PNAS RESEARCH ARTICLE | POPULATION BIOLOGY STATISTICS OPEN ACCESS

Comparative evaluation of behavioral epidemic models using COVID-19 data

Nicolò Gozzi[a,b], Nicola Perra[b,c,d], and Alessandro Vespignani[a,b,1]

Edited by Bernice Pescosolido, Indiana University, Bloomington, IN; received October 24, 2024; accepted May 8, 2025

Characterizing the feedback linking human behavior and the transmission of infectious diseases (i.e., behavioral changes) remains a significant challenge in computational and mathematical epidemiology. Existing behavioral epidemic models often lack real-world data calibration and cross-model performance evaluation in both retrospective analysis and forecasting. In this study, we systematically compare the performance of three mechanistic behavioral epidemic models across nine geographies and two modeling tasks during the first wave of COVID-19, using various metrics. The first model, a Data-Driven Behavioral Feedback Model, incorporates behavioral changes by leveraging mobility data to capture variations in contact patterns. The second and third models are Analytical Behavioral Feedback Models, which simulate the feedback loop either through the explicit representation of different behavioral compartments within the population or by utilizing an effective nonlinear force of infection. Our results do not identify a single best model overall, as performance varies based on factors such as data availability, data quality, and the choice of performance metrics. While the Data-Driven Behavioral Feedback Model incorporates substantial real-time behavioral information, the Analytical Compartmental Behavioral Feedback Model often demonstrates superior or equivalent performance in both retrospective fitting and out-of-sample forecasts. Overall, our work offers guidance for future approaches and methodologies to better integrate behavioral changes into the modeling and projection of epidemic dynamics.

COVID-19 | behavioral epidemic models | behavioral changes | epidemiology

Significance

Modeling the interplay between human behavior and infectious disease transmission remains one of the key challenges in Epidemiology. In this study, we evaluate the performance of three mechanistic behavioral epidemic models designed to address this issue. We compare data-driven and analytical approaches across the first COVID-19 wave, spanning nine diverse locations and two modeling tasks. While the optimal model may vary depending on factors such as data availability and geography, our findings show that approaches explicitly modeling behavioral feedback mechanisms often outperform data-driven approaches, even when considering data quality and the increased numbers of free parameters of these models.

During the COVID-19 Pandemic, epidemic models became central tools for providing situational awareness, scenario analysis, and forecasts. Specifically, mathematical and computational models were used to characterize the initial outbreak phases (1–6), assess policy interventions (6–12), evaluate risks from new virus strains (13–17), and estimate outcomes of various vaccination strategies (18–23). Achieving the required level of realism in these models necessitated incorporating population-level behavioral changes due to epidemic awareness and mandated or recommended nonpharmaceutical interventions (NPIs) (24–26). However, capturing the feedback loop between the transmission of infectious diseases and human behavior has long been regarded, and still remains, as a major challenge in Epidemiology (24, 27–30).

In this context, we identify two major classes of mechanistic modeling approaches: Data-Driven Behavioral Feedback Models and Analytical Behavioral Feedback Models. Data-Driven Behavioral Feedback Models integrate real-world data on behavioral changes, such as mobility patterns, social distancing measures, or other proxies into epidemic simulations (13, 20, 30–48). Hence, these models rely on empirical data to simulate how behaviors change. Analytical Behavioral Feedback Models, instead, use theoretical frameworks to incorporate nonlinear mechanisms describing how individual behaviors change in response to the epidemic's progression (49–56). These models do not rely on real-world data but rather on mechanistic rules that capture the feedback loop between behavior and epidemic dynamics.

Data-driven approaches have been prevalent in the COVID-19 literature for several reasons. First, using empirical data can drastically reduce the number of free parameters and explicit mechanisms needed to capture human behavior. Additionally, most models in the analytic class were developed before the COVID-19 pandemic and often lacked empirical validation (27). Using an explicit behavioral model rather than data, however, has the potential to accurately capture the interplay between human behavior and the spread of infectious diseases, enabling more precise projections and forecasts. Furthermore, data-driven models are not necessarily simpler than their analytical counterparts. In fact, they often integrate large amounts of temporal data, rely on methodologies that involve several assumptions, and may be prone to biases as well as other data collection

Author affiliations: [a]Institute for Scientific Interchange Foundation, Turin 10126, Italy; [b]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA 02115; [c]Centre for Complex Systems, School of Mathematical Sciences, Queen Mary University, London E1 4NS, United Kingdom; and [d]The Alan Turing Institute, London NW1 2DB, United Kingdom

Author contributions: N.G., N.P., and A.V. designed research; N.G., N.P., and A.V. performed research; N.G. analyzed data; and N.G., N.P., and A.V. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: a.vespignani@northeastern.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2421993122/-/DCSupplemental.

Published June 12, 2025.

issues. These considerations stress the need for a systematic analysis of the performance and calibration of these model classes to understand their reliability and usefulness in informing decision-making processes.

Here, we present a systematic comparison of the performance of different behavioral feedback models during the first wave of the COVID-19 pandemic, spanning nine geographies and two modeling tasks. Specifically, we investigate three mechanistic models: i) the Data-Driven Behavioral Model exemplifies data-driven approaches, leveraging mobility data to estimate effective changes in contact patterns; ii) the Compartmental Behavioral Feedback Model simulates the feedback loop by explicitly representing different behavioral classes within the population; iii) the Effective Force of Infection Behavioral Feedback Model employs an effective nonlinear forcing to adjust the infection rate based on the epidemic's progression and the resulting behavioral changes. We first quantitatively assess the performance of these models through a retrospective analysis of their ability to capture the dynamics of the first COVID-19 wave across nine diverse geographical areas. Using the same datasets, we also evaluate their out-of-sample forecasting performance over a rolling four-week horizon in these regions. Remarkably, in the retrospective analysis, the Data-Driven Behavioral Model, which integrates mobility data, does not consistently outperform the Analytical Behavioral Feedback Models. In forecasting performance, similar results are observed, indicating that Analytical Behavioral Feedback Models, though largely neglected during the COVID-19 pandemic, can often provide superior, or comparable, performance by capturing the interplay between human behavior changes and disease progression.

Overall, our results highlight how data-driven behavioral models might not necessarily lead to the best performance. On the contrary, our findings pave the way for a broader use of analytical behavioral approaches in epidemic modeling and forecasting contexts. Modeling choices should therefore consider factors such as data availability and quality, target metrics, and geographic scope. It is important to note that, despite similar performances in retrospective and forecasting analyses, the models often offer different characterizations of disease dynamics, exemplified by different effective reproductive numbers time series. This evidence reinforces the indication of interpreting disease dynamics in the context of the model's structure rather than as intrinsic properties of the pathogen. While our results focus on COVID-19, they hold broad relevance for the analysis and forecasting of respiratory and other transmissible diseases.

Results

We consider three mechanistic models in which the disease progression is described via an age-structured Susceptible-Exposed-Infected-Recovered (SEIR) disease dynamic with the addition of compartments accounting for COVID-19 deaths and their delayed reporting (Fig. 1). Each model differs in the way population behavioral changes are integrated into the dynamic:

- Data-Driven Behavioral (DDB) model. This model integrates data from the COVID-19 Community Mobility Report published by Google LLC (57) to derive a time-varying contact reduction coefficient.
- Compartmental Behavioral Feedback (CBF) model. This model introduces behavioral changes explicitly through a new compartment S^B for susceptible individuals who are risk averse. These individuals experience a relative reduction $r < 1$ in the force of infection. Additional parameters of the model characterize the transitions to and out of the risk-averse compartment.
- Effective Force of Infection Behavioral Feedback (EFB) model. This model integrates the behavior changes in the population with an explicit modulation of the transmissibility (55). More precisely, we consider a nonlinear function (that saturates as the number of reported deaths grows) characterizing the rate at which susceptible individuals acquire infection (i.e., the force of infection).

