# Biostatistics Project

Survival Analysis

# Survival analysis in patients with monoclonal gammopathy of undetermined significances

**Roberto Rey Sieiro**

**Abstract**

Monoclonal gammopathy of undetermined significance also known as MGUS is a plasma cell dyscrasia in which plasma cells or other types of antibody-producing cells secrete a myeloma protein. Research has shown that several factors may indicate the progression to a plasma cell malignancy (PCM) or even a premature death. This study aims to detect which of those factors may lead to death, which could be helpful in order to provide a closer monitoring to those patients and act as soon as possible.

## 1. Introduction

MGUS or monoclonal gammopathy of undetermined significance is a condition that causes the body to create an abnormal protein. This protein is the monoclonal protein or M protein. It is composed by blood cells called plasma cells in the body's bone marrow.

This disease is usually not serious but diagnosed patients may have a slighy increased risk of developing blood cancers like multiple myeloma. When the bone marrow gets crowded with very large amounts of M proteins the tissue can be damaged. Aside from blood cancers it can also produce bone fractures, blood clots and kidney problems, all this problems may lead to death too.

Our data is found in the library *survival* and the data is loaded with the function, data(mgus). We are working with the *mgus2* data it has 1384 observations and the following 11 variables:

- id: subject identifier.
- age: age at diagnosis, in years.
- sex: a factor with two levels female (F) and male (M).
- dxyr: year of diagnosis.
- hgb: hemoglobin (g/dl).
- creat: creatinine (g/dl).
- mspike: size of the monoclonal serum splike.
- ptime: time until progression to a plasma cell malignancy (PCM) or last contact, in months.
- pstat: occurrence of PCM: 0=no, 1=yes.
- futime: time until death or last contact, in months.
- death: occurrence of death: 0=no, 1=yes.

Our goal is to fit a proportionals hazards model in order to obtain the effect of a unit increase in a covariate to conclude which of the covariates may lead to death.

## 2. Data Preprocessing

Despite the data being obtained via a R library, the data still needs some preprocessing. Several outliers and NAs were found among the observations.

### 2.1. Dealing with NAs

There were some observations which had some NAs, so our first step was to calculate which was the proportion of NAs in order to decide our next step. The NAs were less than the 4% of the total observations of our dataset, so removing these observations with NAs was the approach we decided to take.

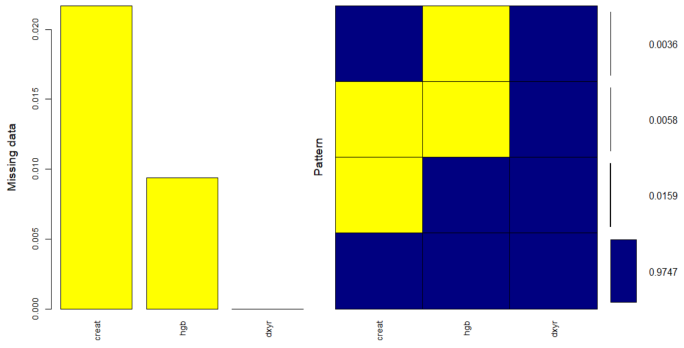


Figure 1: NAs proportion among the columns

Creatinine was the column with more NAs, creatinine is a breakdown product of creatine phosphate from muscle and protein metabolism and is used as an indicator of renal function. As we already stated MGUS can lead to renal failure, so it is probably going to be an important variable during the project.

### 2.2. Outlier detection

During the project we encountered several problems with some outliers, these problems were reflected when we plotted the residuals. So to deal with the outliers we used the *Cook's distance*. The Cook's distance is a commonly used

estimate of the influence of a data point when performing a least-squares regression analysis.

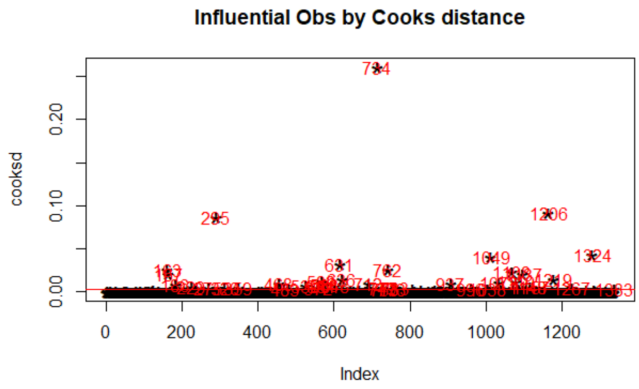


Figure 2: Cook's Distance using the traditional 4/n criterion

According to this criterion 40 outliers were found and removed from the dataset. We still have 1300 observations.

