A Project Report

on

COVID 19 DATA ANALYSIS AND PREDICTION FOR INDIA

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology in

Computer Science & Engineering

Under The Supervision of
Dr. Ravindra kumar chahar
Associate Professor

Submitted By

Shivansh dwivedi / 18SCSE1010141 / 18021011388
Mayank Srivastav / 18SCSE1010252 / 18021011490

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

DECEMBER, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project, entitled **"Covid19 Data Analysis and Prediction of India."** in partial fulfillment of the requirements for the award of the B.Tech submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of September,2021 to December and 2021, under the supervision of Dr. Ravindra Kumar Chahar Associate Professor, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by us for the award of any other degree of this or any other places.

Shivansh Dwivedi/18SCSE1010141

Mayank Srivastav/18SCSE1010252

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Ravindra Kumar Chahar

Associate Professor

CERTIFICATE

The Final Project Viva-Voce examination of Shivansh Dwivedi/18SCSE1010141 and Mayank Srivastav/18SCSE1010252 has been held on _____ and his/her work is recommended for the award of B. tech in Computer Science & Engineering .

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021

Place: Greater Noida

Abstract

The outbreak of COVID-19 in different parts of the world is a major concern for all the administrative units of respective countries. India is also facing this very tough task for controlling the virus outbreak and has managed its growth rate through some strict measures. This analysis presents the current situation of coronavirus spread in India along with the impact of various measures taken for it. With the help of data sources (till 10th of June) from various state units of India and Ministry of Health and Family Welfare, Government of India, this study presents various trends and patterns.

This study answers six different research questions in a comprehensive manner.

It has been reported that growth rate of infected cases has been controlled with the help of National Lockdown, however some uncontrolled mass level events had negatively impacted the infected cases.

It appears that only essential services should be open for the citizens of India and the na- tional lockdown should be carried on for next 2-4 months. This study will be useful for the Government of India and various states of India, Administrative Units of India, Frontline health workforce of India, researchers and scientists. This study will also be favorable for the administrative units of other countries to consider various aspects related to the control of COVID-19 outspread in their respective regions.

COVID-2019 has been recognized as a global threat, and several studies are being conducted in order to contribute to the fight and prevention of this pandemic. This work presents a scholarly production dataset focused on COVID-19, providing an overview of scientific research activities, making it possible to identify countries, scientists and research groups most active in this task force to combat the coronavirus disease. The dataset is composed of number of records records of articles' metadata collected from Scopus, PubMed, arXiv and bioRxiv databases from January 2019 to July 2020. Those data were extracted by using the techniques of Python Web Scraping and preprocessed with Pandas Data Wrangling. In addition, the pipeline to preprocess and generate the dataset are versioned with the Data Version Control tool (DVC) and are thus easily reproducible and auditable.

Acknowledgement

We take this opportunity to express our sincere gratitude to Prof. Dr. Ravindra kumar chahar Associate Professor, Department of CSE GALGOTIAS UNIVERSITY, GREAETER NOIDA. Deep Knowledge & keen interest of our supervisor in the field of “Data analysis and machine learning” to carry out this project. His endless patience, scholarly guidance, strong motivation, continual encouragement, constant and energetic supervision, constructive criticism , valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project. We would like to express our heartiest gratitude other faculty member and the staff of CSE department of GALGOTIAS UNIVERSITY, GREAETER NOIDA to finish our project. We would like to thank our entire course mate in GALGOTIAS UNIVERSITY, GREAETER NOIDA, who took part in this discuss while completing the course work.

Table of Contents

Acronyms

B.Tech.	Bachelor of Technology
M.Tech.	Master of Technology
BCA	Bachelor of Computer Applications
MCA	Master of Computer Applications
B.Sc. (CS)	Bachelor of Science in Computer Science
M.Sc. (CS)	Master of Science in Computer Science
SCSE	School of Computing Science and Engineering

Chapter-1 : Introduction

1.1 Introduction

Data analysis and Machine learning has been widely employed in pattern recognition, a fairly simple talent for people but a difficulty for machines, as a strong technique to achieving Artificial Intelligence. With the advancement of computer technology, pattern recognition has become a necessary and significant part of technology. In the realm of Artificial Intelligence, this is a technique. It is possible to recognize patterns using pattern recognition software. It can recognize letters, images, voices, and other items, as well as their condition or any other abstractions. The Iris flower is the subject of this project. Iris has three separate classifications in its data set: Setosa, Versicolor, and Virginica. These three separate types of Irises will be distinguished by the developed recognition mechanism. The entire machine learning operation should go seamlessly. Users do not need to inform the computer which class the Iris belongs to; the computer is capable of recognizing all of them on its own.

1.2 Formulation of Problem

Machine learning is the process of allowing a computer to build learning skills on its own using pre-programmed knowledge. Pattern recognition can be compared to a computer's ability to recognize many types of items. As a result, pattern recognition and machine learning are inextricably linked. Goal is to build and implement a machine learning-based identification of Iris Flower species using Python and the Scikit-Learn biology module. It is crucial in the fields of artificial intelligence and image processing. 1

1.3 Tool and Technology Used

This project has been made using VS Code editor incorporating the Flask web frame work. We have made use of the Machine Learning patter recognition technology. Pattern recognition may be broken down into two categories: supervised classification and unsupervised classification. They have a lot in common with supervised and unsupervised learning. Unsupervised classification, unlike supervised classification, does not require a teacher to provide the categorization of samples. The process of feature selection is so important that it can have a great effect on the result of pattern recognition. Algorithms like logistic regression, KNN and Support vector machines (SVMs) which are a class of

supervised learning methods for classification, regression, and detection of outliers. All these have been used to produce more accurate result model for flower detection using IRIS dataset in machine learning.

1.4 Python

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991. Python has a dynamic type system and memory management that is automated. It contains a wide standard library and supports several programming paradigms, including object-oriented, imperative, functional, and procedural. For a wide range of operating systems, Python interpreters are available. C Python, like nearly all of Python's other implementations, is open-source software with a community-based development strategy. The Python Software Foundation, a non-profit organisation, oversees Python and C Python.

1.5 VS Code Editor

Visual Studio Code is a source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. Anaconda

1.6 Anaconda

Anaconda is a Python and R programming language distribution for scientific computing that promises to make package management and deployment easier. Data-science packages for Windows, Linux, and macOS are included in the release.

1.7 Scikit-Learn

Scikit-learn is a Python-based machine learning library that is available for free. It includes support vector machines, random forests, gradient boosting, kmeans, and DBSCAN, among other classification, regression, and clustering techniques, and is designed to work with the Python numerical and scientific libraries NumPy and SciPy.

COVID-19, or more popularly known as Novel Corona Virus, is associated with the respiratory disorder in humans which has been declared as a global epidemic and pandemic in the first quarter of the year 2020 by the World Health Organization. As per the latest data (10 th June 2020) by John Hopkins University and other tracking websites, there are currently more than 7.3 million people infected by the Novel Corona Virus all around the world and close to 109 thousand deaths reported from different parts of the world. The top 10 countries with maximum number of infected cases are the United States of America, India, Italy, Germany, France, China, Iran, United Kingdom, Turkey and Switzerland. The top countries with maximum number of reported deaths are Italy, India, United States of America, France and United Kingdom. With respect to the recovered patients list, China is at the top of the list followed by India, Germany, Italy, Iran and the United States of America.
India was first placed comfortably out the list of infected nations by huge margins, but recent events led to its rise to 2nd position which is a point of concern. The mortality rate is controlled at less than 3% right now, which is better than the ~5.5% mortality rate of world, but the model of spread is slowly moving towards an exponential trend which can lead to massive loss of lives and infrastructure.

India is being looked upon by various nations now as a World Leader and even WHO acknowledged that world is looking towards Indian strategies to contain the outbreak of this epidemic.
India accounts for almost one-fifth of the world's population and is second leading country in terms of population in the world. India contributes heavily to the world's GDP and is amongst the most prominent developing country in the world with fairly strong economic growth percentages. India's good camaraderie with majority of the nations in the world and its helpful nature makes it a perfectally for other countries. Therefore, the analysis of COVID-19 outbreak in Indian region is closely watched and monitored by the World and there is a need of comprehensive analytical studies based on different strategies taken by Indian administrators from time to time. India has been following a nationwide lockdown since 22-March-2020, which was a one-day lockdown, followed by a 21-day lockdown after two days. Every activity in India since then has been happening with permission from various administration units and almost all the domestic and international travels have been either banned or monitored closely. India is already got into the third phase of COVID-19 outbreak i.e. the community outbreak as seen by various countries around the world, but the cases have been rising continuously. India's lockdown period has been impacted by two major events in the recent days which were related to the mass exodus of laborers and workers from one state to other states (especially from Delhi to neighboring states) and conduction of a religious event in Delhi which led to spike in the number of cases in various states of India. During this time, the Indian Prime Minister has been trying to connect with Indian citizens through innovative

strategies and coming up with various engagement activities which are impacting the whole nation.

