An Evaluation of Prospective COVID-19 Modeling in the US: From Data to Science Translation

Kristen Nixon[1], Sonia Jindal[1], Felix Parker[1], Nicholas G. Reich[2], Kimia Ghobadi[1], Elizabeth C. Lee[3], Shaun Truelove[3], Lauren Gardner[1]

[1] Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore MD, USA
[2] University of Massachusetts–Amherst, School of Public Health and Health Sciences, Amherst, Massachusetts, United States of America
[3] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

To whom correspondence should be addressed; E-mail: l.gardner@jhu.edu

Summary

Infectious disease modeling can serve as a powerful tool for situational awareness and decision support for policy makers. However, COVID-19 modeling efforts faced many challenges, from poor data quality to changing policy and human behavior. To extract practical insight from the large body of COVID-19 modeling literature, we provide a narrative review with a systematic approach to quantitatively assess prospective, data-driven modeling studies of COVID-19 in the US. We focus on the aspects of models that are critical to decision-makers. We found that a significant fraction of papers neglect evaluating performance (25%), expressing uncertainty (50%), and stating limitations (36%). We also document the forecasting window, methodology, prediction target, datasets used, and geographic resolution for each study. 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 outbreak science.

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 million deaths globally as of May 2022[1]. To aid planning and response efforts during a pandemic, mathematical modeling of current and future trends of 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, while long-term projections and scenario modeling can shed light on outcomes that may result from a set of assumptions. Insights from modeling can educate individuals on how to mitigate their own risks, while also providing decision support for policy makers seeking to minimize 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 produced: 125,000 within 10 months of the first confirmed case, 30,000 of which are preprints[2]. In this noisy publication landscape, journals prioritized quickly sharing COVID-19 information, but there is a trade-off between speeding up peer review and ensuring high quality research[3]. Preprints also played an important role in spreading COVID-19 research. Preprints were often covered in the media, had large audiences on social media platforms like Twitter, and in some cases were misunderstood in consequential ways [2]. For COVID-19 modeling specifically, the utility of models for informing response efforts was criticized largely due to 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 synthesize COVID-19 modeling work that have been published up until August 2021 form an incomplete, piecemeal understanding of modeling work, largely due to the rapid pace of publication on preprint servers and in publications. Most existing reviews are either systematic but only cover a short time span, 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 modeling that are crucial for translation.

To build on previous work, we provide a narrative review with a systematic approach, which handles the challenges presented in synthesizing an enormous body of work using objective criteria to obtain the most representative and informative sample of papers possible. Our review covers publications up until August 20, 2021, which captures eight months of 2021 that have not been covered by other reviews. We focus on factors of modeling that have been neglected in the existing literature, namely input data, uncertainty, performance evaluation, and stated limitations, which are critical 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 stronger and more justified conclusions about trends and areas in need of improvement, with respect to modeling COVID-19 and future pandemics.

Methods: Search Strategy and Selection Criteria

There are three main types of COVID-19 disease spread modeling: retrospective analysis, nowcasting, and prospective modeling. Retrospective modeling, or backward-looking analysis, has been applied throughout the outbreak to explore a variety of key questions such as inferring basic epidemiological characteristics like R_0, incubation period, and fatality rate, reveal factors driving transmission, and assess the effectiveness of different interventions[15–17]. Nowcasting focuses on understanding the current situation, like inferring the true number of cases in light of underreporting[18,19]. Prospective modeling 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, while projections span over a much longer future time window, and thus must make assumptions about how the factors driving COVID-19 will change in the future. Scenario analyses produce multiple projections that explore the impacts of different sets of assumptions that vary factors like transmission rates and interventions.

Due to the magnitude of the COVID-19 modeling literature, we had to impose significant constraints on the scope of this review to enable us to conduct a systematic, quantitative, and timely assessment of the relevant literature. Therefore, this work is a narrative review with systematic approach. Specifically, the following inclusion criteria defined our review scope:

1) Prospective modeling work on population-level dynamics of COVID-19: we include papers that provide future predictions for a specific location, including forecasting, projections, and future scenario analysis. We exclude retrospective modeling studies. Papers that only fit a model without providing out-of-sample predictions were not included.
2) Data-driven: we broadly define this as papers that incorporate COVID-19 data into the setup or fitting of the model. Those which only use parameters from the literature or rely on data from other viruses were excluded.
3) Geographic restriction: we only included papers that implement forecasting or projections for US counties, states, or at the national level, which restricts our analysis to papers working with the same data issues and in a similar context.
4) Journal restriction: We include only papers from peer-reviewed journals, as defined by Scopus' context curation standards[20]. In addition, we restrict to papers from journals ranked in the top 10% for their respective field based on the Scopus CiteScore. While we recognize this 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 level and still obtaining the most representative sample of papers possible. For our final sample of peer-

reviewed papers, a table showing the number of papers from each journal and each journal's top category and percentile according to Scopus CiteScore is available in the appendix (p 1).

The Scopus query we developed based on these criteria is also included in the appendix (p 2). To minimize the chance of our search missing relevant papers, we searched PubMed with the equivalent query. *Figure 1* outlines our scoping process and shows the number of papers screened out at each step.

