National and subnational short-term forecasting of COVID-19 in Germany and Poland, early 2021

J. Bracher[1,2,*], D. Wolffram[1,2,3], J. Deuschel[1], K. Görgen[1], J.L. Ketterer[1], A. Ullrich[4], S. Abbott[5], M.V. Barbarossa[6], D. Bertsimas[7], S. Bhatia[8], M. Bodych[9], N.I. Bosse[5], J.P. Burgard[10], J. Fiedler[11], J. Fuhrmann[12], S. Funk[5], A. Gambin[13], K. Gogolewski[13], S. Heyder[14], T. Hotz[14], Y. Kheifetz[15], H. Kirsten[15], T. Krueger[9], E. Krymova[16], N. Leithäuser[11], M.L. Li[17], J.H. Meinke[18], B. Miasojedow[13], J. Mohring[11], P. Nouvellet[19], J.M. Nowosielski[20], T. Ozanski[9], M. Radwan[20], F. Rakowski[20], M. Scholz[15], S. Soni[7], A. Srivastava[21], T. Gneiting[2,22], M. Schienle[1,*]

November 5, 2021

Abstract

We report on the second and final part of a pre-registered forecasting study on COVID-19 cases and deaths in Germany and Poland. Fifteen independent research teams provided forecasts at lead times of one through four weeks from January through mid-April 2021. Compared to the first part (October–December 2020), the number of participating teams increased, and a number of teams started providing subnational-level forecasts. The addressed time period is characterized by rather stable non-pharmaceutical interventions in both countries, making short-term predictions more straightforward than in the first part of our study. In both countries, case counts declined initially, before rebounding due to the rise of the B.1.1.7 variant. Deaths declined through most of the study period in Germany while in Poland they increased after a prolonged plateau. Many, though not all, models outperformed a simple baseline model up to four weeks ahead, with ensemble methods showing very good relative performance. Major trend changes in reported cases, however, remained challenging to predict.

* Correspondence to: Johannes Bracher (johannes.bracher@kit.edu), Melanie Schienle (melanie.schienle@kit.edu)

[1]Chair of Statistics and Econometrics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2]Computational Statistics Group, Heidelberg Institute for Theoretical Studies (HITS), Germany
[3]HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[4]Robert Koch Institute (RKI), Berlin, Germany
[5]London School of Hygiene and Tropical Medicine, London, UK
[6]Frankfurt Institute for Advanced Studies, Frankfurt, Germany
[7]Sloan School of Management, Massachusetts Institute of Technology, Cambridge MA, USA
[8]MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), Imperial College London, London, UK
[9]Wroclaw University of Science and Technology, Wroclaw, Poland
[10]Economic and Social Statistics Department, University of Trier, Trier, Germany
[11]Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany
[12]Institute for Applied Mathematics, University of Heidelberg
[13]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland
[14]Institute of Mathematics, Technische Universität Ilmenau
[15]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany
[16]Swiss Data Science Center, ETH Zurich and EPFL, Lausanne, Switzerland
[17]Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA
[18]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany
[19]School of Life Sciences, University of Sussex, Brighton, UK
[20]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland
[21]Ming Hsieh Department of Computer and Electrical Engineering, University of Southern California, Los Angeles, CA, USA
[22]Institute for Stochastics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

1 Introduction

Short-term forecasts of infectious diseases and longer-term scenario projections provide complementary perspectives to inform public health decision making. Both have received considerable attention during the COVID-19 pandemic and are increasingly embraced by public health agencies. This is illustrated by the US COVID-19 Forecast (1; 2) and Scenario Modelling Hubs (3), supported by the US Centers for Disease Control and Prevention, as well as the more recent European COVID-19 Forecast Hub (4), supported by the European Center for Disease Prevention and Control (ECDC). The Forecast Hub concept, building on pre-pandemic collaborative disease forecasting projects like *FluSight* (5), the *DARPA Chikungunya Challenge* (6) or the *Dengue Forecasting Project* (7) aims to provide a broad picture of existing short-term projections in real time, making the agreement or disagreement between different models visible. Also, it forms the basis for a systematic evaluation of performance. This is a prerequisite for model consolidation and improvement, and a need repeatedly expressed (8). In the German-speaking public discourse, the need for well-designed prospective and pre-registered studies in the field of disease modelling has been highlighted (9).

We here report on the second part of such a study, pre-registered on 8 October 2020 (10) and including forecasts made between 11 January 2021 and 29 March 2021 (with last observed values running through April; twelve weeks of forecasting). It is based on the German and Polish COVID-19 Forecast Hub (https://kitmetricslab.github.io/forecasthub/), which gathers and stores forecasts in real time. This platform was launched in close exchange with the US COVID-19 Forecast Hub in June 2020. In April 2021 it was largely merged into the European COVID-19 Forecast Hub, shortly after the latter had been launched by ECDC. During our study period, fifteen independent modelling teams provided forecasts of cases and deaths by reporting date, based on data either from national health authorities (Robert Koch Institute, RKI or the Polish Ministry of Health, MZ; the primary data source) or the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE; (11)). As in the first part of our study ((12), October 2020–December 2020), we focus on forecasts one and two weeks ahead. As non-pharmaceutical interventions were more stable in the second than in the first period, we give more attention to the evaluation of three- and four-week-ahead forecasts, but acknowledge that forecasts (as opposed to scenarios) are most meaningful for short time horizons.

The time series of cases and deaths in both countries are displayed in panels (a) and (b) of Figure 1. The study period covered in this paper is marked in dark grey, while the light grey area represents the time span addressed in (12). Our study period contains the transition from the original wild type variant of the virus to the B.1.1.7 variant (later called *Alpha*); panel (c) of Figure 1 shows the weekly percentages of all cases which were due to the B.1.1.7 variant in the two countries in calendar weeks 4–12 (taken from (13) for Germany and (14), (15) for Poland). Panel (d) shows the Oxford Coronavirus Government Response Tracker (OxCGRT) Stringency Index (16). It can be seen that compared to the first part of our study, the level of non-pharmaceutical interventions was rather stable at a high level during the second period. We note, however, that on 27 March a new set of restrictions was added in Poland (closure of daycare centers, hair salons and sports facilities, among others), which is not reflected very strongly in the stringency index. The start of vaccination rollout in both countries coincides with the start of our study period. However, by its end only roughly one sixth of the population of both countries had received a first dose, and roughly one twentieth had received two doses (with the role of the one-dose Johnson and Johnson vaccine negligible in both countries); see panel (e).

We find that averaged over the second evaluation period, most though not all of the compared models were able to outperform a naïve baseline model. Heterogeneity between forecasts from different models was considerable. Combined ensemble forecasts achieved very good performance relative to single-model forecasts. However, most models, including the ensemble, did not anticipate changes in trend well, in particular for cases. Pooling results over both evaluation periods we find that ensemble forecasts for deaths were well-calibrated even at longer prediction horizons and clearly outperformed baseline and individual models, while for cases this was only the case for one- and to a lesser degree two-week-ahead forecasts.

