USAID
FROM THE AMERICAN PEOPLE

WADHWANI AI

COVID Modelling & Decision Support Engagement

A Flexible Data-Driven Framework for COVID-19 Case Forecasting Deployed in a Developing-world Public Health Setting

COVID Modelling & Decision Support Engagement

A Flexible Data-Driven Framework for COVID-19 Case Forecasting Deployed in a Developing-world Public Health Setting

TABLE OF CONTENTS

Why Did We Make This Guide?

The coronavirus disease (COVID-19) has resulted in a worldwide pandemic. One of the worst aspects of the devastation caused by the pandemic is the severe toll it has extracted on public health systems. This is especially true in the developing world, where resources are few, information flow pipelines are weak, medical care is spotty and unevenly spread, and the ability to contain infection spread is limited by dense populations and poor compliance.

As an Institute committed to serving the underserved through the development and use of AI technologies, we engaged with Mumbai city government and the state government of Jharkhand to aid in pandemic response through a combination of predictive modelling and data analytics.

The primary objectives were as follows:

a. to provide accurate estimates of future case loads via epidemiological models that are simple enough to accommodate the type of data that was collected on the ground, yet complex enough to reasonably capture the infection dynamics

b. to develop and implement analytics tools to guide testing and interventions in areas where they would have the most impact.

This guide summarises the work we did from a technical and public health perspective, with pointers to resources that may prove beneficial to various stakeholders.

Who Should Read This Guide?

This practical guide has been designed for data science practitioners and epidemiologists who may utilise our methodology and codebase in their respective research areas.

Moreover, public health professionals and government officials may glean from the data pipelines, modelling framework and analytics capabilities to forecast disease spread in communities.

Glossary

Age Stratification

Age stratification is the process of splitting a population by their age group, anticipating that each group will be analysed or modelled separately.

Agent-Based Models

Agent-based epidemiological models model every potential disease carrier as a separate individual or agent, along with the interactions between individuals and possibly their mobility patterns. By contrast, compartmental models do not operate at an individual level but rather at a population level.

Basic Reproduction Number or R_0

In epidemiology, in a given population of susceptible people, the basic reproduction number R_0 is the number of people in the population who will be directly infected by one infected person. It is commonly understood that if $R_0 > 1$, the infection will spread throughout the population, while if $R_0 < 1$, the infection will eventually die out. The specific value of R_0 for a certain disease within a certain population plays a crucial role in determining the proportion of people who need to be immunised in order to eradicate the disease.

Bayesian Model Averaging

Bayesian Model Averaging (BMA) involves making parameter inferences on the basis of all models within a class of models. The parameter(s) to be estimated is computed as a weighted average over all models in the class using the posterior distribution of the parameter. This approach typically results in robust estimates and explicitly takes into account model uncertainty.

Compartmental Models

Compartmental epidemiological models are models for the spread of a pandemic in which each compartment corresponds to a group of people or a population in a specific disease state. Individuals move between compartments as they transition from one disease state to another.

Confidence Intervals

Confidence intervals arise in situations where a mathematical variable is described by a probability distribution. We refer to the value of the variable as lying within an interval of values with a certain probability, or “confidence”.

Convex Function

A mathematical function is termed convex if the line segment between any two points on the graph of a function does not lie below the graph between the two points. It is strictly convex if the line segment always lies above the graph. Functions that are strictly convex over a domain are important because they have a single minimum value in that domain.

Data Smoothing

Data smoothing is a mathematical or computational procedure to smoothen out kinks in the data. This procedure is typically used in the context of time series data.

Ensembles

An ensemble of models is a set of models of the same type that carry out the same task, possibly with each model having a different set of parameter values. An ensemble of models taken together is often more accurate than any one model. For example, in the context of epidemiological models, an ensemble could be a family of compartmental models such that each model in the family has a different set of parameters. Forecasts from the different models in the ensemble are then combined in various ways to form a consensus forecast of the ensemble. A common way of combining these forecasts is to do a weighted average of the forecasts from each model in the ensemble. This is referred to as a *weighted ensemble.*

Epidemiology

Epidemiology is the study of the causes and distribution of health states across individuals in a population. It also involves the application of this study to the control of health problems. An *epidemiological model,* as used in this playbook, is a mathematical model that describes dynamical transitions between various health states in a population.

Fitting Durations

These are time periods, specified as a start and end date, over which epidemiological model parameters are fit to data.

Glossary

IHME CurveFit Model

This is a model developed at the Institute for Health Metrics and Evaluation. The model consists of a specific parameterized function that is fit to COVID-19 case count data.

Incubation Period

The incubation period for an infection is the time elapsed between being infected by a disease and showing symptoms of the disease. In the context of epidemiological modelling, it is often assumed to be the time from infection to hospitalisation or isolation.

Identifiability

A parameter of an epidemiological model is said to be identifiable if its value can be inferred from data.

Interpretability

A parameter of an epidemiological model is said to be interpretable if its values can be assigned meaning in terms of epidemic spread.

Loss Function

A loss function is a mathematical function that measures the error between the estimated value of a variable and its true value.

Markov Chain Monte–Carlo (MCMC)

MCMC is a computational method for sampling from a distribution without knowing all of the distribution's mathematical properties. It is implemented by sequentially sampling values from proposal distributions that are easy to sample from, and then accepting or rejecting these samples with a probability that depends on the original distribution. The sequential samples thus obtained ultimately converge to samples from the original distribution. The name MCMC combines two properties: Monte–Carlo and Markov chain. The set of sequential random variables form a Markov chain; Monte–Carlo is the practice of estimating the properties of a distribution by examining random samples from the distribution.