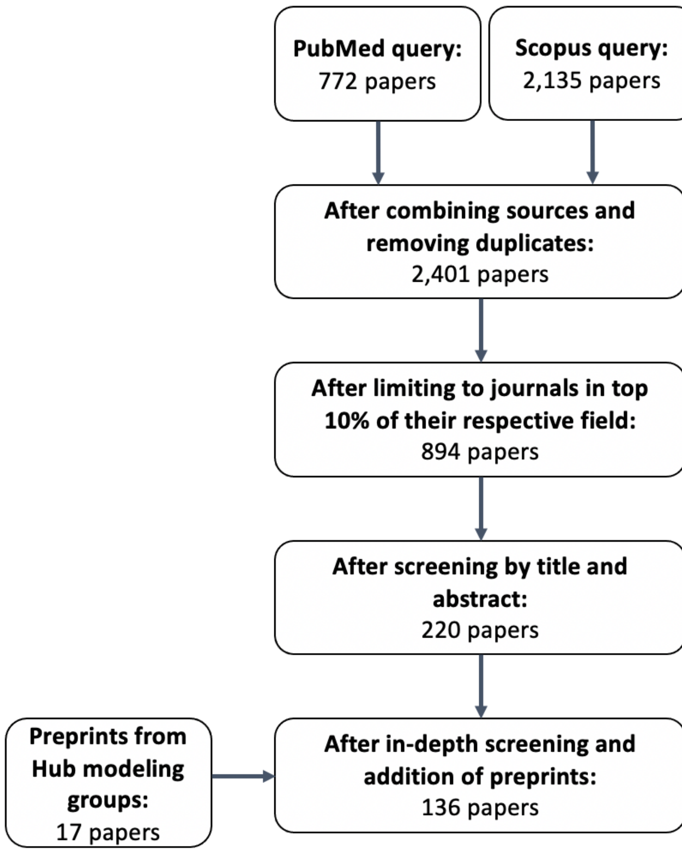

Figure 1. Scoping Process

A schematic of our scoping process and the number of papers left after each step is shown.

The searches of Scopus and PubMed were carried out on August 20, 2021, and our final selection of papers was distributed from March 2020 to August 2021 (*Figure 2*). Notably, the top 10% criteria only reduced the number of papers to 37% of the original size, from 2,401 to 894 papers. Papers were screened individually by KN, SJ, and FP, with a mechanism for double-checking with another individual if a paper's eligibility for inclusion was unclear. For the data collection, categorizations were done individually by the same authors, and all categorizations were double checked with one individual covering all papers for a particular category, so that categorizations were applied consistently across all papers. After screening steps, we narrowed down to 119 peer-reviewed papers.

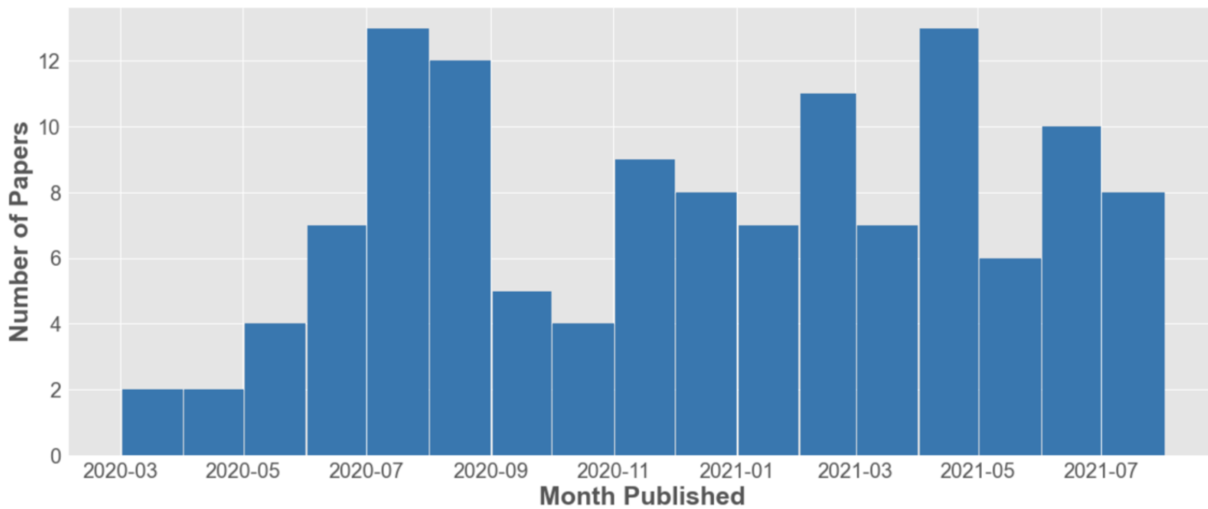

Figure 2. Number of Papers in our Analysis by Month of Publication.
This histogram shows the publication date by month of all papers included in our analysis.

We additionally considered preprints from authors known to be engaged in real-time modeling work. We included preprints from modelers participating in the US COVID-19 Forecast Hub[2]. We also attempted to include the Scenario Modeling Hub[21–23], but no preprints met our criteria for the time window considered. While these papers do not have the validation that comes with peer-review, these models were used in real-time by a national public health agency, which we believe justifies their inclusion in this analysis. We found 17 preprints in the metadata provided by the modeling 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 significant portion of real-time COVID-19 modeling work that exists on preprint servers and on the websites of modeling groups.

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

Role of the funding source
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 full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

To conduct a quantitative analysis on the substance and quality of these studies, for each paper we classified the following features: forecasting window, methodology, prediction target, datasets used, geographic resolution, quantitative uncertainty, performance evaluation, and stated limitations. We acknowledge that some of these categorizations are subjective and/or difficult to consistently extract from papers, especially the performance evaluation and stated limitations category. Thus, we narrowly define our categories and transparently discuss these definitions in this section. While we acknowledge that some of the categorizations we made could be disputed, we are confident that the overall conclusions still hold. The classification of papers for each category is shown in the appendix (p 3). *Figure 3* visualizes the relative size of each category and the most common connections between categories. Each line through the figure represents the categorizations of a single paper, so the thicker the line between two categories, the more often papers tend to fall into both of those categories. The width of the lines are weighted such that in cases where a paper falls into more than one category, like using both cases and deaths data, a line with half of the normal width is assigned to each category.

