EasyChair Preprint

№ 9596

Data Science Approaches to Public Health: Case Studies Using Routine Health Data from India

Arun Mitra, Biju Soman, Rakhal Gaitonde, Tarun Bhatnagar, Engelbert Nieuhas and Sajin Kumar

January 20, 2023

Data science approaches to public health: case studies using routine health data from India

Arun Mitra[1[0000-0001-6742-4033]], Biju Soman[1*[0000-0003-0748-0839]], Rakhal Gaitonde[1[0000-0003-4046-8380]], Tarun Bhatnagar[2[0000-0002-5603-1418]], Engelbert Nieuhas[3] and Sajin Kumar[4[0000-0002-6461-8112]]

[1] Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, India
[2] ICMR - National Institute of Epidemiology, India
[3] RPTU University of Kaiserslautern-Landau, Germany
[4] University of Kerala, Thiruvananthapuram, India
arunmitra@sctimst.ac.in
bijusoman@sctimst.ac.in
rakhalgaitonde@sctimst.ac.in
tarunbhatnagar@nie.gov.in
niehaus@rptu.de
sajinks@keralauniversity.ac.in
*Corresponding Author

Abstract. The promise of data science for social good has not yet percolated to public health, where the need is most, but lacks priority. The lack of data use policy or culture in Indian health information systems could be one of the reasons for this. Learning from global experiences on how routine health data has been used might benefit us as a newcomer in the field of digital health. The current study aims to demonstrate the potential of data science in transforming publicly available routine health data from India into evidence for public health decision-making. Four case studies were conducted using the expanded data sources to integrate data and link various sources of information. Implementing these data science projects required developing robust algorithms using reproducible research principles to maximize efficiency. They also led to new and incremental challenges that needed to be addressed in novel ways. The paper successfully demonstrates that data science has immense potential for applications in public health. Additionally, data science approach to public health can ensure transparency and efficiency while also addressing systemic and social issues such as data quality and health equity.

Keywords: health information systems, data science, public health, decision-making, deep learning, digital health, ICT4D.

1 Introduction

The dawn of the digital era has transformed our society in many ways. Rapid advances in technology and innovation have brought about unprecedented changes in our daily

lives. These changes brought about by the digital revolution have also touched all sectors and industries. Healthcare likewise has evolved and is well launched down this digitalization journey.(1–3) Innovations like drug discovery, vaccine research, diagnostic modalities, and therapeutic interventions have significantly impacted human life expectancy and how humans live their lives. In addition, evolving habits, interactions and lifestyles brought about by technological innovations have directly and indirectly affect human health.

Over the last few decades, the emergence of the Information Age has accelerated this transformation opening windows of opportunities to new possibilities. It has become increasingly clear that data collected in the healthcare industry has become more complex, huge in volume and generated faster than ever before.(4,5) The types of data generated through health information systems have also increased significantly, moving from paper-based systems to machine-generated data such as medical imaging, wearable technology, IoT devices, remote sensing sensors, satellite imagery etc.(6,7)

Data on health is considered a national asset for every country. It helps track the performance of the health system. It provides reliable information for decision-making across all building blocks. It is also vital for monitoring the overall objective of health systems, such as the changes in health status and outcomes over time. Therefore, health information is necessary to take many decisions at all levels, such as the local, regional and national levels. A national health information system typically collects crucial health-related data for various purposes. The health information systems are integral to the health system providing specific information support to the decision-making process at multiple levels. Just like a health system, the health information system is not static and must amend itself to the changes that occur with time. A health information system's primary functions are to monitor trends in health outcomes and services, take decisions relating to public health quickly and efficiently, identify the best strategies for public health interventions, and ensure the coordination and equity of health services. At the same time, ensuring that data are trustworthy while also managing resources for optimal use and benefit. The routine health information systems constitutes data that is collected and coallated at regular intervals at different levels from health facilities and health programs including public, private, and community-level. This data provides a picture of health status of the population, health services delivered by the health system, and health resources available for utilization. The data from the routine health information system is both actively and passively captured by healthcare providers as they go about their work, by their supervisors for the purposes of monitoring, and through routine health facility surveys for the purposes of health governance and administration. The role of a routine health information system is critical to achieving Universal Health Coverage (UHC) and it is of paramount importance for decision making in public health.

Traditional health data has been limited to clinical health information systems and data from the vertical national health programs have been used for surveillance and monitoring purposes. Digitalization of paper-based systems, implementation of electronic medical/health records and integration of multiple databases including sociodemographic, climatic, environmental and economic data routine health information

systems have ensured that there is a need to expand the RHIS into an expanded RHIS which is more dynamic, responsive and proactive than the traditional RHIS. However, one of the major challenges faced by the expanded RHIS is issue of transforming data into evidence for decision-making in public health. Existing data analysis methods and techniques have their limitations, owning to not just the volume and velocity of the generated data but also the evidence generation needs to be done in a timely and transparent fashion to earn public trust. The true challenge is transforming data into information and knowledge for public health benefit. It has become evident that innovative solutions to generate robust, near-real-time, high-quality, actionable evidence for public health decision-making are the need of the hour.

With this background the current study aims to demonstrate the utility and potential of data-science as a tool to address the challenge of transforming routinely available data into evidence for informed decision making in public health.

2 Review of Literature

2.1 What is routine health data?

Routine health data often refers to non-targeted data obtained passively from different health information systems. They are collected continuously at various periods (daily, weekly, monthly, annually etc.) and can be collected individually patient by patient or aggregated at the family, facility or geographic level. They come from the existing national or regional health information system and its subsystems that are collected as part of an ongoing routine.

It can be categorized as (a) demographic data, (b) health (disease) events data, (c) population-based health-relevant information and (d) data from other disciplines such as climatic data, meteorological data etc. One other way of classification of routine data could be as (a) administrative data, (b) clinically generated data, (c) patient-generated data, and (d) machine-generated data.(8)

2.2 Sources of routine health data?

There can be many sources of routine health data. They can be classified as standard sources or expanded sources.(6,9)

Standard. Standard sources are traditional sources of health data and can be divided into three different sub-divisions again – (i) health services, (ii) public health, and (iii) research. The sources of health services data could be clinical records, electronic health records, electronic medical records, prescription data, diagnostic data, laboratory information and data on health insurance. The sources relating to public health could be data

on disease surveillance, immunization records, public health reporting data, vital statistics, disease registries, civil registration systems etc. Research data sources could include demographic surveillance sites, omics or genomics data, clinical trial registries, bio-banks, medical devices, and the internet of medical things (IoMT).

Expanded. Expanded data sources are newly identified sources that were traditionally not considered sources of health data. These sources can be further classified into (i) environmental, (ii) lifestyle & socio-economic, and (iii) behavioural and social. Environmental sources include climate, meteorological, transport, pollution, forest cover and animal health. The data sources on lifestyle and socio-economic could be location-tracking information, financial data, education and data from various relevant mobile applications. Similarly, the behavioural and social data sources are data on wellness, fitness, internet use, social media, self-monitoring devices, wearables and IoT sensors.

2.3 Potential of routine health data

There is a lot of potential for using routine data in informed decision-making in public health. Some of these are it improves the utilization of health information being collected. It saves time by leveraging already available data and also adds no or low additional costs to the health system. It helps generate new hypotheses, testing these hypotheses and comparing across populations and high-risk groups. It also provides a baseline estimate of the expected levels of health and disease in a specific population and allows for comparisons across geographic regions. As it is passively collected information, it is possible to conduct natural experiments which are unbiased and aid in informed decision-making processes. It is also comprehensive and considers the forward motion of time, allowing for longitudinal analysis of health outcomes and states.

2.4 Indian Context

It is argued that there are three distinct stages of digital health evolution, and India has crossed the first stage of digitalization, where most paper-based systems are being captured in their electronic forms, including x-rays and CT films. As a country, India has now stepped into the second stage of digitalization where new forms of digital data are being collected, such as mobile health and health monitoring devices such as wearables, the capture of health information in the forms of electronic health records and electronic medical records, and telemedicine consultations including both public and private health facilities. To achieve complete digitalization, the healthcare systems (both public and private) must embrace digital workflows. This stage also gives rise to new opportunities and business models, which have already started in some metropolitan cities but are yet to percolate to the rural and grass-root level where the need is most. The final stage of the digital evolution is digital transformation, where the good quality data

from the health system (including routine health information systems, clinical health information systems, and both public and private health facilities) is envisioned to be leveraged through government initiatives like the National Health Stack, National Digital Health Mission, Electronic Health Record Standards, Integrated Health Information Portal Health ID, Universal Payment Interface etc.(6,10,11)

However, the current state of the Indian health information system reveals multiple challenges that must be addressed before achieving the ultimate digital transformation. Currently, routine health data is not utilized to its full potential in India. Just five percent of all healthcare data is collected in India, out of which only a tiny fraction is being used for public health decision-making. Some of these are the governmental focus on centralizing health information systems with a limited emphasis on supporting local action, the use of proprietary platforms, and working in silos. Some issues that might arise could be the probability of data breach, data storage and ownership, data quality and standardization.

