BMJ Open

Comprehensive compartmental model and calibration algorithm for the study of clinical implications of the population-level spread of COVID-19: a study protocol

Brandon Robinson,[1] Jodi D Edwards,[2,3] Tetyana Kendzerska,[3,4,5] Chris L Pettit,[6] Dominique Poirel,[7] John M Daly,[8] Mehdi Ammi,[9] Mohammad Khalil,[10] Peter J Taillon,[11] Rimple Sandhu,[12] Shirley Mills,[13] Sunita Mulpuru,[4,5] Thomas Walker,[1] Valerie Percival,[14] Victorita Dolean,[15,16] Abhijit Sarkar [1]

To cite: Robinson B, Edwards JD, Kendzerska T, *et al*. Comprehensive compartmental model and calibration algorithm for the study of clinical implications of the population-level spread of COVID-19: a study protocol. *BMJ Open* 2022;**12**:e052681. doi:10.1136/bmjopen-2021-052681

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-052681).

Received 23 April 2021
Accepted 13 January 2022

For numbered affiliations see end of article.

Correspondence to
Professor Abhijit Sarkar; abhijit.sarkar@carleton.ca

ABSTRACT

Introduction The complex dynamics of the coronavirus disease 2019 (COVID-19) pandemic has made obtaining reliable long-term forecasts of the disease progression difficult. Simple mechanistic models with deterministic parameters are useful for short-term predictions but have ultimately been unsuccessful in extrapolating the trajectory of the pandemic because of unmodelled dynamics and the unrealistic level of certainty that is assumed in the predictions.

Methods and analysis We propose a 22-compartment epidemiological model that includes compartments not previously considered concurrently, to account for the effects of vaccination, asymptomatic individuals, inadequate access to hospital care, post-acute COVID-19 and recovery with long-term health complications. Additionally, new connections between compartments introduce new dynamics to the system and provide a framework to study the sensitivity of model outputs to several concurrent effects, including temporary immunity, vaccination rate and vaccine effectiveness. Subject to data availability for a given region, we discuss a means by which population demographics (age, comorbidity, socioeconomic status, sex and geographical location) and clinically relevant information (different variants, different vaccines) can be incorporated within the 22-compartment framework. Considering a probabilistic interpretation of the parameters allows the model's predictions to reflect the current state of uncertainty about the model parameters and model states. We propose the use of a sparse Bayesian learning algorithm for parameter calibration and model selection. This methodology considers a combination of prescribed parameter prior distributions for parameters that are known to be essential to the modelled dynamics and automatic relevance determination priors for parameters whose relevance is questionable. This is useful as it helps prevent overfitting the available epidemiological data when calibrating the parameters of the proposed model. Population-level administrative health data will serve as partial observations of the model states.

Ethics and dissemination Approved by Carleton University's Research Ethics Board-B (clearance ID: 114596). Results will be made available through future publication.

Strengths and limitations of this study

- ► New compartments and parameters are introduced to model more complex disease dynamics and to capture clinically relevant quantities of interest.
- ► The increased complexity of the mechanistic model is complemented by a non-linear sparse Bayesian learning algorithm for model calibration to help avoid overfitting the available data.
- ► Population-level modelling averages across potentially highly varying demographics of different communities within the region of interest and lacks the spatial resolution for capturing localised activity.

INTRODUCTION

Since first being identified in December 2019, the coronavirus disease 2019 (COVID-19) has spread across the world, creating a global health crisis. To date (7 December 2021), 266 457 039 confirmed cases have been recorded worldwide with 5 262 849 deaths.[1] It has become critically important to have reliable methods to model and predict the transmission of COVID-19 to inform policy decisions and forecast health system resource utilisation.

Value added

As the pandemic progresses, we are learning more about the transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and about the clinical effects of COVID-19 on individuals. The fundamental

Kermack-McKendrick susceptible–infectious–recovered (SIR) model[2] and adaptions to this model have been used to try to understand and predict both short- and long-term case counts, to strategically manage healthcare resources, and to inform public health policies designed to control the spread of the virus. The simplicity of these models makes them convenient tools from a mathematical perspective,[3–5] and allows them to capture salient trends in disease progression and project short-term growth[6] and assess critical quantities of interest such as the reproduction number (an indicator of the transmissibility of infectious and/or parasitic agents[7]). However, their simplicity limits their utility for the objectives of the current protocol as they lack the refinement to account for the specific clinically distinct classes of individuals we seek to quantify and may oversimplify the complex dynamics and global nature of the COVID-19 pandemic reducing reliability of long-term forecasts.[8] Tuite *et al*[9] presented an elaborate model of the transmission of COVID-19 in the province of Ontario, Canada. It consists of 16 compartments stratified by age and comorbidity, representing the largest number of unique compartments used in the study of population-level transmission of COVID-19. We propose to expand upon this model, increasing the model complexity in two ways: (i) the addition of six new compartments as depicted in figure 1, to incorporate the effects of vaccination, asymptomatic carriers (quarantining and not), inadequate access to hospital or intensive care unit (ICU) resources, recovery with long-term health complications and post-acute COVID-19 and (ii) within the above described 22-compartment framework, we can incorporate more information as data become available through stratification of the model, allowing for population demographics (age, comorbidity, socioeconomic status, sex and geographic location) and clinically relevant information (vaccination status, variants of COVID-19) to be reflected through the model parameters. Beyond allowing for clinically relevant quantities of interest to be accounted for explicitly within the model through additional compartments, the increased resolution of the 22-compartment model also allows for inter-compartment dynamics and interactions to be captured.

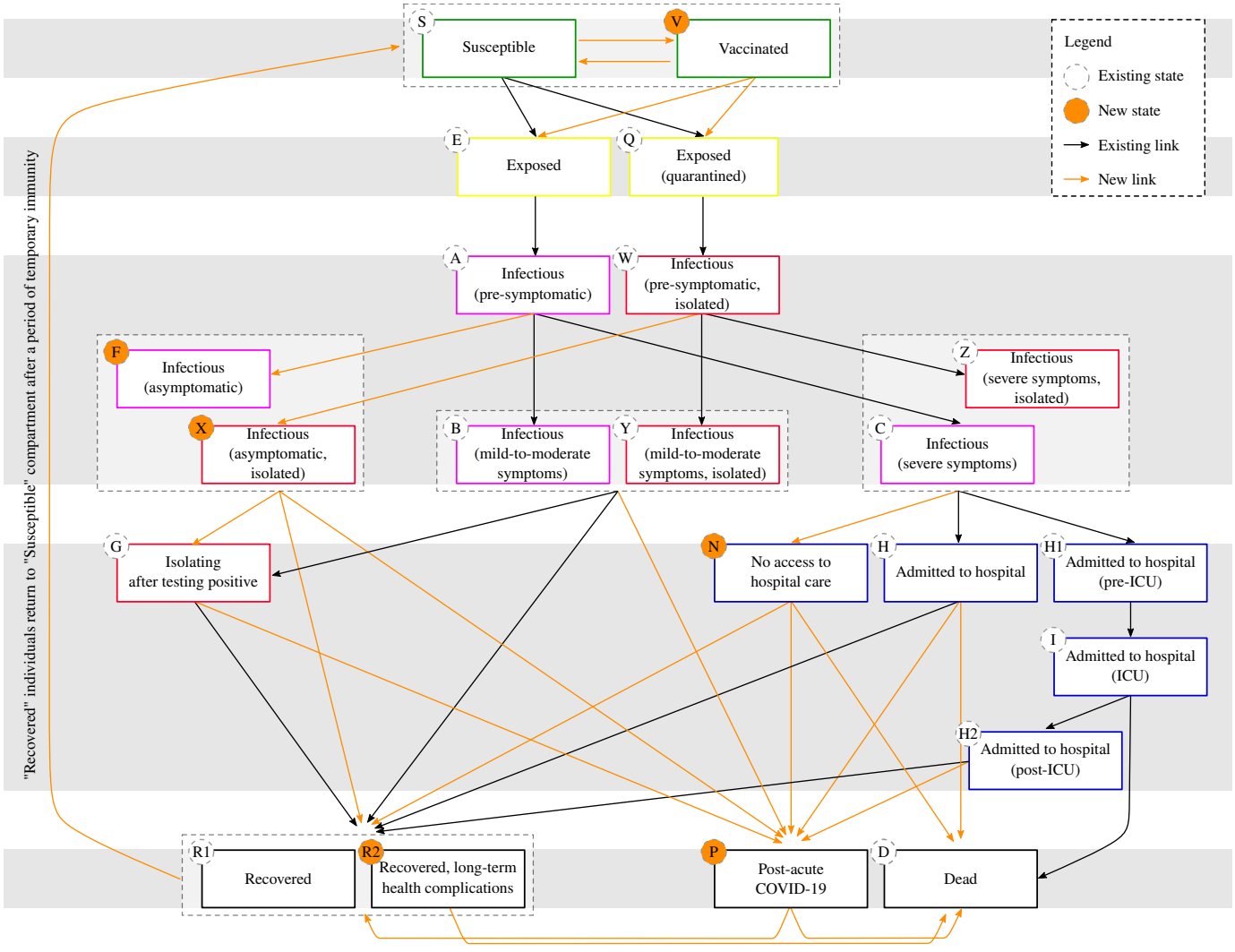

Figure 1 Flowchart of the proposed 22-compartment model, highlighting the extensions to the 16-compartment model by Tuite *et al*.[9]

