Report - Classical and modern methods of data analysis

Reynaldi Ikhsan Kosasih & Sara Roose

Master of Epidemiology - University of Antwerp

Introduction

In this report, we provide answers to the questions of the assignment about the dataset on hospital length of stay following the COVID-19 infection in Belgium. We describe the results of performed statistical analyses and interpret and discuss our findings. For all computed analyses, we used R Studio (R Foundation for Statistical Computing, version 4.0.5).

The dataset we worked with (COVID19 LoS 14042021.txt) contains information on hospital length of stay together with individual-specific characteristics collected in 12 different Belgian hospitals during the ongoing COVID-19 pandemic. Variables collected are patient ID, hospital ID, length of stay at the hospital, age of patients, gender of patients, cycle threshold (Ct) value at hospital admission, COVID-19 wave, and hospital mortality.

Question 1. Belgian COVID-19 hospital data

A. Explore the variables included in the Belgian COVID-19 hospital dataset

1) Patient identification number (patient_id)
There are 9984 unique patient numbers in the dataset.

2) Hospital identification number (hosp_id)
The dataset contains data collected in 12 different hospitals. The number of subjects is rather equal among the different hospitals with on average 835 patients/hospital. Therefore, all hospitals are equally represented in the dataset. This is as well the case within each of the three different waves.

3) Length of stay in the hospital (LoS) (days)
The variable length of stay is continuous. 191 subjecs have missing values for LoS. The mean is 8.412 days, median is 7.672, minimum is 1.288 days, and maximum 37.686 days. For female patients the average days of hospitalization was 7.62 days and for male patients 9.12 days.
The histogram of LoS showed that this variable is not normally distributed, as the distribution appears to have a right-skewed distribution with long right tail. A transformation to the log scale showed a normal distribution.

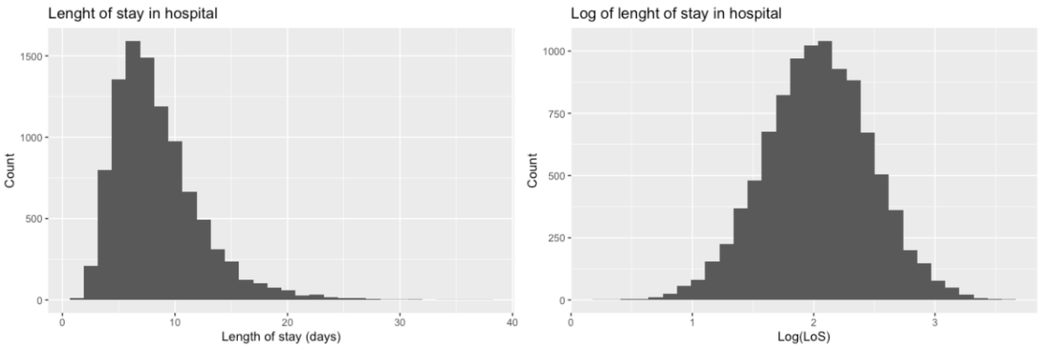


Fig 1. Distribution of the variable LoS (left) and its logarithmic transformation (right)

4) Age of the individual
The youngest hospitalized patient was 22 years and the oldest 100. The average age is 64.9, with 64.7 for female patients and 65.1 for male patients.

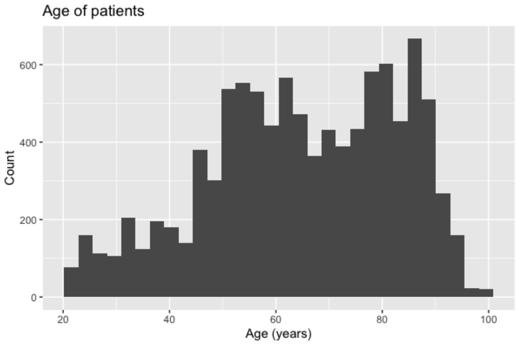


Fig 2. Distribution of the age of hospitalized patients in the dataset.

5) Gender of the individual (1: male, 0: female)

There are 5265 male and 4719 female patients registered.

6) Ct-value

Ct-values relate inversely with the amount of viral RNA in the sample, thus high viral load corresponds to low Ct-values. The minimum Ct value is 11.25 and maximum 42.90, on average patients have a Ct value of 25.49. 201 patients have missing values for this variable.

7) COVID-19 wave

There were more patients registered in the first (3600) and second (4860) wave than in the third (1524) wave. The second wave accounts for 48.68% of the total observations.

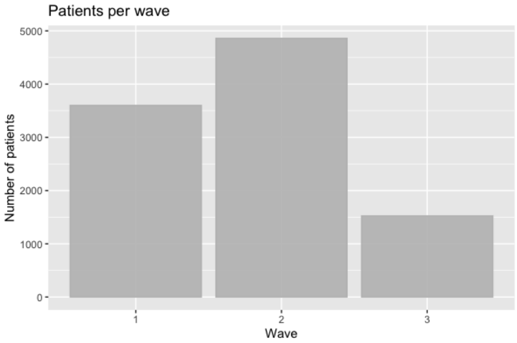


Fig 3. Number of observations (patients) per wave in the complete dataset.

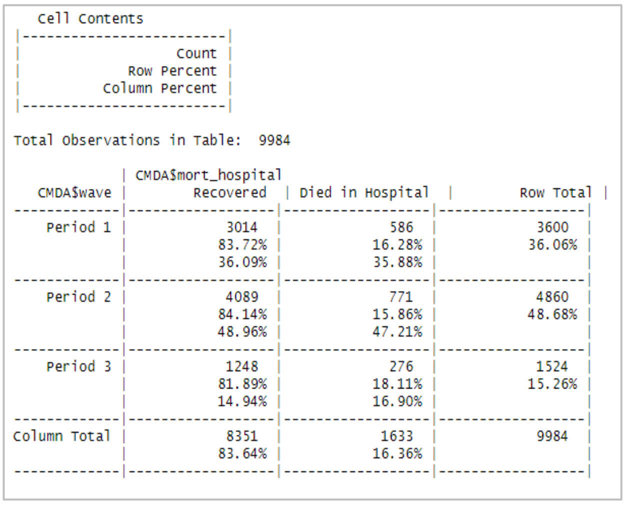
8) Hospital mortality indicator (1: died in hospital, 0: recovered)

The overall mortality is 16.4%, 1633 of the 9984 patients died in the hospital. Mortality ranged from 15.9 % (second wave) to 18.1 % (third wave). The mortality was slightly less for female (14.1%) than for male (18.4%) patients.

B. Using a contingency table analysis, compare the death rates in hospital across the different periods (waves)

From the total of 9984 observations, we found that 1633 (16.36%) patients died in the hospital. We identified that the third wave had the highest mortality rate (18.11%) compared to the first wave (16.28%) and the second wave (15.86%). The highest absolute number of deaths (771) occurred during the second wave of COVID-19 in Belgium.