With so much happening in India right now, it becomes imperative that we study the current situation and impact of various such events in India through data analysis methods and come up with different plans for future which can be helpful for the Indian administrators and medical professionals.

The current study explores various aspects associated with the COVID-19 outbreak in India and the various regions situated in India. The specific research questions (RQ) explored in this study are as follows.

1. **Which state has most number of Confirmed Cases, Cured Cases and Death Cases ?**
2. **Mortality rate and Recovery Rate of Maharastra?**
3. **Which Age group has Heigher and Lower chance of getting infected?**
4. **Which state has better Public Health care facilities?**
5. **Which state has heigher number of Public Beds?**
6. **Which state has Heigher number of Beds avilable in urban areas and Rural areas?**
7. **what are the number of people hospitalized, recoverd, Decased and Migrated?**
8. **Which state has most number of sample testing on 12/07/2020?**
9. **What percent of people are tested positive and negative?**

CHAPTER 2 : LITERATURE REVIEW

As per different papers available in literature, there are a few studies that focus on the trend analysis and forecasting for Indian region.

The studies on Indian region presents long term and short term trend, respectively.

These studies use time series data from John Hopkins University database and present forecasting using ARIMA model, Exponential Smoothing methods, SEIR model and Regression Model.

However network modelling and pattern mining are not attempted in these versions of the studies and that too at the regional level, hence the current study attempts to do that. Also, the studies in Indian region from the past are more focused on presenting time series analysis based on the overall data for Indian region rather than covering other sources of information apart from just considering the number of infected patients, so the need to analyze the patients background and information is required for the authorities to get better insight about the situation.

Similarly, there are other mathematical models that were developed for analyzing the trends of COVID-19 outbreak in India. A model for studying the impact of social distancing on age and gender of the patients in India was presented.

It compared the country de mographics amongst India, Italy and China and suggested the most vulnerable age categories and gender groups amongst all the nations.

The study also predicted the rise of infected cases in India with different lockdown periods. Similarly, a network structure approach was used by one of the study to see whether any specific node clusters were getting formed.

But only travel data nodes were considered by the authors to check which the prominent regions are af fecting Indian travelers coming back to the India. Also, the study presented the SIR model to see the rate of spread of the Corona Virus amongst patients in India.

Analysis on the testing labs and infrastructure was also presented by earlier authors. Work of medical doctors and frontline health workers was also presented by some studies [9]. It was found that in India, the role of health workers was less stressed as the spread stage of corona virus was still in phase two or the phase of local transmission rather than the community transmission as compared to other nations like Italy, Spain and USA.

However, it was also claimed that Indian healthcare infrastructure is not very strong as per the WHO guide lines and in case of community spread, the Indian government may find it difficult to manage the spread. Some detailed discussion on the nature of the Corona Virus was also presented by some studies.

Apart from India, a few models are also available for other countries primarily for China,Italy and USA as the number of infected patients was high. Studied like worked on various mathematical models to determine the spread of the disease, predict the number of in fected patients, commenting on the preparedness for each country in

tackling COVID-19 spread and finding the patterns of flattening curve in different conditions. A lot of researches are still in preprint stage for the world level and are yet to be peer reviewed. With respect to the research activities conducted in the Indian region, the studies are yet to work on the impact of different policies working towards containment of the corona virus.

Even in the preprint databases, there are fewer evidences available which worked in the Indian region with more granularities and came up with analysis that can support the decision making of the various administrators in India to curb the lockdown and work on future strategies. Therefore, this study attempts to work on a comprehensive level to analyze the COVID 19 spread in India and impact of various strategies imposed by the Government at both state level and central level.

Chapter 3: Questions answered in Analysis

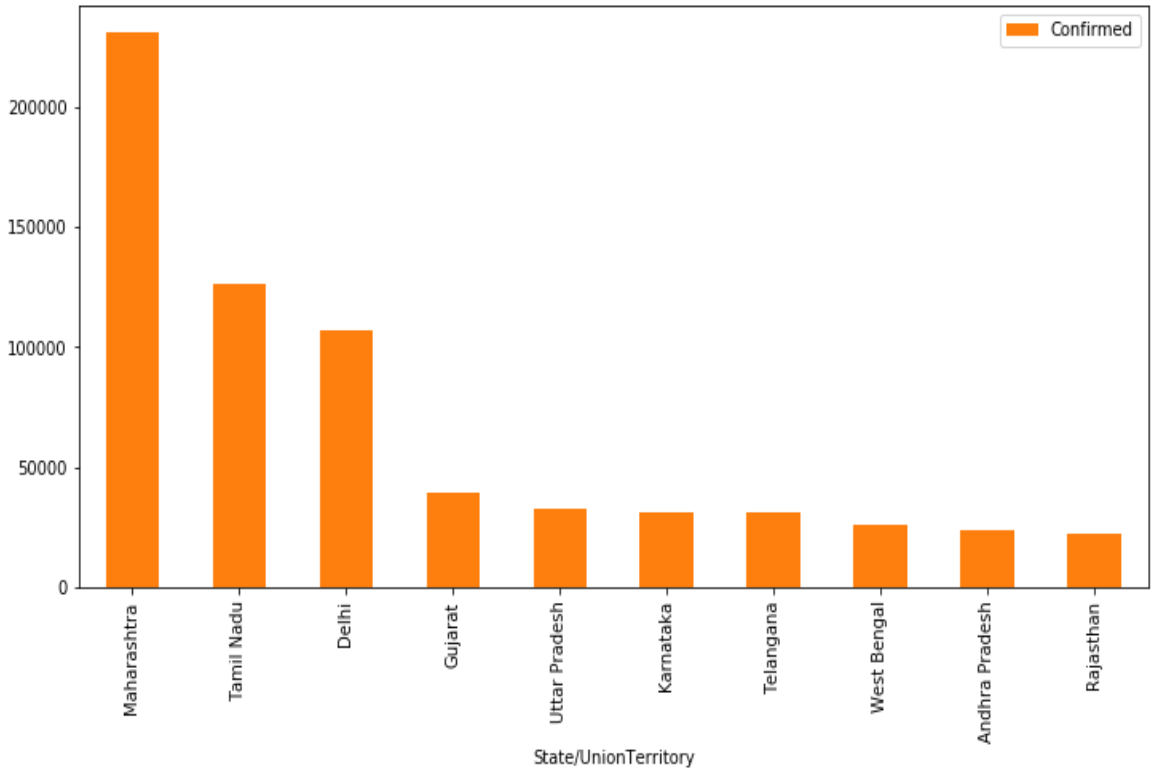
The current study explores various aspects associated with the COVID-19 outbreak in India and the various regions situated in India. The specific research questions (RQ) explored in this study are as follows.