This increase in model complexity is complemented by the proposed use of the non-linear sparse Bayesian learning[10 11] algorithm for parameter calibration and model selection. Using epidemiological data from a given region, model parameters may be calibrated using a traditional Bayesian statistical framework. Many model parameters have a clinical interpretation; hence, approaching the problem from a Bayesian perspective will permit this knowledge to be reflected through informative parameter prior distributions. In an extensive comparison of 22 individual models, Cramer[12] showed the Bayesian compartmental model considered in the study[13] better captured the true case counts within its probability intervals, and was among the models with the lowest mean absolute error in its predictions. Beyond a standard Bayesian approach, we propose an automatic optimal model discovery process, using sparse and noisy observations, to identify a low-dimensional model that is nested under a potentially overparameterised COVID-19 compartmental model. This discovery process takes place in the presence of model error (imperfection), and little (or no) prior information may be available on some parameters. The inference procedure, powered by the non-linear sparse Bayesian learning algorithm for non-linear dynamics, has the goal of optimally balancing average data fit and model complexity (Occam's razor[14 15]) to avoid overfitting sparse data. The goal is to obtain a comprehensive compartmental model that will generalise beyond the timeframe of the observed data to provide reliable predictions with reduced uncertainty compared to standard Bayesian approaches. Sparse learning in epidemiological models was previously approached from a non-Bayesian perspective using sparse identification of non-linear dynamical systems (SINDy).[16] Horrocks and Bauch,[17 18] used the SINDy approach for an SIR model with modified transmission dynamics and data sets for measles, varicella and rubella.

Objective

To avoid oversimplifying the epidemiological dynamics and to account for structural errors or imperfections that are inherent in any model of complex systems, we have increased the complexity of the underlying mechanistic model, which is designed to capture more of the system dynamics. This is complemented by the addition of an explicit model error term (as described in the online supplemental material 1) whose characteristics can be inferred by Bayesian inference algorithms. We hypothesise that the low-dimensional model (nested within the proposed stratified stochastic 22-compartment model) informed by heterogeneous data will have better predictive capabilities (less bias and uncertainty as demonstrated by Sandhu *et al*[10] for engineering systems). Obtaining the data-optimal model will help advance our understanding of the mechanics of the COVID-19 pandemic at a population-level scale, by identifying various critical time-varying and time-invariant parameters that drive the spread such as the reproduction number (see the approach described in Allenman *et al*[19] Diekmann *et al*[20] for a stratified model with multiple infectious compartments). It will also allow for the estimation of clinically relevant quantities of interest and for the forecasting of various what-if scenarios to predict the short- and long-term demand on healthcare systems.

METHODS AND ANALYSIS

Model updates reflecting the evolving knowledge of COVID-19 and its dynamics will occur in two stages. First, new compartments, new connections between compartments and model stratification are incorporated into the mechanistic compartmental model, such that the observed dynamics may be replicated. Second, model parameters are to be calibrated using the data collected to date and subsequently continuously updated as new data become available for real-time forecasting. In this section, we introduce a new proposed mechanistic model and model stratifications, and subsequently discuss the algorithmic development and data sources that will be used to implement this model for real-time prediction.

Mechanistic model framework

There are a number of variations of the SIR model, each designed ad hoc to evaluate a specific phenomenon relevant to a disease outbreak of interest. The 16-compartment model from Tuite *et al*[9] combines many control measures such as physical distancing and quarantining, as well as modelling the burden on hospital and ICU resources, and it effectively addresses many pressing challenges that were present during the first wave of the pandemic. The proposed 22-compartment model builds on this model, so for consistency, the same labels and symbols are used wherever possible (particularly in the online supplemental material 1). Critically, the proposed model introduces new phenomena to the system: (i) vaccination, (ii) reinfection with COVID-19 or a new variant of concern, (iii) asymptomatic carriers, (iv) inadequate access to hospital resources (accounting for deaths occurring outside of hospitals, individuals in long-term care or the scenario wherein demand for ICU resources and ventilators exceeds capacity) and (v) recovery with long-term health complications and post-acute COVID-19.

The flowchart in figure 1 depicts the proposed compartmental model, identifying all 22 compartments; the arrows indicate the pathways by which individuals may flow between compartments. The flowchart provides an explicit visual representation of the model equations, outlined in the online supplemental material 1, along with a summary of the model states and parameters. Readers are encouraged to refer to Tuite *et al*[9] to see the foundational model; however, for convenience, the extensions to the model are indicated by orange highlights. Six new compartments are proposed (indicated with an orange symbol in the top-left corner): vaccinated (V), infectious asymptomatic (F), infectious asymptomatic, isolated (X), no access to hospital care (N), post-acute COVID-19 (P)

and recovered with long-term health complications (R2). Additionally, the orange-coloured arrows denote new connections between compartments.

The model considers two compartments of individuals (top row) who may become infected with COVID-19. Upon being exposed (infected but not yet infectious), individuals from these two compartments will flow to one of two exposed compartments (second row) and enter one of two tracks depending on whether they are isolating. The inclusion of an isolation track extends the susceptible–exposed–infectious–recovered model to incorporate information on the effectiveness of contact tracing and other measures for preventing transmission[9 21] as in the six-compartment (susceptible, exposed, exposed and quarantined, infectious, infectious and quarantined, recovered) model inspired by the SARS outbreak.[22] After the viral incubation period, individuals are considered infectious and can now transmit the disease; hence, they proceed into one of two infectious pre-symptomatic compartments (third row) along their current isolation track. Following a pre-symptomatic infectious period, the infectious individuals will be separated into three classes along their current isolation track, based on the severity of symptoms: asymptomatic, mild to moderate and severe. Symptoms are deemed to be severe if they warrant hospitalisation; otherwise, the symptoms are categorised as mild to moderate. Individuals with severe symptoms will proceed to a hospital track, entering one of the three compartments (fourth row). After various periods of time, the portion of the population with acute COVID-19 will proceed directly to one of two recovery compartments or the dead compartment (fifth row). Individuals whose symptoms persist beyond the typical symptomatic period will proceed to an intermediate compartment accounting for post-acute COVID-19 prior to transitioning to the recovered or dead compartments.[23] Individuals who are asymptomatic or experience mild-to-moderate symptoms will recover and enter one of the two recovery compartments (full recovery or recovery with long-term complications) or the post-acute COVID-19 compartment. Those who were previously not on the isolating track may enter the isolating track after testing positive once symptoms arise. The key dynamics that are introduced in this model are discussed in the subsections that follow.

Temporary immunity

Recent evidence suggests the possibility of reinfection with COVID-19 after recovery,[24] and so temporary immunity is modelled by the same mechanics of a simple susceptible–infectious–recovered–susceptible (SIRS) model. After entering one of the recovered compartments (R1 or R2), the individual will be returned to the susceptible compartment (S) according to the average duration of temporary immunity. The recovered compartment is, therefore, no longer a final state; hence, in long-term forecasts, these compartments may not necessarily increase monotonically. Note that the dead compartment (D) is now the only final compartment.

Vaccination (V)

Vaccination resulting in temporary immunity is modelled by removing individuals from the susceptible (S) compartment and placing them in the vaccinated (V) compartment according to the rate of vaccination. This rate parameter may vary in time due to the availability of the vaccines and government policies for vaccine roll-out, and it may accordingly vary based on age, comorbidity or other factors addressed in the Stratification by characteristics of the population section. Vaccinated individuals should be reintroduced into the susceptible compartment at a rate determined by the average duration of protection from vaccination. The framework allows for an imperfect vaccine (providing less than 100% immunity) to be modelled,[25] enabling vaccinated individuals to become infected, and, therefore, to proceed through the flowchart due to inefficacy of the administered vaccine. Vaccine models as simple as the three-compartment (susceptible–vaccinated–infectious) model[26] have been used to study the influence of vaccination on disease control and have been used previously in the COVID-19 literature for studying the control of the disease.[27] The effects of having multiple vaccines with different clinical properties being administered to the public can be modelled through the stratification of the model as outlined in the Vaccination and COVID-19 variants section.