Mean Absolute Percent Error (MAPE)

MAPE is defined as the absolute difference between true and estimated value of a variable, expressed as a percentage of the true value.

Model-Agnostic Evaluation

This is a method or class of methods for evaluating the predictions of a model in a manner that does not depend on the details of the specific model but only on the specific prediction that is being evaluated.

Quantiles

A quantile, as referring to a data distribution, refers to a cutoff value that represents a range of values in a data distribution that lie below the cutoff value. The cutoff value is specified based on the probability that the data lies below that value. For example, the 0.5-quantile is a value such that half of all the data lies below it; thus, the 0.5-quantile is the median value of the data.

ReichLab Forecasting Hub

The ReichLab Forecasting Hub is a COVID-19 forecast resource that serves as a central repository of COVID-19 case forecasts and predictions from over 50 international research groups. It was founded in 2020 by the lab of Nicholas Reich at the University of Massachusetts, Amherst.

SEIR Model

The SEIR model is a compartmental epidemiological model describing transitions between disease states or compartments corresponding to susceptible (S), exposed (E—infected but not infectious), infectious (I) and removed (R) population cohorts. Individuals move between these compartments in sequence as they become exposed, infected and infectious during disease progression until recovery.

Sequential Model-Based Optimization (SMBO)

SMBO is a global hyperparameter optimization method. It iterates between fitting and interpolation model to an already sampled set of hyperparameters, and using the model to decide on the next set of hyperparameters to investigate. It is one of a class of black box algorithms that are used when objective functions are difficult to evaluate.

01

Background

TOPICS COVERED IN THIS SECTION

- Introduction to Epidemiological Forecasting Models
- Our Model
- Areas of Epidemiological Modelling
- Salient Contributions

Introduction to Epidemiological Forecasting Models

Coronavirus disease (COVID-19) was declared a public health emergency of international concern in January 2020 by the World Health Organization (WHO). Globally, as of December 2021, there have been over 250 million confirmed cases of COVID-19, including over 5 million deaths, reported to the WHO. The nature of the crisis is unprecedented. The long-term effects on human capital, productivity and behaviour may be long-lasting in addition to the repeated health and economic shocks. These effects are even more pronounced in the developing world.

AN UNPRECEDENTED CRISIS

> 250 million
confirmed cases of COVID-19

> 5 million
confirmed deaths due to COVID-19

Globally, as of December 2021, there have been over 250 million confirmed cases of COVID-19, including over 5 million deaths, reported to the WHO.

During the course of the pandemic, governments across the world have worked assiduously to limit the human cost and economic disruption as well as adjust preventive policies to mitigate the disease spread. It is the most consequential set of public policy and mass behaviour change actions most of us have seen in our lifetimes. Robust policy development and planning necessitates pandemic models that can accurately predict the number of COVID-19 cases and deaths sufficiently far into the future, as such models would allow governmental policy makers to examine the effects of different preventive policies.

The ongoing COVID-19 pandemic has spurred intense interest in epidemiological forecasting models. The need for robust pandemic response and planning has been especially pressing in dense populations across the developing world, with their limited health resource availability, limited data to anticipate outbreaks, and long lead times for addressing shortfalls. It is paramount to ensure adequate capacity availability of critical health care resources to reduce mortality. There is a need for forecasting reported infections at a local level to inform capacity planning, model the effects of policy changes and prepare for potential scenarios. We deployed a forecasting framework that was used in Mumbai, India, one of the most densely populated cities in the world, as well as in other resource-constrained regions such as the state of Jharkhand, India, during the first Covid infection wave. The partners for data and usage of the solution include the Brihanmumbai Municipal Corporation (BMC) and the Integrated Disease Surveillance Programme, respectively, at the two locations.

THE NEED FOR FORECASTING MODELS

The ongoing COVID-19 pandemic has spurred an intense interest in epidemiological forecasting models. The need for robust pandemic response and planning has been especially pressing in dense populations across the developing world, with their limited health resource availability, limited data to anticipate outbreaks, and long lead times for addressing shortfalls. It is paramount to ensure the availability of an adequate capacity of critical health care resources to reduce mortality.

Our Model

Forecasting infection case counts and estimating accurate epidemiological parameters are critical components of managing the response to a pandemic. We have devised a flexible modelling framework and demonstrated its value for epidemic forecasting taking into consideration the kind of case count aggregate data that is typically available in a constrained public health setting. The deployed system was used to drive decision making and planning with good accuracy (worst case Mean Absolute Percent Error < 20%) during the COVID-19 pandemic in Mumbai and Jharkhand, India. Our framework allows rapid forecasting with uncertainty estimates and is extensible to other model families and to different types of loss or error functions. Furthermore, it enables the optimisation of hyperparameters such as fitting durations and ensemble weights. We motivated the choice of the specific compartmental model used through identifiability of the underlying parameters in the light of the data constraints.

Areas of Epidemiological Modelling

An epidemiological modelling framework consists of four important components. They are as follows:

FORECASTING

The COVID-19 pandemic and the concerned global forecasting challenges have prompted new research on modelling infectious disease spread. There are three broad classes of models.

Compartmental models assume that individuals in a population at any given time are assigned to one of several states known as compartments. As the disease progresses, individuals transition between these compartments. Over the last year, variants of the SEIR compartmental model have been widely used to study the healthcare burden brought about by the pandemic. Many of these models also incorporate aspects such as age-stratification, asymptomatic transmission and effects of social distancing measures.

Agent-based models simulate interactions and disease stage transitions of individual agents or disease carriers.