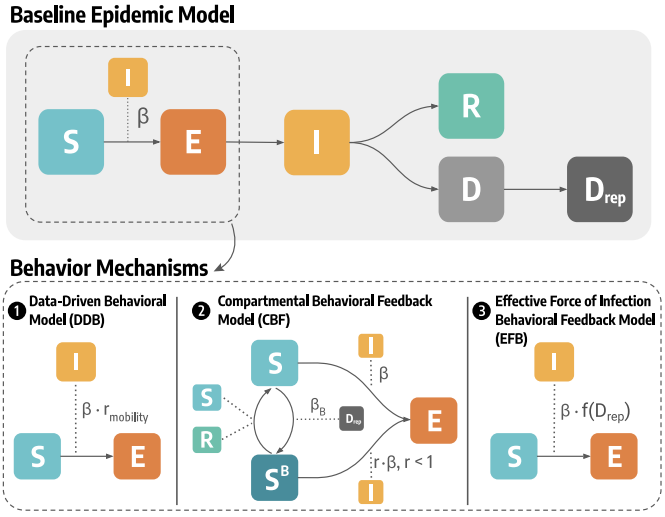

Fig. 1. Flow diagrams of the behavioral epidemic models considered. The *Top* row shows the baseline SEIR-like epidemic model. The *Bottom* row shows the epidemic-behavior mechanisms for the three models considered. In particular, β indicates the transmission rate, r_{mobility} is the contact reduction parameter estimated on mobility data, β_B regulates the behavioral transitions in the Compartmental Behavioral Feedback model and r is the relative reduction in the force of infection of risk-averse individuals, and $f(D_{rep})$ is a nonlinear function of the number of reported deaths that modulates the force of infection in the Effective Force of Infection Behavioral Feedback model.

In Fig. 1, we show a schematic depiction of the compartmental structure and the transitions among compartments of each model. Full details of the models are provided in *Materials and Methods* and *SI Appendix*. It is important to note that none of the models distinguishes between spontaneous and mandated behavioral changes (24, 27–29). In all cases, increased risk aversion among individuals and the resulting reduction in contacts account for all causes leading to behavioral changes.

Retrospective Model Inference. We calibrated the three models to fit the initial wave of COVID-19 deaths across nine distinct geographical areas: metropolitan areas such as Bogotá, Chicago, Jakarta, London, Madrid, New York, Rio de Janeiro, and Santiago de Chile, as well as a larger administrative region, such as Gauteng in South Africa. This selection captures a diverse range of epidemiological, sociodemographic, and socioeconomic contexts from both the global North and South. In each region, models are calibrated to weekly deaths using an Approximate Bayesian Computation–sequential Monte Carlo (ABC-SMC) algorithm (58) (details are reported in *Materials and Methods* and *SI Appendix*). In Fig. 2, we present the fitted curves (median and 90% predictive intervals) for these nine locations using the three models. Overall, all models successfully replicate the shape of the observed epidemic curves. However, we observe inferior fit quality in specific cases. For instance, the models exhibit lower performance in the epidemic tail for Gauteng and Rio de Janeiro,

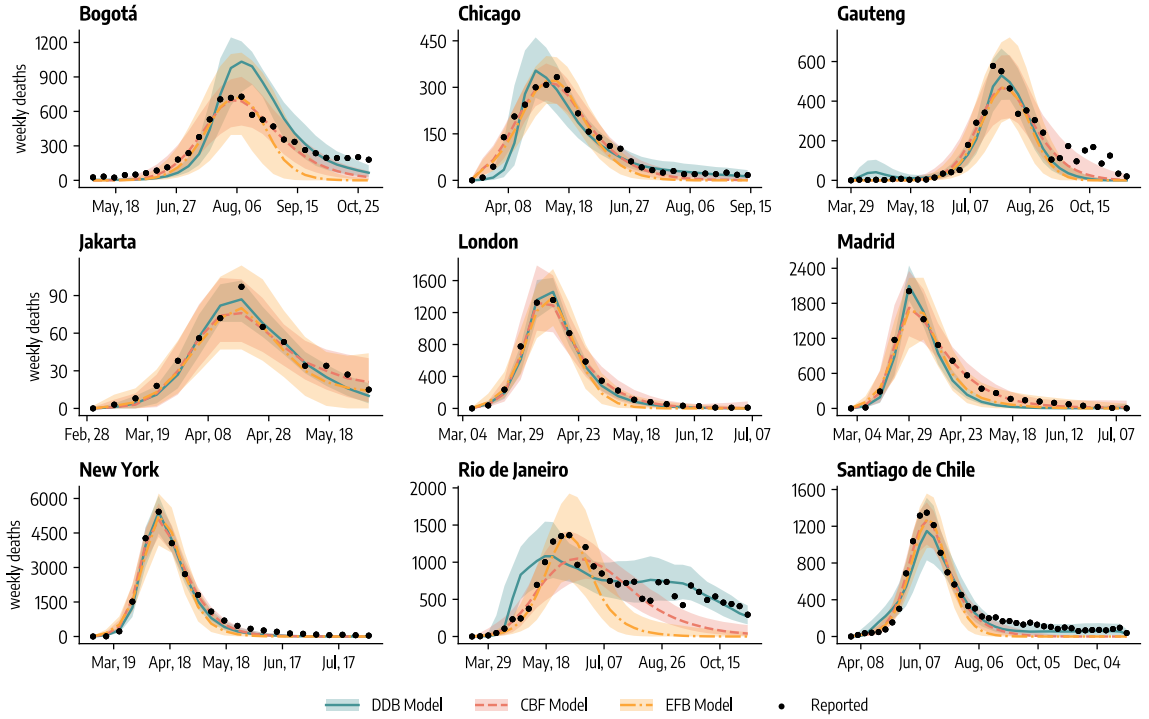

Fig. 2. Fitted curves (median, 90% predictive intervals obtained from 1,000 stochastic trajectories) of weekly deaths during the COVID-19 initial wave across the nine geographies considered and three behavioral epidemic models. DDB stands for Data-Driven Behavioral model, CBF for Compartmental Behavioral Feedback model, and EFB for Effective Force of Infection Behavioral Feedback model.

possibly due to factors such as the completeness and reporting time of epidemiological data in those settings. To quantitatively characterize the model's performance, in Table 1, we report the normalized mean absolute error (nMAE), the normalized weighted interval score (nWIS), and the Bayesian Information Criterion (BIC) weight of each model.

When considering the nMAE of the median, the Compartmental Behavioral Feedback model is the top performer in all geographies, except for Rio de Janeiro and Santiago de Chile, where the Data-Driven Behavioral model performs better. The WIS measures the effectiveness of predictive intervals in bounding reported data, with normalization enabling comparisons across different geographies. Analyzing the nWIS, the Compartmental Behavioral Feedback model is the top performing model in 6 geographies (Bogotá, Chicago, Gauteng, London, Madrid, and New York), although there is more variability in performance. The Data-Driven Behavioral model is the top performer in 3 geographies (Jakarta, Rio de Janeiro, and Santiago de Chile).

It is important to note that the models have different structures and number of free parameters that obfuscate the simple comparison through goodness of fit. To have a more unbiased estimator, we consider the Bayesian Information Criterion that discounts the number of estimable parameters and calculate the BIC weights of each model in each location. The BIC weight can be interpreted as the probability that any given model is the best model (i.e., likelihood of the model given the data) among those considered. According to the BIC weights, the Compartmental Behavioral Feedback model is the most probable in 4 cases (Bogotá, Chicago, London, and Madrid), while the

Table 1. Comparison of models performance in retrospective modeling task (indicates probabilities <0.01%)**

	nMAE			nWIS			BIC weights		
	DDB	CBF	EFB	DDB	CBF	EFB	DDB	CBF	EFB
Bogotá	0.41	**0.18**	0.35	0.29	**0.14**	0.27	**	**99.99**	**
Chicago	0.20	**0.14**	0.20	0.12	**0.09**	0.14	**	**99.90**	0.09
Gauteng	0.32	**0.25**	0.34	0.25	**0.18**	0.26	**94.20**	5.73	0.07
Jakarta	0.14	**0.13**	0.14	**0.09**	**0.09**	0.11	**97.92**	0.35	1.73
London	0.12	**0.07**	0.18	0.07	**0.06**	0.11	4.57	**95.42**	**
Madrid	0.28	**0.10**	0.24	0.21	**0.08**	0.14	**	**99.99**	**
New York	**0.13**	**0.13**	0.21	0.09	**0.08**	0.15	**98.97**	1.02	**
Rio de Janeiro	**0.23**	0.33	0.52	**0.16**	0.25	0.43	**99.99**	**	**
Santiago de Chile	**0.24**	0.25	0.32	**0.14**	0.19	0.26	**63.80**	36.17	0.03

DDB stands for Data-Driven Behavioral model, CBF for Compartmental Behavioral Feedback model, and EFB for Effective Force of Infection Behavioral Feedback model. Bold texts indicate the best performance based on values rounded to two decimal places; when multiple models share the same rounded value, all are highlighted. In the text, comparisons among models are based on full-precision values for greater accuracy.