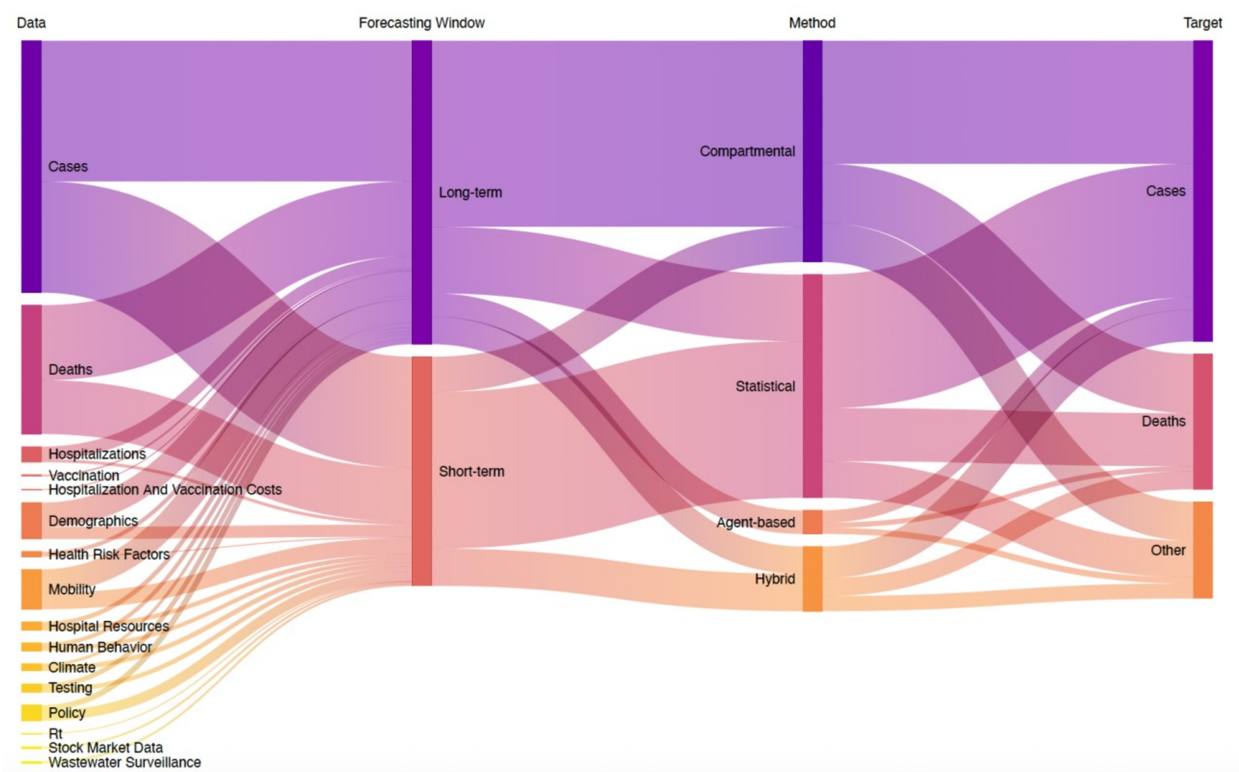

Figure 3. Sankey Diagram of the Connections Between Categorizations of our Analysis.
This diagram shows the relative cooccurrence of categories within papers in our analysis. Thicker lines between categories indicate that those categories are more likely to occur in the same paper.

Model Objective and Prediction Horizon

Forecasts are unconditional in the sense that they attempt to predict what will actually happen in the near future, while *projections* and *scenarios* are conditioned on the model's assumptions about the future in order to extend the prediction horizon. We were unable to reliably categorize models into forecasts or projections due to inconsistent use of these terms and a lack of clear communication in papers on which approach was used. Therefore, as a proxy for model objective, we categorized papers into short-term predictions (namely, forecasts), or long-term predictions (namely, projections). To remain consistent with the COVID-19 Forecast Hub and COVID-19 Scenario Hub, which represent best practice for prospective COVID-19 modeling, we categorize studies making predictions for four weeks or less as short-term (46%), and studies making predictions with a horizon that extends beyond 4 weeks as long-term (60%). There were a small number of papers which produced both long-term and short-term predictions[24–27]. Note that because papers often fall into multiple categories, percentages in this analysis will 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. This could include modeling scenarios with 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 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 modeling beyond classification into three broad categories: compartmental models (SIR and variations), statistical models (machine learning, deep learning, ARIMA, etc.), and hybrid (a combination of compartmental and statistical models). Since most compartmental models in our sample used statistical methods to fit parameters, in order 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 using 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[28]. A model that only uses statistical methods to fit parameters for a compartmental model was classified as compartmental. We found that 47% of papers used a compartmental model, 43% used a statistical model, and 12% used a hybrid model. There were a few papers which showed both a compartmental model and a statistical model[24–27]. We also noted if models used agent-based methods (9%).

Target Variables

The most common target prediction variables were cases (89%), deaths (52%), hospitalizations (10%), and R_t (9%). Some of the lesser used target variables included growth rate, peak cases, and ICU admissions. 38% of papers had only one target variable, 43% of papers had two target variables, and 19% had more than two.

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[29].

Data Categories

Next, we quantified the categories of input data used to inform models. *Table 1* shows how we defined the data categories, including an in-depth look at the datasets used by papers in our analysis that attempt to capture COVID-19 behaviors.