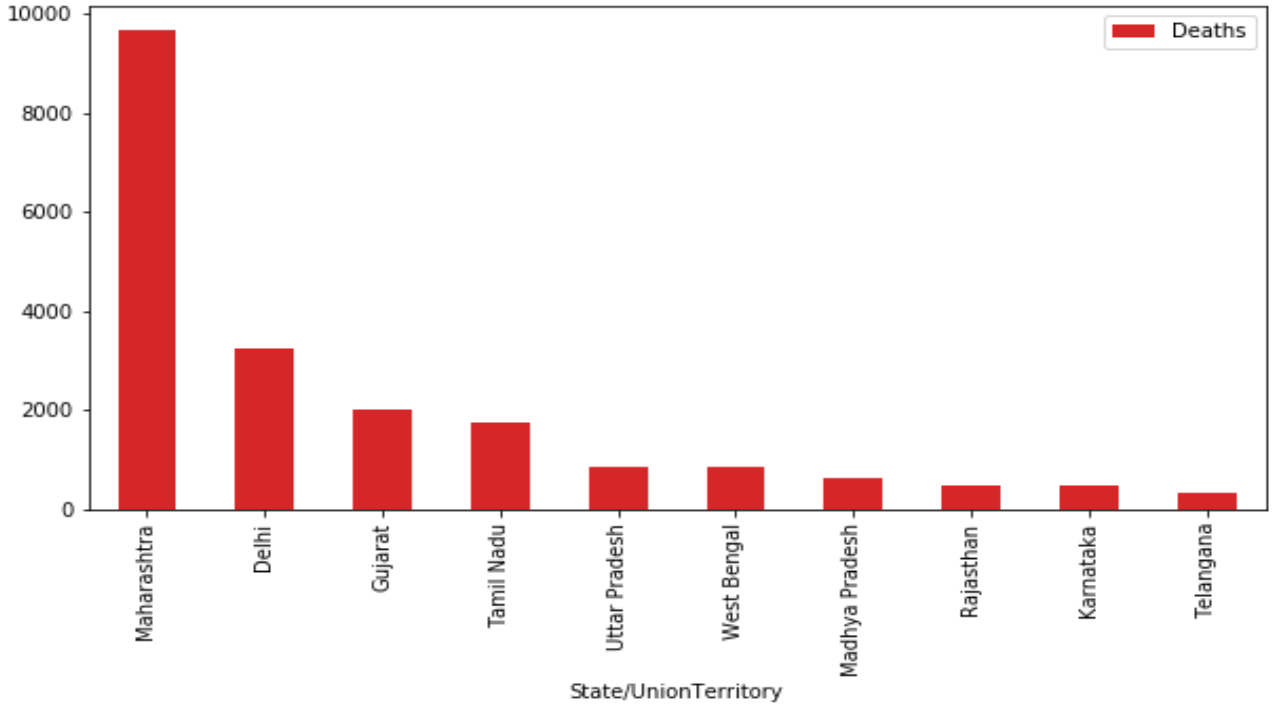
1. **Which state has most number of Confirmed Cases, Cured Cases and Death Cases ?**
2. **Mortality rate and Recovery Rate of Maharastra?**
3. **Which Age group has Heigher and Lower chance of getting infected?**
4. **Which state has better Public Health care facilities?**
5. **Which state has heigher number of Public Beds?**
6. **Which state has Heigher number of Beds avilable in urban areas and Rural areas?**
7. **what are the number of people hospitalized, recoverd, Decased and Migrated?**
8. **Which state has most number of sample testing on 12/07/2020?**
9. **What percent of people are tested positive and negative?**

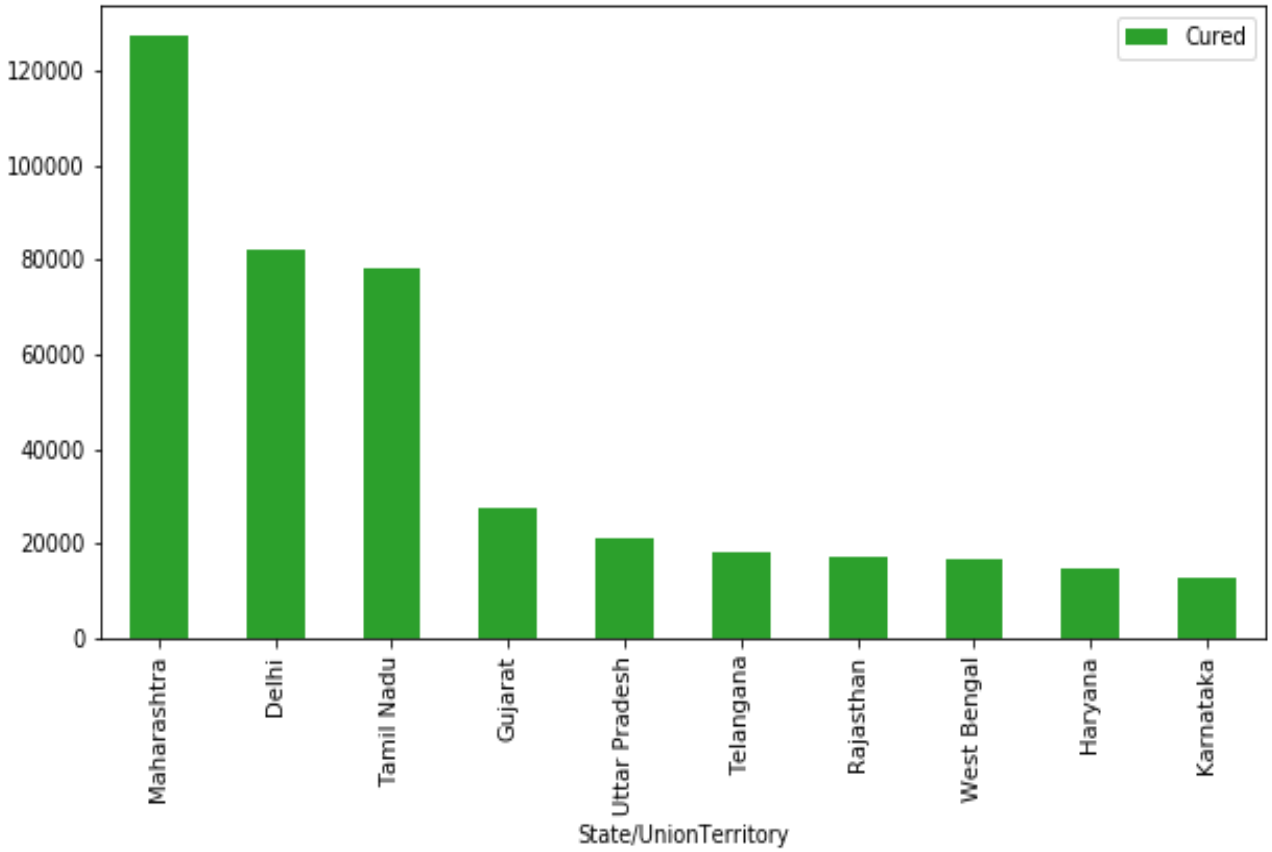
Chapter 4: Findings and discussion

Q1 . Which state has most number of Confirmed Cases, Cured Cases and Death Cases ?

1. Maharashtra has the most number of confirmed cases 730599 on 12/10/2020.
2. Maharashtra has the most number of Deaths cases 29667 on 12/10/2020.
3. Maharashtra has the most number of Cured cases 562401 on 12/10/2020.

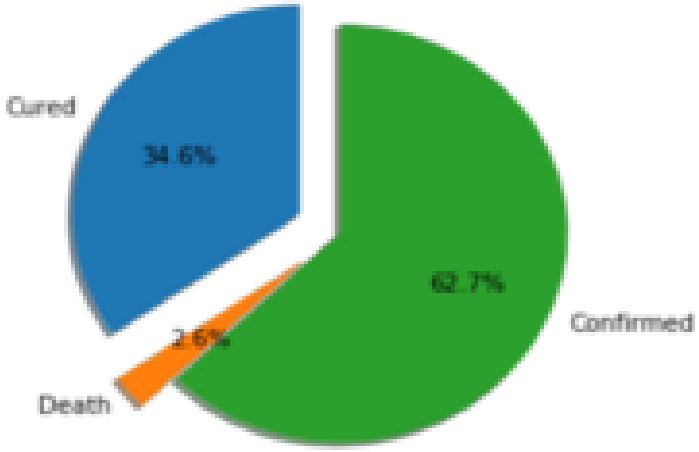



Deaths
10000
8000
6000
4000
2000
0
Maharashtra
Delhi
Gujarat
Tamil Nadu
Uttar Pradesh
West Bengal
Madhya Pradesh
Rajasthan
Karnataka
Telangana
State/UnionTerritory


Cured
120000
100000
80000
60000
40000
20000
0
Maharashtra
Delhi
Tamil Nadu
Gujarat
Uttar Pradesh
Telangana
Rajasthan
West Bengal
Haryana
Karnataka
State/UnionTerritory