2 Results

Figures 2 and 3 show the forecasts made by the median ensemble (`KIT-median_ensemble`; our pre-specified main ensemble approach), a naïve last-observation-carried-forward model (`KIT-baseline`) and seven con-

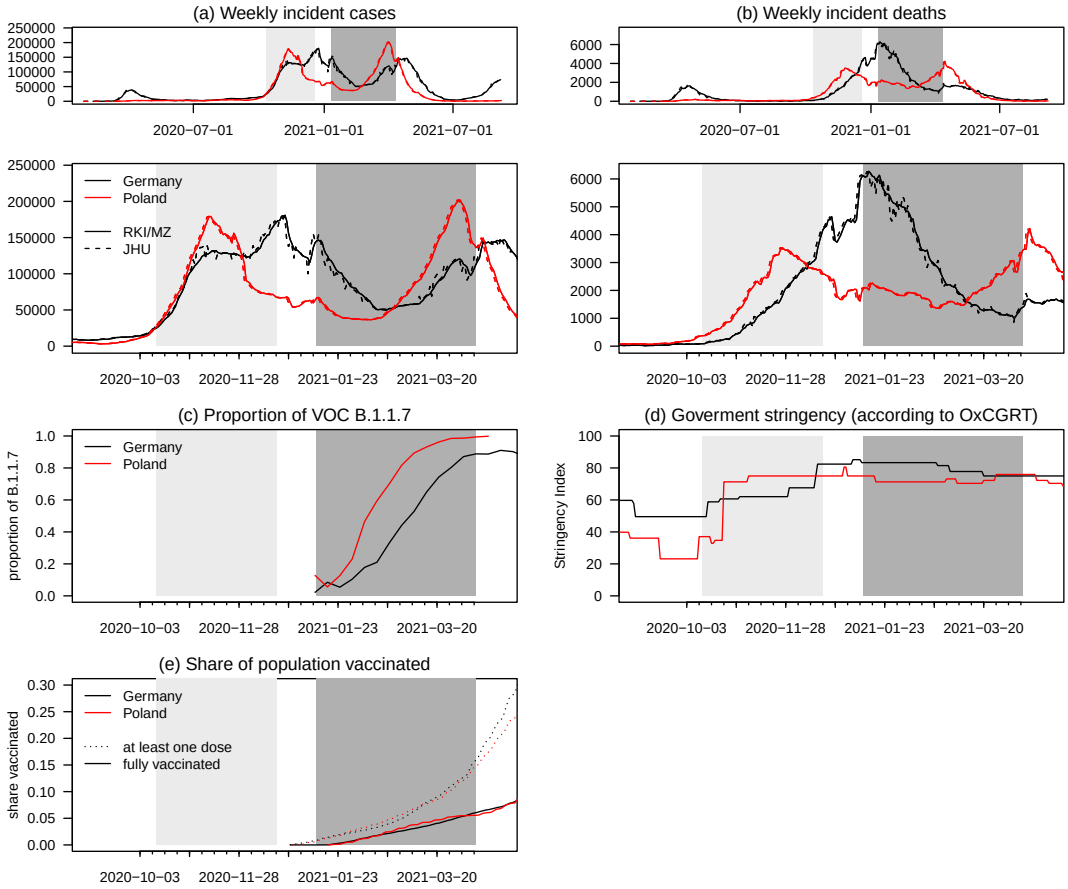

Figure 1: Reported cases (a) and deaths (b) in Germany and Poland according to Robert Koch Institute, the Polish Ministry of Health (MZ) and Johns Hopkins CSSE. Additional panels show (c) the share of cases due to the B.1.1.7 (Alpha) variant, (d) the overall level of non-pharmaceutical interventions as measured by the Oxford Coronavirus Government Response Tracker (OxCGRT) Stringency Index, and (e) the population shares having received at least one vaccination dose and complete vaccination. The dark grey area indicates the period addressed in the present manuscript, the light grey area the one from (12).

tributed models with above-average overall performance across locations and targets. The forecasts are probabilistic, and we illustrate the 50% and 95% prediction interval (PI) along with the respective median. Forecasts by the remaining teams are illustrated in Supplementary Figures 8 and 9, and forecasts at horizons of three and four weeks are shown in Supplementary Figures 10–13. In the following, we discuss the performance of these forecasts, starting with a formal statistical evaluation before directing attention to the behaviour at inflection points. Additional information on the submitted, baseline and ensemble models can be found in Sections 4.3 and 4.4.

2.1 Formal evaluation, January–April 2021

Table 1 and Figure 4 summarize the performance of the submitted, baseline and ensemble models over the twelve-week study period. Performance is measured via the average weighted interval score (WIS, (17)), an error measure for probabilistic forecasts, and the absolute error of the predictive median. For both measures lower values indicate better predictive performance, and the WIS can be decomposed into components representing underprediction, forecast spread and overprediction, see Section 4.2. Detailed results in tabular form at horizons of three and four weeks ahead can be found in Supplementary Table 4. As specified in the study protocol, we also provide results for cumulative cases and deaths (Supplementary Tables 6 and 7) and based on JHU rather than RKI/MZ data (Supplementary Tables 8 and 9; evaluation

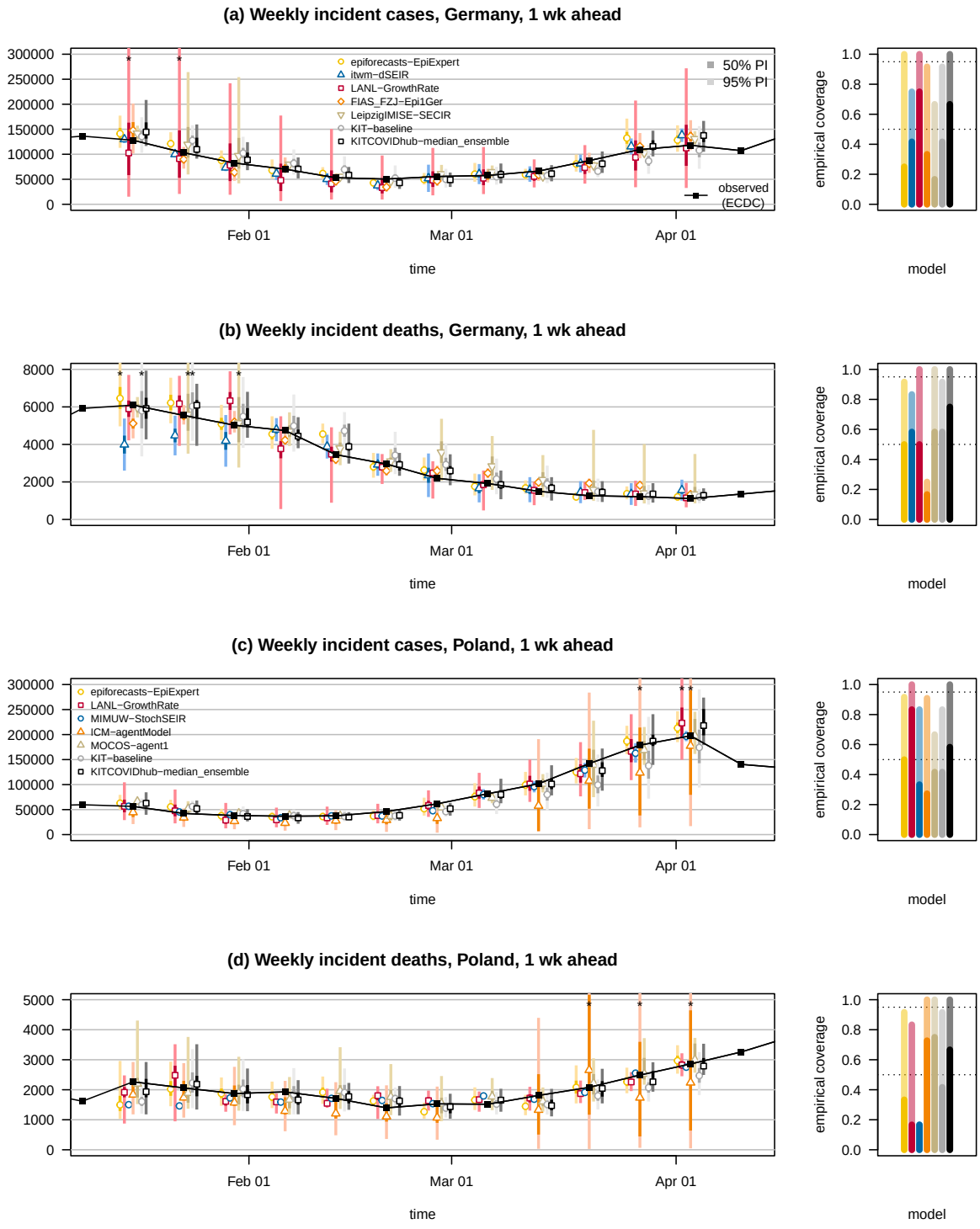

Figure 2: One-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. The figure shows forecasts from a baseline model, the median ensemble of all submissions and a subset of submitted models with above-average performance. Asterisks mark prediction intervals exceeding the upper plot limit. The remaining submitted models are displayed in Supplementary Figure 8.