Asymptomatic carriers (F and X)

Two compartments have been introduced to model asymptomatic carriers, who are undergoing isolation (X) and who are not (F). The inclusion of explicit compartments to quantify asymptomatic carriers has also been used previously, such as in the study of influenza[28] and COVID-19,[29] respectively. Due to the non-linear interaction of the susceptible and infectious classes (see equation (A.23) in the online supplemental material 1), these additional infectious compartments could have a significant influence on the model output. Furthermore, these compartments allow the model to project the influence of government policies on the testing of asymptomatic individuals, or to retrospectively study the effect asymptomatic carriers had on case counts through undetected community transmission.

Inadequate access to hospital resources (N)

The distinction between mild-to-moderate and severe symptomatic infections was defined as whether cases warrant hospitalisation. The inclusion of a compartment that accounts for inadequate access to hospital resources provides a mechanism to account for severe cases that result in death, but that are not accounted for in hospital or in ICU statistics. This compartment also provides a mechanism to assess worst-case scenarios, where the demand for hospital and ICU resources exceeds capacity.

Post-acute COVID-19 (P) and recovery with long-term health complications (R2)

The compartmental model includes a compartment for post-acute symptomatic COVID-19, and two distinct

recovery compartments: one that assumes a full recovery (R1) and a second in which individuals recover but are subjected to long-term health complications (R2). This second recovery compartment allows for health service utilisation, and for deaths resulting from long-term health complications to be modelled for long-term forecasts. For model stratifications that consider pre-existing health conditions (see discussion on comorbidity in the Stratification by characteristics of the population section), this also allows for a mechanism whereby individuals may be transferred between health states when returning to the susceptible compartment under the assumption of temporary immunity (see equation (A.1) in the online supplemental material 1).

Model stratification

We now discuss how this base model will consider a combination of the effects of age, comorbidity, sex, socioeconomic status, geographical location, multiple variants of COVID-19 and different vaccines, without requiring any further modification to the 22-compartment model's structure. In the most general sense, this is achieved by stratifying the model to capture the desired effects, such that a number of coupled 22-compartment models will exist in parallel for each possible combination of modelled effects. Tuite *et al*'s 16-compartment model[9] is stratified by age into 16 age groups with equal widths of 5 years and includes a second stratification indicating whether an individual has a pre-existing health condition. These two model stratifications allow for clinically relevant information to be explicitly modelled, and for age- and health-specific model predictions to be obtained, as parameters are multidimensional arrays. We propose further use of stratification to account for additional demographic and clinical phenomena.

Stratification by characteristics of the population

The population can be optimally stratified by age to reflect age-dependent differences in COVID-19 transmission, clinical outcomes and policy decisions that affect specific demographic groups (eg, age-based vaccination priority). Grouping the population based on specific pre-existing health conditions that are known to be relevant to COVID-19 (respiratory diseases, cardiovascular diseases, autoimmune diseases, etc) is also important in forecasting the outcomes of infections at the population level. Further stratifications based on socioeconomic status and sex are possible. For example, these model stratifications can be leveraged to model outbreaks among long-term care residents or the increased exposure of individuals of lower socioeconomic status, whose occupations may result in more daily interactions than people who are able to work from home.

One may even account for geographical location in a rudimentary sense, using a multiregional discrete model as in Zakary *et al*.[30] This is achieved by assigning a specific index to population centres and accounting for the travel between these locations through a coupling term. A more formal account of these effects would require a partial differential equation model (such as in Viguerie *et al*[31–33]), which effectively extends the current ordinary differential equation framework by accounting for population densities and the spatio-temporal movement of individuals by means of a diffusion term. One would need algorithmic developments that allow for the propagation of uncertainty in the large-scale problem by leveraging high-performance computing platforms and domain decomposition methods[34] like those outlined by Desai *et al*.[35]

Vaccination and COVID-19 variants

Mutations of SARS-CoV-2 into new variants[36] and the subsequent modelling of human-to-human transmission of these variants can also be achieved through the introduction of an additional index that stratifies the model further, as in an n-strain model.[37] This approach would assign an index to each distinct strain of the virus and allow for parameter values to vary according to the clinical characteristics of that particular strain. The inclusion of a model stratification for multiple variants has future implications as well, as the emergence of escape variants may cause the pandemic to persist despite widespread vaccination efforts.[38 39]

The model includes a compartment (V) to account for the vaccinated population, but to model vaccines that do not provide 100% immunity against infection, an additional stratification could be introduced that accounts for the vaccination status of individuals who become infected. This could be of use when modelling multi-dose vaccines and vaccine boosters, respectively. The additional index would also allow modellers to reflect how a vaccinated person's experience with the disease may differ from an unvaccinated infected individual (eg, reduced probability of severe infection) by modifying the associated parameter values for the given index. As more data become available, this additional index could also allow for transmission-related differences between mRNA vaccines and viral vector vaccines to be modelled.[40]

Bayesian calibration of the proposed 22-compartment COVID-19 model

The data available for model calibration from testing and public health databases represent incomplete measurements of the model states. Hence, adopting a Bayesian framework for the calibration of the model allows for more reliable long-term forecasting as it allows the modeller to impose known transmission dynamics through the model, rather than relying on patterns in the data alone. Prior knowledge of the model parameters is included through the assignment of parameter prior distributions. This prior knowledge is then updated based on the available data to obtain a parameter posterior distribution, which is used for forecasting. The probabilistic representation of the parameters allows for the uncertainty in the states and parameters to be propagated through the model to obtain predictions with associated uncertainty intervals.

Despite the current effort to extend the compartmental model to include more relevant disease dynamics, there are certainly unmodelled phenomena that will contribute to the transmission of the disease. To address concerns of model inadequacy (stemming from the lack of knowledge, unmodelled dynamics and reduced order modelling), an additive white or coloured noise model discrepancy term will be added to the dynamics, resulting in a stochastic compartmental model.

As we increase the complexity of the model to capture more phenomena, we must be mindful that predictions obtained using overparameterised models calibrated with limited data can exhibit large uncertainty due to overfitting. Furthermore, an inappropriate choice of parameter prior distribution by the modeller may introduce bias and lead to erroneous predictions, as probabilistic predictions are sensitive to the choice of priors in the case of sparse observations. For model calibration, non-informative priors are often assigned to some model parameters for which there is little or no prior information. The selection of priors for such parameters can be handled through the concept of automatic relevance determination (ARD),[41 42] thereby extending the Bayesian parameter estimation and model selection framework outlined by Sandhu *et al*.[43–46] This addresses the two concerns above, as it induces sparsity in the unknown parameter space during model calibration, helping to prevent overfitting, and it removes the onus of choosing parameter priors for parameters that are not well understood from the individual modeller and instead relies on data-informed priors. Assigning a combination of ARD priors and known priors, this approach performs automatic model reduction in nonlinear dynamics using a hybrid scheme to prune redundant parameters.[47] As a result, one or a few nested models (under the more complex model) are identified that balance average data fit and model complexity. Through Bayesian model selection aided by ARD, the data-optimal dynamical model and model error will be simultaneously identified.

Data

The province of Ontario, Canada, represents an interesting case study owing to heterogeneity in the demographics and population density across the region. For Bayesian analysis, we will use data from linked health administrative databases housed at ICES[48] and public health data[49] from the province of Ontario, the largest province in Canada. Ontario has high-density, ethnically and socioeconomically diverse metropolitan regions as well as large low-density rural areas with more homogeneous demographics.

It is important to note that data will not exist for each of the 22 unique model compartments, and compartments that are observable will largely consist of biased and noisy measurements. In reference to figure 1, we anticipate that data will be available concerning: vaccination (V), isolation after testing positive (G), the four compartments relating to hospital care (H, H1, I and H2), and the dead compartment (D). Various parameters may also be informed by systematic review, such as the demographics within the region of interest, and various clinical parameters that need not be inferred from the data. Other compartments are hidden variables that will need to be determined through a combination of the mechanics of the stochastic compartmental model and the data, using non-linear filters such as the extended Kalman filter, ensemble Kalman filter or particle filter for state estimation.[50]

Patient and public involvement

No patients involved. ICES has a public engagement team, which advises researchers and staff who are interested in engaging with the public. We will leverage the ICES Public Advisory Council to provide perspectives from public members.

Planned start and end date for the study

Data collection through ICES will tentatively run for 2 years beginning in September 2022, with an additional year anticipated thereafter for the analysis and summary of findings.

DISCUSSION

The increased complexity of the proposed 22-compartment model will allow for a more comprehensive account of the underlying dynamics of the pandemic, which we hypothesise will provide a means to obtain more accurate predictions than previous models. The proposed Bayesian framework addresses concerns of overfitting, model error and the estimation of time-varying parameters using available public health data. The data-optimal sparse representation of the observed dynamics allows for predictions with less uncertainty than models calibrated using standard Bayesian approaches.