Data-Driven Behavioral model is most probably the best model in the remaining 5 (Gauteng, Jakarta, New York, Rio de Janeiro, and Santiago de Chile). We note how, in Data-Driven Behavioral models, the contact reduction values estimated on mobility data [i.e., $r_{\text{mobility}}(t)$] are discounted from the number of free parameters as they are not subject to calibration. In *SI Appendix*, we also report the models' accuracy in reproducing both the intensity and timing of epidemic peaks. Interestingly, we find that the Effective Force of Infection model and the Data-Driven model are more accurate than the Behavioral Compartmental model when evaluating these target quantities. This highlights that the definition of the best model often depends on the specific metric being evaluated. Another approach to model comparison is the use of Bayes Factors. While BIC explicitly penalizes model complexity, Bayes Factors inherently favor simpler models by integrating over the parameter space. However, this effect is highly sensitive to prior specifications for each model. For completeness, in *SI Appendix* we present the results of an alternative ABC-SMC calibration scheme that directly estimates Bayes Factors for model comparison (59), assuming a flat prior over the three models. The results indicate that the CBF model is favored in most locations, further strengthening the statistical evidence for the competitive performance of analytical behavioral epidemic models.

Estimates of Transmission Potential. Although the three models offer a similar fit of the epidemic trajectories, it is important to quantify the difference in the disease dynamic emerging from them. First, we considered the posterior distribution of the basic reproductive number (i.e., R_0) provided by the three models across the nine geographies. This quantity is defined as the number of secondary infections, due to a single infectious individual, in an otherwise susceptible population (60). In all three models, R_0 is defined as $R_0 = \rho(\tilde{\mathbf{C}})\beta/\mu$, where $\tilde{C}$ is the contact matrix weighted by age group populations, $\rho(\cdot)$ is the spectral radius, β is the transmission rate, and μ is the inverse of the infectious period (61). Notably, in each behavioral model, R_0 remains the same as in a SEIR model. Indeed, in a fully susceptible population, such as during the epidemic's early phases, behavioral effects are negligible.

As shown in Fig. 3, we observe comparable posterior distributions in some instances, as in the case of London, where the three models estimate median and 90% CI for R_0 values of 2.54 [2.29, 2.90] (DDB), 2.77 [2.43, 3.20] (CBF), and 2.38 [2.09, 2.68] (EFB). The posterior distributions of the three models are similar (to different extents) also for Jakarta, Madrid, and New York. However, we also find significant variations. For instance, in Santiago de Chile, the Data-Driven Behavioral model estimates a R_0 of 4.20 [3.96, 4.42], whereas the estimates from the Compartmental and the Effective Force of Infection Behavioral Feedback models are notably lower, at 1.87 [1.69, 2.04] and 1.77 [1.59, 1.97], respectively. To quantify the similarity among R_0 distributions, we employ the Wasserstein distance. On average, we find that the distributions projected by the two Analytical Behavioral Feedback Models (CBF and EFB) are closer to each other compared to the distribution projected by the Data-Driven Behavioral model, which tends to exhibit greater dissimilarity. We refer the reader to *SI Appendix* for full details on R_0 distributions and related analysis.

While R_0 is of epidemiological significance, the effective reproductive number R_t emerged as a crucial and closely monitored metric during the COVID-19 pandemic. Unlike R_0, R_t accounts for fluctuations in transmissibility attributed to seasonality,

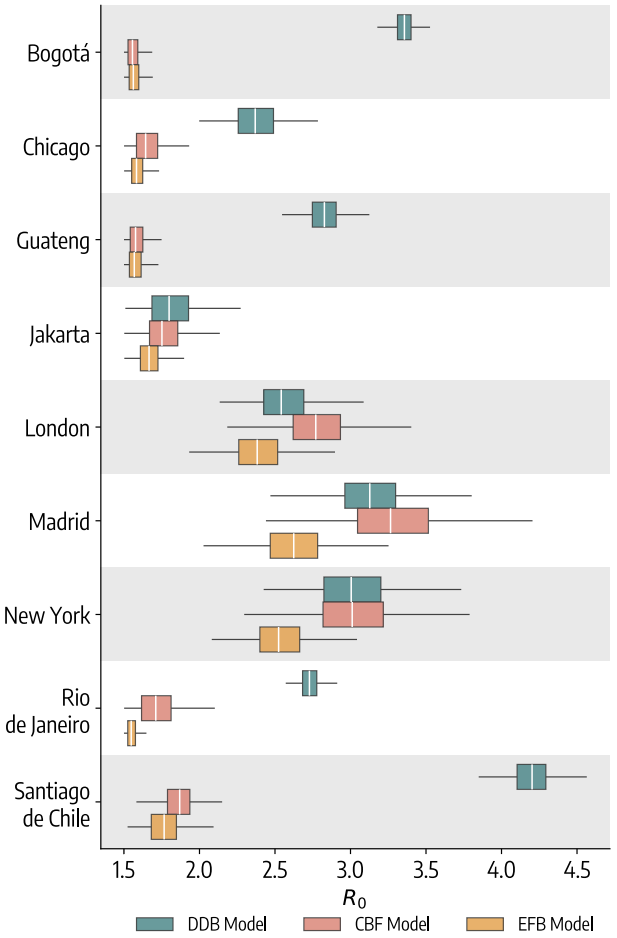

Fig. 3. Boxplot of R_0 posterior distributions according to the three models considered across the nine geographies, considering 1,000 posterior samples. The box boundaries represent the interquartile range (IQR) between the first and third quartiles (Q1 and Q3), the line inside the box indicates the median and the *Upper* (*Lower*) whisker extends to the last datum less (greater) than $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$). DDB stands for Data-Driven Behavioral model, CBF for Compartmental Behavioral Feedback model, and EFB for Effective Force of Infection Behavioral Feedback model.

changes in susceptibility, and also behavioral changes. For this reason, in Fig. 4, we compare the R_t estimated using the method described in ref. 62 from the three models' data. Generally, we observe a consistent trend in their evolution. Noteworthy is the close alignment of tipping points (i.e., instances where R_t crosses 1) projected by all three models. However, deviations are evident in the cases of Bogotá and Gauteng, where the Data-Driven Behavioral model predicts an early tipping point in March/April 2020. This divergence can be attributed to decreased mobility in those regions during the early months of 2020, prompted by global emergency measures, despite a subsequent rise in cases and deaths. This underscores some limitations of mobility data in accurately estimating the impact of NPIs (*Discussion*). As an additional analysis, we calculate pairwise correlations between one-step changes in R_t as estimated by the three models (shown in the *Inset* of Fig. 4). We generally observe positive and statistically significant correlations, with a few exceptions. For instance, the evolution of R_t in the Data-Driven Behavioral model does not show a significant correlation with the corresponding quantity estimated by the Compartmental Behavioral Feedback model and the Effective Force of Infection Behavioral Feedback model in Gauteng. Similarly, the R_t of the Data-Driven Behavioral model is not significantly correlated with that of the Compartmental

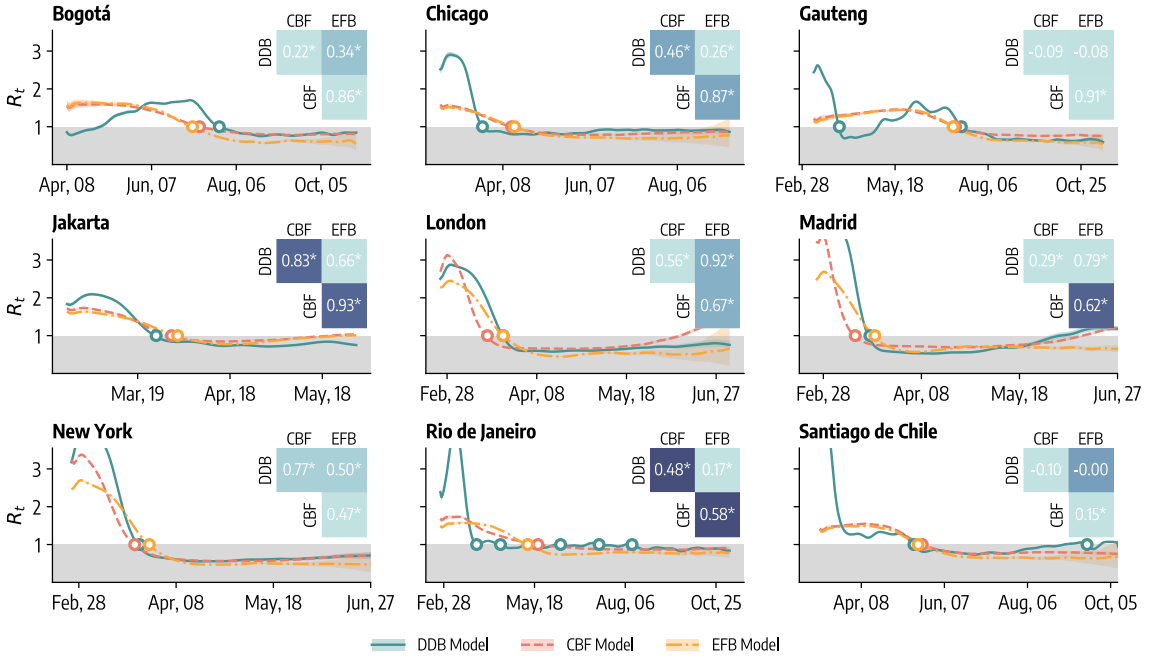

Fig. 4. Effective reproductive number (R_t) analysis. Effective reproductive number (median, 90% predictive intervals) projected by three models across nine geographical regions considered. The shaded grey area indicates $R_t < 1$, while dots indicate points in time where R_t went below 1. In the *Inset* of each figure, we show pairwise Pearson correlation coefficients between one-step changes in R_t as estimated by the three models. Asterisks indicate correlations significant at the 5% significance level. DDB stands for Data-Driven Behavioral model, CBF for Compartmental Behavioral Feedback model, and EFB for Effective Force of Infection Behavioral Feedback model.