Curve-fitting models fit parameterised curves to data. The examples include the exponential growth model and the IHME CurveFit model.

In recent times, there has been an increased emphasis on the practical aspects of model fitting like choice of training duration and identifiability issues. Model-agnostic evaluation of forecasts is another related area of interest. Whilst our deployment was primarily based on compartmental models, the techniques we used and developed are largely agnostic to model class and choice of loss function.

In recent times, there has been an increased emphasis on the practical aspects of model fitting like choice of training duration and identifiability issues.

PARAMETER ESTIMATION (WITH UNCERTAINTY)

Compartmental epidemiological models are highly interpretable in terms of their model parameters, which carry meaning independent of the specific forecasts. A case in point is the so-called R0 parameter that roughly measures the tendency of a disease to spread within a population. In addition to enabling high quality estimates of forecasts, it is thus incumbent upon a sound epidemiological modelling framework to provide robust estimates of the underlying parameters and the uncertainties in these estimates.

MODELLING WITH DATA LIMITATIONS

Epidemiological modelling in the developing world is beset by the dearth of data and quality issues. Multiple studies have concentrated on understanding transmission dynamics in such limited data settings. We too faced some of these challenges, and attempted to resolve them via appropriate model choices and data preprocessing.

PUBLIC HEALTH DEPLOYMENT

The practical use of epidemiological models in public health response demands a holistic view of government priorities, policy levers and processes. The initiatives in economic epidemiology are targeted towards supporting decision-making related to interventions and policy choices. Currently, several organisations share automated COVID case forecasts with relevant public health authorities. However, the forecasts are not always customised for decision-making. Our deployment involved a two-way partnership with the government thereby providing precise capacity planning guidance and insights into the pandemic dynamics per requirements.

Salient Contributions

Our framework employs aggregate case count data that is collected by government officials from health facilities. It outputs predicted case counts that are utilised by public health authorities for subsequent planning of personnel and supplies. Our work sheds light on the cardinal elements of a forecasting framework and contributes to the following areas.

SYSTEM AND PROCESS DESIGN

This refers to a practical, modular and extensible learning-based epidemic case forecasting system that is customisable to locales and application scenarios. The system consists of modules for data ingestion, preprocessing and exploratory analysis, model fitting, scenario-conditioned forecasting and application-specific report generation.

MODELLING METHODOLOGY

This refers to techniques for model and loss-agnostic estimation of parameters via sequential model based optimisation (SMBO). Combining SMBO sampling with Bayesian model averaging allows fast approximate quantification of forecast uncertainty. Through the research, we have demonstrated that this method is empirically comparable to a more rigorous Markov Chain Monte Carlo (MCMC) approach but computationally faster. Additionally, we have developed smoothing methods to handle data issues arising from delays in reporting.

EPIDEMIOLOGICAL MODEL CHOICE

We have presented arguments for constraints of interpretability warranting a simple variant of the SEIR compartmental model especially when only confirmed, active, recovered and deceased case counts are observed.

INTERPRETABILITY AND IDENTIFIABILITY

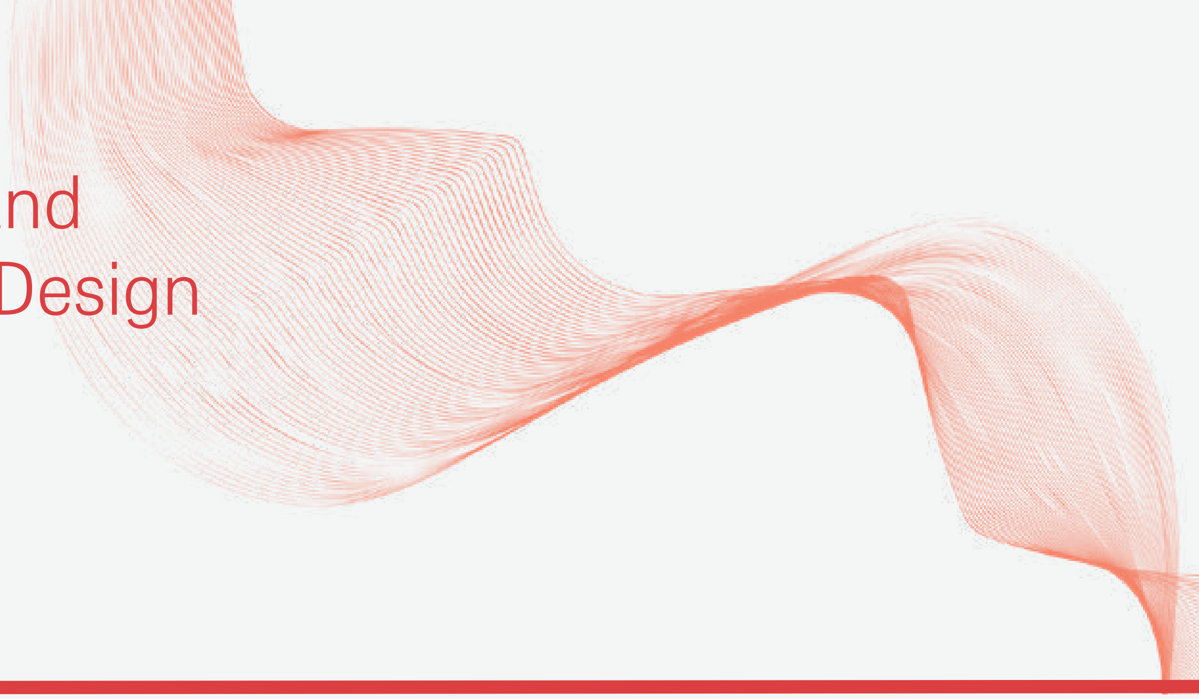
We developed practical notions of identifiability of epidemiological parameters, expressed in terms of the underlying parameter uncertainty, in the specific context of SEIR type models.