In the short term, the proposed research effort will allow for the calibration of the model within a probabilistic setting, which will then lend itself to forecasting case counts and the associated anticipated burden on healthcare resources under uncertainty. The long-term implications of this research will extend beyond the height of the current pandemic. From a clinical perspective, relevant quantities of interest that the model framework seeks to capture include: the influence of asymptomatic carriers (compartments F and X), vaccination (compartment V, and through model stratification), deaths occurring outside of hospital (compartment N) and long COVID-19 (compartments P and R2), as well as potential implications of temporary immunity and COVID-19 variants. The relevance of age, comorbidity, socioeconomic status or sex to the predicted clinical outcomes may also be quantified. Furthermore, many retrospective analyses may be performed. For example, estimates of the true case counts, obtained through state estimation, may be used to study the effectiveness of testing efforts as well as the effectiveness of policy-based control measures in

mitigating community transmission. Finally, through machine learning techniques like transfer learning,[51] the calibrated model for the COVID-19 pandemic can be methodically used to inform parameter priors as appropriate for the modelling of future epidemics and pandemics.

Ethics and dissemination

This study was approved by Carleton University's Research Ethics Board-B (clearance ID: 114596). Results will be made available through future publication.

Author affiliations
[1]Department of Civil and Environmental Engineering, Carleton University, Ottawa, Ontario, Canada
[2]School of Epidemiology and Public Health, University of Ottawa and University of Ottawa Heart Institute, Ottawa, Ontario, Canada
[3]ICES, Ottawa, Ontario, Canada
[4]The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
[5]Department of Medicine, Faculty of Medicine, Division of Respirology, University of Ottawa, Ottawa, Ontario, Canada
[6]US Naval Academy, Aerospace Engineering Department, Annapolis, Maryland, USA
[7]Royal Military College of Canada, Department of Mechanical and Aerospace Engineering, Kingston, Ontario, Canada
[8]Independent Control Systems Engineer, Ottawa, Ontario, Canada
[9]School of Public Policy and Administration, Carleton University, Ottawa, Ontario, Canada
[10]Sandia National Laboratories, Livermore, California, USA
[11]Schaffen Research Inc, Ottawa, Ontario, Canada
[12]National Renewable Energy Laboratory, Golden, Colorado, USA
[13]School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada
[14]School of International Affairs, Carleton University, Ottawa, Ontario, Canada
[15]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, Scotland
[16]Laboratoire J.A. Dieudonné, CNRS, Université Côte d'Azur, Nice, France

Acknowledgements The first author acknowledges the support of a Natural Sciences and Engineering Research Council of Canada's Graduate Scholarship–Doctoral scholarship. This work was authored (in part) by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the US Department of Energy (DOE) under contract number: DE-AC36-08GO28308. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. The views expressed in the article do not necessarily represent the views of the DOE or the US Government. The US Government retains and the publisher, by accepting the article for publication, acknowledges that the US Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work or allow others to do so, for US Government purposes.

Contributors BR drafted the protocol, derived the equations in the supplemental material and will conduct and report the research findings. JDE, TK and AS contributed substantially to the conception of the study. JDE, TK and SMulpuru provided clinical/epidemiological context. CLP, DP, MK, RS, JMD and TW contributed to the algorithmic development and implementation. JDE, TK and SMills contributed to the data acquisition and utilisation perspective. PJT and VD contributed to computational/software aspect. MA and VP provided critical insight on the role of modelling in public policy. All listed authors contributed to revision of the protocol for intellectual content and have approved the final version.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD
Abhijit Sarkar http://orcid.org/0000-0002-8427-8901

REFERENCES

1 Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020;20:533–4.
2 Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics proceedings of the Royal Society of London. *series a, containing papers of a mathematical and physical character* 1927;115:700–21.
3 He S, Peng Y, Sun K. SEIR modelling of the COVID-19 and its dynamics. *Nonlinear Dyn* 2020;100:1667–80.
4 Tornatore E, Maria Buccellato S, Vetro P. Stability of a stochastic SIR system. *Physica A* 2005;354:111–26.
5 Zhang Z, Gul R, Zeb A. Global sensitivity analysis of COVID-19 mathematical model. *Alexandria Engineering Journal* 2021;60:565–72.
6 Postnikov EB. Estimation of COVID-19 dynamics "on a back-of-envelope": Does the simplest SIR model provide quantitative parameters and predictions? *Chaos, Solitons & Fractals* 2020;135, :109841.10.1016/j.chaos.2020.109841
7 Diekmann O, Heesterbeek JA, Metz JA. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *J Math Biol* 1990;28:365–82.
8 Moein S, Nickaeen N, Roointan A, *et al*. Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan. *Sci Rep* 2021;11:pp. 1–9.
9 Tuite AR, Fisman DN, Greer AL. Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. *CMAJ* 2020;192:E497–505.
10 Sandhu R, Khalil M, Pettit C, *et al*. Nonlinear sparse Bayesian learning for physics-based models. *J Comput Phys* 2021;426:109728.
11 Sandhu R. *Model comparison and sparse learning of nonlinear physics-based models using Bayesian inference*. Ottawa, ON: PhD Thesis, Carleton University, 2020.
12 Cramer E, Ray EL, Lopez VK, *et al*. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *medRxiv* 2021.
13 Gibson GC, Reich NG, Sheldon D. Real-Time mechanistic Bayesian forecasts of COVID-19 mortality. *medRxiv* 2020. doi:10.1101/2020.12.22.20248736. [Epub ahead of print: 24 Dec 2020].
14 MacKay DJC. Bayesian interpolation. *Neural Comput* 1992;4:415–47.
15 Oden JT. Adaptive multiscale predictive modelling. *Acta Numerica* 2018;27:353–450.
16 Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc Natl Acad Sci U S A* 2016;113:3932–7.
17 Horrocks J. "*Sparse identification of epidemiological models from empirical data,*" *MS Thesis*. University of Waterloo, 2018.
18 Horrocks J, Bauch CT. Algorithmic discovery of dynamic models from infectious disease data. *Sci Rep* 2020;10:1–18.
19 Allenman T, Vergeynst J, Torfs E, *et al*. A deterministic, age-stratified, extended SEIRD model for investigating the effect of non-pharmaceutical interventions on SARS-CoV-2 spread in Belgium.