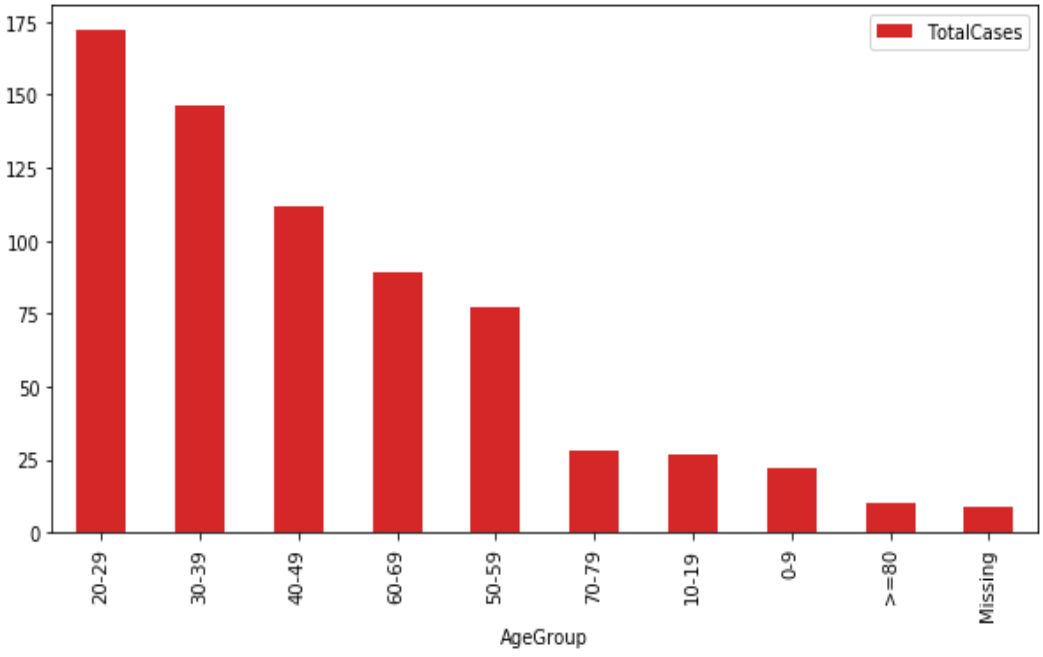
Q2 . Mortality rate and Recovery Rate of Maharastra?

Death percentage is quite less as compared to Cured patients. Thanks to Doctors and working staff in India who take care of all the corona patients 34.6 % of patients are cured of the corona.



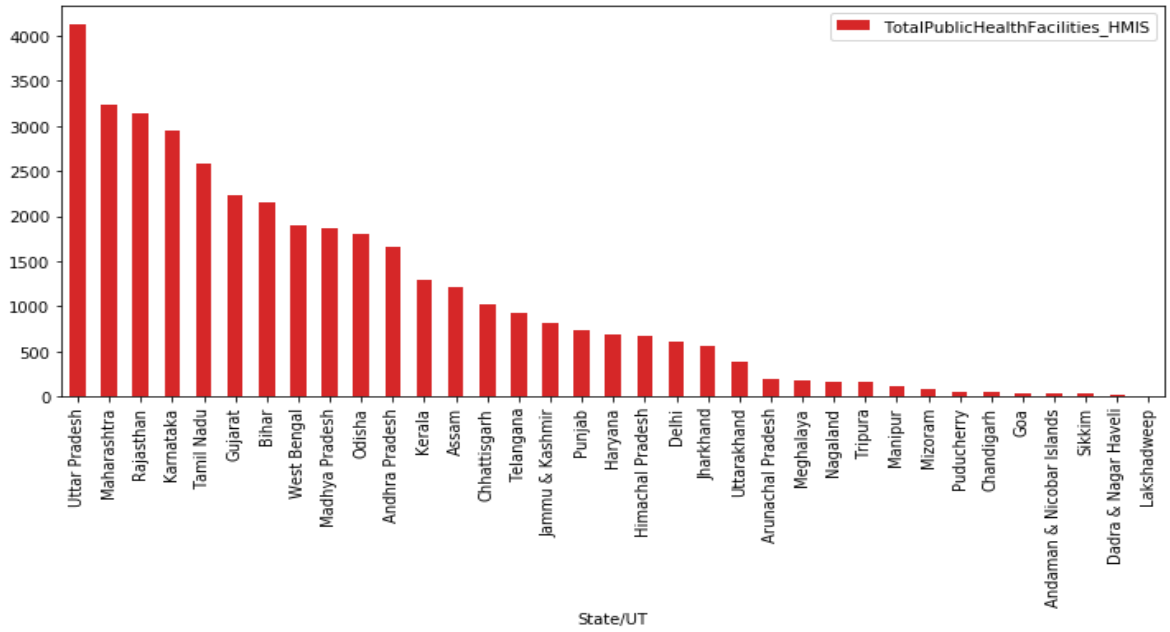
Q3 . Which Age group has a Higher and Lower chance of getting infected?

1. From the data Age group between 20–29 has a higher chance of getting infected
2. From the data Age group more than or equals to 80 has a lower chance of getting infected.



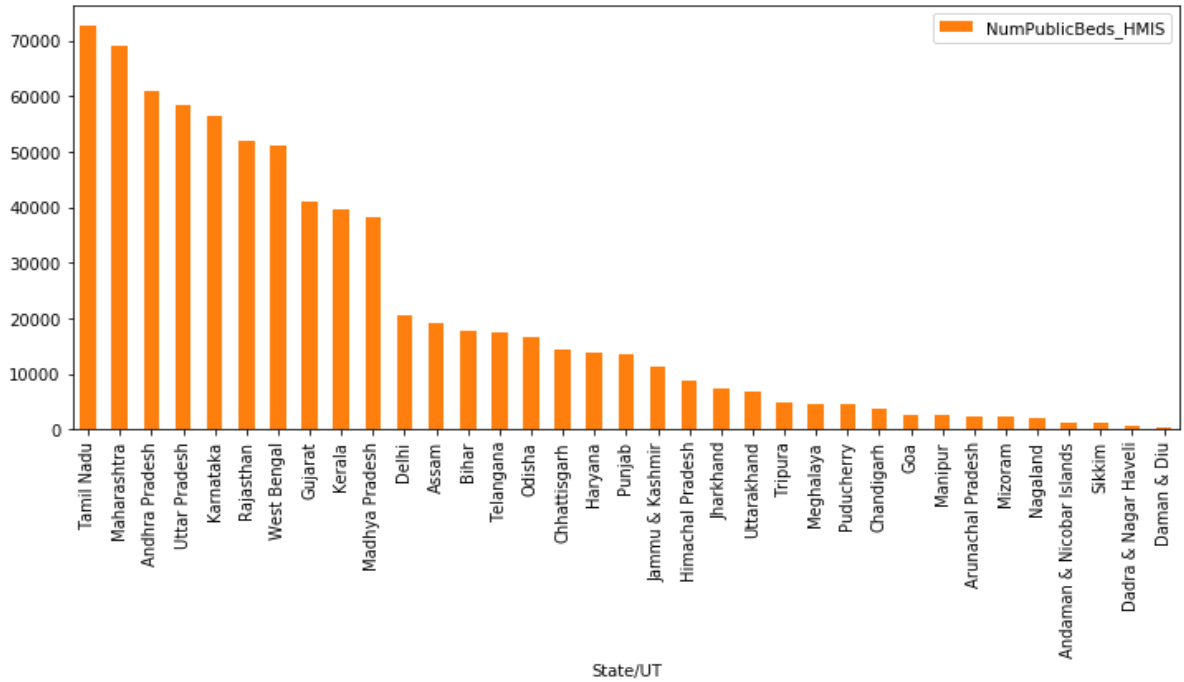
Q4. Which state has better Public Health care facilities?

Uttar Pradesh is top in terms of Total Public Health Care has 4122. According to the second edition of NITI Aayog's Health Index released



