Translating Data in a Pandemic 1

An evaluation of prospective COVID-19 modelling studies in the USA: from data to science translation

Kristen Nixon, Sonia Jindal, Felix Parker, Nicholas G Reich, Kimia Ghobadi, Elizabeth C Lee, Shaun Truelove, Lauren Gardner

Infectious disease modelling can serve as a powerful tool for situational awareness and decision support for policy makers. However, COVID-19 modelling efforts faced many challenges, from poor data quality to changing policy and human behaviour. To extract practical insight from the large body of COVID-19 modelling literature available, we provide a narrative review with a systematic approach that quantitatively assessed prospective, data-driven modelling studies of COVID-19 in the USA. We analysed 136 papers, and focused on the aspects of models that are essential for decision makers. We have documented the forecasting window, methodology, prediction target, datasets used, and geographical resolution for each study. We also found that a large fraction of papers did not evaluate performance (25%), express uncertainty (50%), or state limitations (36%). To remedy some of these identified gaps, we recommend the adoption of the EPIFORGE 2020 model reporting guidelines and creating an information-sharing system that is suitable for fast-paced infectious disease outbreak science.

Lancet Digit Health 2022; 4: e738–47

This is the first in a **Series** of two papers about translating data in a pandemic. All papers in the Series are available at www.thelancet.com/series/translating-data-in-a-pandemic

Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD, USA (K Nixon BS, S Jindal BS, F Parker BS, K Ghobadi PhD, L Gardner PhD); **School of Public Health and Health Sciences, University of Massachusetts Amherst, Amherst, MA, USA** (Prof N G Reich PhD); **Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA** (E C Lee PhD, S Truelove PhD)

Correspondence to:
Dr Lauren Gardner, Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
l.gardner@jhu.edu

Introduction

The COVID-19 pandemic has become an unprecedented public health crisis in its prolonged impact on health and its disruption to economic and social life, with more than 6·5 million reported deaths globally as of Sept 7, 2022.[1] To aid planning and response efforts during a pandemic, mathematical modelling of current and future trends of infectious disease outbreaks has historically served as a valuable tool. Nowcasting and forecasting models can improve situational awareness of the current and near future states of disease spread, whereas long-term projections and scenario modelling can shed light on outcomes that might result from a set of assumptions. Insights from modelling can educate individuals on how to mitigate their own risks, while also providing support for decision making for policy makers seeking to minimise harm to an entire population.

These insights are historically provided though peer-reviewed published literature, which can serve as an invaluable tool for communicating state of the art science. During the COVID-19 pandemic, an extremely large volume of research articles have been published: about 125 000 within 10 months of the first confirmed case, approximately 30 000 of which are preprints.[2] In this noisy publication landscape, journals prioritised the quick sharing of COVID-19 information, but there is a trade-off between speeding up peer review and ensuring high-quality research.[3] Preprints also had an important role in disseminating COVID-19 research. Preprints were often covered in the media, had large audiences on social media platforms such as Twitter, and in some cases were misunderstood in consequential ways.[2] For COVID-19 modelling specifically, the use of models for informing response efforts was criticised largely because of a few particularly erroneous projections at the start of the outbreak and poor communication on what insight models can and cannot provide.[4–6]

Literature reviews that attempt to synthesise COVID-19 modelling studies, published up until the time of this Series paper, form an incomplete, fragmental understanding of modelling work, largely due to the rapid pace of publication on preprint servers and in peer-reviewed journals. To the best of our knowledge, most existing reviews are either systematic but only cover a short time span (eg, up until July, 2020),[7–9] or use a narrative approach and do not develop a method to examine a representative set of papers.[10–12] The only exceptions we found are one systematic review covering 242 papers up until November, 2020,[13] and one narrative review that covered 50 of the most cited papers.[14] Only one review included preprints,[13] and all are limited to papers published before August, 2020,[7–10] or in 2020.[12,13] Many of these reviews are focused on model objectives and methodology,[8,9,12] and neglect other aspects of modelling that are crucial for science translation to decision makers and the public.

In this Series paper, to build on previous work, we provide a narrative review with a systematic approach, which handles the challenges presented in synthesising an enormous body of work with objective criteria to obtain the most representative and informative sample of papers possible. Our review covers publications up until Aug 20, 2021, which captures 8 months of 2021 that have not been covered by other reviews. We focus on factors of modelling that have been neglected in the existing literature, namely input data, uncertainty, performance evaluation, and stated limitations, which are crucial for science translation and enable models to provide insight for decision makers and the public. We provide a quantitative evaluation of each of these elements, which enables strong and justified conclusions about trends and areas in need of improvement, with respect to modelling COVID-19 and future pandemics.

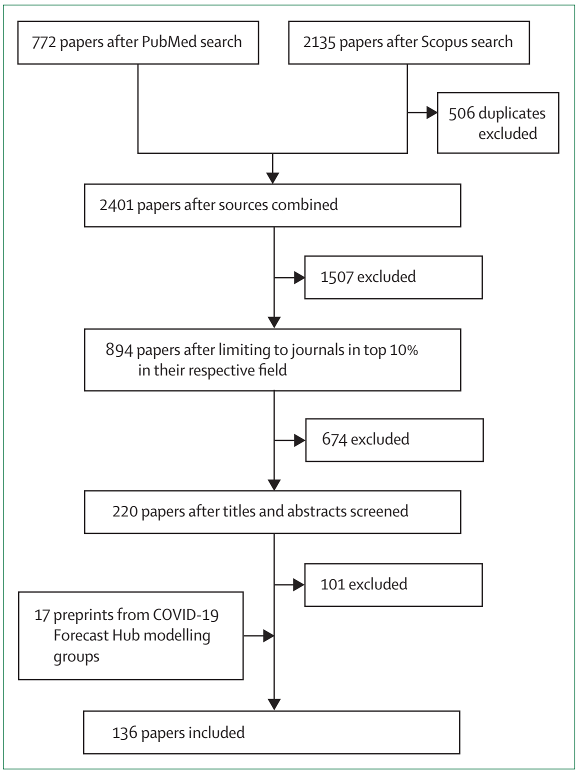
772 papers after PubMed search
2135 papers after Scopus search
506 duplicates excluded
2401 papers after sources combined
1507 excluded
894 papers after limiting to journals in top 10% in their respective field
674 excluded
220 papers after titles and abstracts screened
101 excluded
17 preprints from COVID-19 Forecast Hub modelling groups
136 papers included

Figure 1: Study selection

Methods

Search strategy and selection criteria

There are three main types of COVID-19 disease-spread modelling: retrospective modelling, nowcasting, and prospective modelling. Retrospective modelling, or backward-looking analysis, has been applied throughout the outbreak to explore a variety of key questions such as inferring basic epidemiological characteristics (eg, the basic reproductive rate [R_0], incubation period, and fatality rate), revealing factors driving transmission, and assessing the effectiveness of different interventions.[15–17] Nowcasting focuses on understanding the current situation, like inferring the true number of cases in light of under-reporting.[18] Prospective modelling is forward-looking, and includes forecasts, projections, and future scenario analysis. Forecasting aims to predict near-term epidemiological dynamics, often relying on data-driven methods and assuming that there will be minimal changes during the forecast period, whereas projections span over a much longer future time window, and thus have to make assumptions about how the factors driving COVID-19 will change in the future. Future scenario analyses produce multiple projections that explore the effects of different sets of assumptions that vary factors such as transmission rates and interventions.