Data Category	**Description**	**Examples**
Cases, Deaths	Epidemiological data on the number of cases or deaths and corresponding metrics.	Daily cases/deaths, cumulative cases/deaths, reproduction number, growth rate

Hospitalizations	Data related to hospitalization of COVID-19 patients.	Daily hospitalizations, active hospitalizations, ICU occupancy, hospital capacity
Testing	Data pertaining to COVID-19 testing in a population or location.	Daily tests, test positivity rate
Climate	Data describing the climate or any meteorological variables pertaining to a specific location, timeseries or static.	Daily precipitation, daily temperature, average temperature
Demographic	Demographic or socio-demographic information about the population of a specific location.	Population, age, race, income, rural/urban ratio
Health Risk Factors	Data which quantifies the health risk factors of the population in the context of COVID-19.	Prevalence of comorbidities, use of preventative services (doctor visits)
Mobility	Data which quantifies the movement of a population.	Google Mobility Trends (residential, grocery & pharmacy stores, parks, retail & recreation, workplaces, transit stations)[30] Unacast social distancing scoreboard (average mobility, nonessential visits, encounters density)[31] SafeGraph (trip counts at a census block group resolution)[32] Apple Mobility Trends (trends in apple maps routing requests)[33] Facebook Movement Range Maps (change in movement compared to baseline percent of population that stays home)[34] Flight data
Human Behavior	Data which quantifies the behavior or beliefs of a population in the context of COVID-19, excluding data on the mobility of a population.	Google search trends[35] Mask use per capita[36] Facebook's COVID-19 Trends and Impact Survey (timeseries of self-reported mask use and other social distancing behaviors)[37] New York Times Mask-Wearing Survey data (static)[38]

		Sentiment index constructed from COVID-19 news[39]
Policy	Data pertaining to policies relating to COVID-19.	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)[40] State Level Social Distancing Policies: tracks dates and details of policies including emergency declarations, gathering restrictions, closures, stay-at-home orders, travel restrictions, isolation orders, and mask mandates [41]

Table 1. Data Categories.

The most frequently used data categories were cases, deaths, mobility, demographics, and hospitalizations (*Table 2*). 20% of papers used only one category of data, 39% of papers used two categories, 16% used three categories, and 25% used four or more categories.

Data Category	**Occurrences**	**Percent of Papers**
Cases	126	93%
Deaths	79	62%
Mobility	34	26%
Demographics	30	25%
Hospitalizations	15	12%
Policy	13	12%
Testing	11	9%
Hospital Resources	10	8%
Climate	8	7%
Human Behavior	8	7%
Health Risk Factors	4	5%

Table 2. Top 10 Data Categories.

The data sources informing predictions in our analysis were dominated by case and death data. Data used in 2 or less papers include vaccinations, R_t, wastewater surveillance, and economic data. 51% of modelers only used epidemiological data sources (cases, deaths, hospitalizations). The most often 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[42–44]. 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. Of particular interest due to the increasing impact of new variants on epidemiological dynamics, none of the papers in our sample utilized variant prevalence data. In the US, this data suffers from low sample size, sampling bias, and is difficult to use as a signal for predictive modeling.

Geographic Resolution

We noted the geographic scale at which predictions were made, categorizing papers as national, state, or county-level and lower. 54% out of 136 of papers included a national level prediction, 36% at the state-level, and 34% at the county-level or smaller scale. Half of the models in our analysis were at the national level, which tends to be the easiest resolution to predict and the least useful for decision-making, which must often occur at the local level.

Uncertainty

We analyzed which papers included a quantitative expression of uncertainty of their predictions, excluding those which only did so for model parameters. We found that half of papers (50% out of 136) did not express any quantitative uncertainty. 49% of papers included some form of confidence or prediction intervals. A sensitivity analysis was performed in 13% of papers.
Half of the papers studied did not express any quantitative uncertainty around the forecasts, despite the highly uncertain and consequential nature of COVID-19 dynamics. The utility of forecasts for decision-makers depends on clear communication of uncertainty[45], 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 determine whether 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 preexisting agenda.

Performance Evaluation

We categorized the type of performance evaluation used for each model. We chose to conduct this analysis only for the subset of papers implementing short-term prediction models, which can be fairly evaluated against truth data. In 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 using standard error metrics is not possible since these models make assumptions about the future that do not match reality.
For timeseries forecasts, the setup of train and test data should be representative of real-time forecasting conditions. Since the utility 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 using data up until a certain cutoff date and evaluated on data after that date. This preserves the fundamental challenge of forecasting: not knowing future data or trends. Within the subset of short-term studies considered, 75% of papers used some sort of performance evaluation metric to compare future-blind, out-of-sample predictions to ground truth data, and 25% did not conduct a performance evaluation on their forecasts. Ground truth data used is usually reported cases or deaths, and sources used in our sample include JHU CSSE[1], The COVID Tracking Project[46], and WHO[47]. The most common metrics to compare predictions to ground truth were mean absolute error, root mean square error,

mean absolute percentage error, R^2, mean square error, and coverage rate of prediction intervals. Out of the papers that did conduct a metric-based evaluation, only 13% evaluated the accuracy of confidence intervals. Within the group of 47 papers which conducted a future-blind performance evaluation, 34% evaluated only one model, 55% compared performance metrics across multiple internal models, and 19% compared the performance metrics of their model against those of other models in the COVID-19 Forecast Hub. 15% of evaluated models used a baseline model for comparison.

Most modelers (75%) quantified the performance of their model relative to truth data, but most did not evaluate their model on predictions made across a timespan that included varying epidemiological dynamics. In order to quantify this, we counted how many dates papers showed predictions from. For example, if a paper presents a model prediction using data up until September 1st and predicts future case counts on the 8th, 15th, 22nd, and 29th, this would be a prediction made from a single date. If this paper adds another prediction made from October 1st (using 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 (September 1st to October 1st). We defined the category this way in order to make sure we could reliably extract this data from each paper. Our analysis found that among short-term models, more than half (55%) only showed a prediction made from a single date, 28% of papers showed predictions made from multiple dates over a timespan that was less than 2 months long, while 17% 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[48]. Therefore, failing to evaluate a model in a variety of epidemiological dynamics severely limits the generalizability of the performance evaluation and the ability to make fair comparisons between models. In addition, one-third of papers (34%) that completed a quantitative performance evaluation did not compare their model to a baseline or any other models, so it is unclear whether the model provides any improvement over a naïve model. The COVID-19 Forecast Hub uses a baseline model that assumes no change in incidence over the next four weeks. According to historical error metrics calculated by the Forecast Hub and CMU Delphi on September 8th, 2021, only 25% of models outperformed the baseline model for cases while 75% outperformed the baseline for deaths by relative mean absolute error and weighted interval score[49]. Thus, comparison to a baseline model provides context that provides important information about the utility 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 review. 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 it was available at the date predictions are made from, which is available through the CMU Delphi API[50,51]. Without thorough performance evaluation, the broader scientific community will be unable to determine which approaches are working and build knowledge on best practices.

