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.

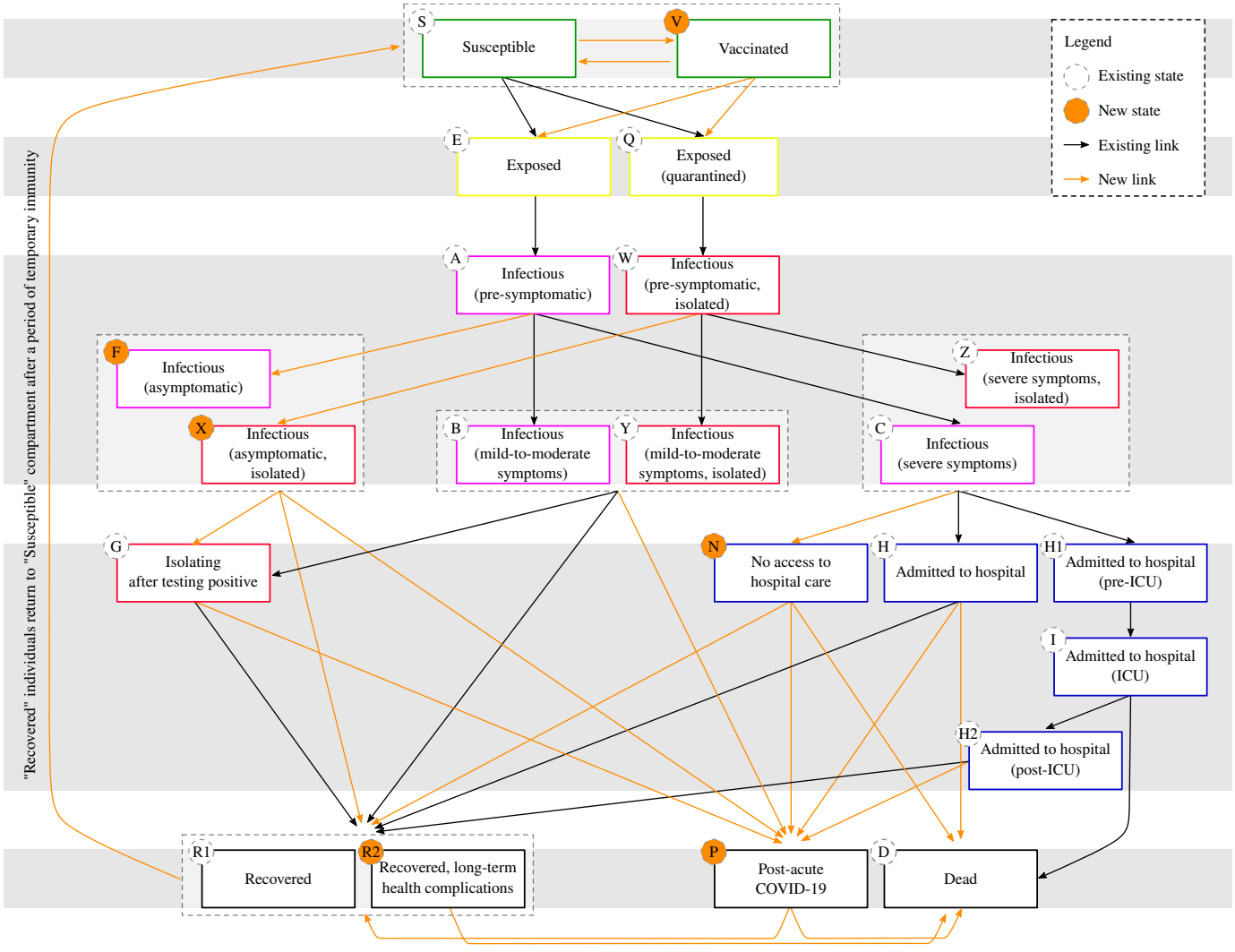


**Figure 1** Flowchart of the proposed 22-compartment model, highlighting the extensions to the 16-compartment model by Tuite *et al*.[9]

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.