Analysis of Longitudinal Data - Effect of Lung Surgery for Severe Emphysema

Inês Fortes

last update: 15/07/2022

1. Introduction

Lung emphysema is a clinical condition where the air sacs in the lungs (alveoli) are damaged, causing shortness of breath [1]. Severe emphysema can seriously decrease the quality of life by limiting the type of activities that patients can do.

There are several treatments available for this condition. For the less severe cases physical therapy is used, with the main objective of halting the progressive decline in lung function [2]. Also, medication may be used to reduce symptoms and increase the quality of life [2]. However, these treatments are more palliative and none of these seems to truly improve the respiratory function. For that reason, and supported by the idea that emphysema patients have larger lungs [2], lung surgery to reduce lung capacity has been implemented. However, the scientific evidence for its efficacy was doubtful [2,3,4]. To study the efficacy of different treatments for this disease, in the 90s the National Emphysema Treatment Trial (NETT) joined more than 17 clinical centers from the USA. This trial followed more than 1000 patients, that were assessed across time to monitor their lung capacity and quality of life. More details about the original trial can be found in [2] and [3].

In the present study a sub-sample of the NETT study was obtained. The main objective of the present analysis was to determine whether respiratory function is improved with surgery in comparison with medical therapy only.

2. Exploratory data analysis

2.1 Exploring the database and choosing the dependent variable

The database is composed of 120 patients, that were divided into two groups: surgical (N = 60) and non-surgical (N = 60). The study had a balanced design, that is, patients should be measured at baseline, prior to randomization between groups (time = 0), and then 6, 12, 24, 36, 48, and 60 months (5 years) after randomization. The majority of patients were white (95%), and 62.5% were men.

Even though the study had a balanced design, it resulted in an unbalanced data set because patients (1) stopped being followed, (2) missed one visit and then returned on the following one, or (3) missed one particular measurement (i.e., they were present at a given time but they were only measured in some of the variables). Because there is no indication for the cause of missing data (e.g., intermittent missing, loss to follow-up for random reasons, or dropout), we will consider all data, assuming that missing data are uncorrelated with the variables under study (the possible implications are discussed in the Conclusion).

Figure 1 shows the percentage of patients with missing data at each moment (left panel) and the percentage of patients assessed at each moment or later (right panel).

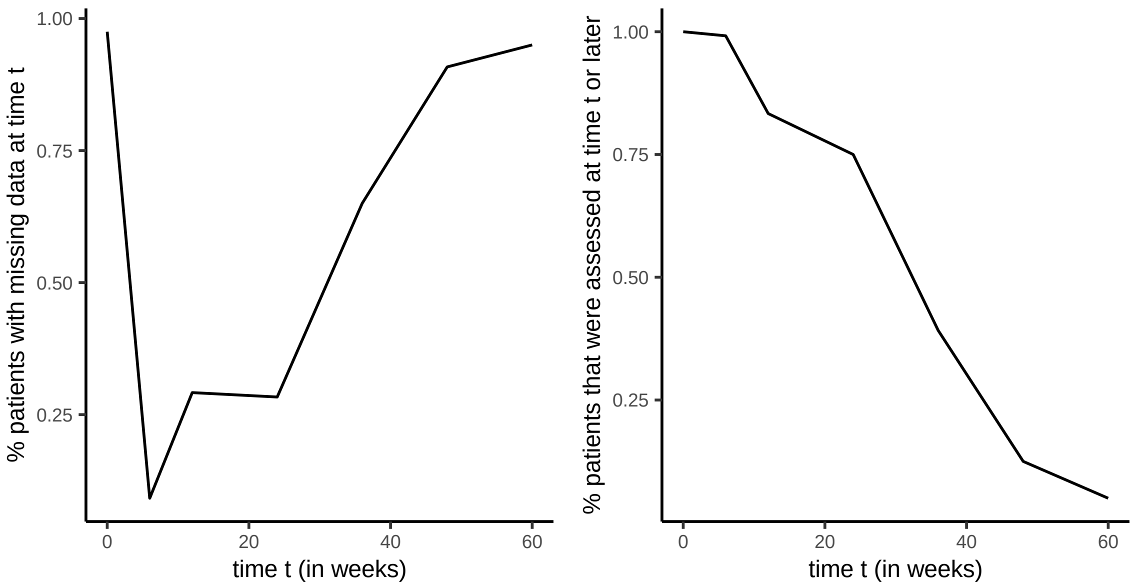

Figure 1: Percentage of patients with missing data at time t (left panel) and Percentage of patients assessed at time t of after time t (right panel).

