Universitat Oberta de Catalunya (UOC)

Master's degree in Data Science

MASTER'S FINAL PROJECT

AREA: MEDICAL

Coronavirus in children, bio-markers, infection severity and hospitalization.

Author: Guillermo Argüello González

Tutors: Xavier Paolo Burgos Artizzu and Elisenda Bonet Carne

Professor: Ferran Prados Carrasco

Barcelona, January 8, 2021

Copyright

FINAL THESIS SHEET

Title of the Thesis:	Coronavirus in children, bio-markers, infection severity and hospitalization
Name of the author:	Guillermo Argüello González
Name of the teaching collaborator:	Xavier Paolo Burgos Artizzu and Elisenda Bonet Carne
Name of the PRA:	Ferran Prados Carrasco
Delivering date (mm/aaaa):	01/2021
Titulation or program:	Data Science
Area of work of the Final Thesis:	Medical
Language:	English
Key-words	Children, COVID-19, risk factors, interferon, cytokines

A mathematician, like a painter or poet, is a maker of patterns.

G. H. Hardy

Dedication

To my parents. And for Tania, of course.

Aknowledgements

I would like to thank my advisors Elisenda Bonet Carne and Xavier Paolo Burgos Artizzu for their support during this project, accurate guidance, and confidence.

I am also thankful to Hospital San Juan de Deu doctors Monica Girona Alarcon, Iolanda Jordan Garcia and Laia Alsina Manrique De Lara, for their support and help in interpreting and understanding the results of this project.

Finally, I would like to thank to Hospital Sant Joan de Déu of Barcelona, particularly to Kids Corona program that has allowed me to have great data to work with. This study has been funded by Stavros Niarchos Foundation (SNF), Banco Santander and other private donors of Kids Corona.

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection is typically very mild and often asymptomatic in children. Also, children and adolescents are less susceptible to the infection. Anyway, some children require hospitalization and rarely develop a severe multisystem inflammatory syndrome. Risk factors are widely studied in adults but there are not studies with large databases in children. Regarding IFN signature there is also little knowledge in its behaviour in children. Our goal is to bring new results in these two lines.

Using generalized additive models we have found that risk factors in children are different than in adults. Risk factors in adults are comorbid conditions often considered acquired or probably related with unhealthy lifestyle but in children are congenital conditions. There are no significant gender differences in children.

SOCS1 (suppressor of cytokine signaling 1) and CIITA (major histocompatibility complex class II transactivator) gene have an inverse relationship between them. We can use SOCS1 y CIITA levels, using a decision tree, to classify cases between mild and severe: low levels of CIITA and high of SOCS1 are related with severe cases. Finally, we observed that CIITA levels are higher in children than in adults. This could provide children protection against the infection and could open a new research line for treatments.

Keywords: Children, COVID-19, risk factors, interferon signature, cytokines

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and justification of the interest of the proposal

The novel coronavirus disease, COVID-19, caused by SARS-CoV-2 was originated in Wuhan, Hubei Province, China. The virus has quickly spread all around the world, as of now nearly 80 million cases and 2 million deaths have been confirmed worldwide. [1]

In Spain, the first case of COVID-19 infection was reported on 31st of January 2020. The pandemic has had a fast spread in the country generating more than 800 000 cases and 30 000 deaths.

COVID-19 disease can affect people of all ages however, we already know that children and adolescents are less susceptible to the infection. The prevalence study published in the ENE-COVID19 report [2], which has performed more than 60 000 IgG antibody tests across the country, has found children under the age of 10 with lower prevalence than adults: mean values: 3.7(<1 year), 3.7(1-4 years), 2.9(5-9 years) versus 5.2 for the general population.

We also know that age is a key factor in contracting a severe type of disease: children develop milder types of the disease and with better prognosis than adults. [3]. At present, SARS-CoV-2 infection disease in pediatric age has not been well categorized. Data on clinical course and prognosis markers are missing in children admitted [4].

This work seeks to provide new data on the aggravation of disease and hospitalization in children population. Also, we will study interferon pathway in children with COVID-19.

1.2 Main objectives

This work aims to bring new knowledge to the development of COVID-19 in children. Following topics will be studied:

- How cytokines induced by interferon impacts in the aggravation of the disease and in hospitalization. Differences between children and adult population.
- Risk factors for severe COVID-19 in children. Analyze if are the same as per adults.

1.3 Methodology

A database collected in Barcelona´s Sant Joan de Deu Hospital as part of the Kids Corona project will be the main source of information [5]. A public database from the Mexican government with clinical data of COVID cases will be also used to enrich the analysis. The project will be addressed as an advanced analytical project: first step is to understand, to describe and to prepare the data for further analysis. Subsequently, different statistics and machine learning techniques will be applied (statistical inference, regression, decision tress...) to extract the maximum value to the data.

1.4 Planning followed

All deliverables will be completed incrementally and iteratively until the final version of the work is completed, having periodic reviews with the tutors

- First contact with tutors and project planning. Commitment date: mid-September
- Availability of BBDD and first exploratory analysis of the data. Fundamental to understanding what data we have and what we can get. Commitment Date: end of September
- Literature review: papers published on COVID in children analysis in order to provide an innovative approach. Commitment date: end of October.
- Descriptive analysis and database cleanup: this point will be done in parallel with the previous one. Commitment date: end of October.
- Advanced analytics: application of different techniques on the data set to obtain the highest possible value. Engagement date: mid-December.

- Memory writing: it will be done in parallel with the previous points and incrementally. Commitment date: mid/late December.

- Last review with tutors and preparation of the presentation for the defense. Commitment date: early January.

1.5 Summary of the final results

Regarding hospitalization risk factors in children, after studding a large database, we have found several differences with well known risk factors in adults: risk factors in adults are comorbid conditions often considered acquired or probably related with unhealthy lifestyle but in children are congenital conditions. There are no significant gender differences in children.

Regarding cytokines response in children we have found that two cytokines (CIITA and SOCS1) are the most relevant to classify children with COVID-19 in severe and mild cases. We have also found some differences with adults behavior in this cytokines. Finally, we observed that CIITA levels are higher in children than in adults, this could provide children protection against the infection

1.6 Summary of the rest of the Chapters

- Chapter 2: State of art. Literature review of previous knowledge about COVID-19 in children.

- Chapter 3: Hospitalization risk causes in children. Analysis of clinical risk factors in severe COVID-19 in children and it´s differences between adults.

- Chapter 4: Interferon signature and cytokines response in children with COVID-19. Differences in cytokines response between children with mild and severe symptoms. Comparison with adults.

Chapter 2

State of art

Scientific community is making a huge effort in provide knowledge novel coronavirus disease. More than 20 000 papers have been published since the pandemic outbreak. This volume of scientific work has no precedents. In this project we are going to focus on how the disease affect children, particularly to clinical characteristics of the infection.

2.1 Evidence before this study

Through this section we are going to summarize available knowledge regarding severity and prevalence of the disease in children and the reasons behind it. The Spanish National Research Council has made a great effort in resuming actual evidence [6], we are going to use this work as basic guide.

2.1.1 COVID-19 disease is less severe in children than in adults

Novel COVID-19 disease varies extremely with age. Mortality data across all countries shows that mortality increases dramatically with age. In Spain, regarding RENAVE public COVID report, 4 deaths have been reported in 0-15 ages and 21 children were in ICU, from 75 146 cases since 10th May, but deaths in elderly people are in order of tens of thousands. This behavior has been observed in all countries, nevertheless a small number of children develop very severe disease as described in a multicenter Spanish Pediatric Intensive Care Units servey [7]

Less severe disease in children is also reported in one of the biggest studies with children in China [8] with 728 (34.1%) laboratory-confirmed cases and 1407 (65.9%) suspected cases. The median age of all patients was 7 years (interquartile range: 2–13 years), and 1208 case patients (56.6%) were boys. More than 90% of all patients had asymptomatic, mild, or moderate cases.

We also know that pediatric population with symptoms, have similar ones than adults. In the meta-analysis made by an Italian team [9], they found that subjects frequently have fever and cough and rarely diarrhea, nasal congestion and dyspnea.

Finally, regarding risk factors, a Greek team have found an underlying medical condition in some studies in hospitalized children, in contrast with those not hospitalized. The most common comorbidities were chronic lung disease (including asthma), cardiovascular disease, and immune suppression. Although this is not confirmed in all studies and is unclear whether male gender can also be considered as risk factor, because of insufficient data [10].

2.1.2 COVID-19 disease is less frequent in children than in adults

The seroprevalence study published in Spanish ENE-COVID19 report [2], which has performed more than 60 000 IgG antibody tests across the country, has found children under the age of 10 with lower seroprevalence than adults: mean values: 3.7(<1 year), 3.7(1-4 years), 2.9(5-9 years) versus 5.2 for the general population. For ages between 10 and 20 years less prevalence is no longer significant.

There are some biases in above affirmation that we should manage. Maybe less PCRs have been done to children because they have less severe cases. Also, it could be that because of having less severe disease they generate less antibodies misrepresenting seroprevalence studies. Finally, children innate inmute could protect them from COVID-19, this immunity doesn't generate specific antibodies also misrepresenting seroprevalence studies.

2.1.3 Pediatric multisystem inflammatory syndrome

We have seen that children have mild COVID-19 disease, but serious syndrome have been described in a very few cases. Multisystem inflammatory syndrome associated with SARS-CoV-2 pandemics has recently been described in children (MIS-C), partially overlapping with Kawasaki disease (KD). Clinical presentations, regarding the WHO [11], are fever (38–40°C), rash or bilateral non-purulent conjunctivitis or muco-cutaneous inflammation signs (oral, hands or feet), hypotension or shock, features of myocardial dysfunction, pericarditis, valvulitis, or coronary abnormalities, evidence of coagulopathy, acute gastrointestinal problems and elevated markers of inflammation.

This MIS-C have similarities with KD but also can be compatible with toxic shock syndrome, macrophage activation syndrome or septic shock.

A significant increase in Kawasaki-like illness have been observed in countries affected by SARS-CoV-2 pandemic [12]. The root cause of KD is not clear yet, but some authors have related it with common flu coronavirus. Further research could explain if coronavirus are the root cause of this similar diseases.

2.1.4 Reasons behind less severity and frequency in children

The reasons behind less severe infections in children and probably less prevalence have been discussed widely last moths in the literature. There are two main biological theories: first one based on viral receptor ACE2 and another one based on differences between immune system in children and adults [6]

- ACE2: Lethality of COVID-19 is much higher in the elderly than in young people, and in men than in women. The preprint [13] assumes that the variation in the lethality of COVID-19 infection with age and sex is due to the variation in the expression of the ACE2 protein, the receptor of the virus in cells.

- Trained innate immunity: It has been proposed that children might be more resistant to COVID-19 because their innate immunity, which does not develop specific antibodies, is stronger than in adults, which could explain lower frequency of children with SARS-CoV-2 antibodies because they could have been infected and resolved the infection with innate immunity before it has caused serious consequences [14]

  According to some theories, the greater strength of the innate immune system in children may be due to trained immunity, favored by vaccines against other pathogens. Vaccines unleash specific immunity, but also enhance the body's response by stimulating innate immunity [15].

2.1.5 Neutralizing auto-antibodies against type I interferon

The immune system works as a coordinated army, with different levels of action and response. It´s goal is defending us against a wide variety of pathogens. When a new virus such as SARS-Cov-2 attacks our body, infected cells release an alarm signal (interferon -IFN- type I), which capture other neighboring cells, alerting them about invader's entry, in order to prevent virus replication. At this point, innate, non-specific and rapid immunity is launched, in which soluble components such as interleukins, type I or type III IFNs and cells such as macrophages, neutrophils, dendritic cells and natural cytotoxic cells are responsible for delaying the progression of the virus even avoiding infection and/or onset of symptoms. [6]

More than 10% of severe COVID-19 patients produce a type of antibodies that, instead of protecting them from the virus, worsens infection neutralizing immune system. Antibodies that neutralize interferon type 1 were found in 10.2% of severe patients, regardless of age, race and sex. It is the result of two international studies [16] [17] that may explain why some people develop asymptomatic or very mild disease and others severe or fatal. Three major risk factors have been known before this research: being a man, being older, and having previous illnesses. Also, the study adds a fourth risk factor: congenital genetic defects that can prevent the immune system from fighting and removing the virus. This type of problem seems to be much more common in men than in women.