medRxiv 2020:2020.07.17.20156034.10.1101/2020.07.17.20156034 Now published in Epidemics doi: 10.1016/j.epidem.2021.100505
20 Diekmann O, Heesterbeek JAP, Roberts MG. The construction of next-generation matrices for compartmental epidemic models. *J R Soc Interface* 2010;7:873–85.
21 Kretzschmar ME, Rozhnova G, Bootsma MCJ, *et al*. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *Lancet Public Health* 2020;5:e452–9.
22 Gumel AB, Ruan S, Day T, *et al*. Modelling strategies for controlling SARS outbreaks. *Proc Biol Sci* 2004;271:2223–32.
23 Nalbandian A, Sehgal K, Gupta A, *et al*. Post-Acute COVID-19 syndrome. *Nat Med* 2021;27:601–15.
24 Gousseff M, Penot P, Gallay L, *et al*. Clinical recurrences of COVID-19 symptoms after recovery: viral relapse, reinfection or inflammatory rebound? *J Infect* 2020;81:816–46.
25 Gandon S, Mackinnon M, Nee S, *et al*. Imperfect vaccination: some epidemiological and evolutionary consequences. *Proc Biol Sci* 2003;270:1129–36.
26 Kribs-Zaleta CM, Velasco-Hernández JX. A simple vaccination model with multiple endemic states. *Math Biosci* 2000;164:183–201.
27 Perkins TA, España G. Optimal control of the COVID-19 pandemic with non-pharmaceutical interventions. *Bull Math Biol* 2020;82:1–24.
28 Longini IM, Halloran ME, Nizam A, *et al*. Containing pandemic influenza with antiviral agents. *Am J Epidemiol* 2004;159:623–33.
29 Aguilar JB, Faust JS, Westafer LM, *et al*. Investigating the impact of asymptomatic carriers on COVID-19 transmission. *medRxiv* 2020.
30 Zakary O, Rachik M, Elmouki I. On the analysis of a multi-regions discrete SIR epidemic model: an optimal control approach. *Int J Dyn Control* 2017;5:917–30.
31 Viguerie A, Lorenzo G, Auricchio F, *et al*. Simulating the spread of COVID-19 via a spatially-resolved susceptible–exposed–infected–recovered–deceased (SEIRD) model with heterogeneous diffusion. *Appl Math Lett* 2021;111:106617.
32 Viguerie A, Veneziani A, Lorenzo G, *et al*. Diffusion–reaction compartmental models formulated in a continuum mechanics framework: application to COVID-19, mathematical analysis, and numerical study. *Comput Mech* 2020;66:1131–52.
33 Jha PK, Cao L, Oden JT. Bayesian-based predictions of COVID-19 evolution in Texas using multispecies mixture-theoretic continuum models. *Comput Mech* 2020;66:1055–68.
34 Dolean V, Jolivet P, Nataf F. *An introduction to domain decomposition methods: algorithms, theory and parallel implementation*. Society for Industrial and Applied Mathematics, 2015.
35 Desai A, Khalil M, Pettit C, *et al*. Scalable domain decomposition solvers for stochastic PDEs in high performance computing. *Comput Methods Appl Mech Eng* 2018;335:194–222.
36 Public Health Ontario. COVID-19 Variants of Concern ({VOC} s). Available: https://www.publichealthontario.ca/en/diseases-and-conditions/infectious-diseases/respiratory-diseases/novel-coronavirus/variants [Accessed 14 Mar 2021].
37 Andreasen V, Lin J, Levin SA. The dynamics of cocirculating influenza strains conferring partial cross-immunity. *J Math Biol* 1997;35:825–42.
38 Li Q, Wu J, Nie J, *et al*. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* 2020;182:e9:1284–94.
39 Garcia-Beltran WF, Lam EC, St Denis K, *et al*. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* 2021;184:2372–83.
40 Government of Ontario. COVID-19 vaccines for Ontario. Available: https://covid-19.ontario.ca/covid-19-vaccines-ontario [Accessed 14 Mar 2021].
41 Oh CK, Beck JL, Yamada M. Bayesian learning using automatic relevance determination prior with an application to earthquake early warning. *Journal of Engineering Mechanics* 2008;134:1013–20.
42 Sandhu R, Pettit C, Khalil M, *et al*. Bayesian model selection using automatic relevance determination for nonlinear dynamical systems. *Comput Methods Appl Mech Eng* 2017;320:237–60.
43 Sandhu R, Khalil M, Sarkar A, *et al*. Bayesian model selection for nonlinear aeroelastic systems using wind-tunnel data. *Comput Methods Appl Mech Eng* 2014;282:161–83.
44 Bisaillon P, Sandhu R, Khalil M, *et al*. Bayesian parameter estimation and model selection for strongly nonlinear dynamical systems. *Nonlinear Dyn* 2015;82:1061–80.
45 Khalil M, Sarkar A, Adhikari S, *et al*. The estimation of time-invariant parameters of noisy nonlinear oscillatory systems. *J Sound Vib* 2015;344:81–100.
46 Sandhu R, Poirel D, Pettit C, *et al*. Bayesian inference of nonlinear unsteady aerodynamics from aeroelastic limit cycle oscillations. *J Comput Phys* 2016;316:534–57.
47 Robinson B, Sandhu R, Edwards JD, *et al*. A sparse Bayesian model selection algorithm for forecasting the transmission of COVID-19. *American Thoracic Society 2021 International Conference, Virtual Event (originally San Diego)*, 2021.
48 ICES. ICES (formerly the Institute of clinical Evaluative sciences). Available: www.ices.on.ca/Data-and-Privacy/ [Accessed 07 Jan 2021].
49 Government of Ontario. Novel coronavirus data Catalogue, 2019. Available: https://data.ontario.ca/en/group/2019-novel-coronavirus [Accessed 07 Jan 2021].
50 Khalil M, Sarkar A, Adhikari S. Nonlinear filters for chaotic oscillatory systems. *Nonlinear Dyn* 2009;55:113–37.
51 Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2009;22:1345–59.

Supplemental Material:

Comprehensive compartmental model and calibration algorithm for the study of clinical implications of the population-level spread of COVID-19: a study protocol

Brandon Robinson[1], Jodi D. Edwards[2,3], Tetyana Kendzerska[3,4,5], Chris L. Pettit[6], Dominique Poirel[7], John M. Daly[8], Mehdi Ammi[9], Mohammad Khalil[10], Peter J. Taillon[11], Rimple Sandhu[12], Shirley Mills[13], Sunita Mulpuru[4,5], Thomas Walker[1], Valerie Percival[14], Victorita Dolean[15,16], and Abhijit Sarkar *[1]

[1]Department of Civil and Environmental Engineering, Carleton University, Ottawa, ON, Canada K1S 5B6
[2]School of Epidemiology and Public Health, University of Ottawa and University of Ottawa Heart Institute, Ottawa, ON, Canada K1Y 4W7
[3]ICES, Ottawa, ON, Canada K1Y 4E9
[4]The Ottawa Hospital Research Institute, Ottawa, ON, Canada
[5]Department of Medicine, Faculty of Medicine, Division of Respirology, University of Ottawa, Ottawa, ON, Canada K1H 8L6
[6]US Naval Academy, Aerospace Engineering Department, Annapolis, MD 21402, United States
[7]Royal Military College of Canada, Department of Mechanical and Aerospace Engineering, Kingston, ON, Canada K7K 7B4
[8]Independent Control Systems Engineer, Ottawa, ON, Canada K1S 4H6
[9]School of Public Policy and Administration, Carleton University, Ottawa, ON, Canada K1S 5B6
[10]Sandia National Laboratories†, Livermore, CA 94550, United States
[11]Schaffen Research Inc. Ottawa, ON, Canada K1H 7S7
[12]National Renewable Energy Laboratory, Golden, CO 80401, United States
[13]School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada K1S 5B6
[14]School of International Affairs, Carleton University, Ottawa, ON, Canada K1S 5B6
[15]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, Scotland G1 1XQ
[16]Laboratoire J.A. Dieudonné, CNRS, Université Côte d'Azur, Nice, France 06108

The 22-compartment model described in the paper consists of a number of extensions to the detailed 16-compartment model by Tuite et al. [1], including the addition of multiple compartments, model stratifications, and model discrepancy. This supplemental material contains the equations of motion for the model described in "Study protocol: a comprehensive compartmental model and calibration algorithm for the study of clinical implications of the population-level spread of COVID-19." The mathematical form of the model is described, along with the inclusion of model error to extend the deterministic model to a stochastic compartmental model. Subsequently, the pertinent details of the proposed nonlinear sparse Bayesian learning (NSBL) algorithm are provided. The necessary modifications to the ordinary differential equation model capturing only the temporal variation of COVID-19 case

*Corresponding author; email: abhijit.sarkar@carleton.ca

†Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

counts to include the spatio-temporal variation of disease dynamics through a partial differential equation model are discussed.

A Modelling population-level spread through ordinary differential equations

The model by Tuite et al. [1] is stratified by age into 16 age groups with equal widths of five years, and includes a second stratification indicating whether or not an individual has a pre-existing health condition. In this case, each compartment is indexed by $i = (1,\ldots,16)$ for age group, and $j = (1,2)$ for comorbidity. The proposed model considers the stratification more generally; each compartment in Figure 1 would be indexed by $i = (1,\ldots,N_i)$ for age group, $j = (1,\ldots,N_j)$ for comorbidity ($j = 1$ for no comorbidity, and $j = 2,\ldots,N_j$ for specific pre-existing health conditions). Furthermore, in the equations below, an additional index $k = (1,\ldots,N_k)$ is introduced that has a distinct index for unvaccinated individuals and $N_k - 1$ indices different vaccines (see discussion in Section 2.2.2). The total number of states in the model is $(22 \times N_i \times N_j \times N_k)$. Clearly, as the number of phenomena modelled by stratification increases, the number of states in the model increases exponentially.

Note that the other phenomena described in Sections 2.2.1 and 2.2.2 may be modelled through further stratification (by introducing additional indices) with minimal modification to the system of equations below. The equations are shown for the three above-mentioned stratifications as the model includes cross-terms that allow individuals to move from one index to others in time (i.e., by aging out of their current age index, i, by developing long-term health complications resulting in a change in comorbidity index, j, and by becoming vaccinated, or having the period of vaccine effectiveness elapse resulting in a change in vaccination index, k).

The system of equations (Eqs. (A.1 - A.22), derived by extending the equations in [1]) is a direct mathematical representation of the flowchart in Figure 1, hence, each state (summarized in Table 1) relates to one of the compartments of that flowchart. Likewise, the coefficients (rate parameters or probabilities, summarized in Table 2) are represented by the arrows in the flowchart, as these parameters dictate the flow of individuals between adjacent compartments. Each equation represents the net flow of the population into the respective compartment. The indices of the parameters are omitted from Eqs. (A.1 - A.22) for simplicity. Note the model discrepancy terms in red for each model equation. These consist of a strength parameter, $q_S,\ldots,q_D$, and correlated noise processes, $\xi_S,\ldots,\xi_D$, thereby extending the model equations to a stochastic form. Note that the model errors must be statistically correlated in order to maintain the conservation of the population in the mechanistic model [2]. The model error strength parameters will be estimated in the inference through the NSBL algorithm.