EMPIRICAL RESULTS

This refers to extensive empirical analyses detailing the optimisation of relevant hyperparameters, field predictive performance for the city of Mumbai as well as comparison with other state-of-the-art models hosted by ReichLab in the USA.

DEPLOYMENT LESSONS

We summarise the critical lessons from deployment of our modelling framework in Mumbai and Jharkhand. The audience for this work consists of applied researchers working on practical forecasting and public health officials.



02

System and Process Design

TOPICS COVERED IN THIS SECTION

- Data Operations
- Generic Modelling Framework
- Reporting Results

The end-to-end epidemic forecasting system developed consisted of the three major components outlined below. This system generically applies to any data-driven public health response to pandemics.

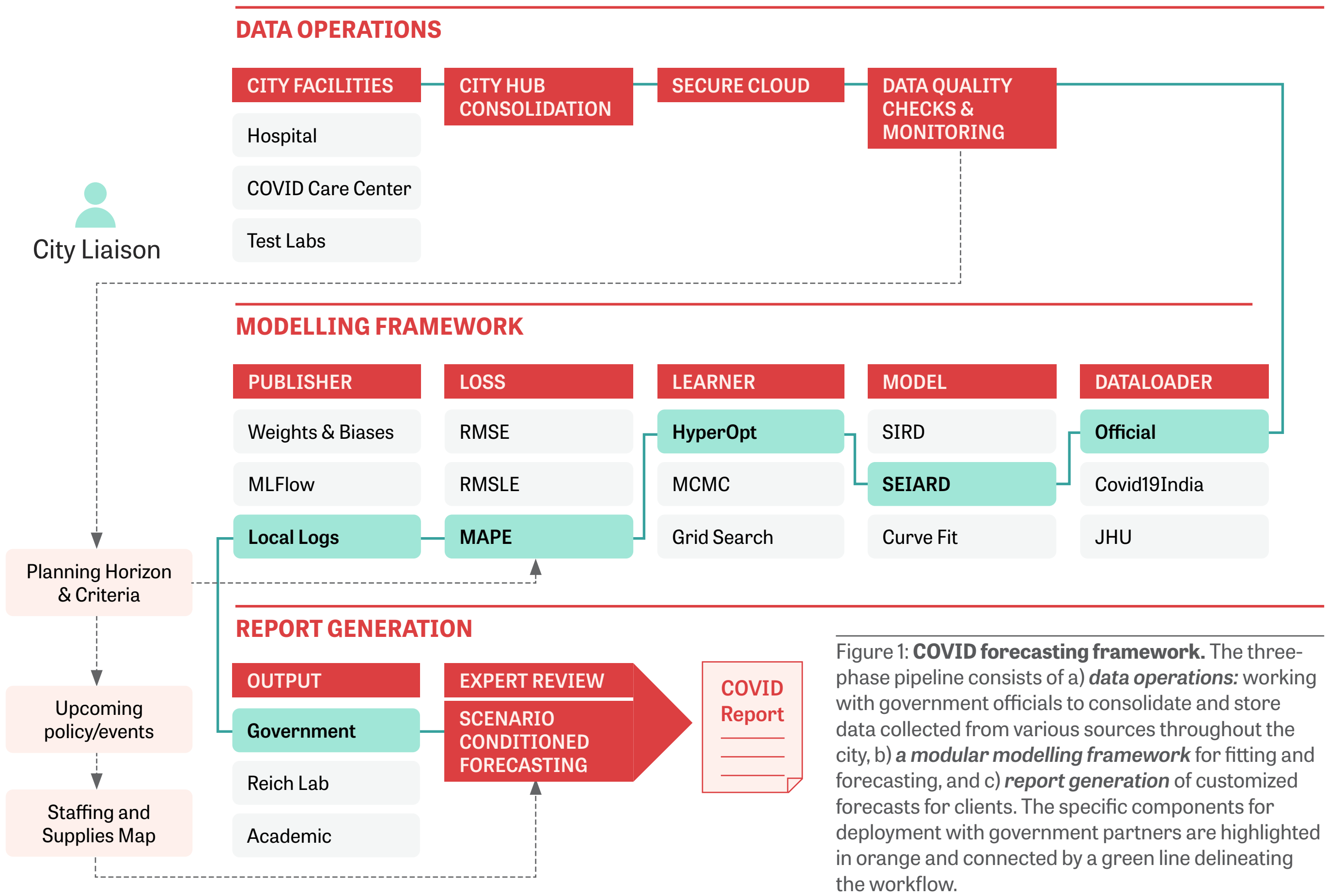


Figure 1: **COVID forecasting framework.** The three-phase pipeline consists of a) ***data operations:*** working with government officials to consolidate and store data collected from various sources throughout the city, b) ***a modular modelling framework*** for fitting and forecasting, and c) ***report generation*** of customized forecasts for clients. The specific components for deployment with government partners are highlighted in orange and connected by a green line delineating the workflow.