See Online for appendix

Due to the magnitude of the COVID-19 modelling literature, we had to impose substantial constraints on the scope of this Series paper to enable us to do a systematic, quantitative, and timely assessment of the relevant literature. Therefore, this Series paper comprises a narrative review with a systematic approach. Specifically, these four inclusion criteria defined our review scope. (1) Prospective modelling work on population-level dynamics of COVID-19: we included papers that provided future predictions for a specific location, including forecasting, projections, and future scenario analysis. We excluded retrospective modelling studies and nowcasting. Papers that only fit a model without providing out-of-sample predictions were not included. (2) Data-driven: we broadly defined this as papers that incorporated COVID-19 data into the setup or fitting of the model. Papers that only used parameters from the literature or only used data from other viruses were excluded. (3) Geographical restriction: we only included papers published in English that implemented forecasting or projections (including future scenario analyses) for US counties, states, or at the national level, which restricted our analysis to papers working with the same data issues and in a similar context. (4) Journal restriction: we only included papers from peer-reviewed journals, as defined by Scopus' context curation standards,[19] or preprints from modellers that contribute to the US COVID-19 Forecast Hub.

For papers published in peer-reviewed journals, we restricted papers to those from journals ranked in the top 10% in their respective field on the basis of the Scopus CiteScore. Although we recognise this restriction will exclude important work, this criterion was the best option available to apply a systematic approach to reducing the set of papers to a manageable number while still obtaining the most representative sample of papers possible. For our final sample of peer-reviewed papers, the number of papers from each journal, and each journal's top category and rank percentile according to Scopus CiteScore, is shown in the appendix (pp 1–2). We developed a Scopus query on the basis of these criteria (appendix p 2). To minimise the chance of our search missing relevant papers, we searched PubMed with the equivalent query (figure 1).

We searched Scopus and PubMed on Aug 20, 2021, and our final selection of papers was distributed from March 23, 2020, to Aug 16, 2021 (figure 2). Notably, the top 10% criteria only reduced the number of papers to 37% of the original size, from 2401 to 894 papers. Papers were screened individually by KN, SJ, and FP, and could be confirmed by another screener if a paper's eligibility for inclusion was unclear. For the data collection, categorisations were done individually by the same authors, and confirmed on a second pass, with one individual covering all papers for a particular category to ensure consistent categorisation. 119 peer-reviewed papers were included (figure 1).

We additionally considered preprints from authors known to be engaged in real-time modelling work. We included preprints from modellers participating in the US COVID-19 Forecast Hub, which focuses on 1-week to

4-week predictions.[2] We also attempted to include the Scenario Modelling Hub,[20–22] but no preprints met our criteria for the time window considered. Although these papers do not have the validation that comes with peer-review, these models were used in real-time by the Centers for Disease Control and Prevention (CDC), which we believe justifies their inclusion in this analysis. We found 17 preprints in the metadata provided by the modelling teams contributing to the Forecast Hub. Thus, 136 papers in total are included in our analysis. Despite our efforts, we acknowledge that we will miss a substantial portion of real-time COVID-19 modelling work that exists on preprint servers and on the websites of modelling groups.

We have designed our process to obtain the most objective and representative sample possible, given the challenges of synthesising an enormous body of work in a useful, timely manner. Despite the limitations of our scoping process, we are confident that our analysis can provide valuable insight on the state of published COVID-19 work and highlight areas for improvement.

For more on the **CDC** see https://www.cdc.gov/

Categorisation analysis

To conduct a quantitative analysis on the substance and quality of these studies, for each paper we classified eight features: model objective and prediction horizon, methodology, target variables, data categories, geographical resolution, uncertainty, performance evaluation, and model limitations (appendix pp 3–6). We acknowledge that some of these categorisations are subjective or difficult to consistently extract from papers, especially the performance evaluation and stated limitations categories. Thus, we narrowly defined our categories and transparently discuss these definitions in the Results.

Since many of the existing COVID-19 review papers go into more detail on methodology,[8,9,12] we opted not to cover this aspect of modelling beyond classification into three broad categories: compartmental models (eg, susceptible, infectious, and recovered [known as SIR] and variations), statistical models (eg, machine learning, deep learning, and ARIMA), and hybrid (a combination of compartmental and statistical models).

To capture meaningful data on performance evaluation, we made an a priori decision to report on the performance evaluation only for the subset of papers implementing short-term prediction models, which can be fairly evaluated against truth data. By contrast, the purpose of long-term projections is to compare multiple plausible scenarios of the future, not to predict what will happen. Therefore, a fair performance evaluation with standard error metrics is not possible since these models make assumptions about the future that do not match reality.

To understand the multidisciplinary nature of the COVID-19 literature, we provide the most common journal subject areas, as defined by Scopus, in our set of papers. Additionally, we provide a breakdown of how the COVID-19 Forecast Hub papers compare with the entire set of papers on expressing uncertainty, conducting a thorough performance evaluation, and discussing limitations.

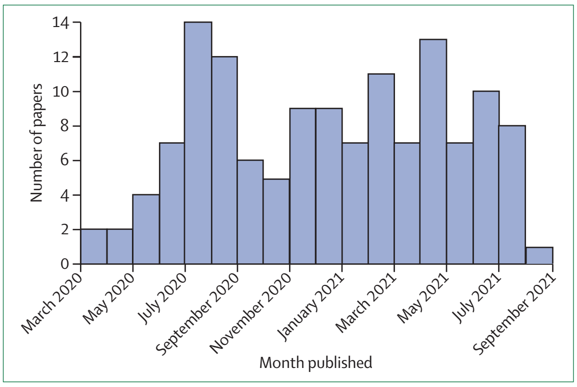

Figure 2: **Histogram of the number of papers in our analysis by month of publication**

Results and discussion

We visualised the relative size of each category and the most common connections between categories (figure 3). Each line through the figure represents the categorisations of a single paper. The width of the lines is weighted such that in cases of a paper being in more than one category, such as with both cases and deaths data, a line with half of the normal width is assigned to each category.