Model Limitations

Authors 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 limited generalizability (8%). We define unknowable factors as those that cannot be known at the time predictions were made, like future implementation of non-pharmaceutical interventions or the emergence of new variants during the prediction horizon. In contrast, disregarded factors have some relevant data or information available at the time of the analysis, but the authors choose to disregard it, like the demographic breakdown of populations or healthcare 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 named limitations. In most cases, all of these types of limitations are relevant to COVID-19 models. In addition, our categorization does not give information about how thoroughly these limitations categories were discussed. For COVID-19 applications, clearly stating model limitations is critical to help the public understand the appropriate way to interpret results.

Multidisciplinary Nature of the COVID-19 Literature

The highly consequential nature of the COVID-19 pandemic has attracted modeling 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, multidisciplinary, general physics and astronomy, general mathematics, and statistical and nonlinear physics. Notably, 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, and Fractals*, *PLOS One*, and *Scientific Reports* (*Figure 4*). We were unable to conduct a thorough analysis on the contributions to COVID-19 modeling from different fields due to the difficulty of classifying papers into different disciplines based on their journal and the inherent interdisciplinarity of this work. However, we completed a sub-analysis on the group of papers from Forecast Hub modelers.

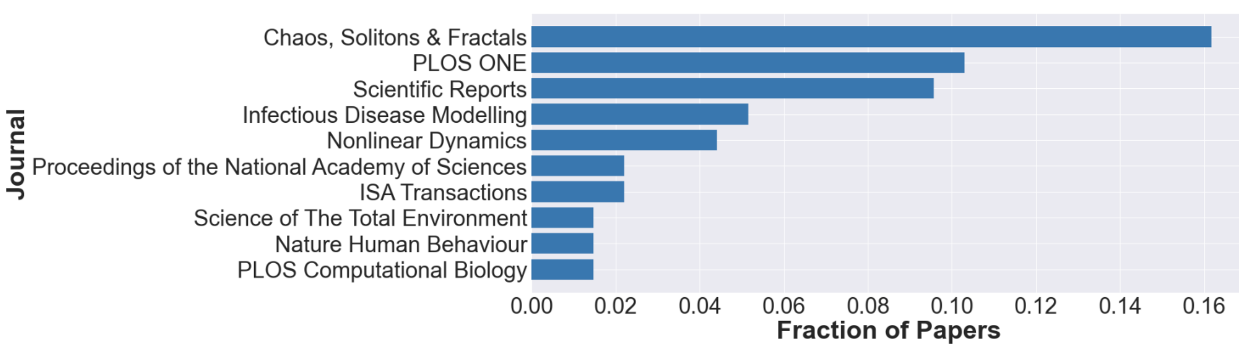

Figure 4. Top 10 Journals in the final set

This graph shows the most frequent journals in our analysis and what fraction of our sample of papers each journal published.

The set of papers written by authors that contributed to the COVID-19 Forecast Hub includes 17 preprints[43,52–67] and 3 papers published in peer-reviewed journals[68–70]. 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, which focuses on one to four week predictions, these preprints are not necessarily on the exact model and application that was submitted to the 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, which is shown in *Table 3*. In addition, Forecast Hub papers were significantly more likely to show and evaluate predictions made from several dates over a timespan greater than 2 months (50% versus 17% for all papers). A significant advantage of the hub approach is that it encourages good practices in terms of uncertainty, evaluation, and high geographic resolution. Additionally, the real-time sharing of forecasts ensures that predictions were truly future-blind.

Categories	**All Papers (N=136)**	**Forecast Hub Papers and Preprints (N=20)**
Prediction Horizon		
Short-term Predictions	46%	70%
Long-term Predictions	60%	40%
Methodology		
Compartmental	47%	35%
Statistical	43%	45%
Hybrid	12%	20%
Agent-based	9%	5%
Geographic Level		
National	54%	25%
State	36%	65%
County or Smaller	34%	55%
Uncertainty		
Expressed Quantitative Uncertainty	50%	55%
Sensitivity Analysis	13%	5%
Performance Evaluation		
Comparison to Ground Truth (out of short-term models only)	75%	86%
Only Made Predictions from One Date	55%	7%
Made Multiple Predictions over a Timespan greater than 2 months	17%	50%
Limitations		
Authors discussed limitations	64%	65%

Table 3. Comparison of All Papers and Forecast Hub Preprints

Discussion

Our analysis found significant gaps in model transparency in the literature, especially on reporting aspects of models that are crucial for 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 the public is paying attention to scientific papers[71,72]. In addition, poor reporting limits the ability of literature reviews to synthesize insights from the research to determine best practices. In response to these concerns, the EPIFORGE 2020 guidelines were developed and recommend consistent terminology, a clear definition of study purpose and model targets, identification of prospective versus retrospective work, comparison to 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 generalizability[73]. Consistent sharing of this information for epidemiological predictions would improve the consistency, reproducibility, comparability, and quality of epidemic forecasting papers, in addition to minimizing the potential for the public to misunderstand or misuse the research.

Another obstacle to maximizing the knowledge gained from epidemic forecasting is the suitability of the information sharing system. Since it is not standard practice for modeling papers to report on translational work, this review can only comment on the translation potential of papers based on their reporting practices, not on how models were actually used during the outbreak. In addition, the volume and variable quality of the literature forced us to adopt stringent and limiting scoping criteria in order to obtain a manageable sample of literature to analyze. Other reviews adopted their own narrow scope, creating a body of COVID-19 modeling literature reviews which amount to a piecemeal, incomplete picture of the efforts of researchers. This illustrates the difficulty of building knowledge from the COVID-19 literature through the traditional information sharing system: peer-reviewed literature and 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

The main limitations of this review are the result of the difficult nature of synthesizing 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 US and to papers published in the top 10% of journals based on Scopus' CiteScore. 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 with this analysis, not the state of all real-time work, some of which is not and may not ever be represented in the literature. In addition, some of the categorizations we made were subjective and/or difficult to extract consistently, so we implemented quality control mechanisms as discussed in the Methods section, and we are confident in our overall conclusions. Despite these limitations, we have studied the most representative sample of papers possible and obtained findings that are informative for improving epidemic modeling in the future.