Behavioral and of the Effective Force of Infection Behavioral Feedback model in Santiago de Chile. Similar to the findings regarding R_0 distributions, we note that the evolution of R_t projected by the two Analytical Behavioral Feedback Models generally shows higher correlations.

Overall, our results highlight an important point. Differences in the estimated values of R_0 and R_t across models are influenced by approaches used to describe the force of infection. Hence, they are affected by different underlying assumptions and approximations. Ensemble approaches, which average across models, might be used to provide more reliable estimations of the real value of such quantities (63). While these methods are typically used in out-of-sample forecasts, they have been also used to provide in-sample consensus estimates across models for R_t (64).

Forecasting Performance. As a final step in comparing the three models, we use them to forecast the number of weekly deaths during the first wave in the nine geographies under consideration. Specifically, we calibrate each model up to time t, forecast the subsequent four weeks, then shift the calibration window to $t+1$ and repeat the process. We assess forecasting performance using two metrics: the Weighted Interval Score (WIS) and the mean absolute error (MAE) of the median. Normalization of metrics is not required in this context, as forecasting performance is assessed relative to a baseline model, as explained below. For simplicity, we present analysis concerning the WIS as a performance metric in the main text, while MAE results are reported in *SI Appendix*. The main findings remain consistent across both metrics.

In Fig. 5*A*, we present the ratio, for all forecasting rounds, between the average WIS of each model over the 4-wk horizon and the average WIS of a baseline model. The latter is defined as a model that consistently predicts, as median value, the last data point within the calibration period and whose predictive intervals are estimated on past data. Similar baseline models have become a standard neutral benchmark providing a simple reference for all models in the context of collaborative forecasting hubs, such as the US and the European COVID-19 Forecast Hub (63, 65). We refer the reader to *Materials and Methods* for more details and the definition of the baseline model. It follows that, values below (above) 1 indicate superior (inferior) performance compared to the baseline. We observe heterogeneous forecasting performance among the geographies under consideration. Notably, in Chicago, Gauteng, London, Madrid, New York, and Santiago de Chile, all models statistically outperform the baseline. However, in other locations, some models significantly underperform compared to the baseline; for instance, the Effective Force Behavioral Model in Rio de Janeiro and the Data-Driven Behavioral Model in Bogotá. We use the Wilcoxon signed-rank test (with outliers removal) to statistically compare the performance of different models. The null hypothesis of this test is that the two groups come from the same distribution. In Fig. 5*A*, we report the statistical significance of the tests comparing different pairs of models as follows: ****: $P_{\text{value}} \leq 10^{-4}$, ***: $10^{-4} < P_{\text{value}} \leq 10^{-3}$, **: $10^{-3} < P_{\text{value}} \leq 10^{-2}$, *: $10^{-2} < P_{\text{value}} \leq 0.05$, and otherwise blank if $P_{\text{value}} > 0.05$. The Data-Driven Behavioral Model performs best in terms of median relative WIS in Chicago, Jakarta, London, New York, Rio de Janeiro, and Santiago de Chile. However, its performance distribution is statistically different from that of the Compartmental Behavioral Model only in three cases (Chicago, London, and New York). The Compartmental Behavioral Model is the median top performer in the remaining three locations. Its performance distribution is statistically different from that of the Data-Driven model in all these three cases, namely Bogotá, Gauteng, and Madrid.

To provide an overall performance assessment, in Fig. 5*B*, we show the distribution of the relative WIS with respect to the baseline model for the three models. To provide an overall view of models' performance, we combine results from all geographies and all forecasting points. The analysis shows that

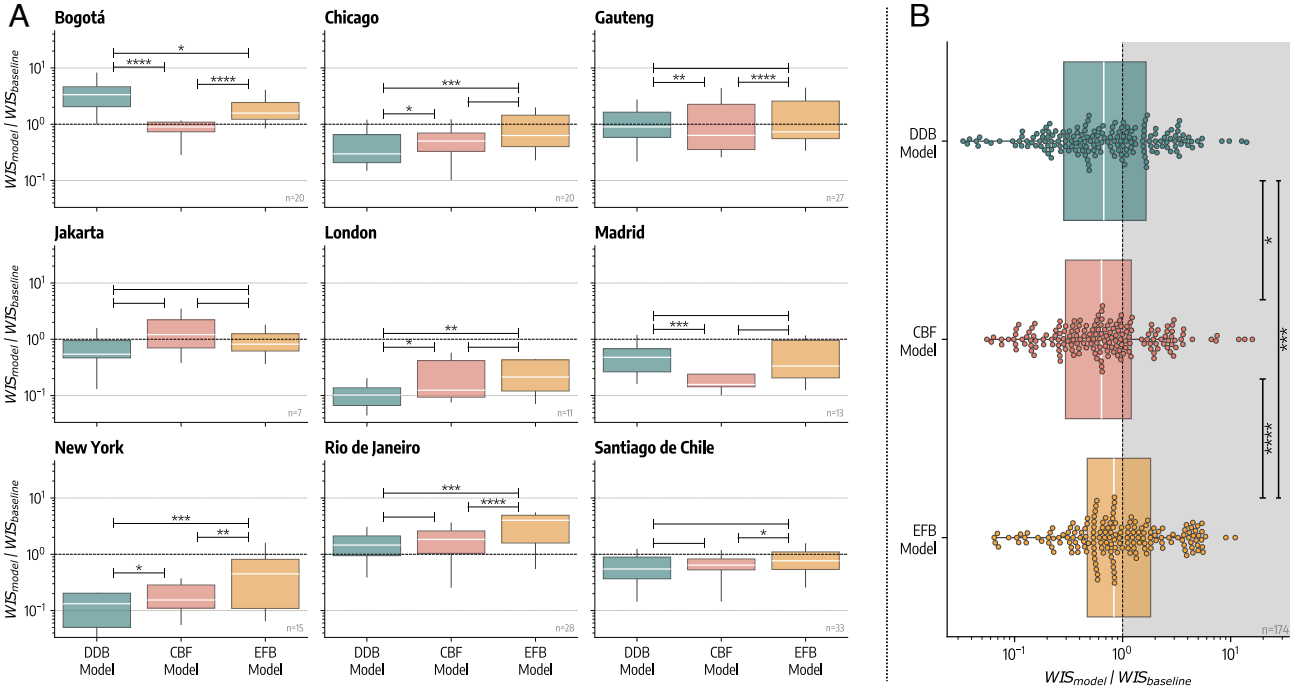

Fig. 5. Forecasting performance (WIS). (*A*) Relative WIS computed over all forecasting rounds for the three behavioral epidemic models across the nine geographical regions considered. Values below 1 indicate better performance with respect to the baseline forecasting model. Each data point underlying the boxplot represents the relative WIS averaged over the four-week horizon of the corresponding forecasting round. In the *Bottom Right* of each plot, we report the number of forecasting rounds for each location. (*B*) Boxplot and swarmplot of relative WIS for different models pooling together results from all rounds and geographies. The box boundaries represent the interquartile range (IQR) between the first and third quartiles (Q1 and Q3), the line inside the box indicates the median and the *Upper* (*Lower*) whisker extends to the last datum less (greater) than $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$). DDB stands for Data-Driven Behavioral model, CBF for Compartmental Behavioral Feedback model, and EFB for Effective Force of Infection Behavioral Feedback model. In both panels, we report the statistical significance of the Wilcoxon test comparing different forecasting performances as follows: ****: $P_{value} \leq 10^{-4}$, ***: $10^{-4} < P_{value} \leq 10^{-3}$, **: $10^{-3} < P_{value} \leq 10^{-2}$, *: $10^{-2} < P_{value} \leq 0.05$, and otherwise blank if $P_{value} > 0.05$.

the Compartmental Behavioral Feedback model is statistically the top performer with an overall median relative WIS of 0.64, closely followed by the Data-Driven Behavioral model (0.67) and the Effective Force of Infection Behavioral Feedback model (0.83). The reported significance of the Wilcoxon test confirms the statistical difference in performance between the Compartmental Behavioral Model and the other two models. Furthermore, for this model, nearly 70% of the forecasts are better than the baseline compared to only the 60% and 58% for the Data-Driven and the Effective Force of Infection Behavioral Feedback model, respectively.

