*Assignment 4: Quantitative research plan*

# A deep learning based quantitative analysis for predicting medical diagnosis in healthcare

Muthupriya Shankaran (mush8206)

Department of Computer and Systems Sciences

*Empirical Research Methodology for Computer and Systems Sciences*

Stockholm University

Autumn term 2021

Supervisor: Name

Swedish title: XXX – not mandatory

# Abstract

The paper is about the deep learning based quantitative model for predicting medical diagnosis with accurate result has been conducted. It uses deductive approach. Among several of medical diagnosis this paper specifically would compare the severity of covid-19 and respiratory studies. The objective of this quantitative research study is to develop and employ mathematical models, theories and hypotheses pertaining to phenomena. Using quantitative data analysis various results can be obtained which would be useful to get the expected predicted value during diagnostic process.

Keywords: Quantitative analysis, Deep learning (DL), medical diagnosis, covid-19, respiratory study.

# Table of Contents

# List of Figures

# List of Tables

# 1 Quantitative research plan

Quantitative research plan provides the clear and detailed foresight before the research begins. In this paper with the base of sampling and population related data are collected. Once the data is collected, survey and test data may need to be transformed from words to numbers then I used statistical analysis to answer my research questions. Before survey, study plan is explained in Gantt chart which gives the detail of the task and deliverables with specific timeline description. Finally, with the help of data collection, various statistical test (Regression & correlation) has to be performed to predict the expected outcome.

## 1.1 Research question

The research question for this paper is"To what extend can deep learning model predict medical diagnosis in healthcare?" Providing accurate and accessible diagnosis is a fundamental challenge for global healthcare systems. Therefore, aimed to establish a deep learning (DL) model based on quantitative analysis and initial clinical features to predict the severity of COVID-19 and respiratory study.

## 1.2 Sampling and population

To ensure reliable and valid inferences from a sample, probability sampling technique is used to obtain unbiased results. A stratified random sample is one in which the population is first divided into relevant strata or subgroups and then, using the simple random sample method, a sample is drawn from each strata. The samples of 115 with 244 separate patient cohorts report on diagnostic accuracy of DL on respiratory disease. Lung nodules were largely identified on CT scans, whereas chest X-rays (CXR) were used to diagnose a wide spectrum of conditions from simply being 'abnormal' to more specific diagnoses, such as pneumothorax, pneumonia and covid-19.Here a high or unclear risk of bias was seen in 93/115 (81%) of respiratory studies.Therefore it shown sampling issues,this was largely due to missing information about patients not receiving the index test or whether all patients received the same reference standard. This challenge can be overcome by proper information and determination of sample size should be measured correctly.

## 1.3 Place of study

In geographical area,research on medical diagnosis is being effectively carried out in various areas such as in Brazil belonged to the World Health Survey,Latin American countries and the caribbean are few more region where this study has been conducted .And in locality,the place of study would be in medical laboratory or clinical laboratory where tests are carried out on clinical specimens to obtain information about the health of a patient to aid in diagnosis, treatment, and prevention of disease.

## 1.4 Time plan

Time plan plays a major role when it comes to research plan. For this Quantitative Research method from below Gantt chart we may come to know the time period that how long will take to complete this task. This research requires high samples of data collection which make use of validity and reliability measurement therefore time planning and time management plays very important. Below (Table 1) are the steps can be planned on weekly basis