In-depth studies on the use of data in the Indian health sector suggest that the decision-making process is central to the health system. To function efficiently, the decision-making process should incorporate an iterative cycle of generating data demand, data collection and analysis, information availability, and the information used for decision-making. Implementing this will lead to improved health decisions and accountability.

Some of the recommendations arising from the research are to train health care professionals on the importance of data use and encourage its use in the decision-making process. Additionally, support all stakeholders, involve civil society groups, conduct comprehensive needs assessments, implement data standards, and conduct regular data audits.

2.5 Challenges of routine health data

Some of the critical challenges and questions to consider when using any data collected by the routine health information system to infer population health are as follows:

Data Quality. Is the data of good quality? To what extent is the data accurate in capturing the natural phenomenon? Are there any inherent data quality issues or biases present in the data?
Precision. Can we comment on the level of uncertainty in the data? Can we provide a confidence interval or any other measure of uncertainty along with the estimate?
Completeness and Coverage. Is the data from the routine health information system complete? Is it representative of the population at hand? Is there any missing data? If yes, how much? Are there any patterns of missingness?
Timeliness. When was the data in question collected? Is it still relevant? Are any newer, more recent data sources that might be more relevant?
Analysis. What are the different analytical approaches that have been applied to the data? Has the data from the routine health information system been analyzed adequately? Has there been any standardization done to this data? Is there adequate

documentation mentioning the preprocessing done? Has it been done in a way that can be reproducible and peer-reviewed?
Accessibility. Who owns the data at hand? Who has access to this data? Who controls access to this data?
Confidentiality. Is the data at hand confidential? Does it contain any sensitive information? Can individuals be identified from this data? Has it been anonymized?
Original purpose of collection. Some of the data may contain personal information and may not be used for any other purpose unless informed consent.
Motivation. The other major challenge is that all the stakeholders involved in the data generation process of a routine health information system may not be adequately motivated.

Some additional challenges associated with the use of routine health information systems are given below.(12)
Technical Determinants. Examples of these are the lack of explicit information such as gender, age etc.; different definitions in different formats (paper vs electronic); data on emerging infections and diseases lacking, and lack of integration with data from other sources at the community level.
Behavioural Determinants. Some of the behavioural determinants are misclassification of conditions by healthcare workers; incomplete data collection; incorrect input due to factors like recall bias by the healthcare worker; delay in submission of data which may sometimes be due to the delay in salaries; data errors due to computational mistakes; difficulty in understanding feedback from the central level.
Environmental / Organizational Determinants. Shortage in precious resources, including time, monetary and human resources; lack of trained personnel; non-prioritization of local needs in decision-making; inadequate data analysis and interpretation, lack of inter and intra-departmental coordination on data sharing and many more constitute the environmental/organizational determinants.

Some of the lessons we can learn from the experiences of other countries are outlined in Table 1.

Table 1: Lessons learned from use cases of routine health data globally

Countries	Lessons learned
Australia(13,14)	Suggest the importance of **continually engaging and incentivizing national and local stakeholders.** Within the Australian government structure, there are **defined roles and responsibilities for the different layers of government, advisory committees, and research centers.** Department of Health and Ageing allows these groups to **collaborate better and support each other in decision-making activities**

Europe(14–16)	**Strong political will** and support for the collection of credible, population-based data **Efforts of non-governmental organizations** to improve data collection and measurement allowed Ensuring the data used for making decisions were **valid, reliable, and current**. **Data transparency** – the success of the UK's hospital waiting lists policy reform. Research using hospital waiting list data - regularly reported in the media and leading medical journals raising public and health professional awareness. Although the government understands the importance of collecting data and measuring policy effectiveness, of policies - must **re-compete for priority** in every political election cycle. Another barrier to long-term reform may be its sustainability. **Eventually, the costs involved may become higher than the incremental gains.** **Maintaining data collection over time** - important for **evaluating temporal trends and for determining whether a shift in policy focus will be required**
USA(13,14)	Retrospective medical record databases not only provide readily available data for timely decision making but the fact that the **same data are analyzed by policy decision-makers, academic researchers and the pharmaceutical industry enhances the credibility and transparency of the findings.** **Funding** for this research falls short of the demand. Researchers trained in pharmacoepidemiology, drug safety, and risk management are needed in the USA **to increase research capacity for this important policy-relevant work**.
Scandinavian Countries(17,18)	Improving **user-friendliness**, and adding even more health indicators to monitor the state of **social inequalities** in health.
Developing Countries	Reporting requirements must be able to **change over time.** Programme reporting requirements must be integrated in order to ensure the **development of coherent information** Need for a **hierarchy** of information needs Additional information can be collected through **specific programme surveys**

3 Case Studies

The authors chose to demonstrate the potential of public health data science as four case studies, each using publicly available data from different sources, formats and levels.

- Case Study 1 – using crowd-sourced data on COVID-19 in India at the national level
- Case Study 2 – using mortality data from the civil registration system at the panchayat level

- Case Study 3 – using periodic survey data to inform policy on maternal health at the district level
- Case Study 4 – using medical image data and design innovative solutions with applications in tele-medicine and tele-ophthalmology.

3.1 Case Study 1 – the AMCHSS COVID-19 Dashboard

For this case study, a translational research design with a data science approach was chosen. The data sources used for this case study include:

- Crowd-sourced database and website maintained at https://covid19bharat.org.
- State Level Daily Medical Bulletins (for the state of Kerala)
- District level population density based on 2020 projections by NASA Socio-economic Data and Applications Center (SEDAC) NASA and the Unique Identification Authority of India, Govt India.(19,20)
- Population Mobility Data from Google (https://www.google.com/covid19/mobility/)
- Indian COVID-19 Genome Surveillance Data - (https://clingen.igib.res.in/covid19genomes/)

The tools used were open source statistical programming language R and the RStudio IDE.(21,22) The study was implemented adhering to the tidy data principles and the tenets of reproducible research recommended by the scientific community.(23–27) The packages used were available from CRAN and from the R Epidemics Consortium (ReCon).(28,29) The fill list of packages is available below (see Table. 2). The detailed methodology is described elsewhere.(30–32)

Table 2. Packages used for the implementation of the COVID-19 Dashboard

Clean, Tidy, Link	Epidemiology	Visualization	Dashboard
`tidyverse`	`incidence`	`ggplot2`	`flexdashboard`
`dplyr`	`R0`	`plotly`	`shiny`
`fuzzyjoin`	`projections`	`dygraphs`	`shinybulma`

The case study results can be viewed as a live dashboard which is hosted at https://amchss-sctimst.shinyapps.io/covid_dashboard/ for public viewing (see Table. 3).

Table 3. Intended target audience for the COVID-19 Dashboard case study

Intended Users	Indended Purpose	Mode
District Level Program Manager	Evidence informed decision making; feedback on future improvements	Interactive Dashboard
Decision Makers	Evidence informed decision making	Dashboard
Academia and Reserachers	Discussion on methods; peer-review; academic discourse	Dashnoard + Methodology
General Public	Citizen involvement, public discourse, media and information professionals	Website / Blog

.

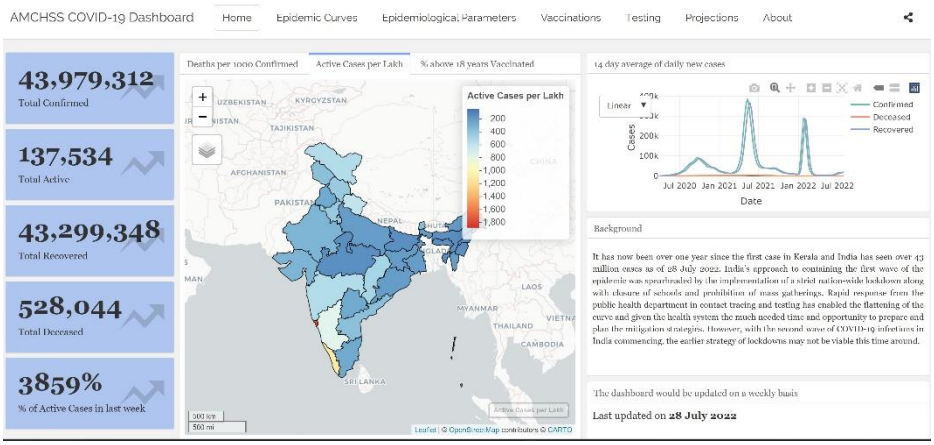

Fig. 1. Screenshot of the AMCHSS COVID-19 Dashboard (https://amchss-sctimst.shinyapps.io/covid_dashboard/).