We used a chi-square test to analyze the contingency table formed by the two categorical variables mortality and wave. We defined a X-squared of 4.30 and thus a non-significant p-value of 0.1163. We therefore concluded that the death rates across the different periods are not significantly different.



```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|          Column Percent |
|-------------------------|

Total Observations in Table:  9984

             | CMDA$mort_hospital
   CMDA$wave |          Recovered  | Died in Hospital  |          Row Total |
-------------|-------------------|-------------------|-------------------|
    Period 1 |              3014 |               586 |              3600 |
             |            83.72% |            16.28% |            36.06% |
             |            36.09% |            35.88% |                   |
-------------|-------------------|-------------------|-------------------|
    Period 2 |              4089 |               771 |              4860 |
             |            84.14% |            15.86% |            48.68% |
             |            48.96% |            47.21% |                   |
-------------|-------------------|-------------------|-------------------|
    Period 3 |              1248 |               276 |              1524 |
             |            81.89% |            18.11% |            15.26% |
             |            14.94% |            16.90% |                   |
-------------|-------------------|-------------------|-------------------|
Column Total |              8351 |              1633 |              9984 |
             |            83.64% |            16.36% |                   |
-------------|-------------------|-------------------|-------------------|
```

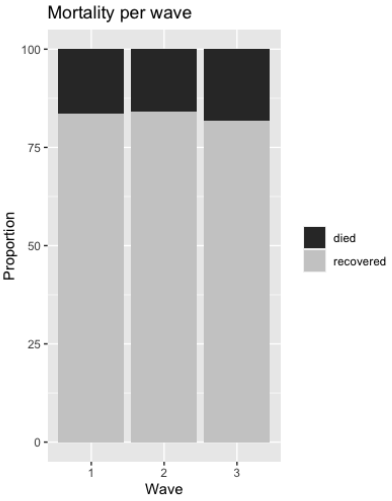
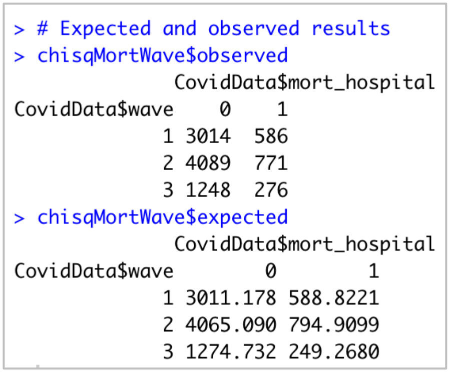


Fig 4. 2x3 contingency table for mortality and wave (left), proportion of deaths per wave (right)

We also did some exploration of the observed and expected contingency tables and found that the values are indeed very close to each other.



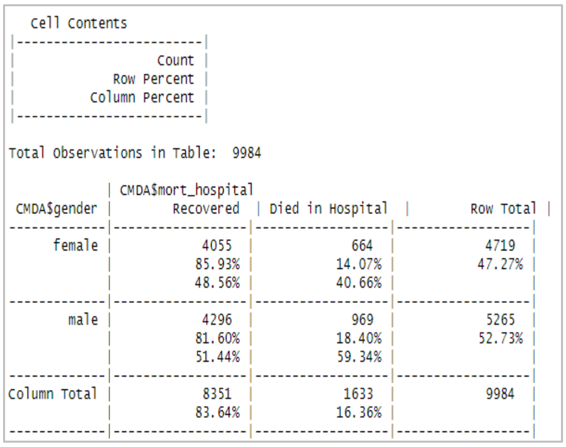
```
> # Expected and observed results
> chisqMortWave$observed
               CovidData$mort_hospital
CovidData$wave    0    1
             1 3014  586
             2 4089  771
             3 1248  276
> chisqMortWave$expected
               CovidData$mort_hospital
CovidData$wave        0        1
             1 3011.178 588.8221
             2 4065.090 794.9099
             3 1274.732 249.2680
```

Fig 5. Expected versus observed contingency tables

C. Compare the probabilities of dying in the hospital for males and females. Is there a significant difference in mortality between these two groups?

The proportion of male patients who died in the hospital (18.40%) is higher compared to female patients (14.07%).

We used a chi-square test to analyze the contingency table formed by the two categorical variables mortality and gender. We defined a X-squared of 33.847 and a significant p-value of 5.067098e-9. We therefore concluded that the mortality between the two genders is significantly different.



```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|          Column Percent |
|-------------------------|

Total Observations in Table:  9984

             | CMDA$mort_hospital
 CMDA$gender |          Recovered  | Died in Hospital  |          Row Total |
-------------|-------------------|-------------------|-------------------|
      female |              4055 |               664 |              4719 |
             |            85.93% |            14.07% |            47.27% |
             |            48.56% |            40.66% |                   |
-------------|-------------------|-------------------|-------------------|
        male |              4296 |               969 |              5265 |
             |            81.60% |            18.40% |            52.73% |
             |            51.44% |            59.34% |                   |
-------------|-------------------|-------------------|-------------------|
Column Total |              8351 |              1633 |              9984 |
             |            83.64% |            16.36% |                   |
-------------|-------------------|-------------------|-------------------|
```

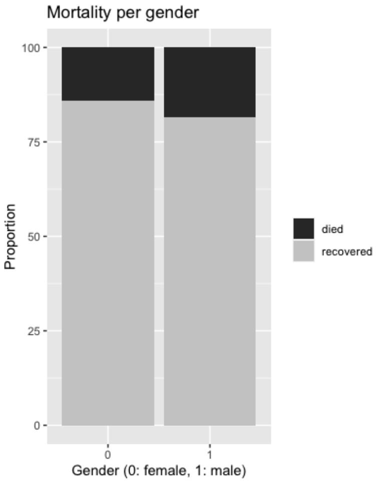
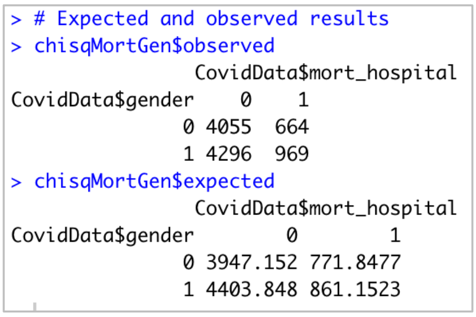


Fig 6. 2x2 contingency table for mortality and gender (left), proportion of deaths per gender (right)

In this case, the expected and observed values in the contingency tables are different.



```
> # Expected and observed results
> chisqMortGen$observed
                CovidData$mort_hospital
CovidData$gender    0    1
               0 4055  664
               1 4296  969
> chisqMortGen$expected
                CovidData$mort_hospital
CovidData$gender        0        1
               0 3947.152 771.8477
               1 4403.848 861.1523
```

Fig 7. Expected versus observed contingency tables

Since the mortality between the two genders is significantly different, we calculated two measures of association that give an indication about the size of the difference; i.e. the risk ratio and the odds ratio. The risk ratio male to female is 1.31, male patients have thus 31% more chance to die in the hospital than female patients (confidence interval RR: 1.19 to 1.43). We found that the odds ratio of male to female is 1.37. This means that the odds of dying in the hospital due to COVID-19 among male patients is 1.37 times the odds of female patients.