It is important to note that in assessing forecast performance, the role of model complexity and the number of parameters remains unclear. Especially at the early stage of an epidemic wave, limited available data can disadvantage more complex models, which may be outperformed by more parsimonious approaches. To gather a better understanding on this point, in *SI Appendix*, we compare the forecasting performance of the three behavioral epidemic models against two simpler forecasting frameworks: a generalized logistic growth model (GLM) and a two-subepidemic model where each subepidemic follows GLM dynamics (66). Across the board, we find behavioral epidemic models to consistently outperform these approaches.

In *SI Appendix*, we present the forecasting performance analysis for two ensemble models that combine the forecasts of individual behavioral epidemic models. Ensemble forecasts have consistently demonstrated greater accuracy and reliability over time in various epidemiological forecasting contexts (63, 65). In the first ensemble approach, all models are weighted equally, while in the second, weights are proportional to past forecasting performance. Interestingly, we find that ensemble models outperform individual behavioral epidemic models in most cases across different metrics. Although the performance-weighted ensemble shows slight improvements over the equally weighted ensemble, these improvements are generally marginal and not statistically significant.

Discussion

Modeling the interplay between human behavior and the spread of infectious diseases is still considered a hard problem in epidemiology (24, 27, 28, 67, 68). One of the main obstacles to solving this challenge has been the lack of data to validate the theoretical models developed. A review of studies published between 2010 and 2015 found that only about 15% utilized data to parameterize and/or validate the proposed epidemic-behavior mechanisms (27). The COVID-19 pandemic has significantly altered this landscape. The abundance of novel data sources and the scale of the emergency made incorporating behavioral data into modeling studies possible. Most computational models, however, have leveraged data to include behavioral changes as an exogenous factor, as seen in the Data-Driven Behavioral Feedback Model used here. While this approach benefits from a transparent data integration process, it has limited the use and validation of general classes of models that explicitly simulate the feedback loop between the spread of infectious diseases and human behaviors.

Interestingly, our results show that Analytical Behavioral Feedback Models, developed well before the COVID-19 pandemic, often provide comparable or superior performance to data-driven approaches in both retrospective analysis and forecasting.

Although we observe variability depending on the task, geographical context, and the metrics considered, our findings suggest that purely data-driven methodologies to model behavior change may not always represent the best modeling solution.

While it may seem counterintuitive to suggest that data-driven approaches based on mobility data are less reflective of the actual dynamics of an epidemic, several factors contribute to this conclusion. In many cases, there is limited knowledge about the underlying population, or the data generation process, complicating the assessment of data representativeness and potential biases. Moreover, how variations in mobility data translate into changes in the contact structure relevant to infection dynamics, such as SARS-CoV-2 transmission, remains unclear. For forecasting purposes, real data can only be used under specific assumptions about the future -often with a status quo assumption- which may not accurately capture the true dynamics of the population. Finally, there are limited mobility data available for many low and middle-income countries, and when available, the data quality can be poor (69–72). This can easily deteriorate the modeling results (73). These issues can be mitigated by using analytical behavioral feedback models, which derive the epidemic-behavior interplay in a self-consistent way through calibration and parameterization of behavioral mechanisms. However, purely model-based approaches can misinterpret errors in reporting, the emergence of more transmissible virus strains, or other factors influencing the epidemic's progression as changes in behavior. Moreover, the high-dimensional space characterizing complex analytical behavioral feedback models could lead to challenges such as parameter identifiability, interpretability, and overfitting (74). Nevertheless, as shown in *SI Appendix*, all behavioral models (including the data-driven framework) do not show signs of practical nonidentifiability within the ranges of parameters explored.

Our results stress the critical importance of systematically evaluating different modeling approaches and considering the use of ensemble modeling techniques. Indeed, as shown in *SI Appendix*, by integrating the three models, ensemble methods generally improve forecast accuracy and generate a broader range of potential epidemic trajectories. Since no single model can completely capture the feedback mechanisms of epidemic dynamics and human behavior, ensembling multiple models allows for the integration of different mechanisms and factors, potentially resulting in a more reliable representation of the epidemic-behavior interplay (75, 76).

As with all modeling approaches, this study is not without limitations. First, we could not consider all the possible approaches used to model COVID-19 epidemic-behavior interplay. For instance, we did not test the performance of survey-driven models (77–80) and semimechanistic models (81, 82). Second, our results may be influenced by the chosen calibration framework. To address this, in *SI Appendix*, we conduct extensive analyses assessing the robustness of our approach via parameter recovery experiments. The findings demonstrate a high degree of robustness and reliability in the ABC-SMC framework used. Furthermore, translating mobility changes into contact reductions remains an open challenge. Therefore, the performance of the data-driven model may vary depending on different approaches to effective contacts rescaling (83–87). Finally, our forecasts are based on data reported as of today, not addressing the challenge of retrospective data adjustments (i.e., backfilling) very common for epidemiological datasets. While this could impact forecast performance, our primary goal is to compare models. Therefore, assuming that all models would be equally affected, this issue may not significantly impact our comparative analysis.

In light of the assumptions and limitations, our study offers a clear path forward for the application of analytic feedback behavioral models in both retrospective analysis and epidemic forecasting (65, 88, 89). The strong performance of the Compartmental Behavioral Feedback model in both tasks suggests that the epidemic-behavior interplay can be mechanistically captured in a parsimonious way, substantially improving accuracy. Interestingly, its superior performance over the other analytical behavioral model (i.e., the EFB model) appears to stem from the functional form of its behavioral mechanism, which offers greater flexibility in capturing behavioral responses. This increased adaptability enables it to reproduce a wider range of epidemic trajectories more effectively (see *SI Appendix* for more details). Moreover, our study paves the way for more systematic use of analytical feedback behavioral models in operational forecasting efforts. By demonstrating their capability to account for the dynamic relationship between human behavior and disease spread, these models can improve epidemic forecasts and projections without relying on explicit assumptions about the possible evolution of population behavior. This approach can be particularly valuable in anticipating the effects of public health interventions and supporting more informed decision-making.

Materials and Methods

Epidemic Models.

Compartmental and age structure. In all models studied, we adopt a SEIR compartmentalization setup. Healthy and susceptible individuals are placed in the compartment S. Through interactions with infectious, they transition to the compartment of the exposed E. Individuals in the E compartment get infectious only after the latent period (ϵ^{-1}) when they transition to the compartment I. Finally, after the infectious period (μ^{-1}) individuals in the I compartment transition to the compartment of the recovered R. We assume the population to be stratified into 10 different age groups ($[0-9, 10-19, 20-24, 25-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+]$). We introduce the contact matrix $\mathbf{C} \in \mathbb{R}^{K\times K}$, whose element $C_{i,j}$ is the average number of daily contacts that an individual in age group i has with individuals in age group j (90). The rate at which susceptible individuals acquire infection, namely the force of infection, is $\lambda(k,t) = s(t)\beta \sum_{k'=1}^{K} C_{kk'} \frac{I'_k(t)}{N'_k}$ (where $s(t)$ is a seasonality modulation term and $N_{k'}$ is the number of individuals in age group k' such that $N = \sum_{k'=1}^{K} N_{k'}$). The basic reproductive number for this model is $R_0 = \rho(\tilde{\mathbf{C}})\beta/\mu$, where β is the transmission rate of the disease, $\tilde{C}_{ij} = C_{ij}N_i/N_j$ and $\rho(\tilde{\mathbf{C}})$ is the spectral radius of the matrix. Our implementation of the model is stochastic and the number of individuals transitioning among compartments is simulated via chain binomial processes. Additionally, we model COVID-19 deaths by applying age-stratified infection fatality rates (91) to the number of individuals transitioning from I_k to R_k and accounting for a lag Δ between such transition and actual death due to isolation, hospitalization, and reporting delays. More details on the model definition are provided in *SI Appendix*.