Figure 1 shows that several patients were not assessed at the beginning of the clinical trial (97.5%, time = 0) and at 48 and 60 weeks (90.8% and 95%, respectively, see left panel). On the right panel, we can see that the proportion of patients being followed decreases. For instance, at 48 weeks only 12.5% of patients were alive for sure, because they were either measured at 48 weeks or 60 weeks (i.e., they may have missed the 48-week appointment but were assessed at 60 weeks or the other waya around). The other 87.5% of patients were not measured neither at 48, nor at 60 weeks. This may be due to death, or other unrelated reason.

Table 1 shows all the variables included in the data set.

Table 1: Description of variables in the data set

Variable	Description
NEWNETT	Patient identification
VISIT	Visit times: 0, 6, 12, 24, 36, 48, and 60 months from randomization
MEDID	Group: Non-surgical or Surgical
GENDER	Gender: Male or Female
ETHNIC	Ethnic: White or Other
PREDFVC	Forced vital capacity (decrease = improvement)
PREDFEV1	Forced expiratory volume in one second (increase = improvement)
TLC	Total lung capacity (decrease = improvement)
RV	Residual volume (decrease = improvement)
PAO2	Arterial oxygen (increase = improvement)

Table 1 shows that besides the time-independent variables such as group (MEDID), gender and ethnic, there were five dependent variables, all associated with lung function. To assess the effect of surgery on lung function, the first step was to select one of these five variables. For that, we chose the one with (1) more observations (i.e., with less NA or missing data) and (2) an approximately normal distribution.

All variables had a similar number of NA data (min = 490 for PREDFVC and PREDFEV1; max = 494 for PAO; Note that the maximum NA was 7 visits * 120 patients = 840). Thus, we resorted to the second criterion, and the variable with a distribution closer to normal is PAO (potential arterial oxygen).

Figure 2 shows the histogram of PAO2 (left panel) and its distribution as a function of treatment (right panel).

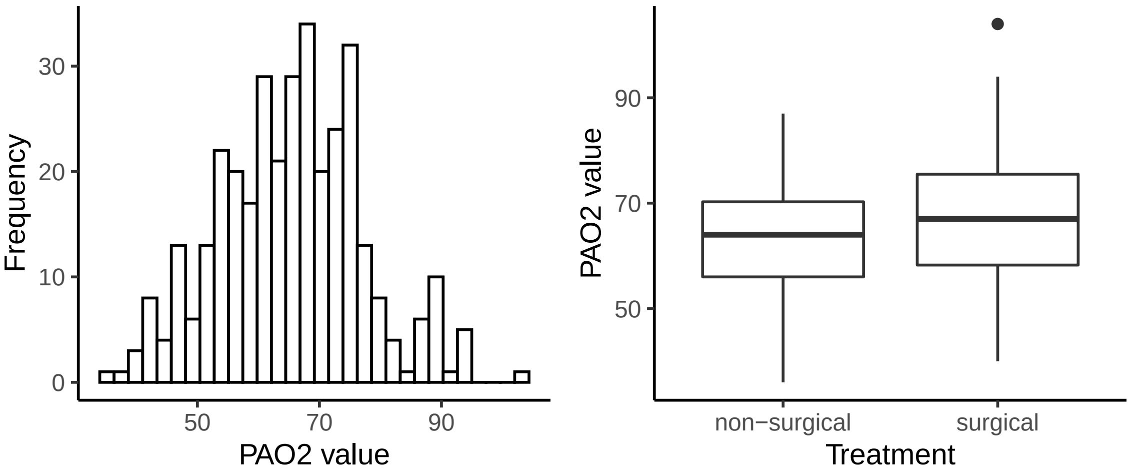

Figure 2: Histogram of PAO2 (left panel) and PAO2 boxplot as a function of treatment (right panel).