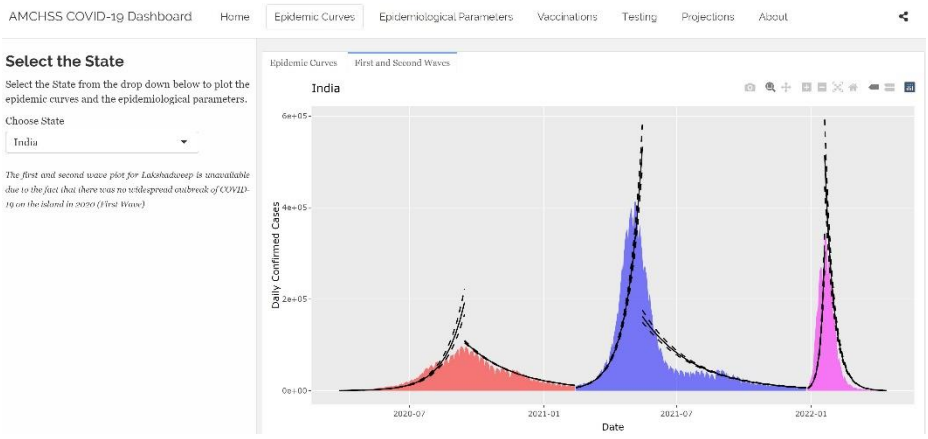

Fig. 2. Modeling epidemiological parameters based on different COVID-19 waves in India.

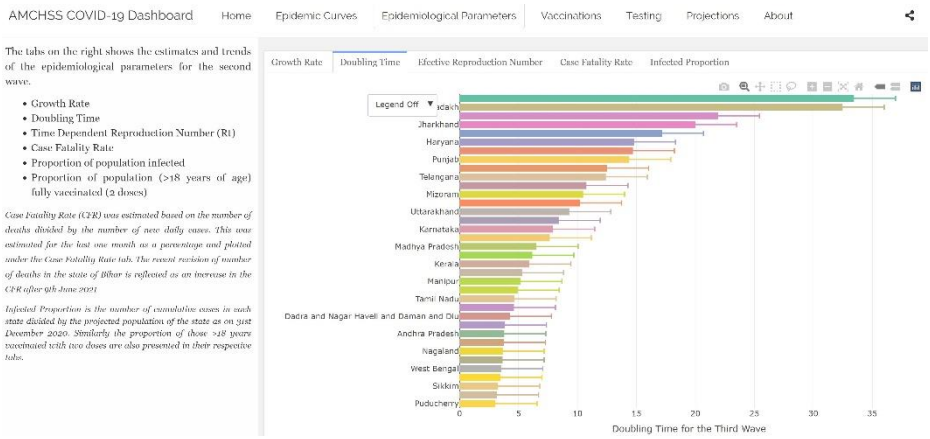

Fig. 3. Interactive visualization of the state-wise comparison of doubling time of number of COVID-19 cases in India during the third COVID-19 wave.

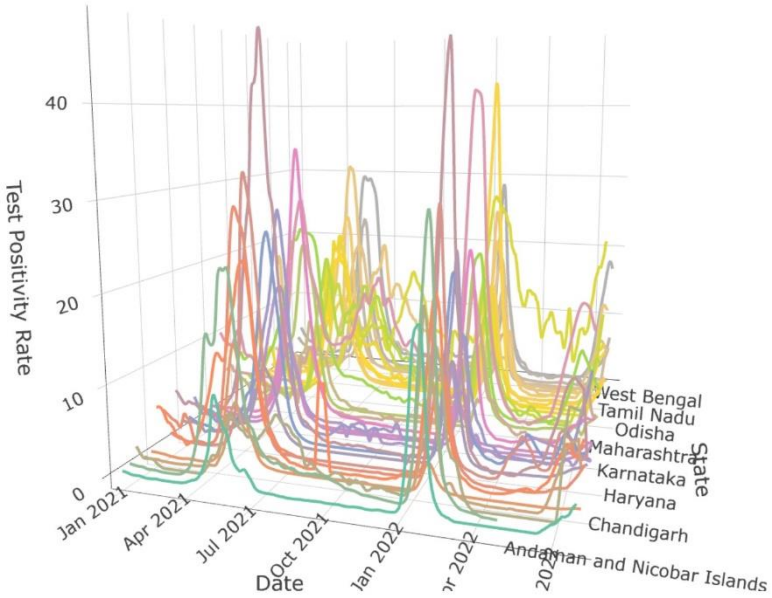

Fig. 4. 3D visualization of the trends in test positivity rate (TPR) across different states between January 2021 till July 2022.

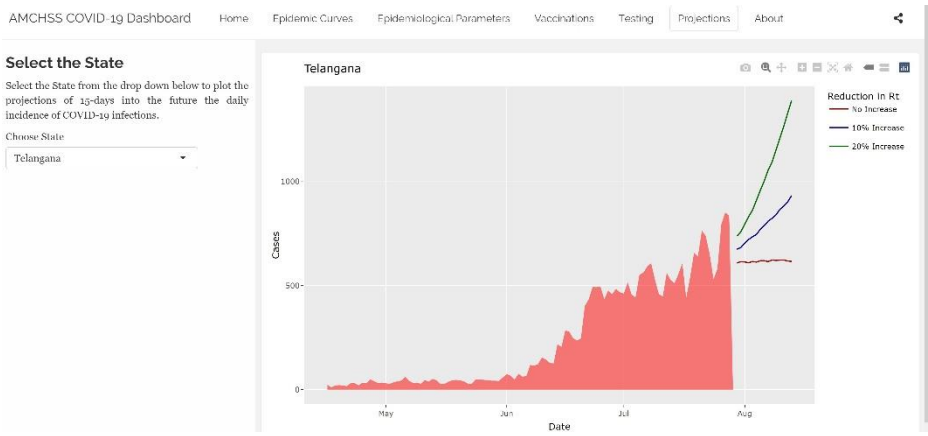

Fig. 5. Forecasting 15-day daily future incidence of COVID-19 cases based on the trajectory of time dependent reproduction number (projections for three different scenarios for the state of Telangana shown for illustration).

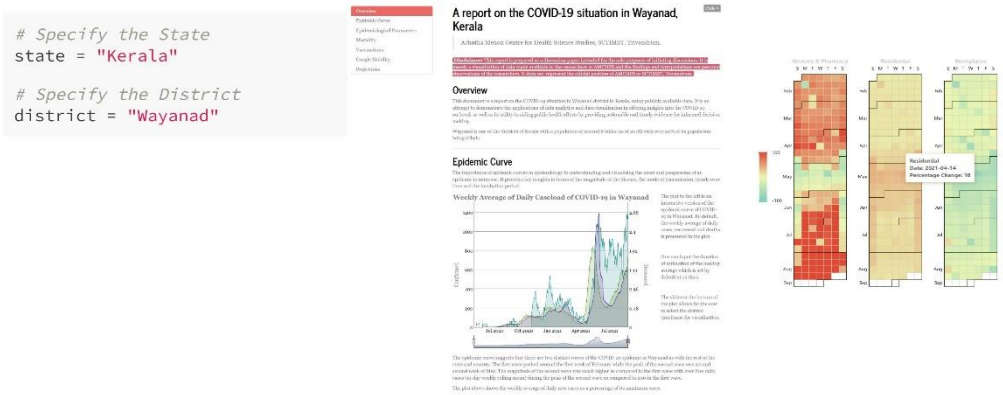

Fig. 6. An illustrative example of semi-automated report generation using reproducible algorithms for interactive and timely information for decision making in public health.

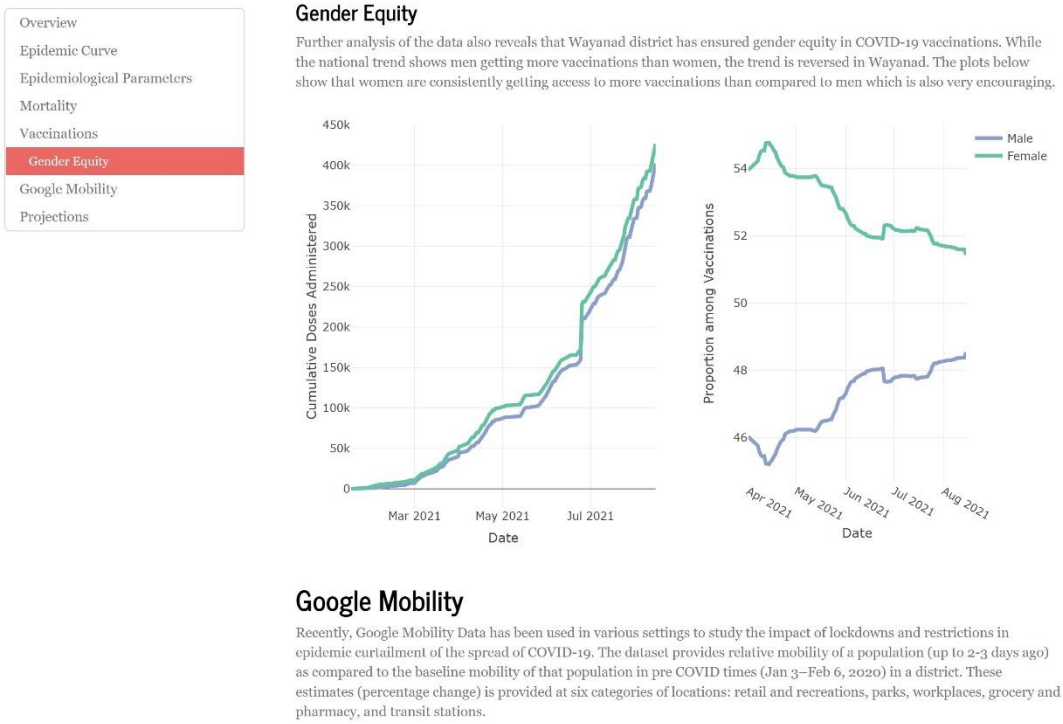

Fig. 7. A section detailing an in-depth analysis addressing issue of gender equity in access to COVID-19 vaccination using data from the COWIN portal.