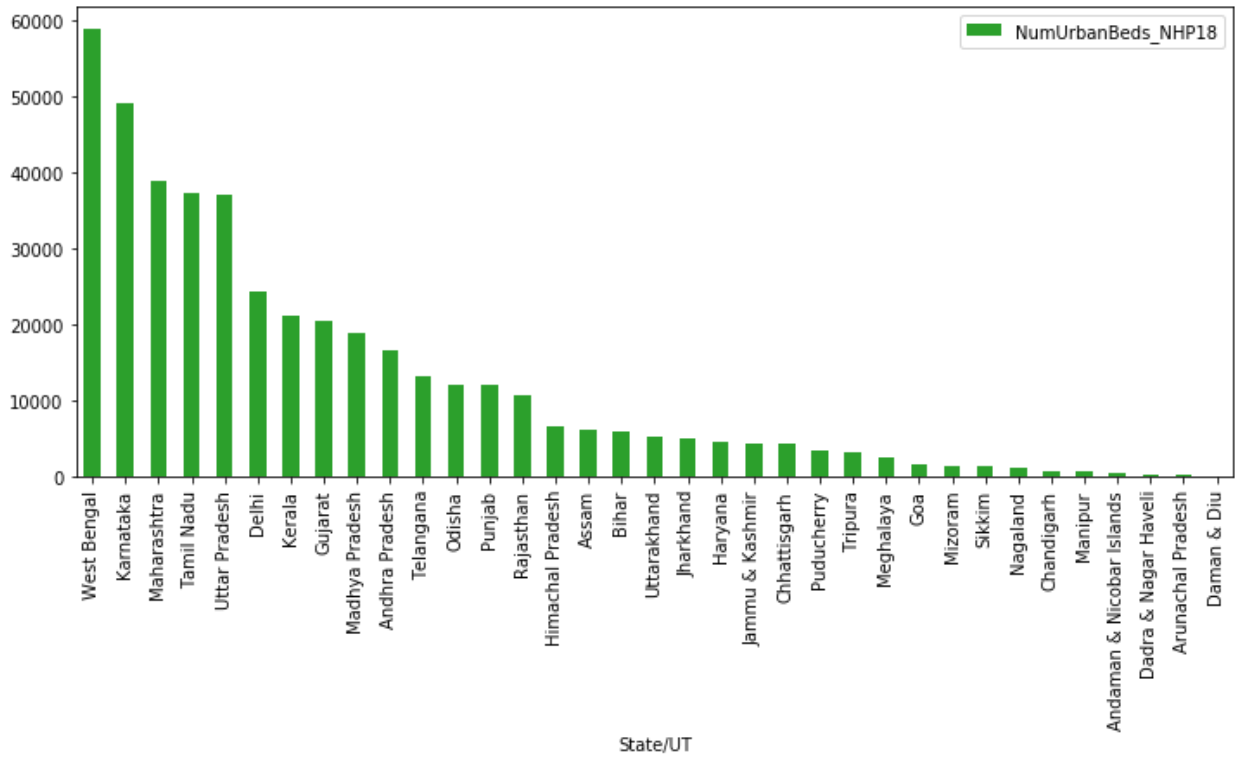
Q5. Which state has a higher number of Public Beds?

Tamil Nadu is top in terms of the Number of Public Beds has a total of 72616 beds.

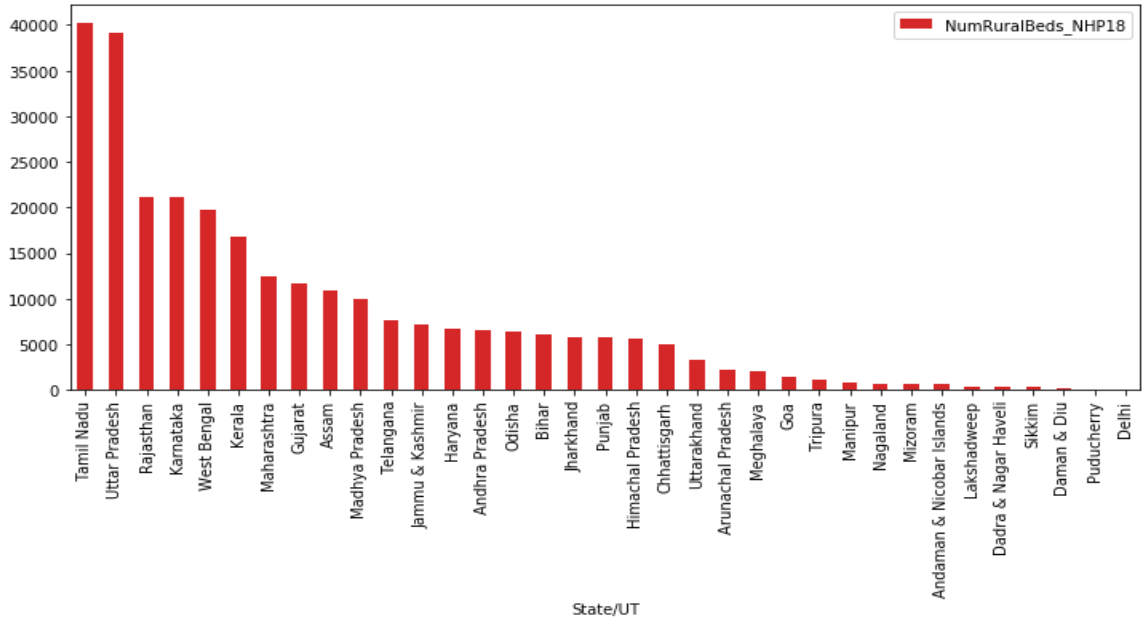


Q6. Which state has a Higher number of Beds available in Urban areas & Rural areas?

West Bengal is at the top in terms of the Number of Urban Beds that have 58882 beds.

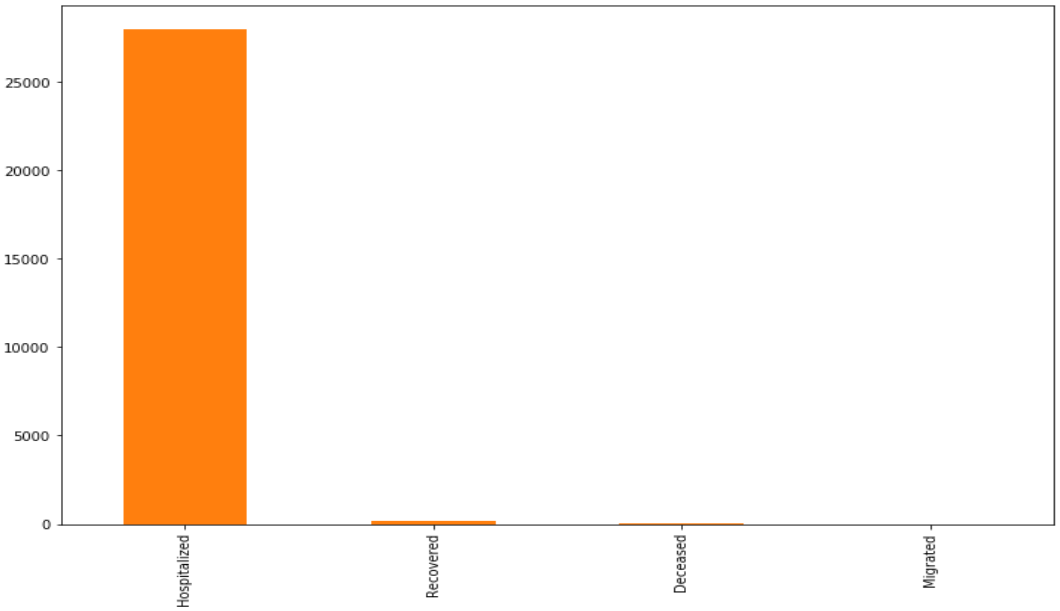


Q7. Tamil Nadu is at the top in terms of the Number of Rural beds has 40179 beds.



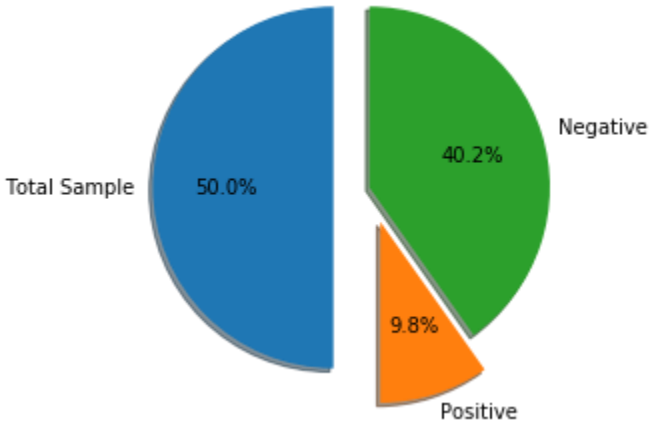
Q8. What are the number of people hospitalized, recovered, Deceased, and Migrated?

1. According to DataSet 27953 are Hospitalized.
2. According to DataSet 182 are Recovered.
3. According to DataSet 46 are Deceased.
4. According to DataSet 1 are Migrated.



Q9. Which state has the most number of sample testing on 12/07/2020?

1. Maharashtra did most number of sample testing which is 1321715.0 total sample on 12/07/2020.
2. The maximum total Positive sample for Maharastra on date 12/07/2020 is 259037.0.
3. The maximum total Negative sample for Maharastra on date 12/07/2020 is 1062678.0.



Q10. What percent of the chance that people are tested positive and negative in Maharastra?

