



Research article

A Six Sigma DMAIC methodology as a support tool for Health Technology Assessment of two antibiotics

Alfonso Maria Ponsiglione¹, Carlo Ricciardi², Giovanni Improta^{3,4,*}, Giovanni Dell’Aversana Orabona⁵, Alfonso Sorrentino⁵, Francesco Amato^{1,4,+} and Maria Romano^{1,+}

¹ Department of Electrical Engineering and Information Technology (DIETI), University of Naples “Federico II”, Naples, Italy

² Department of Advanced Biomedical Sciences, University of Naples “Federico II”, Naples, Italy.

³ Department of Public Health, University of Naples “Federico II”, Naples, Italy

⁴ Interdepartmental Center for Research in Healthcare Management and Innovation in Healthcare (CIRMIS), University of Naples “Federico II”, Naples, Italy

⁵ Maxillofacial Surgery Unit, Department of Neurosciences, Reproductive and Odontostomatological Sciences, University Hospital of Naples “Federico II”, Naples, Italy

* **Correspondence:** Email: ing.improta@gmail.com; Tel: +390817473096.

† These authors contributed equally to this work.

Supplementary

S1. Multiple regression model

Prior to estimating the final values of the coefficient of the Multiple Regression model, the following assumptions are checked for both antibiotics:

- **Linearity:** to verify if a linear relationship exist between the dependent variable and each predictors of the model;
- **Independence of the residuals:** to verify if the errors of the model are independent;
- **Collinearity:** to verify if the predictors are not linearly correlated with each other’s;
- **Outliers:** to verify if there are influential cases biasing the model;
- **Normality of the residuals:** to verify if the errors of the model are normally distributed;

- Homoscedasticity: to verify if the variance of the errors of the model is constant.

S1.1. Linearity

Scatter plots of the dependent variable vs the independent variable are commonly used in regression models. However, in the case of multiple regression, they do not take into account the effect of the other independent variables in the model. Therefore, partial regression plots are used to assess possible nonlinearity in the data.

Here, partial regression plots are used to identify the linear relationship between the Length of Stay (LOS), which is the dependent variable, and each of the 8 selected predictors (independent variables). Partial regression plots show on the x-axis the residuals of the independent variable, i.e. the errors obtained in modelling a given variable using all the other predictors, and on the y-axis the residuals of the LOS modelled excluding that specific independent variable.

The plots for the two antibiotics are displayed in Figures S1 and S2.

The plots show a weak linear relationship between the LOS and the selected predictors. In the case of Ceftriaxone, ASA Score, Diabetes and Oral Hygiene have the highest R2; while in the case of Cefamezyn plus Clyndamicin, ASA Score, Diabetes and Infections exhibit the highest R2.

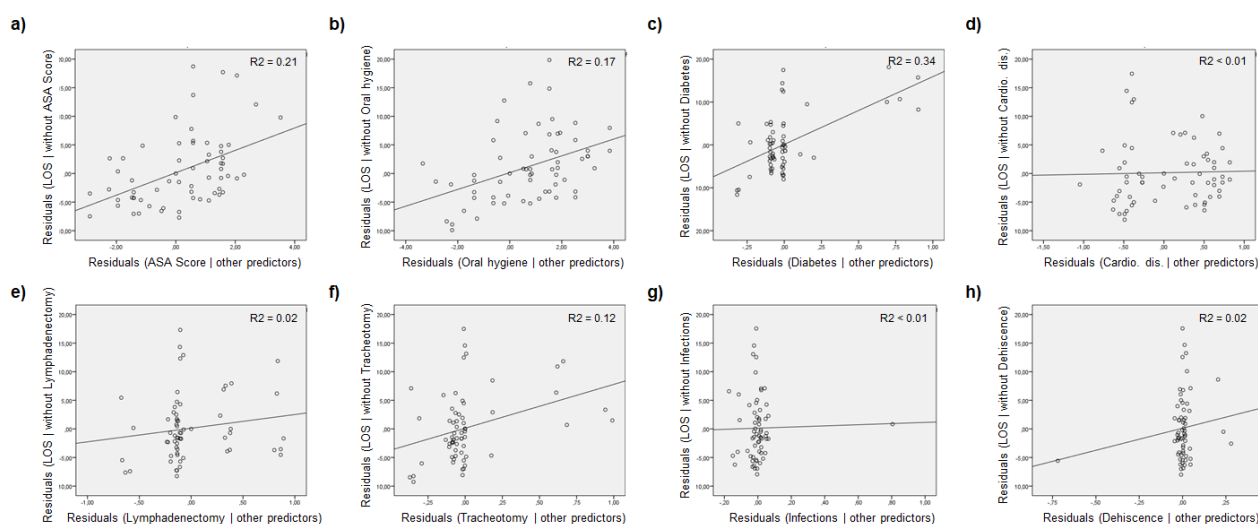


Figure S1. Partial regression plots for Ceftriaxone displayed on each predictor: a) ASA Score; b) Oral hygiene; c) Diabetes; d) Cardiovascular disease; e) Lymphadenectomy; f) Tracheotomy; g) Infections; h) Dehiscence. Determination coefficient R2 is also reported for each subplot.