Model objective and prediction horizon

Forecasts are unconditional in the sense that they attempt to predict what will happen in the near future, whereas projections and scenarios are conditioned on the model's assumptions about the future to extend the prediction horizon. We were unable to reliably categorise models into forecasts or projections due to inconsistent use of these terms and a scarcity of clear communication on which approach was used in the papers. Since papers did not consistently state the precise objective of their model (unconditional forecast or assumption-based projection), we report a proxy for model objective: short-term predictions (ie, forecasts), or long-term predictions (ie, projections). To remain consistent with the COVID-19 Forecast Hub and COVID-19 Scenario Hub, which represent best practice for prospective COVID-19 modelling, we categorised studies that made predictions for 4 weeks or less as short-term (46%, n=63), and studies making predictions with a horizon that extended beyond 4 weeks as long-term (60%, n=82). There were a few papers that produced both long-term predictions and short-term predictions.[23–26] Because papers often fall into multiple categories, percentages in this analysis do not always add up to 100%. Within the category of papers conducting long-term projections, we also tagged papers with multiple scenarios, which provided multiple predictions based on different sets of assumptions. For

For more on the the **COVID-19 Scenario Hub** see https://covid19scenariomodelinghub.org/

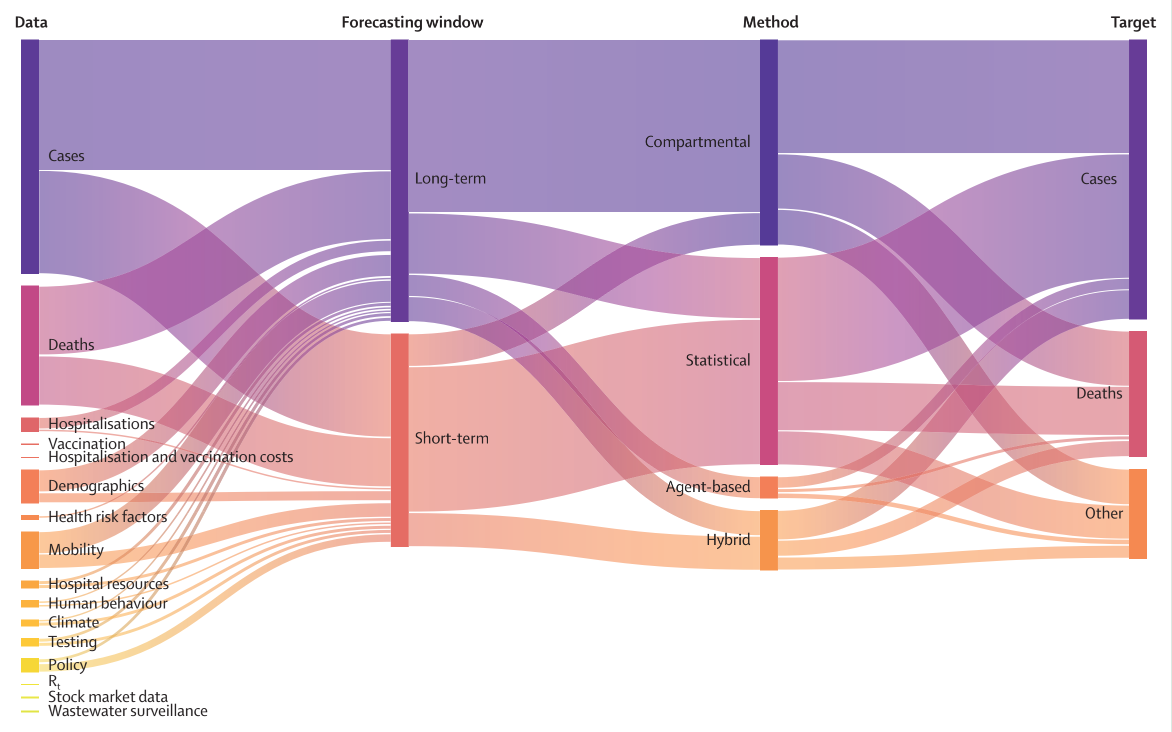

Figure 3: **Sankey diagram of the connections between categorisations of our analysis**
This diagram shows the relative co-occurrence of categories within papers in our analysis. Thicker lines between categories indicate that those categories are more likely to occur in the same paper. R_t=effective reproductive number.

example, modelling scenarios could explore the impact of different reopening speeds, non-pharmaceutical interventions, and vaccination rates. Of the 82 papers in the long-term projections category, 54 papers (66%) considered multiple scenarios.

Methodology

Since most compartmental models in our sample used statistical methods to fit parameters, to retain informative categories we adopted a stringent definition of a hybrid model, requiring both compartmental and statistical layers of the model that go beyond the use of statistical approaches to fit parameters. For example, one paper classified as hybrid used deep learning to infer a time-dependent reproduction number, which was then fed into a compartmental model.[27] A model that only uses statistical methods to fit parameters for a compartmental model was classified as compartmental. We found that 47% of papers (n=64) used a compartmental model, 43% (n=59) used a statistical model, and 13% (n=17) used a hybrid model. A few papers developed and showed both a compartmental model and a statistical model.[23–26] We also noted when models used agent-based methods (9%, n=12; figure 3; appendix pp 3–6).

Target variables

The most common target prediction variables were cases (89%, n=121), deaths (52%, n=71), hospitalisations (10%, n=14), and effective reproductive number (R_t; 9%, n=12). Some of the lesser used target variables included growth rate, peak cases, and intensive care unit admissions. 38% (n=52) of papers had only one target variable, 43% (n=59) of papers had two target variables, and 18% (n=25) had more than two (appendix pp 3–6).

The target prediction variables were dominated by absolute numbers of cases and deaths, which aligns with the goals of the US COVID-19 Forecast Hub. Despite the continued desire for these targets from across the field of public health, government, industry, and the public, accurate prediction of them remains challenging.[28]

Data categories

Next, we quantified the categories of input data used to inform models. We defined the data categories (table 1), including an in-depth look at the datasets used by papers in our analysis that attempt to capture COVID-19 behaviours.