### 2.3. Correlation

As a final preprocessing step we have to check the correlation between our variables, correlation may be good for predicting models but in our case we want to build an explicative model. If there is lot correlation between the variables we may have some kind of overfitting or colinearity which we try to avoid.

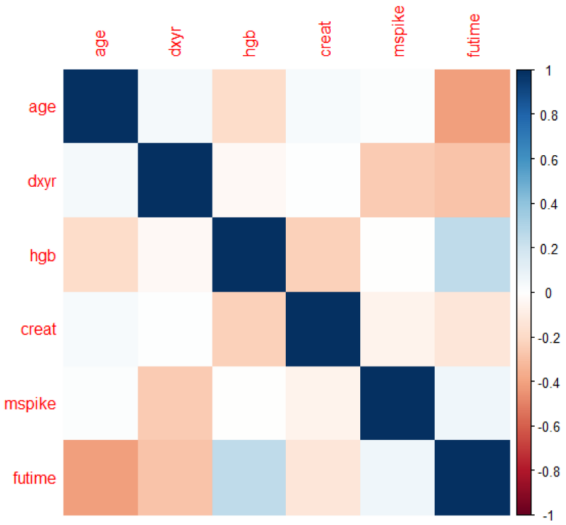


Figure 3: Correlation plot of the numerical variables

There is no significant correlation between the variables, the highest one is the correlation between age and futime although is smaller than 0.6 in absolute value.

## 3. Model building process

Once we have preprocessed our data we are going to start building our model. To determine which model we are going to use, we have to check which kind of data we have. First of all, we can see that we have covariates that have fixed values which we are going to consider them as constant and we also have a time dependent covariate called *pstat* and indicates that a pacient has developed a PCM. Also, it is associated with *ptime* which is the month at which the patient developed the PCM.

### 3.1. Model Selection

We are interested in building a model which measures the hazards of dying from this disease. The most used multivariate approach for analyzing survival time data in R is the Cox model. However, before building our Cox model, we are going to use the Kaplan-Meier estimate which is a nonparametric estimate of the survival function.

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

After comparing the survival functions, we are going deeper in the details with the Cox model. Our Cox model has a time dependent covariate, so it is going to be like:

$$h(t, \mathbf{X}(t)) = h_0(t) e^{\mathbf{X}(t)\beta}$$

We are going to have two periods for each pacient that develops a PCM, pre-PCM or post PCM.

### 3.2. Model Building

Before building our model we have to modify our data a bit. This is because we have a time dependent covariate, so we have to create both time intervals for patients with PCM, patients without PCM will not be modified. To create the start-stop format we can use the function *tmerge()*:

| id | death | tstart | tstop | PCM |
|---|---|---|---|---|
| 230 | 0 | 0 | 58 | 0 |
| 231 | 0 | 0 | 101 | 0 |
| 231 | 1 | 101 | 122 | 1 |
| 232 | 1 | 0 | 78 | 0 |
| 234 | 0 | 0 | 261 | 0 |
| 235 | 1 | 0 | 175 | 0 |

After modifying the data we are going to build survival curves based on the events (death or no). Also, we are going to compare survival curves for the categorical data, although we will check the cox model for further insight.
In order to build a good cox model we take into account the data, when interactions are included, and there are too many levels in a covariate it can lead to 0 variance coefficients. However, in our model we have two categorical variables aside from the status, which are sex and pstat, both of them have only 2 levels and the distribution of individuals for each factor is not very biased to a certain level. In addition we do not have many variables, so we can check every interaction with a conventional computer, so we are going to run a full model and perform backward elimination with BIC criterion. We chose BIC criterion versus AIC because we are willing to find a more explicative model. According to the covariates the only covariate that we did not take into

account was the id which is the subject identifier, the rest of the variables seemed to be relevant a priori.

Using the BIC criterion we obtained the following cox model:

- age + sex + hgb + creat + PCM + age:hgb + hgb:creat.

### 3.3. Model diagnostics

We check if the model is valid, one of the validation tools in our hand is the residuals analysis. They take value in $(\infty, 1]$ and have mean value 0. We use them to check the functional form of the covariates (continuous). So first we have to fit the model without the covariate, calculate the martingale residuals and plot the residuals versus the covariates.