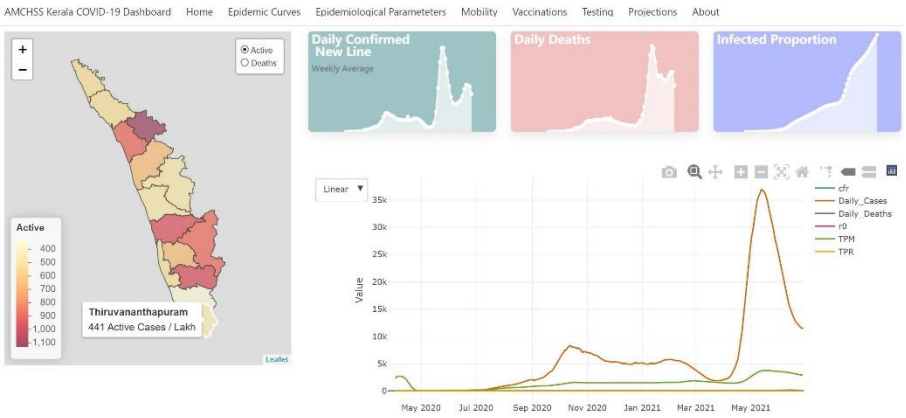

Fig. 8. An example of customizing the code for creating state-level dashboards for monitoring the status of COVID-19 at the district level (Kerala state for illustration).

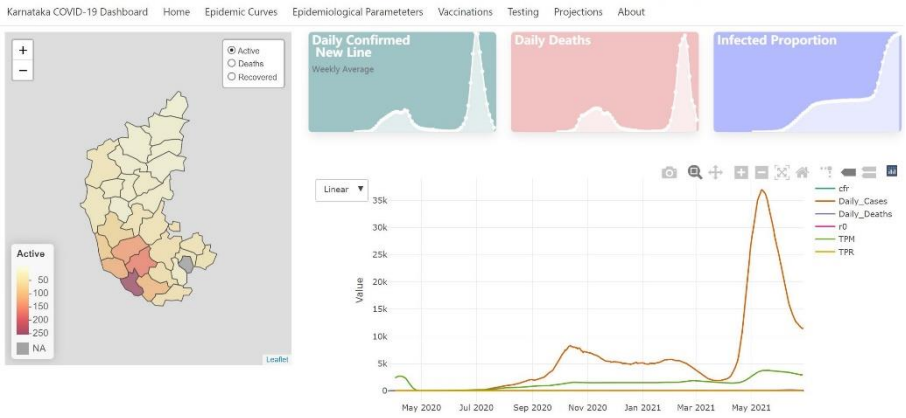

Fig. 9. A screenshot the dashboard for Karnataka state demonstrating the scalability of the algorithms to quickly create decision support tools for local action at the district level.

Fig. 1 - Fig. 9 illustrate the salient features of the dashboard. The figure description provides more insight into the feature and are self explanatory.

3.2 Case Study 2 – Cause of Death

The Civil Registration System (CRS) may be defined as a unified process of continuous, permanent, compulsory and universal recording of the vital events and characteristics thereof, per the country's legal requirements.(33)

CRS in India dates back to the mid-19th century. It started with the registration of deaths with a view to introducing sanitary reforms for control of pestilence and diseases. CRS in Kerala came into force on 1st April 1970. As of today, CRS in Kerala is computerized in corporations, municipalities and rural registration units (gram

panchayat). The registry is being supported by the 'Sevana' Civil Registration Software developed by Information Kerala Mission (IKM), set up for computerization of local bodies (https://cr.lsgkerala.gov.in). The CRS records prior to the date of computerization have been digitized and the issue of certificates is also computerized.
For this case study we chose CRS data from the Trivandrum Corporation. The study area is presented below (see Fig. 10).

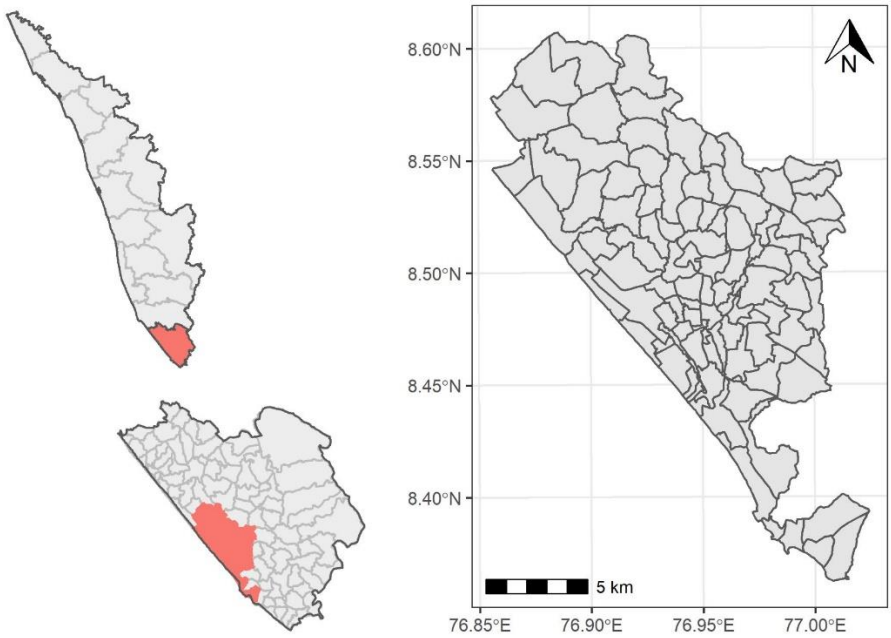

Fig. 10. Study Area: Kerala State, Thiruvananthapuram District, Trivandrum Corporation (inset).

Though data is being collected in real time, the analysis is being performed and being reported only once a year as part of the Annual Vital Statistics Report.(34,35) Some of the challenges with the current CRS in Kerala are as follows:

- Pre-occupation of Registrars and other functionaries with other duties.
- Excessive Delay in Registrations and Reporting.
- Inability to use the data for local action.
- Large proportion of infant and child deaths missed (esp. in rural areas).
- Data from various LSGs are collected, but unclear on the data use.
- Lack of user-friendly of information systems for the decision makers.
- Data is available for decision makers only as a PDF file which cannot be used readily for data analysis.

We sought to explore the use of data science address some of the challenges faced. The data extraction pipeline is presented in the figure below (see Fig. 11).

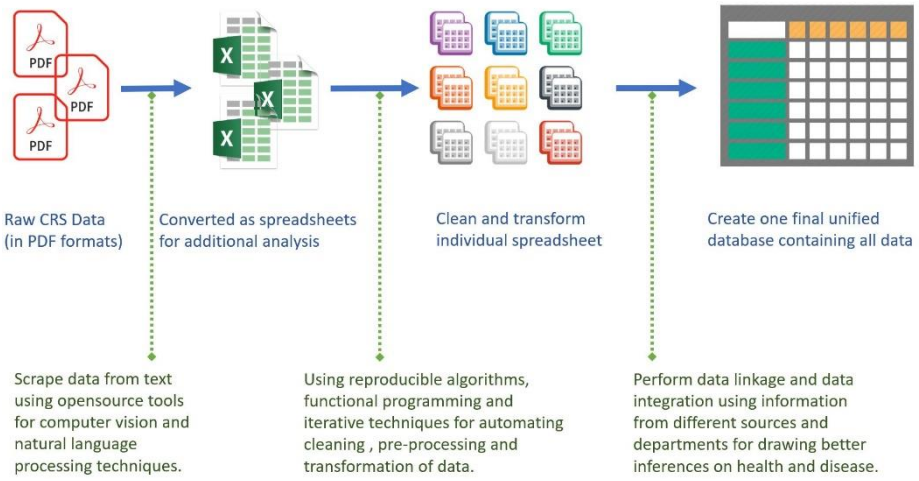

Fig. 11. The Data Extraction and Preparation Schema.