Moreover it seems that patients with COVID-19 might have a low or delayed IFN response, which may lead to a high citokines release, ending up with a hiperinflammatory state. [18]

There is still little knowledge about the role of the IFN pathway in adult patients with COVID-19, even if it is under suspicion of being determinant for the severity of the disease, but in children there is a complete ignorance about its function in SARS-CoV-2.

2.2 Current research

As we have seen before, a few children develop severe cases of COVID-19, so the first issue is the lack of information.

Most of the papers reviewed perform descriptive statistics and inference of the data, but no advanced analytic techniques are being applied.

Many of the studies use data collected during the first epidemiological wave of COVID-19. It is necessary to extend the analysis with larger samples including patients of the second wave, to ensure that there are no selection or other kind of biases. It´s also important know if the disease is changing it´s behavior between waves.

In this study we will try to provide new knowledge on the following topics:

- Try to confirm risk factors in children with new data. Are comorbidities like chronic lung disease, cardiovascular disease, and immune suppression risk factors?
- Are risk factors for severe COVID-19 in children the same as per adults?
- Are there any risk factors for Intensive Care Unit admission at the moment of hospitalization in children?
- Are there any differences in cytokines generated by type I IFNs between age groups?

2.3 Methods

Algorithms selection in a data science problem is one of the key points, so I'd like to discuss this point in detail using Professor Bradley Efron's ideas in a fantastic talk about Prediction, Estimation, and Attribution. [19]

Two major families of algorithms are currently coexisting. First one, the traditional prediction methods coming from Gauss–Galton tradition, wildly developed during the twentieth century, regression ideas were adapted to a variety of important statistical tasks: the prediction of new cases, the estimation of regression surfaces, and the assignment of significance to individual predictors. Second family, “pure prediction algorithms” (neural nets, deep learning, boosting, support vector machines, random forests), that rise with the twenty-first century and are able to operate at immense scales, with millions of data points and even more millions of predictor variables.

Both families have advantages and disadvantages, also there are several differences of scientific philosophy and goals. I´m going to highlight most important ones for the purpose of this study. Traditional regression methods aim to extract underlying truth from noisy data: perhaps not eternal truth but at least some takeaway message transcending current experience. Pure prediction algorithms are great in empirical prediction accuracy but due it´s lack of explainability and interpretability are not suitable for the purpose of discover scientific truth.

In this study we are going to analyze datasets with a small number of observations, with the objective of understanding better risk factors in severe COVID-19 disease in children. Also we want to find how cytokines induced by Inferferon behave in COVID-19 cases and if there is any differences between severity and age groups. These problems are both classification problems, first of them binary classification problem and the second one multiclass classification. For all the above, traditional prediction methods will be used, specifically Logistic GAMs and decision trees that we’re going to comment below.

2.3.1 Binary classification

Most common models for binary classification problems are regression models, but these traditional models often fail because in real life effects are not linear. So, we are going to use generalized additive models (GAM), more flexible statistical methods that can identify and characterize nonlinear regression effects. GAM has the interpretability advantages of GLMs (generalized linear model) where the contribution of each independent variable to the prediction is clearly encoded. However, it has substantially more flexibility because the relationships between independent and dependent variable are not assumed to be linear.

The additive logistic regression model replaces each linear term by a more general functional form

$$log(\frac{\mu(X)}{1-\mu(X)}) = \alpha + f_1(X_1) + ... + f_p(X_p)$$

where X is the dependent variable (i.e., what we are trying to predict), $X_1, ..., X_p$ are the predictor variables and each f_j is an unespecified smooth function [20].

2.3.2 Multiclass classification

For multiclass classification we are going to use tree-based methods, in particular we are going to apply a popular method for regression and classification problems called CART. These methods partition the feature space into a set of rectangles, and then fit a simple model in each one. They are conceptually simple yet powerful. A key advantage of the recursive binary tree representation is it interpretability. The feature space partition is fully described by a single tree. This representation is extremely popular among medical scientists: the tree stratifies the population into strata of high and low outcome, based on patient characteristics.

Our data consists of p inputs and a response, for each of N observations: that is (x_i, y_i) for $i = 1, 2, ..., N$ with $x_i = (x_i1, ..., x_ip)$. The algorithm needs to automatically decide on the splitting variables and split points, and what topology the tree should have. Suppose first that we have a partition into M regions $R_1, ..., R_M$, and we model the response as a constant c_m in each region:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

There are different criteria for minimization the error of above function, for example, misclassification error, gini index, cross-entropy or deviance. Having found the best split, we partition the data into the two resulting regions and repeat the splitting process on each of the two regions. A large tree might overfit the data, while a small tree might not capture the important structure. Tree size is a tuning parameter governing the model´s complexity, and the optimal tree size should be adaptively chosen from the data [20].

Chapter 3

Hospitalization risk factors in children

Main objective of this chapter is to find risk clinical factors in children COVID-19 hospitalization and analyze if are the same as per adults. We are going to work with two databases, one large from Mexican government and other small from San Joan De Deu Hospital in Barcelona and try to apply results from Mexican data to Barcelona's one.

3.1 Data

3.1.1 Mexican National Epidemiological Surveillance database

We have seen that the disease is less frequent in children so is hard to find large COVID-19 children databases. The largest database regarding COVID with enough detail and clinical data we have found is from Mexican National Epidemiological Surveillance System [21]. In this database we can find daily updates in COVID suspicious cases in the whole country. Data has been taken on 15th November, at the time there were 656782 cases, 201737 have tested positive for SARS-CoV-2. In table 3.1 is the distribution of cases by age (pediatric/adult), if they have required hospitalization and admission to ICU.

Age	*Hospitalized*	*ICU*	*Count*	*Percent*
ADULT	NO	NO	160871	79.742933
ADULT	YES	NO	30805	15.269881
ADULT	YES	YES	2858	1.416696
PEDIATRIC	NO	NO	6695	3.318677
PEDIATRIC	YES	NO	394	0.195304
PEDIATRIC	YES	YES	114	0.056509

Table 3.1: *Mexican database cases distribution.*

The dataset contains 92 variables, on the symptoms of patients, the hospital in which they have been treated, the origin and the clinical characteristics. As our objective is to study clinical risks factors, we are going to work only with **10 features**: *Gender*, *Chronic kidney disease*, *Diabetes*, *Hypertension*, *COPD*, *Obesity*, *Age*, *Coronary heart disease*, *Immunosuppression*, *Other condition*. Most of these variables are known risk factors in adults for severe COVID-19. We will work with this data to find risk factors in children and analyse the differences with adults.

3.1.2 Sant Joan De Deu Hospital hospitalization data

To try to validate the model that we are going to train with the public database of the Mexican government we have a small dataset collected in San Joan de Deu Hospital during the first wave of the pandemic. In SJD dataset there are 27 children (<= 15 years), all of them tested positive for COVID-19, 11 of them have been hospitalized and 15 didn´t required hospitalization. Regarding variables the database is extracted from the clinical records of the patients so there many variables about other diseases, medications, hospitalizations, physical characteristics... As we want to use this data for validate the model, we will only use the same 10 clinical characteristics we are going to use for training it detailed previously.

3.2 Analysis

With the data we have just presented first we will analyse if the risk factors of adult hospitalization match those described in the literature, just to be sure that this data is aligned with currently knowledge. Secondly, let’s see if the risk factors in children are the same as in adults. Finally, to validate the model trained with Mexican National Epidemiological Surveillance database we are going to use data collected in Sant Joan de Déu Hospital in Barcelona.

3.2.1 Risk factors for hospitalization COVID-19 in adults

Risk factors for severe COVID-19 disease are well known in adults. We have seen in all countries worldwide older age along with male gender are the main risk factors, being old mans the group with high risk of severity and death. Since March 2020 in China [22] it has been found that severity risk is increased with comorbidities such as hypertension, type II diabetes, coronary heart disease, active cancer, chronic kidney disease, obesity and COPD. [6]

As we are going to work with Mexican government database, first of all we are going to check if the database agrees with the evidence we know of literature. Let´s check if risk factors in adults for hospitalization are the same as those described in literature that generate severity.

The target variable is whether a positive COVID-19 case will be hospitalized or not due the disease, and the features are several clinical characteristics. As we detailed before we are going to work with *Gender*, *Chronic kidney disease*, *Diabetes*, *Hypertension*, *COPD*, *Obesity*, *Age*, *Coronary heart disease*, *Immunosuppression*, *Other condition*. Only age is a continuous variable, all the others are binary (0 means not having the comorbidity and 1 have it), in gender 0 means woman and 1 man. We are going to analyze the correlation between the features, then we will apply a GAM model to the data to find which variables are important to explain hospitalization due COVID in adults, and finally interpret the results with partial dependency plots.

CORRELATION ANALYSIS

First of all we are going to analyze the correlation between all features. Correlation coefficients are used to measure how strong a relationship is between two variables, here are several types of correlation coefficient we are going to use Pearson's correlation, the most popular one.

To show the correlation between all features we will use the correlation matrix, a type of visualization that allows us to see in the same plot all correlation coefficients. Correlation coefficients are between -1 and 1. The extreme values -1 and 1 indicate a perfectly linear relationship where a change in one variable is accompanied by a perfectly consistent change in the other. A coefficient of zero represents no linear relationship. When the value is in-between 0 and $+1/-1$, there is a relationship, as the coefficient approaches -1 or 1, the strength of the relationship increases. Positive values indicate that if when the value of one variable increases, the value of the other variable also tends to increase. For negative values when a variable increases the other decrease.

In figure 3.1 we can find the correlation matrix between all the features. As we want to know how comorbidities affect the risk of hospitalization, we focus in the first row. Some of the features as age, gender, diabetes or hypertension have strength correlation with the target feature but as correlation does not imply causation we need to apply other techniques or models to find if this correlation is significant.

GENERALIZED ADDITIVE MODELS (GAM)

We are going to use in this chapter generalized additive models (GAM), GAMs have been introduced in Chapter 2. The response variable is hospitalization in COVID positive cases. The features are the clinical characteristics we have detailed previously. In GAMs models we can define different types of functions for each feature because the model will fit a different function for each of the features, and the functions could be not lineal, this is the main difference with