The most frequently used data categories were cases, deaths, mobility, demographics, and hospital admissions (table 2). 20% (n=27) of papers used only one category of data, 39% (n=53) of papers used two categories, 16% (n=22) used three categories, and 25% (n=34) used four or more categories.

The data sources that informed predictions in our analysis were dominated by case and death data

	Description	Examples
Cases or deaths	Epidemiological data on the number of cases or deaths and corresponding metrics	Daily cases or deaths , cumulative cases or deaths, reproduction number, and growth rate
Hospital admissions	Data related to hospitalisation of patients with COVID-19	Daily hospitalisations, active hospitalisations, and intensive care unit occupancy
Testing	Data pertaining to COVID-19 testing in a population or location	Daily tests and test positivity rate
Climate	Data describing the climate or any meteorological variables pertaining to a specific location; time series or static data	Daily precipitation, daily temperature, and average temperature
Demographics	Demographic or sociodemographic information about the population of a specific location	Population, age, race, income, and rural to urban ratio
Hospital resources	Data on the amount of certain resources available in hospitals	Number of beds and intensive care unit beds
Health risk factors	Data that quantifies the health risk factors of the population in the context of COVID-19	Prevalence of comorbidities and use of preventative services (eg, doctor visits)
Mobility	Data that quantifies the movement of a population	Google Mobility Trends (residential, grocery and pharmacy stores, parks, retail and recreation, workplaces, and transit stations),[29] Unacast social distancing scoreboard (average mobility, non-essential visits, and encounters density),[30] SafeGraph (trip counts at a census block group resolution),[31] Apple Mobility Trends (trends in Apple Maps routing requests),[32] Facebook Movement Range Maps (change in movement compared with baseline percentage of population who stays home),[33] and flight data
Human behaviour	Data that quantifies the behaviour or beliefs of a population in the context of COVID-19, excluding data on the mobility of a population	Google search trends,[34] mask use per capita,[35] Facebook's COVID-19 Trends and Impact Survey (time series of self-reported mask use and other social distancing behaviours),[36] *New York Times* Mask-Wearing Survey data (static),[37] and sentiment index constructed from COVID-19 news[38]
Policy	Data pertaining to COVID-19 policies	Oxford COVID-19 Government Response Tracker (ordinal scale on stringency of many types of COVID-19 policies, including containment and closure policies, economic policies, health system policies, and vaccination policies),[39] state-level social distancing policies (dates and details of policies including emergency declarations, gathering restrictions, closures, stay-at-home orders, travel restrictions, isolation orders, and mask mandates)[40]

Table 1: Data categories

(figure 3). Data used in two or less papers include vaccinations, R_t, wastewater surveillance, and economic data. 51% (n=70) of modelling studies only used epidemiological data sources (ie, cases, deaths, and hospital admissions). The most frequently used non-epidemiological sources were mobility and demographic data. The models that did use other data sources tended to incorporate a large number and variety of input data.[41–43] Some factors that have been shown to be associated with COVID-19 dynamics, such as demographics, health risk factors, and climate, rarely appeared in our sample, although little research has been done to rigorously test for whether these factors can improve predictive performance. Despite the increasing effect of new variants on epidemiological dynamics, none of the papers in our sample used variant prevalence data. In the USA, these data have a low sample size, sampling bias, and are difficult to use as a signal for predictive modelling.

Geographical resolution

We noted the geographical scale at which predictions were made, categorising papers as national, state, or county level and smaller. 54% of 136 papers included a national-level prediction, 36% (n=49) at the state level, and 34% (n=46) at the county level or smaller scale (table 3). Half of the models in our analysis were at the national level. This resolution tends to be the easiest to predict and the least useful for decision making, which must often occur at the local level.

Uncertainty

We established which papers included a quantitative expression of uncertainty of their predictions, excluding those that only did so for model parameters. We found that half of the papers (50%, n=68; table 3) did not express quantitative uncertainty around the predictions, despite the highly uncertain and consequential nature of COVID-19 dynamics. 49% of papers (n=67) included some form of confidence or prediction intervals. A sensitivity analysis was performed in 13% of papers (n=18; appendix pp 3–6).

The use of forecasts for decision makers is dependent on clear communication of uncertainty,[44] especially since point estimate predictions will rarely match ground-truth data. Well calibrated expressions of uncertainty help stakeholders assess future risk and decide how to respond. For example, the difference between a 1% chance of exceeding hospital capacity versus a 25% chance could establish whether or not certain preparatory actions are taken. Additionally, expressing uncertainty is especially important to prevent harmful, incorrect interpretations of COVID-19 models. Clearly communicating uncertainty around predictions weakens the ability of actors to use a study in a misleading way to support their pre-existing agenda.

Performance evaluation

We categorised the type of performance evaluation used for each short-term model, which can be fairly evaluated on ground truth data. When defining our performance

	Occurrences, n (%; n=136)
Cases	126 (93%)
Deaths	79 (58%)
Mobility	34 (25%)
Demographics	30 (22%)
Hospital admissions	15 (11%)
Policy	13 (10%)
Testing	11 (8%)
Hospital resources	10 (7%)
Climate	8 (6%)
Human behaviour	8 (6%)
Health risk factors	4 (3%)

Numbers exceed 136 as categories overlap between papers.

Table 2: **Papers in the top data categories**

	All papers (n=136)	Forecast Hub papers and preprints (n=20)
Prediction horizon		
Short-term predictions	63 (46%)	14 (70%)
Long-term predictions	82 (60%)	8 (40%)
Methodology		
Compartmental	64 (47%)	7 (35%)
Statistical	59 (43%)	9 (45%)
Hybrid	17 (13%)	4 (20%)
Agent-based	12 (9%)	1 (5%)
Geographical level		
National	74 (54%)	5 (25%)
State	49 (36%)	13 (65%)
County or smaller	46 (34%)	11 (55%)
Uncertainty		
Expressed quantitative uncertainty	68 (50%)	11 (55%)
Sensitivity analysis	18 (13%)	1 (5%)
Performance evaluation (out of short-term models only)		
Comparison to ground truth	47/63 (75%)	12/14 (86%)
Number of predictions (out of short-term models only)		
Only made predictions from one date	39/63 (62%)	1/14 (7%)
Made multiple predictions over a timespan less than 2 months	10/63 (16%)	6/14 (43%)
Made multiple predictions over a timespan greater than 2 months	14/63 (22%)	7/14 (50%)
Limitations		
Authors discussed limitations	87 (64%)	13 (65%)

