Artery.Disease, Liver.Disease, and Transplant_Type (**Figure 14**). These predictors were used to fit a logistic regression model with Had.Transfusion as the outcome (**Table 3**).

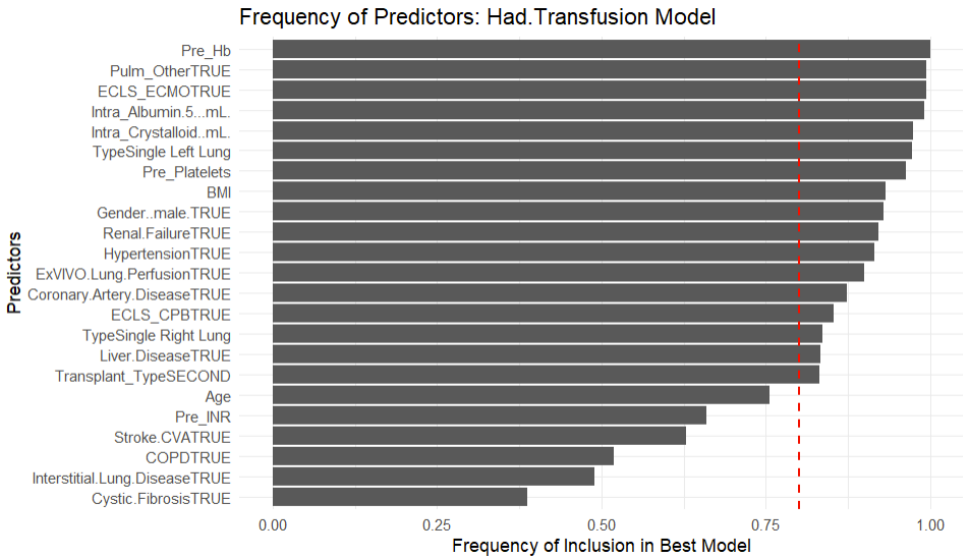


Figure 14: Frequency plot displaying the frequency of predictors' inclusion in the Had.Transfusion Lasso classifier model associated with the optimal lambda, over 2000 bootstraps. Dashed red line denotes frequency cutoff for 'most important' predictors (80%).

Table 3: Beta coefficients, t-values, and p-values associated with each predictor in the logistic regression model for Had.Transfusion.

Logistic Regression Model: Had.Transfusion			
Predictor	**Coefficient**	**T-value**	**P-value**
β_0 (Intercept)	8.478	3.288	0.001
Pre_Hb	-0.08	-4.527	5.99e-6
ECLS_ECMO (TRUE)	1.468	2.806	0.005
Pulm_Other (TRUE)	3.316	2.540	0.011
Intra_Albumin.5.mL	0.001	2.794	0.005
Intra_Crystalloid.mL	0.0004	2.369	0.018
Type (Single Left Lung)	-1.331	-1.496	0.134
Type (Single Right Lung)	-0.917	-1.092	0.274
Pre_Platelets	0.005	2.237	0.025
BMI	-0.096	-1.457	0.145
Gender.Male (TRUE)	-0.994	-1.946	0.051

Renal.Failure (TRUE)	15.87	0.009	0.993
Hypertension (TRUE)	0.974	1.859	0.063
ExVIVO.Lung.Perfusion (TRUE)	-0.726	-1.500	0.133
Coronary.Artery.Disease (TRUE)	-0.950	1.674	0.094
ECLS_CPB (TRUE)	19.54	0.004	0.996
Liver.Disease (TRUE)	1.515	1.052	0.293
Transplant_Type (SECOND)	16.14	0.011	0.991

The interpretation of the beta coefficients in **Table 3** is the same as that of the Massive.Transfusion model. Interestingly, many more predictors were present in >80% of the tuned models than in the two previous analyses, suggesting that the need of *any* transfusion in lung transplant patients may be influenced by a wider array of factors. Interestingly, Pre_Hb, ECLS_ECMO, Intra_Albumin, Intra_Crystalloid, Pre_Platelets, Gender.Male, ExVIVO.Lung.Perfusion, ECLS_CPB, and Transplant_Type all emerged as important predictors in at least one of the previous bootstrapped Lasso models, suggesting their general importance as predictors for transfusion.

The remaining predictors are unique to this model and highlight some additional factors that may influence the need for RBC, FFP, Plt or Cryo transfusion. The presence of other pulmonary issues increases the probability of requiring some form of transfusion, which is expected as various pulmonary conditions have different effects on the body and may affect surgery. The presence of hypertension emerged as possibly increasing the probability of transfusion, reflecting the literature. Having a single lung transplant (whether left or right) appears to lower the probability of transfusion compared to a double transplant, which also reflects the literature. The presence of renal failure and liver disease are both associated with a higher probability of requiring transfusion; pre-operative renal failure has a known association with RBC transfusion in cardiac surgery, while liver diseases have not been studied as a predictor in these contexts but may be on the causal pathway for blood coagulation issues, highlighting an interesting result. Finally, higher BMI and the presence of coronary artery disease appear to lower the probability of requiring a transfusion, contradicting prior assumptions.

Furthermore, three predictors have significantly larger beta coefficients: Renal.Failure, ECLS_CPB, and Transplant_Type. While these coefficients might suggest a strong association with the probability of requiring transfusion, their magnitude raises concerns about potential issues in the model. Large beta coefficients can indicate problems such as data separation, where these variables almost perfectly predict the outcome, leading potentially unreliable estimates. Consequently, additional analyses or model validation strategies may be required to confirm these findings and understand their implications fully.

3.3 Research Question 2 - Impact of Transfusion on Patient Outcomes

The analysis of transfusion effects on postoperative outcomes, using simple regression models, yielded mixed results. **Table 4** presents these findings, which indicate a significant relationship between the volume of transfused RBCs and both short-term and long-term patient outcomes. Specifically, for patients who require an additional unit of RBCs transfused, there was a significant increase in hospital length of

stay (Coefficient = 1.8029, p = 0.00318), and a decrease in the probability of being alive at 30 days (Coefficient = -0.03655, p = 1.27E-10), at 90 days (Coefficient = -0.02557, p = 1.54E-15), and at 12 months (Coefficient = 0.02872, p = 4.74E-14) post-surgery. Furthermore, the need for reoperation due to bleeding within 24 hours was also significantly associated with the volume of RBCs transfused (Coefficient = 0.20609, p = 3.48E-08). Lastly, this analysis revealed a statistically significant association between the amount of RBCs transfused within 24 hours and the death variable. This significant correlation with death was a critical factor in selecting total RBC transfusion as the variable of interest for Kaplan-Meier survival analysis.