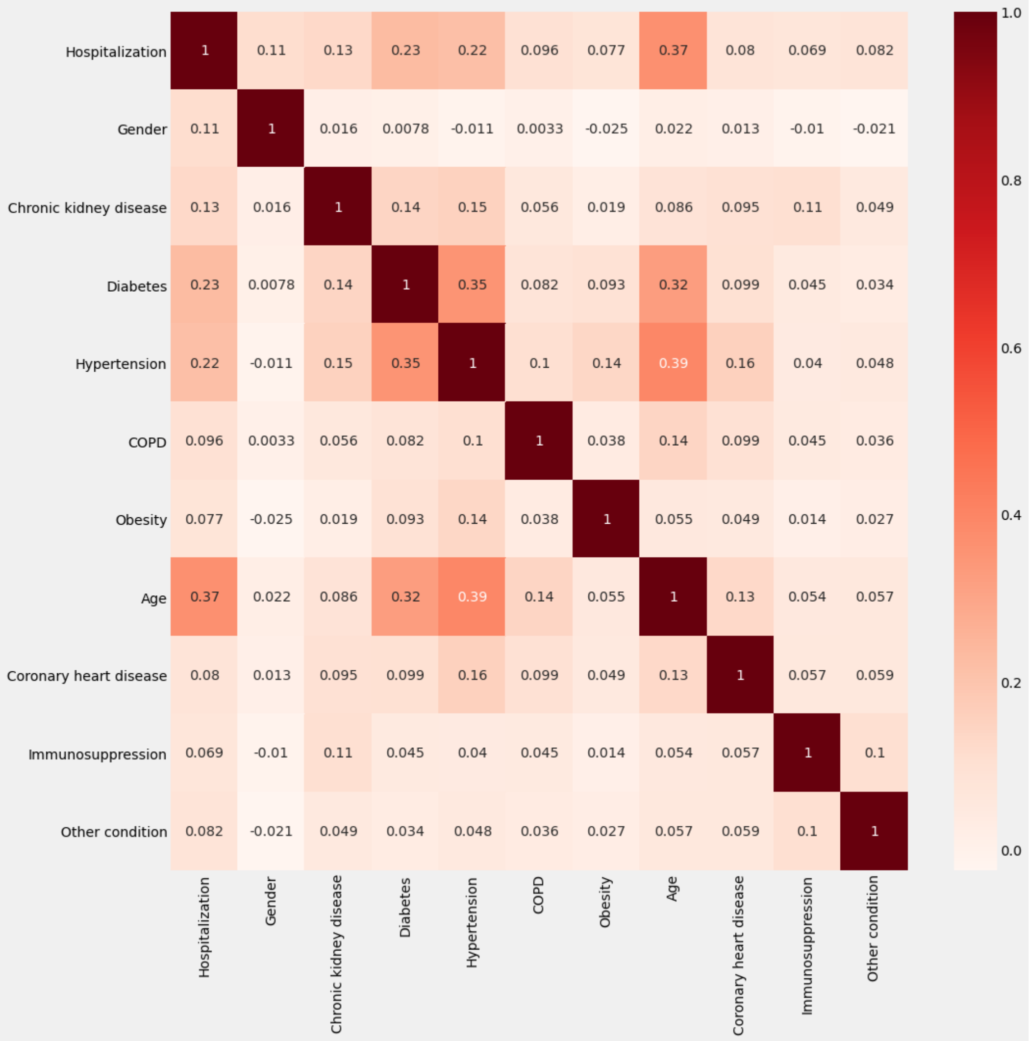
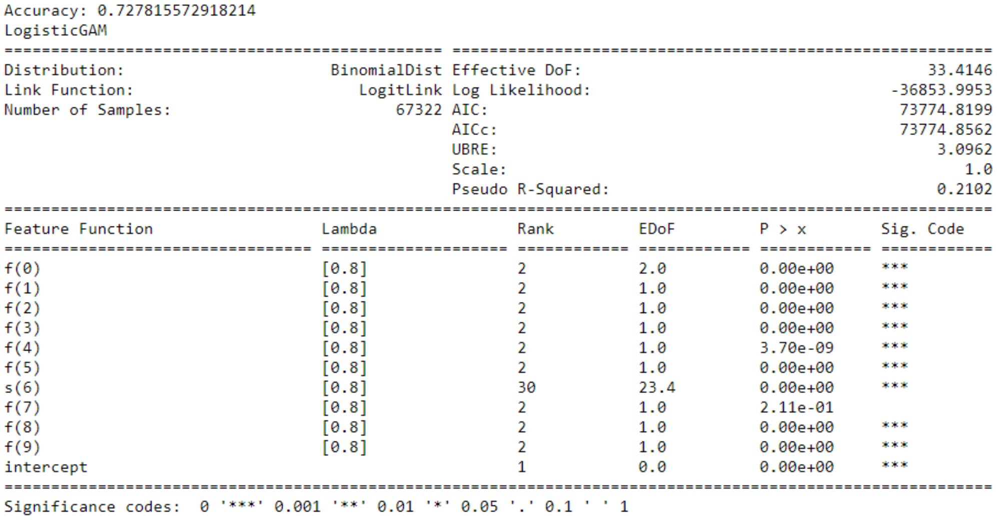


Figure 3.1: Correlation matrix for adults risks factors features in severe COVID-19.

lineal models. For the continuous variable (age) we are going to use a Spline. A Spline is defined by a set of control points and a set of basic functions that interpolate (fit) the function between these points. A spline of degree k is a piecewise polynomial that is continuously differentiable $k-1$ times. By choosing to have no smoothing factor we force the final spline to pass through all the points. If, on the other hand, we set a smoothing factor, our function is more of an approximation with the control points as ”guidance”. In following models K is called rank, and the smoothing factor, lambda. For the binary variables we will select a step function, it will take a value in 0 and another value in 1.

We train the model with all the features and the model will decide which of them are significant to explain the target. In figure 3.2 is the summary of the model.