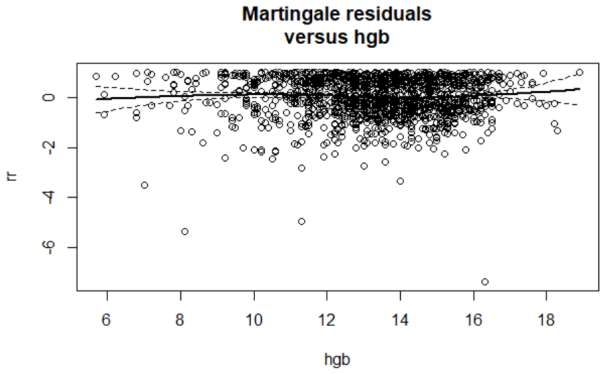


Figure 4: Martingale residuals vs hemoglobin

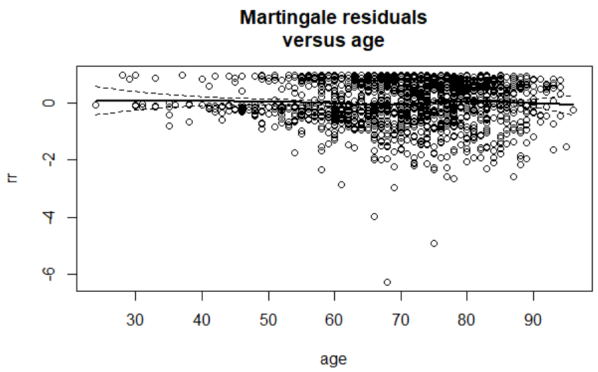


Figure 5: Martingale residuals vs age

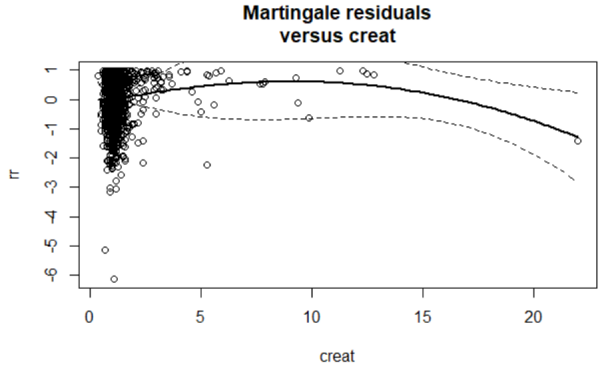


Figure 6: Martingale residuals vs creatine

As we can see the residuals look fairly linear except for the creatine, checking the residuals we could say that log-hazard ratio might not be linear. So trying to fit a penalized spline in order to reflect this non linear relationship, looking at the graph looks like we need a second order spline. Once we introduce this into the model, we obtain that the non linear component is relevant because the p value is 0.04. The linear component is also relevant as we obtain a 1.7e-04 p value.

### 3.4. Checking the proportional hazard assumption

In this subsection we are going to check if the proportional hazard assumtion hold.

$$\lambda(t) = \lambda_0(t)e^{\beta X(t)}$$
$$\lambda(t) = \lambda_0(t)e^{\beta(t) X}$$

In the first equation we have the model that we are looking for, in the second equation we have a time dependent coefficient. These models are much less common, the proportional hazard assumption is precisely that the coefficient does not change over time. With the function *cox.zph* we are going to obtain an estimate of $\beta(t)$. The null hypothesis of this function is that the coefficients are not time dependent.

| | chiqsq | df | p |
|---|---|---|---|
| age | 24.9857 | 0.99 | 5.7e-07 |
| sex | 0.2434 | 0.99 | 0.61845 |
| hgb | 12.8925 | 0.99 | 0.00032 |
| pspline(creat, df = 2) | 0.7980 | 2.06 | 0.68401 |
| PCM | 0.0932 | 1.00 | 0.76009 |
| age:hgb | 31.0859 | 1.00 | 2.4e-08 |
| hgb:creat | 0.0767 | 0.91 | 0.74767 |
| GLOBAL | 51.9623 | 7.94 | 1.6e-08 |

According to the test we can see that age, hgb and the interaction between both coefficients are time dependent. We are going to check the plot of each one of them to see how the time dependecy goes for each one of them.