Susceptible:

$$\frac{dS^{ijk}}{dt} = -\lambda^{ijk}S^{ijk} + \sum_{\psi=1}^{N_k}(-\gamma_V^{\psi}S^{ijk} + \gamma_T^{\psi}\Psi^{\psi k}V^{ij\psi}) + \gamma_R(R1^{ijk} + \sum_{\upsilon=1}^{N_j}\Upsilon^{\upsilon j}R2^{i\upsilon k}) - \mu S^{ijk} + \mu S^{(i-1)jk} + q_S\xi_S \quad \text{(A.1)}$$

Vaccinated:

$$\frac{dV^{ijk}}{dt} = -(1-r_V)\lambda^{ijk}V^{ijk} + \gamma_V S^{ijk} - \gamma_T V^{ijk} - \mu V^{ijk} + \mu V^{(i-1)jk} + q_V\xi_V \quad \text{(A.2)}$$

Exposed:

$$\frac{dE^{ijk}}{dt} = (1-\delta_S)\lambda^{ijk}S^{ijk} + (1-\delta_V)(1-r_V)\lambda^{ijk}V^{ijk} - \gamma_E E^{ijk} - \mu E^{ijk} + \mu E^{(i-1)jk} + q_E\xi_E \quad \text{(A.3)}$$

Exposed, isolating:

$$\frac{dQ^{ijk}}{dt} = \delta_S\lambda^{ijk}S^{ijk} + \delta_V(1-r_V)\lambda^{ijk}V^{ijk} - \gamma_E Q^{ijk} - \mu Q^{ijk} + \mu Q^{(i-1)jk} + q_Q\xi_Q \quad \text{(A.4)}$$

Infectious, presymtomatic:

$$\frac{dA^{ijk}}{dt} = \gamma_E E^{ijk} - \gamma_P A^{ijk} - \mu A^{ijk} + \mu A^{(i-1)jk} + q_A \xi_A \tag{A.5}$$

Infectious, pre-symptomatic, isolating:

$$\frac{dW^{ijk}}{dt} = \gamma_E Q^{ijk} - \gamma_P W^{ijk} - \mu W^{ijk} + \mu W^{(i-1)jk} + q_W \xi_W \tag{A.6}$$

Infectous, asymptomatic:

$$\frac{dF^{ijk}}{dt} = \sigma_A \gamma_P A^{ijk} - \gamma_A F^{ijk} - \gamma_{DA} F^{ijk} - \mu F^{ijk} + \mu F^{(i-1)jk} + q_F \xi_F \tag{A.7}$$

Infectious, mild-to-moderate symptoms:

$$\frac{dB^{ijk}}{dt} = (1-\sigma_A)(1-\sigma_S)\gamma_P A^{ijk} - \gamma_M B^{ijk} - \gamma_{DM} B^{ijk} - \mu B^{ijk} + \mu B^{(i-1)jk} + q_B \xi_B \tag{A.8}$$

Infectious, severe symptoms:

$$\frac{dC^{ijk}}{dt} = (1-\sigma_A)\sigma_S \gamma_p A^{ijk} - \gamma_{S1} C^{ijk} - \mu C^{ijk} + \mu C^{(i-1)jk} + q_C \xi_C \tag{A.9}$$

Infectious, asymptomatic, isolating:

$$\frac{dX^{ijk}}{dt} = \sigma_A \gamma_P W^{ijk} - \gamma_A X^{ijk} - \gamma_{DA} X^{ijk} - \mu X^{ijk} + \mu X^{(i-1)jk} + q_X \xi_X \tag{A.10}$$

Infectious, mild-to-moderate symptoms, isolating:

$$\frac{dY^{ijk}}{dt} = (1-\sigma_A)(1-\sigma_S)\gamma_P W^{ijk} - \gamma_M Y^{ijk} - \gamma_{DM} Y^{ijk} - \mu Y^{ijk} + \mu Y^{(i-1)jk} + q_{Y2} \xi_Y \tag{A.11}$$

Infectious, severe symptoms, isolating:

$$\frac{dZ^{ijk}}{dt} = (1-\sigma_A)\sigma_S \gamma_P W^{ijk} - \gamma_{S1} Z^{ijk} - \mu Z^{ijk} + \mu Z^{(i-1)jk} + q_Z \xi_Z \tag{A.12}$$

Infectious, isolating after testing positive:

$$\frac{dG^{ijk}}{dt} = \gamma_{DA}(F^{ijk} + X^{ijk}) + \gamma_{DM}(B^{ijk} + Y^{ijk}) - \gamma_I G^{ijk} - \mu G^{ijk} + \mu G^{(i-1)jk} + q_G \xi_G \tag{A.13}$$

Inadequate access to health care resources:

$$\frac{dN^{ijk}}{dt} = (1-\sigma_H)\gamma_{S1}(C^{ijk} + Z^{ijk}) - \gamma_{S2} N^{ijk} - \mu N^{ijk} + \mu N^{(i-1)jk} + q_N \xi_N \tag{A.14}$$

Hospital:

$$\frac{dH^{ijk}}{dt} = \sigma_H(1-\sigma_C)\gamma_{S1}(C^{ijk} + Z^{ijk}) - \pi_H H^{ijk} - \mu H^{ijk} + \mu H^{(i-1)jk} + q_H \xi_H \tag{A.15}$$

Pre-ICU:

$$\frac{dH1^{ijk}}{dt} = \sigma_H \sigma_C \gamma_{S1}(C^{ijk} + Z^{ijk}) - \pi_A H1^{ijk} - \mu H1^{ijk} + \mu H1^{(i-1)jk} + q_{H1} \xi_{H1} \tag{A.16}$$

ICU:

$$\frac{dI^{ijk}}{dt} = \pi_A H1^{ijk} - \pi_B I^{ijk} - \mu I^{ijk} + \mu I^{(i-1)jk} + q_I \xi_I \tag{A.17}$$

Post-ICU:

$$\frac{dH2^{ijk}}{dt} = (1-\kappa_I)\pi_B I^{ijk} - \pi_C H2^{ijk} - \mu H2^{ijk} + \mu H2^{(i-1)jk} + q_{H2}\xi_{H2} \tag{A.18}$$

Recovered:

$$\begin{aligned}\frac{dR1^{ijk}}{dt} = (1-\phi_C)\Big((1-\phi_M)\left(\gamma_I G^{ijk} + \gamma_A(F^{ijk}+X^{ijk}) + \gamma_M(B^{ijk}+Y^{ijk})\right) \\ + (1-\phi_S)\left((1-\kappa_N)\gamma_{S2}N^{ijk} + (1-\kappa_H)\pi_H H^{ijk} + \pi_C H2^{ijk}\right)\Big) \\ +(1-\phi_P)(1-\kappa_P)\gamma_C P^{ijk} - \gamma_R R1^{ijk} - \mu R1^{ijk} + \mu R1^{(i-1)jk} + q_{R1}\xi_{R1}\end{aligned} \tag{A.19}$$

Recovered with long-term health complications:

$$\begin{aligned}\frac{dR2^{ijk}}{dt} = (1-\phi_C)\Big(\phi_M\left(\gamma_I G^{ijk} + \gamma_A(F^{ijk}+X^{ijk}) + \gamma_M(B^{ijk}+Y^{ijk})\right) \\ + \phi_S\left((1-\kappa_N)\gamma_{S2}N^{ijk} + (1-\kappa_H)\pi_H H^{ijk} + \pi_C H2^{ijk}\right)\Big) \\ +\phi_P(1-\kappa_P)\gamma_C P^{ijk} - \gamma_R R2^{ijk} - \gamma_L R2^{ijk} - \mu R2^{ijk} + \mu R2^{(i-1)jk} + q_{R2}\xi_{R2}\end{aligned} \tag{A.20}$$

Post-acute COVID-19:

$$\begin{aligned}\frac{dP^{ijk}}{dt} = \phi_C\left(\gamma_I G^{ijk} + \gamma_A(F^{ijk}+X^{ijk}) + \gamma_M(B^{ijk}+Y^{ijk}) + (1-\kappa_N)\gamma_{S2}N^{ijk} + (1-\kappa_H)\pi_H H^{ijk} + \pi_C H2^{ijk}\right) \\ -\gamma_C P^{ijk} - \mu P^{ijk} + \mu P^{(i-1)jk} + q_P\xi_P\end{aligned} \tag{A.21}$$

Death:

$$\frac{dD^{ijk}}{dt} = \kappa_H\pi_H H^{ijk} + \kappa_I\pi_B I^{ijk} + \kappa_N\gamma_{S2}N^{ijk} + \gamma_L R2^{ijk} + \kappa_P\gamma_C P^{ijk} + q_D\xi_D \tag{A.22}$$

Eq. (A.1) accounts for the rate at which people flow into and out of the susceptible (S) compartment. Individuals leave the compartment as they become infected as a result of the interaction with infectious individuals through the force of infection term, λ^{ijk}. This term introduces nonlinearity to the model through the multiplicative nonlinear coupling between susceptible and infectious compartments (B, C, F, X, Y, Z), as in Eq. A.23.