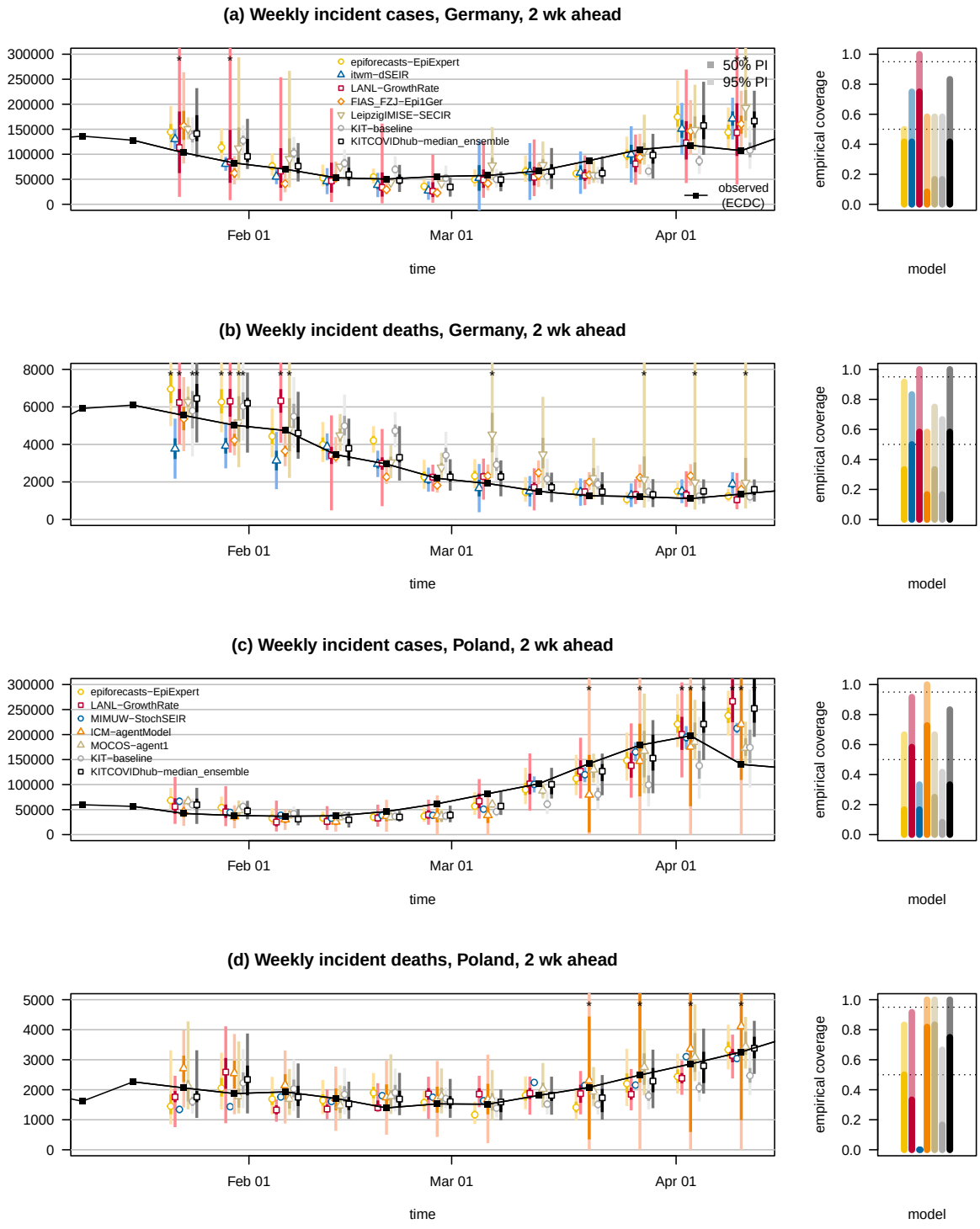

Figure 3: Two-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. The figure shows forecasts from a baseline model, the median ensemble of all submissions and a subset of submitted models. Asterisks mark prediction intervals exceeding the upper plot limit. The remaining submitted models are displayed in Supplementary Figure 9.

against JHU data leads to slightly higher WIS and absolute errors, but quite similar relative performance of models). Both for incident cases and deaths, a majority, but not all models outperformed the naïve baseline model `KIT-baseline` (a model outperforms the baseline for a given target whenever its stacked bar chart in Figure 4 does not reach into the grey area). As one would expect, the performance of all models considerably deteriorated for longer forecast horizons. The pre-specified median ensemble (see Materials and Methods) was consistently among the best-performing methods, outperforming most individual forecasts for all targets. The `KIT-extrapolation_baseline` model shows quite reasonable relative performance for cases in both countries. This model extrapolates exponential growth or decrease if the last three observations are monotonically increasing or decreasing, with a weekly growth rate equal to the one observed between the last and second to last week; if the last three observations are not ordered, it predicts a plateau. Predictive quantiles are obtained by assuming a negative binomial observation model with a dispersion parameter estimated via maximum likelihood from five recent observations (see Supplementary Note 2 of (12) for a detailed description). Given the relatively long stretches of continued upward or downward trends, this simple heuristic was not easy to beat. For deaths, on the other hand, the ensemble approaches achieve quite substantial improvement relative to this baseline.

The most striking cases of individual models outperforming the ensemble occurred for longer-range case forecasts in Poland. Here, the two microsimulation models `MOCOS-agent1` and `ICM-agentModel` performed considerably better. These two models were arguably among the ones which were most meticulously tuned to the specific national context. It seems that this yielded benefits for longer horizons, while at shorter horizons the ensemble and some considerably simpler models were at least on par (the best performance at the one week horizon being achieved by the compartmental model `MIMUW-StochSEIR`).

There were considerable differences in the forecast uncertainty of the different models. This can be seen from the quite variable forecast interval widths in Figures 2 and 3, and resulted in large differences in the empirical coverage rates of 50% and 95% prediction intervals (Table 1). The ensemble methods performed quite favourably in terms of coverage, typically with slight undercoverage for cases and slight overcoverage for deaths. The differences in forecast dispersion are also reflected by the components of the weighted interval score shown in Figure 4 (see Materials and Methods for an explanation of the decomposition). Some models, most strikingly `ITWW-county_repro`, issued very sharp predictions, leading to very small dispersion components of the weighted interval score (the darkest block in the middle of the stacked bar). In turn, this model received rather large penalties for both over- and underprediction. Other models, like `LANL-GrowthRate`, `epiforecasts-EpiNow2` and `ICM-agentModel` issued comparatively wide forecasts, leading to WIS values with large dispersion components. While there is no clear rule on what the score decomposition of an "ideal" forecast should look like, comparisons of the components provide useful indications on how to improve a model (e.g., the `ITWW-county_repro` model might benefit from widening the uncertainty intervals).