It outputs a 73% of accuracy, is not great but is fair enough. Model also show, in p-value column, that all risk factors that are well documented in the literature for explaining severe



```
Accuracy: 0.727815572918214
LogisticGAM
=============================================== ==========================================================
Distribution:                      BinomialDist Effective DoF:                                     33.4146
Link Function:                        LogitLink Log Likelihood:                                -36853.9953
Number of Samples:                        67322 AIC:                                            73774.8199
                                                AICc:                                           73774.8562
                                                UBRE:                                               3.0962
                                                Scale:                                                 1.0
                                                Pseudo R-Squared:                                   0.2102
==========================================================================================================
Feature Function                  Lambda               Rank         EDoF         P > x        Sig. Code   
================================= ==================== ============ ============ ============ ============
f(0)                              [0.8]                2            2.0          0.00e+00     ***         
f(1)                              [0.8]                2            1.0          0.00e+00     ***         
f(2)                              [0.8]                2            1.0          0.00e+00     ***         
f(3)                              [0.8]                2            1.0          0.00e+00     ***         
f(4)                              [0.8]                2            1.0          3.70e-09     ***         
f(5)                              [0.8]                2            1.0          0.00e+00     ***         
s(6)                              [0.8]                30           23.4         0.00e+00     ***         
f(7)                              [0.8]                2            1.0          2.11e-01                 
f(8)                              [0.8]                2            1.0          0.00e+00     ***         
f(9)                              [0.8]                2            1.0          0.00e+00     ***         
intercept                                              1            0.0          0.00e+00     ***         
==========================================================================================================
Significance codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3.2: Logistic GAM model summary for adults risks factors in severe COVID-19.

COVID-19 cases are significant, except Coronary heart disease. The summary of the model also provide us some more statistical information as pseudo R-squared, AIC, AICc and log-likelihood, this are measures to compere models between them, these values have been used to find the best parametrization.

PARTIAL DEPENDENCY PLOTS

To graphically analyze the resulting model, we are going to use partial dependency plots. These plots are extremely useful because they are highly interpretable and easy to understand. The partial dependence plot shows the marginal effect one or two features have on the predicted outcome of a machine learning model. A partial dependence plot can show whether the relationship between the target and a feature is linear, monotonic or more complex [23]. The y-axis of a partial dependence plot for regression represents the marginal impact of the independent variable to the dependent variable. If the line is at 0, then for that value of the independent variable, there is 0 impact to the dependent variable. For binary features the plot is a stpe function, for all of them we see that 0 value is lower than 1, this means that having any of these comorbidities affect positively in the risk of being hospitalized.For positive values there is positive impact in the target feature and viceversa for negative values. In our case each of the plots are showing how the risk of being hospitalized due COVID-19 changes depending on the comorbidities, the gender and age.

Analyzing dependency plots 3.3 and 3.4 we can tell that there is a very strong relationship between all features, except Coronary heart disease, and the response variable. Reader may find surprising that at the most advanced ages (over 90) the risk of hospitalization appears to decrease, is because there are far fewer observations in that age group.

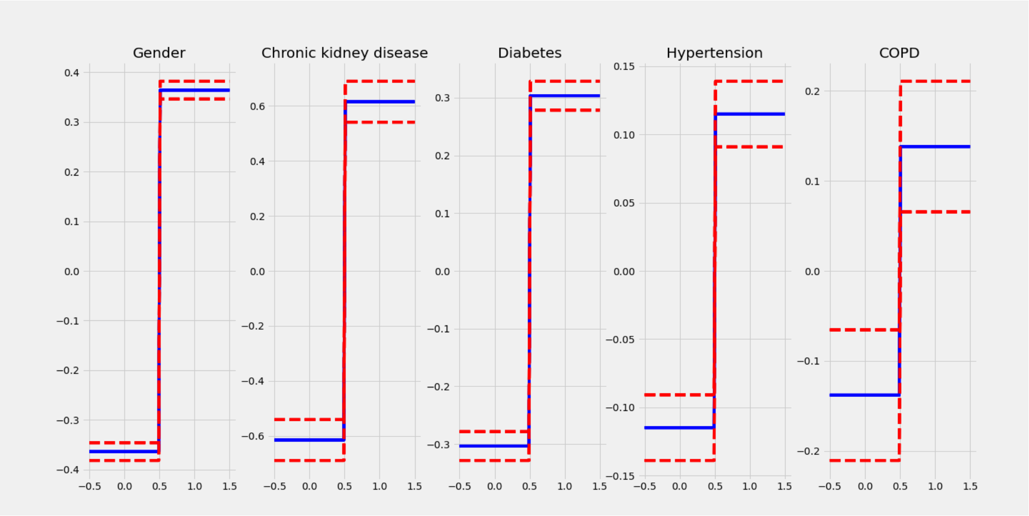


Figure 3.3: Partial dependence plots for adults risks factors in severe COVID-19 I.

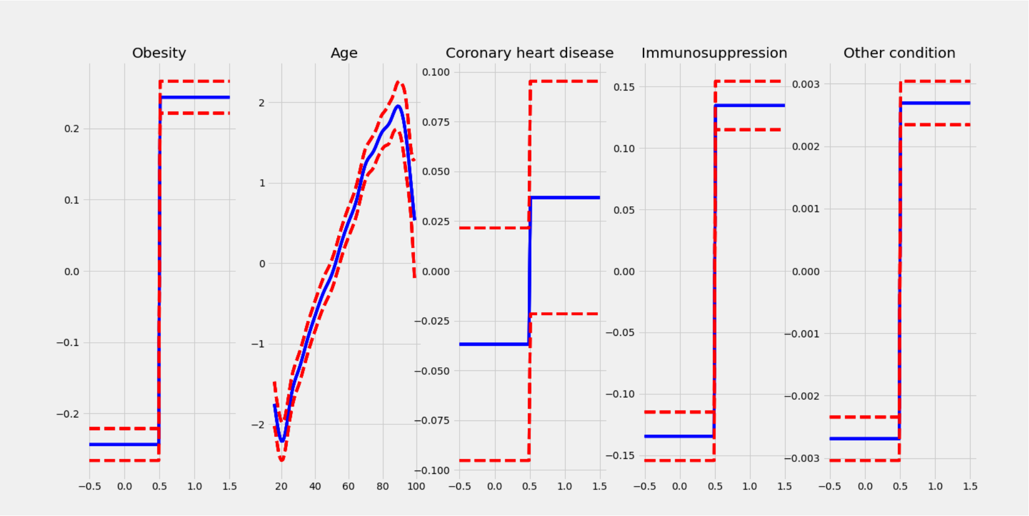


Figure 3.4: Partial dependence plots for adults risks factors in severe COVID-19 II.

With this first outcomes we can conclude that the database that we are using is appropriate: it´s behavior is as expected with the available evidence.

3.2.2 Risk factors for hospitalization COVID-19 in children

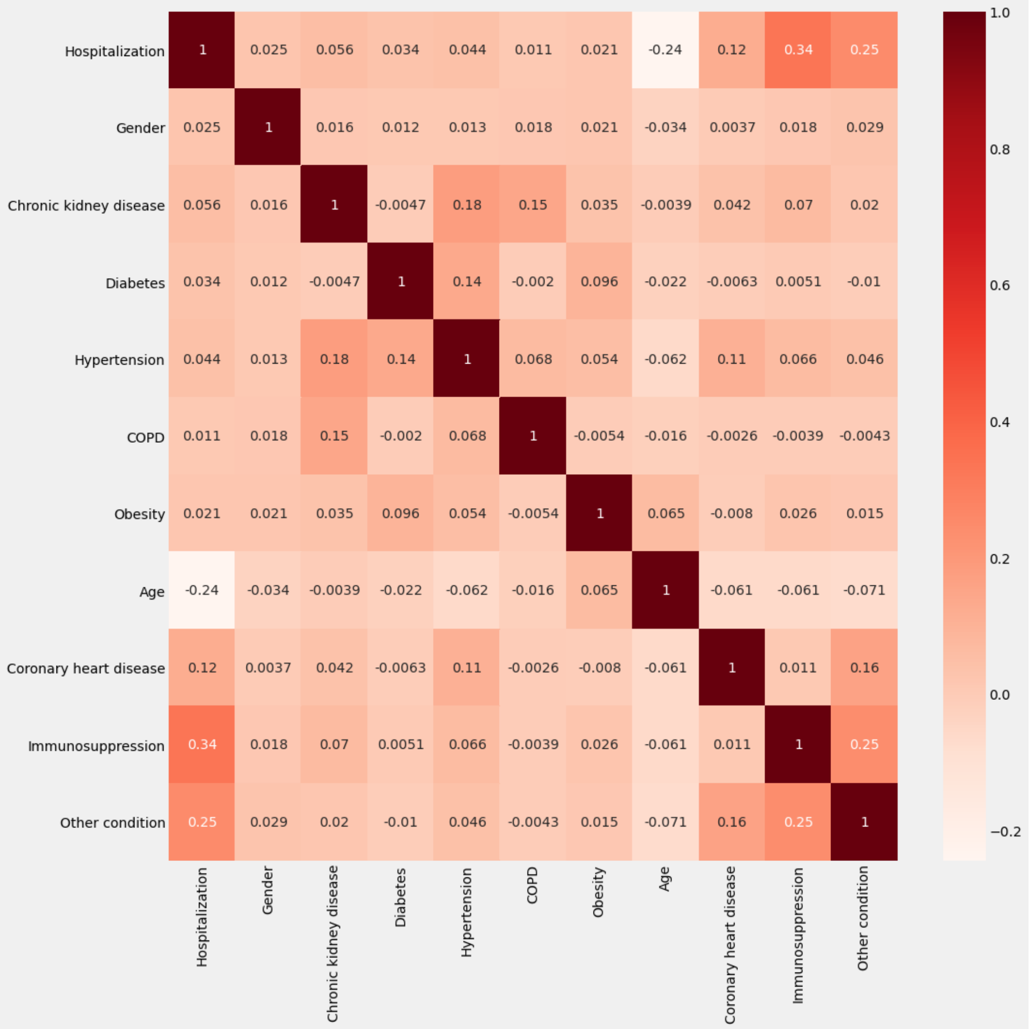
Once we know that the database we are using, behaves as expected from what is known about COVID19 in adults. Now we are going to repeat previous analysis but in children population. We filter the database for children with 15 years or less and with the same analysis as before we are going to check if risk factors for hospitalization in children are the same as per adults. There are 7203 children in the database that have been tested positive for COVID-19, 3435 are girls and 3768 boys. In table 3.2 are the distribution for all clinical variables that we are using.

		Grouped by Hospitalization			
		Overall	0	1	P-Value
n		7203	6695	508	
Gender, n (%)	0	3435 (47.7)	3216 (48.0)	219 (43.1)	0.036
	1	3768 (52.3)	3479 (52.0)	289 (56.9)	
Age, n (%)	0	386 (5.4)	233 (3.5)	153 (30.1)	< 0.001
	1	282 (3.9)	237 (3.5)	45 (8.9)	
	10	490 (6.8)	469 (7.0)	21 (4.1)	
	11	523 (7.3)	505 (7.5)	18 (3.5)	
	12	607 (8.4)	584 (8.7)	23 (4.5)	
	13	759 (10.5)	727 (10.9)	32 (6.3)	
	14	793 (11.0)	768 (11.5)	25 (4.9)	
	15	929 (12.9)	904 (13.5)	25 (4.9)	
	2	226 (3.1)	200 (3.0)	26 (5.1)	
	3	226 (3.1)	197 (2.9)	29 (5.7)	
	4	252 (3.5)	227 (3.4)	25 (4.9)	
	5	262 (3.6)	236 (3.5)	26 (5.1)	
	6	327 (4.5)	315 (4.7)	12 (2.4)	
	7	338 (4.7)	316 (4.7)	22 (4.3)	
	8	378 (5.2)	362 (5.4)	16 (3.1)	
	9	425 (5.9)	415 (6.2)	10 (2.0)	
Diabetes, n (%)	0	7167 (99.5)	6666 (99.6)	501 (98.6)	0.012
	1	36 (0.5)	29 (0.4)	7 (1.4)	
COPD, n (%)	0	7197 (99.9)	6690 (99.9)	507 (99.8)	0.355
	1	6 (0.1)	5 (0.1)	1 (0.2)	
Asma, n (%)	0	6991 (97.1)	6488 (96.9)	503 (99.0)	0.010
	1	212 (2.9)	207 (3.1)	5 (1.0)	
Immunosuppression, n (%)	0	7072 (98.2)	6656 (99.4)	416 (81.9)	< 0.001
	1	131 (1.8)	39 (0.6)	92 (18.1)	
Hypertension, n (%)	0	7169 (99.5)	6669 (99.6)	500 (98.4)	0.002
	1	34 (0.5)	26 (0.4)	8 (1.6)	
Other condition, n (%)	0	7050 (97.9)	6619 (98.9)	431 (84.8)	< 0.001
	1	153 (2.1)	76 (1.1)	77 (15.2)	
Coronary heart disease, n (%)	0	7146 (99.2)	6662 (99.5)	484 (95.3)	< 0.001
	1	57 (0.8)	33 (0.5)	24 (4.7)	
Obesity, n (%)	0	6960 (96.6)	6476 (96.7)	484 (95.3)	0.105
	1	243 (3.4)	219 (3.3)	24 (4.7)	
Chronic kidney disease, n (%)	0	7172 (99.6)	6673 (99.7)	499 (98.2)	< 0.001
	1	31 (0.4)	22 (0.3)	9 (1.8)	

Table 3.2: *Mexican database children clinical variables.*

CORRELATION ANALYSIS

Analyzing the first row of the correlation matrix 3.5, this is, the correlation of the features with the target variable, we can see that many of them are very little correlated, age is correlated but negatively. Only inmunosupression seam to be strongly correlated with COVID-19 hospitalization in children.



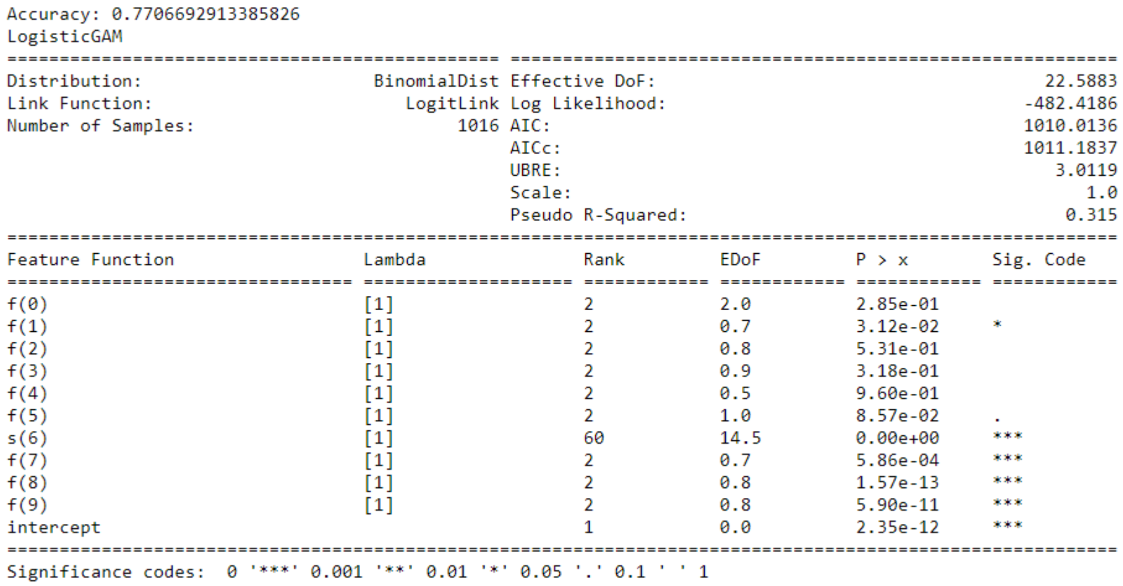
	Hospitalization	Gender	Chronic kidney disease	Diabetes	Hypertension	COPD	Obesity	Age	Coronary heart disease	Immunosuppression	Other condition
Hospitalization	1	0.025	0.056	0.034	0.044	0.011	0.021	-0.24	0.12	0.34	0.25
Gender	0.025	1	0.016	0.012	0.013	0.018	0.021	-0.034	0.0037	0.018	0.029
Chronic kidney disease	0.056	0.016	1	-0.0047	0.18	0.15	0.035	-0.0039	0.042	0.07	0.02
Diabetes	0.034	0.012	-0.0047	1	0.14	-0.002	0.096	-0.022	-0.0063	0.0051	-0.01
Hypertension	0.044	0.013	0.18	0.14	1	0.068	0.054	-0.062	0.11	0.066	0.046
COPD	0.011	0.018	0.15	-0.002	0.068	1	-0.0054	-0.016	-0.0026	-0.0039	-0.0043
Obesity	0.021	0.021	0.035	0.096	0.054	-0.0054	1	0.065	-0.008	0.026	0.015
Age	-0.24	-0.034	-0.0039	-0.022	-0.062	-0.016	0.065	1	-0.061	-0.061	-0.071
Coronary heart disease	0.12	0.0037	0.042	-0.0063	0.11	-0.0026	-0.008	-0.061	1	0.011	0.16
Immunosuppression	0.34	0.018	0.07	0.0051	0.066	-0.0039	0.026	-0.061	0.011	1	0.25
Other condition	0.25	0.029	0.02	-0.01	0.046	-0.0043	0.015	-0.071	0.16	0.25	1

Figure 3.5: Correlation matrix for children risks factors features in severe COVID-19.

GENERALIZED ADDITIVE MODELS (GAM)

As we have done in previous section, we are going to apply Logistic GAM model to children data, to find which variables can explain hospitalization and how. We are going to use the same features as before with the objective of compare if the behavior is the same as in adults. The model is the same that we have applied before for adults but for children. In figure 3.6 is the summary of the model.

It outputs nearly a 78% of accuracy. But now we can observe, looking at p-value column, that many of the features are not significant. This is, they is not a strong relationship between



```
Accuracy: 0.7706692913385826
LogisticGAM
=============================================== ==========================================================
Distribution:                      BinomialDist Effective DoF:                                     22.5883
Link Function:                        LogitLink Log Likelihood:                                  -482.4186
Number of Samples:                         1016 AIC:                                             1010.0136
                                                AICc:                                            1011.1837
                                                UBRE:                                               3.0119
                                                Scale:                                                 1.0
                                                Pseudo R-Squared:                                    0.315
==========================================================================================================
Feature Function                  Lambda               Rank         EDoF         P > x        Sig. Code   
================================= ==================== ============ ============ ============ ============
f(0)                              [1]                  2            2.0          2.85e-01                 
f(1)                              [1]                  2            0.7          3.12e-02     *           
f(2)                              [1]                  2            0.8          5.31e-01                 
f(3)                              [1]                  2            0.9          3.18e-01                 
f(4)                              [1]                  2            0.5          9.60e-01                 
f(5)                              [1]                  2            1.0          8.57e-02     .           
s(6)                              [1]                  60           14.5         0.00e+00     ***         
f(7)                              [1]                  2            0.7          5.86e-04     ***         
f(8)                              [1]                  2            0.8          1.57e-13     ***         
f(9)                              [1]                  2            0.8          5.90e-11     ***         
intercept                                              1            0.0          2.35e-12     ***         
==========================================================================================================
Significance codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3.6: Logistic GAM model summary for children risks factors in severe COVID-19 .

all features and the response variable.

PARTIAL DEPENDENCY PLOTS

Let´s analyze features importance in partial dependency plots 3.7 and 3.8

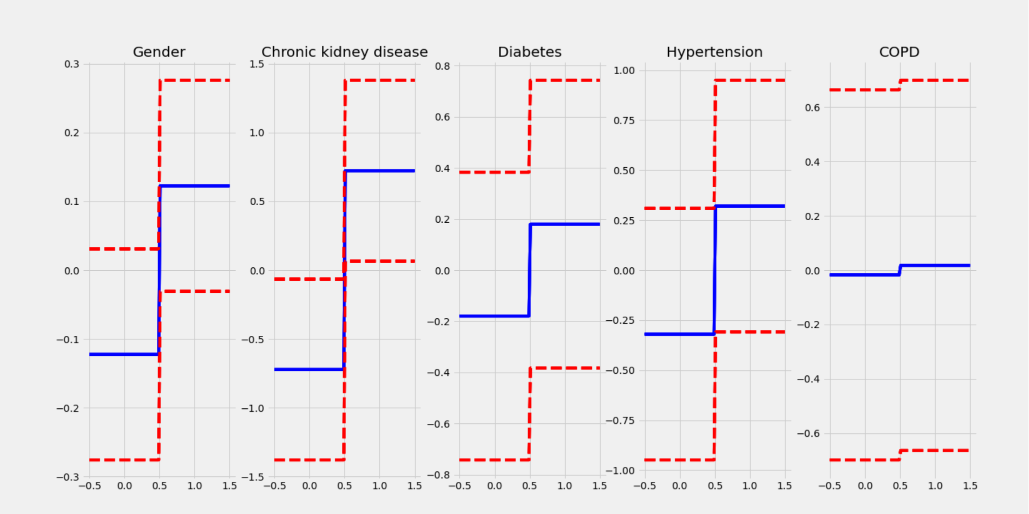


Figure 3.7: Partial dependence plots for children risks factors in severe COVID-19 I.