D. Compare the length of stay in the hospital, for recovered individuals, between males and females, and across periods

Our outcome/dependent variable (length of stay in the hospital) is continuous. We were asked to compare this outcome variable between the levels of two independent variables, which are gender (male and females) and COVID-19 wave (1, 2 and 3). For these analyses, there are parametric and non-parametric tests available (Table 1). We decided to perform all possible statistical analyses and therefore use both non-transformed and log-transformed outcome data.

Table 1. Possible statistical analyses to answer the research question stated in question 1D.

Independent Variable	**Dependent Variable**	**Possible Statistical Analyses**
Gender (Factor with 2 levels: male and female)	Length of stay (continuous variable)	Parametric Test: Unpaired t-test Non-parametric Test: unpaired two-samples Wilcoxon test
Wave of COVID-19 (Factor with 3 levels: wave 1, wave 2, wave 3)	Length of stay (continuous variable)	Parametric Test: One-way ANOVA Non-parametric Test: Kruskal-Wallis test

<u>Gender and length of stay in hospital</u>

By using descriptive statistics, we identified that recovered male patients have a higher mean (8.86 days) and median (8.09 days) of length of stay in the hospital than female patients (mean 7.39, median 6.73). We therefore want to know whether these differences are statistically significant.

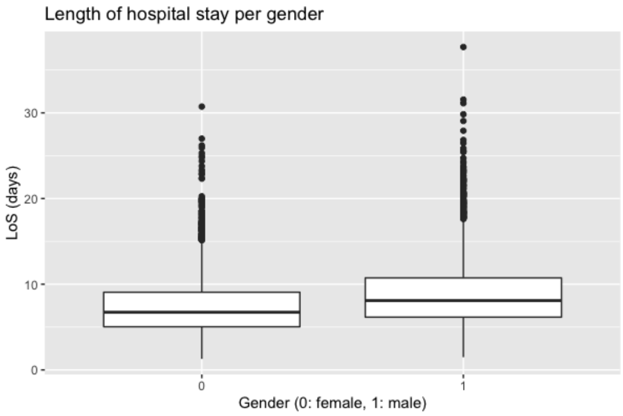


Fig 4. Boxplot of length of stay in the hospital for female and male patients.

Option 1. Using a non-parametric test

First, we checked whether the outcome data length of stay in the hospital is normally distributed using a normal probability plot (Fig 5) and concluded that our variable is not normally distributed. The Shapiro-Wilk normality test is as well significant for both males and females (p-value for both tests < 2.2e-16). Therefore, we conclude that the normality assumption is violated.

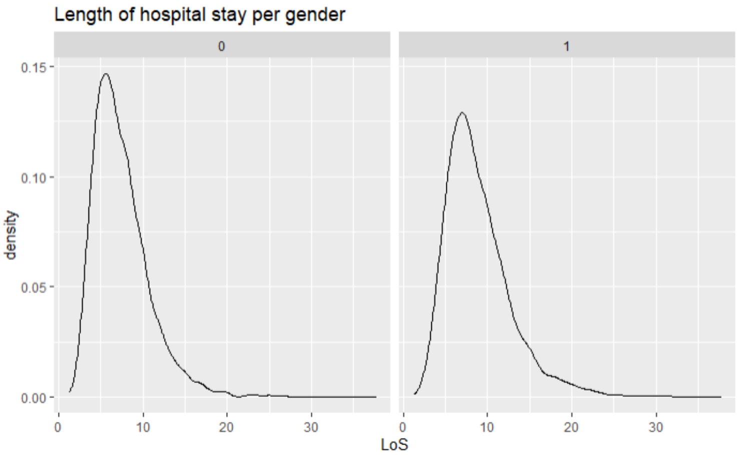
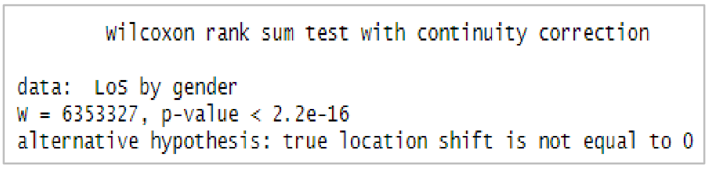


Fig 5. Density plot of length of hospital stay per gender (0: female, 1: male). The normality assumption is violated.

We ran the F-test to check homogeneity in variances between male and female patients. We found a significant p-value of < 2.2e-16. Therefore, we reject the null hypothesis and conclude that there is a significant difference between the two variances. The assumption of equal variance is also violated.

Since both the normality assumption and the assumption of equal variance do not hold for the non-transformed length of stay data, it is not appropriate to run a parametric test. Therefore, we performed a non-parametric test; i.e. the unpaired two-samples Wilcoxon test.
The p-value of this test is < 2.2e-16, which is less than 0.05 (α). We concluded that there is a significant difference in length of stay between male and female patients.



```
	Wilcoxon rank sum test with continuity correction

data:  LoS by gender
W = 6353327, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Fig 6. R output of the two-samples Wilcoxon test with continuity correction

Option 2. Using a parametric test after data transformation

We could also opt to transform the data using a logarithmic transformation and see whether the assumptions for a parametric test are now valid.
After transforming the data, Shapiro-Wilk p-values are 0.1047 (male) and 0.8797 (female), both tests are not significant. We concluded that the transformed data is normally distributed. The F-test to compare two variances (alternative hypothesis: true ratio of variances is not equal to 1) also gives a non-significant p-value.

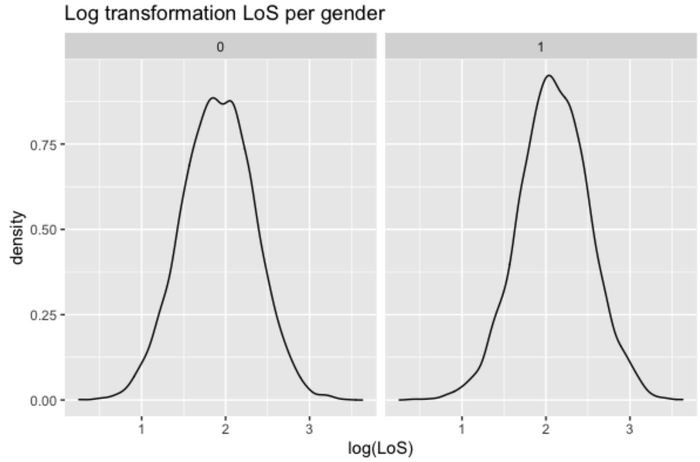
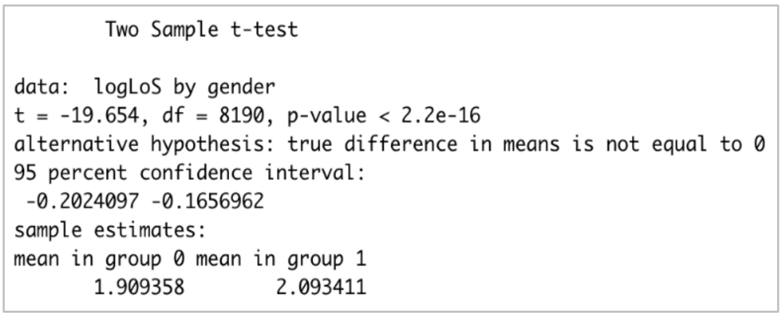


Fig 7. Density plot of the log-transformed length of hospital stay per gender (0: female, 1: male). The data is normally distributed.

Since the transformed data is not violating the assumptions for a parametric test, we can run a two sample unpaired t-test. The results showed that there is a significant difference in LoS between the two genders.



