Assessment 1: Research Project Proposal

Predicting the Progress of the Vaccination using Machine Learning

|  |  |
| --- | --- |
| **Name**: Jaishree Pandey | **Supervisor**: Edlira Vakaj |
| **University ID**:20101813 |  |
| **Word Count**: 3431 |  |
| **Date**: 08/04/2021 |  |

Table of Contents

[Table of Figures 3](#_Toc68816508)

[Abstract 4](#_Toc68816509)

[Introduction 4](#_Toc68816510)

[Background 5](#_Toc68816511)

[Aim 5](#_Toc68816512)

[Research Questions 5](#_Toc68816513)

[Ethical Considerations 5](#_Toc68816514)

[Project Timeline 6](#_Toc68816515)

[Methodology 6](#_Toc68816516)

[Literature Review 7](#_Toc68816517)

[Project Design 7](#_Toc68816518)

[Experimentation 7](#_Toc68816519)

[Dataset 8](#_Toc68816520)

[Machine Learning 8](#_Toc68816521)

[Project Evaluation 8](#_Toc68816522)

[Literature Review 9](#_Toc68816523)

[Methodology 9](#_Toc68816524)

[Results 10](#_Toc68816525)

[The role of inflammation and frailty in identifying prognostic risk factors for poor outcome in in-center haemodialysis patients with COVID-19 disease. 11](#_Toc68816526)

[Methodology 12](#_Toc68816527)

[Results 12](#_Toc68816528)

[The effect of COVID-19 on treatment for breast cancer. 12](#_Toc68816529)

[Methodology 12](#_Toc68816530)

[Results 12](#_Toc68816531)

[Specimens in COVID-19: a Systematic Review & Meta-Analysis Comparing saliva and nasopharyngeal swab in detection. 12](#_Toc68816532)

[Methodology 13](#_Toc68816533)

[Results 13](#_Toc68816534)

[Taking care of fragile rectal cancer patients during a pandemic COVID-19: an Italian experience in one-centre. 13](#_Toc68816535)

[Methodology 13](#_Toc68816536)

[Results 13](#_Toc68816537)

[Conclusion 14](#_Toc68816538)

[References 15](#_Toc68816539)

# Table of Figures

[Figure 1 - Project Timeline using Gantt chart 6](#_Toc68810309)

[Figure 2 - Risk Management 6](#_Toc68810310)

[Figure 3 - LR Workflow 7](#_Toc68810311)

[Figure 4 - Characterisation by writers, country, method of research and number of participants Studies analysed (children and adolescents), Brazil(2020) 10](#_Toc68810312)

[Figure 5 - The epidemiological profile of children and teenagers (age group, likely pollution, preexisting disease/condition and hospitalization). 11](#_Toc68810313)

# Abstract

The virus is a threat to global health due to its fast transmission of COVID-19. In order to avoid further COVID-19 dissemination, early identification of suspicious patients will play an important part (Nasiri and Dimitrova, 2021). Coronavirus pandemic (COVID-19) has strongly infected comorbidity and fragility patients who are unable to self-isolate, such as people with hemodialysis (Hendra et al., 2021). COVID-19 has not only affected lives of humans all around the globe, but has also affected businesses and in turn the global economy.

This research uses the prowess of Machine Learning (ML) and Statistical Data Analysis to predict the progress of the development and spread of the Vaccination for COVID-19 around the world. There will also be detailed data visualizations that show the trends associated with the vaccine. ML will be applied to the dataset on COVID-19 collected from Our World in Data and is currently available on [Kaggle](https://www.kaggle.com/gpreda/covid-world-vaccination-progress).

Machine Learning enables researchers and doctors involved in the COVID-19 domain to automate and increase the efficiency of their research significantly by involving computers that can digest and learn data to make predictions and provide statistics based on that data. The output of this research project will be the statistics on the progress of the vaccination for COVID-19 around the world as well as predictions regarding the progress of the vaccine.

**Keywords:** Machine Learning (ML), Statistical Analysis, Our World in Data, COVID-19

# Introduction

In December 2019, the disease in Hubei Province, China, started to develop an unclear respiratory virus which soon posed a threat to global health because of its easy transmission. It was classified as a new extreme acute respiratory syndrome coronavirus 2 after detailed studies on the virus (SARS-CoV-2). The WHO later called the virus ′′Corona Virus Diseases 2019′′ in February2020 (COVID-19). Various steps including social distance and lock-downs have been taken in many parts of the world to contain the rapid transmission of the virus, affecting people's living and working conditions (Nasiri and Dimitrova, 2021).