Table 3: **Comparison of category occurrences in all papers and Forecast Hub papers and preprints**

evaluation categories, we considered that for timeseries forecasts, the setup of training and testing data should be representative of real-time forecasting conditions. Since the use of a model is based on its ability to predict future dynamics, randomly excluded out-of-sample evaluation methods do not adequately describe performance. Instead, models should be trained with data up until a certain cutoff date and evaluated with data after that date. This future-blind approach preserves the fundamental challenge of forecasting: not knowing future data or trends. Within the subset of short-term studies considered (N=63), 75% (n=47) of papers used performance evaluation metrics to compare future-blind, out-of-sample predictions to ground truth data. Ground truth data are usually reported cases or deaths, and sources used in our sample include the Center for Systems Science and Engineering at Johns Hopkins University (Baltimore, MD, USA),[1] the COVID Tracking Project, and WHO Dashboard. The most common metrics to compare predictions to ground truth were mean absolute error, root mean square error, mean absolute percentage error, coefficient of determination, mean square error, and coverage rate of prediction intervals. Of the papers that did a metric-based evaluation, only 13% (n=6) evaluated the accuracy of confidence intervals (table 3). Within the group of 47 papers that conducted a future-blind performance evaluation, 34% (n=16) evaluated only one model, 55% (n=26) compared performance metrics across multiple internal models, and 19% (n=9) compared the performance metrics of their model against those of other models in the COVID-19 Forecast Hub. 15% (n=7) of evaluated models used a baseline model for comparison (appendix pp 3–6).

Although most of the 63 modelling studies (75%, n=58) quantified the performance of their model relative to ground truth data, 78% (n=49) did not evaluate their model on predictions made across a timespan that included varying epidemiological dynamics. To quantify the frequency of these practices, we counted the number of dates from which papers showed predictions. For example, if a paper presents a model prediction with data up until Sept 1 and predicts future case counts on Sept 8, 15, 22, and 29, this prediction would be made from a single date. If this paper adds another prediction made from Oct 1 (with data up until this date) and predicts weekly values for the next 4 weeks, this paper would be showing predictions made from two dates, which cover a month-long timespan (Sept 1 to Oct 1). We defined the category this way to ensure we could reliably extract these data from each paper. Our analysis found that among short-term models, more than half (62%, n=39) only showed a prediction made from a single date, 16% (n=10) of papers showed predictions made from multiple dates over a timespan that was less than 2 months long, and 22% (n=14) covered a timespan longer than 2 months. From the COVID-19 Forecast Hub, we know that predictive accuracy of models varies widely over time, especially with respect to epidemiological trends.[45] Therefore, not evaluating a model across a variety of epidemiological dynamics severely limits the generalisability of the performance evaluation and the ability to make fair comparisons between models. In addition, a third of papers (34%, n=16) that completed a quantitative performance evaluation did not compare their model to a baseline or any other models, so whether the model provides any improvement over a naive model is unclear. The COVID-19 Forecast Hub uses a baseline model that assumes no change in incidence over the next 4 weeks. According to historical error metrics calculated by the Forecast Hub and Carnegie Mellon University (CMU; Pittsburgh, PA, USA) Delphi, on Sept 8, 2021, only 25% of models outperformed the baseline model for cases, whereas 75% outperformed the baseline for deaths by relative mean absolute error and weighted interval score.[46] Thus, comparison with a baseline model provides context and thereby important information about the usefulness of a model.

Many papers did not cover the specific methodology of their performance evaluation, which limited our ability to provide more specific analyses in this Series paper. Authors should clearly state the dates of the training period, the dates predictions were made from, how error metrics were computed and aggregated, and whether metrics are computed in-sample or out-of-sample. In addition, models that aim to contribute to real-time forecasting efforts should use input data as they were available at the date predicitons are made from; these data are available from the CMU Delphi's COVIDcast Epidata.[47,48] Without thorough performance evaluation, the broader scientific community will be unable to identify which approaches are working and build knowledge on best practices.

Model limitations

Authors of the papers stated six main categories of limitations: disregarded factors (39%), data quality (28%), unknowable factors (26%), limitations specific to the methods used (22%), data availability (16%), and poor generalisability (8%). We define unknowable factors as those that cannot be known at the time predictions were made, such as future implementation of non-pharmaceutical interventions, or the emergence of new variants during the prediction horizon. By contrast, disregarded factors have some relevant data or information available at the time of the analysis, but the authors of the papers chose to disregard it, like the demographic breakdown of populations or health-care capacity of different regions. A third of the papers in our analysis (36%) did not list any limitations in an accessible section of the paper, which we considered to be in the discussion, conclusion, or in a separate section called limitations. In most cases, all these types of limitations are relevant to COVID-19 models. Unfortunately, our categorisation does not give information about how thoroughly these limitation categories were discussed.

Multidisciplinary nature of the COVID-19 literature

The highly consequential nature of the COVID-19 pandemic has attracted modelling experts from a variety of different fields. The top five journal subject areas represented in our final set of papers, in order from most to least frequent, are applied mathematics (30%, n=41), multidisciplinary (22%, n=30), general physics and astronomy (21%, n=29), general mathematics (18%, n=24), and statistical and non-linear physics (16%, n=22). Note that the assignment of journals to subject areas was done by Scopus. Public health did not appear in the top five subject areas. Our final set of papers represented 52 journals. The most common journals were *Chaos, Solitons & Fractals*; *PLoS One*; and *Scientific Reports* (figure 4). We were unable to conduct a thorough analysis on the contributions to COVID-19 modelling from different fields due to the difficulty of classifying papers into distinct disciplines solely on the basis of the journal they were published in and the inherent interdisciplinarity of this work. However, we completed a subanalysis on the group of papers from COVID-19 Forecast Hub modellers.