```
        Two Sample t-test

data:  logLoS by gender
t = -19.654, df = 8190, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2024097 -0.1656962
sample estimates:
mean in group 0 mean in group 1
       1.909358        2.093411
```

Fig 8. R output of the two sample unpaired t-test

Wave and length of stay in hospital

Following the same logic of previous analysis on gender and length of stay in hospital, we start by looking at the descriptive statistics. We can see that people stayed longer in the hospital during the third period of COVID-19. The mean LoS is 8.40 days in wave 1, 7.37 days in wave 2 and 10.08 in wave 3. Is this difference statistically significant?

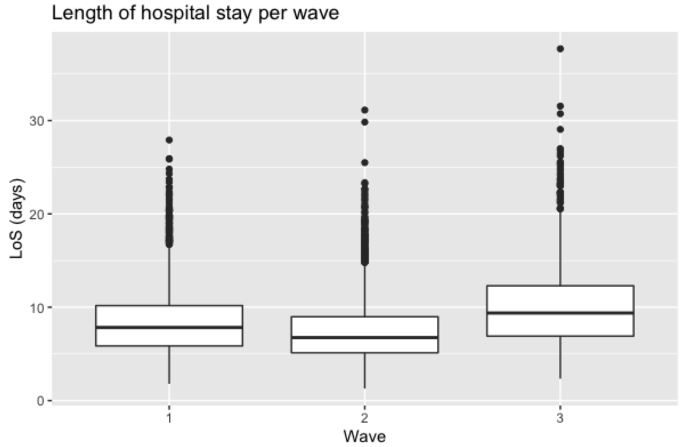


Fig 9. Boxplot of length of stay in the hospital across the three waves.

Option 1. Using a non-parametric test

In a second step, we evaluated the assumptions needed to perform parametric analyses (homogeneity of variances and normality). The results are as follows:

The residuals versus fitted plot showed a sign of violation to homogeneity of variances (Fig 10). To confirm this, we ran the Levene's test that gave us a p-value smaller then 2.2e-16. This means that the first assumption of homogeneity of variances is violated.

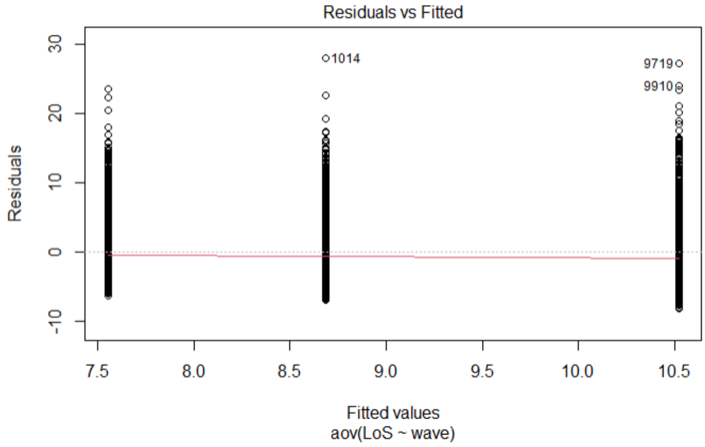


Fig 10. Test for homogeneity of variances using residuals vs fitted plot

We then ran the normality plot of residuals (Fig 11) to check the assumption of normality. In the plot below, the quantiles of the residuals are plotted against the quantiles of the normal distribution. A 45-degree reference line is also plotted. The result clearly showed that the assumption of normality is violated.

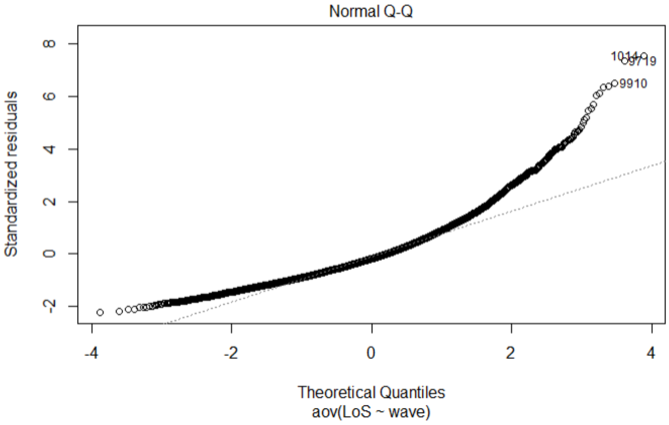
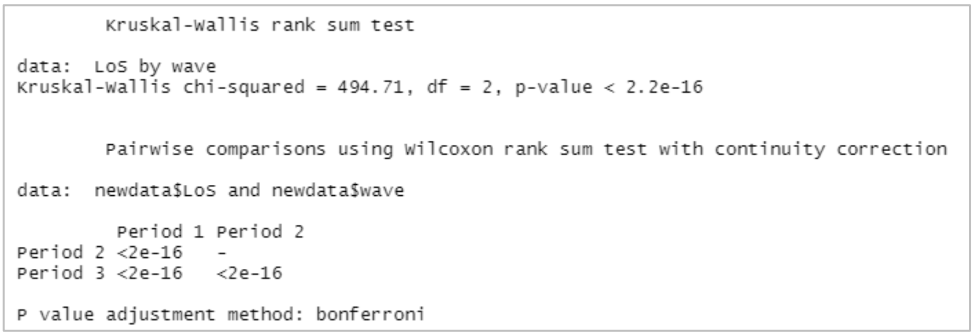


Fig 11. Test for normality using the normality plot of residuals

Since both the normality assumption and the assumption of equal variance do not hold for the non-transformed length of stay data, it is not appropriate to run a parametric test. Therefore, we performed the non-parametric Kruskal-Wallis test, to assess the difference in means between the three COVID-19 waves. We also used a test of multiple pairwise-comparison between the groups to identify which pair of groups were different from each other.