Data-driven behavioral model. In the Data-Driven Behavioral model, we use the COVID-19 Community Mobility Report published by Google LLC (57) to modulate the force of infection. This dataset reports percentage changes in mobility to specific locations on a given day and geography. Our models do not consider multiple locations, so we derive an overall mobility change percentage $m(t)$ as the average of mobility changes toward all locations (excluding mobility toward parks due to its anomalous behavior). Finally, $m(t)$ is turned into a contacts reduction parameter as follows: $r_{\text{mobility}}(t) = (1 - |m(t)|/100)^2$. The intuition is that, under the homogeneous mixing assumption, the number of contacts will be proportional to the square of the number of individuals. Then, we use $r_{\text{mobility}}(t)$ to modulate the rate at which susceptible becomes infected as a consequence of behavior change, namely, we modify the force of infection as

$\lambda'(k,t) = r_{\text{mobility}}(t)\lambda(k,t)$. In short-term forecasting, we assume that future $r_{\text{mobility}}(t)$ will be equal to the last observed contacts reduction parameter in the calibration window (i.e., status quo assumption).

Compartmental behavioral feedback model. In the Compartmental Behavioral Feedback model, we introduce an additional compartment S_k^B of susceptible individuals that adopt behavior change and thus get infected at a lower rate $r\lambda(k,t)$, where $r < 1$ is a parameter that describes the efficacy of preventive measures. The transition from S_k to S_k^B happens at rate $\beta_B\left(1 - e^{-\gamma D_{rep}(t-1)}\right)$, where the β_B and γ regulate the behavioral response, and $D_{rep}(t-1)$ is the total number of reported deaths in the previous day. We also assume that S_k^B individuals can relax their behavior and transition back to S_k at a rate $\mu_B\left(\sum_{k'=1}^{K} S_{k'}(t) + R_{k'}(t)\right)/N$, where μ_B sets the tendency of susceptibles to drop safer behaviors (92).

Effective force of infection behavioral feedback model. In the Effective Force of Infection Behavioral Feedback model, we consider the following function

$$f(D_{rep},\xi,\psi,t) = \frac{1}{1 + \xi D_{rep}(t-1) + \psi\sum_{t'=1}^{t-1} D_{rep}(t')}, \quad [1]$$

where, as above, $D_{rep}(t-1)$ is the number of reported deaths at time $t-1$, $\sum_{t'=1}^{t-1} D_{rep}(t')$ is the cumulative number of reported deaths up to $t-1$, ξ and ψ are parameters that set the behavioral reactivity of individuals (55). This function multiplies the force of infection and serves as a proxy for the modulating effect of behavioral changes. Specifically, it takes into account new reported deaths in the last time step, capturing short-term effects of recent epidemiological conditions on behavior, as well as cumulative reported deaths, capturing the long-term effects of past epidemiological conditions on current behavior.

Models Calibration. Models are calibrated using an ABC-SMC algorithm (58, 93). The ABC-SMC is an extension of the more simple rejection algorithm, which works as follows. The modeler needs to choose prior distribution $\pi(\boldsymbol{\theta})$ for the free parameters $\boldsymbol{\theta}$ of the model, a distance metric $d(\cdot)$, a tolerance δ, and a population size P. Then, the model is run iteratively sampling at each step a parameters set $\boldsymbol{\theta}_i$ from the prior distribution $\pi(\boldsymbol{\theta})$. At each iteration, an output quantity produced by the model y_i (i.e., simulated deaths) is compared to the corresponding real quantity y_{obs} using the distance metric $d(y_i, y_{obs})$. If $d(y_i, y_{obs}) < \delta$ then $\boldsymbol{\theta}_i$ is accepted, otherwise it is rejected. This process continues until P parameter sets are accepted. The main limitation of this approach is that the acceptance criterion remains fixed, causing slow convergence. Additionally, finding an appropriate tolerance value δ beforehand is challenging, especially with multiple models and geographies. The ABC-SMC algorithm addresses these issues by using a sequence of T rejection steps (i.e., generations) with decreasing tolerance. Each generation's prior distribution is the posterior distribution from the previous one perturbed via a kernel function. This method starts with high error tolerances and broad prior distributions, progressively refining the parameter space. The final generation's accepted $\boldsymbol{\theta}_i$ distribution approximates the true posterior distribution of the parameters. Here, we consider 10 generations, 1,000 parameter sets accepted at each step, weekly deaths as output quantity, and the weighted mean absolute percentage error (wMAPE) as a distance metric, defined as

$$wMAPE_i = \frac{\sum_{t=1}^{t_n} |y_i(t) - y_{obs}(t)|}{\sum_{t=1}^{t_n} |y_{obs}(t)|}, \quad [2]$$

where y_{obs} is the vector of actuals and y_i of model estimates, and t_n is the number of weeks considered. We use the ABC-SMC implementation of the *pyabc* Python package (94). In *SI Appendix*, we report additional information on the calibration, including prior distributions.

Performance Metrics. The MAE of the median is defined as

$$MAE = \frac{\sum_{t=1}^{t_n} |y_{obs}(t) - M(t)|}{t_n}, \quad [3]$$

Where y_{obs} is the vector of actuals and M the model's medians. Its normalized version is simply defined as $nMAE = MAE/mean(y_{obs})$.

The WIS is a score that approximates the continuous ranked probability score (95). For a given a prediction interval $(1-\alpha)\times 100\%$ (i.e., 90% interval) of a model's estimate, the interval score (IS_α) is defined as

$$IS_\alpha = (u-l) + \frac{2}{\alpha}(l - y_{obs})\mathbb{I}(y_{obs} < l) + \frac{2}{\alpha}(y_{obs} - u)\mathbb{I}(y_{obs} > u), \quad [4]$$

Where u (l) is the upper (lower) limit of the prediction interval, y_{obs} is the actual outcome, and $\mathbb{I}(c)$ is an indicator function that equals 1 if condition c is met and 0 otherwise. Looking at IS_α we see that its first term captures how wide is the prediction interval, while the second and third terms are the penalization for under- and overprediction. Indeed, they are different from 0 only if the actual data point y_{obs} is below or above the interval limits. The *WIS* is an extension of the *IS* and takes into account multiple prediction intervals at once. It is defined as

$$WIS_{\alpha_{0:K}} = \frac{1}{K+\frac{1}{2}}\left(w_0|y_{obs} - M| + \sum_{k=1}^{K} w_k IS_{\alpha_k}\right), \quad [5]$$

Where K is the number of prediction intervals considered, M is the model's median, and w_k are the nonnegative weights of the different intervals. We can see that the first term in parenthesis measures how much the model's central estimate M differs from the actual data y_{obs}, while the second term is a weighted sum of the different interval scores IS_{α_k}. Following a common approach, we set $w_0 = 1/2$, $w_k = \alpha_k/2$, and we consider 11 prediction intervals ($\alpha_k = 0.02, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90$). Its normalized version is simply defined as $nWIS = WIS/mean(y_{obs})$.

The BIC is a metric that evaluates the model's estimates based on the accordance with real data and on the model complexity (96). For a model q it is defined as

$$BIC_q = k_q ln(t_n) - 2ln(L_q), \quad [6]$$

Where k_q is the number of model's free parameters, and L_q is the model's likelihood which we define it here as

$$L_q = \frac{1}{t_n}\sum_{t=1}^{t_n}\left(y_{obs}(t) - M_q(t)\right)^2 \quad [7]$$

that is the mean squared error between actual (y_{obs}) and median predicted values (M_q). We note that, for simplicity, we assume a likelihood function equivalent to a Gaussian error model. Intuitively, the model reaching the lowest BIC is the best, since it guarantees the minimum deviation from observed data with the minimum number of parameters. In this sense, BIC favors both accordance with real data and the model's parsimony; however, its values lack an immediate interpretation. For this reason, we consider BIC weights defined as

$$w_q(BIC) = \frac{e^{-\frac{1}{2}\Delta_q(BIC)}}{\sum_{q'} e^{-\frac{1}{2}\Delta_{q'}(BIC)}}, \quad [8]$$

Where $\Delta_q(BIC) = BIC_q - \min_{q'}(BIC_{q'})$. These weights express the relative probability of a model over the others.

Forecasting. In each forecasting round, we calibrate the three models using data up to time t and we forecast weekly deaths in the next four weeks. In the next round, we move our window up to $t+1$ and we repeat the calibration and forecasting procedures. This process is performed iteratively until the end of the epidemic curve, starting with at least 4 data points for model calibration. Instead of the ABC-SMC algorithm, for forecasting, we adopt a modified version of the rejection algorithm where, instead of setting a predefined tolerance, we calibrate models by selecting top 1,000 simulations out of a total of $1M$ simulations obtained through sampling from the prior distributions. In the case of forecasting, we also consider as distance metric a generalized version of the wMAPE which gives more importance to more recent data points defined as $\sum_{t=1}^{t_n}\left(w(t)\frac{|y_{obs}(t) - y_i(t)|}{|y_{obs}(t)|}\right) / \sum_{t=1}^{t_n} w(t)$, where $w(t) = 1/((t_n+1)-t)$.