Table 4: Analysis of the impact of transfusion of total red blood cells in 24 hours on patient outcomes using simple regression.

Outcome Variable	Coefficients - Intercept	Coefficients - Total.24hr.RBC	P-value
Hospital Length of Stay	28.7828	1.8029	0.00318*
Alive 30 Days After Surgery	4.2614	-0.03655	1.27E-10*
Alive 90 Days After Surgery	3.08902	-0.02557	1.54E-15*
Alive 12 Months After Surgery	1.9246	0.02872	4.74E-14*
Death	-1.51036	-0.0419	1.58E-11*
Need for Reoperation for Bleeding Within 24h	-5.52949	0.20609	3.48E-08*

When examining the impact of massive transfusion, as shown in **Table 5**, the results suggest a considerable increase in hospital length of stay (Coefficient = 44.56, p = 0.00188). However, the relationship between massive transfusion and survival at 30, 90, and 12 months, as well as death and the need for reoperation for bleeding within 24 hours, was not statistically significant (p-values ranging from 0.65 to 0.995).

Table 5: Analysis of the impact of massive transfusion on patient outcomes using simple regression.

Outcome Variable	Coefficients - Intercept	Coefficients - Massive.Transfusion1	P-value
Hospital Length of Stay	31.51	44.56	0.00188*
Alive 30 Days After Surgery	4.0943	14.4717	0.995
Alive 90 Days After Surgery	2.9618	15.6042	0.994
Alive 12 Months After Surgery	1.94	15.626	0.991
Death	-1.5899	-0.4895	0.65
Need for Reoperation for Bleeding Within 24h	-22.57	22.34	0.995

As for the Cox proportional hazards model (shown in **Table 6**) exploring the impact of total RBCs transfused within 24 hours post-surgery, the model did not indicate a statistically significant association with patient mortality hazard (Coefficient = -0.07753, Hazard Ratio = 0.9254, p = 0.4771). Type of lung transplant surgery, specifically single left lung transplants, showed a significant negative coefficient,

suggesting a lower hazard of mortality compared to other types such as bilateral transplant which is reduced to be the reference category here (Coefficient = -2.32715, Hazard Ratio = 0.09757, p = 0.0139). Age presented a borderline significant positive coefficient, indicating a slight increase in the hazard of mortality per year of age (Coefficient = 0.04973, Hazard Ratio = 1.05098, p = 0.0570), hinting at increased mortality risk with advancing age. Other variables included in the model, such as gender, BMI, and various comorbid conditions like COPD, cystic fibrosis, and liver disease, did not demonstrate a statistically significant impact on survival. The analysis of the proportional hazards assumption through a cloglog plot confirmed that the assumption was not violated, suggesting that the time-to-event model adequately reflects the proportional hazards of the covariates over time (**Figure 15**).

Table 6: Cox proportional hazards model results for total 24-hour RBC transfusion and controlling predictors.

Cox Proportional Hazards Model			
Predictor	**Coefficient**	**Hazard Ratio (Exp(Coefficient))**	**P-value**
Total.24hr.RBC	-0.07753	0.9254	0.4771
Type (Single Left Lung)	-2.32715	0.09757	0.0139*
Type (Single Right Lung)	-0.61047	0.5431	0.34
Gender (Male)	-0.32567	0.72204	0.5262
Age	0.04973	1.05098	0.0570*
BMI	-0.00606	0.99396	0.9474
COPD (True)	-1.67605	0.18711	0.1911
Cystic Fibrosis (True)	-0.34373	0.70912	0.8572
ILD (True)	0.22663	1.25436	0.8653
Pulmonary Other (True)	0.73108	2.07733	0.6582
Coronary Artery Disease (True)	-0.58537	0.5569	0.4489
Hypertension (True)	0.43309	1.54201	0.4853
Renal Failure (True)	-0.53283	0.58694	0.6848
Stroke/CVA (True)	-2.24413	0.10602	0.0791
Liver Disease (True)	-0.98456	0.3736	0.5992

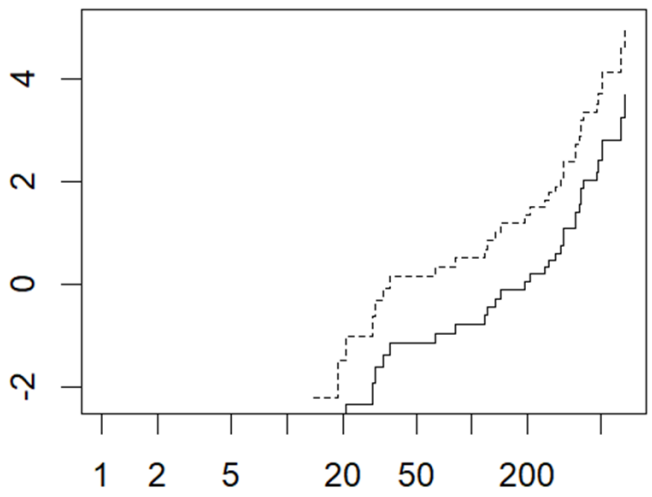


Figure 15: Cox Hazard cloglog Plot The relationship between survival probabilities and time is demonstrated, aiding in the evaluation of the proportional hazards assumption in the Cox model.

Furthermore, in an attempt to further assess mortality, the Kaplan-Meier curves stratified by the volume of RBCs transfused within 24 hours, suggested no significant differences in survival probabilities among between patients receiving regular versus massive transfusions ($p = 0.12$). Although the visual trend was apparent, the statistical analysis did not demonstrate a significant difference in survival probabilities between these group (**Figure 16**).