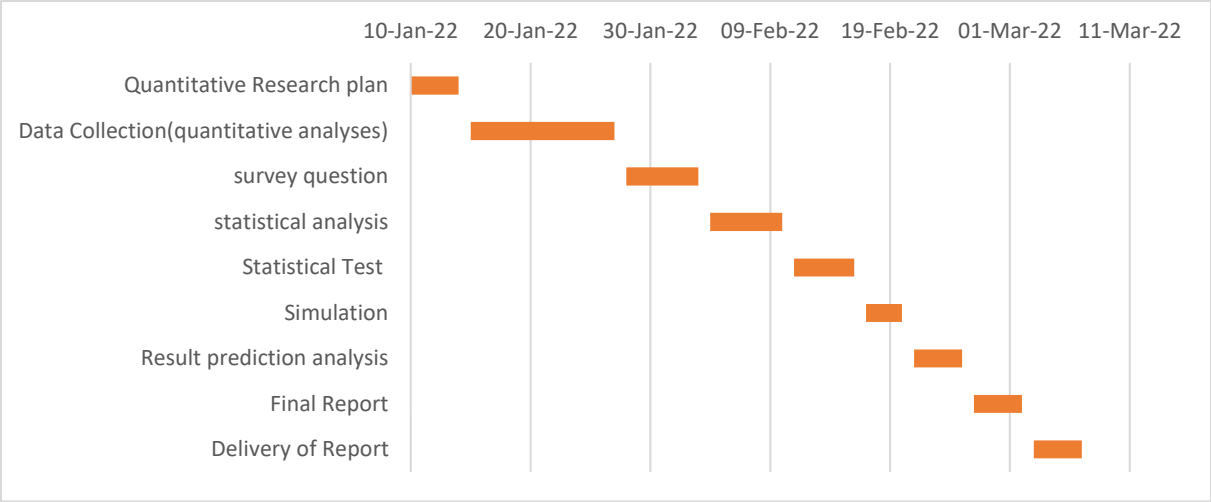


Table 1: Planned Gantt chart (Duration is calculated on weekly basis).

## 1.5 Plan for data collection

To predict the dependent variable to this research, data collection is crucial to perform statistical analysis and statistical test. Data can be collected for medical diagnosis in many ways such as primary data, secondary data or experimental design. To this research question combination of experimental design and secondary data surveys are performed, because secondary data is data gathered from studies, surveys, or experiments that have been run by other people are other researchers. Also, in healthcare, secondary data collection can plan methods in two forms such as government publications and public records (Electronic health records (EHR)). Moreover, public records are taken for this quantitative analysis method it's because they may some problems occur while getting private database as it may contains sensitive and confidential data which might come under GDPR. Furthermore, in public dataset the plan itself contains different steps including type of data, variables (independent and dependent variable) which would help in predict the accurate result.

### 1.5.1 Kind of data needed

Secondary data, public records from online (EHR) websites data are collected for this research question. A quantitative analysis was performed in electronic databases, health institutions websites, and references of papers retrieved and contact with authors.

Electronic databases were: PubMed, Lilacs, Embase, Web of Science and the Scientific Electronic Library Online (SCIELO).

The institutional websites were: the World Health Organization http://www.who.int;
Brazilian National School of Public Health http://www.ensp.fiocruz.br,
Brazilian Health Ministry http://www.saude.gov.br,
Pan American Health Organization http://www.paho.org and
The Management Sciences for Health http://www.msh.org/seam.

### 1.5.2 Data guide

Using secondary data and experimental design- data preparation, statistical analysis, test, inferential technique and discussion has been conducted on the basis of research question to estimate the measurement in medical diagnosis using DL in terms of statistical test to predict the accuracy rate.

Following steps are carried out for Quantitative data analysis:

**Survey Question**: How will you measure medical diagnosis?

**Data Preparation**: To answer survey question data guidelines and statistical test to be performed with the base of sampling and population. In healthcare, for diagnostic process, Covid-19, pneumonia and respiratory study was taken to analyze.According to the COVID-19 Guidelines (the fifth version) set by the National Health Commission of the People's Republic of China (3), patients with COVID-19 can be divided into four subtypes: mild, common, severe and critically ill. As the mild subtype with no pneumonia was excluded, the patients enrolled in this study were divided into non-severe (common subtype, 151 cases) and severe (severe and critically ill subtypes, 45 cases). Among the severe group, 28 patients had severe pneumonia, and 2 developed shock. The clinical and laboratory data were reviewed and classified. The pneumonia severity index (PSI) was calculated for all patients (4).

**Statistical analysis & Test/Inferential Technique:** The continuous data are expressed as the median and interquartile range (IQR, 25th and 75th percentiles) because a majority of the data did not follow a normal distribution. The Wilcoxon rank-sum test and Fisher's exact test were used to compare quantitative and categorical variables, respectively, between the two groups. Variables found to be significant in univariate analysis were inputted into the least absolute shrinkage and selection operator (LASSO) logistic regression analysis to determine the optimal subset of clinic-radiological features for prediction. By incorporating these significant predictors, a nomogram model was established by five-fold cross-validation. The nomogram was calibrated by performing a calibration curve analysis. Receiver operating characteristic (ROC) analyses were conducted, and the areas under the receiver operating characteristic curve (AUCs) of the nomogram model, quantitative CT parameters that were significant in univariate analysis, and PSI were compared using the Delong test to evaluate the effects of the classifier when comparing the severe group with the non-severe group.