```
        Kruskal-Wallis rank sum test

data:  LoS by wave
Kruskal-Wallis chi-squared = 494.71, df = 2, p-value < 2.2e-16


        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  newdata$LoS and newdata$wave

         Period 1 Period 2
Period 2 <2e-16   -
Period 3 <2e-16   <2e-16

P value adjustment method: bonferroni
```

Fig 12. R output of the Kruskal-Wallis rank rum test and pairwise comparisons test

The Kruskal-Wallis test gave us a p-value of <2.2e-16. This means that the differences observed are statistically significant. Furthermore, the multiple pairwise-comparison test told us that all groups are significantly different from each other.

Option 2. Using a parametric test after data transformation

We could also opt to transform the data using log transformation and see whether the assumptions for a one-way analysis of variance (ANOVA) are now valid. This test is an extension of independent two-samples t-test for comparing means in a situation where there are more than two groups (in this case three waves).

We have three possibilities to check the homogeneity of variance: (1) residuals versus fits plot (Fig. 13), (2) Levene's test for homogeneity of variance and (3) a rule of thumb. All three possibilities give us the same results: our log-transformed data does not violate the homogeneity of variance. The P-value of the Levene's test is 0.5134 and thus this test is not significant. The rule of thumb defines that the variances cannot differ more than a factor of five, which is indeed not the case: the logLoS variances are 0.172 (wave 1), 0.177 (wave 2) and 0.183 (wave 3).

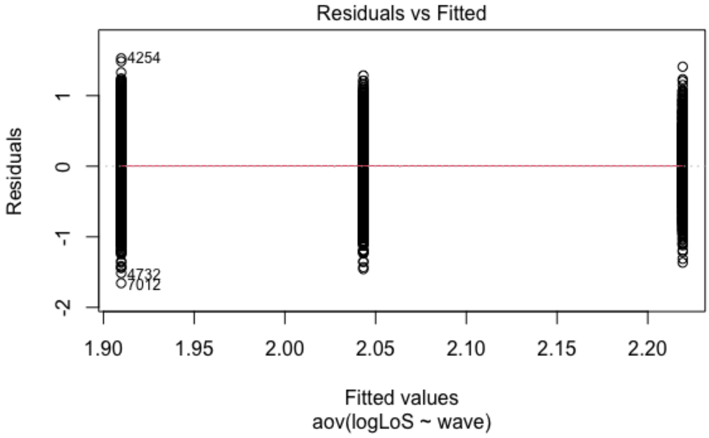


Fig 13. Test for homogeneity of variances using residuals vs fitted plot (transformed data)

We run the normality plot of residuals to test for normality (Fig 14). The transformed data is normally distributed.

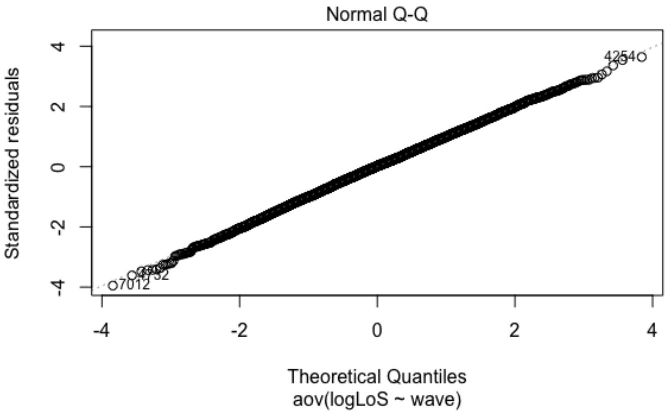
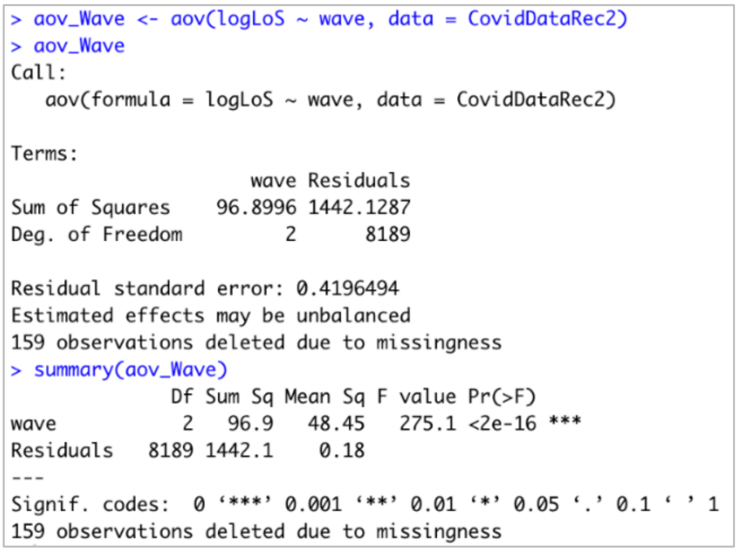


Fig 14. Test for normality using the normality plot of residuals (transformed data)

It is safe to use a parametric one-way ANOVA since for the log-transformed data every assumption holds. Results of the one-way ANOVA showed that there is a significant difference in LoS between the three waves. Thereafter, we performed multiple pairwise-comparisons. A Tukey (Honest Significant Differences) test showed that all three waves are significantly different from each other (adjusted p-value: 0).



```
> aov_Wave <- aov(logLoS ~ wave, data = CovidDataRec2)
> aov_Wave
Call:
   aov(formula = logLoS ~ wave, data = CovidDataRec2)

Terms:
                    wave Residuals
Sum of Squares   96.8996 1442.1287
Deg. of Freedom        2      8189