1. The chance of being tested Positive is 19.5985 %.
2. The chance of being tested Negative is 80.401 %.

Value of Data :

• This dataset can be used by other researchers to implement automatic mechanisms (through Natural Language Processing, for instance) to extract insights contained on the metadata (e.g., abstracts and keywords) of scholarly studies; • This dataset can also be used together with other datasets (e.g. Publon and Google Scholar) in order to get a more accurate overview of research related to COVID-19 and identify possible research gaps that have not yet been explored to combat COVID-19;
• Several insights can be extracted from the relationships among various entities (e.g. drugs, researchers and their affiliations) applying techniques of Complex Network Analysis in this dataset;
• The most influential researchers or research groups can be identified to initiate new possible collaborations or task forces to combat COVID-19 pandemic.

1. **Data Description**

The dataset available in this paper is composed of 40,212 records of metadata about the publications related to COVID-19. Such data were collected from Scopus [7] , PubMed [6] , arXiv [4] and bioXiv/medRxiv [5] databases, and correspond to productions whose publication year is 2019 or 2020, and were published, indexed or made available until 07/02/2020 (date of data collection).

2. **Experimental Design, Materials and Methods**:

The process of data collection can be seen in Fig. 1 . For each database, a specific dataset was generated using Web Scraping tools and techniques [1] to collect its respective metadata. Next, using Pandas tool [8] , all specific datasets were preprocessed (data cleaning and normalization,

Methods

In March 2020, the World Health Organization (WHO) declared COVID-19 a pandemic, caused by the novel SARS-CoV-2 virus. Following the call from the WHO to immediately assess available data to learn what care approaches are most effective and evaluate the effects of therapies, this collection aims to report on original peer-reviewed research articles in methodological approaches to medical research related to COVID-19.

Data collection

Monitoring the health situation, trends, progress and performance of health systems requires data from multiple sources on a wide variety of health topics. A core component of WHO's support to Member States is to strengthen their capacity to collect, compile, manage, analyze and use health data mainly derived from population-based sources (household surveys, civil registration systems of vital events) and institution-based sources (administrative and operational activities of institutions, such as health facilities).

Searches and sources

We used databases that were created for the Zika Open Access Project (ZOAP) [11] and COVID-19 Open Access Project (COAP) [12]. Both databases are maintained by the authors and are used to conduct living systematic reviews [13, 14]. For each pathogen, we ran daily automated searches to index and deduplicate records of articles about Zika virus (from January 1, 2016) and SARSCoV-2 research (from January 1, 2020) in EMBASE via OVID, MEDLINE via PubMed, and the preprint server bioRxiv (for SARS-CoV-2 we also searched medRxiv). These data have been collected and deduplicated daily for several living systematic reviews and detailed methodology is described elsewhere [11, 13–16]. We specify the search terms in the Additional file 1 Text 1.

Annotation of records with study design

We screened the title and abstract, or full text when the first was insufficient, and annotated each record with its study design. For weeks where the volume SARS-CoV-2 of research was over 400 publications, starting midMarch, we drew a random sample of 400 publications with the R 'sample' function, without replacement. The number of selected publications was a pragmatic decision that balanced an adequate sample size and manageable workload for the number of crowd-volunteers. The annotation of the Zika virus dataset was performed for previous systematic reviews (from January 1, 2016 to December 31, 2016) [13, 15]. We first classified publications into the broad groups "epidemiology" or "basic research", "non-original" articles (editorials, viewpoints, and commentaries) and "other". These are groups that we used in an earlier study about Zika virus [7], so for this comparative study, we applied them to the publications about

SARS-CoV-2. We subdivided epidemiological and basic research further, based on their study design. We provide details on the classification of the study designs in the Additional file 1 Table S1 and in an online annotation guide [17].

Reported number of cases

To compare the number of publications against the number of reported cases, we used open-source data on Zika virus and SARS-CoV-2 from https://github.com/ andersen-lab/zika-epidemiology/tree/master/pahocase_ numbers and https://ourworldindata.org/covid-cases., see Availability of data and materials

Date that a publication becomes available

We defined the date at which a publication became available as the date it was indexed in the MEDLINE or EMBASE database, or when it appeared on the preprint server.

Data analysis

First, we described the evolution of reported cases and publications over time. Second, we described the proportions of study designs, by week, for SARS-CoV-2 and by month for Zikavirus, due to the differences in research volume. We omitted the first two weeks of 2020 for SARS-CoV-2 because there were only four publications, making the proportions unstable.
To take into account the random sampling of the SARS-CoV-2 research, we provided the Wilson score 95% confidence intervals (CI) for the proportions. Third, we quantified the timing and speed of the accumulation of publications of different study designs: We plotted the time elapsed between the first and twentieth occurrence of publications of each study design. Last, we described the proportion of evidence that was published on preprint servers during the two epidemics, and by study design. All analyses were conducted in R 4.0.1.

Chapter 5: Results and Conclusion

Between week one and week 21 (up to May 24) 2020, we indexed 21,990 publications, and a crowd of 25 contributors annotated a sample of 5294 (24%) publications on SARS-CoV-2. For the Zikavirus research, we annotated all 2286 identified publications for 2016. Both the volume of the weekly reported cases and number of publications were 30–50-fold higher for SARS-CoV-2 than for Zikavirus .

The proportion of different study designs

In both epidemics, a substantial and reasonably stable proportion of the publications were non-original research. The overall proportion of non-original publications was higher for Zika virus (55%, (Additional file 1 Table 2)) than for SARS-CoV-2 (34% [95% CI: 33–35], (Additional file 1 Table 3)). For publications of original research, the proportion of basic research publications increased over time for Zika virus, but decreased for SARS-Cov-2 research

Within the epidemiological study designs, mathematical modelling studies had a larger role at the beginning of the SARS-CoV-2 pandemic (10.1%, [95% CI: 9.3– 11.0]) and compared to the Zika virus outbreak (3.2%). Many of these were published as preprint publications. When we excluded preprint publications, the evolution of evidence over time became more similar between the two epidemics .
Case reports and case series accounted for approximately 10% of the total body of evidence; 10.7% [95% CI: 9.9–11.6] for SARS-CoV-2 and 9.7% for Zikavirus research. Analytical epidemiological study designs became more prevalent

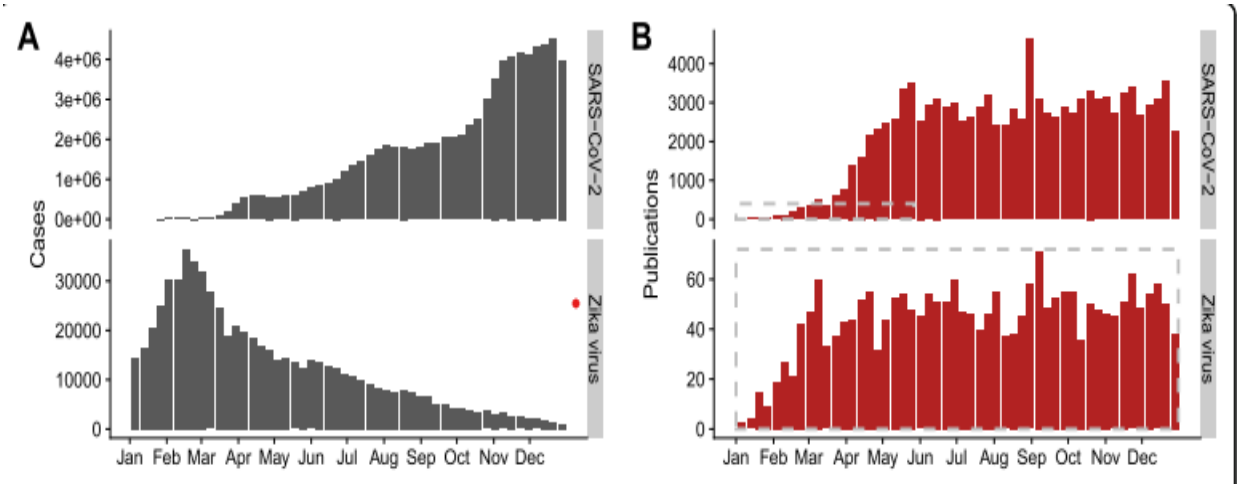


Fig. 1 The global number of reported cases (**a**), and the number publications (**b**) by week for SARS-CoV-2 infections in 2020 and Zika virus infections (ZIKV) in 2016. In panel B, the dashed grey boxes contain the period and number of publications for which the study design was annotated. The vertical scales differ for each infection. SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

Broadly speaking, Open Science aims to optimize scientific conduct and communication by exposing the scientific process, and results thereof, to the scientific community and broader public.

This idea is implemented concretely through a number of core Open Science practices [8, 11, 12]: Open Access, Open Source, Open Data and Open Peer-Review. The best-known of those, Open Access, consists of making all scholarly communications freely available with full re-use rights. Open Access also encompasses early dissemination of manuscripts in the form of preprints (articles not yet published in scientific journals).