From Figure 2 (left panel) we can see that PAO2 follows a normal distribution, which was confirmed by a Shapiro-Wilk normality test (W = 0.992, p = 0.079). The right panel shows that the distribution is similar between treatments, but the patients undergoing surgery show higher levels of arterial oxygenation (i.e., a sign of lung function improvement).

2.2 The Arterial Oxygen variable

Figure 3 shows the individual levels of PAO (lighter lines), as well as the average mean (darker line).

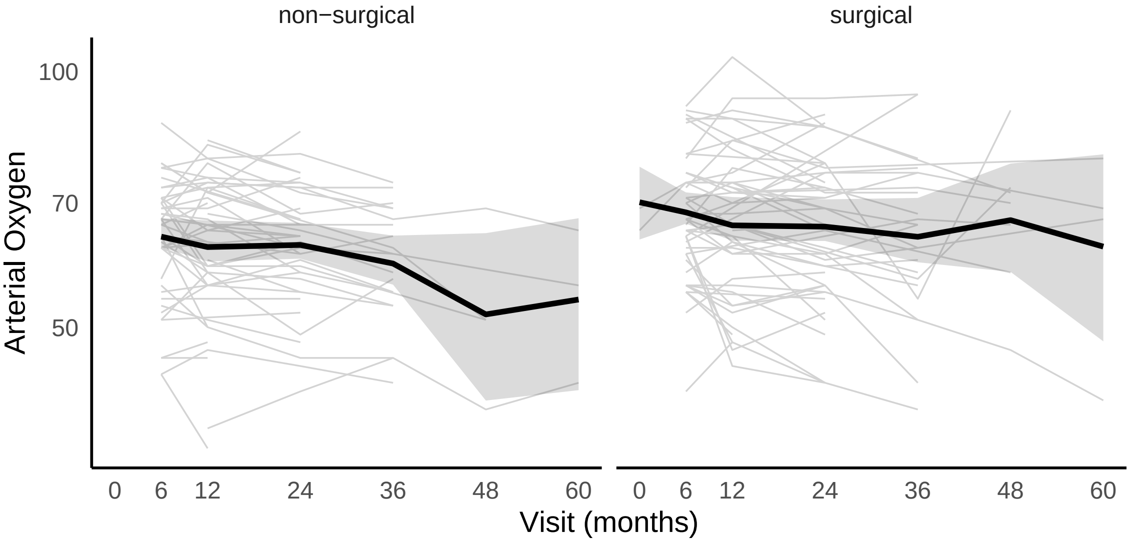

Figure 3: Arterial Oxygen (PAO) leves for each individual at each visit. The darker line represents the mean. The grey shadow shows the 95% confidence interval for the mean.

As expected, Figure 3 shows slightly larger values of arterial oxygen for the surgical group in comparison with the non-surgical group. Then, at the beginning and after 36 weeks, several patients stopped being followed. However, for those that were not assessed after 36 weeks, there is no clear pattern in the results: some seem to be decreasing, others seem to be increasing and others seem stable. Thus, it is reasonable to assume that missing data is not associated with a tendency in levels of arterial oxygen. After 36 weeks there are few observations, which is reflected on the confidence interval that increases in range. Also, there might be a tendency for a decrease in arterial oxygen over time, especially in the non-surgical treatment.

3. Linear Regression assuming independent errors

Because patients were repeatedly measured over time, it is reasonable to assume that their lung function at a given time t is correlated with the lung function at time t+k.

Figure 4 shows the predicted values as a function of observed values assuming independent errors.