Residual standard error: 0.4196494
Estimated effects may be unbalanced
159 observations deleted due to missingness
> summary(aov_Wave)
              Df Sum Sq Mean Sq F value Pr(>F)
wave           2   96.9   48.45   275.1 <2e-16 ***
Residuals   8189 1442.1    0.18
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
159 observations deleted due to missingness
```

Fig 15. R output of the one-way ANOVA

E. Formulate and fit an appropriate model to describe the relation between the length of stay in the hospital and other variables in the dataset

We decided to only include information from recovered patients, a total number of 8351 observations, with the purpose to avoid bias due to underlying health conditions and comorbidities that may played a role in patient's mortality and thus have a reflection in the length of stay of the patient.

We considered the dependent variable (length of stay (LoS) in the hospital) as a continuous variable. The original dataset had length of stay in the hospital as a continuous variable, so we proceed by keeping the integrity of the data without converting the dependent variable to a discrete variable. It might however be possible to deal with this research question by using the length of stay as count data (after discretizing) and therefore model a Poisson regression.

For this model, we decided to fit a multiple linear regression model to describe the relation between the length of stay in the hospital and other variables in the dataset. As discussed in previous questions of this report, the normality assumption for LoS is violated and therefore we use a logarithmic transformation of the data (logLoS) to fit the assumptions of the linear regression model (linear relationship, homoscedascity, equal variances, and normal distribution of variables)

Linear regression model

We used a forward selection method to decide which independent variables to include in our model. R^2 was used to check if our model had a good fit. This measure represents the proportion of variance in the outcome variable that may be predicted by knowing the value of the x variables.

A simple linear regression with variable age in the model showed that the relationship between LoS and age was significant, and this simple linear regression accounted for 17.46% of the data variability (Adjusted R-squared: 0.1745).

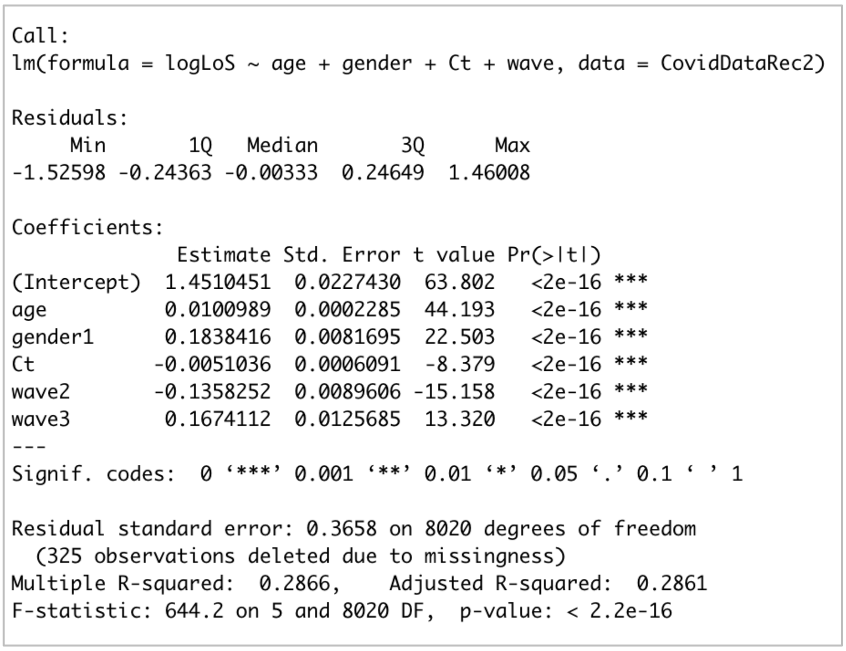
We then checked whether the relationship between on the one hand gender, Ct-value, wave and on the other hand LoS was also statistically significant. All three simple linear regression models gave p-values smaller than 0.01 for the relationships. The covariable gender explains 4.50% of the variability, the covariable wave accounts for 6.30% and the variable Ct only explains little extra variability (0.68%). We decided to take these three values, next to age, into account when building our model.

The multiple linear regression model showed that the relationships found between all four variables and LoS were still significant, this means that none of the significant results in the simple linear regressions was due to an effect of one of the other covariates. The adjusted R^2 for this model is 0.2861, and thus information about age, gender, Ct-value and waves can only explain 28.61% of the variability in LoS. In summary, even when including four possible covariates, the model still quite poorly fits the data.

Since gender and wave are factor data, R will use one of the levels as reference to compare the other levels to. We kept female gender and wave 1 as reference levels.

The estimated regression equation is:

LôS = 1.45 + 0.01 x age + 0.18 x gender1 – 0.005 x Ct – 0.14 x wave2 + 0.17 x wave3.

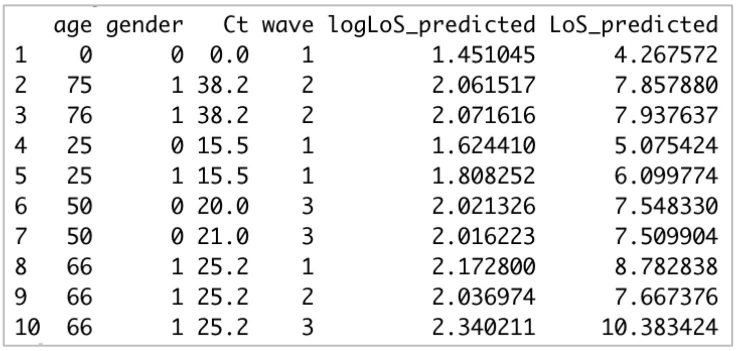


```
Call:
lm(formula = logLoS ~ age + gender + Ct + wave, data = CovidDataRec2)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.52598 -0.24363 -0.00333  0.24649  1.46008 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.4510451  0.0227430  63.802   <2e-16 ***
age          0.0100989  0.0002285  44.193   <2e-16 ***
gender1      0.1838416  0.0081695  22.503   <2e-16 ***
Ct          -0.0051036  0.0006091  -8.379   <2e-16 ***
wave2       -0.1358252  0.0089606 -15.158   <2e-16 ***
wave3        0.1674112  0.0125685  13.320   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3658 on 8020 degrees of freedom
  (325 observations deleted due to missingness)
Multiple R-squared:  0.2866,	Adjusted R-squared:  0.2861 
F-statistic: 644.2 on 5 and 8020 DF,  p-value: < 2.2e-16
```

Fig 16. R output of multiple linear regression to describe the length of stay in hospital (log-transformed data)

We investigated whether to include interactions between the four covariates. The interaction between age and wave 3 had a p-value of 0.01495. There is thus some evidence that the effect of wave 3 and age are not entirely independent of one another. However, the contribution of this interaction to a better fitting model was very low, the adjusted R^2 only changed from 0.2861 to 0.2865 (and the AIC from 6640 to 6637).

In a last step we did some predictions based on the multiple linear regression model without interaction to better understand the estimates calculated by our model. We built a data frame with the variables age, gender, Ct, and wave for ten fictious patients (fictive data) and predicted the length of stay in the hospital for each of these patients. We made changes only in one of the four covariates and investigated the change in response. These predictions illustrate the meaning of our intercept and estimates for the different covariates.