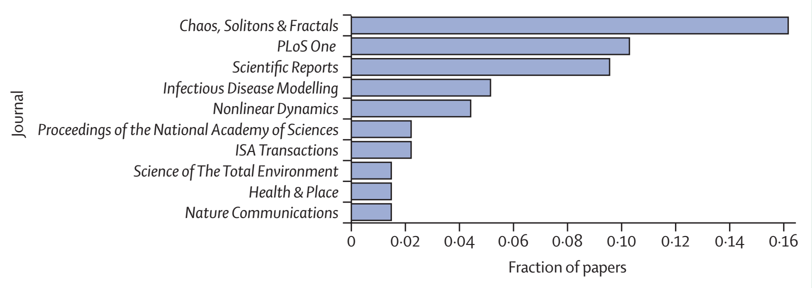

Figure 4: Top 10 journals in the final set

The set of papers written by authors that contributed to the COVID-19 Forecast Hub includes 17 preprints[42,49–64] and three papers published in peer-reviewed journals.[65–67] 70% of these papers made short-term predictions and 40% of these papers made long-term predictions. Although these papers were cited by teams in the metadata of their submissions to the COVID-19 Forecast Hub, these preprints are not necessarily on the exact model and application that was submitted to the COVID-19 Forecast Hub. Despite being mostly preprints with many serving to provide a brief explanation of a model being used in real-time, these papers were more likely to express uncertainty, have forecasts for state and county levels, and conduct performance evaluation than the full set of papers (table 3). In addition, COVID-19 Forecast Hub papers were substantially more likely to show and evaluate predictions made from several dates over a timespan greater than 2 months (50% *vs* 22% for all papers). A great advantage of the COVID-19 Forecast Hub approach is that it encourages good practices in terms of uncertainty, evaluation, and high geographical resolution. Additionally, the real-time sharing of forecasts ensures that predictions were truly future-blind.

Concluding remarks

Our analysis found substantial gaps in COVID-19 model transparency in the literature, especially on reporting aspects of models that are crucial for science translation.

Papers did not consistently state the precise objective of their model (unconditional forecast or assumption-based projection), detail their methodology, express uncertainty, evaluate performance across a long, varied timespan, and clearly list their limitations. Without this information, studies are more vulnerable to misinterpretation, which can have serious consequences during a global health crisis in which decision makers and the public rely on scientific papers for critical guidance.[68,69] In addition, poor reporting limits the ability of literature reviews to synthesise insights from the research to establish best practices. In response to these kind of concerns, the EPIFORGE 2020 model reporting guidelines[70] were developed, primarily for epidemic forecasting studies, but but the concepts apply to other types of modeling as well. These guidelines recommend consistent terminology, a clear definition of study purpose and model targets, identification of prospective versus retrospective work, comparison with a baseline model, a non-technical summary of results, and full documentation of: data sources, data availability, data processing, methods, assumptions, code, model validation, forecast accuracy evaluation, uncertainty, limitations, interpretation, and generalisability.[70] Consistent sharing of this information for epidemiological predictions would improve the consistency, reproducibility, comparability, and quality of epidemic forecasting and modelling papers, in addition to minimising the potential for the public to misunderstand or misuse the research.

Another obstacle to maximising the knowledge gained from epidemic modelling is the suitability of the information-sharing system. Since it is not standard practice for modelling papers to report on translational work, this Series paper can only comment on the translation potential of papers on the basis of their reporting practices, not on how models were actually used during the COVID-19 pandemic. In addition, the volume and variable quality of the literature forced us to adopt stringent and limiting scoping criteria to obtain a manageable sample of literature to analyse. Other reviews adopted their own narrow scope, creating a body of COVID-19 modelling literature reviews that amount to a fragmented, incomplete understanding of the efforts of researchers.

The obstacles to completing this literature review illustrate the difficulty of building knowledge from the COVID-19 literature through the traditional information-sharing system: peer-reviewed literature synthesised by systematic literature reviews. Thus, a new information-sharing system that is better suited to the needs of outbreaks is urgently needed, which can handle the pace of publications and strike a balance between the speed and quality of disseminating research findings.

Limitations

For COVID-19 applications, clearly stating model limitations is crucial to help the public understand the appropriate interpretation of results. The main limitations of this Series paper are the result of the difficult nature of synthesising the COVID-19 literature. We had to adopt stringent scoping criteria, which included limiting our analysis to studies that made prospective, data-driven predictions for the USA, and to papers published in the top 10% of journals based on Scopus' CiteScore.[19] The CiteScore is an imperfect metric that relies on the number of citations per study in a journal. However, the CiteScore was the best option we knew of to select for a higher quality sample of papers, since we did not want to introduce a time bias by using each paper's number of citations. Another limitation is that we can only comment on the state of the peer-reviewed literature (and a specially selected sample of the preprint literature) within this analysis, not the state of all real-time work, some of which is not and might never be represented in the literature. In addition, some of the categorisations we made were subjective and difficult to extract consistently, so we implemented quality control mechanisms as discussed in the Methods, and we are confident in our overall conclusions. Despite these limitations, we believe we have studied the most representative sample of papers possible and obtained findings that are informative for improving epidemic modelling in the future.

Conclusions

To conclude, this Series paper examined a subset of the COVID-19 modelling literature, focused on data-driven, prospective modelling, and identified several opportunities to improve the use of outbreak modelling, which are especially relevant to inform the work of the new CDC Center for Forecasting and Outbreak Analytics, for which planning began in August, 2021. In response to considerable scoping challenges, we selected a sample that should represent the best modelling papers and still found them to be inadequate in some of the areas that are most crucial for translating models into useful insight for decision makers and the general public.

The main takeaways of this Series paper are adopting epidemic forecasting standards and creating a suitable information-sharing system. Adopting the EPIFORGE 2020 model reporting guidelines addresses many of the issues identified in this Series paper, including the need to be transparent about the methods, express uncertainty, thoroughly evaluate performance, state limitations, and discuss appropriate interpretations. Additionally, the creation of an information-sharing system suited to the needs of an epidemic would allow the hard work of COVID-19 modellers to be more efficiently synthesised into best practices.

Contributors

LG, KN, and SJ contributed to the conceptualisation and design of the study. KN, SJ, and FP collected and analysed the data. FP and SJ made the figures. KN led the writing of the original draft. NGR, KG, ECL, ST and LG edited the manuscript. LG supervised the study. KN and SJ have verified the underlying data. All authors had full access to the data and approved the manuscript for publication.

Declaration of interests

ECL declares payment for expert testimony from Cohen Ziffer Frenchman & McKenna for a report related to COVID-19 epidemiology. KN, SJ, and LG submitted a model to the US COVID-19 Forecast Hub. NGR is a coauthor on EPIFORGE 2020 model reporting guidelines, is a codirector of the Forecast Hub, has submitted individual models to the Forecast Hub, and served in an advisory role for the US Scenario Modeling Hub. ST is a cofounder and member of the leadership team for the US Scenario Modeling Hub, and has submitted individual models to both the Scenario and Forecast Hubs. ECL has submitted models to both the Forecast and Scenario Hubs. FP and KG declare no competing interests.