$$\begin{aligned}\lambda^{ijk} = \beta\sum_{l=1}^{N_i}\sum_{m=1}^{N_j}\sum_{n=1}^{N_k}\frac{c^{ijklmn}}{N^{lmn}}\Big(rr_C\left(rr_A\left(A^{lmn}+F^{lmn}\right)+B^{lmn}+C^{lmn}\right) \\ + rr_Q\left(G^{lmn}+W^{lmn}+X^{lmn}+Y^{lmn}+Z^{lmn}+N^{lmn}\right) + rr_H\left(H^{lmn}+H1^{lmn}+I^{lmn}+H2^{lmn}\right)\Big)\end{aligned} \tag{A.23}$$

where c^{ijklmn} is a tensor that quantifies the average number of daily interactions between individuals indexed by $\{i,j,k\}$ and those indexed by $\{l,m,n\}$. The terms rr_C and rr_Q are factors accounting for reductions in daily contacts due to social distancing and quarantining policies, respectively, and hence may vary in time as restrictions are imposed and removed as in Tuite et al.[1]. The force of infection term differs from [1], as follows: (i) the force of infection λ may explicitly vary in time to capture the changing reproduction number due to temporal variations in β (for example due to seasonal trends in transmission, human mobility patterns, and the implementation of control measures) whereas in [1], this term was a static value, multiplied by a random process that was meant to induce volatility in the transmission rate in time [3] [4]), (ii) with the inclusion of asymptomatic individuals, a

reduction factor, rr_A, is introduced to allow potential reductions or increases in transmission rates for asymptomatic carriers, and (iii) the above allows for individuals within the hospital track (compartments N, H, $H1$, I, $H2$) to infect susceptible individuals subject to a reduction factor rr_H to account for heightened precautions taken in hospital settings, but allowing for outbreaks to occur. The expression from Tuite et al. [1] can be recovered by setting those factors equal to one and zero, respectively.

The second term in Eq. (A.1) accounts for the individuals who leave the susceptible compartment as they are vaccinated and who return once the effective duration of their vaccine has elapsed. The third term quantifies people who re-enter the susceptible compartment after a period of temporary immunity from having recovered from the disease. The final terms account for individuals who remain susceptible, but are aging in and out of the current age index, i. To ensure that the total population is conserved (i.e., the sum of all 22 compartments across all indices is equal to unity for all time instances), individuals are introduced to the susceptible compartment (S) of age group $i = 1$ at a rate equal to the aging out of members from all compartments of age group $i = N_i$ as in [1]. The interpretation of all other equations follows the same reasoning with the flow in and out of the respective compartments having a one-to-one correspondence with the directed arrows in Figure 1 in the manuscript.

In a discrete state-space form, the model equation can be represented as in Eq. (A.24),

$$\mathbf{u}_{t+1} = \mathbf{g}_t(\mathbf{u}_t, \mathbf{q}_t; \boldsymbol{\phi}), \tag{A.24}$$

where $\mathbf{g}$ is the model operator and $\mathbf{u}$ is the state vector containing all of the above-mentioned $22 \times N_i \times N_j \times N_k$ model states. $\mathbf{q}$ is a vector consisting of the model error associated with each model equation, which corresponds to the red highlighted terms in Eqs.(A.1 - A.22). $\boldsymbol{\phi}$ is a vector of uncertain time-invariant parameters, meaning the subset of parameters from Table 2 that are to be estimated.

Similarly, the model must be accompanied by data; the measurement equation is given in Eq. (A.25),

$$\mathbf{y} = \mathbf{h}_k(\mathbf{u}_{t(k)}, \boldsymbol{\varepsilon}_k; \boldsymbol{\phi}), \tag{A.25}$$

which states that measurements, $\mathbf{y}$, of the state, $\mathbf{u}$, are obtained according to the model operator, $\mathbf{h}$, at time, $t(k)$, and are corrupted by noise, $\boldsymbol{\varepsilon}$. Note that in our case, the dimension of $\mathbf{y}$ will not be equal to that of $\mathbf{u}$, as a number of the states are not directly observable. The above model and measurement equations are necessary for the concurrent estimation of model error, state, and parameters [5, 6, 7, 8].

The initial conditions required for simulations consist of a single snapshot of all model states at a given instant because the model is a system of coupled first-order ordinary differential equations. These initial conditions will be subject to uncertainty, and thus may need to be . As the forcing of the dynamics are driven by the interaction of infectious individuals with susceptible individuals, a nonzero initial condition must be assigned in one of the infectious or exposed compartments (as it is intermediate to susceptible and infectious). This initial condition would model the immigration of a small number of infectious individuals into the region of interest and initiate community transmission through forward iterations of the state-space model.

The increased number of states and the inclusion of time-varying parameters and stochastic source terms has significantly increased the complexity of the model. Thus, it would be beneficial to study the stability of the model itself [9] prior to investigating the inverse problem. Understanding the influence of parametric uncertainty on various quantities of interest by conducting global sensitivity analysis (GSA) [10] would also be beneficial to facilitate the reduction of dimension of the uncertain parameter space.

Table 1: Model states/compartments (indices omitted for brevity)

Symbol	Definition
S	Susceptible
V	Vaccinated
E	Exposed
Q	Exposed, isolating
A	Infectious, pre-symptomatic
W	Infectious, pre-symptomatic, isolating
F	Infectious, asymptomatic
B	Infectious, mild-to-moderate symptomatic (i.e., symptoms not requiring hospitalization)
C	Infectious, severe symptomatic (i.e., symptoms requiring hospitalization)
X	Infectious, asymptomatic, isolating
Y	Infectious, mild-to-moderate symptomatic, isolating
Z	Infectious, severe symptomatic, isolating
G	Infectious, mild-to-moderate symptomatic, isolating but not previously in isolation
N	No access to hospital care
H	Hospitalized, never to be admitted to the intensive care unit (ICU)
$H1$	Hospitalized, to be admitted to the ICU
I	Hospitalized, in the ICU
$H2$	Hospitalized, after being discharged from the ICU
$R1$	Recovered, without long-term health complications
$R2$	Recovered, with long-term health complications
P	Post-acute COVID-19
D	Death

B Nonlinear sparse Bayesian learning algorithm at a glance

The nonlinear sparse Bayesian learning algorithm seeks to obtain parameter posterior distributions in a hierarchical Bayesian setting. The parameter posterior distribution is written as

$$\mathrm{p}(\boldsymbol{\phi}|\mathscr{D},\boldsymbol{\alpha}) = \frac{\mathrm{p}(\mathscr{D}|\boldsymbol{\phi})\mathrm{p}(\boldsymbol{\phi}|\boldsymbol{\alpha})}{\mathrm{p}(\mathscr{D}|\boldsymbol{\alpha})}, \tag{B.1}$$

where $\mathrm{p}(\mathscr{D}|\boldsymbol{\phi})$ is the likelihood function, $\mathrm{p}(\boldsymbol{\phi}|\boldsymbol{\alpha})$ parameter prior distribution, and $\mathrm{p}(\mathscr{D}|\boldsymbol{\alpha})$ is the model evidence. As in Eqs. (A.24 and A.25), $\boldsymbol{\phi}$ is the set of uncertain model parameters, $\mathscr{D}$ represents the data, which is a realization of $\mathbf{y}$ from Eq. (A.25), and $\boldsymbol{\alpha}$ is a hyperparameter, which we will examine next.

This algorithm leverages the concept of automatic relevance determination (ARD) to induce sparsity among the parameters. The set of parameters is decomposed into two distinct subsets, parameters that are known to be relevant to the observed dynamics (denoted $\boldsymbol{\phi}_{-\alpha}$), and parameters whose relevance is questionable (denoted $\boldsymbol{\phi}_{\alpha}$). The parameters that are known to be relevant are assigned prior distributions as in a standard Bayesian approach, whereas questionable parameters are assigned ARD priors which are zero mean normal distributions with precision controlled by the hyperparameter $\boldsymbol{\alpha}$. The ARD priors are independent; hence, the parameter prior can be written as

$$\mathrm{p}(\boldsymbol{\phi}|\boldsymbol{\alpha}) = \mathrm{p}(\boldsymbol{\phi}_{-\alpha})\mathrm{p}(\boldsymbol{\phi}_{\alpha}|\boldsymbol{\alpha}) = \mathrm{p}(\boldsymbol{\phi}_{-\alpha})\mathscr{N}(\boldsymbol{\phi}_{\alpha}|\mathbf{0},\mathrm{diag}(\boldsymbol{\alpha})^{-1}). \tag{B.2}$$

The algorithm optimally selects the values of the elements of $\boldsymbol{\alpha}$, by selecting the values based on the available data, which maximize the model evidence from Eq. (B.1). A parameter whose precision is determined to be large, results in a highly peaked Gaussian distribution centered at zero. As the precision tends to infinity, the prior pdf tends to a Dirac delta function at zero, effectively fixing the parameter to a value of zero and removing it from the analysis. This results in a data-optimal sparse model. The computational efficiency of this algorithm is achieved by substituting the prior pdf from Eq. (B.2) into Eq. (B.1), and creating a Gaussian mixture model of the known