SARS-CoV-2 has been regarded as highly infectious and affected individuals may continue to have asymptomatic or clinical symptoms ranging from minor, to extreme and with a variable incubation time up to 14 days. The following clinical findings were reported for COVID-19 in a reduced incidence of fever (82%), toxicity with or without sputum (61%), muscle discomfort and/or weakness (36%), dysphagia (26%), headache (12%), sore throat (10%), and gastrointestinal symptoms (9%) in a survey performed between January and February 2020 (Bernardino et al., 2021).

14 percent of the individuals afflicted by COVID-19 expect to develop the most extreme type and 5 percent are the most critical, according to the Pan American Health Organization (PAHO). While the mortality rate may fluctuate depending on countries and territories, it is generally correlated with age and comorbidity factors. The more serious cases of asthma, diabetes, smoke, infectious and coronary liver disorders were linked to the age group of elderly patients and comorbidities (Bernardino et al., 2021).

The importance of having a COVID-19 vaccine has skyrocketed and is of the highest priority. But the spread of COVID-19 has to be monitored to efficiently see that the vaccine is spread so that the spread of the virus can be controlled efficiently. This project provides the statistical analysis and predictions required to observe the spread of COVID-19. The output of this research can be compared to the statistical analysis of the spread of COVID-19 and areas more prone to spreading the virus can be targeted for the vaccine first.

## Background

The need to develop a vaccine arose when the new epidemic COVID-19 hit hard. It wasn’t expected that a new epidemic would occur and humanity wasn’t prepared for it in the least. The death toll started increasing and so did the risks of it spreading even faster. The only way to deal with it was to develop a vaccine that to immunize people to COVID-19 once taken. The development of this vaccine however took a year to complete.

The requirement of a powerful assistant that can provide a solution faster than traditional methods and can save lives faster was required. With machine learning, an approach could be built where a large amount of data can be fed into the algorithms and in turn, they spit out trends and predictions regarding COVID-19 attacks and possible sick regions while also maybe stumbling upon a quick, viable solution.

## Aim

The aim for this research project is to predict the spread and progress of the COVID-19 vaccine around the globe. The project will also strive to provide detailed data visualizations that give deep insights of the COVID-19 situation in the countries of the world.

## Research Questions

The following question provides purpose to this research project:

1. What is the progress of COVID-19 vaccination development and spread across the globe?
2. What vaccines are used in which countries?
3. Which of the countries have vaccinated more people?
4. What country has vaccinated a large percentage of population?
5. Which vaccine is the most popular?

## Ethical Considerations

Ethics is a complicated subject that has only become more prominent during the advent of Big Data. The UK Data Service department also provides guidelines for ethical research with specific relation to Big Data. These guidelines will form the basis for this reports ethical approach. Some of the concerns that will be addressed are:

* Maintaining confidentiality in line with Birmingham City University (BCU) and DC guidelines
* Anonymizing information that violates group privacy
* Ensuring transparency in reasons for data collection
* Ensuring data is only used for the direct purpose it has been requested
* Referencing sources for all information used within the research project
* Ensuring all data is stored in the correct location. DC information must remain on DC servers.

As this project encounters any further ethical concerns these will be met within the recommended UK guidelines and with the advice of BCU and DC supervising members.

## Project Timeline

To make research projects more efficient, the projects are divided in modules and sub-modules to keep the focus and concentration alive. Each of those modules and sub-modules are then provided a deadline that needs to be fulfilled. Focusing on one task gives efficiency and speed while the importance of each task is built by creating dependencies.

This process of assigning deadlines to the modules of the project is called creating a project timeline. A Gantt chart is the tool preferred worldwide to visualize the project timeline. The project timeline for this project is shown in *Figure 1*.