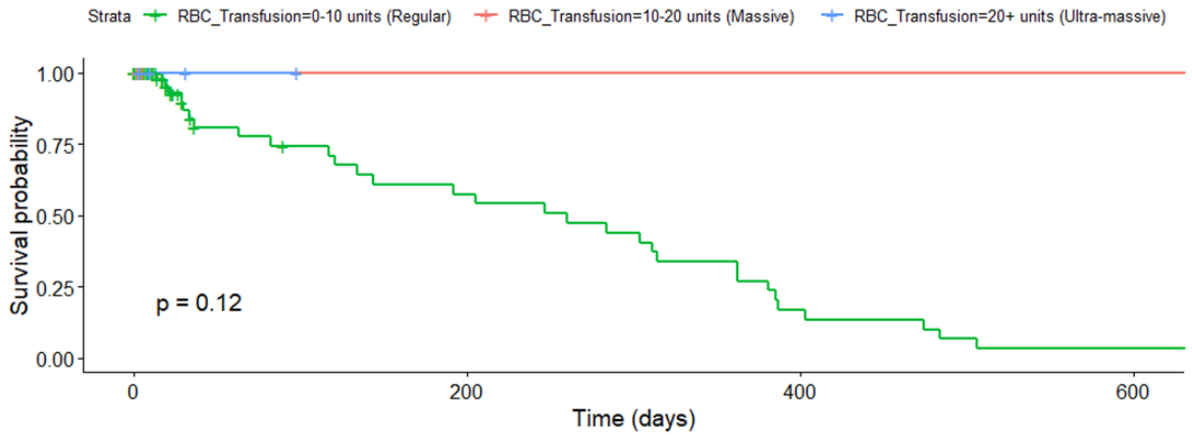


Figure 16. Kaplan-Meier survival curves based on three categories of total red blood cell (RBC) transfused within 24 hours post-surgery: those who received 0-10 units (Regular), 10-20 units (Massive), or more than 20 units (Ultra-Massive). The x-axis measures time in days, while the y-axis represents the survival probability of patients over time. The p-value (0.12) indicates the statistical significance of the difference between the survival curves.

4. Discussion

Overall, this investigation identified some characteristics of lung transplant patients requiring transfusion in this dataset, and highlighted some potential influential factors, or predictors, that determine patients' need for and amount of transfusions (with need represented through the probability of a patient needing a massive transfusion or any transfusion, and amount represented through the amount of RBC units transfused in a 24 hour period). This investigation suggests that different factors may influence patient need for transfusion and the amount of transfusion, with some factors common to both. Furthermore, this investigation identified that the volume of RBCs transfused within 24 hours is significantly associated with an increased hospital length of stay and a decreased probability of survival in the short-term and long-term post-surgery. However, the analysis using the Cox proportional hazards model indicated that the total amount of RBCs transfused in 24 hours did not significantly impact long-term mortality risk, suggesting a more complex interplay of postoperative factors influencing patient survival.

While transfusion should remain an integral part of any intense surgical process, recent evidence has highlighted that it may be linked to poorer patient outcomes in the context of lung transplant surgery. Consequently, understanding which patients may be at greater risk for blood loss or require transfusion can aid in clinical decision making with the goal of improving patient outcomes. However, the specific predictors for blood loss and the need for transfusion in lung transplant are not well understood.

Here, we leveraged a dataset containing demographic, clinical, and treatment information for 193 lung transplant patients across 113 variables, along with the amount of RBCs received by each patient over 24 hours, whether the patient received a massive transfusion, and whether the patient received any transfusion, to identify the characteristics of patients who require transfusions and the factors influencing the need and amount of transfusions. First, a literature review was conducted to narrow the 113 variables down to a set of candidate predictors associated with blood loss or transfusions in the literature. Exploratory data analysis revealed that patients with lower preoperative hemoglobin levels and those undergoing repeat lung transplants were more likely to require blood transfusions. Specifically, individuals receiving a subsequent lung transplant, as opposed to their first, tended to need a higher number of RBC units, suggesting increased transfusion requirements in repeat procedures. Furthermore, a higher incidence of transfusions was observed in patients with specific comorbidities such as COPD and renal failure, indicating a correlation between certain underlying health conditions and the need for transfusion support. Subsequently, bootstrapped Lasso regression and linear/logistic regression was used to perform further variable selection and identify the most important predictors for each outcome variable. For Total.24hr.RBC, these were found to be hemoglobin and platelet count before surgery, gender, transplant type, the use of ECMO, use of CPB, use of ex vivo lung perfusion, and albumin and crystalloid used during surgery. For Massive Transfusions, these were found to be hemoglobin and platelet count before surgery and transplant type. A much larger set of important predictors were found for the need for Any Transfusion, suggesting that it may be influenced by a wider array of factors.

While the majority of important predictors reflect those described in the literature, some new insights were found by this analysis. First, the analyses always found that higher pre-surgery hemoglobin and platelet levels are important predictors for the need and amount of transfusions, highlighting their utility to surgeons as easy-to-measure and strong predictors. Two of the models suggested that the male gender

is associated with a lower need and amount of transfusions, contradicting literature descriptions in other surgical areas that males require more supplemental blood. This indicates that predictors in certain transplant areas may not easily generalize to others, and as such predictors for transfusion should be assessed individually for related groups of transplants if possible. Liver disease emerged as a possible predictor for the need for transfusion; the liver is responsible for several important anticoagulant agents but is not described in the literature as important in this context, warranting further investigation. Finally, the presence of coronary heart disease and higher BMI appear to lower the probability of requiring a transfusion, while these were assumed to have the opposite effect before the analysis. Overall, these results reinforce known predictors for transfusion requirements in lung transplant, but also challenge existing assumptions and introduce new variables, emphasizing the need for a nuanced and context-specific understanding of transfusion predictors across different transplant types.

Next, our variable selection for the transfusion variable to address our second objective was guided by an intent to distinguish between 'total RBCs given within 24 hours' and 'total RBCs given within 72 hours.' It appeared that a significant transfusion event may have occurred within the first 24 hours, which could explain why “the total RBCs given within 24 hours” values exceed those 'total RBCs given within 72 hours”. It is also possible that “the total RBCs given within 24 hours” aggregated data from multiple sources or checkpoints, whereas the 72 hour total one might not.. Lastly, due to excessive missing data (>80%) in the columns from which the total RBCs given within 72 hours variable was drawn from, we grounded our primary analysis in the 'total RBCs given within 24 hours' measure.