A subset of models also provided forecasts at the subnational level (states in Germany, voivodeships in Poland). Table 2 provides a summary of the respective results at the one and two week horizons (results for three and four weeks can be found in Supplementary Table 5). Despite the rather low number of available models, the ensembles generally achieved improvements over the individual models and, with exceptions for case forecasts in Germany, clearly outperformed the baseline model `KIT-baseline`. The mean WIS values are lower for the regional forecasts than for the national-level forecasts in Table 1 primarily because the numbers to be predicted are lower at the regional level; the WIS – like the absolute error – scales with the order of magnitude of the predicted quantity and cannot be compared directly across different forecasting tasks. Coverage of the ensemble forecasts was close to the nominal level for deaths and somewhat lower for cases. Note that in this comparison part of the forecasts from the `FIAS_FZJ-epi1Ger` model were created retrospectively as the team only started issuing forecasts for all German federal states on 22 February 2021.

As specified in the study protocol (10), we also report evaluation results at the national level pooled across the two study periods for those models which covered both. These are summarized in Supplementary Tables 10 and 11. For deaths, ensemble forecasts clearly outperformed individual models, the four-week-ahead horizon in Poland being the only one at which an individual model (`epiforecasts-EpiExpert`) meaningfully outperformed the pre-specified median ensemble. While most contributed and baseline models were somewhat overconfident, the ensemble showed close to nominal coverage even at the four-week-ahead horizon. For cases, the median ensemble achieved good relative performance (comparable to the best individual models) one and two weeks ahead, but was outperformed by a number of other models at three and four weeks. Notably, it failed to beat the naïve last-observation-carried-forward model `KIT-baseline`. Its coverage of prediction intervals was acceptable one week ahead, but substantially below nominal at higher horizons (e.g.,

Table 1: Forecast evaluation for Germany and Poland (incidence scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

Model	1 wk ahead case			2 wk ahead case			1 wk ahead death			2 wk ahead death		
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	9,252	5,415	3/12	12/12	20,233	13,607	5/12	6/12	300	204	6/12	11/12	509	323	4/12	11/12
epiforecasts-EpiNow2	9,676	6,644	8/12	10/12	29,348	21,478	7/12	8/12	300	188	9/12	11/12	581	417	8/12	9/12
FIAS_FZJ-Epi1Ger	10,218	6,294	4/12	11/12	25,662	16,621	1/12	7/12	436	336	2/12	3/12	655	475	2/12	7/12
IHME-CurveFit									516				656			
Imperial-ensemble2									*193	*136	8/10	9/10				
itwm-dSEIR	6,905	4,644	5/12	9/12	18,935	13,626	5/12	9/12	483	326	7/12	10/12	534	354	6/12	10/12
ITWW-county_repro	15,223	12,418	1/12	3/12	31,836	25,851	0/12	2/12	564	527	0/12	0/12	286	236	1/12	3/12
Karlen-pypm	18,532	13,629	6/12	11/12	35,010	25,385	3/12	10/12	380	232	5/12	11/12	628	394	1/12	10/12
LANL-GrowthRate	12,623	10,542	9/12	12/12	15,797	13,945	9/12	12/12	338	222	6/12	12/12	425	265	7/12	12/12
LeipzigIMISE-SECIR	9,161	6,376	2/12	8/12	26,650	19,185	2/12	7/12	370	281	7/12	12/12	874	636	4/12	9/12
MIT_CovidAnalytics-DELPHI	*11,910	* 8,277	6/11	10/11	*22,734	*16,006	4/11	8/11	803	490	5/12	11/12	773	451	7/12	12/12
SDSC_ISG-TrendModel	7,861								436							
USC-SIkJalpha	13,766	9,001	3/12	10/12	25,730	17,681	2/12	7/12	381	255	6/12	10/12	568	348	3/12	10/12
KIT-baseline	12,756	7,953	5/12	11/12	23,785	17,330	2/12	7/12	411	277	7/12	11/12	780	525	2/12	8/12
KIT-extrapolation_baseline	8,823	5,715	6/12	12/12	22,858	14,679	4/12	9/12	456	269	4/12	12/12	806	490	4/12	10/12
KIT-time_series_baseline	15,583	10,281	3/12	9/12	32,306	22,026	3/12	8/12	406	263	8/12	12/12	851	601	6/12	11/12
KITCOVIDhub-inverse_wis_ensemble	8,586	5,294	7/12	12/12	22,000	13,824	6/12	10/12	216	149	9/12	12/12	307	207	9/12	12/12
KITCOVIDhub-mean_ensemble	8,377	5,277	9/12	12/12	21,825	13,662	6/12	11/12	220	152	7/12	12/12	346	219	9/12	12/12
KITCOVIDhub-median_ensemble	7,344	4,660	8/12	12/12	19,296	12,734	5/12	10/12	232	150	9/12	12/12	376	225	7/12	12/12

Poland

Model	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	7,500	4,553	6/12	11/12	25,316	17,408	2/12	8/12	208	137	4/12	11/12	287	181	6/12	10/12
epiforecasts-EpiNow2	7,928	5,906	7/12	11/12	29,762	22,098	5/12	10/12	184	119	7/12	12/12	340	228	7/12	12/12
ICM-agentModel	*23,011	*15,824	3/11	10/11	*26,694	*18,098	8/11	11/11	*488	*294	8/11	11/11	*605	*507	9/11	11/11
IHME-CurveFit									374				520			
Imperial-ensemble2									*188	*138	3/10	7/10				
ITWW-county_repro	20,054	17,364	2/12	3/12	36,651	31,445	2/12	6/12	589	551	0/12	0/12	784	711	0/12	0/12
LANL-GrowthRate	8,129	5,787	10/12	12/12	23,269	15,240	7/12	11/12	229	137	2/12	10/12	347	216	4/12	11/12
MIMUW-StochSEIR	5,705	4,028	4/12	10/12	17,642	15,347	2/12	4/12	237	224	2/12	2/12	288	267	0/12	0/12
MIT_CovidAnalytics-DELPHI	*22,344	*12,912	2/10	9/10	*49,687	*33,033	1/10	7/10	*393	*244	6/11	11/11	*520	*296	4/11	11/11
MOCOS-agent1	5,173	4,978	5/12	8/12	15,022	11,380	3/12	8/12	158	132	9/12	12/12	203	149	10/12	12/12
SDSC_ISG-TrendModel	6,323								265							
USC-SIkJalpha	10,404	6,919	4/12	10/12	32,822	24,436	2/12	6/12	206	133	4/12	11/12	266	168	5/12	11/12
KIT-baseline	16,407	9,736	5/12	10/12	32,182	22,709	1/12	5/12	258	167	5/12	11/12	416	275	2/12	8/12
KIT-extrapolation_baseline	9,448	5,992	6/12	11/12	29,638	22,165	3/12	7/12	269	190	7/12	10/12	404	284	5/12	10/12
KIT-time_series_baseline	10,784	7,787	9/12	10/12	30,359	21,510	6/12	9/12	300	232	8/12	8/12	467	362	6/12	7/12
KITCOVIDhub-inverse_wis_ensemble	7,319	4,689	6/12	12/12	23,418	15,580	5/12	10/12	150	111	9/12	12/12	197	144	9/12	12/12
KITCOVIDhub-mean_ensemble	6,866	4,784	9/12	12/12	23,673	15,573	4/12	10/12	141	114	9/12	12/12	173	152	11/12	12/12
KITCOVIDhub-median_ensemble	7,130	4,403	7/12	12/12	23,027	16,241	4/12	10/12	162	103	8/12	12/12	193	137	9/12	12/12