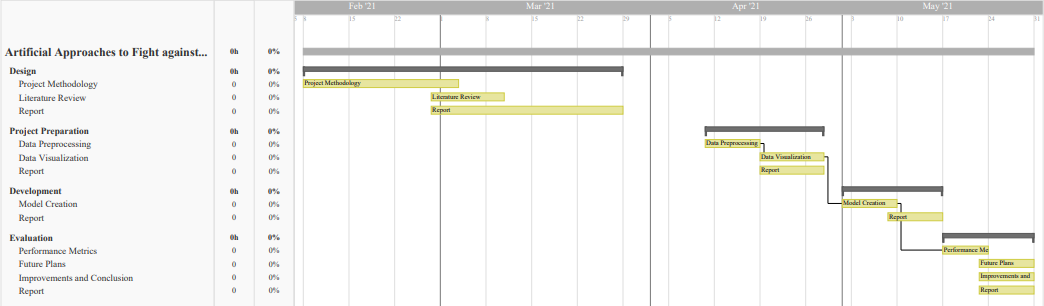


Figure 1 - Project Timeline using Gantt chart

There may be unexpected events that cause difficulty in maintaining the project timeline. These are risks. Contingencies for some expected risks are discussed below in *Figure 2*.

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Risk Impact (1-5)** | **Risk Likelihood (1-5)** | **Contingency** |
| Data Mismanagement | 5 | 1 | Each iteration of the dataset will be synced with Version Control. |
| Personal Time Mismanagement | 3 | 3 | Regular Feedback from Supervisor for Continuous development. |
| Hardware Failure | 1 | 1 | Extra Supply of RAM and SSD. |

Figure 2 - Risk Management

# Methodology

The tools and technologies required to complete this research project and the processes that will implement these tools to do so are mentioned under this heading. The methodology to achieve the completion of this project comes in three stages:

1. Literature Review
2. Project Design
3. Project Evaluation

## Literature Review

Since the advent of COVID-19 and its spread, researchers all around the world have poured their efforts into researching on the disease to provide the world with the data to fight the virus. This has resulted in a lot of data and research already available for use on the internet. To refer these research papers and study them for insight is known as Literature review (LR).

A good LR is necessary for any research project since it provides the knowledge to handle any contingencies that might occur before starting the project. It gives a head-start regarding the project design and proceedings. A strong LR can also be the basis for the validation of the integrity of the research performed. The workflow for the LR in this research is shown in *Figure 3*.

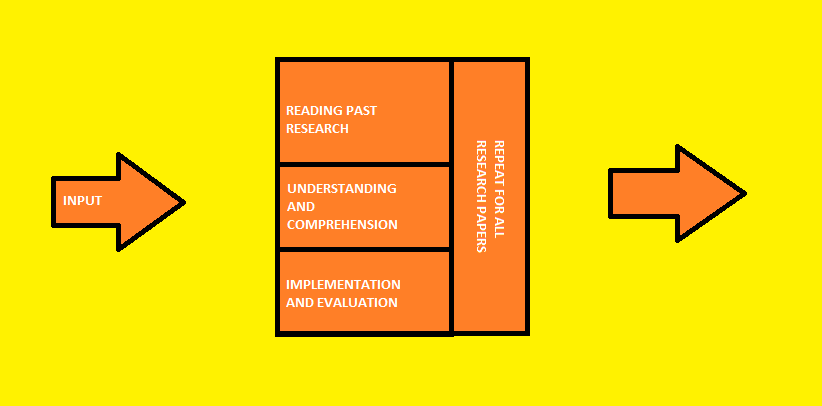


Figure 3 - LR Workflow

Due to the availability of a huge amount of research, it is paramount that the quality of the LR be prioritized. Therefore, a selection criterion is put in place to manage the selection and approval of research papers for this research. The criteria are as follows:

* All the papers must be in English.
* All the papers must be published in journals with a high impact factor.
* Any paper older than 2019 will not be accepted.

## Project Design

Research projects mainly consist of repeated processes of experimentation and data analysis. Although all the experiments and data analysis must be performed in an orderly fashion, there is no need for a Software Development Life Cycle since the output is not an application but rather results of experiments performed in this project.

## Experimentation

The requirement of a good dataset and a pairing ML algorithm to give life to the dataset using predictions go hand-in-hand. This section will shed light on both of these aspects of the project.

### Dataset

Data is collected daily from Our World in Data GitHub repository for covid-19, merged and uploaded and is available on [Kaggle](https://www.kaggle.com/gpreda/covid-world-vaccination-progress). The data contains the following information:

* Country - the country whose vaccination stats are provided.
* Country ISO Code.
* Date - date when the data was collected.
* Total number of vaccinations - total vaccines in the country.
* Total number of people vaccinated - in some cases, due to the immunity setup of people, they might require 2 dosages, the number of vaccines might not match with the number of people vaccinated.
* Total number of people fully vaccinated - total people that have totally been vaccinated.
* Daily vaccinations (raw) - total vaccinations done for a specific data instance.
* Daily vaccinations
* Total vaccinations per hundred (%) - (vaccination number) / (total population up to the date).
* Total number of people vaccinated per hundred (%) - (population immunized) / (total population).
* Total number of people fully vaccinated per hundred (%) - (population fully immunized) / (total population).
* Number of vaccinations per day - the number of vaccines in a day.
* Daily vaccinations per million (ppm) - (vaccination number) / (total population).
* Vaccines used in the country - total vaccines used by the country.
* Source name - information source for the data.
* Source website - website for the information source.