Partial dependence plots are very different from those of the previous section. Red lines are the confidence interval that now is very large for many features, this means that ones are not significant.

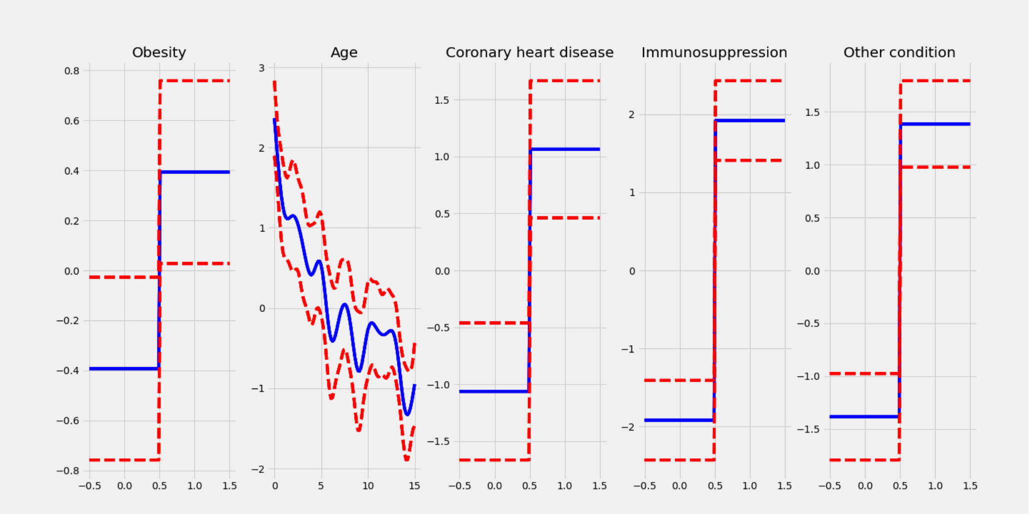


Figure 3.8: Partial dependence plots for children risks factors in severe COVID-19 II.

In adults we have seen that all features except coronary heart disease are significant to explain hospitalization in COVID-19 cases. However in children only age, inmmunosupresion, coronary heart disease, and having another comorbidity are significant. Age have and inverse relation with hospitalization in children: newborn children have more risk in developing a severe case of COVID-19.

3.2.3 Model validation with Sant Joan de Deu hospitalization data

The objective of this section is to check if model built in previous section can be used in children hospitalization data of San Joan de Deu Hospital. This is, we are going to use SJD data as test data set.

If we apply previous model to SJD Hospital data it outputs a 66% accuracy with Table 3.3 confusion matrix.

$n = 27$	*Predicted NO*	*Predicted YES*
Real No	9	5
Real YES	3	7

Table 3.3: *Confusion Matrix.*

Is not a great accuracy but considering testing dataset is completely different from training one we can say that the model seems to have the ability to adapt fair enough to new, previously unseen, data. As disclaimer, we only have 27 observations of children with COVID in SJD Hospital data, with so few records is not possible to extract significant conclusions regarding the generalization of the model, the model should be tested with more data of SJD hospital or from other hospitals just to be sure that the model generalizes, and there are not biases in the data we have used.

3.3 Discussion

Although is well known that children, in general, develop milder COVID-19 disease, some of them required hospitalization due severe outcomes. We have used the Mexican National Epidemiological Surveillance database, with $n = 7203$ children tested positive for COVID-19, 7,05% ($n = 508$) required hospitalization and 1,58% ($n = 114$) admission to ICU. We have analysed risk factors for severe COVID-19 in children, considering severe cases those who require hospitalization, and compare them with adult's risk factors. Prior to this study, no other research has been performed focus on hospitalization risk factors for pediatric COVID-19 patients, with such a large samples, and with data from March to November.

3.3.1 Contributions

Let us list the most important contributions that we will discuss in more detail below.

- Risk factors for severe COVID-19 are different in children than in adults.
- Age (from 0 to 2 years old), coronary heart disease, immunosuppression and having another comorbidity are the most significant risks factors for hospitalization in children.
- There are not significative differences in developing severe COVID-19 between genders in children.

Children hospitalized due to COVID-19, like adults, have a high incidence of comorbid conditions: 38% ($n = 194$) have at least one comorbidity. We have found important differences in risk factors between adults and children. In children **age, coronary heart disease, immunosuppression and having another comorbidity** (other than chronic kidney disease, diabetes, hypertension, COPD, obesity, coronary heart disease or immunosuppression) **are the most significant risks factors for hospitalization** (p-value < 0.001). Is important to note that age has an inverse relation with hospitalization in children: **youngest children (from 0 to 2 years old) have more risk in developing a severe case of COVID-19**. Also, chronic kidney disease is significant, but with a higher p-value (p-value < 0.01). However, other comorbidities like diabetes, hypertension, COPD or obesity that are risk factors in adults are not in children. It is also important to emphasize that there are not significative differences in developing severe COVID-19 between genders in children. This comes to confirm that risk factors in adults are comorbid conditions often considered acquired or probably related with unhealthy lifestyle but in children are congenital conditions, including developmental delay and genetic anomalies, already hypothesized in [24].

As the disease develops mildly in children, most studies have been focused, logically, on analyzing the disease in adults. In addition, with fewer cases in children and milder, there are not many large databases to work with. This part of the study has tried to fill in this gap with the analysis of a large database of children, since knowing the risk factors in the aggravation of COVID-19 in children seems important to us. First, it is still unknown which long-term sequelae severe COVID-19 could have in children, so understanding better the risk factors in the aggravation could help to protect them better. Second, several western countries will start the vaccination campaign against coronavirus soon, so this information can help to establish priority groups.

3.3.2 Future research

There are a few COVID-19 severe cases in children, that develop a multisystem inflammatory syndrome and require admission to ICU, even in a very few cases with a fatal result. We have tried to find risk factors for these cases, but due to the lack of data we have not found relevant variables. More research in these severe cases should be done to understand why some children develop this syndrome.

Chapter 4

Interferon signature and cytokine response

As we have discussed in section 2.1.5, there is still little knowledge about the role of the IFN pathway in adults with COVID-19, but in children there is a complete ignorance.

The main objectives of this chapter are:

- Analyse the interferon signature in children infected with SARS-CoV-2 and its differences between severity groups.
- Compare interferon signature and cytokine responses in children with the ones from adults.

4.1 Data

Data we are going to use for this part of the project has been collected in Sant Joan de Déu Hospital in Barcelona. Data have been classified into different groups depending on the severity of the disease and split between children and adults. These are the different cohorts:

- Children (< 18 years):
  - Asymptomatic
  - Mild symptoms: cough, fever, odynophagia, gastrointestinal symptoms
  - Moderate symptoms: respiratory distress needing oxygen or high flow nasal cannula or other symptoms which required admission to the hospital
  - Severe symptoms: MIS-C, admitted to the Paediatric Intensive Care Unit (PICU)

- Adults:
  - Asymptomatic: healthy controls
  - Mild symptoms: pregnant or puerperal women with cough, fever, odynophagia, gastrointestinal symptoms
  - Severe symptoms: ARDS, admitted to the ICU

Apart of age based and clinical cohorts, the dataset have 32 variables, measuring the expression of the genes that characterize interferon signature in above cohorts. There are 66 patients with SARS-CoV-2, positive by a PCR test. In table 4.1 we can find the number of observations we have in the dataset by cohort and children/adults.

Age	*Clinical*	*Count*	*Percent*
ADULT	Control	12	18.18%
ADULT	Severe	10	15.15%
ADULT	Moderate	1	1.52%
ADULT	Mild	7	10.61%
ADULT	Asymptomatic	8	12.12%
PEDIATRIC	Severe	4	6.06%
PEDIATRIC	Moderate	5	7.58%
PEDIATRIC	Mild	9	13.64%
PEDIATRIC	Asymptomatic	10	15.15%

Table 4.1: *SJD Hospital cohorts distribution.*

In table 4.2 there is a summary of all variables available grouped by age and the cohorts depending on the severity of the disease.