The results from the regression analysis offered insights on the role transfusions play in postoperative outcomes. The significant relationship between the volume of RBCs transfused within 24 hours and both short-term and long-term outcomes was noteworthy. Increased RBC transfusion volumes are associated with longer hospital stays, which could suggest more complex postoperative courses or complications. The inverse relationship with survival at 30 and 90 days and 12 months post-surgery underscores the gravity of transfusion decisions made during the critical postoperative period. The need for reoperation due to bleeding within 24 hours being significantly associated with transfused RBC volumes aligns with clinical expectations. This could reflect more severe intraoperative or postoperative bleeding events, often necessitating more aggressive intervention and potentially leading to poorer outcomes. The marked association between RBC transfusion volume and mortality further emphasizes the potential risks associated with transfusion, reinforcing the importance of judicious blood management in surgical care.

Morever, the lack of statistical significance in the relationship between massive transfusion and various survival endpoints, as well as the need for reoperation, presents a complex picture. While massive transfusion is undoubtedly a marker of severe intraoperative issues and is associated with increased length of stay, its direct impact on survival was less clear. This might indicate that while massive transfusion is a response to critical situations, the underlying conditions necessitating such interventions are likely the actual drivers of patient prognosis. This analysis, while informative, must be contextualized within the broader clinical scenario. The decision to transfuse, particularly in large volumes, is often a response to life-threatening situations, and the outcomes may reflect the severity of the underlying medical condition rather than the transfusion itself.

The Cox proportional hazards model's findings, focusing on the total RBCs transfused within 24 hours, suggested a unique understanding of postoperative care in lung transplant patients. Despite the

expectation of a direct relationship, the data did not demonstrate a significant impact of RBC transfusion volume on mortality risk. This finding aligns with the observed complexity of postoperative recovery, where multiple factors intertwine to determine patient outcomes. The significant impact observed in cases of single left lung transplants and the near-significant influence of age ($p = 0.0570$) further underscored the multifaceted nature of post-transplant recovery.

The Kaplan-Meier survival analysis complemented these insights by visually depicting the survival probabilities based on transfusion volumes. This analysis depicted that patients receiving a 'regular' volume of transfusion (0-10 units) have different survival probabilities compared to those receiving a 'massive and ultra massive' transfusions (10+ units). The survival probability for those who received a massive transfusion volume remained relatively high and stable over time, whereas there is a noticeable decline for patients with regular transfusions, indicating a decrease in survival probability as time progresses. However, the lack of statistical significance ($p = 0.12$) cautioned against drawing definitive conclusions about the impact of transfusion volume on survival from these observations alone. Therefore, while there is a visible trend, we cannot conclusively state that the volume of transfused RBCs within 24 hours has a statistically significant impact on survival based on this analysis alone.

4.2 Limitations

This analysis is subject to several limitations as a result of the methods used and some assumptions made. Firstly, this data may not be representative of the greater lung transplant patient population, representing a single treatment center over one year. The initial removal of a significant number of predictors based on a literature review may have removed predictors that are actually important but have not been previously assessed. Furthermore, the use of stochastic imputation was done under the assumption that the missing data had a MAR pattern; due to the use of machine learning methods, multiple imputation could not feasibly be done, which could have introduced bias to the imputed dataset. Also, several variables were combined into a single Had.Transfusion variable; combining the variables led to a loss of information and it may have been more helpful to investigate each of the four transfusion types individually (although this would take significantly more computational time and several of the transfusion types had low representation in the dataset). The use of VIF as the sole tool to assess multicollinearity is also a limitation, given the arbitrary VIF threshold of 5 and only assessing VIF in the context of Total.24hr.RBC as the outcome.

For Question 1, the approach involving bootstrapped Lasso regression has several limitations. The randomness associated with bootstrapping introduces the risk of certain variables being repeatedly selected due to random fluctuations in the data rather than true predictive power. The models also could have been hyper-specific to the fluctuations present in this dataset, and not generalizable to the greater population. Furthermore, the method used did not result in any useful p-values to determine the significance of any of the resulting predictors. Moreover, the regression models assume a linear relationship between the outcome and the predictors. This assumption was evaluated for some and not all variables, and any non-linear relationships or complex interactions between variables may not be adequately captured by the model.

For Question 2, in assessing the impact of transfusion on patient outcomes in lung transplant recipients, our approach was guided by the need to focus on transfusion as the primary predictor. The utilization of

simple linear and logistic regression models with transfusion as the single predictor was a necessary approach given our research focus. The simplicity of these models is both a strength and a limitation: they offer clarity in understanding the direct relationship between transfusion and outcomes but do not account for the complex nature of postoperative recovery where numerous factors may simultaneously influence patient outcomes. Thus, the use of multiple predictors could complicate the interpretation of transfusion's specific impact. However, when checking for the regression assumptions of variables, it was found that the model had relatively high null deviance, suggesting that the predictor (Total.24hr.RBC) may not explain a significant portion of the variance in the categorical outcomes (**Supplementary Table 4**). Moreover, in survival analysis, the Cox proportional hazards model brings its own challenges. It assumes that the hazard ratios are constant over time, an assumption that may not hold true in all clinical scenarios. Additionally, while the Kaplan-Meier estimator is a useful tool for survival analysis, it does not adjust for other variables that may affect survival, limiting its explanatory power regarding survival time disparities. Additionally, its non-parametric nature precludes the assumption of a specific distribution for survival data, potentially limiting inferential conclusions.

To mitigate these limitations, external validation of the model using datasets from different treatment centers or diverse patient populations would be essential, as it would help confirm the robustness and generalizability of the identified predictors. Also, conducting prospective studies or randomized controlled trials would offer a more robust methodological framework. These approaches allow for the establishment of clearer causal relationships, as they can control for various confounding factors and provide a more structured investigation into the direct effects of transfusion on patient outcomes and vise versa.

5. Conclusion

In conclusion, this investigation into blood transfusion practices for lung transplant patients has revealed several new details about the predictors and outcomes associated with lung transplant transfusion needs. By analyzing data from 193 lung transplant patients, the study has identified key factors such as preoperative hemoglobin levels, transplant type, and specific comorbidities like COPD and renal failure, as significant determinants of the likelihood and volume of blood transfusions. Our results also challenge some existing work in the literature, especially regarding gender differences in transfusion requirements and the role of liver disease as a predictive factor. It was observed that males might have a lower need for transfusions, contradicting existing literature in other surgical contexts. Conditions like liver disease and coronary heart disease, previously not emphasized in this context, emerged as potential influencers of transfusion needs. Furthermore, the study underscores the significant relationship between the volume of RBCs transfused within the first 24 hours and both short-term and long-term outcomes. We noted that increased transfusion volumes are linked to longer hospital stays and lowered survival rates at 30 and 90 days, as well as 12 months post-surgery. This finding points to the critical nature of transfusion decisions during the postoperative period and their potential long-term impact on patient survival. Overall, these findings highlight the need for a nuanced approach to blood management, taking into account individual patient characteristics and surgical contexts, to optimize outcomes and enhance the quality of care for lung transplant patients. Future research should aim to further investigate these predictors with the goal of improving lung transplant patient outcomes.