### Machine Learning

Understanding the problem leads to the conclusion that the paradigm to implement for this dataset should be regression. The data is a time-series data and the objective is to predict the spread of the COVID-19 vaccine. Prediction from a time-series data requires regression algorithms from ML. The algorithm that will be used in this project will be ARIMA. ARIMA is a family of models that 'explains' a given time series dependent its past values, i.e., its own deficiencies and delay in the prediction of future values, so that the equation can be used to predict future values. This algorithm for this project will be implemented in the 7-day period.

# Project Evaluation

The evaluation of this project will be performed with the following steps:

* Performance metrics for the ML algorithms used.
* The accuracy and depth of the data analysis performed.

# Literature Review

Five research papers have been finalized for the literature review. The papers will be studied and summarized. The summary will focus on the methodology of the paper, the experiments and the results of those experiments.

In comparison and research from adolescents and infections, open information on young people's and comorbidities with COVID-19 is more conflicting, making it harder to find the risk factors for uncertainty and mortality of this population age gain. In such a special circumstance, it is important that children and young people who are infected with SARS CoV-2 are aware of the significance of illness for this age group and of the mixing aspects of their disease. The purpose of this research is to prepare the epidemiological profile of COVID-19 young people on the globe.

## Epidemiological profile of children and adolescents with COVID-19: a scoping review.

### Methodology

The on-screen study is planned as a reading audit based on the Joanna Briggs Institute's convention, coordinated by the Preferred detailing Items for Systematic Review and Meta-Analyses expansion for Scoping Reviews (PRISMA-ScR), whose aim is to plan ideas central information on a given subject matter, through a complete inclusion of the writing, to distinguish existing examinations. The examination query was created using PCC technique, which recommends the following mental aids as important components: P - Population, C - Concept, and C - Context. The following components were identified: P (children and adolescents with COVID-19); C (epidemiological profile); and C (world setting).

Following the formulation of the inquiry issue, the process of identifying significant studies was carried out in the Cumulative Index to Nursing and Allied Health Literature (CINAHL), Web of Science, Scopus, Science direct data sets, accessed through the CAPES diaries entry via Café, and Google Scholar access. The search process was divided into three stages: first, the search was limited to the PubMed/MEDLINE data collection, followed by an examination of the most often used terms in the titles and updated works of the documents. In the following point, searches were conducted in all knowledge bases using the catchphrases defined in the previous stage. In the third stage, the reference rundown of the articles gathered was counselled to recognize extra examinations that had not been planned in the past stages.

### Results

The research results are summarized in the following tables.

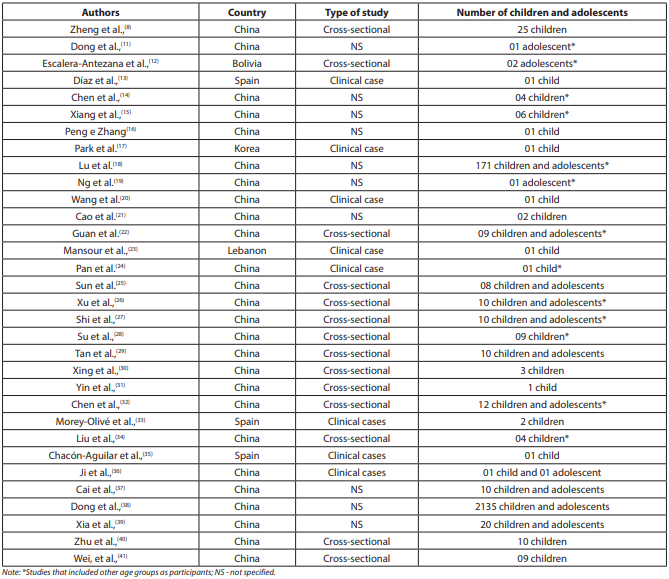


Figure 4 - Characterisation by writers, country, method of research and number of participants Studies analysed (children and adolescents), Brazil (2020)