DESIGN COMPONENTS	NOTABLE ATTRIBUTES
Data Operations	• Standardised data schema for aggregate case counts • Data validation at point-of-ingestion, including de-duplication • Data anonymization, quality checks • Visualisations to manually monitor anomalies • SQL-compliant database exposed to the modelling pipeline
Generic Modelling Framework	• Extensible, scalable, suitable for rapid experimentation in multivariate time series forecasting • Five principal modules: Data loader, Model, Learner, Loss function, Publisher
Reporting Results	• Types of model outputs: time-series forecast format, planning reports, model pushes for external publication • Three projected scenarios reported, reflecting parameter uncertainties: low case scenario, high case scenario, planning scenario • Estimation of future facility-level demand for medical staff and equipment

03

Modelling Methodology

TOPICS COVERED IN THIS SECTION

- Parameter Fitting with Sequential Model-Based Optimization (SMBO)
- Estimation of Forecast Uncertainty
- Data Spike Smoothing
- Epidemiological Model Structure
- Interpretability Of Epidemiological Models
- Notions Of Identifiability
- Non-Identifiability In Epidemiological Models

The ultimate aim of our modelling framework is twofold:

a. to provide reasonable estimates of future case counts, and

b. to estimate the underlying parameters of the epidemiological model, which carry independent meaning and interpretation that is useful for public health decision making.

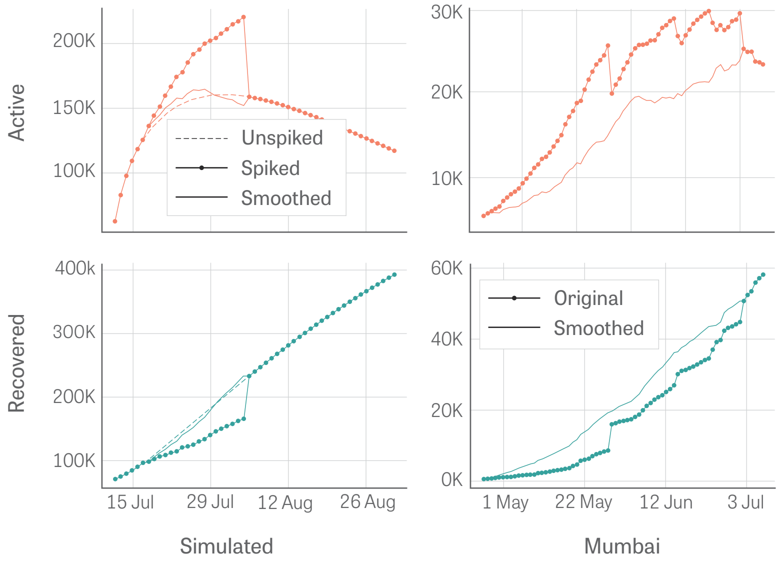
In both these cases, it is also important to provide uncertainties associated with a certain confidence level, since these uncertainties too must be potentially factored into decision making. The framework we currently have (open-sourced here) is general enough to be adapted to any compartmental model. Its technical components are the following.

Data Cleaning

A common issue with case count aggregate data is delays in reporting of infected cases, recoveries and deaths. The effect of these delays is that the reported count on any given day always lags behind the actual count. Usually, after many such days, the missing case counts from past days are consolidated and reported on a single day, resulting in a large spike in reported data. Because the spike results from consolidated reporting of past data, it is artificial and needs to be smoothed out before the data is fed into the modelling pipeline. This is the primary data cleaning operation that needs to be performed, although there may be instances of other types of incorrectly reported data or missing values due to manual errors.

We developed various techniques for data smoothing by redistributing case counts on the "spike" day to previous days in different ways. Our numerical experiments indicated that a "proportional count" method of smoothening, in which the number of case counts attributed to a given day was made proportional to the actual reported count on that day, turned out to be the most faithful method for reproducing true case counts.

Figure 2: **Performance of smoothing algorithm on simulated spike and Mumbai data.** Smoothing is done via the "Proportional counts" method.



Parameter Fitting

Historical case count data, aggregated at the city or district level, is used to estimate epidemiological parameters such as the disease incubation time, the basic reproduction number, and so on. Technically, these parameters are estimated through an iterative sequential sampling process called Sequential Model Based Optimization (SMBO). In each iteration of this process, the estimated parameters provide a better fit to the data. While in principle the fitting process need only be done once to estimate the relevant parameters, in reality the parameters keep changing as a consequence of varying lockdown measures, population migration, and the nature of the virus variant. This necessitates repeated fitting as more data comes in to ensure that any parameter values used for prediction of future case counts are based on fitting to the most recent data.

Forecast Uncertainty Estimation

Uncertainties in case count forecasts are a consequence of uncertainties in the underlying epidemiological parameters, which are themselves estimated through the fitting procedure described above. In principle, if the probability distribution of the parameters given historical case count data were known, then that distribution can be used to estimate uncertainties and confidence intervals for the parameters, and hence for the forecasted values. In practice, however, estimating this distribution is difficult and computationally expensive. It can be done through a Markov Chain Monte Carlo (MCMC) procedure, which is also a repeated sampling procedure.

We developed a fast, approximate procedure for computing uncertainties by sampling from the empirical distribution corresponding to the sequential samples generated by SMBO. These samples are then exponentially down-weighted by the value of the loss function for these samples. In other words, samples with small values of the loss (lower error) get much higher weights than samples with large values of the loss (higher error), and these weights are used to compute a weighted mean parameter value and a weighted variance in parameter values. This approximate procedure of computing uncertainty or variability in parameter values is termed Approximate Bayesian Model Averaging (ABMA).

Furthermore, the ABMA procedure also entails using the SMBO parameter samples to compute forecast trajectories and taking the same weighted averages and variances of these trajectories to compute uncertainties in forecasts as well. A similar procedure is used to generate forecast quantiles corresponding to pre-specified confidence levels.