Conclusion

This analysis examined a subset of the COVID-19 modeling literature, focused on data-driven, prospective modeling, and identified several opportunities to improve the utility of outbreak modeling, which are especially relevant to inform the work of the new CDC Center for Forecasting and Outbreak Analytics. In response to significant scoping challenges, we selected a sample that should represent the best modeling papers and still found them to be substantially lacking 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 literature review are adopting epidemic forecasting standards and creating a suitable information sharing system. Adopting the EPIFORGE 2020 guidelines address many of the issues identified in this review, 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 modelers to be more efficiently synthesized into best practices.

Contributors

LG, KN, and SJ contributed to the conceptualization and design of the study. KN, SJ, and FP collected the data and conducted the analysis. 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 and acquired funding. 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

We declare no competing interest.

Acknowledgements

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). NGR has been supported by the National Institutes of General Medical Sciences (R35GM119582). 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. KG and FP were funded by the SMDM Covid Modeling Accelerator and CDC U01CK000589.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases. 2020 May 1;20(5):533–4.
2. Fraser N, Brierley L, Dey G, Polka JK, Pálfy M, Nanni F, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. PLoS Biology [Internet]. 2021 Apr 2 [cited 2022 Jun 10];19(4). Available from: /pmc/articles/PMC8046348/
3. Horbach SPJM. Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. Quantitative Science Studies [Internet]. 2020 Sep 1 [cited 2022 Jun 10];1(3):1056–67. Available from: https://direct.mit.edu/qss/article/1/3/1056/96126/Pandemic-publishing-Medical-journals-strongly
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: https://doi.org/101177/0272989X21990391 [Internet]. 2021 Feb 3 [cited 2021 Jul 11];41(4):379–85. Available from: https://journals.sagepub.com/doi/full/10.1177/0272989X21990391
5. Press WH, Levin RC. Modeling, post COVID-19. Science (1979) [Internet]. 2020 Nov 27 [cited 2021 Nov 20];370(6520):1015. Available from: https://www.science.org/doi/abs/10.1126/science.abf7914
6. Ioannidis JPA, Cripps S, Tanner MA. Forecasting for COVID-19 has failed. International Journal of Forecasting. 2020 Aug 25;
7. Shankar S, Mohakuda SS, Kumar A, Nazneen PS, Yadav AK, Chatterjee K, et al. Systematic review of predictive mathematical models of COVID-19 epidemic. Medical Journal Armed Forces India. 2021 Jul 1;77:S385–92.
8. Guan J, Wei Y, Zhao Y, Chen F. Modeling the transmission dynamics of COVID-19 epidemic: a systematic review. Journal of Biomedical Research [Internet]. 2020 Nov 1 [cited 2021 Nov 21];34(6):422. Available from: /pmc/articles/PMC7718076/
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 Jan 1;6:324–42.
10. Zawadzki RS, Gong CL, Cho SK, Schnitzer JE, Zawadzki NK, Hay JW, et al. Where Do We Go From Here? A Framework for Using Susceptible-Infectious-Recovered Models for Policy Making in Emerging Infectious Diseases. Value in Health. 2021 Jul 1;24(7):917–24.
11. Adiga A, Dubhashi D, Lewis B, Marathe M, Venkatramanan S, Vullikanti A. Mathematical Models for COVID-19 Pandemic: A Comparative Analysis. Journal of the Indian Institute of Science 2020 100:4 [Internet]. 2020 Oct 30 [cited 2021 Jul 11];100(4):793–807. Available from: https://link.springer.com/article/10.1007/s41745-020-00200-6
12. Rahimi I, Chen F, Gandomi AH. A review on COVID-19 forecasting models. Neural Computing and Applications [Internet]. 2021 Feb 4 [cited 2021 Nov 21];1–11. Available from: https://link.springer.com/article/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 Jan 1;6:258–72.

14. McCabe R, Donnelly CA. Disease transmission and control modelling at the science–policy interface. Interface Focus [Internet]. 2021 Oct 12 [cited 2022 May 26];11(6). Available from: https://royalsocietypublishing.org/doi/full/10.1098/rsfs.2021.0013
15. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science (1979) [Internet]. 2020 Apr 24 [cited 2021 Dec 9];368(6489):395–400. Available from: https://www.science.org/doi/abs/10.1126/science.aba9757
16. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature [Internet]. 2020 Aug 13 [cited 2021 Dec 9];584(7820):257–61. Available from: https://pubmed.ncbi.nlm.nih.gov/32512579/
17. Tian H, Liu Y, Li Y, Wu CH, Chen B, Kraemer MUG, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science (1979) [Internet]. 2020 May 8 [cited 2021 Dec 9];368(6491):638–42. Available from: https://www.science.org/doi/abs/10.1126/science.abb6105
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. The Lancet. 2020 Feb;395(10225):689–97.
19. 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. The Lancet. 2020 Feb 29;395(10225):689–97.
20. Content - How Scopus Works - Scopus - | Elsevier solutions [Internet]. [cited 2022 May 6]. Available from: https://www.elsevier.com/solutions/scopus/how-scopus-works/content
21. Lemaitre JC, Grantz KH, Kaminsky J, Meredith HR, Truelove SA, Lauer SA, et al. A scenario modeling pipeline for COVID-19 emergency planning. Scientific Reports 2021 11:1 [Internet]. 2021 Apr 6 [cited 2021 Jul 24];11(1):1–13. Available from: https://www.nature.com/articles/s41598-021-86811-0
22. Truelove S, Smith CP, Qin M, Mullany LC, Borchering RK, Lessler J, 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. medRxiv [Internet]. 2021 Sep 2 [cited 2022 Apr 9];16:2021.08.28.21262748. Available from: https://www.medrxiv.org/content/10.1101/2021.08.28.21262748v2
23. Borchering RK, Viboud C, Howerton E, Smith CP, Truelove S, Runge MC, et al. Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Nonpharmaceutical Intervention Scenarios — United States, April–September 2021. Morbidity and Mortality Weekly Report [Internet]. 2021 [cited 2022 Apr 9];70(19):719. Available from: /pmc/articles/PMC8118153/
24. 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(2).
25. 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. European Journal of Operational Research. 2021;290(1).
26. Cot C, Cacciapaglia G, Islind AS, Óskarsdóttir M, Sannino F. Impact of US vaccination strategy on COVID-19 wave dynamics. Scientific Reports. 2021;11(1).

27. Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. Proceedings of the National Academy of Sciences [Internet]. 2020 Jul 21 [cited 2021 Nov 2];117(29):16732–8. Available from: https://www.pnas.org/content/117/29/16732
28. Bhouri MA, Costabal FS, Wang H, Linka K, Peirlinck M, Kuhl E, et al. COVID-19 dynamics across the US: A deep learning study of human mobility and social behavior. Computer Methods in Applied Mechanics and Engineering. 2021;382.
29. On the predictability of COVID-19 - International Institute of Forecasters [Internet]. [cited 2021 Dec 6]. Available from: https://forecasters.org/blog/2021/09/28/on-the-predictability-of-covid-19/
30. COVID-19 Community Mobility Reports [Internet]. [cited 2021 Dec 14]. Available from: https://www.google.com/covid19/mobility/
31. Covid-19 Social Distancing Scoreboard — Unacast [Internet]. [cited 2021 Dec 14]. Available from: https://www.unacast.com/covid19/social-distancing-scoreboard
32. SafeGraph | Academics [Internet]. [cited 2021 Dec 14]. Available from: https://www.safegraph.com/academics
33. COVID-19 - Mobility Trends Reports - Apple [Internet]. [cited 2021 Dec 14]. Available from: https://covid19.apple.com/mobility
34. Facebook Data For Good Movement Range Maps [Internet]. [cited 2021 Dec 14]. Available from: https://dataforgood.facebook.com/dfg/tools/movement-range-maps
35. Coronavirus Search Trends - Google Trends [Internet]. [cited 2021 Dec 14]. Available from: https://trends.google.com/trends/story/GB_cu_JSW_pHABAADqAM_en
36. COVID-19 [Internet]. [cited 2021 Dec 14]. Available from: https://covid19.healthdata.org/global?view=mask-use&tab=trend
37. Facebook Data For Good COVID 19 Symptom Survey [Internet]. [cited 2021 Dec 14]. Available from: https://dataforgood.facebook.com/dfg/tools/covid-19-trends-and-impact-survey#methodology
38. covid-19-data/mask-use at master · nytimes/covid-19-data · GitHub [Internet]. [cited 2021 Dec 14]. Available from: https://github.com/nytimes/covid-19-data/tree/master/mask-use
39. Chalkiadakis I, Yan H, Peters GW, Shevchenko P v. Infection rate models for COVID-19: Model risk and public health news sentiment exposure adjustments. PLoS ONE. 2021;16(6 June).
40. Hale T, Angrist N, Goldszmidt R, Kira B, Petherick A, Phillips T, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nature Human Behaviour 2021 5:4 [Internet]. 2021 Mar 8 [cited 2021 Dec 9];5(4):529–38. Available from: https://www.nature.com/articles/s41562-021-01079-8
41. Adolph C, Amano K, Bang-Jensen B, Fullman N, Wilkerson J. Pandemic Politics: Timing State-Level Social Distancing Responses to COVID-19. Journal of Health Politics, Policy and Law. 2021 Apr 1;46(2):211–33.
42. 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 U S A. 2020;117(33).
43. Arik SO, Li CL, Yoon J, Sinha R, Epshteyn A, Le LT, et al. Interpretable Sequence Learning for COVID-19 Forecasting. ArXiv [Internet]. 2020 Aug 3 [cited 2022 Apr 3]; Available from: http://arxiv.org/abs/2008.00646
44. Lee SY, Lei B, Mallick B. Estimation of COVID-19 spread curves integrating global data and borrowing information. PLoS ONE. 2020;15(7 July).

45. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease forecasting to public health: A path forward using influenza forecasting examples. BMC Public Health [Internet]. 2019 Dec 10 [cited 2021 Dec 8];19(1):1–12. Available from: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-7966-8
46. The COVID Tracking Project | The COVID Tracking Project [Internet]. [cited 2022 Jun 1]. Available from: https://covidtracking.com/
47. WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data [Internet]. [cited 2022 Jun 1]. Available from: https://covid19.who.int/
48. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences [Internet]. 2022 Apr 12 [cited 2022 Apr 9];119(15). Available from: https://pnas.org/doi/full/10.1073/pnas.2113561119
49. Home - COVID 19 forecast hub [Internet]. [cited 2021 Dec 7]. Available from: https://covid19forecasthub.org/eval-reports/?state=US&week=2021-09-08
50. Reinhart A, Brooks L, Jahja M, Rumack A, Tang J, Agrawal S, et al. An open repository of real-time COVID-19 indicators. Proc Natl Acad Sci U S A. 2021 Dec 21;118(51).
51. COVIDcast Dashboard | DELPHI [Internet]. [cited 2022 Jun 12]. Available from: https://delphi.cmu.edu/covidcast/
52. 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 [Internet]. 2020 May 25 [cited 2021 Sep 19];2020.05.24.20111989. Available from: https://www.medrxiv.org/content/10.1101/2020.05.24.20111989v1
53. Khan ZS, van Bussel F, Hussain F. A predictive model for Covid-19 spread applied to eight US states. 2020 Jun 10 [cited 2021 Sep 20]; Available from: http://arxiv.org/abs/2006.05955
54. Galasso J, Cao DM, Hochberg R. A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. medRxiv [Internet]. 2021 Sep 14 [cited 2021 Sep 20];2021.05.23.21257689. Available from: https://www.medrxiv.org/content/10.1101/2021.05.23.21257689v2
55. Zhang-James Y, Hess J, Salekin A, Wang D, Chen S, Winkelstein P, et al. A seq2seq model to forecast the COVID-19 cases, deaths and reproductive R numbers in US counties. medRxiv [Internet]. 2021 Apr 20 [cited 2021 Sep 20];2021.04.14.21255507. Available from: https://www.medrxiv.org/content/10.1101/2021.04.14.21255507v1
56. Shi Y, Ban X. Capping Mobility to Control COVID-19: A Collision-based Infectious Disease Transmission Model. medRxiv [Internet]. 2020 Jul 28 [cited 2021 Sep 20];2020.07.25.20162016. Available from: https://www.medrxiv.org/content/10.1101/2020.07.25.20162016v1
57. Wu D, Gao L, Xiong X, Chinazzi M, Vespignani A, Ma YA, et al. DeepGLEAM: A hybrid mechanistic and deep learning model for COVID-19 forecasting. 2021 Feb 12 [cited 2021 Sep 20]; Available from: https://arxiv.org/abs/2102.06684v3
58. Srivastava A, Xu T, Prasanna VK. Fast and Accurate Forecasting of COVID-19 Deaths Using the SIkJalpha Model. 2020 Jul 10 [cited 2021 Sep 20]; Available from: https://arxiv.org/abs/2007.05180v2

59. team IC 19 health service utilization forecasting, Murray CJ. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. medRxiv [Internet]. 2020 Mar 30 [cited 2021 Sep 20];2020.03.27.20043752. Available from: https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1

60. Pei S, Shaman J. Initial Simulation of SARS-CoV2 Spread and Intervention Effects in the Continental US. medRxiv [Internet]. 2020 Mar 27 [cited 2021 Sep 20];2020.03.21.20040303. Available from: https://www.medrxiv.org/content/10.1101/2020.03.21.20040303v2

61. Gibson GC, Reich NG, Sheldon D. Real-time Mechanistic Bayesian Forecasts of COVID-19 Mortality. medRxiv [Internet]. 2020 [cited 2021 Sep 20]; Available from: /pmc/articles/PMC7781348/

62. Wang L, Wang G, Gao L, Li X, Yu S, Kim M, et al. Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States. 2020 Apr 29 [cited 2021 Sep 20]; Available from: https://arxiv.org/abs/2004.14103v4

63. Neural Relational Autoregression for High-Resolution COVID-19 Forecasting [Internet]. [cited 2021 Sep 20]. Available from: https://ai.facebook.com/research/publications/neural-relational-autoregression-for-high-resolution-covid-19-forecasting/

64. Biegel HR, Lega J. EpiCovDA: a mechanistic COVID-19 forecasting model with data assimilation. ArXiv. 2021;

65. Wilson DJ. Weather, Social Distancing, and the Spread of COVID-19. Federal Reserve Bank of San Francisco, Working Paper Series. 2020 Jul 2;01–64.

66. Baxter A, Oruc BE, Keskinocak P, Asplund J, Serban N. Evaluating Scenarios for School Reopening under COVID19. medRxiv [Internet]. 2020 Jul 24 [cited 2021 Dec 14];2020.07.22.20160036. Available from: https://www.medrxiv.org/content/10.1101/2020.07.22.20160036v1

67. Rodriguez A, Tabassum A, Cui J, Xie J, Ho J, Agarwal P, et al. DeepCOVID: An Operational Deep Learning-driven Framework for Explainable Real-time COVID-19 Forecasting. medRxiv [Internet]. 2020 Sep 29 [cited 2021 Nov 2];2020.09.28.20203109. Available from: https://www.medrxiv.org/content/10.1101/2020.09.28.20203109v1

68. Pei S, Kandula S, Shaman J. Differential effects of intervention timing on COVID-19 spread in the United States. Science Advances. 2020 Dec 4;6(49).

69. Rowland MA, Swannack TM, Mayo ML, Parno M, Farthing M, Dettwiller I, et al. COVID-19 infection data encode a dynamic reproduction number in response to policy decisions with secondary wave implications. Scientific Reports 2021 11:1 [Internet]. 2021 May 25 [cited 2021 Jul 11];11(1):1–7. Available from: https://www.nature.com/articles/s41598-021-90227-1

70. Gao J, Sharma R, Qian C, Glass LM, Spaeder J, Romberg J, et al. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. Journal of the American Medical Informatics Association [Internet]. 2021 Mar 18 [cited 2021 Nov 2];28(4):733–43. Available from: https://academic.oup.com/jamia/article/28/4/733/6118380

71. Bagdasarian N, Cross GB, Fisher D. Rapid publications risk the integrity of science in the era of COVID-19. BMC Medicine [Internet]. 2020 Jun 25 [cited 2022 Jun 11];18(1):1–5. Available from: https://link.springer.com/articles/10.1186/s12916-020-01650-6

72. Fraser N, Brierley L, Dey G, Polka JK, Pálfy M, Nanni F, et al. The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science

communication landscape. PLoS Biology [Internet]. 2021 Apr 2 [cited 2022 Jun 5];19(4). Available from: /pmc/articles/PMC8046348/

73. Pollett S, Johansson MA, Reich NG, Brett-Major D, del Valle SY, Venkatramanan S, et al. Recommended reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines. PLOS Medicine [Internet]. 2021 Oct 1 [cited 2021 Dec 6];18(10):e1003793. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003793