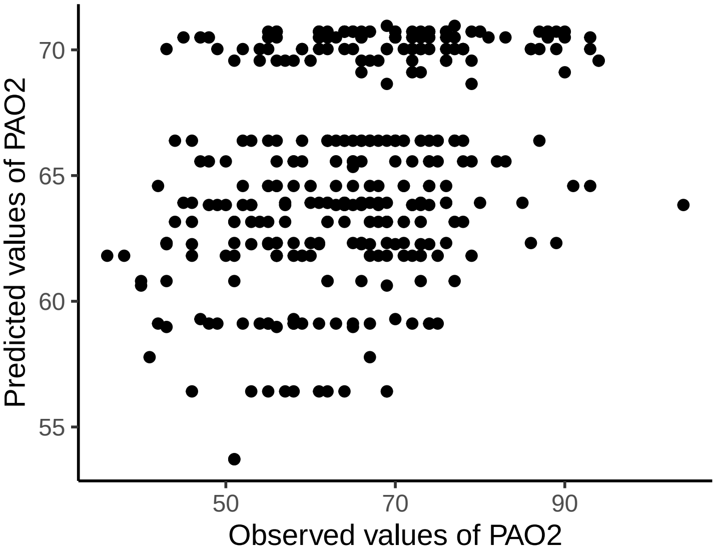

Figure 4: Predicted values as a function of observed values assuming independent errors.

Figure 4 shows that if a linear model with independent errors is fitted to the data, the predicted values do not clearly fit the observed values. Note that if prediction was perfect all these values should fall in the Y=X line.

To explore this possible correlation, we first modeled the data assuming independent errors, i.e., assuming that the lung function does not correlate across time. For this, we tested a saturated model, that is, the most complex model to capture the effect of the covariates, and then analyzed the residuals. For this and later analysis, we modeled PAO with the covariates gender, MEDID, and VISIT (we did not include race because almost all patients were white). The saturated model included the main effects of gender, MEDID, and VISIT, and all the 2x2 interactions: gender x MEDID, gender x VISIT, and MEDID x VISIT.

First we compared the model considering time as a continuous variable with another model considering time as a categorical variable. An ANOVA comparing the two models showed that the two models are not statistically different at a 0.05 significance level (F = 0.3133, p = 0.9922), so we chose the simplest one, with time as a continuous variable.

3.1 Naive vs. Robust Estimators

After modeling the PAO2 data with time as a continuous variable, one can analyze the residuals, that is, the difference between the observed values and the predicted ones by the model. Because the model does not account for correlation across time, if there is a temporal correlation present in the data it should be visible in the correlation structure of residuals (i.e., in the unexplained variability). The model estimators obtained assuming error independence are called naive, because they are "blind" to a possible correlation structure. On the other hand, if we compute the empirical correlation between residuals at different times and use that correlation in the model, we obtain robust estimators.

Table 2: Naive and robust estimators.

Coefficient	Estimate	Naive SE	Naive t	Naive p	Robust SE	Robust t	Robust p
$\beta_{intercept}$	64.508	2.078	31.047	0	2.441	26.427	0
β_{VISIT}	-0.225	0.092	-2.455	0.015	0.072	-3.109	0.002
$\beta_{MEDID=surgical}$	0.835	2.522	0.331	0.741	3.591	0.233	0.816
$\beta_{gender=male}$	2.701	2.416	1.118	0.264	2.942	0.918	0.359
$\beta_{VISIT*MEDID=surgical}$	0.099	0.092	1.069	0.286	0.087	1.131	0.259
$\beta_{VISIT*gender=male}$	0.088	0.094	0.927	0.354	0.089	0.986	0.325
$\beta_{MEDID=surgical*gender=male}$	2.911	2.447	1.19	0.235	4.112	0.708	0.479

If there was no temporal correlation between measures, using the empirical correlations or assuming they are null should yield the same results. Table 2 shows that the standard errors of the estimators are different between naive and robust estimators. This has an impact on the t statistic and consequently on the p-value of testing the null hypothesis of the estimator being equal to zero. In this particular case, assuming a significance level of 5% testing each estimator yielded the same result with naive and robust estimations. In any case, the different values obtained show the importance of exploring the temporal correlation in the data.

4. Longitudinal Data Analysis

4.1 Variogram

To explore the temporal correlation, one might analyze the variogram of the residuals of the saturated model. The variogram shows how residuals correlate as a function of their temporal distance, u. Note that the higher the value in the variogram, the smalller the correlation between residuals with time distance of u. Figure 5 shows the empirical variogram.