References

1. Egan TM, Westerman JH, Lambert CJ, Detterbeck FC, Thompson JT, Mill MR, et al. Isolated lung transplantation for end-stage lung disease: a viable therapy. Ann Thorac Surg. 1992 Apr;53(4):590–5; discussion 595-596.
2. Indications for and Adverse Effects of Red-Cell Transfusion | NEJM [Internet]. [cited 2023 Dec 1]. Available from: https://www.nejm.org/doi/full/10.1056/nejmra1612789
3. Jacques F, El-Hamamsy I, Fortier A, Maltais S, Perrault LP, Liberman M, et al. Acute renal failure following lung transplantation: risk factors, mortality, and long-term consequences. Eur J Cardiothorac Surg. 2012 Jan;41(1):193–9.
4. Oechslin P, Zalunardo MP, Inci I, Schlaepfer M, Grande B. Established and potential predictors of blood loss during lung transplant surgery. Journal of Thoracic Disease [Internet]. 2018 Jun [cited 2023 Dec 1];10(6). Available from: https://jtd.amegroups.org/article/view/21787
5. Diamond JM, Lee JC, Kawut SM, Shah RJ, Localio AR, Bellamy SL, et al. Clinical Risk Factors for Primary Graft Dysfunction after Lung Transplantation. American Journal of Respiratory and Critical Care Medicine. 2013 Mar 3;187(5):527.
6. Grande B, Oechslin P, Schlaepfer M, Seifert B, Inci I, Opitz I, et al. Predictors of blood loss in lung transplant surgery—a single center retrospective cohort analysis. J Thorac Dis. 2019 Nov;11(11):4755–61.
7. Mohite PN, Sabashnikov A, Patil NP, Garcia-Saez D, Zych B, Zeriouh M, et al. The role of cardiopulmonary bypass in lung transplantation. Clin Transplant. 2016 Mar;30(3):202–9.
8. Reducing Need for Blood Transfusion during Heart Surgery Is Focus of New Practice Guideline | STS [Internet]. [cited 2023 Dec 1]. Available from: https://www.sts.org/media/news-releases/reducing-need-blood-transfusion-during-heart-surgery-focus-new-practice-guideline
9. Noda K, Chan EG, Furukawa M, Ryan JP, Clifford S, Luketich JD, et al. Single-center experience of ex vivo lung perfusion and subsequent lung transplantation. Clinical Transplantation. 2023;37(2):e14901.
10. Coster JN, Loor G. Extracorporeal life support during lung transplantation. Indian J Thorac Cardiovasc Surg. 2021 Sep;37(Suppl 3):476–83.
11. Triulzi DJ, Griffith BP. Blood usage in lung transplantation. Transfusion. 1998 Jan;38(1):12–5.
12. Shao W, Huang J, Feng H, Sun H, Ren Y, Wang X, et al. Impact of interstitial lung disease on postoperative morbidity and 90-day mortality after pulmonary resection. Transl Cancer Res. 2020 Feb;9(2):1151–9.
13. Vonk ABA, Meesters MI, van Dijk WB, Eijsman L, Romijn JWA, Jansen EK, et al. Ten-year patterns in blood product utilization during cardiothoracic surgery with cardiopulmonary bypass in a tertiary hospital. Transfusion. 2014 Oct;54(10 Pt 2):2608–16.
14. Ringaitiene D, Puodziukaite L, Vicka V, Gineityte D, Serpytis M, Sipylaite J. Bioelectrical Impedance Phase Angle-Predictor of Blood Transfusion in Cardiac Surgery. J Cardiothorac Vasc Anesth. 2019 Apr;33(4):969–75.
15. Pempe C, Werdehausen R, Pieroh P, Federbusch M, Petros S, Henschler R, et al. Predictors for blood loss and transfusion frequency to guide blood saving programs in primary knee- and hip-arthroplasty. Sci Rep. 2021 Feb 23;11:4386.

16. Amitrano L, Guardascione MA, Brancaccio V, Balzano A. Coagulation disorders in liver disease. Semin Liver Dis. 2002 Feb;22(1):83–96.
17. Wei L, Jing B, Li H. Bootstrapping promotes the RSFC-behavior associations: An application of individual cognitive traits prediction. Human Brain Mapping. 2020;41(9):2302–16.
18. Lazzareschi DV, Fong N, Mavrothalassitis O, Whitlock EL, Chen CL, Chiu C, et al. Intraoperative Use of Albumin in Major Noncardiac Surgery: Incidence, Variability, and Association With Outcomes. Annals of Surgery. 2023 Oct;278(4):e745.
19. Jennings LK, Watson S. Massive Transfusion. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 [cited 2023 Dec 1]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK499929/
20. Diller GP, Lammers AE, Oechslin E. Treatment of adults with Eisenmenger syndrome—state of the art in the 21st century: a short overview. Cardiovasc Diagn Ther. 2021 Aug;11(4):1190–9.
21. Coronary artery disease patients requiring combined anticoagulant and antiplatelet therapy - UpToDate [Internet]. [cited 2023 Dec 1]. Available from: https://www.uptodate.com/contents/coronary-artery-disease-patients-requiring-combined-anticoagulant-and-antiplatelet-therapy
22. Karangwa SA, Adelmeijer J, Burgerhof JGM, Lisman T, de Meijer VE, de Kleine RH, et al. Controlled DCD Liver Transplantation Is Not Associated With Increased Hyperfibrinolysis and Blood Loss After Graft Reperfusion. Transplantation. 2022 Feb 1;106(2):308–17.
23. Arya RC, Wander G, Gupta P. Blood component therapy: Which, when and how much. J Anaesthesiol Clin Pharmacol. 2011;27(2):278–84.
24. Professional Education [Internet]. 2018 [cited 2023 Dec 1]. Albumin. Available from: http://professionaleducation.blood.ca/en/transfusion/clinical-guide/albumin
25. Neal MD, Hoffman MK, Cuschieri J, Minei JP, Maier RV, Harbrecht BG, et al. Crystalloid to packed red blood cell transfusion ratio in the massively transfused patient: when a little goes a long way. J Trauma Acute Care Surg. 2012 Apr;72(4):892–8.