Table 2: Model parameters with definition and reference (indices omitted for brevity)

Symbol	Definition
μ	Rate that individuals pass from age group i to $i+1$ (1/width of age group)
r_V	Vaccine effectiveness (1 indicates 100% immunity, 0 indicates no immunity)
δ_S	Probability that a susceptible individual exposed to the virus will self-isolate (without prior testing)
δ_V	Probability that a vaccinated individual exposed to the virus will self-isolate (without prior testing)
γ_A	1/the average duration of the infectious period for asymptomatic individuals
γ_C	1/the average duration of sub-acute COVID-19
γ_{DA}	Rate of detection among asymptomatic cases
γ_{DM}	Rate of detection among mild-to-moderate cases
γ_E	1/the average incubation period
γ_I	1/the average duration of self-isolation
γ_L	Rate of deaths due to long-term health complications
γ_M	1/the average duration of the infectious period for individuals with mild-to-moderate symptoms
γ_P	1/the average duration of the pre-symptomatic infectious period
γ_R	1/the average effective duration of temporary immunity from having recovered from the virus
γ_{S1}	1/the average duration of severe symptoms before seeking hospitalization
γ_{S2}	1/the average remaining duration of symptomatic period for individuals with severe symptoms
γ_T	1/the average effective duration of temporary immunity from vaccination
γ_V	Rate of vaccination
σ_A	Probability that an infectious individual is asymptomatic
σ_C	Probability that a hospitalized case will be admitted to the ICU
σ_H	Probability that an individual has access to hospital care
σ_S	Probability that a case displaying symptoms will require hospitalization
π_A	1/the average time in hospital prior to ICU
π_B	1/the average time in ICU
π_C	1/the average time in hospital following ICU
π_H	1/the average duration hospitalization (non-ICU track)
ϕ_A	Probability of acute COVID
ϕ_M	Probability of long-term complications for asymptomatic, mild-to-moderate cases
ϕ_S	Probability of long-term complications for severe cases
ϕ_P	Probability of long-term complications for post-acute COVID-19 cases
κ_H	Probability of death among hospital cases
κ_I	Probability of death among ICU cases
κ_N	Probability of death among cases without access to hospital care
κ_P	Probability of death among post-acute COVID-19 cases
Ψ	Parameter controlling the transition of vaccination indices
Υ	Parameter controlling the transition of comorbidity indices

prior $\mathrm{p}(\boldsymbol{\phi}_{-\alpha})$ times the likelihood function $\mathrm{p}(\mathscr{D}|\boldsymbol{\phi})$. This allows for the semianalytical computation of many Bayesian entities and quantities of interest that expedite the optimization of the hyperparameters as summarized in [11]. A detailed derivation of the algorithm is available in [12].

The NSBL algorithm (and more generally, ARD) has some apparent similarities with the aforementioned GSA, particularly as both methods can be used for model reduction. However, they are two distinct concepts with fundamentally different approaches to model reduction which may be complimentary to one another. Some key differentiating features of the two methods are as follows: (i) GSA considers a single output quantity of interest at a time, whereas NSBL may involve assimilating observations of multiple quantities of interest which are functions of the model states, (ii) in GSA, the input parameters' uncertainty is predefined (akin to a prior distribution), whereas NSBL considers both prior knowledge and available data, and (iii) for model reduction, GSA does not consider model complexity, whereas NSBL explicitly handles model complexity through evidence optimization.

C Modelling the spatio-temporal spread through partial differential equations

In order to extend Eqs. (A.1 – A.22) to model the spatio-temporal spread of COVID-19, the model are to be interpreted as population densities. The following nine compartments (indicated by their associated symbol from Table 1) would require the addition of a diffusion term to account for the movement of individuals in those compartments in space: S, V, E, A, F, C, B, P, $R1$, $R2$. This implies that the remaining 12 model compartments do not move in space, which makes use of the fact that compartments Q, W, X, Y, Z, and G on the quarantine track and N, H, $H1$, I, $H2$ on the hospital track will remain fixed in place until they proceed to one of the recovered compartments, or to the death compartment, D, where they will remain fixed permanently.

To illustrate how the form of the ordinary differential equations can be extended to partial differential equations, we take Eq. (A.3), which models infectious, pre-symptomatic individuals as an example. In Eq. (C.1), we see that the form of the equation is largely unchanged, a diffusion term is added (in blue text), and the state variables (uppercase letters) are now population densities in space and time. The model error term, ξ_E, may now also be modelled as a spatio-temporal white or coloured noise process,

$$\frac{\partial E^{ijk}}{\partial t} = (1-\delta_s)\lambda^{ijk}S^{ijk} + (1-\delta_v)(1-r_v)\lambda^{ijk}V^{ijk} - \varepsilon E^{ijk} - \mu E^{ijk} + \mu E^{(i-1)jk} + \nabla\cdot(\nu_E \nabla E) + q_E\xi_E \tag{C.1}$$

where ν_E is a diffusion coefficient for compartment E; one such parameter exists per model state. For more details on partial differential equation-based compartmental models, see [13, 14, 15], which describe spatio-temporal compartmental models for COVID-19. For details on the implementation of similar large-scale problems, see [16, 17].

References

[1] A.R. Tuite, D.N. Fisman, and A.L. Greer. Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. CMAJ, 192(19):E497–E505, 2020.

[2] L.J.S. Allen. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. Infectious Disease Modelling, 2(2):128–142, 2017.

[3] A. Camacho, A. Kucharski, Y. Aki-Sawyerr, M.A. White, S. Flasche, M. Baguelin, T. Pollington, J.R. Carney, R. Glover, E. Smout, et al. Temporal changes in Ebola transmission in Sierra Leone and implications for control requirements: a real-time modelling study. PLoS currents, 7, 2015.

[4] J.O. Lloyd-Smith, S.J. Schreiber, P.E. Kopp, and W.M. Getz. Superspreading and the effect of individual variation on disease emergence. Nature, 438(7066):355–359, 2005.

[5] M. Khalil, A. Sarkar, and S. Adhikari. Nonlinear filters for chaotic oscillatory systems. Nonlinear Dynamics, 55(1):113–137, 2009.

[6] P. Bisaillon, R. Sandhu, M. Khalil, C. Pettit, D. Poirel, and A. Sarkar. Bayesian parameter estimation and model selection for strongly nonlinear dynamical systems. Nonlinear Dynamics, 82(3):1061–1080, 2015.

[7] M. Khalil, A. Sarkar, S. Adhikari, and D. Poirel. The estimation of time-invariant parameters of noisy nonlinear oscillatory systems. Journal of Sound and Vibration, 344:81–100, 2015.

[8] R. Sandhu, D. Poirel, C. Pettit, M. Khalil, and A. Sarkar. Bayesian inference of nonlinear unsteady aerodynamics from aeroelastic limit cycle oscillations. Journal of Computational Physics, 316:534–557, 2016.

[9] E. Tornatore, S.M. Buccellato, and P. Vetro. Stability of a stochastic SIR system. Physica A: Statistical Mechanics and its Applications, 354:111–126, 2005.

[10] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. Global sensitivity analysis: the primer. John Wiley & Sons, 2008.

[11] R. Sandhu, M. Khalil, C. Pettit, D. Poirel, and A. Sarkar. Nonlinear sparse bayesian learning for physics-based models. Journal of Computational Physics, 426:109728, 2021.

[12] R. Sandhu. Model Comparison and Sparse Learning of Nonlinear Physics-Based Models Using Bayesian Inference. PhD thesis, Carleton University, 2020.

[13] A. Viguerie, G. Lorenzo, F. Auricchio, D. Baroli, T.J.R. Hughes, A. Patton, A. Reali, T.E. Yankeelov, and A. Veneziani. Simulating the spread of COVID-19 via a spatially-resolved susceptible–exposed–infected–recovered–deceased (SEIRD) model with heterogeneous diffusion. Applied Mathematics Letters, 111:106617, 2021.

[14] A. Viguerie, A. Veneziani, G. Lorenzo, D. Baroli, N. Aretz-Nellesen, A. Patton, T.E. Yankeelov, A. Reali, T.J.R. Hughes, and F. Auricchio. Diffusion–reaction compartmental models formulated in a continuum mechanics framework: application to COVID-19, mathematical analysis, and numerical study. Computational Mechanics, 66(5):1131–1152, 2020.

[15] P.K. Jha, L. Cao, and J.T. Oden. Bayesian-based predictions of COVID-19 evolution in Texas using multispecies mixture-theoretic continuum models. Computational Mechanics, 66(5):1055–1068, 2020.

[16] V. Dolean, P. Jolivet, and F. Nataf. An Introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation. SIAM, 2015.

[17] A. Desai, M. Khalil, C. Pettit, D. Poirel, and A. Sarkar. Scalable domain decomposition solvers for stochastic PDEs in high performance computing. Computer Methods in Applied Mechanics and Engineering, 335:194–222, 2018.