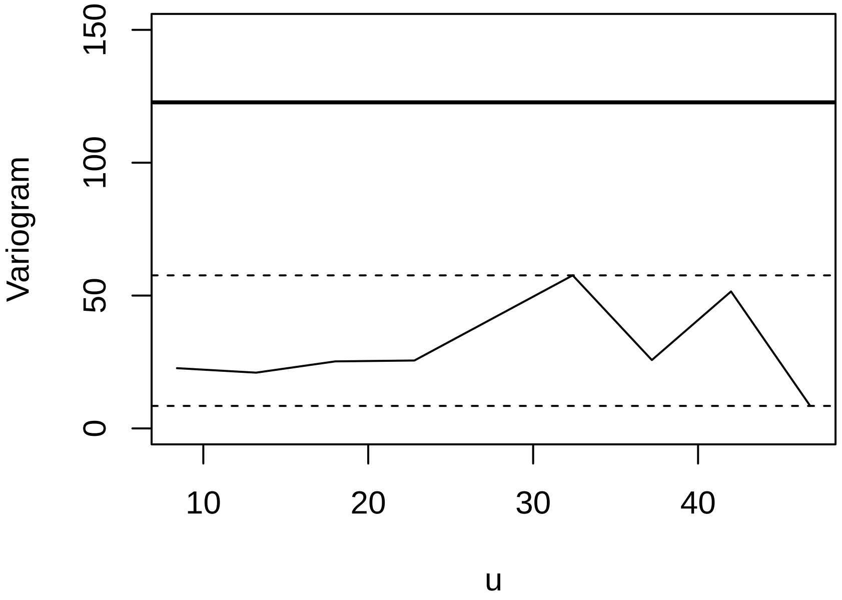

Figure 5: Empirical variogram

The variogram provides insightful information regarding the correlation structure of data. First, it is important to note that there are just a few observations for 48 and 60 times, the correlations for larger intervals of time, u, are based in just a few pairs of data. Then, this variogram should only be analyzed until approximately u = 32.

From the empirical variogram it is possible to explore the various sources of variability in the data: between-subject variability (from the continuous line at v(u)=122.72 to the dashed line approximately at V(u)=60), within-subject variability (from the dashed line at approximately V(u)=60 to the dashed line approximately at V(u)=15), and unexplained variability (from zero to the the dashed line approximately at V(u)=15). From the variogram it is clear that the major source of variability is the between-subject variability, followed by the within-subject variability. Also, as the time distance increases, the correlation between residuals decrease (i.e., V(u) increases). Thus, even when a saturated linear model is fitted to the data, the residuals show temporal correlation. The shape of the variogram (similar to an S-shape) suggests that a gaussian correlation may be present.

4.2 Correlation Structure

To account for the temporal correlations, we can use the gls() or the lme() functions in R. The former method does not discriminate between the various sources of variability (i.e., it fits linear models using generalized least squares), whereas the later does (i.e., it fits linear mixed-effects models). For instance, the lme method allows to distinguish the three sources of variability in the data described in the variogram.

Based on the previous exploratory analysis, we expected a better fit with a model with a gaussian temporal correlation and accounting for random effects (between-subject variability). In any case, we fitted several models, with different correlation structures and components of variability.

Table 3 shows the main results of fitting different saturated models.

Table 3: AIC (Akaike information criterion) and Logliklihood values for each fitted model

Model	Method	AIC	LogLik	df
Independent errors	lm	2659.222	-1321.611	8
First-order Autoregressive (AR1)	gls	2661.222	-1321.611	9
Exponential Correlation	gls	2424.485	-1203.242	9
Gaussian Correlation	gls	2497.653	-1239.827	9
Compound Symmetry	gls	2417.142	-1199.571	9
Random Intercept and Slope	lme	2413.738	-1195.869	11
Random Intercept with Exponential Correlation	lme	2412.211	-1195.105	11
Random Intercept with Gaussian Correlation	lme	2410.664	-1194.332	11

Table 3 shows that the models with independent errors and the model assuming a First-order Autoregressive (AR1) correlation are the ones with higher AIC and lower log likelihood, thus they are not appropriate to model this data. This results strengthens the hypothesis that temporal correlations are present in this data.

Using the gls method, the model with lower AIC and higher log likelihood is the one assuming a compound symmetry correlation structure. This model assumes a constant correlation across time, and it corresponds to a model with a random slope. Thus, even though this model simplifies the correlation structure by setting it to a constant, it captures the between-subject variability that was present in the variogram (see Figure 5).