Acknowledgments

KN, SJ, and LG were funded by the NSF Rapid Response Research grants (award ID 2108526 and 2028604), and the CDC SHEPheRD Project (200-2016-91781). ECL declares support from the National Science Foundation (NSF), a grant from Public Health Institute. NGR has been supported by the National Institutes of General Medical Sciences (NIGMS; R35GM119582), and a grant from the CDC (1U01IP001122). KG and FP were funded by the Rockefeller Foundation Covid-19 Modeling Accelerator and CDC U01CK000589. ST declares a grant from NSF (2127976) and a CDC contract (200-2016-91781). The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The authors had final responsibility for the decision to submit for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIGMS or the National Institutes of Health.

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:** 533–34.

2 Fraser N, Brierley L, Dey G, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS Biol* 2021; **19:** e3000959.

3 Horbach SPJM. Pandemic publishing: medical journals strongly speed up their publication process for COVID-19. *Quant Sci Stud* 2020; **1:** 1056–67.

4 James LP, Salomon JA, Buckee CO, Menzies NA. The use and misuse of mathematical modeling for infectious disease policymaking: lessons for the COVID-19 pandemic. *Med Decis Making* 2021; **41:** 379–85.

5 Press WH, Levin RC. Modeling, post COVID-19. *Science* 2020; **370:** 1015.

6 Ioannidis JPA, Cripps S, Tanner MA. Forecasting for COVID-19 has failed. *Int J Forecast* 2020; **38:** 423–38.

7 Shankar S, Mohakuda SS, Kumar A, et al. Systematic review of predictive mathematical models of COVID-19 epidemic. *Med J Armed Forces India* 2021; **77:** S385–92.

8 Guan J, Wei Y, Zhao Y, Chen F. Modeling the transmission dynamics of COVID-19 epidemic: a systematic review. *J Biomed Res* 2020; **34:** 422.

9 Xiang Y, Jia Y, Chen L, Guo L, Shu B. Long E. COVID-19 epidemic prediction and the impact of public health interventions: a review of COVID-19 epidemic models. *Infect Dis Model* 2021; **6:** 324–42.

10 Zawadzki RS, Gong CL, Cho SK, et al. Where do we go from here? A framework for using susceptible-infectious-recovered models for policy making in emerging infectious diseases. *Value Health* 2021; **24:** 917–24.

11 Adiga A, Dubhashi D, Lewis B, Marathe M, Venkatramanan S, Vullikanti A. Mathematical models for COVID-19 pandemic: a comparative analysis. *J Indian Inst Sci* 2020; **100:** 793–807.

12 Rahimi I, Chen F, Gandomi AH. A review on COVID-19 forecasting models. *Neural Comput Appl* 2021; published online Feb 4. https://doi.org/10.1007/s00521-020-05626-8.

13 Gnanvi JE, Salako KV, Kotanmi GB, Glèlè Kakaï R. On the reliability of predictions on Covid-19 dynamics: a systematic and critical review of modelling techniques. *Infect Dis Model* 2021; **6:** 258–72.

14 McCabe R, Donnelly CA. Disease transmission and control modelling at the science–policy interface. *Interface Focus* 2021; **11:** 20210013.

15 Chinazzi M, Davis JT, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* 2020; **368:** 395–400.

16 Flaxman S, Mishra S, Gandy A, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* 2020; **584:** 257–61.

17 Tian H, Liu Y, Li Y, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* 2020; **368:** 638–42.

18 Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* 2020; **395:** 689–97.

19 Elsevier. How Scopus works: Scopus content. https://www.elsevier.com/solutions/scopus/how-scopus-works/content (accessed May 6, 2022).

20 Lemaitre JC, Grantz KH, Kaminsky J, et al. A scenario modeling pipeline for COVID-19 emergency planning. *Scientific Reports* 2021; **11:** 7534.

21 Truelove S, Smith CP, Qin M, et al. Projected resurgence of COVID-19 in the United States in July–December 2021 resulting from the increased transmissibility of the Delta variant and faltering vaccination. *eLife* 2022; **11:** e73584.

22 Borchering RK, Viboud C, Howerton E, et al. Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios—United States, April–September 2021. *MMWR Morb Mortal Wkly Rep* 2021; **70:** 719–24.

23 Li Q, Bedi T, Lehmann CU, Xiao G, Xie Y. Evaluating short-term forecasting of COVID-19 cases among different epidemiological models under a Bayesian framework. *Gigascience* 2021; **10:** giab009.

24 Nikolopoulos K, Punia S, Schäfers A, Tsinopoulos C, Vasilakis C. Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *Eur J Oper Res* 2021; **290:** 99–115.

25 Cot C, Cacciapaglia G, Islind AS, Óskarsdóttir M, Sannino F. Impact of US vaccination strategy on COVID-19 wave dynamics. *Sci Rep* 2021; **11:** 10960.

26 Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. *Proc Natl Acad Sci USA* 2020; **117:** 16732–38.

27 Bhouri MA, Costabal FS, Wang H, et al. COVID-19 dynamics across the US: a deep learning study of human mobility and social behavior. *Comput Methods Appl Mech Eng* 2021; **382:** 113891.

28 Reich NG, Tibshirani RJ, Ray EL, Rosenfeld R. On the predictability of COVID-19. Sept 28, 2021. https://forecasters.org/blog/2021/09/28/on-the-predictability-of-covid-19/ (accessed Dec 6, 2021).

29 Google. COVID-19 community mobility reports. https://www.google.com/covid19/mobility/ (accessed Dec 14, 2021).

30 Unacast. Social distancing scoreboard. https://www.unacast.com/covid19/covid-19-retail-impact-scoreboard (accessed Dec 14, 2021).

31 SafeGraph. SafeGraph data for academics in the United States. https://www.safegraph.com/academics (accessed Dec 14, 2021).

32 Apple. COVID19 mobility trends reports. https://covid19.apple.com/mobility (accessed Dec 14, 2021).

33 Meta (Facebook). Movement range maps. https://dataforgood.facebook.com/dfg/tools/movement-range-maps (accessed Dec 14, 2021).

34 Google Trends. Coronavirus search trends. https://trends.google.com/trends/story/GB_cu_JSW_pHABAADqAM_en (accessed Dec 14, 2021).

35 Institute for Health Metrics and Evaluation. COVID-19 projections. https://covid19.healthdata.org/global?view=mask-use&tab=trend (accessed Dec 14, 2021).

36 Meta (Facebook). COVID 19 Symptom Survey. https://dataforgood.facebook.com/dfg/tools/covid-19-trends-and-impact-survey#methodology (accessed Dec 14, 2021).

37 GitHub. Mask-wearing survey data. https://github.com/nytimes/covid-19-data/tree/master/mask-use (accessed Dec 14, 2021).