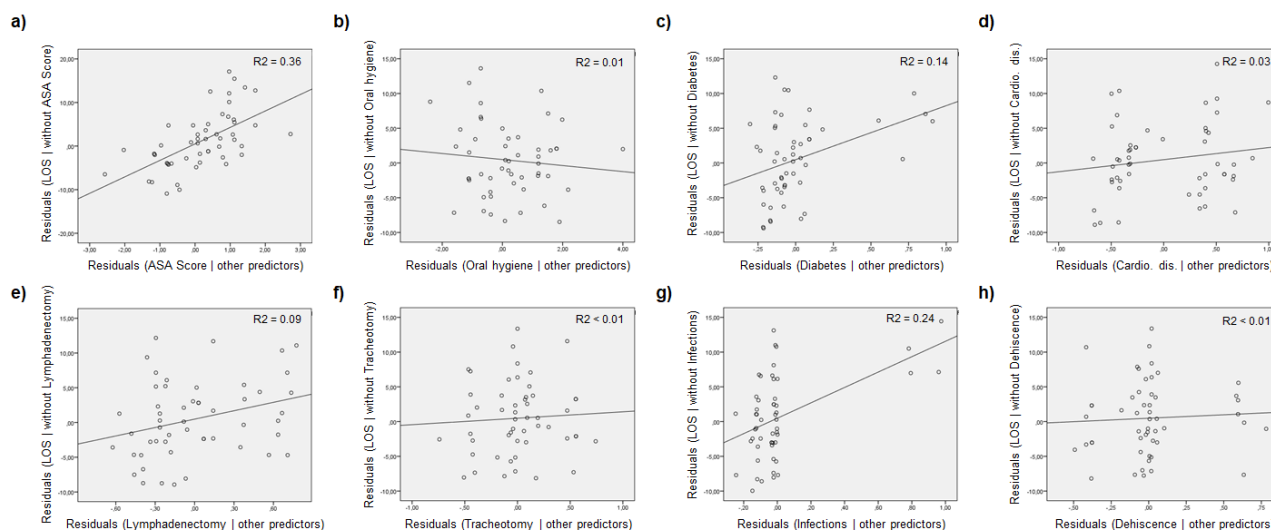


Figure S2. Partial regression plots for Cefamezyn plus Clyndamicin displayed on each predictor: a) ASA Score; b) Oral hygiene; c) Diabetes; d) Cardiovascular disease; e) Lymphadenectomy; f) Tracheotomy; g) Infections; h) Dehiscence. Determination coefficient R^2 is also reported for each subplot.

S1.2. Independence of the residuals

As required by the multiple regression methodology, the Durbin Watson test has been exploited to check if there is autocorrelation in the data. The output of the test can range from 0 to 4. Values around 2 mean no autocorrelation, while values lower than 1 and higher than 3 indicate positive and negative autocorrelation respectively.

Table S1 reports the results of the Durbin Watson test calculated for both sets of data related to the two antibiotics.

Table S1. Durbin Watson test.

Antibiotic	Durbin Watson test results
Ceftriaxone	2.36
Cefamezyn plus Clyndamicin	1.89

Results of the Durbin Watson test are in the acceptable range, so we can assume no autocorrelation in the data, i.e. the values of the residuals of the model are independent.

S1.3. Collinearity

In order to verify that all the predictors are independent each other's (absence of collinearity), Tolerance and Variance Inflation Factors are calculated and reported in Table S2.

Results are in the acceptable range, being the Tolerance higher than 0.2 (slightly lower values are calculated for ASA Score and Oral hygiene) and the Variance Inflation Factors lower than 10, for all the independent variables.

Table S2. Collinearity statistics.

Independent variables	Ceftriaxone		Cefamezyn plus Clyndamicin	
	Tolerance	Variance Inflation Factors	Tolerance	Variance Inflation Factors
ASA Score	0.401	2.49	0.134	7.47
Oral hygiene	0.338	2.96	0.166	6.03
Diabetes	0.795	1.26	0.764	1.31
Cardiovascular disease	0.519	1.93	0.507	1.97
Lymphadenectomy	0.513	1.95	0.278	3.60
Tracheotomy	0.451	2.22	0.376	2.66
Infections	0.201	4.97	0.877	1.14
Dehiscence	0.240	4.17	0.632	1.58

S1.4. Presence of outliers

The presence of influential cases biasing the model, i.e. those having a big effect on the regression equation that can be found especially in small dataset [1], was checked by plotting two adimensional indicators used to identify the outliers of the model. In particular, the Cook's distance is plotted vs the Centered Leverage Value, which is a measure of how much a certain value of the independent variable is different from its mean, as shown in Figure S3.