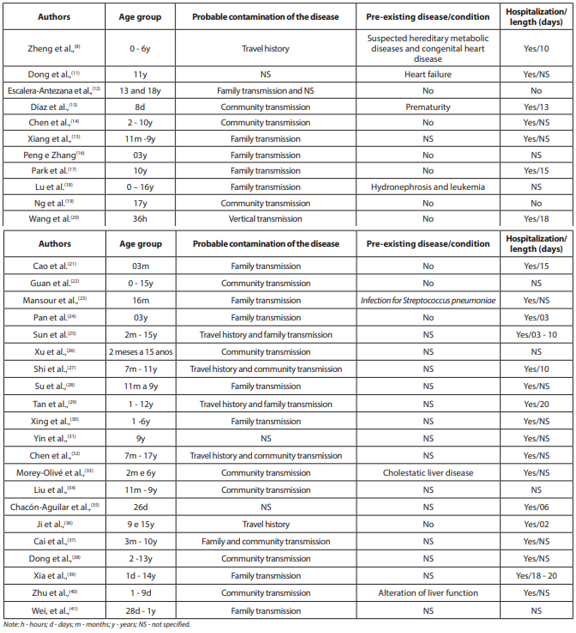


Figure 5 - The epidemiological profile of children and teenagers (age group, likely pollution, pre-existing disease/condition and hospitalization).

## The role of inflammation and frailty in identifying prognostic risk factors for poor outcome in in-centre haemodialysis patients with COVID-19 disease.

The COVID pandemic (COVID-19) has had a significant impact on patients with comorbidities and infirmity that are unable to self-segregate, such as those undergoing haemodialysis. The aim of the study was to identify risk factors for mortality and hospitalization, which may be useful in potential disease outbreaks.

### Methodology

Information was collected reflecting the electronic health reports for all patients receiving COVID-19 findings from the Royal Free London NHS Foundation Trust in London, UK, on eleventh March and ten May 2020. The mandatory outcome was death, and the optional outcome was hospitalization.

### Results

148 suggestive patients were tried positively by RT-PCR for SARS coV-2 and were recalled for examine out of 746 patients under routine haemodialysis. The overall mortality rate was 24.3%. The higher mortality levels (all p value < 0.05) were associated with more developed ages, ischaemic coronary disease, lower systemic circulatory pressure, lower weight file (BMI) and elevated feebleness rates. Higher estimates of WBC, neutrophil tallies, neutrophil to lymphocyte proportions (NLR), C-responsive protein (CRP), bilirubin, ferritin, troponin, and lower serum egg whites level were associated with mortality (all p values 0.05). In the strategic relapse, death was associated with older age and higher CRP, while significant levels of NLR and CRP were associated with the need for hospitalization.

### The effect of COVID-19 on treatment for breast cancer.

### Methodology

Review on information gathered tentatively among April and June 2020. Outpatient participation and surgeries were contrasted with information from a similar stretch in 2019.

### Results

True to form, the quantity of patients treated was lower. Nonetheless, the bosom unit figured out how to keep a fundamental malignancy administration, with the accompanying alteration to the standard working strategies:

* Extra working auditoriums were dispatched at a neighbourhood private medical clinic.
* Oncoplastic and reconstructive methods were halted.
* Oncological medical procedures were restricted to bosom protection and mastectomy.
* Rules for adjuvant radiotherapy were changed.
* Chemical receptor positive malignancy patients were begun on neo-adjuvant endocrine treatment.
* The quantity of new suggestive references dropped, and the bosom screening administration was incidentally suspended.
* The mammogram observation program.
* The mastectomy rate expanded from 28% to 43%, and bosom protection methods dropped from 72% to 57%.

## Specimens in COVID-19: a Systematic Review & Meta-Analysis Comparing saliva and nasopharyngeal swab in detection.

Because of the ease with which COVID-19 spreads, the virus poses a threat to global health. Early detection of suspected patients will play a critical role in preventing further dissemination of COVID-19. The aim of this audit analysis was to compare salivation examples to nasopharyngeal swab (NPS) examples in investigations drawn from various data sets.

### Methodology

To accomplish the target of this examination, an orderly writing search was completed in four information bases, specifically PubMed, Google Scholar, Cochrane Library, and LILACS. The catchphrases ″COVID-19″, ″Nasopharyngeal Swab″, and ″Saliva″ were used by means of Boolean administrators.