Supplementary Information

Supplementary Table 1: Initial Variable Selection based on Literature Review. Variables were retained if the literature review indicated that they were directly found to be predictors of bleeding or transfusion in lung transplant patient or other transplant areas, or if a plausible causal link could be found to transfusion.

Variable	Descripture	Keep/Reject	Reason for Keeping or Rejecting (Literature)
Study ID	Identifies the study referenced.	**Reject**	None used. The ID did not provide any useful information.
TX.DB.ID	Another ID variable used.	**Reject**	None used. The ID did not provide any useful information.
OR.DATE	Date of admission into the OR.	**Reject**	None used. Other dates proved to be more useful for analysis.
TYPE	The type of lung transplant (Bilateral, Single left lung, Single right lung).	**Keep**	Bilateral lung transplants are more often associated with the use of cardiopulmonary bypass (CPB), which is a predictor of more blood loss.[7]
Gender.male	True/False, if TRUE the patient is male.	**Keep**	Male sex has been linked to increased perioperative blood loss requiring transfusion in orthopedic surgery, so this may be the case for lung transplant.[4]
Height	Height of the patient in cm.	**Reject**	Collinear with BMI, therefore removed from analyses.
Weight	Weight of the patient in kg.	**Reject**	Collinear with BMI, therefore removed from analyses.
Age	Age of the patient.	**Keep**	Age is a component of the Charleston Comorbidity Index, which has been linked to increased blood loss in orthopedic surgery.[4]
BMI	BMI of the patient.	**Keep**	Increased BMI is a predictor for bleeding in cardiac surgery.[13]
COPD	Whether the patient has Chronic Obstructive Pulmonary Disease.	**Keep**	A large chunk of the people who have COPD also have anemia. Anemia associated with more

			bleeding/slower wound healing.[15]
alpha1.Antitrypsin.Deficiency	Whether the patient has alpha1 antitrypsin deficiency. It is a condition that raises your risk for lung and other diseases, particularly COPD.	**Reject**	Collinear with COPD therefore removed from analyses. No sources were found to indicate relation with the need for blood transfusion.
Cystic.Fibrosis	Whether a patient has Cystic Fibrosis.	**Keep**	Cystic fibrosis patients more frequently need perioperative blood transfusion, partially because they more often get double lung transplants.[11]
Idiopathic.Pulmonary.Hypertension	Whether a patient has Idiopathic Pulmonary Hypertension. Disorder characterized by elevated blood pressure in the pulmonary arteries without an apparent cause.	**Reject**	Collinear with hypertension, therefore removed from analyses.
Interstitial.Lung.Disease	Whether the patient has interstitial lung disease. It is a lung disorder affecting the interstitium (lung tissue between air sacs), leading to inflammation and scarring.	**Keep**	Patients with ILD have been shown to lose more blood during surgery than patients without it; this may be because of the increased time and difficulty of surgeries in patients with ILD.[12]
Pulm_Other	Whether the patient has any other pulmonary issues.	**Keep**	Included this because other pulmonary diseases like Eisenmenger's syndrome have been associated with increased blood loss during surgery.[20]
Coronary.Artery.Disease	Whether the patient has coronary heart disease.	**Keep**	Patients who have CAD often take **anticoagulants** like warfarin or antiplatelet drugs like aspirin which can increase chances for the need of blood transfusion.[21]
Hypertension	Whether the patient has hypertension.	**Keep**	High systolic blood pressure has been associated with increased need for transfusion, especially

			massive transfusion, in patients with severe injuries.[22]
Diabetes.Insulin	Whether the patient has diabetes that is managed by insulin.	**Reject**	No sources were found on the link between diabetes and blood loss.
Diabetes.diet.OHGs	Whether the patient has diabetes that is managed by diet or Oral Hypoglycemic Agents.	**Reject**	No sources were found on the link between diabetes and blood loss.
GERD.PUD	Whether the patient has Gastroesophageal Reflux Disease or Peptic Ulcer Disease. GERD is a chronic condition where stomach acid frequently flows back into the esophagus, causing irritation and inflammation.	**Reject**	No sources were found on the link between GERD and blood loss.
Renal.Failure	Whether the patient had Renal Failure.	**Keep**	Renal failure is associated with the risk of RBC transfusion in cardiac surgery[14]
Stroke.CVA	Whether the patient had a stroke (cerebral vascular accident).	**Keep**	Preoperative stroke is associated with risk of RBC transfusion in cardiac surgery.[14]
Liver.Disease	Whether the patient has Liver Disease.	**Keep**	Liver diseases are often associated with coagulation disorders. “Liver Disease” is a broad enough category for this to be important.[16]
Thyroid.Disease	Whether the patient has Thyroid Disease.	**Reject**	No sources were found on the link between Thyroid Disease and blood loss.
First.Lung.Transplant	If TRUE, this is their first lung transplant, and Redo Lung Transplant will be FALSE.	**Keep**	See below.

Redo.Lung.Transplant	If TRUE, this is their first lung transplant, and Redo Lung Transplant will be FALSE.	**Keep**	Previous lung transplantation is a risk factor for increased blood loss in lung transplant surgery.[7]
DCD.vs.DBDex	Whether the patient had Donation AFTER circulatory death (i.e. organ donation from individuals who have experienced circulatory death) vs. Donation Before Death.	**Reject**	In the liver and in other disease areas**,** DCD vs. DBD doesn't make a difference for blood loss;[23] the overall evidence is quite limited.
exVIVO.Lung.Perfusion	Whether the patient had Ex vivo lung perfusion, involving the external perfusion and assessment of donor lungs outside the body to evaluate their viability and improve transplant outcomes.	**Keep**	Patients who received lungs screened with EVLP had a higher incidence of PGD3 and longer ICU and hospital stays.[9]
Preoperative.ECLS	Whether the patient was on Extracorporeal Life Support (ECLS) prior to the operation. ECLS is a treatment that uses a machine to take over the work of the lungs and sometimes the heart.	**Reject**	The lung transplantation group at Columbia University compared outcomes between patients who received ECMO support or CPB support during lung transplantation and found that the transfusion requirements were higher in the CPB group than in the ECMO group. Focused on ECMO groups specifically for analyses.[10]
LAS.Score	The Lung Allocation Score (LAS) is a numerical value used to prioritize patients awaiting lung transplants. It takes into account the severity of the patient's lung disease and their likelihood of survival.	**Reject**	No literature was found demonstrating links with blood loss/blood transfusion needs.
Pre_Hb	This is a measure of the hemoglobin level in the patient's blood prior to the surgery. Hemoglobin is a	**Keep**	Low levels would indicate a higher need for blood transfusion.[4]