Then, using the lme method that allows to distinguish the sources of variability improved the model fitting. From those, the one with lower AIC and higher log likelihood is the one with a random intercept with a gaussian correlation. This was expected because when exploring the data we saw that it was important to account for the between-subject variability and that the correlation structure seemed gaussian.

4.3 Fitting the model

By comparing the several models, we have selected the one with a random intercept with a gaussian correlation as the most appropriate one. Then, to select the fixed effects we implemented a step wise method. First, we started with the saturated model, with the three main factors and their 2x2 interactions. Then, step by step, we removed each factor (or interaction) based on non-significant p-value of the corresponding parameter (using a 0.05 significance level). All interactions were removed and the final model was:

$$Y_{ij} = \beta_0 + \beta_1 * VISIT_{ij} + \beta_2 * MEDID_{=surgical} ij + \beta_3 * gender_{=male} ij + W_i(t_{ij}) + Z_{ij}$$

Table 4 presents the obtained results for the final model.

Table 4: Statistics for the selected model

Coefficients	Value	SE	df	t-value	p-value
$\beta_{intercept}$	61.656	1.823	225	33.825	0
β_{VISIT}	-0.125	0.031	225	-4.054	0
$\beta_{gender=male}$	5.77	1.921	117	3.003	0.003
$\beta_{MEDID=surgical}$	4.596	1.867	117	2.462	0.015
τ^2	21.601				
σ^2	29.1				
ν^2	73.004				
ϕ	27.671				

From Table 4 we can see that, everything else being equal, (1) arterial oxygenation decreases 0.125 units per week, (2) males have a higher oxygenation than women, and (3) the surgical group have better oxygenation than the non-surgical group. In other words, being men and having surgery are good predictors of lung function measured by arterial oxygenation, but that may decrease with time.

Congruent with the variogram result, the larger portion of variability comes from random effects ($\nu^2 = 73.004$).The proportion of unexplained variability with this model is about 17% ($\tau^2/(\tau^2 + \sigma^2 + \nu^2)$). Also, the gaussian parameter, ϕ is approximately 28, meaning that from a temporal distance of 28 weeks or longer, the temporal correlation is almost null.

4.4 Model Diagnostics

The selected model is the best one available. However, it may be still not good enough. To evaluate the quality of the model, we analyzed the residuals. In good-fitting models the residuals should (a) have an approximately normal distribution, (b) be centered at zero, (c) have constant variance (homoscedasticity), and (d) be independent. Figure 6 shows the fitted values as a function of observed values (left upper panel), the histogram of residuals (right upper panel), and the residuals as a function of fitted values (left bottom panel).