### Results

14 articles were remembered for this survey study following the qualification standards. In light of information introduced in examinations utilized in the meta-investigation, there was no huge contrast between both example types for discovery of COVID-19. Heterogeneity test showed that I2 esteem was 5.790% (<20%). The impact size (hazard proportion) of the 14 examinations was 0.951 (<1).

## Taking care of fragile rectal cancer patients during a pandemic COVID-19: an Italian experience in one-centre.

During the COVID-19 pandemic, entire countries soon run out of elevated nursing beds due to the presence of fundamentally contaminated patients. Almost all patient care frameworks were placed under intense pressure as a result of this. Elective medical procedures have been discontinued, and intensive consideration medical procedures have been severely restricted. Since the beginning of the pandemic, Minimally Invasive Surgery (MIS) and General Anaesthesia (GA) have been the subject of intense debate because they are all vaporized generating processes that may contribute to disease within operating rooms. Furthermore, as is well known, GA can be associated with delayed recovery following sedation which can result in the patient's admission to the Intensive Care Unit (ICU). This confined the remedial systems. This exploration reports how the issues were taken care of.

### Methodology

In the middle of March and the middle of May 2020 (the Italian lockdown), our medical clinic saw 40 patients. Prior to the surgical treatment, both patients completed a pre-confirmation screening poll to assess the risk of a new illness and had a nasopharyngeal swab for COVID-19 review. Based on the swab reaction, patients should be transferred to "safe" or "Coronavirus committed" wards. Patients who tested negative for COVID-19 were assigned to single or double rooms depending on the preoperative poll. Single rooms were reserved for patients with an uncertain score (question late anamnesis), while twofold rooms were reserved specifically for patients with a "healthy" score (no record of suspect on the new anamnesis). Colorectal cancer was diagnosed in 15 patients (37.5 percent) (10 influenced with rectal malignant growth, 5 influenced with colon malignancy). The average age was 84.3 years old. Both patients were classified as fragile (ASA 3, older than 80 years, and influenced by severe comorbidities) and underwent conscious laparotomy under loco-local sedation (ALLA).

### Results

Medical procedures were conducted under joined spinal-epidural (CSE) anaesthesia in 11 cases, spinal sedation (SA) in three cases, and epidural anaesthesia (EA) in one event. The average employable time was 87 minutes (least an hour, most extreme 165 minutes). In one scenario, the switch to general sedation was critical. In any event, postoperative agony was well-managed. They did not need any postoperative focused consideration support. There were no perioperative confusions or early readmissions after surgical procedures. Both of the nasopharyngeal swabs were negative.

# Conclusion

This proposal has stated the background and methodology of this project. An extensive Literature Review has been performed to understand the effects of COVID-19 and the importance of this research. The limitations of the data available and the methods used to handle tough COVID-19 related situations are discussed here. Previous literature describes the effects of COVID-19 on hospitals and the inconsistencies in the data collected on COVID-19 patients. This provides a validity and necessity for this research.

# References

* Bernardino, F., Alencastro, L., Silva, R., Ribeiro, A., Castilho, G. and Gaíva, M., 2021. Epidemiological profile of children and adolescents with COVID-19: a scoping review. *Revista Brasileira de Enfermagem*, 74(suppl 1).
* Hendra, H., Vajgel, G., Antonelou, M., Neradova, A., Manson, B., Clark, S., Kostakis, I., Caplin, B. and Salama, A., 2021. Identifying prognostic risk factors for poor outcome following COVID-19 disease among in-centre haemodialysis patients: role of inflammation and frailty. *Journal of Nephrology*.
* Khan, S. and Goh, S., 2021. The impact of COVID-19 pandemic on breast cancer treatment. *European Journal of Surgical Oncology*, 47(2), p.e33.
* Nasiri, K. and Dimitrova, A., 2021. Comparing saliva and nasopharyngeal swab specimens in the detection of COVID-19: A systematic review and meta-analysis. *Journal of Dental Sciences*.
* Romanzi, A., Adorni, A., Boleso, N., De Rango, S., Di Palma, G., Gabaglio, M., Galletti, M., Giacci, F., Macchi, L., Moroni, R., Putortì, A., Rongoni, E., Rossi, F., Scolaro, R., Taiana, C., Trevisan, D., Zaccarelli, A., Zanardo, M., Zucchi, D. and Vannelli, A., 2021. Caring for fragile patients with rectal cancer during the COVID-19 pandemic: an Italian single-center experience. *European Journal of Surgical Oncology*, 47(2), pp. e48-e49.