```
   age gender   Ct wave logLoS_predicted LoS_predicted
1    0      0  0.0    1         1.451045      4.267572
2   75      1 38.2    2         2.061517      7.857880
3   76      1 38.2    2         2.071616      7.937637
4   25      0 15.5    1         1.624410      5.075424
5   25      1 15.5    1         1.808252      6.099774
6   50      0 20.0    3         2.021326      7.548330
7   50      0 21.0    3         2.016223      7.509904
8   66      1 25.2    1         2.172800      8.782838
9   66      1 25.2    2         2.036974      7.667376
10  66      1 25.2    3         2.340211     10.383424
```

Fig 17. R output of predicted length of stay

We concluded that age, wave and gender are the most important covariates in our model. The variable Ct did not really contribute much to our model. However, the model still poorly fits the data ($R^2 = 0.2861$). We assumed that there are uncollected covariates that might explain more about the dependent variable LoS.

Model Diagnostics

We diagnosed our model using multiple plots to make sure that the assumptions for multiple linear regression model are not violated.

There are a few assumptions that we can evaluate with the residuals versus fitted plot (Fig 18). First, we observed that the vertical average of the residuals (red line) remains close to the dashed line, this means that the assumption of linear function of the predictors is not violated. The mean of residual value for every fitted value region is close to zero. To evaluate the homoscedascity assumption, we can see that the vertical spread of the residuals remains approximately constant across the x-axis and no systematic trends in the residuals. Therefore, the assumption of homoscedascity is not violated as well. We can also check if there are outliers; these extreme values will be further away from the rest and are indicated by their number.

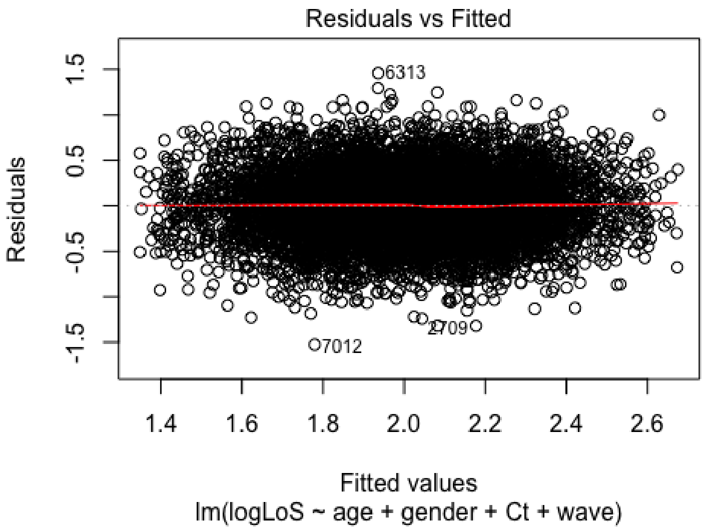


Fig 18. Residuals vs fitted plot of the multiple linear regression

As indicated in the figure 19, we can see that the residuals have an approximate normal distribution, as the histogram has the ideal bell-shaped appearance and the Q-Q plot appears to follow the straight line. This means that the assumption of normality is not violated in our model, built using log-transformed outcome variables (logLoS).

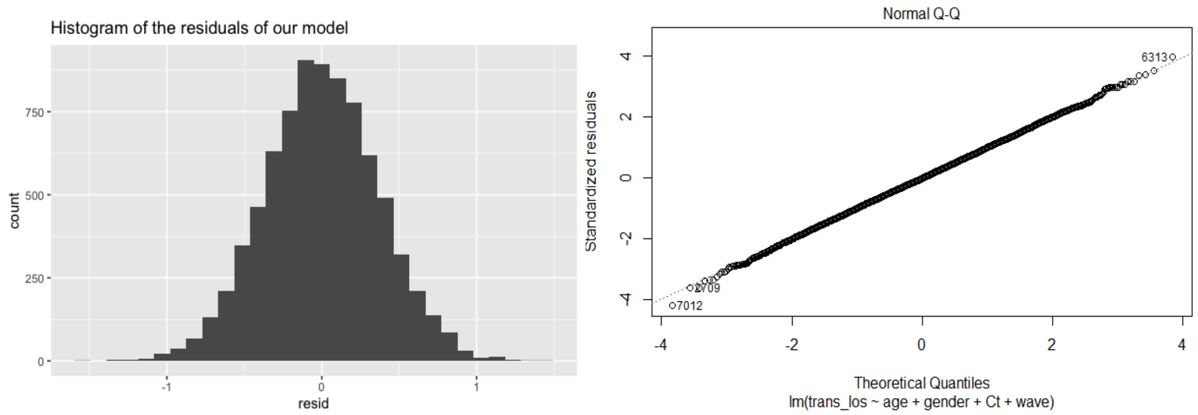


Fig 19. Histogram (left) and normal Q-Q plot (right) of the multiple linear regression

Question 2. Belgian COVID-19 hospital data

A. Formulate and fit a model that accounts for the association between measurements coming from the same hospital; interpret the findings with regard to this association.

To answer this question, we can consider hospital as a grouping factor that could lead to incorrect conclusions when using the model. The data was collected from 12 different hospitals, it is possible that the data from within each hospital are more similar to each other than to the data from different hospitals.

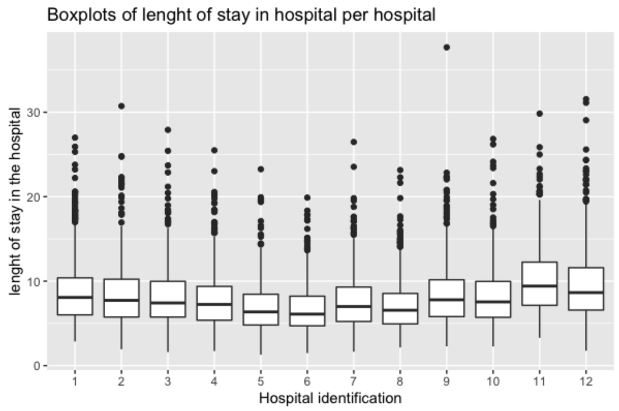
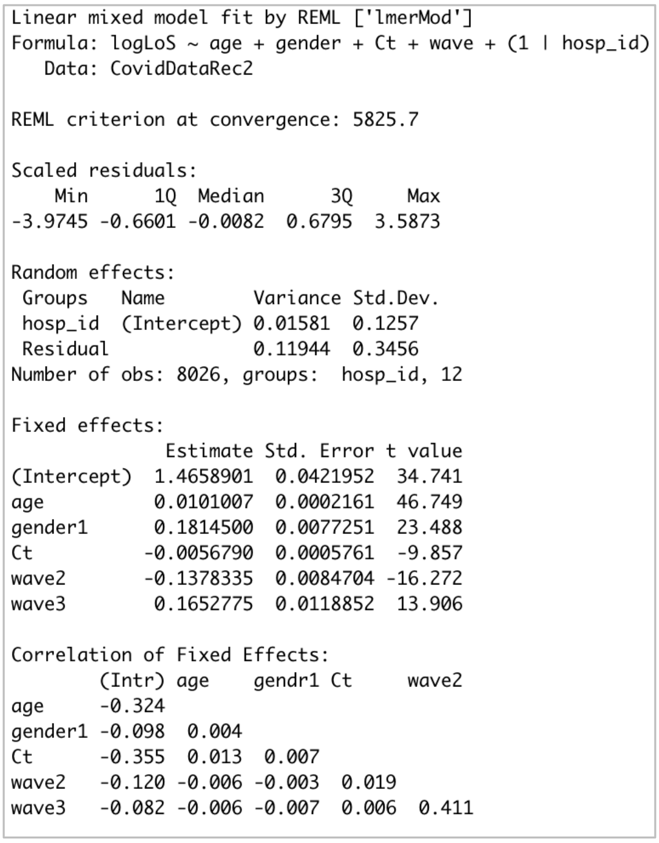


Fig 20. Boxplot of length of stay for each of the hospitals

If we include the variable of hospital in our multiple linear regression model, we will get estimates for each of the 12 hospitals. However, we are not interested in this, we just want to predict how age, gender, Ct and wave influence the length of stay while controlling for the variation coming from hospitals. Therefore, it is appropriate that we run a linear mixed model to account for this possible variability.