Epidemiological Model Choice

The specific choice of epidemiological model in a given situation naturally depends on the type of disease and the nature of the pathogen causing it, but it also depends on a number of practical factors governed by answers to the following questions:

a. Is this the simplest possible model that can reasonably faithfully capture the disease dynamics (i.e., is the model *expressive* enough?)?
b. What is the kind of data that is available in order to estimate model parameters? Complex models with many parameters are often not identifiable (an issue we discuss below) because there simply isn't enough variety in the available data to drive parameter *learnability.*
c. Are the parameters readily *interpretable* so that the disease dynamics can be understood?
d. Is the model *generalizable* to more complex scenarios by incorporation of additional information as it becomes available?

> For the specific case of modelling the COVID-19 pandemic in India, we chose the simplest model compatible with the available data that is expressive enough to capture the disease dynamics.

For the specific case of modelling the COVID-19 pandemic in India, we chose the simplest model compatible with the available data that is expressive enough to capture the disease dynamics. Curve-fit models are not expressive since they do not typically take into account the decreasing susceptibility as the pandemic progresses; when they do, the underlying parameters are not readily interpretable. Agent-based models are typically over-parametrized and complex. A low-complexity compartmental model that we found to be expressive enough, yet generalizable enough to allow for modelling scenarios with changes in testing and lockdown policies, was the SEIARD model, comprising compartments corresponding to Susceptible (S), Exposed (E), Infectious (I), Active (A), Recovered (R), and Dead (D) cases.

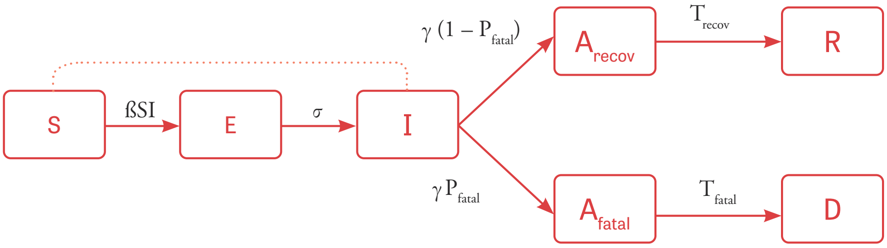


Figure 3: **SEIARD Model Structure.** Individuals transition between states of disease progression: Susceptible (**S**), Exposed (**E**), Infectious (**I**), Recovered (**R**), and Deceased (**D**). Active cases are split between A_{recov} and A_{fatal}.

Identifiability And Interpretability

The SEIARD model, like all compartmental models, has a number of epidemiological parameters that need to be specified in order to generate forecasts. These include the transmission rate, the infectious time period (the product of these is the basic reproduction number R₀), the incubation time period, and so on. In addition, compartments like E and I are *latent,* in the sense that the actual populations in these compartments are never observed (it is never known, for example, how many people are actually exposed to the pathogen, even though everyone may be susceptible). Therefore the initial values of the population in these compartments are also effectively parameters to be estimated. However, in practice, many of these parameters cannot be estimated accurately from case count data alone. These parameters are thus considered (in a sense clarified below) *non-identifiable.* Their values can only be fixed by additional knowledge of the spread of the pandemic that is not captured in aggregate case count data. Patient linelist data, for example, can be used to better pin down the incubation time period and render it effectively identifiable.

Note that identifiability of parameters is key to their interpretability. Without being able to estimate parameter values with reasonable certainty, we cannot interpret or assign meaning to their values. Furthermore, large uncertainty in parameter values leads to reduced accuracy on long-time forecasts, as illustrated in the figure below, where the forecasting errors between a non-identifiable model and its reparameterized version (later shown to be identifiable) are compared.

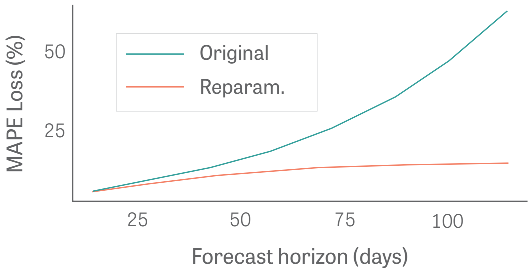


Figure 4: **Forecast Error**

There are many factors that lead to non-identifiability in epidemiological models, including the details of the model dynamics, the loss function and method used to fit parameters, and the quality and quantity of data available. It is common to classify identifiability into two broad classes:

Structural identifiability, which is purely dependent on the nature of the compartmental model and the underlying mathematical equations describing the dynamics of transitions between compartments, and

Practical identifiability, which is a less precise notion, but depends on the details of the parameter fitting process. This is made quantitative by computing confidence intervals for the parameter in questions. Thus we may say that, given the data and fitting process, a certain parameter can be estimated to within a certain range (corresponding to the interval) with a certain level of confidence.

Note that structural identifiability is a prerequisite for practical identifiability: if a parameter is not structurally identifiable, it cannot be estimated even in a practical context. However, there are parameters that may be structurally identifiable in a rigorous sense but are practically very difficult to estimate because of paucity of data or a quirk in the fitting process.

In the specific case of the SEIARD model, the initial active cases are split between those that eventually recover and those that eventually die. This split is structurally non-identifiable: aggregate case count data cannot be used to separately estimate these two types of active cases. A large number of eventually fatal cases, with a long time to mortality, would be indistinguishable (in model fitting) from a small number of eventually fatal cases with short time to mortality. Moreover, the latent compartments also cannot be estimated from case count data alone.

Our contribution to identifiability analysis consisted of making precise a new, intermediate type of identifiability, which we call *statistical identifiability,* a situation in which the loss function used for fitting parameters to data is convex in nature, thus possessing a unique minimum value. Note that this concept is different from structural identifiability because it depends on the choice of loss functions, and also different from the concept of practical identifiability since it does not depend on the width of the minimum of the loss function, only on the shape itself.