Baseline forecasting model. We employ a baseline forecasting model that consistently predicts the median value as the last data point within the calibration period. To compute predictive intervals, we consider the previous 1-step increments. Specifically, we compute 1-step differences up to time t: $\delta = (d_2, d_3, \ldots, d_t)$. To ensure the median forecast aligns with the last calibration point, we symmetrize δ by considering $\delta' = (\delta, -\delta)$. If the maximum horizon is H, we sample H differences from δ'. Finally, predictions at horizon h are computed as $f_h = v_t + \sum_{i=1}^{h} \hat{d}_i$, where v_t represents the last observed data points, and $\hat{d}_i$ are 1-step differences sampled from δ'. Following this process, we generate 10,000 trajectories from which we compute quantiles and predictive intervals.

Data, Materials, and Software Availability. Code and data have been deposited in GitHub (https://github.com/ngozzi/covid-behavior-models) (97).

ACKNOWLEDGMENTS. N.G. acknowledge support from the Lagrange Project of the Institute for Scientific Interchange Foundation funded by Fondazione Cassa di Risparmio di Torino. A.V. acknowledges support from the HHS/CDC-5U01IP000113 and the CDC-RFA-FT-23-0069 cooperative agreement from the Center for Forecasting and Outbreak Analytics of the US Centers for Disease Control and Prevention. The findings and conclusions in this study are those of the authors and do not necessarily represent the official position of the funding agencies. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US Government.

1. N. Imai *et al*., Report 2: Estimating the potential total number of novel Coronavirus cases in Wuhan City, China. *Imp. Coll. Lond.* **22**, 1–7 (2020).
2. P. M. De Salazar, R. Niehus, A. Taylor, C. O. Buckee, M. Lipsitch, Identifying locations with possible undetected imported severe acute respiratory syndrome Coronavirus 2 cases by using importation predictions. *Emerg. Infect. Dis.* **26**, 1465 (2020).
3. M. Gilbert *et al*., Preparedness and vulnerability of African countries against importations of COVID-19: A modelling study. *Lancet* **395**, 871–877 (2020).
4. J. T. Davis *et al*., Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. *Nature* **600**, 127–132 (2021).
5. H. Salje *et al*., Estimating the burden of SARS-CoV-2 in France. *Science* **369**, 208–211 (2020).
6. M. Chinazzi *et al*., The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).
7. N.M. Ferguson *et al*., Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. *ICL* **16**, 1-20 (2020).
8. C. R. Wells *et al*., Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 7504–7509 (2020).
9. S. Flaxman *et al*., Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
10. M. Gatto *et al*., Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10484–10491 (2020).
11. L. Di Domenico, G. Pullano, C. E. Sabbatini, P. Y. Boëlle, V. Colizza, Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. *BMC Med.* **18**, 1–13 (2020).
12. M. U. Kraemer *et al*., The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
13. N. G. Davies *et al*., Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. *Science* **372**, eabg3055 (2021).
14. R. C. Barnard, N. G. Davies, M. Jit, W. J. Edmunds, Modelling the medium-term dynamics of SARS-CoV-2 transmission in England in the Omicron era. *Nat. Commun.* **13**, 1–15 (2022).
15. W. Yang, J. Shaman, COVID-19 pandemic dynamics in India, the SARS-CoV-2 Delta variant and implications for vaccination. *J. R. Soc. Interface.* **19**, 20210900 (2022).
16. M. Cai, G. Em Karniadakis, C. Li, Fractional SEIR model and data-driven predictions of COVID-19 dynamics of Omicron variant. *Chaos Interdiscip. J. Nonlinear Sci.* **32**, 071101 (2022).
17. M. U. Kraemer *et al*., Spatiotemporal invasion dynamics of SARS-CoV-2 lineage b. 1.1. 7 emergence. *Science* **373**, 889–895 (2021).
18. K. M. Bubar *et al*., Model-informed COVID-19 vaccine prioritization strategies by age and serostatus. *Science* **371**, 916–921 (2021).
19. L. Matrajt, J. Eaton, T. Leung, E. R. Brown, Vaccine optimization for COVID-19: Who to vaccinate first? *Sci. Adv.* **7**, eabf1374 (2021).
20. N. Gozzi *et al*., Anatomy of the first six months of COVID-19 Vaccination Campaign in Italy. *PLoS Comput. Biol.* **18**, e1010146 (2022).
21. B. H. Foy *et al*., Comparing COVID-19 vaccine allocation strategies in India: A mathematical modelling study. *Int. J. Infect. Dis.* **103**, 431–438 (2021).
22. P. C. Jentsch, M. Anand, C. T. Bauch, Prioritising COVID-19 vaccination in changing social and epidemiological landscapes: A mathematical modelling study. *Lancet Infect. Dis* **21**, 1097–1106 (2021).
23. N. Saadi *et al*., Models of COVID-19 vaccine prioritisation: A systematic literature search and narrative review. *BMC Med.* **19**, 1–11 (2021).
24. N. Perra, Non-pharmaceutical interventions during the COVID-19 pandemic: A review. *Phys. Rep.* **913**, 1–52 (2021).
25. A. Adiga *et al*., Mathematical models for COVID-19 pandemic: A comparative analysis. *J. Indian Inst. Sci.* **100**, 793–807 (2020).
26. Y. Xiang *et al*., COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models. *Infect. Dis. Model.* **6**, 324–342 (2021).
27. F. Verelst, L. Willem, P. Beutels, Behavioural change models for infectious disease transmission: A systematic review (2010–2015). *J. R. Soc. Interface* **13**, 20160820 (2016).
28. S. Funk *et al*., Nine challenges in incorporating the dynamics of behaviour in infectious diseases models. *Epidemics* **10**, 21–25 (2015).
29. S. Funk, E. Gilad, C. Watkins, V. A. Jansen, The spread of awareness and its impact on epidemic outbreaks. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 6872–6877 (2009).
30. J. Bedson *et al*., A review and agenda for integrated disease models including social and behavioural factors. *Nat. Hum. Behav.* **5**, 834–846 (2021).
31. N. Oliver *et al*., Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Sci. advances* **6**, eabc0764 (2020).
32. L. Alessandretti, What human mobility data tell us about COVID-19 spread. *Nat. Rev. Phys.* **4**, 12–13 (2022).
33. T. Hu *et al*., Human mobility data in the COVID-19 pandemic: Characteristics, applications, and challenges. *Int. J. Digit. Earth* **14**, 1126–1147 (2021).
34. E. Prestige *et al*., Estimating social contact rates for the COVID-19 pandemic using Google mobility and pre-pandemic contact surveys. *Epidemics p.* **51**, 100830 (2025).
35. Y. Zhou *et al*., Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: A modelling study using mobile phone data. *Lancet Digit. Health* **2**, e417–e424 (2020).
36. S. Lai *et al*., Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature* **585**, 410–413 (2020).
37. N. Gozzi *et al*., Estimating the effect of social inequalities on the mitigation of COVID-19 across communities in Santiago de Chile. *Nat. Commun.* **12**, 2429 (2021).
38. J. T. Wu, K. Leung, G. M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *Lancet* **395**, 689–697 (2020).
39. M. D. Patel *et al*., Association of simulated COVID-19 vaccination and nonpharmaceutical interventions with infections, hospitalizations, and mortality. *JAMA Netw. Open* **4**, e2110782–e2110782 (2021).
40. A. Aleta *et al*., Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nat. Hum. Behav.* **4**, 964–971 (2020).
41. J. M. Caldwell *et al*., Understanding COVID-19 dynamics and the effects of interventions in the Philippines: A mathematical modelling study. *Lancet Reg. Health West. Pac.* **14**, 100211 (2021).
42. R. Dutta, S. N. Gomes, D. Kalise, L. Pacchiardi, Using mobility data in the design of optimal lockdown strategies for the COVID-19 pandemic. *PLoS Comput. Biol.* **17**, e1009236 (2021).
43. W. Yang, J. L. Shaman, COVID-19 pandemic dynamics in South Africa and epidemiological characteristics of three variants of concern (Beta, Delta, and Omicron). *eLife* **11**, e78933 (2022).
44. L. Di Domenico *et al*., Adherence and sustainability of interventions informing optimal control against the COVID-19 pandemic. *Commun. Med.* **1**, 57 (2021).
45. S. Chang *et al*., Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2021).
46. K. A. Pawelek, A. Oeldorf-Hirsch, L. Rong, Modeling the impact of twitter on influenza epidemics. *Math. Biosci. Eng.* **11**, 1337–1356 (2014).
47. S. Collinson, K. Khan, J. M. Heffernan, The effects of media reports on disease spread and important public health measurements. *PLoS One* **10**, e0141423 (2015).
48. M. Springborn, G. Chowell, M. MacLachlan, E. P. Fenichel, Accounting for behavioral responses during a flu epidemic using home television viewing. *BMC Infect. Dis.* **15**, 1–14 (2015).
49. T. Usherwood, Z. LaJoie, V. Srivastava, A model and predictions for COVID-19 considering population behavior and vaccination. *Sci. Rep.* **11**, 1–11 (2021).
50. C. Maji, Impact of media-induced fear on the control of COVID-19 outbreak: A mathematical study. *Int. J. Differ. Equ.* **2021**, 1–11 (2021).
51. M. Cabrera, F. Córdova-Lepe, J. P. Gutiérrez-Jara, K. Vogt-Geisse, An SIR-type epidemiological model that integrates social distancing as a dynamic law based on point prevalence and socio-behavioral factors. *Sci. Rep.* **11**, 1–16 (2021).
52. J. Fernández-Villaverde, C. I. Jones, Estimating and simulating a SIRD model of COVID-19 for many countries, states, and cities. *J. Econ. Dyn. Control* **140**, 104318 (2022).
53. T. Kolokolnikov, D. Iron, Law of mass action and saturation in SIR model with application to Coronavirus modelling. *Infect. Dis. Model.* **6**, 91–97 (2021).
54. M. Kochańczyk, F. Grabowski, T. Lipniacki, Dynamics of COVID-19 pandemic at constant and time-dependent contact rates. *Math. Model. Nat. Phenom.* **15**, 28 (2020).
55. J. S. Weitz, S. W. Park, C. Eksin, J. Dushoff, Awareness-driven behavior changes can shift the shape of epidemics away from peaks and toward plateaus, shoulders, and oscillations. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 32764–32771 (2020).
56. S. M. Garba, J. M. S. Lubuma, B. Tsanou, Modeling the transmission dynamics of the COVID-19 Pandemic in South Africa. *Math. Biosci.* **328**, 108441 (2020).
57. Google, COVID-19 community mobility report (2020). https://www.google.com/covid19/mobility/. Accessed 1 August 2021.
58. A. Minter, R. Retkute, Approximate Bayesian Computation for infectious disease modelling. *Epidemics* **29**, 100368 (2019).
59. T. Toni, D. Welch, N. Strelkowa, A. Ipsen, M. P. Stumpf, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface.* **6**, 187–202 (2009).
60. M. J. Keeling, P. Rohani, *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2008).
61. O. Diekmann, J. Heesterbeek, M. G. Roberts, The construction of next-generation matrices for compartmental epidemic models. *J. R. Soc. Interface* **7**, 873–885 (2010).
62. A. Cori, N. M. Ferguson, C. Fraser, S. Cauchemez, A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178**, 1505–1512 (2013).
63. E. Y. Cramer *et al*., Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2113561119 (2022).
64. H. Manley *et al*., Combining models to generate a consensus effective reproduction number R for the COVID-19 epidemic status in England. *Epidemiol. & Infect.* 152, e59 (2024).