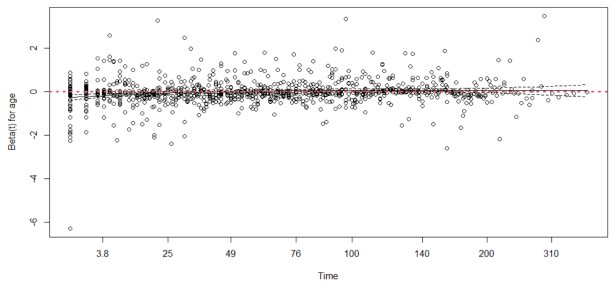


Figure 7: Age coefficient vs Time

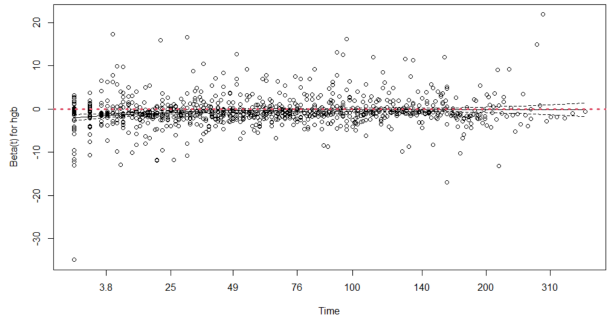


Figure 8: HGB coefficient vs Time

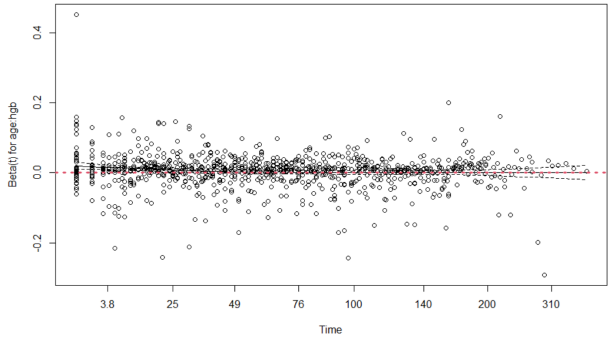


Figure 9: age:hgb coefficient vs Time

In the 3 of them we drew a red dotted line at height zero, which is the value that we are expecting. As we can see all of them look constant around 0, we can see high variability although we cannot see any relevant time dependency, so we decided to continue with this model. So our final model is:

- age + sex + hgb + pspline(creat, df=2) + PCM + age:hgb + hgb:creat.

## 4. Results

We are going to plot the survival curve of our data, this survival curve was obtained with the Kaplan-Meier estimator.

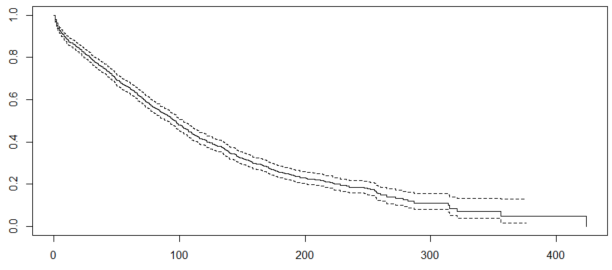


Figure 10: Survival Curve

The median survival time is 100 months for every person that has the disease.
Although, we were considering all the samples, now we compare the differences between having a PCM and not having a PCM. We see both of the survival curves below.

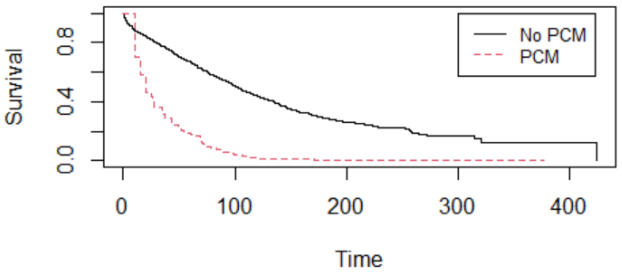


Figure 11: Survival Curves with PCM and without PCM

As we can observe, the difference is very noticeable they look like two different curves, the median survival time for PCM lies around 20 days and the one with no PCM lies around 120 days.
The next curves that we are going to compare for the sex coefficient, we see the curve for females and another curve for males.