Table 1: Notions of identifiability

	STRUCTURAL	STATISTICAL	PRACTICAL
Model Form	✓	✓	✓
Loss function		✓	✓
Observation interval		✓	✓
Noisy data			✓
Fitting method			✓

In the SEIARD model, structural non-identifiability manifests as statistical non-identifiability of the main parameters.

In the SEIARD model, structural non-identifiability manifests as statistical non-identifiability of the main parameters. However, if parameters like the incubation period, the infectious period, and the time to death can be estimated from patient linelists, the model becomes structurally and statistically identifiable. Note that practical identifiability is still not guaranteed since a large amount of data may be required to pin the parameters down to confidence intervals that are narrow enough for the parameter values to be interpretable and meaningful. ■

04

Empirical Results

TOPICS COVERED IN THIS SECTION

- Performance of our framework on Mumbai city forecasts
- Comparison with SOTA Models

The value of our framework was demonstrated by the following:

- Empirical results validating choices of hyperparameters, uncertainty estimation, and data preprocessing
- Field performance and impact of the deployed system in Mumbai, including translation of case forecasts to capacity requirements in hospitals
- Empirical comparison of our approach with other state-of-the-art (SOTA) models on COVID-19 case data from the USA

Performance of our framework on Mumbai city forecasts

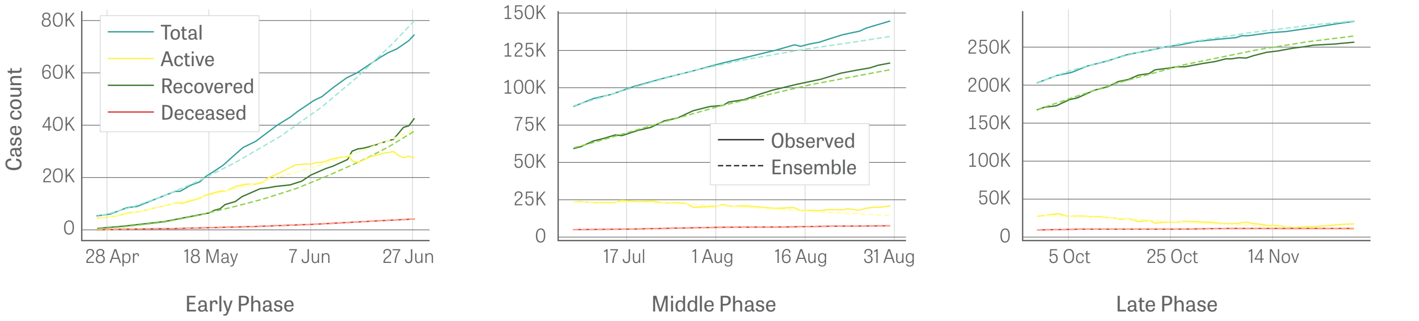


Figure 4: **Performance of ABMA on Predicting Mumbai Caseload.** Predicted and ground truth case counts for Mumbai city across compartments and phases of the pandemic. Here, the parameter fitting period is T_θ= 30 days and the fitting period for the hyperparameter α is $T\alpha$= 3 days (see Supplement). Forecasts are shown 30 days beyond the end of the $T\alpha$ fitting period. All dates refer to the year 2020.

PHASE	PARAMETER VALUES								COMPARTMENT MAPE LOSS (%)			
	R_0	T_{inf}	T_{inc}	T_{recov}	T_{fatal}	P_{fatal}	E_{active_ratio}	I_{active_ratio}	C	A	R	D
Early	1.18	3.59	4.29	30.14	14.70	0.07	0.95	0.12	6.91	11.33	12.91	3.78
Middle	0.83	3.86	4.17	19.32	24.32	0.05	0.30	0.18	3.10	9.04	2.39	1.05
Late	0.81	3.72	4.42	11.75	11.87	0.02	0.35	0.37	1.20	17.54	2.39	1.43

Table 2: **Quantitative Performance of ABMA on Mumbai.** Ensemble mean (ABMA) *parameters,* and test MAPE loss (%) for each *compartment* of the ABMA forecasts. Note that the ABMA forecast is the mean forecast, not the forecast of the mean parameters.

The ABMA framework was used to provide actionable insights for the city of Mumbai, India from May 2020 to October 2020, as well as in the state of Jharkhand, India. The figure and table above confirm the high accuracy and interpretability of our methodology across different phases of the pandemic. Recommendations made using this forecasting framework helped increase Mumbai's ICU bed capacity by over 1200 units, with 95% utilisation of ICUs when hospitalisations peaked. Moreover, over the deployment period, no absolute shortfall of critical health care resources became apparent, demonstrating the value of recommendations.

Comparison with SOTA Models

RANK	MODEL	MEDIAN MAPE (%)
1	UMass-MechBayes	1.21
2	Karlen-pypm	1.32
3	SteveMcConnel-CovidComplete	1.33
4	**ABMA**	**1.38**
5	YYG-ParamSearch	1.44
6	UCLA-SuEIR	1.49
7	PSE-DRAFT	1.65
8	DDS-NBDS	1.70
9	CEID-Walk	1.77
10	COVIDhub-baseline	1.71

Table 3: **Performance of top 10 Reichlab models submitting forecasts for at least 45 regions on cumulative death counts.** All models were fitted on data from 18 Aug to 19 Sept 2020. Based on hyperparameter optimisation for Mumbai, we chose T_θ= 30, $T\alpha$= 3. Model forecasts four weeks into the future starting 20 Sept 2020, aggregated every week, were evaluated by computing their MAPE values for every region, and then taking the median value across all regions (reported here).