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

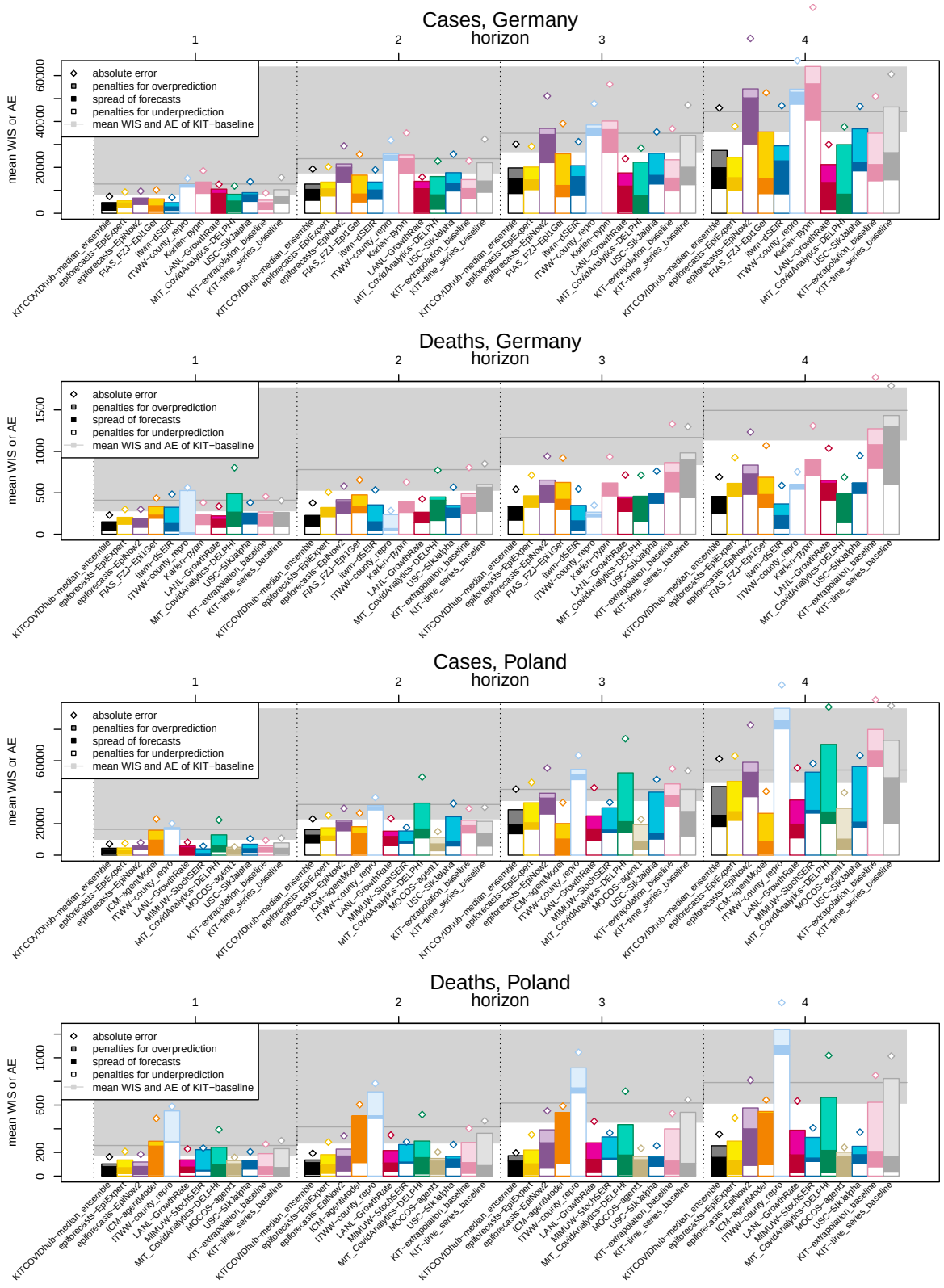

Figure 4: Average weighted interval score and absolute error achieved by models across countries, targets and forecast horizons. The grey area represents the performance of the baseline model `KIT-baseline`. WIS values are decomposed into components for forecast spread and penalties for overprediction and underprediction.

13/19 and 10/19 four weeks ahead in Germany and Poland, respectively, at the 0.95 level), which reflects the severe difficulties in predicting cases in Fall 2020 as discussed in (12).

2.2 Behaviour at inflection points

From a public health perspective, there is often a specific interest in how well models anticipated major inflection points. We therefore specifically discuss these instances. However, we note that, as will be detailed in the discussion, post-hoc conditioning of evaluation results on the occurrence of unusual events comes with important conceptual challenges.

Shift from wild type to B.1.1.7 variant The renewed increase in cases in both Germany and Poland ("third wave") in late February 2021 was due to the shift from the wild-type variant of the virus to the B.1.1.7 (or Alpha) variant, see Figure 1, panel (c) for estimated shares of the new variant over time. Given earlier observations about the spread of the B.1.1.7 variant in the UK (18) and Denmark, there was public discussion about the likelihood of a re-surgence, but there was considerable uncertainty about the timing and strength (see (19) for a German newspaper article from early February 2021). This was largely due to the limited availability of representative sequencing data. In certain regions of Germany, specifically the city of Cologne (20) and the state of Baden-Württemberg (21), large-scale sequencing had been adopted by late January, but results were considered difficult to extrapolate to the whole of Germany. An updated RKI report on virus variants from 10 February 2020 (22) described a "continuous increase in the share of the VOC B.1.1.7", but cautioned that the data were "subject to biases, e.g., with respect to the selection of samples to sequence" (our translation).

Given the limited available data, and the fact that many approaches had not been designed to accommodate multiple variants, only two of the teams submitting forecasts for Germany opted to account for this aspect (a question which was repeatedly discussed during coordination calls). These exceptions were the `Karlen-pypm` and `LeipzigIMISE-SECIR` models, which starting from 1 March 2021 explicitly accounted for the presence of two variants. As a result, most models did not anticipate the change in trend well and only reacted implicitly once the change became apparent in the data on 27 February 2021. Figure 5 shows the case forecasts of all submitted models and the median ensemble from 15 February, 22 February and 1 March 2021. We also show the two short time series of shares of the B.1.1.7 variant available from Robert Koch Institute at the respective prediction time points.

The `ITWW-county_repro` model was the only one to anticipate a change in trend on 15 February (though slower than the observed one), and adapted quickly to the upward trend in the following week. This model extrapolates recently observed growth or decline at the county-level and aggregates these fine-grained forecasts to the state or national level. Therefore it may have been able to catch a signal of renewed growth, as a handful of German states had already experienced a slight increase in cases in the previous week (e.g., Thuringia and Saxony-Anhalt, see panel (b) of Supplementary Figure 14). However, as illustrated in panel (a) of the same Figure, the ITWW model had also predicted turning points earlier during the same phase of decline in cases, and might generally have a tendency to produce such patterns. Another noteworthy observation in this context is the change in the predictions of the `Karlen-pypm` model. After the extension of the model to account for the B.1.1.7 variant on 1 March, its forecasts changed from the most optimistic to the most pessimistic among all included models (panels b and c of Figure 5).

In Poland, availability of sequencing data was very limited during our study period; as indicated in (14), the GISAID database (15) only contained 2271 sequenced samples for Poland by 29 March 2021. Nonetheless, the `ICM-agentModel` and `MOCOS-agent1` models explicitly took this aspect into account to the degree possible. Again, the `ITWW-county_repro` model was the first to predict a change in overall trends (in this case without having predicted turning points already in the preceding weeks; see Supplementary Figure 8).