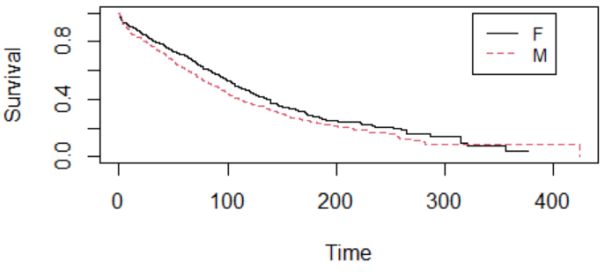


Figure 12: Survival Curves Males vs Females

Both of them look pretty similar, we can appreciate that male's curve is a bit lower nearly until the end. We tried to apply the log-rank test but we got an error 'Right censored data only', so we are going to check them in the cox model.

| | coef | exp(coef) | se(coef) | Chisq | p-value |
|---|---|---|---|---|---|
| age | -0.010069 | 0.989981 | 0.019685 | 0.26 | 6.1e-01 |
| sexM | 0.392060 | 1.480027 | 0.071063 | 29.79 | 4.8e-08 |
| hgb | -0.559673 | 0.5713959 | 0.114596 | 23.04 | 1.6e-06 |
| pspline(creat, df = 2), l | -0.481171 | 0.618059 | 0.109827 | 14.19 | 1.7e-04 |
| pspline(creat, df = 2), n | | | | 4.34 | 4.0e-02 |
| PCM | 1.601883 | 4.962368 | 0.115376 | 192.95 | 7.2e-44 |
| age:hgb | 0.005305 | 1.005319 | 0.001522 | 12.10 | 5.1e-04 |
| hgb:creat | 0.044719 | 1.045734 | 0.011452 | 7.96 | 4.8e-03 |

Here we have the final cox model coefficients and p-values, as we can observe every p-value seems relevant except the age covariate, but age has a relevant interaction with hemoglobin. We also obtained some information about the splines of the creatinine but we are not going to go into further detail. The next plot that we are going to do, is the coefficients (hazard ratios) plot. This plot does not work with splines nor interactions so they are not included in the graph.

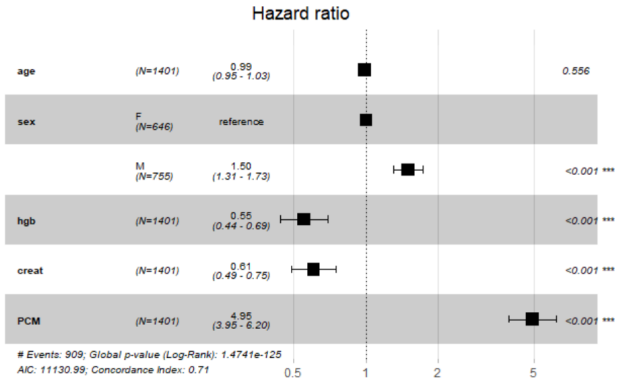


Figure 13: Hazard ratios

In order to interpret the coefficients, we check the exponential of the coeficient which is going to indicate how the hazard ratio changes for a unit increase in the covariates. First we discuss the categorical variables which are sex and PCM. According to sex we cannot interpret as a change in one unit, we have the 'female' category as a reference so we have the coefficient according to the males, when the patient is a male the hazard ratio increases by 48%. According to the PCM, our reference level is when the patient does not have a PCM, so the coefficient means that when patient develops a PCM the hazard ratio increases by 396%. Now we are going to analyze the numerical variables, age does not seem relevant according to its p-value. Looking at the hemoglobin, the hazard ratio for a unit change is 0.57. This means that each unit in hemoglobin decreases the hazard ratio by 43%. The last numerical variable is creatine, it has a hazard ratio for a unit change of 0.62 which means that each unit in creatinine decreases the hazard ratio by 38%. It seems that high levels of hemoglobin and low levels of creatine are risk factors to take into account for patients which might derive in a bad progression of their disease.
The last two variables that we have to interpret are the interactions. On the one hand, the first interaction is age:hgb, it indicates that there is a relationship between age and hemoglobin which is significant to the model. This interaction increases the hazard ratio by 0.5%. On the other hand, the other interaction is hgb:creat, there seems to be a relationship between hemolobin and creatine, each unit in this interaction increases the hazard ratio by 4.5%.