	protein in red blood cells that carries oxygen.		
Pre_Hct	This measures the proportion of blood that is made up of red blood cells. It's used to diagnose and monitor conditions like anemia.	**Keep**	Low levels would indicate a higher need for blood transfusion.[4]
Pre_Platelets	The number of platelets in the patient's blood prior to surgery. Platelets are important for blood clotting.	**Keep**	Low levels would indicate a higher need for blood transfusion.[4]
Pre_PT	A test that measures how long it takes for blood to clot. It's used to check for bleeding or clotting disorders.	**Reject**	Collinear with INR, therefore removed.
Pre_INR	This is a standardized way to report PT results, particularly important for patients on anticoagulation therapy.	**Keep**	FFP is indicated to correct microvascular bleeding in the presence of elevated (>1.5-times normal) prothrombin time (PT) or partial thromboplastin time (PTT).[23]
Pre_PTT	Another test to measure blood clotting, particularly assessing the intrinsic pathway of coagulation.	**Reject**	Collinear with INR, therefore removed.
Pre_Creatinine	A measure of creatinine levels in the blood, used to assess kidney function.	**Reject**	No literature was found demonstrating links with blood loss/blood transfusion needs.
Intraoperative _ECLS	Indicates whether Extracorporeal Life Support was used during the operation.	**Reject**	Reference Preoperative_ECLS.
ECLS_ECMO	Refers to the use of Extracorporeal Membrane Oxygenation, a more specific form of ECLS, during surgery. ECMO is used in severe cases of respiratory or cardiac failure.	**Keep**	Reference Preoperative_ECLS.

ECLS_CPB	Indicates whether the patient was on cardiopulmonary bypass, a technique that temporarily takes over the function of the heart and lungs during surgery.	**Keep**	Cardiopulmonary Bypass leads to extra risks for blood transfusion needs in patients.[8]
Intra_Albumin .5…mL	The amount of 5% albumin solution administered during surgery. Albumin is used to increase blood volume and pressure.	**Keep**	Change in blood pressure and volume as a result of albumin administration can lead to increased chance for blood transfusion.[24]
Intra_Crystalloid..mL	The volume of crystalloid solutions (like saline) given to the patient during surgery. These are used for fluid replacement.	**Keep**	An increased crystalloid ratio could be associated with increased morbidity and poor outcome after MT.[25]
Intra_Cell.Saver.returned..mL	Refers to the volume of blood returned to the patient using a cell saver machine, which collects and processes the patient's blood lost during surgery for reinfusion.	**Reject**	Preoperative values are more valuable when assessing need for transfusion.
Intra_Fresh.Frozen.Plasma	The amount of fresh frozen plasma administered during the operation, used for blood clotting factor replacement.	**Reject**	Preoperative values are more valuable when assessing need for transfusion.
Intra_Packed.Cells	The volume of packed red blood cells given to the patient intraoperatively. This is typically used to treat or prevent anemia.	**Keep**	Preoperative values are more valuable when assessing need for transfusion.
Intra_PCC.Octaplex	Refers to the administration of Prothrombin Complex Concentrate or similar products, used to quickly restore blood clotting factors.	**Reject**	No literature was found demonstrating links with blood loss/blood transfusion needs.

Intra_Platelets	The amount of platelet transfusions given during the operation.	**Reject**	Preoperative values are more valuable when assessing need for transfusion.
Intra_Cryoprecipitate	The volume of cryoprecipitate transfused, which is rich in clotting factors, particularly fibrinogen.	**Reject**	Preoperative values are more valuable when assessing need for transfusion.
Blood.Loss	The amount of blood lost during the surgery.	**Keep**	Blood loss is associated with blood transfusion. The more blood that is lost, the more the need for transfusion.
Urine.Output	The volume of urine produced by the patient during surgery, an important indicator of kidney function and fluid balance.	**Reject**	No literature was found demonstrating links with blood loss/blood transfusion needs.
Fluid.Balance	Represents the net fluid balance, considering the input (like IV fluids) and output (like urine) during surgery.	**Reject**	No literature was found demonstrating links with blood loss/blood transfusion needs.
Tranexamic.Acid.Used	Indicates whether tranexamic acid, a medication used to reduce bleeding, was used during the operation.	**Reject**	No literature was found demonstrating links with blood loss/blood transfusion needs.
ICU.Admission.Date.Time	Indicates the date and time of admission to the ICU.	**Reject**	Said date was not valuable for our analyses.
ICU.Discharge.Date.Time	INdicates the date and time of discharge from the ICU.	**Reject**	Said date was not valuable for our analyses.
Duration.of.ICU.Stay.day	Duration of ICU stay in days.	**Keep**	Included but could have high risk of collinearity with HOSPITAL_LOS,
Date.of.Extubation	Extubation refers to the removal of the endotracheal tube from your lungs. It's done when mechanical ventilation is no longer needed because you can breathe on your own.	**Reject**	No literature was found demonstrating links with blood loss/blood transfusion needs.

ALIVE_30DAYS_YN	Indicates whether the patient was alive 30 days after the transplant.	**Keep**	Important for measuring outcome of transfusion, specifically when performing survival analyses.
ALIVE_90DAYS_YN	Indicates whether the patient was alive 90 days after the transplant.	**Keep**	Important for measuring outcome of transfusion, specifically when performing survival analyses.
ALIVE_12MTHS_YN	Indicates whether the patient was alive 1 year after the transplant.	**Keep**	Important for measuring outcome of transfusion, specifically when performing survival analyses.
ICU_LOS	A duplicate of Duration.of.ICU.Stay..days. and can be used for the same purposes.	**Reject**	Risk of multicollinearity with hospital length of stay based on VIF results.
HOSPITAL_LOS	The total length of stay in the hospital.	**Keep**	Used to assess impact of transfusion on patient hospital length of stay.
PostImmediate_PTT	Partial Thromboplastin Time immediately after surgery.	**Reject**	The PT test, also known as the PT/INR test, looks at the extrinsic pathway of coagulation (meaning coagulation that occurs after blood escapes a blood vessel). Collinear with the INR test.
PostImmediate_Creatinine	Creatinine levels immediately after surgery.	**Reject**	Collinear with PostDay1_Creatinine
PostDay1_Hb	Hemoglobin levels one day after surgery.	**Keep**	Low levels would impact the outcome of patients.
PostDay1_Hct	Hematocrit levels one day after surgery.	**Keep**	Low levels would impact the outcome of patients.
PostDay1_Platelets	Platelet levels one day after surgery.	**Keep**	Low levels would impact the outcome of patients.
PostDay1_PT	A test that measures how long it takes for a clot to form in a blood sample.	**Reject**	Irrelevant to study objectives.