Results of the case study are presented as illustrations below (see Fig. 12 – Fig. 15).

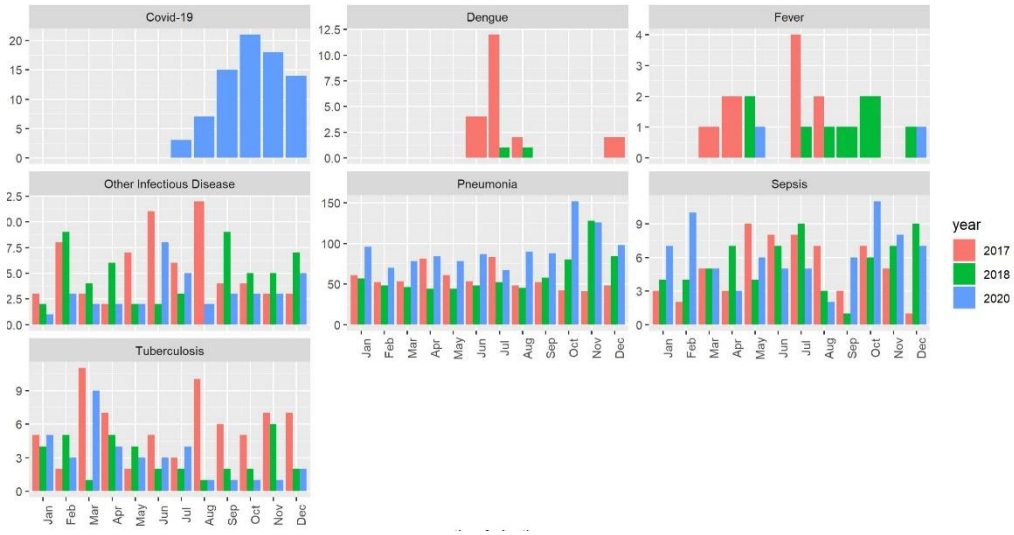

Fig. 12. Distribution of the infectious causes of death registered in Trivandrum Corporation in the years 2017, 2018 and 2020.

Age Distribution of the Deceased

Cause of Death	N	Mean	Std.Dev	Histogram	Median	IQR	Boxplot
NCDs	12868	61.18	16.16		63.0	53 - 72	
Heart Disease/Heart Attack	11449	65.50	14.53		66.0	57 - 75	
Miscellaneous	11040	52.19	24.67		58.0	38 - 70	
Infections	3067	65.82	15.70		68.0	58 - 76	
Senility	897	81.20	9.36		82.0	76 - 88	
Road Traffice Accidents	308	45.01	17.72		46.5	30 - 59	
MCH issues	36	35.28	30.34		35.0	4 - 62	
Suicide	18	46.22	16.03		47.0	38 - 56	

Number of Deaths

Cause of Death	2017	2018	2020
Heart Disease/Heart Attack	3712	3693	4033
Others	2829	2233	1651
Pneumonia	668	727	1101
Neoplasm	1236	1179	843
Chronic Respiratory Diseases	1038	2087	712
Diabetes	371	473	416
Chronic Kindey Diseases	551	467	398
Senility	237	389	269
Chronic Liver Diseases	517	403	242
Stroke	640	515	212

Fig. 13. Some summary statistics of the cause of death segregated by age and number of deaths (the color scheme represents the ranking, blue representing the top cause of death that year).

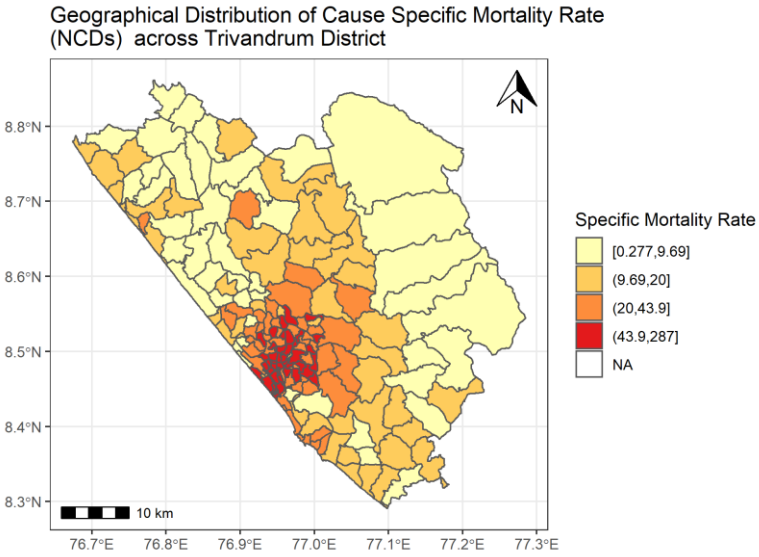

Fig. 14. Geographical distribution of Cause Specific Mortality Rate of non-communicable diseases across Trivandrum District. One can appreciate the spatial relationship with higher rates (in red color) around the central part of the district which is home to the urban population.

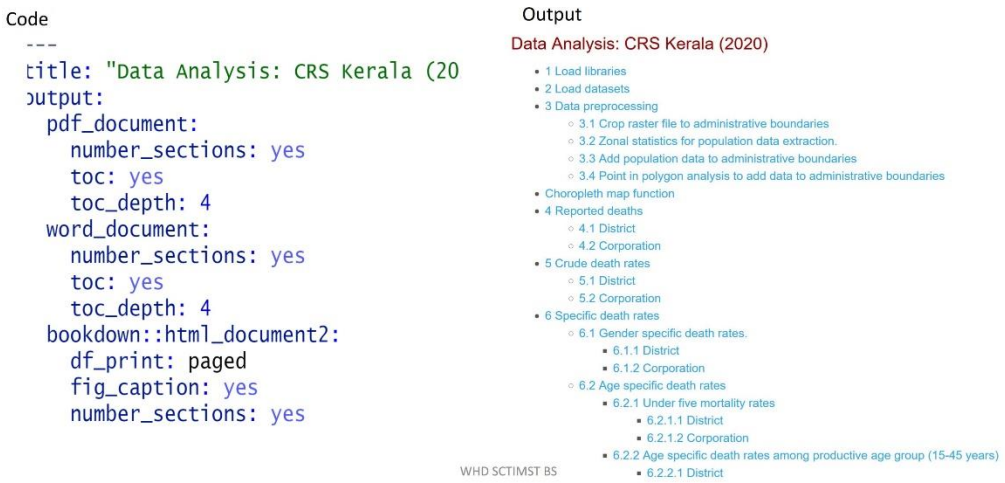

Fig. 15. Reproducible and scalable algorithms to generate automated reports for improving efficiency and promoting the using cause of death data from the CRS in Trivandrum Corporation in the years 2017, 2018 and 2020.

Some of possible benefits of using the data science approach are as follows:

- Encourage Data-use policy
- Improve data collection, processing, validation, efficiency of the system
- Integrate Data from different sources
- Timely and Actionable evidence for public health interventions
- Allows for spatial and temporal analysis (use of Geocoded information)
- Transformation of text data from PDF into analysis ready formats
- Automated Report Generation

3.3 Case Study 3 – Increasing Trends of Caesarean Sections in India

Globally, it is estimated that one in five children are born by caesarean deliveries and this is going to increase to three in five by the year 2030. The continued rise in caesarean sections has been a cause of concern in low- and middle-income countries (LMICs) like India for many reasons, including post-partum bleeding, infection, complications or even death to the mother and the child. Previous studies show the growing disparities in access to quality maternal healthcare as well as inequitable distribution of these services across different geographical regions and sociocultural contexts contribute to caesarean sections. However, evidence on these complex relationships is still emerging and need for an in-depth analysis in the Indian context is critical. Geospatial techniques using data from large scale national surveys like the National Family Health Survey can unearth crucial evidence into the patterns of medically unnecessary and potentially harmful caesarean sections in India. This much needed exploration has the potential to inform public health policy and provides opportunities for course correction in maternal and child health service delivery in India.

The objectives of this case study were to study the patterns of caesarean section at the state and district level in India and investigate spatial clustering of caesarean sections across districts of India.

The data extraction methodology used reproducible algorithms that was used in similar situations with minimal modification and customization (see Fig. 16 and Fig 17).