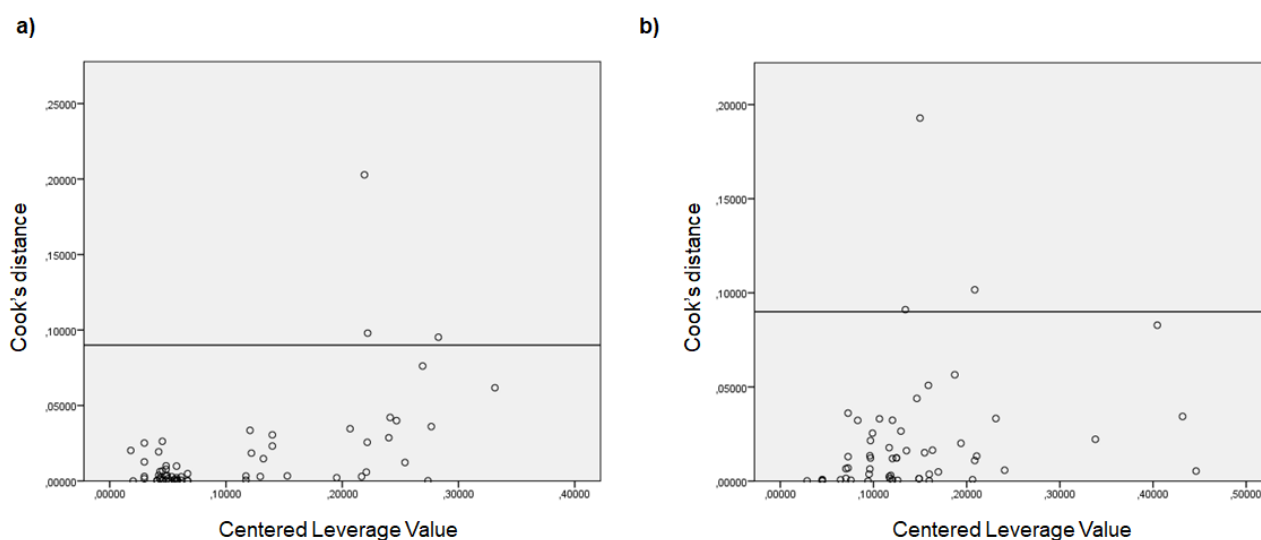


Figure S3. Cook's distance is plotted vs the Centered Leverage Value for: a) Ceftriaxone; b) Cefamezyn plus Clyndamicin. A horizontal line at 0.09 is set in both graphs as threshold to remove the most influential points.

The higher the Cook's distance, the more influential the point is. We identified 3 cases per antibiotic considering a Cook's distance equal to 0.09 as threshold to remove the most influential points.

Since the selected outliers (3 cases per antibiotic) represent a very small percentage of the population (4.5% for Ceftriaxone and 5.5% for Cefamezyn plus Clyndamicin), they are removed from the model. In both cases, indeed, we verified that the LOS values of the identified potential outliers are more than 1-2 standard deviations above/below the mean and therefore we deleted them from the models in order to reduce the influence of extreme values in the parameter estimates.

S1.5. Normality distribution of the residuals

The normality distribution of the residuals, which is particularly relevant in multiple regression models [2], is checked with a Probability-Probability plot (empirical vs theoretical cumulative distribution function), as reported in Figure S4.