	Missing	Overall	Adult Asymptomatic	Control Group	Adult Severe	Adult Mild	Adult Moderate	Pediatric Asymptomatic	Pediatric Severe	Pediatric Mild	Pediatric Moderate	P-Value
n		66	8	12	10	7	1	10	4	9	5	
CXCL10, mean (SD)	0	4.4 (14.4)	4.5 (10.5)	0.7 (1.2)	10.5 (25.4)	0.2 (1.5)	33.4 (0.0)	1.5 (2.5)	1.9 (1.5)	9.1 (25.0)	0.9 (1.3)	0.344
DDX60, mean (SD)	0	1.3 (3.9)	1.1 (3.1)	-0.1 (0.3)	2.6 (6.0)	0.3 (1.3)	12.3 (0.0)	0.9 (2.6)	2.0 (4.3)	2.4 (5.8)	-0.4 (0.5)	0.098
EPSTI1, mean (SD)	0	2.3 (4.7)	2.0 (3.3)	-0.2 (0.4)	5.3 (8.0)	1.1 (2.0)	16.4 (0.0)	1.4 (2.0)	4.0 (5.6)	2.9 (5.5)	1.5 (1.1)	0.013
GBP1, mean (SD)	0	2.2 (6.5)	5.6 (13.8)	0.2 (0.9)	4.6 (7.7)	0.5 (2.2)	14.2 (0.0)	-0.2 (0.5)	5.7 (7.9)	1.0 (3.8)	0.9 (2.0)	0.153
HERC5, mean (SD)	0	0.4 (2.5)	0.4 (2.0)	-0.4 (0.3)	1.3 (3.7)	-0.3 (0.6)	6.5 (0.0)	-0.2 (0.8)	0.5 (2.0)	1.1 (4.4)	-0.5 (0.2)	0.182
HERC6, mean (SD)	0	1.3 (4.7)	-0.0 (0.7)	0.2 (0.4)	2.7 (7.7)	0.3 (1.8)	16.5 (0.0)	1.0 (4.0)	-0.3 (1.0)	3.0 (7.2)	-0.4 (0.8)	0.046
IFI27, mean (SD)	0	14.3 (51.4)	1.1 (3.1)	-0.3 (0.2)	45.4 (112.9)	8.1 (22.4)	95.3 (0.0)	8.5 (22.6)	0.9 (0.9)	24.5 (57.6)	4.2 (5.7)	0.379
IFI44, mean (SD)	0	1.2 (3.8)	1.2 (3.1)	-0.2 (0.2)	3.2 (6.5)	0.2 (1.0)	13.2 (0.0)	0.5 (1.9)	1.6 (3.4)	1.9 (4.9)	-0.1 (0.4)	0.032
IFI44L, mean (SD)	0	1.6 (5.5)	0.5 (1.9)	-0.4 (0.2)	4.0 (9.1)	0.1 (1.2)	21.6 (0.0)	1.2 (3.7)	1.2 (3.0)	3.2 (7.8)	-0.2 (0.4)	0.009
IFI6, mean (SD)	0	1.1 (4.3)	0.1 (1.4)	-0.5 (0.4)	2.9 (6.6)	-0.2 (1.2)	13.4 (0.0)	0.6 (3.0)	3.3 (6.5)	2.1 (6.0)	-0.3 (0.6)	0.045
IFIT1, mean (SD)	0	1.0 (3.9)	0.3 (1.1)	-0.2 (0.3)	2.3 (5.9)	0.2 (1.2)	14.0 (0.0)	0.2 (1.5)	0.8 (2.1)	2.4 (7.0)	-0.2 (0.4)	0.033
IFIT2, mean (SD)	0	1.3 (4.2)	2.2 (4.2)	0.4 (1.2)	2.7 (6.1)	-0.1 (0.9)	7.7 (0.0)	-0.2 (1.0)	3.1 (6.2)	2.4 (7.3)	-0.3 (0.2)	0.437
IFIT3, mean (SD)	0	1.0 (3.8)	1.1 (3.1)	-0.3 (0.4)	3.0 (6.4)	0.2 (1.4)	13.8 (0.0)	-0.2 (1.0)	2.0 (4.3)	1.4 (4.6)	-0.2 (0.6)	0.015
IFIT5, mean (SD)	0	0.8 (2.9)	1.2 (2.4)	-0.1 (0.5)	1.9 (5.0)	0.3 (1.2)	10.1 (0.0)	-0.0 (1.2)	1.2 (3.0)	1.1 (3.2)	-0.6 (0.3)	0.031
ISG15, mean (SD)	0	1.7 (6.4)	0.3 (1.0)	-0.3 (0.3)	3.4 (8.6)	0.4 (1.9)	18.0 (0.0)	1.0 (3.1)	0.6 (2.1)	5.5 (12.8)	-0.1 (0.5)	0.104
LAMP3, mean (SD)	0	2.3 (10.4)	2.3 (7.0)	0.0 (0.5)	2.5 (6.3)	-0.0 (0.9)	10.6 (0.0)	1.4 (3.2)	-0.0 (1.1)	9.9 (26.2)	-0.5 (0.4)	0.558
LY6E, mean (SD)	0	1.6 (5.4)	-0.2 (0.6)	-0.5 (0.3)	4.5 (9.6)	0.5 (2.0)	22.7 (0.0)	1.1 (2.7)	2.5 (5.5)	2.3 (5.5)	0.4 (1.3)	0.002
MX1, mean (SD)	0	0.8 (3.1)	0.5 (1.6)	-0.3 (0.4)	1.5 (4.0)	0.1 (0.9)	7.7 (0.0)	0.6 (2.0)	0.5 (1.5)	2.3 (6.2)	-0.4 (0.3)	0.232
OAS1, mean (SD)	0	1.3 (3.6)	1.1 (2.8)	-0.2 (0.7)	3.6 (6.0)	0.7 (1.7)	11.2 (0.0)	0.4 (2.0)	2.3 (4.7)	1.6 (4.4)	0.0 (0.6)	0.041
OAS2, mean (SD)	0	2.2 (7.1)	0.5 (2.4)	-0.1 (0.4)	4.0 (9.9)	0.3 (1.8)	25.0 (0.0)	2.4 (5.3)	1.5 (3.9)	5.4 (12.2)	-0.5 (0.9)	0.030
OAS3, mean (SD)	0	1.1 (3.7)	0.8 (2.0)	-0.3 (0.2)	2.2 (5.5)	0.3 (1.1)	13.8 (0.0)	0.7 (2.1)	1.2 (2.9)	2.3 (6.2)	-0.2 (0.4)	0.029
OASL, mean (SD)	0	0.8 (3.4)	0.2 (1.3)	-0.7 (0.3)	3.1 (5.0)	-0.0 (0.9)	13.9 (0.0)	-0.2 (1.4)	1.5 (3.0)	1.4 (5.0)	0.2 (0.4)	0.001
RSAD2, mean (SD)	0	1.6 (5.6)	0.7 (2.1)	-0.3 (0.3)	4.1 (9.6)	0.4 (1.7)	19.0 (0.0)	0.4 (1.7)	2.1 (4.3)	3.3 (8.9)	-0.1 (0.3)	0.047
RTP4, mean (SD)	0	1.1 (5.2)	0.8 (4.3)	-0.6 (0.4)	3.2 (8.4)	-0.7 (1.3)	19.5 (0.0)	0.1 (1.6)	2.5 (5.5)	2.1 (6.7)	-0.4 (1.1)	0.012
SIGLEC1, mean (SD)	0	2.4 (8.4)	0.1 (1.0)	-0.2 (0.3)	5.5 (12.9)	0.8 (3.2)	35.3 (0.0)	2.1 (6.5)	-0.5 (0.3)	5.1 (11.9)	0.1 (1.6)	0.003
SOCS1, mean (SD)	0	0.8 (4.4)	1.0 (5.3)	-1.0 (0.7)	2.4 (5.7)	-1.1 (0.6)	7.3 (0.0)	-0.4 (1.0)	4.6 (6.8)	1.7 (7.0)	1.1 (1.6)	0.199
SPATS2L, mean (SD)	0	0.7 (4.8)	0.1 (1.7)	-0.9 (0.6)	3.1 (7.5)	-0.4 (1.1)	22.1 (0.0)	-0.3 (2.1)	-0.2 (2.7)	1.8 (6.1)	-0.8 (1.1)	< 0.001
USP18, mean (SD)	0	4.0 (15.9)	-0.2 (0.5)	-0.5 (0.3)	5.9 (14.2)	0.2 (1.6)	34.4 (0.0)	4.2 (9.9)	-0.0 (1.1)	15.1 (37.0)	-0.2 (1.0)	0.214
CIITA, mean (SD)	0	-1.0 (1.2)	-1.2 (0.7)	-1.0 (0.5)	-1.8 (1.3)	-1.3 (0.8)	-2.2 (0.0)	0.2 (1.7)	-1.7 (0.3)	-0.6 (1.1)	-1.3 (0.9)	0.007
CXCL9, mean (SD)	0	0.7 (2.2)	2.1 (3.8)	-0.1 (0.8)	2.3 (3.3)	-0.3 (0.7)	1.1 (0.0)	0.0 (1.3)	-0.0 (0.2)	0.3 (1.0)	0.8 (0.8)	0.067
IFNA2, mean (SD)	0	-0.2 (1.5)	-0.3 (1.1)	-0.5 (0.6)	0.6 (2.6)	-0.7 (0.4)	5.1 (0.0)	-0.8 (0.5)	0.1 (0.5)	0.1 (2.1)	-0.4 (0.9)	0.017
STAT1, mean (SD)	0	1.2 (3.5)	2.8 (6.5)	-0.2 (0.9)	2.9 (3.9)	0.1 (2.2)	8.5 (0.0)	0.0 (1.3)	2.5 (3.4)	1.0 (3.7)	-0.1 (1.3)	0.090

Table 4.2: *Cytokines summary statistics grouped by age and cohorts.*

4.2 Analysis of cytokines response between mild and severe symptoms in children

Dataset provided and collected by medical team of SJD Hospital has, apart of the age and the cohorts, 32 variables as detailed in Table 4.2, all these variables are cytokines scored with the z-score over healthy controls. Our first objective is to understand they behavior through a correlation matrix. Then we will select the most significant ones to analyze the differences in cytokine response between mild and severe symptoms in children. Since we don´t have many observations, we will group severe and moderate in one category and mild and asymptomatic in another. The goal is to find differences in the immune response between these two groups.

4.2.1 Correlation analysis

The analysis of correlation in Figure 4.1 is very interesting. The dark red indicate correlations nearly to 1, this is, the variables have a behaviour similar between them. We can see roughly only dark red squares; this means that almost all variables behaves remarkably similar. To create descriptive or predictive models we should select features not very correlated, because if the are correlated the model will be more complex and no more information will be added. Having this correlation matrix we can select the feature that better explain the target variable and some of the variables not in dark red if they are important to explain the target.

4.2.2 Feature selection

Feature selection or feature importance techniques are one of the key concepts in all data science projects. Feature Selection is the process for select those features which contribute most to target variable. These techniques allow us to reduce overfitting, improve accuracy and reduce training time because having redundant or irrelevant data as input of our model increase the probability of making decision based on noise. Feature importance scores can be calculated for problems that involve predicting a numerical value, called regression, and those problems that involve predicting a class label, called classification. Our case is a classification problem, we want to classify children in severity cohorts by its biomarkers. As we are going to use a decision tree model, we will use CART Decision Tree Classifier implemented in python scikit-learn library. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples.

By applying this method of feature selection, we get the two cytokines most important to

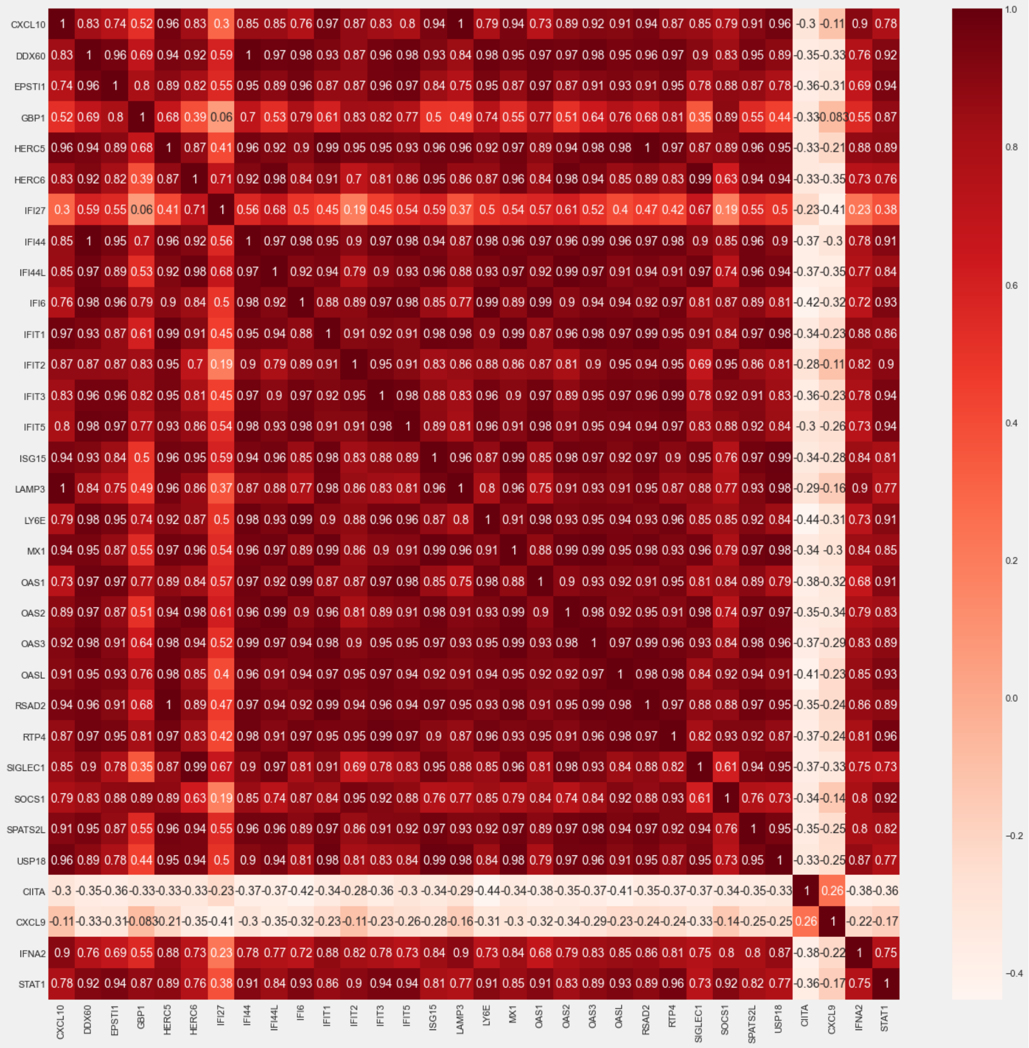


Figure 4.1: Correlation matrix for cytokines.

explain the differences between severe and mild cases are SOCS1 and CITTA. As we have seen in section 4.2.1 SOCS1 behaves like the other cytokines, only CIITA have a negative correlation and a different behavior. In figure 4.2.2 box plots for these two features and the study groups.

In this box plots we observe that apparently low levels of CITTA are related with severe cases of COVID-19, however SOCS1 behaves the other way round, higher levels seem to be related with severe cases. If we change SOCS1 by any of the other cytokines the plot will look very similar due the high correlation between them.