Peak of the third wave (cases) In Poland, the third wave reached its peak in the week ending on 3 April 2021. Despite the fact that it coincided with the Easter weekend and thus somewhat unclear data quality, this turnaround was predicted quite well by two Poland-based teams, `MOCOS-agent1` and `ICM-agentModel`. As can be seen from Figure 6, the trajectory of these two models differed substantially from those of most other models, including the ensemble, which predicted a sustained increase. This successful prediction of the turning point was in large part responsible for the good relative performance of `MOCOS-agent1` and

Table 2: Forecast evaluation at the regional level, Germany and Poland (incidence scale, based on RKI/MZ data). Results are averaged over the different regions (states in Germany, voivodeships in Poland). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
Model	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiNow2	694	448	99/192	172/192	1,663	1,120	75/192	147/192	36	24	103/192	177/192	65	43	75/192	161/192
FIAS_FZJ-Epi1Ger	801	518	57/192	146/192	1,732	1,173	48/192	127/192	41	34	21/192	57/192	54	41	24/192	86/192
IHME-CurveFit									47				53			
ITWW-county_repro	996	671	69/192	141/192	2,053	1,392	76/192	149/192	41	34	29/192	89/192	28	20	49/192	133/192
Karlen-pypm	1,251	804	124/192	183/192	2,300	1,432	93/192	177/192	35	22	120/192	180/192	48	28	90/192	183/192
LeipzigIMISE-SECIR			1/12	7/12			2/12	4/12			1/12	5/12			1/12	5/12
USC-SIkJalpha	941	631	86/192	165/192	1,614	1,094	62/192	142/192	32	21	88/192	174/192	43	28	64/192	159/192
KIT-baseline	859	544	80/192	170/192	1,536	1,085	53/192	125/192	38	25	98/192	176/192	57	38	67/192	149/192
KIT-extrapolation_baseline	785	504	91/192	171/192	1,616	1,073	66/192	145/192	49	31	88/192	168/192	76	50	82/192	163/192
KIT-time_series_baseline	1,010	665	76/192	153/192	1,954	1,320	53/192	130/192	59	36	70/192	176/192	93	59	64/192	162/192
KITCOVIDhub-inverse_wis_ensemble	868	523	94/192	186/192	1,744	1,098	73/192	170/192	28	18	108/192	187/192	39	25	96/192	181/192
KITCOVIDhub-mean_ensemble	848	515	96/192	186/192	1,693	1,044	83/192	175/192	28	18	107/192	186/192	41	26	97/192	184/192
KITCOVIDhub-median_ensemble	769	485	109/192	183/192	1,661	1,028	80/192	172/192	29	19	110/192	181/192	36	23	95/192	183/192

Poland

	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
Model	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiNow2	632	426	111/192	171/192	2,081	1,500	78/192	149/192	27	18	94/192	165/192	47	33	85/192	157/192
ITWW-county_repro	1,351	956	70/192	135/192	2,532	1,867	76/192	142/192	45	38	24/192	60/192	56	44	36/192	80/192
USC-SIkJalpha	853	622	77/192	173/192	1,841	1,305	48/192	123/192	20	13	98/192	182/192	27	18	87/192	172/192
KIT-baseline	1,085	647	65/192	161/192	2,028	1,451	29/192	99/192	26	17	97/192	174/192	34	23	70/192	151/192
KIT-extrapolation_baseline	702	457	91/192	171/192	2,069	1,485	66/192	134/192	29	19	89/192	170/192	39	25	82/192	171/192
KIT-time_series_baseline	804	562	102/192	159/192	2,123	1,459	80/192	141/192	32	21	79/192	153/192	44	31	62/192	135/192
KITCOVIDhub-inverse_wis_ensemble	658	399	115/192	188/192	1,786	1,155	83/192	158/192	19	12	101/192	179/192	25	16	97/192	178/192
KITCOVIDhub-mean_ensemble	604	380	117/192	190/192	1,663	1,061	83/192	168/192	19	13	102/192	178/192	26	17	99/192	180/192
KITCOVIDhub-median_ensemble	615	385	123/192	187/192	1,801	1,125	81/192	160/192	20	13	107/192	179/192	26	17	96/192	180/192

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

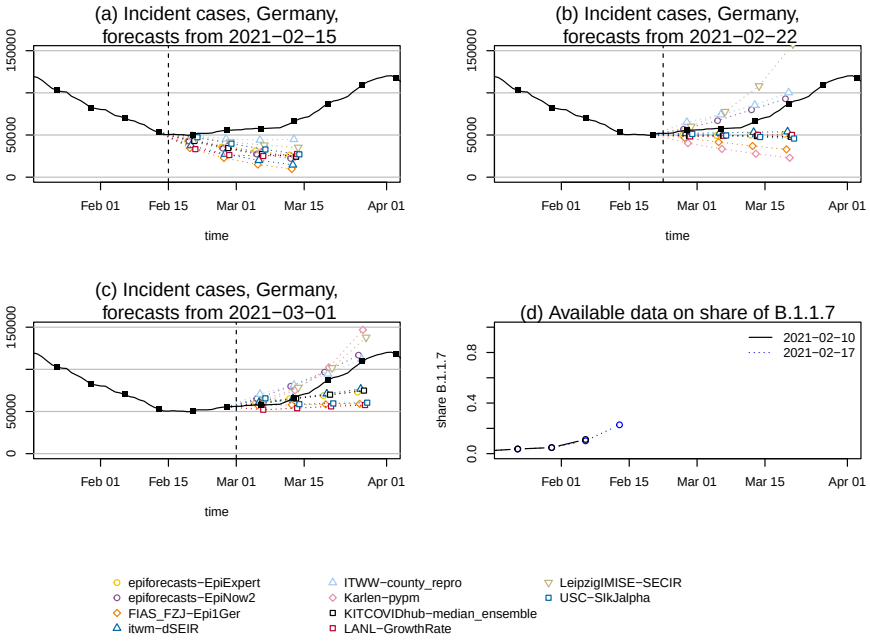

Figure 5: Panels (a)–(c) Point forecasts of cases in Germany, as issued on 15 February, 22 February and 1 March 2021. These dates mark the start of a renewed increase in overall case counts due to the new variant of concern B.1.1.7. Panel (d): Data by RKI on the share of the B.1.1.7 variant as available on the different forecast dates (the next data release by RKI occurred on 3 March).

`ICM-agentModel` at longer horizons reported in Section 2.1. In retrospective discussions, the respective teams noted that the tightening of non-pharmaceutical interventions (NPIs) on 27 March (which they had anticipated) in combination with possible seasonal effects had led them to expect a downward turn.

For Germany, the peak of the third wave occurred only after the end of our pre-specified study period, but we note that numerous models showed strong overshoot as they expected the upward trend to continue. The exact mechanisms underlying the turnaround remain poorly understood (a new set of restrictions referred to as the *Bundesnotbremse* was introduced too late to explain the change on its own).

Changes in trend of deaths In Germany, the study period coincided almost perfectly with a prolonged period of decline in deaths. In Figure 7, panels (a) and (b) show the behaviour of the median ensemble at the beginning and end of this phase. The ensemble had already anticipated a downward turn on 4 January, two weeks before it actually occurred. Following the unexpected strong increase in the following week, it went to extending the upward tendency, before switching back to predicting a turnaround. It seems likely that the irregular pattern in late December and early January is partly due to holiday effects in reporting, and forecast models may have been disturbed by this aspect.

At the end of the downward trend in late March, the ensemble again anticipated the turnaround to arrive earlier than it did, and predicted a more prolonged rise than was observed. Nonetheless, in both cases the ensemble to some degree anticipated qualitative change, and the observed trajectories were well inside the respective 95% prediction intervals (with the exception of the forecast from 4 January; however, this forecast had prospectively been excluded from the analysis as we anticipated reporting irregularities).