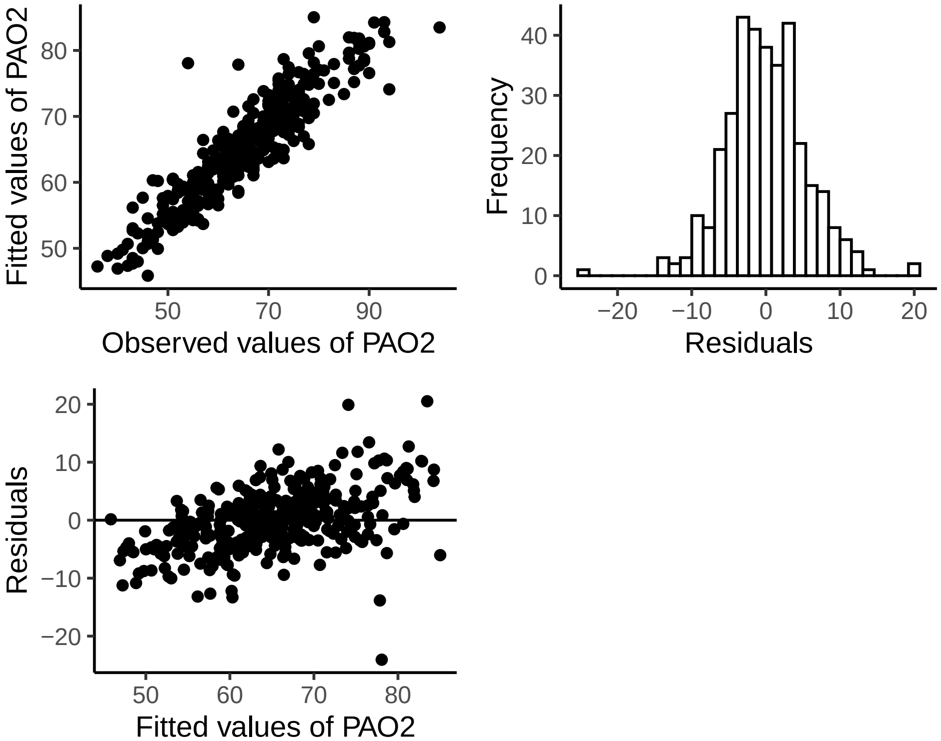

Figure 6: Fitted values as a function of observed values (left upper panel), histogram of residuals (right upper panel), and residuals as a function of fitted values (left bottom panel).

From the left upper panel of Figure 6 we can see that the fitted values increase linearly with the observed values, as expected. The right upper panel shows that the residuals follow a normal distribution around zero. Then, the bottom left panel shows that the residuals' variance seems constant, but there is a tendency for the model to underestimate at smaller PAO2 values and overestimate at larger PAO2 values. The failure to meet this assumption may mean that the structure of the selected model does not account for all the variability. A different structure or even a non-linear model may be necessary (not within the scope of this study). Compared with the other models, this still seems to be the one that accounts for more variability. Therefore, we keep this model taking into account that the results may be biased for the non-assumption of independence of errors.

5. Conclusions

The main objective of the present study was to determine whether surgery improved lung function in patients with lung emphysema. Because patients were followed across time, the data assessing lung function was correlated across time, and simple methods assuming independence were not viable. After accounting for temporal correlations, the selected model showed that in fact, surgery improved lung capacity. However, these results have to take into account some limitations.

First, there were several missing data, specially at the beginning and at weeks 48 and 60. Thus, in practical terms the model was highly influenced the values obtained in weeks 6, 12, 24 and 36. For a better model more data should be obtained at later times.

Second, it would be important to ascertain that at time = 0 both groups were similar in terms of lung function. In the present data set only 3 participants were assessed at baseline, and they were all from the surgical group. Therefore, it was not possible to test for differences in baseline, which, if present, could bias the conclusions.

Third, other variables (e.g., age) could be important to explain the between-subject variability. It could be important to add more covariates in the model.

Fourth, the reasons for missing data were not explicit, and in case of dropouts associated with the lung function, other techniques should be implemented (survival and longitudinal joint models).

Fifth, another design feature that could ideally be implemented was to have a group with only the surgical treatment (without physical therapy). In the present study, both groups received physical therapy and the surgical group received an additional treatment (surgery). To truly assess the effect of surgery it could be interesting to have this variable isolated. However, clinically this may not be possible because therapy may be necessary for a good recovery from surgery.

The data analysis also presented limitations. Mainly, as reported above the independence of residuals was not met, which could influence predictions and inference about the results. Probably, if the times when less data is available were eliminated from the analysis, better results would be obtained. However, the model would be less general and the results probably less clinically interesting. A better option would be to increase the data set at all different times.

At last, other variables assessing lung function should be studied, or a combination of variables could be used as a general metric of lung function. The results reported here only focused on the arterial oxygen, but other variables can be more associated with lung function and show a more clear effect of surgery.

6. References

[1] Thurlbeck, W. M., & Müller, N. L. (1994). Emphysema: definition, imaging, and quantification. AJR. American journal of roentgenology, 163(5), 1017-1025.

[2] Trial, G. T. N. E. T. (1999). Rationale and design of The National Emphysema Treatment Trial: a prospective randomized trial of lung volume reduction surgery. Chest, 116(6), 1750-1761.

[3] National Emphysema Treatment Trial Research Group. (2003). A randomized trial comparing lung-volume–reduction surgery with medical therapy for severe emphysema. New England Journal of Medicine, 348(21), 2059-2073.

[4] Lim, E., Ali, A., Cartwright, N., Sousa, I., Chetwynd, A., Polkey, M., . . . & Goldstraw, P. (2006). Effect and duration of lung volume reduction surgery: mid-term results of the Brompton trial. The Thoracic and cardiovascular surgeon, 54(03), 188-192.