PostDay1_INR	International Normalized Ratio: a ratio that tells you how long it takes for your blood to clot.	**Reject**	The type of calculation based on the PT test. Therefore, it is sufficient to include PT only if relevant to study objectives.
PostDay1_PTT	Similar to the immediate PostImmediate_PTT, but one day after.	**Reject**	Irrelevant to study objective and collinear with PT, therefore removed.
PostDay1_Creatinine	Indicative of kidney function one day after surgery.	**Reject**	Used to assess impact of transfusion on kidney function.
RBC.72hr.Total	Total RBCs transfused within 72 hrs.	**Reject**	Determined that the 24 hrs RBC variable is more valuable for analyses and it is only used to create a new group that assesses if a patient had any type of transfusion.
FFP.72hr.Total	Fresh Frozen Plasma transfused within 72 hrs in units.	**Reject**	Only used to create a new group that assesses if a patient had any type of transfusion.
Plt.72hr.Total	Platelet transfused in units within 72 hrs.	**Reject**	Only used to create a new group that assesses if a patient had any type of transfusion.
Cryo.72hr.Total	Total Cryoprecipitate transfused in units within 72 hrs.	**Reject**	Only used to create a new group that assesses if a patient had any type of transfusion.
Need.for.reoperation.for.bleeding.within.24h	Indicates if a patient needed reoperation for bleeding within 24 hours	**Keep**	Important to determine the outcome of the patients.
Total.24hr.RBC	Total RBC transfused within the 24 hrs.	**Keep**	Important to determine impact of transfusion on patient outcomes.
Massive.Transfusion	Indicates whether a massive transfusion (over 10 RBC units) was performed.	**Keep**	Important to determine the outcome of the patients.

Supplementary Table 2: VIF results for Pre_df, with Total.24hr.RBC as outcome. VIF scores >5 are considered indicative of problematic multicollinearity.

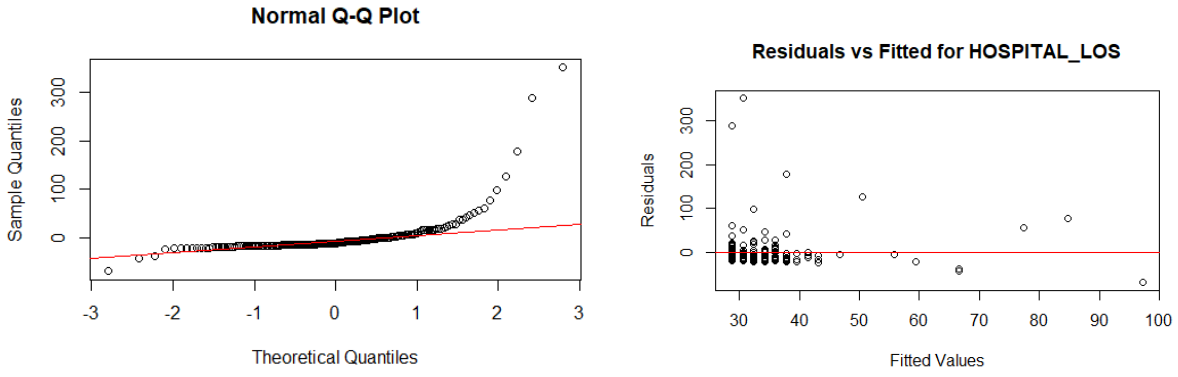
Variable	VIF		Variable	VIF
Type	2.2887		Pre_Platelets	1.4458
Gender.Male	1.5098		Pre_INR	1.2078
Age	3.1354		ECLS_ECMO	1.8628
BMI	1.6750		ECLS_CPB	1.3779
COPD	3.1155		Intra_Albumin	1.7016
Cystic.Fibrosis	3.8073		Intra_Crystalloid	1.5642
Interstitial.Lung.Disease	4.3502		Intra_Packed.Cells	11.194
Pulm_Other	1.6996		Blood.Loss	7.4603
Coronary.Artery.Disease	1.3694		Massive.Transfusion	4.4029
Hypertension	1.5343		RBC.72hr.Total	10.864
Renal.Failure	1.2920		FFP.72hr.Total	2.8593
Stroke.CVA	1.2060		Plt.72hr.Total	8.9411
Liver.Disease	1.1817		Cryo.72hr.Total	3.4948
ExVIVO.Lung.Perfusion	1.4735		Transplant_Type	1.9842
Pre_Hb	16.060		Had.Transfusion	2.4275
Pre_Hct	15.660			

Supplementary Table 3: Proportion of patients received different transfusion types.

Transfusion Type	No. Patients (out of 192)	Percent of Patients
At least one	67	8.7
Blood Transfusion	62	8.1
Plasma Transfusion (FFP + Cryo)	11	1.4
Platelet Transfusion	11	1.4

Supplementary Table 4: Checking for regression assumption for categorical variables against Total.RBC.24hrs

Outcome Variable	Null Deviance	DF Null	Log Likelihood	AIC	BIC	Deviance	DF Residual
Alive 30 Days YN	30.9	191	-15.4	34.7	41.3	30.7	190
Alive 90 Days YN	72.7	191	-36.2	76.5	83	72.5	190
Alive 12 Months YN	141	191	-70.2	144	151	140	190
Death Binary	173	191	-86.1	176	183	172	190
Need for Reoperation for Bleeding Within 24h	38.9	191	-10.5	24.9	31.5	20.9	190



Supplementary Figure 5: Checking for regression assumption for the continuous variable hospital length of stay against Total.RBC.24hrs