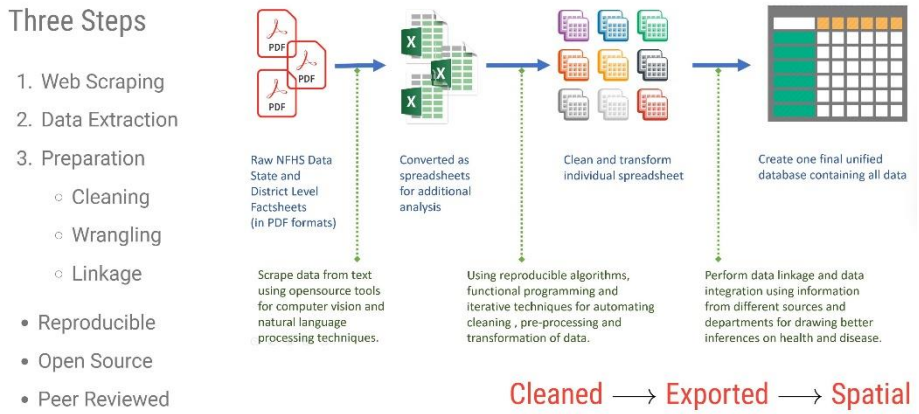

Fig. 16. The Data Extraction and Preparation Schema.

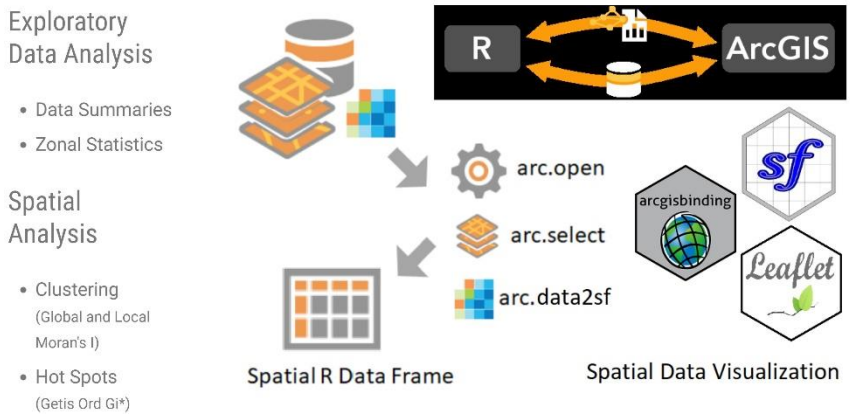

Fig. 17. Summary of the data analysis and visualization methods and tools used for the study.

Results:

There has been as steady increase in caesarean sections in India over the last 15 years (see Fig. 18). The proportion of caesarean sections in India rose from 17.2% in NFHS-4 (2015-16) to 21.5% in NFHS-5 (2019-21). As a result, 21 states and union territories had C-section births proportion greater than the national rural average (>17.6%) while 17 states and union territories were greater than the national urban average (32.3%).

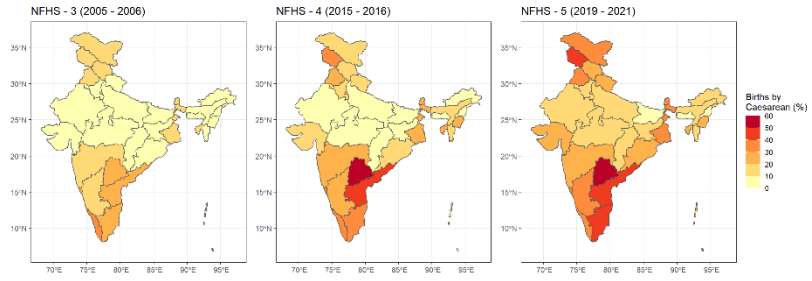

Fig. 18. Spatial Distribution of caesarean births across the three rounds of National Family Health Surveys.

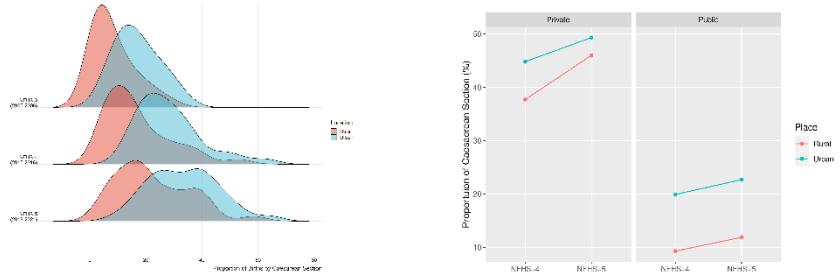

Fig. 19. Distribution and increase of the proportion of caesarean births across different types health facilities in rural and urban India.

The proportion of caesarean births were higher among private hospitals as compared to public hospitals in NFHS-5 (see Fig. 19). Additional investigation for the evidence of clustering in districts with high caesarean sections revealed highly significant clustering the global level with the Moran's I, p-value < 0.01. Similarly, local levels (Getis-Ord General G, p-value < 0.01) in the public and private hospitals (Table. 4).

Table 4. Summary of the clustering analysis districts with high rates of caesarean sections by type of facility and NFHS round.

Facility Type	NFHS Round	G Statistic	Z Score	p-value
Overall	NFHS-4	0.01099	18.67214	<0.0001
Overall	NFHS-5	0.01081	19.13736	<0.0001
Public	NFHS-4	0.0081	8.28497	<0.0001
Public	NFHS-5	0.00808	9.00839	<0.0001

Private	NFHS-4	0.01082	19.18451	<0.0001
Private	NFHS-5	0.00749	17.28355	<0.0001

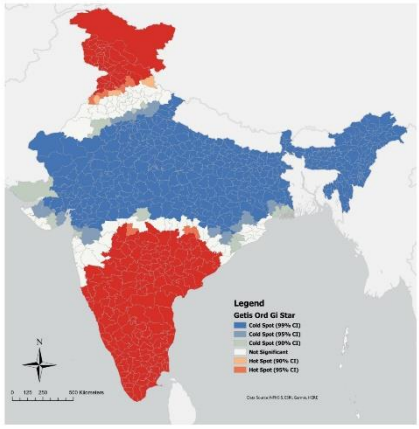

Fig. 20. Optimized Hotspot analysis of caesarean sections (%) in India.

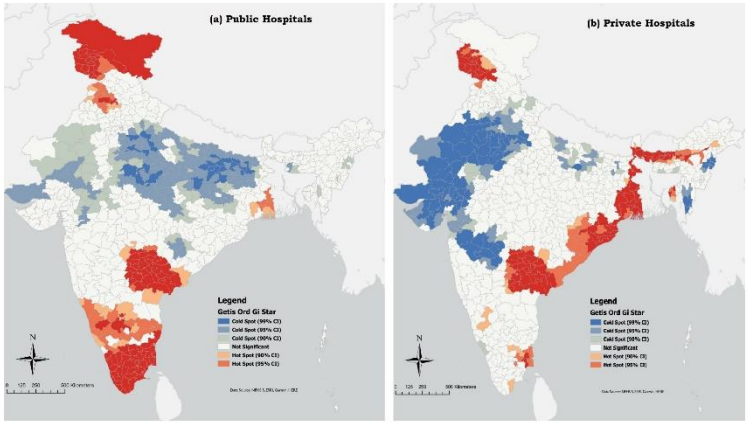

Fig. 21. Hotspot analysis of clusters of districts with high and low caesarean sections (%) by type of health facility (public vs private).

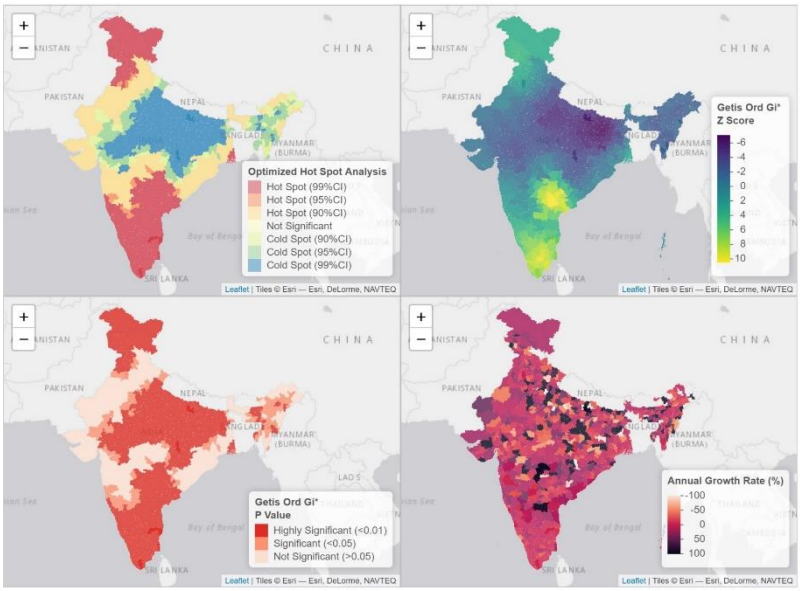

Fig. 22. Local Indicators of Spatial Association (LISA) for the cluster analysis of districts with high caesarean rates.

The urban-rural differential in quality maternal care may be widening with unnecessary and medically unwarranted caesarean sections being performed in rural hospitals. Much of these caesarean sections are being brought about by the recent outcrop of private medical facilities in rural India (see Fig. 20 – Fig. 22).