```
Linear mixed model fit by REML ['lmerMod']
Formula: logLoS ~ age + gender + Ct + wave + (1 | hosp_id)
   Data: CovidDataRec2

REML criterion at convergence: 5825.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9745 -0.6601 -0.0082  0.6795  3.5873

Random effects:
 Groups   Name        Variance Std.Dev.
 hosp_id  (Intercept) 0.01581  0.1257
 Residual             0.11944  0.3456
Number of obs: 8026, groups:  hosp_id, 12

Fixed effects:
              Estimate Std. Error t value
(Intercept)  1.4658901  0.0421952  34.741
age          0.0101007  0.0002161  46.749
gender1      0.1814500  0.0077251  23.488
Ct          -0.0056790  0.0005761  -9.857
wave2       -0.1378335  0.0084704 -16.272
wave3        0.1652775  0.0118852  13.906

Correlation of Fixed Effects:
        (Intr) age    gendr1 Ct     wave2
age     -0.324
gender1 -0.098  0.004
Ct      -0.355  0.013  0.007
wave2   -0.120 -0.006 -0.003  0.019
wave3   -0.082 -0.006 -0.007  0.006  0.411
```

Fig 21. Result of linear mixed model in R accounting for hospitals

In our linear mixed models, it showed that the variance for hospital was 0.01581. To see whether this number explain a lot of variation, we can simply divide the number to the total of variance (0.1581 / (0.01581 + 0.11944)). The result of this simple mathematical equation is 0.1169, this means that the differences between hospitals explains about 11.7 % of the variance in logLoS that was left over after the variance explained by our fixed effects.

B. There is some missingness in the data (i.e. the data are incomplete). Discuss (no additional analyses required):

The extent to which the methods (tests, models,) chosen in the foregoing are valid, given that data are incomplete.

Whether or not the methods chosen in the foregoing are valid, depend on the type of missingness in the dataset. There are three types of missing data, which are Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). One should avoid MNAR at all cost, as this might indicates that the study was poorly designed. When MCAR is ideally the type of missing that we desired, this is rarely the case especially in epidemiological studies. Therefore, in most studies, MAR is the most likely to be the type of missingness (1).

In our tests and models, we performed a complete-case analysis, in which every row with missing columns were removed from the analysis. To compare the length of stay in the hospital, for recovered individuals, between males and females, and across periods we used two different approaches, i.e. parametric analyses with log-transformed data and non-parametric analyses with non-transformed data. 191 Subjects have missing values for length of stay (LoS), our outcome/dependent variable, missingness in the data is therefore an important issue. In the analyses with log-transformed data, i.e. the t-test and ANOVA, missing values (NA's) are removed by R before calculation. Also, in the non-parametric tests missing values are ignored. These four analyses have therefore been conducted as complete-case analyses.

We computed two models. First, a linear regression model that describes the relation between the length of stay in the hospital and four variables of the dataset (age, gender, wave and Ct-value). Second, a linear mixed model that accounted for the association between measurements coming from the same hospital. Both models deleted 325 observations of the 8351 recovered subjects due to missingness and we thus performed a complete-case analysis.

If our data are MCAR, our statistical analysis still yield unbiased and valid estimation (1). However, recovering missing data can improve the power of our statistical analysis. The data are considered as MCAR when the probability that a variable is missing is unrelated to the value of the variable itself or the value of any other variables.

However, if our data are MAR, the complete case analysis might become problematic, as the result can be either unbiased or biased (1). Even when the complete-case analysis of MAR yields valid estimation, this result could be inefficient because data on incomplete cases were removed. The data are considered as MAR If the probability of data being missing is the same only within groups defined by observed data.

Which methods one could choose to increase the validity of the above analyses with incomplete data?

When aiming to increase the validity of the above analyses with incomplete data we can choose between different model frameworks, i.e. selection models (SEM), pattern-mixture models (PMM) and shared-parameter models (SPM) (2). For our report we will discuss the selection model for MAR, since MCAR is in most cases very unlikely. It will allow us to study missingness that depends on outcome or covariates but only through the observed data. We could use multiple imputation to increase the validity of the above tests (2). This method consists of three consecutive steps. First, there is an imputation task that will create M (which is more than 1) complete data sets. The imputed datasets are identical for the observed data but differ in the imputed values (the data that was first missing). The results of the analyses will differ because the datasets differ. Second, each of these datasets are analysed by the standard analyses we performed to answer the questions above, this is called the analysis task. Third, the results from the M analyses are combined into a final estimate and its variance, called the inference task.

What routes would you think are useful to conduct a sensitivity analysis towards missing data.

Sensitivity analysis is defined as the study which defines how the uncertainty in the output of a model can be allocated to the different sources of uncertainty in its inputs (3).

When performing a sensitivity analysis, additional (different) assumptions about the reasons for the missing data are made, and these assumptions are then applied on the primary analysis. We use sensitivity analysis to evaluate the impact of deviations from our MAR assumption within models because MNAR can significantly impact our model and it is not possible to test MNAR against MAR (4). In summary, this sensitivity analysis can thus evaluate the robustness of the results to the deviations from the MAR assumption. Sensitivity analysis can also be used to handle MNAR data, in which the extreme scenarios will produce bounds for the impact of missing data (1).

References

1. Perkins NJ, Cole SR, Harel O, Tchetgen Tchetgen EJ, Sun B, Mitchell EM, et al. Principled Approaches to Missing Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):568-75.
2. National Research Council Panel on Handling Missing Data in Clinical T. The Prevention and Treatment of Missing Data in Clinical Trials. Washington (DC): National Academies Press (US)Copyright 2010 by the National Academy of Sciences. All rights reserved.; 2010.
3. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64(5):402-6.
4. Beunckens C, Molenberghs G, Kenward MG. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. Clin Trials. 2005;2(5):379-86.