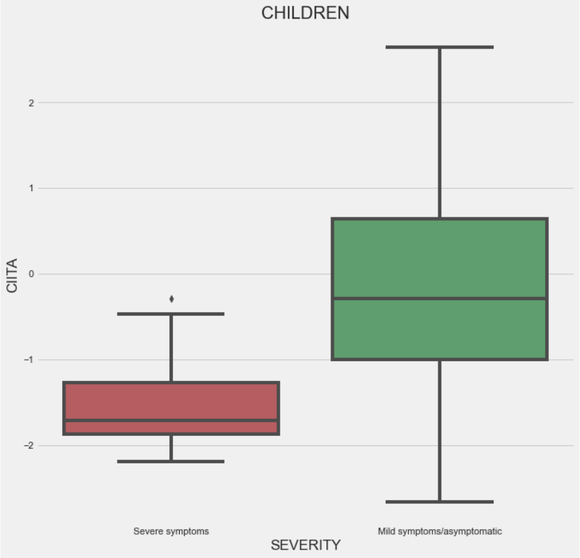


Figure 4.2: Box plot for CIITA

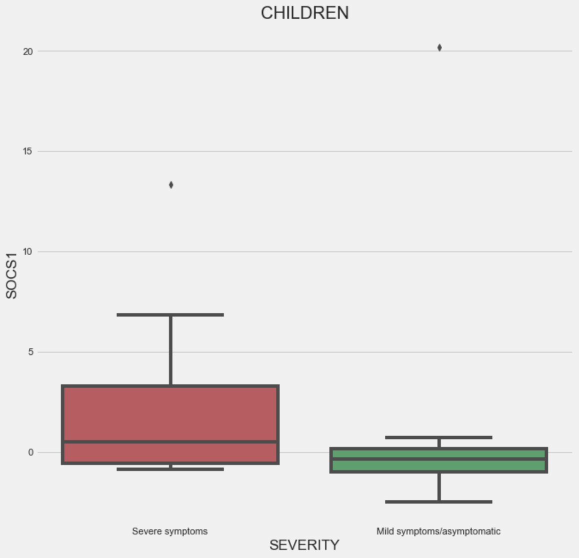


Figure 4.3: Box plot for SOCS1

4.2.3 Differences in cytokines response between mild and severe symptoms

Now, we are going to apply a decision tree method, these methods have been introduced in section 2.3.2. The objective is to classify using both features selected previously (CITTA and SOCS1) severe and mild cases. One of the key advantages of decision trees is that they have a natural visualization. In 4.4 our resultant tree classifies properly 27 from the 28 observations, this is a 96% of accuracy. As we have only 28 observations, to avoid as much as possible overfitting, the maximum depth of the tree has been set to 2, also he minimum number of samples required to be at a leaf node has also fixed to 3 and only two features have been used.

The interpretation of the tree is very simple: if CITTA is between -2.375 and -1.205 or CIITA is greater than -1.205 and SOC1 greater than 0.45 child with COVID is classified as severe. In other case child will be classified as mild. We are going to plot this decision tree in a bidimensional figure 4.5 to clarify it more. We can see that all severe and moderate cases are inside the red region that define the tree. This confirms what we have seen previously: low levels of CIITA and high levels of SOCS1 seem to be indicators of severity. Is also important to note that asymptomatic children seam to have high levels of CIITA, this can indicate that can provide some kind of protection against the virus.

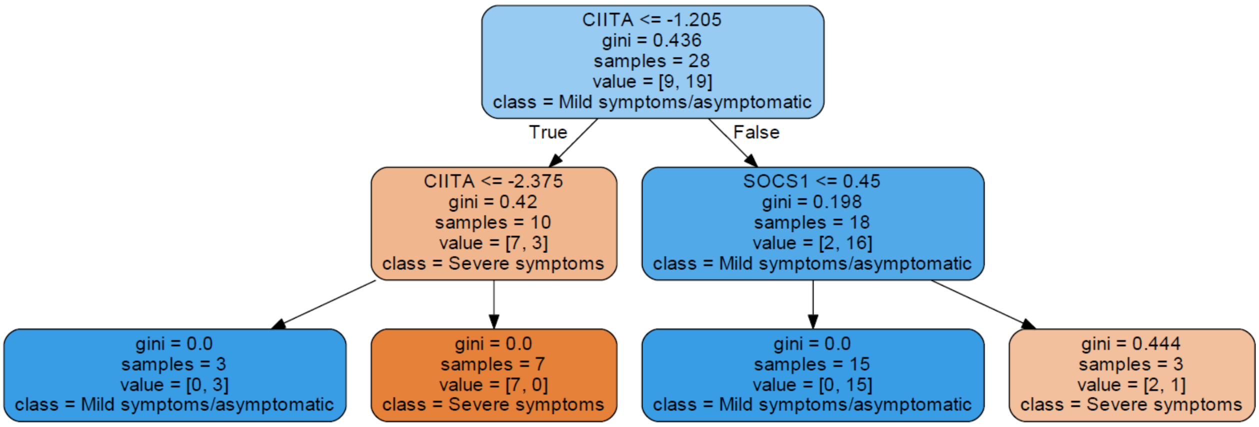


Figure 4.4: Decision tree for classification severe COVID-19 in children

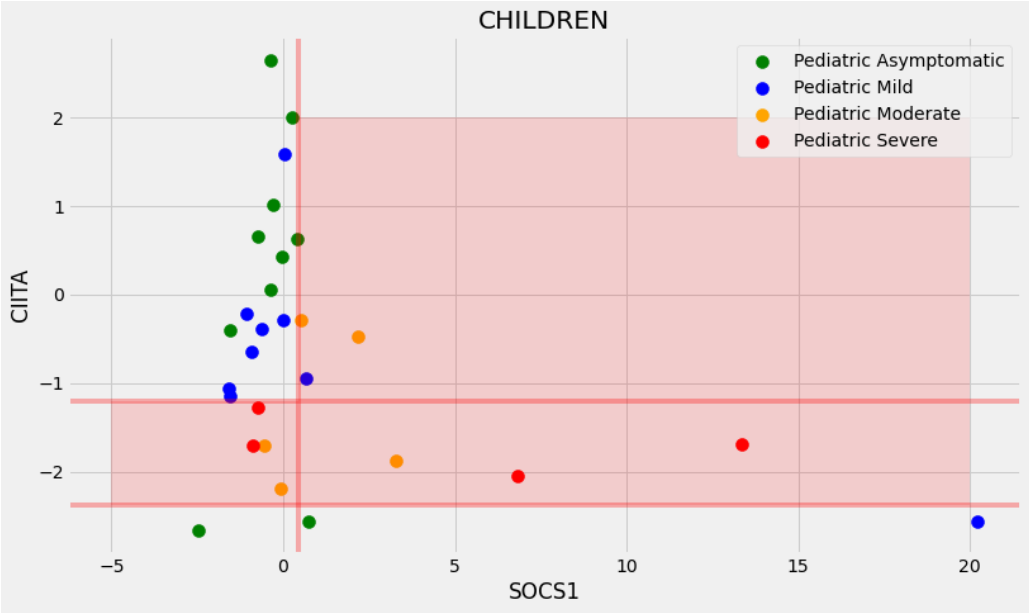


Figure 4.5: Decision tree space partition for classification severe COVID-19 in children

4.3 Compare cytokines responses between adults and children

In Figure 4.6 is the same partition of the space for adult population of the dataset. For adults we can see that the behavior is very different. In the red region defined by the tree for severe cases in children we can´t find the severe adult cases. So we can conclude that the behavior in adults for this to cytokines is different than in children. Also important to note that CIITA values are lower in adults than in children.

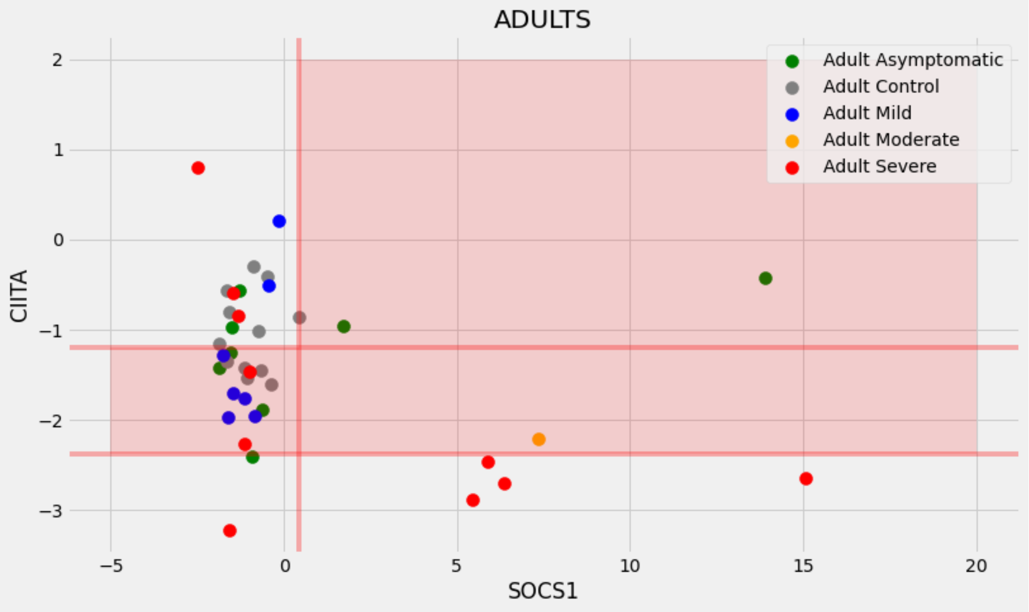


Figure 4.6: Decision tree space partition using previous classification in adults.

4.4 Discussion

There is still little knowledge about the role of the IFN pathway in children with COVID-19, in this chapter we analyzed interferon signature in children. We have used a database collected in Hospital San Joan de Deu with $n = 66$ observations, $n_{children} = 28$ are pediatric patients with COVID-19, $n_{adults} = 26$ and $n_{control} = 12$ cases. For all the observations we have 32 features regarding interferon signature, measured by z-score over healthy controls.

4.4.1 Contributions

Let us list the most important contributions that we will discuss in more detail below.

- All genes that describe interferon signature are very correlated between them.
- CIITA and SOCS1 genes are the most relevant to explain COVID-19 severity in children. These variables have an inverse relation. We can use them to classify cases and predict aggravation. These behavior is different in adults.
- Children seem to have higher levels of CIITA than adults, these high levels are related with mild and asymptomatic cases, so CIITA could protect us from virus entrance.

First, we have found that all interferon-stimulated genes that define interferon signature are very correlated between them. This is, when a patient has high levels of an indicator the others are also high and vice versa. SOCS1 (suppressor of cytokine signaling 1) is the most important feature for explain severity. CIITA (major histocompatibility complex class II transactivator) gene, that is not part of interferon signature, has an inverse correlation with interferon signature variables.

So, using a decision tree with SOCS1 and CIITA, we have found that high values of SOCS1 and low levels of CITTA is an indicator of severe cases of COVID-19. This inverse correlation between SCOS1 and CIITA is a known way of control inflammation described for other diseases as multiple sclerosis [25]. Both indicators could be used together to predict severe cases of COVID-19 in children.

Secondly, CIITA and SOCS1 behavior in adults is different than in children. High values of SOCS1 and low levels of CITTA don´t explain infection severity in adults as well as they do for children.

Finally, we have found that CIITA levels are statistically significant different ($p - value$ 0.009) between ages: children have higher levels of CIITA than adults. It's not actually really clear if CIITA levels decrease with age, but there is a meta-study that goes in that direction [26]. We have found remarkably interesting that children with higher levels of CIITA coincide with mild and asymptomatic cases and that these levels are higher than in adults. CIITA could explain less frequency and less severity observed in children. **This result comes to confirm what has been hypothesized in [27], where authors have found that in in-vitro Ebola virus CIITA prevents viral fusion and entry, and they hypothesized that this behavior could be the same for coronavirus**. This results could also open a new research path in treatments against COVID-19, as highlined in [27]: therapeutic strategies aimed to increase CIITA may be particularly suitable to reduce brain damage and to prevent long-term neurological symptoms observed in COVID-19.

4.4.2 Future research

Fortunately, as we discussed throughout the project, few children develop severe cases of COVID-19, so we have worked with a small number of observations. The results obtained in this study should be confirmed with other studies with more samples. Regarding therapeutic strategies to increase artificially CIITA levels to try to prevent virus entrance, a lot of work should be done, and side effects should be studied.

We have also seen that children with MIS-C seem to have high levels of cytokines response, probably higher than in adults, but as we only have 3 samples in our database the outcomes are not significative.