In Poland, deaths started to increase in early March after a prolonged period of decay. As can be seen in panel c of Figure 7, the median ensemble had anticipated this change (22 February 2021), but in terms of its point forecast did not initially expect a prolonged upward trend as later observed. Nonetheless, the observed trajectory was contained in the relatively wide 95% prediction intervals (Figures 2 and 3).

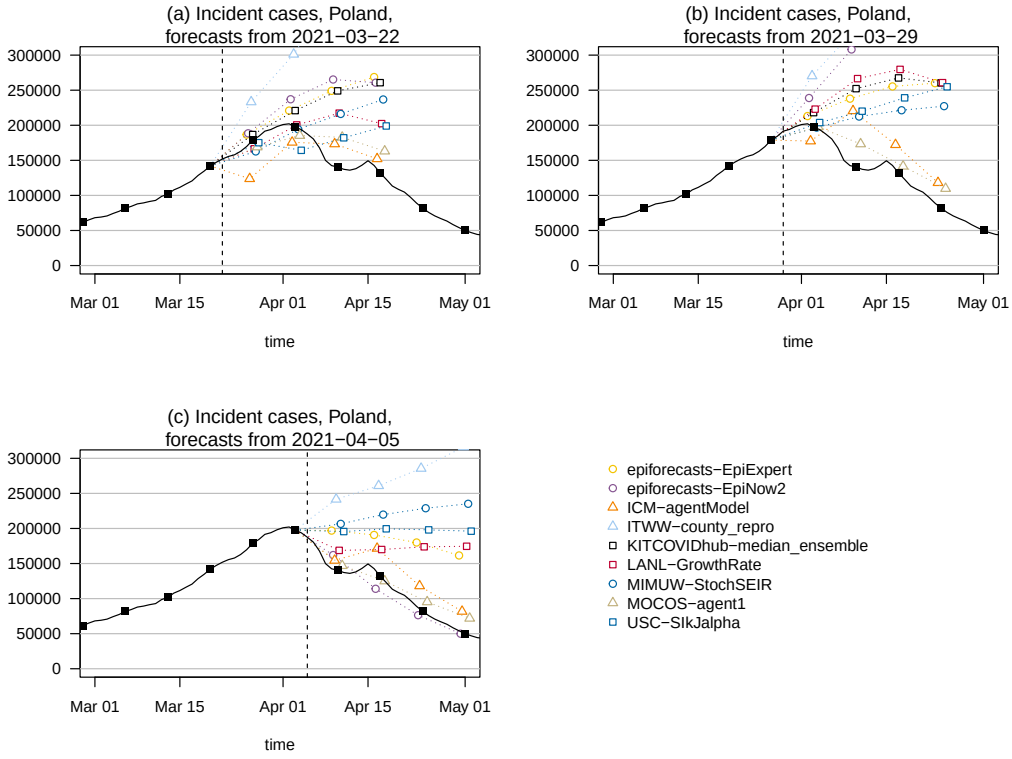

Figure 6: Point forecasts of cases in Poland from 22 March, 29 March and 5 April 2021, surrounding the peak week.

3 Discussion

We presented results from the second and final part of a pre-registered forecast evaluation study conducted in Germany and Poland (January–April 2021). During the period covered in this paper, ensemble approaches yielded very good performance relative to contributed individual models and baseline models. The majority of contributed models was able to outperform a simple last-observation-carried-forward model for most targets and forecast horizons up to four weeks.

The results in this manuscript differ in important aspects from those for our first evaluation period (October–December 2020), when most models struggled to meaningfully outperform the `KIT-baseline` model for cases. Fall 2020 was characterized by rapidly changing non-pharmaceutical intervention measures, making it hard for models to anticipate the case trajectory. Pooled across both study periods, we found ensemble forecasts of deaths to yield satisfactory reliability and clear improvement over baseline models. For cases, however, coverage was clearly below nominal from the two-week horizon onward, and in terms of mean weighted interval scores the ensemble failed to outperform the `KIT-baseline` model three and four weeks ahead. This strengthens our previous conclusion (12) that meaningful case forecasts are only feasible at very short horizons. It also agrees with recent results from the US COVID-19 Forecast Hub (23), which led the organizers to suspend ensemble case forecasts beyond the one-week horizon.

The differences between our two study periods illustrate that performance relative to simple baseline models is strongly dependent on how good a fit these are for a given period. Cases in Germany plateaued during November and early December 2020, making the last-observation-carried-forward strategy of `KIT-baseline` difficult to beat. The second evaluation period was characterized by longer stretches of continued upward or downward trends, making it much easier to beat that baseline. In this situation, however, many models did not achieve strong improvements over the extrapolation approach `KIT-extrapolation_baseline`. Ideally one would wish complex forecast models to outperform each of these different baseline models. However, there are many ways of specifying a "simple" baseline (24), and post-hoc at least one of them will likely be

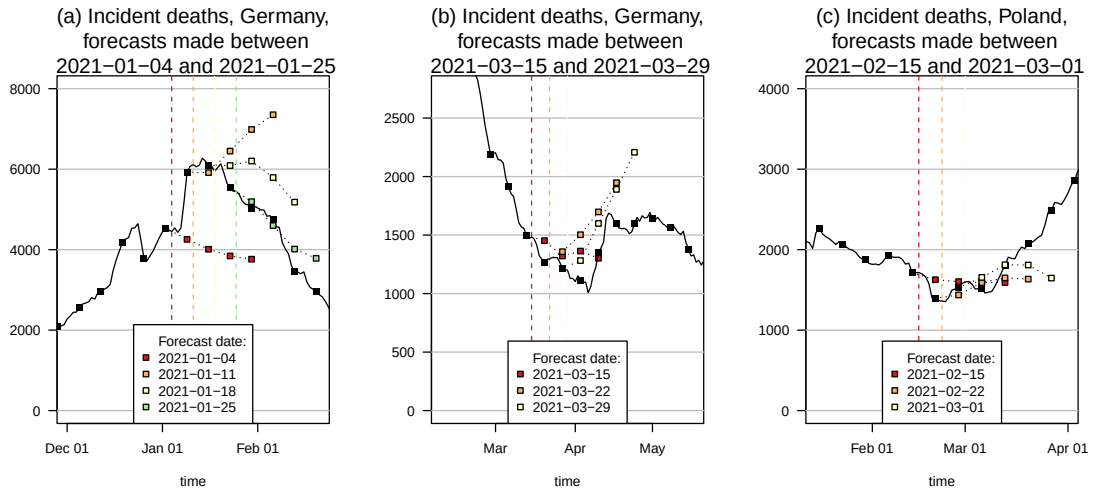

Figure 7: Point forecasts of the median ensemble during changing trends in deaths. Panel (a): Downward turn in Germany, January 2021. Panel (b): Upward turn in Germany, March 2021. Panel (c): Upward turn in Poland, February/March 2021. Different colours represent forecasts made at distinct time points.

in acceptable agreement with the observed trajectory. While the choice of the most meaningful reference remains subject to debate, we believe that the use of a small set of pre-specified baselines as in the present study is a reasonable approach.

An observation made for both the first and the second part of our study is that predicting changing trends in cases is very challenging; turnarounds in death counts are less difficult to anticipate. This finding is shared by other works on short-term forecasts of COVID-19 in real time, see (25) for the UK and (26) for the US. To interpret these insights we note that, in principle, there are two ways of forecasting epidemiological time series:

(i) Applying a mechanistic model to project future spread based on recent trends and other relevant factors like NPIs, population behaviour or vaccination. Models can then predict trend changes based on classical epidemiological mechanisms (depletion of susceptibles) or observed/anticipated changes in surrounding factors, which depending on the model may be treated as exogenous or endogenous.