Geospatial analysis can reveal crucial insights into where the disproportionate increase in caesarean sections is being undertaken, which may aid in designing tailored public health interventions and come up with policy decisions necessary for course corrective steps. One limitation is that we did not take into account factors such as total fertility rate, health insurance, the possibility of repeat caesarean sections, etc. These factors need to be kept in mind in future research works.

To conclude, the case study demonstrates the utility of spatial data science methods to using large survey data to reveal inequities in maternal health care with potential public health policy implications.

3.4 Case Study 4 – Retinal Disease Classification

A number of years have passed since telemedicine and tele-ophthalmology were introduced, however they have become more important in recent years as a result of the COVID-19 pandemic. A recent survey (n = 1180) done by All India Ophthalmological Society reveals that currently 17.5% of the ophthalmologists use tele-ophthalmology in their

practice. 98.6% showed increased interest to incorporate tele-ophthalmology in their practice and 98.8% of the practicing ophthalmologists view mobile based applications as having huge potential in tele-ophthalmology solutions.

Technological advances and the ability to directly visualize and image the eye has made ophthalmology an ideal specialty for the use of telemedicine. The application of tele-ophthalmology is not a new method of care or technique, however, the implementation of artificial intelligence (AI) is set to lead to a great transformation in its use. Deep convolutional neural networks have shown promise in predicting and classifying retinal diseases and has immense public health implications, especially in low-resource settings like India.

The objectives of this case study were to classify diseases of the retina (CNV, DME, Drusen, and normal) based on optical coherence tomography (OCT) images using deep neural networks and to build a full connected network/convolution neural network and compare the results with existing model architectures and to deploy the final model to production as a mobile application.

The dataset contains 84,495 OCT images in *.JPEG format which were taken from various cohorts of adult patients from 5 hospitals between July 1, 2013 and March 1, 2017.(36) There were four categories of retinal diseases such as:

- Choroidal New Vascularization (CNV)
- Diabetic Macular Edema (DME)
- Drusen
- Normal

The computational resources used were as follows:

- RStudio and Google Colab were used for building and training the model
- CPU: Intel Xeon CPU@ 2.3 GHz
- GPU: Nvidia Tesla P100 (16 GB)
- PyTorch library in Python and torch package in R

The methods for the case study include dividing the images into test, training and validation datasets randomly in a proportion of 70%, 20%, and 10% respectively. Data was explored for class imbalance and discrepancies were rectified using up-sampling method. The images were then classified according to the four classes of retinal diseases using a custom deep convolutional neural network. Subsequently, a transfer learning approach was adopted to train two models, ResNet-18 and MobileNetV2. The final model was selected using standard metrics such as accuracy, precision and F1Score.(37)

We also implemented model interpretability using Captum and Gradcam visualization for explainablility of the artificial intelligence algorithm (see Fig. 23). The final

model was developed into an Android mobile application using the torchscript library. Validation of model predictions was done using new set of OCT images.

The MobileNetV2 model performed better than ResNet-18 by achieving a training accuracy of about 97% and a test accuracy of 95%. The ResNet-18 model performed better in the precision and recall with an average F1 score of 0.94 vs 0.92. The average training time and inference time for both the models were around 1 hour 30 mins and >0.02 seconds, respectively (see Table. 5).

Table 5. Comparison of model parameters, hyper-parameters, model metrics of the three deep learning models.

Details	ConvNet	ResNet-18	MobileNetV2
Convolution Layers	3	20	52
Parameters	728,184	11,178,564	3,504,872
Flops (GFLOPS)	0.05	1.83	0.31
Optimizer	Adam	Adam	Adam
Hyperparameters			
Epochs	25	30	25
Learning Rate	0.001	0.001	0.001
Computation Time			
Training	~47 mins	~1hr 40 mins	~1hr 27 mins
Inference	~0.09 secs	~0.003 secs	~0.02 secs
Accuracy (%)			
Training	98.0	96.4	97.1
Validation	65.9	94.6	94.1
Test	43.3	94.1	95.5

The accuracy of a typical ophthalmologist in classifying an OCT image is ~85% and our best model (MobileNetV2) was able to perform with an accuracy of ~96%. A typical ophthalmologist requires ~2 secs to classify one image while our model required only ~0.03 secs for the same. Upon increasing the training data, the accuracy of the model increased (~94% with 8,000 images vs. ~97% with 80,000 images).

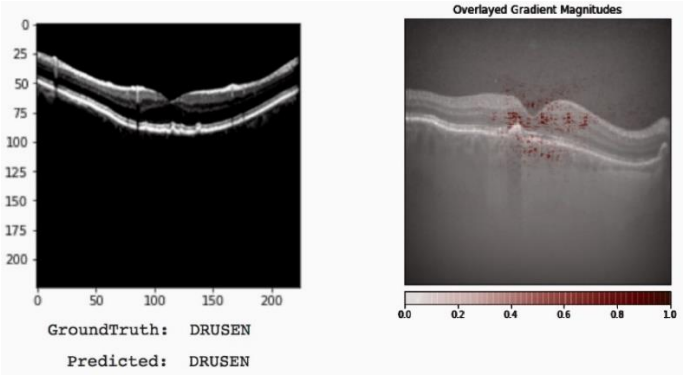

Fig. 23. Incorporating model interpretability.

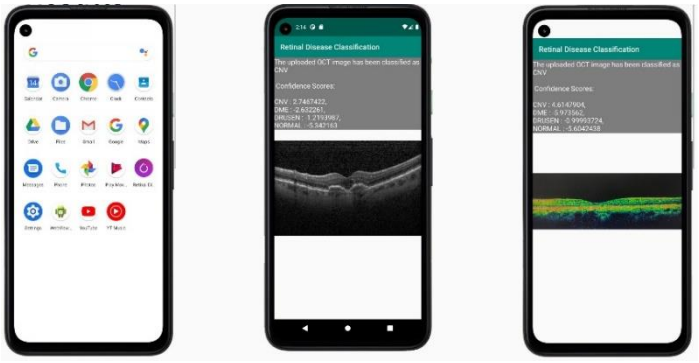

Fig. 24. Screenshots of the deployed mobile application.

Through this case study we demonstrate the application of deep neural networks in tele-ophthalmology using open-source tools and publicly available data. Leveraging explainable artificial intelligence approach provides an edge over many existing 'black-box' neural networks as it aids the clinician in decision making. Mobile based deep CNN models have huge potential to address many issues of public health importance. With adequate training, peripheral health workers can be utilized for delivering tele-ophthalmology services in hard-to-reach rural areas (see Fig. 24).

4 Discussion

We would like to reflect on the challenges faced in each of the case study. The first case study (COVID-19 Dashboard) had brought challenges of data preprocessing and linkage where data from multiple sources had to be integrated with a spatial attribute. This challenge was overcome by developing robust reproducible algorithms that could be reused and customized for research needs. For example, algorithms written for comparing between first two waves of COVID-19 have been repurposed with minimal effort and time during the onset of the third wave. Another challenge we faced was the lack

of data availability. As the covid19india.org tracker website had to suspend operations as it was driven by volunteers with no funding. Fortunately, efforts by activists and data journalists helped revive the ongoing effort of citizen involvement through the new covid19india.org website which filled the shoes of its predecessor.

The second case study on using CRS data to gain public health insights into cause of death and mortality patterns brought new challenges as data was not in analysable formats, it required a lot of data prepration and geocoding of addresses. Other challenge was the organizational barriers in providing access to data to public health stakeholders including researchers. One way some of these challenges were address was through stakeholder engagement and earning trust.

The third case study on caesarean section deliveries had unique challenges such as broken links to publicly available district level factsheets. The second challenge was the complexities involved in computational extraction of data in tidy formats from the PDF documents. Also, the cleaning of data required a lot manual checks to prevent misinterpretation of information. Additionally, one challenge that was particularly difficult to solve was geocoding the data as the district names are not uniformly written across different datasets including spatial and sociodemographic.

The fourth case study required considerable expertise and knowledge on the fundamentals of deep learning techniques and image processing and analysis. The case study also necessitated additional computational resources that may not be readily available to all researchers. Addressing the issue of clinical interpretably, explainability and visualization of where the weights are being picked up through Captum and Gradcam was difficult to implement. The most difficult challenge for the author was to develop an app that can be deployed on an Android Smart Phone especially because of lack of prior experience or lack of formal training in app development.

Additional challenges faced in all the case studies included the issue of data quality. Considerable time was spent on all the casestudies to address this challenge adequately. Some of the data quality issues faced were missing data, outliers, erroneous entries, and entering data in inconsistent formats. Some of these challenges and the approaches takent to address these were discussed in some of the previous work by the authors.(38–41) Apart from this, the lack of adequate computational infrastructure for performing the analysis was persistent across most casestudies.