The term “preprint” can refer to either pre-peer-review manuscript or peer-reviewed manuscript before formatting. Hereafter, we use the wording “preprint” to refer to pre-peer-review manuscript posted on archival websites.

Even though preprints are not yet peer-reviewed and thus could contain mistakes which may have been identified through an independent review process, they contribute to a more transparent and open scholarly publication system, accelerating re reviewing and communication within the scientific community Open Source and Open Data aim at ensuring that materials such as questionnaires, forms, procedures, collected data, metadata, and source code are shared to foster replication studies, increase data -use, and facilitate the peer-reviewing process .

Indeed, reviewers have the material at hand to verify the findings or detect any issues that could not be otherwise identified from the manuscript itself and to provide comprehensive peer-review reports. Then, following the Open Peer-Review principle, these peer-review reports should be publicly and transparently shared, along with the authors’ response. The scientific discussions between authors and reviewers are inherent to the process of creation of knowledge . In addition, Open Peer-Review helps maintain high reviewing quality and reduces the risk of concealed conflicts of interest. Therefore, the adoption of Open Science principles in the last decade has been particularly helpful in increasing the rigour, reliability and reproducibility of scientific results across research field .

There is evidence suggesting that the COVID-19 pandemic has served as a catalyst in the adoption of certain Open Science principles. For instance, major publishers such as Elsevier and Springer Nature have made newly written COVID-19 related articles freely accessible to all (Open Access). Furthermore, authors have shared their preprints more systematically than in previous pandemics and reviews have been posted on external platforms

Specific initiatives, such as OpenSAFELY , have emerged to make data available to researchers while complying with the legislation regulating the use of medical data. Nevertheless, there have been many instances where these principles were ignored. One notorious example is the

lack of transparency and sharing of the data provided by Surgisphere , which led to the retraction of the publication in The Lancet .

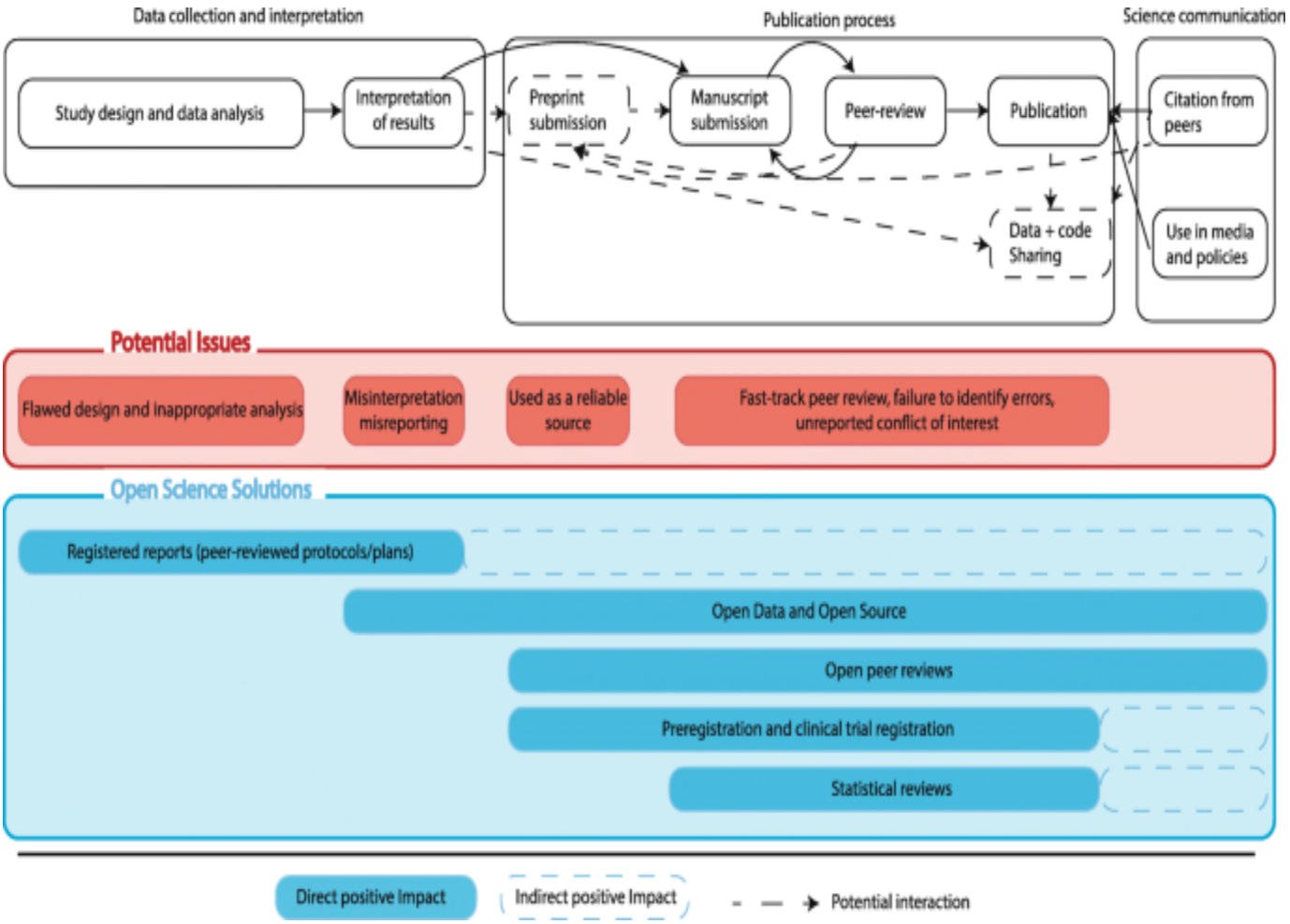
In other instances, some of the Open Science principles were adopted but misused. For example, news agencies have reported unreliable results based on the misuse of unreviewed preprints and some open reviews took place on separate platforms (for example Pubpeer), and were thus not directly available to readers.

While we recognize that the faster embracing of Open Science during the pandemic is a step towards more accessible and transparent research, we also express concerns about the adoption of these practices for early and non-validated findings. Furthermore, embracing only some of these principles, while excluding others can have serious unintended consequences that may be as detrimental as not adopting open practices in some instances. The aim of the present paper is twofold.

First, we identify the issues the scientific community has faced with regard to the publication process since the beginning of the pandemic. To do so we analyzed data collected on preprints and published COVID-19 research articles, as well as on retracted COVID-19 publications, in order to quantify issues related to reviewing time, conflicts of interest, and inappropriate coverage in the media. In light of this analysis, we then discuss how a wider adoption of Open Science principles could have potentially minimized these issues and mitigated their impact on the scientific community and broader public.

The structure of this article follows the stages of the publication process shown.

We first discuss issues arising at the data collection and interpretation stage (before the dissemination of the results). Then, we review the dysfunctions observed during the publication process (between the submission and the publication of research articles), before investigating the misuses of research outputs during science communication (after publication). We provide recommendations based on Open Science principles for each stage of the publication process, which we hope will contribute to better research practices in the future.



Outline of the publication process with its potential issues and our proposed solutions

While previously deplored, waste of scientific effort has been particularly prominent during the COVID-19 pandemic and has been more visible than ever before. In this section, we show that this waste has its roots in the early stages of the research process – at the data collection and interpretation stage – and discuss how study preregistration, registered reports, adherence to reporting guidelines and Open-Source principles could help to minimize waste in research.

Identified flaws:

Methodological and statistical issues

Conducting research during a pandemic is known to pose particular challenges but scientists have raised concerns about methodological flaws in the design and analysis of various COVID-19 pharmacological studies.

To better understand whether inappropriate study designs or statistical analyses contributed to the reasons behind the retraction of articles, we looked at the 29 COVID-19-related papers that had been identified on the COVID19 Retraction Watch dabatase as retracted or subject to expressions of concerns since January 2020 . While the list is still updated by RetractionWatch, our original list of articles (both preprints and peer-reviewed) and the results of our analyses are available on the repository of this project: https://osf.io/renxy/. Of the 29 identified publications, 8 (27.6%) were retracted (or had an expression of concern from the editorial board) based on their data analysis or study design.

More specifically, among these 8 publications, 2 (25.0%) papers were retracted, at the authors' request, in order to conduct further data analyses and 6 (75.0%) were retracted because the methodology or the data analysis was wrong. Out of these 8 retracted papers, 3 were preprints and did not undergo peer-review, however 2 of them stand out from the peer-review data we could extract: one was peer-reviewed in less than 3 days and the second displayed editorial conflict of interest, thus highlighting the need for a more transparent review process.