## 5. Discussion

All things considered we have only done a computational analysis, so in this section we are going to go a bit further and try to obtain some conclusions and justify the results we obtained during the project.
Taking a closer look at the coefficients, one of our hypothesis was, 'age is going to be a relevant factor', aging is accompanied by an impairment of the physiological systems, including the immune system (la Fuente, 2008). Assuming that this statement is true, which has been proved several times, we could say that ageing is not an important factor at least directly, this could be because the causes of death of the MGUS are not related to the strength of the inmune system.
Hemoglobin is a two-way respiratory carrier, transporting oxygen from the lungs to the tissues and facilitating the return transport of carbon dioxide (Marengo-Rowe, 2006). Low levels of hemoglobin are related with kidney failure, bone marrow problems (replacement of bone marrow by cancer) (Cuglievan et al., 2016) which are the main causes of death by MGUS, so a combination of MGUS and low-hemoglobin leads to fatal results. Although we cannot say that high hemoglobin levels are benefitial, it just reduces the hazard of dying with MGUS, very high hemoglobin levels produce certain tumors or lung diseases.
According to the interaction between hemoglobin and age, it was something that we were expecting. Hemoglobin variability is significantly correlated with age (Bal et al., 2018), which is one of the main conclusion based in many scientifical articles. High-amplitude fluctuation predicts high mortality; on the contrary low-amplitude fluctuations is related to better survival.
Creatinine was a predictor which did not have a linear dependency, we had to include a spline in order to reflect this relationship. The hazard decreases when the levels of creatinine increase, despite the fact that high-levels of creatinine can lead to kindey-failure, previous studies have described the association of increased mortality with lower creatinine levels in patients on chronic dialysis, but the implications of a low serum creatinine in critically ill patients are less well known (Ostermann et al., 2016).
The interaction between hemoglobin and creatine was also a relevant predictor. Several studies showed that very high hemoglobin levels and chronic kidney disease independently predict substantially increased risks of death and hospitalization in heart failure (Go et al., 2006). This interaction acts as the diminishing returns of high levels of creatinine and hemoglobin.
PCM was the most relevant factor (as expected), plasma cell myeloma is a type of cancer that begins in plasma cells (white blood cells that produce antibodies). It usually appears when there is a high presence of an M protein in the bone narrow. Less than 9% of the patients developed a PCM.
The year of diagnosis was not relevant, we though it could be relevant because of advances in biomedical techonolgies

and techniques but it was not the case. The size of the monoclonal serum splike was not a relevant factor either.

### 5.1. Issues and limitations

One of the main issues of the project were the flaws of the survival package. Despite the fact that the survival package is an excellent package for survival analysis in R, the use of splines and a time dependent covariate modifies the output of the summary of the model or the cox.phz in a less intuitive way. Furthermore, the comparision between survival curves is only visual because the the survival pack low-rank test does not work when we have interval censored data. In addition, the ggforest plot (from the survminer) does not display the effects of the interactions nor the splines effects. The data we used had several outliers and NAs, nevertheless after dealing with this preprocessing the results and the model looks consistent. Additionally we could not interpret well the splines because we did not find enough information among the class notes nor in the internet.

We did not study if any of the variables caused a PCM or increased the chances of developing a PCM. This could be an interesting extension of the project. We only had information about if the patient had a PCM or not, so the severity of the PCM would have been an interesting factor. The *cox.zph* reflected that there were some time dependent coefficients, checking the graphs we could not see any time relationship and concluded that it was related to variability, although this part may require further analyisis.

## 6. Bibliographical References


Bal, Z., Demirci, B. G., Karakose, S., Tutal, E., Uyar, M. E., Acar, N. O., and Sezer, S. (2018). Factors influencing hemoglobin variability and its association with mortality in hemodialysis patients. *The Scientific World Journal*.

Cuglievan, B., DePombo, A., and Angulo, G. D. (2016). Aplastic anemia: the correct nomenclature matters. *Haematologica*.

Go, Yang, Ackerson, Robbins, Massie, and Shlipak. (2006). Hemoglobin level, chronic kidney disease, and the risks of death and hospitalization in adults with chronic heart failure. *Circulation*.

la Fuente, M. D. (2008). Role of the immune system in aging. *Universidad Complutense de Madrid*, 1:1–16.

Marengo-Rowe, A. J. (2006). Structure-function relations of human hemoglobins. *Bayl Univ Med Cent*, 1:239–245.

Ostermann, M., Kashani, K., and Forni, L. G. (2016). The two sides of creatinine: both as bad as each other? *Journal of Thoracic Disease*.