65. K. Sherratt *et al*., Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *eLife* **12**, e81916 (2023).
66. G. Chowell *et al*., An ensemble n-sub-epidemic modeling framework for short-term forecasting epidemic trajectories: Application to the COVID-19 pandemic in the USA. *PLoS Comput. Biol.* **18**, e1010602 (2022).
67. P. Manfredi, A. D'Onofrio, *Modeling the Interplay Between Human Behavior and the Spread of Infectious Diseases* (Springer Science & Business Media, 2013).
68. H. Manley et al., Combining models to generate a consensus effective reproduction number R for the COVID-19 epidemic status in England. *Epidemiol. & Infect.* **152**, e59 (2024).
69. S. Milusheva, A. Lewin, T. B. Gomez, D. Matekenya, K. Reid, Challenges and opportunities in accessing mobile phone data for COVID-19 response in developing countries. *Data Policy* **3**, e20 (2021).
70. J. Wardle, S. Bhatia, M. U. Kraemer, P. Nouvellet, A. Cori, Gaps in mobility data and implications for modelling epidemic spread: a scoping review and simulation study. *Epidemics* **42**, 100666 (2023).
71. S. Milusheva, D. Bjorkegren, L. Viotti, "Assessing bias in smartphone mobility estimates in low income countries" in *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS '21* (Association for Computing Machinery, New York, NY, 2021), pp. 364–378.
72. M. Tizzoni *et al*., Addressing the socioeconomic divide in computational modeling for infectious diseases. *Nat. Commun.* **13**, 1–7 (2022).
73. T. Ramiadantsoa *et al*., Existing human mobility data sources poorly predicted the spatial spread of SARS-CoV-2 in Madagascar. *Epidemics* **38**, 100534 (2022).
74. G. Chowell, P. Skums, Investigating and forecasting infectious disease dynamics using epidemiological and molecular surveillance data. *Phys. Life Rev.* **51**, 294–327 (2024).
75. K. Shea *et al*., Multiple models for outbreak decision support in the face of uncertainty. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2207537120 (2023).
76. N.G. Reich *et al*., Collaborative hubs: making the most of predictive epidemic modeling. *Am. J. Public Heal.* **112**, 839–842 (2022).
77. J. Zhang *et al*., Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* **368**, 1481–1486 (2020).
78. F. Trentini *et al*., Investigating the relationship between interventions, contact patterns, and SARS-CoV-2 transmissibility. *Epidemics* **40**, 100601 (2022).
79. A. J. Kucharski *et al*., Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: A mathematical modelling study. *Lancet. Infect. Dis.* **20**, 1151–1160 (2020).
80. A. Koher, F. Jørgensen, M. B. Petersen, S. Lehmann, Epidemic modelling of monitoring public behavior using surveys during pandemic-induced lockdowns. *Commun. Med.* **3**, 80 (2023).
81. S. Funk, A. Camacho, A. J. Kucharski, R. M. Eggo, W. J. Edmunds, Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics* **22**, 56–61 (2018).
82. G. C. Gibson, N. G. Reich, D. Sheldon, Real-time mechanistic Bayesian forecasts of COVID-19 mortality. *AOAS.* **17**, 1801 (2023).
83. N. Kishore *et al*., Evaluating the reliability of mobility metrics from aggregated mobile phone data as proxies for SARS-CoV-2 transmission in the USA: A population-based study. *Lancet Digit. Health* **4**, e27–e36 (2022).
84. S. Jewell *et al*., It's complicated: Characterizing the time-varying relationship between cell phone mobility and COVID-19 spread in the US. *NPJ Digit. Med.* **4**, 1–11 (2021).
85. H. S. Badr, L. M. Gardner, Limitations of using mobile phone data to model COVID-19 transmission in the USA. *Lancet Infect. Dis.* **21**, e113 (2021).
86. O. Gatalo, K. Tseng, A. Hamilton, G. Lin, E. Klein, Associations between phone mobility data and COVID-19 cases. *Lancet Infect. Dis.* **21**, e111 (2021).
87. F. Delussu, M. Tizzoni, L. Gauvin, The limits of human mobility traces to predict the spread of COVID-19: A transfer entropy approach. *PNAS nexus* **2**, pgad302 (2023).
88. N. G. Reich *et al*., A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 3146–3154 (2019).
89. E. Howerton *et al*., Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty. *Nat. Commun.* **14**, 7260 (2023).
90. D. Mistry *et al*., Inferring high-resolution human mixing patterns for disease modeling. *Nat. Commun.* **12**, 1–12 (2021).
91. R. Verity *et al*., Estimates of the severity of coronavirus disease 2019: A model-based analysis. *Lancet Infect. Dis.* **20**, 669–677 (2020).
92. N. Perra, D. Balcan, B. Gonçalves, A. Vespignani, Towards a characterization of behavior-disease models. *PLoS One* **6**, e23084 (2011).
93. M. Sunnåker *et al*., Approximate Bayesian computation. *PLoS Comput. Biol.* **9**, e1002803 (2013).
94. Y. Schälte, E. Klinger, E. Alamoudi, J. Hasenauer, pyABC: Efficient and robust easy-to-use approximate Bayesian computation. *J. Open Sour. Softw.* **7**, 4304 (2022).
95. J. Bracher, E. L. Ray, T. Gneiting, N. G. Reich, Evaluating epidemic forecasts in an interval format. *PLoS Comput. Biol.* **17**, e1008618 (2021).
96. D. Anderson, K. Burnham, Model selection and multi-model inference. *Second. NY Springer Verlag* **63**, 10 (2004).
97. N. Gozzi, N. Perra, A. Vespignani, Comparative evaluation of behavioral epidemic models using COVID-19 data. GitHub. https://github.com/ngozzi/covid-behavior-models. Deposited 13 May 2025.