Despite these challenges, the authors were successful in demonstrating that public health data science has immense potential in transforming data from the routine health information systems into evidence for public health. Data science approach to public health has varied applications including health equity, pandemic or outbreak management, medical imaging, health policy and maternal and child health. Empowered with data science tools and FAIR Principles for data management (Find-able, Accessible, Interoperable and Reusable),(42) the way forward for public health can be more equitable, acceptable, patient centric, community oriented, participatory, and trusted.

5 Conclusion

To conclude, this study has many value additions:

- Generates value to data generated from the routine large-scale national-level surveys.
- Adds to the quality of evidence generated from routine health information systems by integrating it with data from multiple sources. This data linkage and integration allows for a comprehensive view of the situation, which is not otherwise possible.
- Captures the ubiquitousness of data triangulation and geographical information systems (GIS), spatial and spatiotemporal methods. This adds a new spatial dimension to the data, greatly enhancing the information and its interpretation by the decision-makers.
- Leverages the applications of novel methods like data science and spatial techniques by providing robust and reproducible frameworks for evidence generation. This becomes especially important when the data volume is vast and resources are low.
- Provides a framework to engage with the program managers/stakeholders in near-real-time to gain insights into patterns and dynamics of issues of public health concern.
- This exploration also can enhance data quality and encourage data use policy in public health.
- It will help the routine public health surveillance system through watchful observation and undertaking corrective action.
- It will enable researchers to study the epidemiological trends in individual factors and geospatial and region-specific disease patterns.
- Potential to help in better resource allocation, resource reallocation, and mobilization.
- It can aid in decision support and inform policy decisions by providing robust epidemiological evidence, which is crucial for optimal use of limited resources and improving the overall health system efficiency.
- It can help us understand the emerging public health threats and create an enabling environment for precision public health.

Acknowledgement

I gratefully acknowledge the financial subsistence provided by the Science for Equity, Empowerment and Development (SEED) Division, Department of Science and Technology, Govt. of India through the project entitled ‘Extending benefits of biomedical science and technology of SC and ST communities through all level participatory engagement - ST Components’ (F. No. SEED/TITE/2020/134).

References

1. A O, S B, W R, N AM, A S. Public health digitalization in Europe. European journal of public health. 2019 Oct;29:28–35.

2. Chiolero A, Buckeridge D. Glossary for public health surveillance in the age of data science. J Epidemiol Community Health. 2020 Jun;74(7):612–6.

3. Ford E, Boyd A, Bowles JKF, Havard A, Aldridge RW, Curcin V, et al. Our data, our society, our health: A vision for inclusive and transparent health data science in the United Kingdom and beyond. Learn Health Syst. 2019 Jul;3(3):e10191.

4. Benke K, Benke G. Artificial Intelligence and Big Data in Public Health. Int J Environ Res Public Health. 2018 Dec 10;15(12):E2796.

5. Belle A, Thiagarajan R, Soroushmehr SMR, Navidi F, Beard DA, Najarian K. Big Data Analytics in Healthcare. BioMed Research International. 2015;2015:1–16.

6. Sahay S, Sundararaman T, Braa J. Public health informatics: designing for change-a developing country perspective. Oxford University Press; 2017.

7. Haneef R, Delnord M, Vernay M, Bauchet E, Gaidelyte R, Van Oyen H, et al. Innovative use of data sources: a cross-sectional study of data linkage and artificial intelligence practices across European countries. Arch Public Health. 2020;78:55.

8. Deeny SR, Steventon A. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. BMJ Qual Saf. 2015 Aug;24(8):505–15.

9. Vayena E, Dzenowagis J, Brownstein JS, Sheikh A. Policy implications of big data in the health sector. Bull World Health Organ. 2018 Jan 1;96(1):66–8.

10.Zodpey SP, Negandhi HN. Improving the quality and use of routine health data for decision-making. Indian Journal of Public Health. 2016 Jan 1;60(1):1.

11.Pandey A, Roy N, Bhawsar R, Mishra RM. Health information system in India: issues of data availability and quality. Demography India. 2010;39(1):111–28.

12.Hung YW, Hoxha K, Irwin BR, Law MR, Grépin KA. Using routine health information data for research in low- and middle-income countries: a systematic review. BMC Health Services Research. 2020 Aug 25;20(1):790.

13.Bomba B, Cooper J, Miller M. Working towards a national health information system in Australia. Medinfo MEDINFO. 1995;8:1633–1633.

14.Morrato EH, Elias M, Gericke CA. Using population-based routine data for evidence-based health policy decisions: lessons from three examples of setting and evaluating national health policy in Australia, the UK and the USA. Journal of Public Health. 2007;29(4):463–71.

15.Houston TK, Sands DZ, Jenckes MW, Ford DE. Experiences of patients who were early adopters of electronic communication with their physician: satisfaction, benefits, and concerns. Am J Manag Care. 2004;10(9):601–8.

16.Tull K. Designing and implementing health management information systems. 2018;

17.Trewin C, Strand BH, Grøholt EK. Norhealth: norwegian health information system. Scandinavian journal of public health. 2008;36(7):685–9.

18.Ringard Å, Sagan A, Sperre Saunes I, Lindahl AK, Organization WH, others. Norway: health system review. 2013;

19.Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11

[Internet]. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC); 2018. Available from: https://doi.org/10.7927/H49C6VHW

20.Government of India. Unique Identification Authority of India [Internet]. 2021 [cited 2021 May 19]. Available from: https://uidai.gov.in/images/state-wise-aadhaar-saturation.pdf

21.R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2021. Available from: https://www.R-project.org/

22.RStudio Team. RStudio: Integrated development environment for r [Internet]. Boston, MA; 2021. Available from: http://www.rstudio.com/

23.Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686.

24.Wickham H. Advanced R, Second Edition. Advanced R. :604.

25.Wickham H. Mastering Shiny. O'Reilly Media, Inc.; 2021. 395 p.

26.Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nature human behaviour. 2017;1(1):1–9.

27.Peng RD, Hicks SC. Reproducible Research: A Retrospective. Annu Rev Public Health. 2021 Apr 1;42:79–93.

28.The Comprehensive R Archive Network [Internet]. [cited 2022 Sep 30]. Available from: https://cran.r-project.org/

29.RECon. R Epidemics Consortium [Internet]. 2021 [cited 2021 May 25]. Available from: http://reconhub.github.io/

30.Mitra A, Soman B, Gaitonde R, Singh G, Roy A. Data Science Methods to Develop Decision Support Systems for Real-time Monitoring of COVID-19 Outbreak. Journal of Human, Earth, and Future. 2022 Jun 1;3(2):223–36.

31.Mitra A, Pakhare AP, Roy A, Joshi A. Impact of COVID-19 epidemic curtailment strategies in selected Indian states: An analysis by reproduction number and doubling time with incidence modelling. PLOS ONE. 2020 Sep 16;15(9):e0239026.

32.Mitra A, Soman B, Singh G. An interactive dashboard for real-time analytics and monitoring of COVID-19 outbreak in india: A proof of concept. In: arXiv preprint arXiv:210809937 [Internet]. Norway: International Federation for Information Processing; 2021. Available from: https://arxiv.org/ftp/arxiv/papers/2108/2108.09937.pdf

33.United Nations. Principles and recommendations for a vital statistics system: Revision 3 [Internet]. UN; 2014 [cited 2022 Sep 30]. (Statistical Papers (Ser. M)). Available from: https://www.un-ilibrary.org/content/books/9789210561402

34.MCCD Division. Report on Medical Certification of Cause of Death - 2018. Thiruvananthapuram: Dept of Economics & Statistics; 2018.

35.Vital Statistics Division. Annual Vital Statistics Report - 2019. Thiruvananthapuram: Dept of Economics & Statistics; 2019.

36.Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018 Feb 22;172(5):1122-1131.e9.

37.Mitra A. Retinal Disease Classification Using Deep Learning Networks : applications in teleophthalmology. In: Proceedings of the 17th International Conference of Telemedicine Socitey of India. India: TSI, India; 2021.

38.Mitra A, Soman B, Gaitonde R, Singh G, Roy A. Tracking and monitoring COVID-19 in Kerala: Development of an interactive dashboard. In: Health System Research. Cochin, Kerala; 2021.

39.Singh G, Soman B, Mitra A. A Systematic Approach to Cleaning Routine Health Surveillance Datasets: An Illustration Using National Vector Borne Disease Control Programme Data of Punjab, India. arXiv:210809963 [cs] [Internet]. 2021 Aug 23 [cited 2021 Oct 2]; Available from: http://arxiv.org/abs/2108.09963

40.Joshi A, Mitra A, Anjum N, Shrivastava N, Khadanga S, Pakhare A, et al. Patterns of Glycemic Variability During a Diabetes Self-Management Educational Program. Medical Sciences. 2019 Mar;7(3):52.

41.Saoji A, Nayse J, Deoke A, Mitra A. Maternal risk factors of caesarean delivery in a tertiary care hospital in Central India: A case control study. People's Journal of Scientific Research. 2016;9(2):18–23.

42.Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3(1):160018.