Appendices

Appendix A

Code

Code for this whole project have been written in python. For commodity I have used Jupyter notebooks. These notebooks, one for risk factors (Chapter 3) and another one for interferon signature analysis (Chapter 4) can be found in my GitHub:

https://github.com/guille-arguello/covid-children-tfm

Bibliography

[1] World Health Organzation. WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/table, 10 2020.

[2] Ministerio de Sanidad. Gobierno de España. Estudio Nacional de Sero-Epidemiología de la Infección por SARS-CoV-2 en España. Informe Final. https://www.mscbs.gob.es/ciudadanos/ene-covid/docs/ESTUDIO_ENE-COVID19_INFORME_FINAL.pdf, 7 2020.

[3] Bi Q, Wu Y, Mei S, Ye C, Zou X, and Zhang Z et al. Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen china: a retrospective cohort study. *Lancet Infect Dis.*, 20, August 2020.

[4] Pilar Storch de Gracia, Inés Leoz-Gordillo, David Andina, Patricia Flores, Enrique Villalobos, Silvia Escalada-Pellitero, and Raquel Jiménez. Espectro clínico y factores de riesgo de enfermedad complicada en niños ingresados con infección por sars-cov-2. *Anales de Pediatría*, 2020.

[5] Barcelona Sant Joan de Déu Hospital. Project kids corona. a platform to understand covid-19 in children and pregnancy, 2020.

[6] M. Victoria et al Moreno-Arribas. Una visión global de la pandemia covid-19: qué sabemos y qué estamos investigando desde el csic. Consejo Superior de Investigaciones Científicas (España), 2020.

[7] R. González Cortés, A. García-Salido, D Roca Pascual, and et al. A multicenter national survey of children with sars-cov-2 infection admitted to spanish pediatric intensive care units. *Intensive Care Med*, 2020.

[8] Yuanyuan Dong, Xi Mo, Yabin Hu, Xin Qi, Fan Jiang, Zhongyi Jiang, and Shilu Tong. Epidemiology of covid-19 among children in china. *Pediatrics*, 145(6), 2020.

[9] Alessandro Mantovani, Elisabetta Rinaldi, Chiara Zusi, Giorgia Beatrice, Marco Deganello Saccomani, and Andrea Dalbeni. Coronavirus disease 2019 (covid-19) in children and/or adolescents: a meta-analysis. *Pediatric Research*, pages 1–6, 06 2020.

[10] Sophia Tsabouri, Alexandros Makis, Chrysoula Kosmeri, and Ekaterini Siomou. Special article: Risk factors for severity in children with coronavirus-19 disease (covid-19): A comprehensive literature review. *Pediatric Clinics of North America*, 07 2020.

[11] World Health Organization. Multisystem inflammatory syndrome in children and adolescents temporally related to covid-19, 2020.

[12] Lucio Verdoni, Angelo Mazza, Annalisa Gervasoni, Laura Martelli, Maurizio Ruggeri, Matteo Ciuffreda, Ezio Bonanomi, and Lorenzo D'Antiga. An outbreak of severe kawasaki-like disease at the italian epicentre of the sars-cov-2 epidemic: an observational cohort study. *The Lancet*, 395, 05 2020.

[13] Ugo Bastolla. Mathematical model of sars-cov-2 propagation versus ace2 fits covid-19 lethality across age and sex and predicts that of sars, supporting possible therapy, 2020.

[14] Fabio Midulla, Luca Cristiani, and Enrica Mancino. Will children reveal their secret? the coronavirus dilemma. *European Respiratory Journal*, 55(6), 2020.

[15] Asociación Española de Pediatria (AEP). Manual de vacunas en línea de la aep. capítulo 46 – inmunología y vacunas. adyuvantes: moduladores de la inmunidad innata., 2019.

[16] Qian Zhang, Paul Bastard, Zhiyong Liu, Jérémie Le Pen, Marcela Moncada-Velez, Jie Chen, Masato Ogishi, Ira K. D. Sabli, Stephanie Hodeib, Cecilia Korol, Jérémie Rosain, Kaya Bilguvar, Junqiang Ye, Alexandre Bolze, Benedetta Bigio, Rui Yang, Andrés Augusto Arias, Qinhua Zhou, Yu Zhang, Fanny Onodi, Sarantis Korniotis, Léa Karpf, Quentin Philippot, Marwa Chbihi, Lucie Bonnet-Madin, Karim Dorgham, Nikaïa Smith, William M. Schneider, Brandon S. Razooky, Hans-Heinrich Hoffmann, Eleftherios Michailidis, Leen Moens, Ji Eun Han, Lazaro Lorenzo, Lucy Bizien, Philip Meade, Anna-Lena Neehus, Aileen Camille Ugurbil, Aurélien Corneau, Gaspard Kerner, Peng Zhang, Franck Rapaport, Yoann Seeleuthner, Jeremy Manry, Cecile Masson, Yohann Schmitt, Agatha Schlüter, Tom Le Voyer, Taushif Khan, Juan Li, Jacques Fellay, Lucie Roussel, Mohammad Shahrooei, Mohammed F. Alosaimi, Davood Mansouri, Haya Al-Saud, Fahd Al-Mulla, Feras Almourfi, Saleh Zaid Al-Muhsen, Fahad Alsohime, Saeed Al Turki, Rana Hasanato, Diederik van de Beek, Andrea Biondi, Laura Rachele Bettini, Mariella DAngio, Paolo Bonfanti, Luisa Imberti, Alessandra Sottini, Simone Paghera, Eugenia Quiros-Roldan, Camillo Rossi, Andrew J. Oler, Miranda F. Tompkins, Camille Alba, Isabelle Vandernoot, Jean-Christophe Goffard, Guillaume Smits, Isabelle Migeotte, Filomeen Haerynck, Pere Soler-Palacin, Andrea Martin-Nalda, Roger Colobran, Pierre-Emmanuel Morange, Sevgi Keles, Fatma Cölkesen, Tayfun Ozcelik, Kadriye Kart Yasar, Sevtap Senoglu, ¸Semsi Nur

Karabela, Carlos Rodríguez Gallego, Giuseppe Novelli, Sami Hraiech, Yacine Tandjaoui-Lambiotte, Xavier Duval, Cédric Laouénan, Andrew L. Snow, Clifton L. Dalgard, Joshua Milner, Donald C. Vinh, Trine H. Mogensen, Nico Marr, András N. Spaan, Bertrand Boisson, Stéphanie Boisson-Dupuis, Jacinta Bustamante, Anne Puel, Michael Ciancanelli, Isabelle Meyts, Tom Maniatis, Vassili Soumelis, Ali Amara, Michel Nussenzweig, Adolfo García-Sastre, Florian Krammer, Aurora Pujol, Darragh Duffy, Richard Lifton, Shen-Ying Zhang, Guy Gorochov, Vivien Béziat, Emmanuelle Jouanguy, Vanessa Sancho-Shimizu, Charles M. Rice, Laurent Abel, Luigi D. Notarangelo, Aurélie Cobat, Helen C. Su, and Jean-Laurent Casanova. Inborn errors of type i ifn immunity in patients with life-threatening covid-19. *Science*, 2020.

[17] Paul Bastard, Lindsey B. Rosen, Qian Zhang, Eleftherios Michailidis, Hans-Heinrich Hoffmann, Yu Zhang, Karim Dorgham, Quentin Philippot, Jérémie Rosain, Vivien Béziat, Jérémy Manry, Elana Shaw, Liis Haljasmägi, Pärt Peterson, Lazaro Lorenzo, Lucy Bizien, Sophie Trouillet-Assant, Kerry Dobbs, Adriana Almeida de Jesus, Alexandre Belot, Anne Kallaste, Emilie Catherinot, Yacine Tandjaoui-Lambiotte, Jeremie Le Pen, Gaspard Kerner, Benedetta Bigio, Yoann Seeleuthner, Rui Yang, Alexandre Bolze, András N. Spaan, Ottavia M. Delmonte, Michael S. Abers, Alessandro Aiuti, Giorgio Casari, Vito Lampasona, Lorenzo Piemonti, Fabio Ciceri, Kaya Bilguvar, Richard P. Lifton, Marc Vasse, David M. Smadja, Mélanie Migaud, Jérome Hadjadj, Benjamin Terrier, Darragh Duffy, Lluis Quintana-Murci, Diederik van de Beek, Lucie Roussel, Donald C. Vinh, Stuart G. Tangye, Filomeen Haerynck, David Dalmau, Javier Martinez-Picado, Petter Brodin, Michel C. Nussenzweig, Stéphanie Boisson-Dupuis, Carlos Rodríguez-Gallego, Guillaume Vogt, Trine H. Mogensen, Andrew J. Oler, Jingwen Gu, Peter D. Burbelo, Jeffrey Cohen, Andrea Biondi, Laura Rachele Bettini, Mariella DAngio, Paolo Bonfanti, Patrick Rossignol, Julien Mayaux, Frédéric Rieux-Laucat, Eystein S. Husebye, Francesca Fusco, Matilde Valeria Ursini, Luisa Imberti, Alessandra Sottini, Simone Paghera, Eugenia Quiros-Roldan, Camillo Rossi, Riccardo Castagnoli, Daniela Montagna, Amelia Licari, Gian Luigi Marseglia, Xavier Duval, Jade Ghosn, John S. Tsang, Raphaela Goldbach-Mansky, Kai Kisand, Michail S. Lionakis, Anne Puel, Shen-Ying Zhang, Steven M. Holland, Guy Gorochov, Emmanuelle Jouanguy, Charles M. Rice, Aurélie Cobat, Luigi D. Notarangelo, Laurent Abel, Helen C. Su, and Jean-Laurent Casanova. Auto-antibodies against type i ifns in patients with life-threatening covid-19. *Science*, 2020.

[18] Blanco-Melo D, Nilsson-Payant BE, Liu WC, Uhl S, Hoagland D, Møller R, Jordan TX, Oishi K, Panis M, Sachs D, Wang TT, Schwartz RE, Lim JK, and Albrecht RA and. Imbalanced host response to sars-cov-2 drives development of covid-19. *Cell*, 2020.

[19] Bradley Efron. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115(530):636–655, 2020.

[20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2017.

[21] Sistema Nacional de Vigilancia Epidemiologica. Gobierno de México. Base del sistema nacional de vigilancia epidemiologica para el seguimiento a posibles casos de covid-19. https://datos.cdmx.gob.mx/explore/dataset/base-covid-sinave/table/, 11 2020.

[22] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiuyang Xu, Shengjin Tu, Yi Zhang, Hua Chen, and Bin Cao. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The Lancet*, 395(10229):1054 – 1062, 2020.

[23] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of statistics 1189-1232*, 2001.

[24] Jonathan Baruch Steinman, Fok Moon Lum, Peggy Pui-Kay Ho, Naftali Kaminski, and Lawrence Steinman. Reduced development of covid-19 in children reveals molecular checkpoints gating pathogenesis illuminating potential therapeutics. *Proceedings of the National Academy of Sciences*, 117(40):24620–24626, 2020.

[25] Rebecca Zuvich, William Bush, Jacob Mccauley, Ashley Beecham, Philip De Jager, Adrian Ivinson, Alastair Compston, David Hafler, Stephen Hauser, Stephen Sawcer, Margaret Pericak-Vance, Lisa Barcellos, Douglas Mortlock, and Jonathan Haines. Interrogating the complex role of chromosome 16p13.13 in multiple sclerosis susceptibility: Independent genetic signals in the ciita-clec16a-socs1 gene complex. *Human molecular genetics*, 20:3517–24, 06 2011.

[26] Alexandra Gyllenberg, Samina Asad, Fredrik Piehl, Maria Swanberg, Leonid Padyukov, B Yserloo, Elizabeth Rutledge, B McNeney, Jinko Graham, M Orho-Melander, Eero Lindholm, Caroline Graff, C Forsell, Kristina Akesson, Mona Landin-Olsson, Annelie Carlsson, Gun Forsander, S Ivarsson, Helena Elding Larsson, and Ingrid Kockum. Age-dependent variation of genotypes in mhc ii transactivator gene (ciita) in controls and association to type 1 diabetes. *Genes and immunity*, 13:632–40, 10 2012.

[27] Rafal Butowt, Krzysztof Pyrc, and Christopher S. von Bartheld. Battle at the entrance gate: Ciita as a weapon to prevent the internalization of sars-cov-2 and ebola viruses. *Signal Transduction and Targeted Therapy*, 5(1), December 2020.