We used the ReichLab forecasting hub for US states as a basis for comparing the performance of our forecasting framework with other methods, since no similar forecasting hub was available in India where models could transparently be compared. The source of data for the forecasts is the John Hopkins University CSSE data. We evaluated our method solely on regions where all four primary case counts were available (44 states plus Washington D.C.). The basis for comparison with other models is the MAPE value on deceased case counts. We found that 26 models submitted to ReichLab had submissions for at least 45 regions over the duration studied, with a range of median MAPE values between 1.21% and 4.26%. Our forecasts have low error and compare well to other models without the need for any special customisation of hyperparameters or fitting method.

05

Insights Acquired

TOPICS COVERED IN THIS SECTION

- Insights We Acquired From Our Research
- Lessons for Deployment

Dynamic data entails humans in the loop: Data collection and management activities during a pandemic are adversely impacted due to severe demands on public health authorities, leading to data discrepancies.

The data-related challenges prompted us to rely heavily on humans in the loop for tracking evolving data definitions and carrying out semi-automated quality checks. Additionally, data versioning and pre-processing prior to modelling were essential.

Model interpretability and identifiability is paramount: Our choice of the SEIARD model over other model families was motivated by planning needs and policy choices. Hence, the parameters had to be interpretable, independently verifiable where possible, and robustly estimable from the available data.

Application needs should dictate modelling choices: We focused on capacity planning, which impacted policies on capacity that took about a month to implement. We therefore customised our model fitting with loss computed over this time horizon. Furthermore, we adapted models in accordance with the information provided on upcoming policy changes and events (for instance, festivals) to generate accurate forecasts.

Insights must be actionable: Model insights had to be translated to concrete action guidance to enable smooth planning. Uncertainty estimation allowed us to provide three relevant scenarios which included a planning scenario, a high case scenario, and a low case scenario. Localised testing levels, evolving severity of cases, and sero-surveillance information to comprehend the state of the pandemic were pivotal factors informing the selection of planning scenarios.

Capacity gaps at the last mile are hard to anticipate: We recognise that while the use of our framework mitigated capacity shortfalls, the ability of a critical patient to access these resources is mediated by other factors. These include access to information on the availability of beds, local emergency transportation, the ability to pay for treatment, and other equity considerations, all of which need to be addressed within a larger framework of pandemic response. ■

Conclusions And Ensuing Work

We have presented a flexible modelling framework and demonstrated its value for epidemic forecasting utilising the case count aggregate data available in a constrained public health setting. The deployed system was used to drive decision making and planning with good accuracy (worst case MAPE < 20%) during the COVID-19 pandemic in Mumbai and Jharkhand respectively.

Our framework enables rapid forecasting with uncertainty estimates and is extensible to other model families and to different loss functions. Additionally, it allows the optimisation of hyper parameters such as fitting durations and ensemble weights. We motivated the choice of the specific compartmental model opted for via identifiability of the underlying parameters given the data constraints. Empirical comparison of our methods with other advanced models in the ReichLab hub on real-world data in the USA further points to their efficacy.

We also clarified notions of identifiability of parameters, specifically the interplay between structural identifiability, statistical identifiability, and practical identifiability intervals. We present these ideas and empirical results in the specific context of the SEIARD compartmental model. In the future, we plan to explore connections between different types of identifiability and empirically analyse the identifiability of multiple SEIR variants on real and synthetic data. This framework may be applied to the estimation of case burden in other infectious diseases such as Tuberculosis which are widespread across the developing world.

Acknowledgements

We would like to express our sincere gratitude to our scientists: Sansiddh Jain, Avtansh Tiwari, Nayana Bannur, Ayush Deva, Siddhant Shingi, Vishwa Shah, Mihir Kulkarni, Namrata Deka, Keshav Ramaswami, Vasudha Khare, Harsh Maheshwari, Soma Dhavala, Jithin Sreedharan, Jerome White, Srujana Merugu and Alpan Raval for their valuable contributions.

We would like to thank Janak Shah and the Wadhwani AI Programs team for their unparalleled support with data and deployment processes.

We would like to thank Rusha Sen from the Communications team for drafting this guidebook.

We would like to thank Arun Kale from the Communications team for designing this guidebook.

This study is made possible by the generous support of the American People through the United States Agency for International Development (USAID). The work described in this playbook was implemented under the TRACE-TB Project, managed by Wadhwani Institute for Artificial Intelligence (WIAI) under the terms of Cooperative Agreement Number 72038620CA00006. The contents of this manuscript are the sole responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government.

This work is co-funded by the Bill and Melinda Gates Foundation, Fondation Botnar and CSIR - Institute of Genomics and Integrative Biology. It was enabled by several partners through the informal COVID-19 data science consortium, most notably Shreyas Shetty, Anupama Agarwal, and a group of volunteer data scientists under the umbrella group "Data Science India vs. COVID".

We owe a debt of gratitude to AmazonWeb Services for Cloud credits and support; the Smart Cities Mission for providing early impetus to this work; and the Brihanmumbai Municipal Corporation (BMC) and Integrated Disease Surveillance Program (Jharkhand) for assistance whilst facing unprecedented challenges.

WADHWANI AI

wadhwaniai.org

USAID
FROM THE AMERICAN PEOPLE

DISCLAIMER: This document is made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents of this brochure are the sole responsibility of Wadhwani AI and do not necessarily reflect the views of USAID or the United States Government.