38 Chalkiadakis I, Yan H, Peters GW, Shevchenko PV. Infection rate models for COVID-19: model risk and public health news sentiment exposure adjustments. *PLoS One* 2021; **16:** e0253381.

39 Hale T, Angrist N, Goldszmidt R, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat Hum Behav* 2021; **5:** 529–38.

40 Adolph C, Amano K, Bang-Jensen B, Fullman N, Wilkerson J. Pandemic politics: timing state-level social distancing responses to COVID-19. *J Health Polit Policy Law* 2021; **46:** 211–33.

41 Duque D, Morton DP, Singh B, Du Z, Pasco R, Meyers LA. Timing social distancing to avert unmanageable COVID-19 hospital surges. *Proc Natl Acad Sci USA* 2020; **117:** 19873–78.

42 Arik SO, Li CL, Yoon J, et al. Interpretable sequence learning for COVID-19 forecasting. *arXiv* 2020; published online Jan 13, 2021. https://doi.org/10.48550/arXiv.2008.00646 (preprint).

43 Lee SY, Lei B, Mallick B. Estimation of COVID-19 spread curves integrating global data and borrowing information. *PLoS One* 2020; **15:** e0236860.

44 Lutz CS, Huynh MP, Schroeder M, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health* 2019; **19:** 1659.

45 Cramer EY, Ray EL, Lopez VK, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc Natl Acad Sci USA* 2022; **119:** e2113561119.

46 COVID 19 ForecastHub. COVID-19 US forecast evaluation report. Sept 8, 2021. https://covid19forecasthub.org/eval-reports/?state=US&week=2021-09-08 (accessed Dec 7, 2021).

47 Reinhart A, Brooks L, Jahja M, et al. An open repository of real-time COVID-19 indicators. *Proc Natl Acad Sci USA* 2021; **118:** e2111452118.

48 Carnegie Mellon University Delphi Group. COVIDcast dashboard. https://delphi.cmu.edu/covidcast/ (accessed June 12, 2022).

49 Zou D, Wang L, Xu P, Chen J, Zhang W, Gu Q. Epidemic model guided machine learning for COVID-19 forecasts in the United States. *medRxiv* 2020; published online May 25. https://doi.org/10.1101/2020.05.24.20111989 (preprint).

50 Khan ZS, van Bussel F, Hussain F. A predictive model for Covid-19 spread applied to eight US states. *arXiv* 2020; published online June 10. https://doi.org/10.48550/arXiv.2006.05955 (preprint).

51 Galasso J, Cao DM, Hochberg R. A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. *Chaos Solitons Fractals* 2022; **156:** 111779.

52 Zhang-James Y, Hess J, Salekin A, et al. A seq2seq model to forecast the COVID-19 cases, deaths and reproductive R numbers in US counties. *medRxiv* 2021; published online April 20. https://doi.org/10.1101/2021.04.14.21255507 (preprint).

53 Shi Y, Ban X. Capping mobility to control COVID-19: a collision-based infectious disease transmission model. *medRxiv* 2020; published online July 28. https://doi.org/10.1101/2020.07.25.20162016 (preprint).

54 Wu D, Gao L, Xiong X, et al. DeepGLEAM: a hybrid mechanistic and deep learning model for COVID-19 forecasting. *arXiv* 2021; published online Feb 12. https://arxiv.org/abs/2102.06684v3 (preprint).

55 Srivastava A, Xu T, Prasanna VK. Fast and accurate forecasting of COVID-19 deaths using the SIkJalpha model. *arXiv* 2020; published online July 10. https://arxiv.org/abs/2007.05180v2 (preprint).

56 IHME COVID-19 health service utilization forecasting team, Murray CJL. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US in the next 4 months. *medRxiv* 2020; published online March 30. https://doi.org/10.1101/2020.03.27.20043752 (preprint).

57 Pei S, Shaman J. Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. *medRxiv* 2020; published online March 27. https://doi.org/10.1101/2020.03.21.20040303 (preprint).

58 Gibson GC, Reich NG, Sheldon D. Real-time mechanistic Bayesian forecasts of COVID-19 mortality. *medRxiv* 2020; published online Dec 24. https://doi.org/10.1101/2020.12.22.20248736 (preprint).

59 Wang L, Wang G, Gao L, et al. Spatiotemporal dynamics, nowcasting and forecasting of COVID-19 in the United States. *arXiv* 2020; published online April 29. https://doi.org/10.48550/arXiv.2004.14103 (preprint).

60 Meta (Facebook). Neural relational autoregression for high-resolution COVID-19 forecasting. Sept 23, 2020. https://ai.facebook.com/research/publications/neural-relational-autoregression-for-high-resolution-covid-19-forecasting/ (accessed Sept 20, 2021).

61 Biegel HR, Lega J. EpiCovDA: a mechanistic COVID-19 forecasting model with data assimilation. *arXiv* 2021; published online May 12. https://doi.org/10.48550/arXiv.2105.05471 (preprint).

62 Wilson DJ. Weather, social distancing, and the spread of COVID-19. July, 2020. https://doi.org/10.24148/wp2020-23 (accessed Dec 14, 2021).

63 Baxter A, Oruc BE, Keskinocak P, Asplund J, Serban N. Evaluating scenarios for school reopening under COVID19. *BMC Public Health* 2020; **22:** 496.

64 Rodriguez A, Tabassum A, Cui J, et al. DeepCOVID: an operational deep learning-driven framework for explainable real-time COVID-19 forecasting. *Proc Conf AAAI Artif Intell* 2020; **35:** 15393–400.

65 Pei S, Kandula S, Shaman J. Differential effects of intervention timing on COVID-19 spread in the United States. *Sci Adv* 2020; **6:** eabd6370.

66 Rowland MA, Swannack TM, Mayo ML, et al. COVID-19 infection data encode a dynamic reproduction number in response to policy decisions with secondary wave implications. *Sci Rep* 2021; **11:** 10875.

67 Gao J, Sharma R, Qian C, et al. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. *J Am Med Inform Assoc* 2021; **28:** 733–43.

68 Bagdasarian N, Cross GB, Fisher D. Rapid publications risk the integrity of science in the era of COVID-19. *BMC Med* 2020; **18:** 192.

69 Fraser N, Brierley L, Dey G, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS Biol* 2021; **19:** e3000959.

70 Pollett S, Johansson MA, Reich NG, et al. Recommended reporting items for epidemic forecasting and prediction research: the EPIFORGE 2020 guidelines. *PLoS Med* 2021; **18:** e1003793.