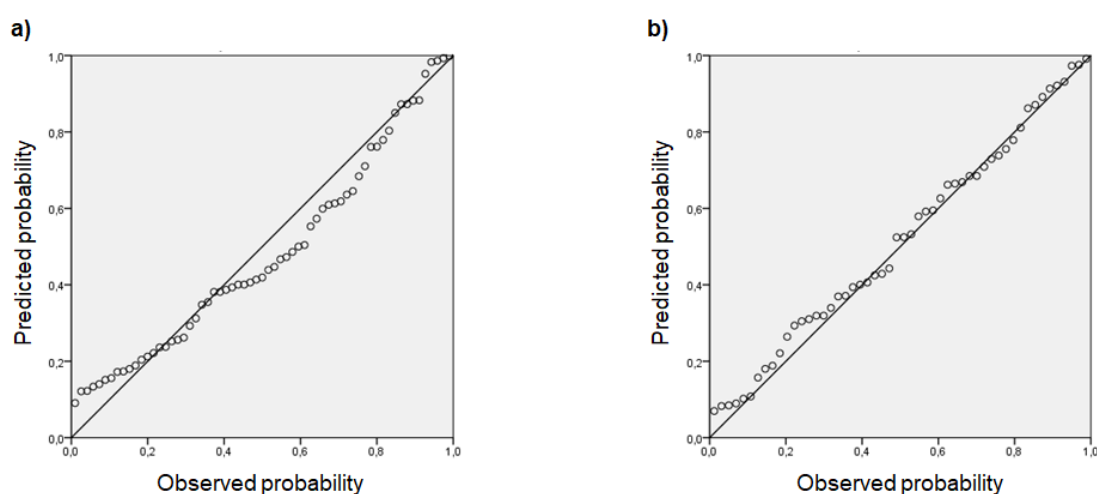


Figure S4. Probability-Probability plot for: a) Ceftriaxone; b) Cefamezyn plus Clyndamicin.

Basically, both plots show data following the diagonal line. Actually, since no severe violation of the normality is present, we can assume that residuals of the model are normally distributed for both antibiotics.

S1.6. Homoscedasticity of the data

As also showed in other studies [3], the robustness of the MR model was verified by checking the homoscedasticity of the data by plotting the standardized residuals against the standardized predicted values obtained from the model, as shown in Figure S5.

Ideally, homoscedasticity is verified if the residuals are randomly scattered around 0 (the horizontal line), providing a relatively even distribution [4]. Since the variation of the residuals is almost constant around the mean and only a few points fall out of the 95% confidence band (upper and lower confidence limits around the mean are indicated in each graph), no significant heteroscedasticity is detectable.

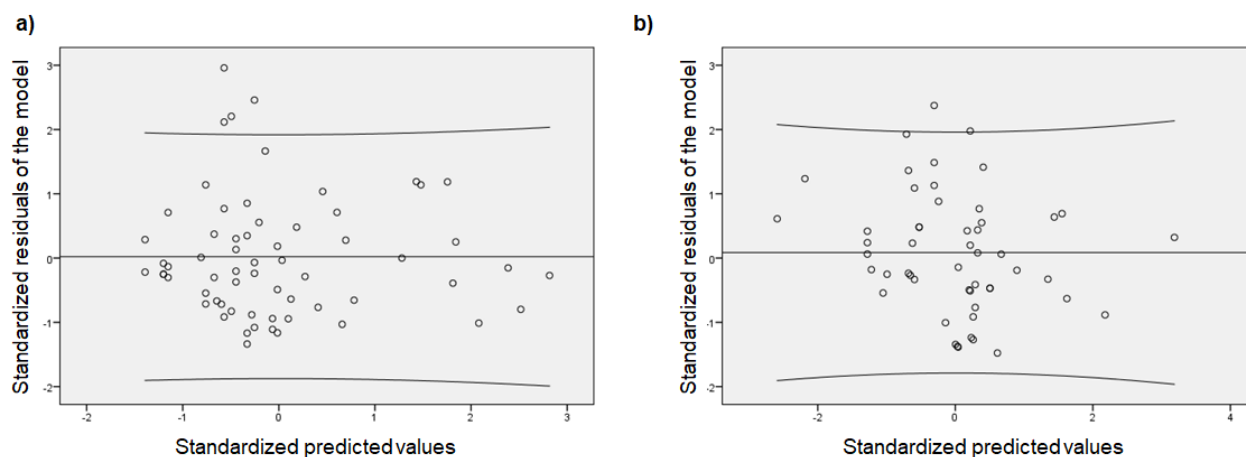


Figure S5. Standardized residuals vs standardized predicted values for: a) Ceftriaxone; b) Cefamezyn plus Clyndamicin. Mean and 95% confidence band are included in each graph.

References

1. C. Combes, F. Kadri, S. Chaabane, Predicting hospital length of stay using regression models: Application to emergency department. In 10eme Conference Francophone de Modelisation, Optimisation et Simulation-MOSIM'14, 2014.
2. M. N. Williams, C. A. G. Grajales, D. Kurkiewicz, Assumptions of multiple regression: Correcting two misconceptions. *Pract. Assess. Res. Eval.*, **18** (2012), 11.
3. T. R. Moss, Hospital length of stay and healthcare costs among African American women due to obesity and diabetic conditions in United States: a model for correlation studies comparing ethnicity, co-morbidities and hospital resources, *Chronic Dis. Transl. Med.*, **4** (2018), 244–253.
4. J. W. Osborne, E. Waters, *Multiple Regression Assumptions*. ERIC Digest., ERIC Publications, 2002.



AIMS Press

©2021 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)