The cut-off values were defined based on the maximal Youden index.

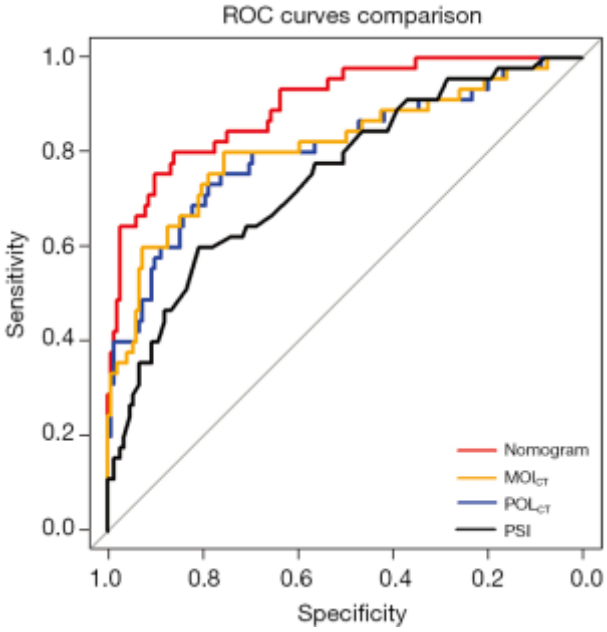


Figure 1: severity by Delong test

**Discussion**: In this study, the parsimonious model containing five features (age, LDH, CRP, $CD4^+$ T cell count, and $MOI_{CT}$) was an ideal measure to predict the severity of COVID-19. LASSO logistic regression analysis not only works better than the conventional method of choosing predictors on the basis of the intensity of their univariate association with outcome, but it also allows researchers to combine the selected features into a model. This is because the model combines DL-based quantifiable computed tomography (CT) parameters of lesions with clinical laboratory indicators to comprehensively assess the severity of the disease, rather than as a partial assessment of each patient's condition.

However, our study has several limitations, the nomogram model was established by cross-validation and had a good predictive performance, the proportion of patients with severe disease was relatively low, and there was an imbalance between the sizes of the severe and non-severe group, which may impact the statistical analysis. More data, especially from different geographic areas, are needed to validate the robustness of the model to further improve its prediction accuracy.

**Conclusion**: Quantitative CT parameters and the PSI can well predict the severity of COVID-19. The DL-based quantitative CT model containing five clinic-radiological features can serve as a more efficient tool for prediction than individual quantitative CT parameters and PSI. Therefore this empirical study research would help for future work.

# References


1. World Health Organization. 2019-nCoV outbreak is an emergency of international concern. Available online: http://www.euro.who.int/en/health-topics/health-emergencies/international-health-regulations/news/news/2020/2/2019-ncov-outbreak-is-an-emergency-of-international-concern. Published 31 January, 2020. Acessed 24 February, 2020.
2. Carvalho MF, Pascom AR, Souza-Junior PR, Damacena GN, Szwarcwald CL: Utilization of medicines by the Brazilian population, 2003. 2005, Cad Saúde Pública, 21 (Suppl): 100-108.
3. China National Health Commission. Diagnosis and treatment of pneumonitis caused by new coronavirus (trial version 5). Available online: http://www.nhc.gov.cn/yzygj/s7653p/202002/3b09b894ac9b4204a79db5b8912d4440.shtml. Updated 5 February, 2020. Accessed 5 February, 2020.
4. Kao KC, Chang KW, Chan MC, et al. Predictors of survival in patients with influenza pneumonia-related severe acute respiratory distress syndrome treated with prone positioning. Ann Intensive Care 2018;8:94. [Crossref] [PubMed].
5. Das KM, Lee EY, Enani MA, et al. CT correlation with outcomes in 15 patients with acute Middle East respiratory syndrome coronavirus. AJR Am J Roentgenol 2015;204:736-42. [Crossref] [PubMed].

# Appendix A – An Appendix