Duplication of research

Another concern is the increased risk of research waste due to duplication. Many studies that aimed to assess the efficacy of hydroxychloroquine were conducted in parallel: 218 registered trials were ongoing or already completed as of 26 th April 2020 . Many comparative effectiveness studies – randomised or not – were conducted without preregistration

however, meaning that the broader research community only became aware of these studies at the time of the release of the results. This illustrates the general lack of cooperation between research teams, putting more patients at risk by exposing them to potentially harmful treatments in multiple underpowered studies, and also leading to a waste of research time and financial and human resources . Given the additional workload for healthcare workers and clinical researchers these trials require, it may have contributed to the disruptions in the conduct of clinical trials during the pandemic . Other studies have been pre-registered but conducted and reported with major deviations from the preregistration record without

justification: for example, outcome measures and their timing of assessment reported in the aforementioned study by Gaiter et al. we're not those listed on the EU Clinical Trials Register

Ethical concerns

Ethical concerns have also arisen during the pandemic. While the research community needs to find ways to provide timely solutions to the COVID-19 crisis, it should not be at the detriment of good research and clinical practice. Among possible ethical risks. identified over-recruitment in trials, the conduct of human vaccine studies before the completion of animal studies, and the neglect of adverse effects in drugs studies.

An example of the last is the little consideration given to the known cardiotoxicity of the combination of hydroxychloroquine and azithromycin early on in the pandemic . Issues surrounding patients' participation in clinical studies have also been observed: in her analysis of COVID-19 papers unsuitable for publication, Bramstedt identified issues surrounding informed consent as the second most common source of concerns . In addition to the ethical problems this poses, it could also weaken the trust that patients and the broader community afford researchers, with detrimental consequences for public health in the long term.

Open science solutions

Here, we argue that the adoption of certain Open Science principles could have helped to detect or avoid the issues in data collection and interpretation described above . Two methods seem to be particularly relevant:

Study preregistration

First, study preregistration on dedicated platforms (e.g., ClinicalTrials.gov, OSF, or As Predicted), with a thorough description of the study design, ethical approval, methods for data collection and data analysis, can help prevent some of the issues identified above .

Indeed, study preregistration may reduce the amount of unnecessary duplication of research as researchers will be able to check whether specific studies are ongoing and design theirs to address complementary questions. Finally, study preregistrations can be used by Institutional Review Board for ethics approval and to fulfil the ethical obligation to transparently inform both the public about ongoing trials as well as the research community.

Another goal of preregistration is for readers and reviewers to make sure that a published study has been conducted and analysed as planned, thus limiting the risks of changes to the design, methods or outcomes in response to the data obtained other than the flexibility allowed by the protocol (in case of interim analyses of adaptive designs).

Researchers should register studies prior to data collection. On the platform ClinicalTrials.gov, retrospective registrations or updates to the study protocol are flagged. Depending on the level of methodological details in the record, standardized study preregistration may help in limiting questionable research practices such as Harking , p-hacking and p-fishing and eventually lead to better subsequent reporting .

As **COVID-19** was a new disease, there was no standardized diagnostic criteria or clinical outcomes.

This led to a multiplication of different outcomes studies in the articles participating in the difficulty to replicate and compare results.

Study preregistration could help researchers adopt the same criteria and outcome measurements and promote the use of validated international standardized criteria for variable and outcome measurements.

However, such preregistrations have two major limitations. First, they do not fully prevent duplication. While replication (defined as a deliberate effort to reproduce a study to validate the findings) is an important step of the research process, duplication (an inadvertent repetition of the research) contributes to research waste .

This waste has been noted among COVID-19 research , with a strikingly high amount of duplication despite study preregistration. Second, whereas preregistrations allow the detection of questionable research practices, they do not help prevent methodological issues before data collection since the preregistration is not itself peer-reviewed and the statistical analysis section of these records is often very brief.

Therefore, standardized study preregistrations are necessary, since they encourage researcher to outline the study design and analysis strategy, but not sufficient to avoid the excessive waste of scientific resources.

Registered report

Peer-reviewed study protocols, also called registered reports [41, 42], can also have a major impact on the reduction of wasted resources.

They essentially consists in articles with a two-stage peer-review, and provide details about the research question, hypotheses, methodology, statistical analyses and reporting strategy.

Since protocols are peer-reviewed before the enrollment of participants and data-collection, potential omissions or mistakes in the proposed methodology can be corrected before any substantial resources are used, thereby limiting scientific waste .

In these reports, researchers are also encouraged to provide details about the resources used, using for instance Research Resource Identifiers

when applicable, and specify the reporting guidelines that will be used (e.g., CONSORT [45], STROBE [46]). Registered reports can therefore contribute to higher quality research, with a reduced risk of bias and increased generalizability. One disadvantage of registered reports is that their reviewing takes time, while preregistrations are immediately available. However, some platforms for the submission of registered reports put in place measures to guarantee a timely review of COVID-19 protocols: stage 1 review of registered reports at Royal Society Open Science are performed within 7 days .

Furthermore, since they reduce the risk of publication bias, they also reduce the number of submissions needed to publish one's results and should ultimately save both resources and time .Both pre-registration and registered reports contribute to a better visibility of ongoing research, and should be used at institution levels to coordinate research projects at an international level in a more efficient way, in order to optimize resources.

Beyond registration: open methodology and reforming the publication system:

Preregistrations and registered reports are necessary but not sufficient to conduct reliable transparent research.

Open Methodology goes further. It consists in transparently sharing all the necessary details to allow replications of the research.

In other words, Open Methodology relies on the authors to not withhold any details of their research project so that any outsider could exactly replicate it. While one might assume that

preregistrations and registered reports are sufficiently detailed to allow replications, numerous past studies have shown that replicating research work is excruciating or impossible .

Journals should therefore support fully Open Methodology better by enforcing that all submitted articles must be fully reproducible before the manuscript is published. Beyond Open Methodology and registrations, researchers have suggested to reform scientific communication so that they would be less story-telling oriented and more focused on the methodology . Every step of the research methodology (research-question formulation, hypothesis-making, data-collection plan, data analysis, interpretation...) is a smaller paper that builds onto the previous one and all of them are open for comments and reviews in order to make insights more robust and foster collaborations.

Adopting this would, however, require a complete change of the scientific publishing system and further proof is needed to show its benefits.

During the COVID-19 emergency, IRST IRCCS, an Italian cancer research institute and promoter of no profit clinical studies, adapted its activities and procedures as per European and national guidelines to maintain a high standard of clinical trials, uphold participant safety and guarantee the robustness and reliability of the data collected.

This study presents the measures adopted by our institute with the aim of providing information that could be useful to other academic centers promoting clinical trials during the pandemic.

After an in-depth analysis of European and Italian guidelines and consultation and analysis of publications regarding the actions implemented by international no profit clinical trial promoters during the emergency, we monitored the way in which the institute managed clinical trials, verifying compliance with regulatory guidelines and clinical procedures, and evaluating screening and recruitment trends in studies.

During the pandemic, our center activated a new clinical trial for the treatment of patients with COVID-19. A number of procedural changes in clinical trials were also authorized through notified amendments, in accordance with Italian Medicines Agency (AIFA) guidelines.

Patient screening and enrolment was not interrupted in any site participating in multicenter interventional clinical trials on drugs.

The institute provided clear indications about essential procedures to be followed, identifying those that could be postponed or carried out by telephone/teleconference. All external sites were monitored remotely, avoiding on-site visits. Although home-working was encouraged, the presence of staff in the central office was also guaranteed to ensure the continuity of promoter activities.

Conclusion

I. Don’t touch your eyes, nose, or mouth.

II. Clean your hands often. Use soap and water, or an alcohol-based hand rub.

III. Maintain a safe distance from anyone who is coughing or sneezing.

IV. Wear a mask when physical distancing is not possible.

V. Cover your nose and mouth with your bent elbow or a tissue when you cough or sneeze.

VI. Stay home if you feel unwell.

References

1. Methodologies for COVID-19 research and data analysis - https://www.biomedcentral.com/collections/COVID19-research-methodology

2. Statistical Analysis of Covid-19(SARS-Cov-2) Patients Data of Karnataka, India - https://www.researchsquare.com/article/rs-72912/v1

3. A Comprehensive Analysis of COVID-19 Outbreak situation in India - https://www.medrxiv.org/content/10.1101/2020.04.08.20058347v1.full.pdf