(ii) Establishing a statistical relationship (often with a mechanistic motivation) to a leading indicators, i.e. a data stream which is informative on the trajectory of the quantity of interest, but available earlier. Changes in the trend of the leading indicator can then help anticipate future turning points in the time series of interest.

Death forecasts belong into the realm of category (ii), with cases and hospitalizations serving as leading indicators. This prediction task has been addressed with considerable success. Case forecasts, on the other hand, typically are based on approach (i), which largely reduces to trend extrapolation, unless models are carefully tuned to changing NPIs (see Table 3). Theoretical arguments on the limited predictability of turning points in such curves have been brought forward (27; 28), and empirical work including ours confirms that this is a very difficult task. The success of the two microsimulation models `MOCOS-agent1` and `ICM-agentModel` in anticipating the downward turn in cases in Poland is encouraging, but remains a rather rare exception. Potential leading indicators to improve case forecasts could be trajectories in other countries (29) or additional data streams on e.g., mobility, insurance claims or web searches. However, the benefits of such data for short-term forecasting thus far have been found to be modest (30). Changes in dominant variants may make changes in overall trends predictable as they arise from the superposition of adverse but stable trends for the different variants. The availability of sequencing data has improved considerably since our study period, but in practice the associated delays may still limit predictability in crucial periods.

We have extensively discussed the difficulties models encountered at turning points, both upward and downward. In the aftermath of such events, epidemic forecasts typically receive increased attention in the general media (see e.g. (31) for coverage of the rapid downward turn in cases in Germany in May 2021). While

important from a subject-matter perspective, this is not without problems from a formal forecast evaluation standpoint. Major turning points are rare events and as such difficult to forecast. Focusing evaluation on solely these instances will benefit models with a strong tendency to predict change, and adapting scoring rules to emphasize these events in a principled way is not straightforward. This problem is known as the *forecaster's dilemma* (32) in the literature and likewise occurs in, e.g., economics and meteorology (see illustrations in Table 1 from (32)).

The present paper marks the end of the German and Polish COVID-19 Forecast Hub as an independently run platform. In April 2021, the European Center for Disease Prevention and Control (ECDC) announced the launch of a European COVID-19 Forecast Hub (4), which has since attracted submissions from more than 30 independent teams. The German and Polish COVID-19 Forecast Hub has been synchronized with this larger effort, meaning that all forecasts submitted to our platform are forwarded to the European repository, while forecasts submitted there are mirrored in our dashboard. In addition, we still collect regional-level forecasts, which are not currently covered in the European Forecast Hub. The adoption of the Forecast Hub concept by ECDC underscores the potential of collaborative forecasting systems with combined ensemble predictions as a key output, along with continuous monitoring of forecast performance. We anticipate that this closer link to public health policy making will enhance the usefulness of this system to decision makers. An important step will be the inclusion of hospitalization forecasts. Due to unclear data access, these had not been tackled in the framework of the German and Polish COVID-19 Forecast Hub, but have recently been added in the new European version.

4 Materials and Methods

The methods described in the following are largely identical to those in the first part (12) of our study, but are presented in abridged form to ensure self-containedness of the present work.

4.1 Targets and submission system

Teams submitted forecasts for weekly incident and cumulative confirmed cases and deaths from COVID-19 via a dedicated public GitHub repository (https://github.com/KITmetricslab/covid19-forecast-hub-de). For certain teams running public dashboards, software scripts were put in place to transfer forecasts to the Forecast Hub repository. Weeks were defined to run from Sunday through Saturday. Each week, teams were asked to submit forecasts using data available up to Monday, with submission possible until Tuesday 3pm Berlin/Warsaw time (the first two daily observations are thus already available at the time of forecasting). Forecasts could either refer to the time series provided by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE; (11)) or those from Robert Koch Institute and the Polish Ministry of Health. All data streams were aggregated by time of reporting, see Supplementary Note 4 of (12) for details. Submissions consisted of a point forecast and 23 predictive quantiles (1%, 2.5%, 5%, 10%, ..., 95%, 97.5%, 0.99) for the incident and cumulative weekly quantities. As in (12), we focus on the targets on the incidence scale. These are easier to compare across the different data sources than cumulative numbers which sometimes show systematic shifts.

4.2 Evaluation metrics

As forecasts were reported in the form of 11 nested central prediction intervals (plus the predictive median), a natural choice for evaluation is the interval score (33). For a central prediction interval $[l, u]$ at the level $(1-\alpha)$, thus reaching from the $\alpha/2$ to the $1-\alpha/2$ quantile, it is defined as

$$\mathrm{IS}_\alpha(F, y) = (u-l) \; + \; \frac{2}{\alpha} \times (l-y) \times \chi(y < l) \; + \; \frac{2}{\alpha} \times (y-u) \times \chi(y > u), \tag{1}$$

where χ is the indicator function and y is the realized value. Here, the first term characterizes the spread of the forecast distribution, the second penalizes overprediction (observations fall below the prediction interval) and the third term penalizes underprediction. To assess the full predictive distribution we use the weighted interval score (WIS; (17)). The WIS is a weighted average of interval scores at different nominal levels and

Table 3: Forecast models contributed by independent external research teams. Abbreviations: NPI: Does the forecast model explicitly account for non-pharmaceutical interventions? Test: Does the model account for changing testing strategies? Variants: Does the model accommodate multiple variants? Age: Is the model age-structured? DE, PL: Are forecasts issued for Germany and Poland, respectively? Regional: Were regional-level forecasts for at least one country submitted? Truth: Which truth data source does the model use? Pr: Are forecasts probabilistic (23 quantiles)? Detailed descriptions of the different models can be found in (12), Supplementary Note 3 and in Supplement A of this article.

Category	Model	NPI	Test	Variants	Age	DE	PL	Regional	Truth	Pr
Agent-based	`ICM-agentModel`	✓	✓	✓	✓		✓		MZ	✓
	`MOCOS-agent1`	✓	✓	✓	✓		✓		JHU	✓
Compartment	`CovidAnalytics-DELPHI`	✓				✓	✓		JHU	✓
	`FIAS_FZJ-Epi1Ger`					✓		✓	RKI	✓
	`itwm-dSEIR`				✓	✓			RKI	✓
	`Karlen-pypm`			✓		✓		✓	RKI	✓
	`LeipzigIMISE-SECIR`	✓	✓	✓		✓		(✓)	RKI	✓
	`MIMUW-StochSEIR`						✓		JHU	✓
	`USC-SIkJalpha`					✓	✓	✓	RKI/MZ	✓
Growth rate/ renewal eq.	`epiforecasts-EpiNow2`					✓	✓	✓	RKI/MZ	✓
	`SDSC_ISG-TrendModel`					✓	✓		JHU	
	`ITWW-county_repro`				✓	✓	✓	✓	RKI/MZ	✓
	`LANL-GrowthRate`					✓	✓		JHU	✓
Human judgement	`epiforecasts-EpiExpert`	(✓)	(✓)	(✓)	(✓)	✓	✓		RKI/MZ	✓
Forecast ensemble	`Imperial-ensemble2`[7]					✓	✓		RKI	✓

the absolute error. For eleven prediction intervals it is defined as

$$\mathrm{WIS}(F, y) = \frac{1}{11.5} \times \left(\frac{1}{2} \times |y - m| \; + \; \sum_{k=1}^{11} \left(\frac{\alpha_k}{2} \times \mathrm{IS}_{\alpha_k}(F, y) \right) \right), \tag{2}$$

where m is the predictive median. The WIS is a well-known approximation of the continuous ranked probability score (CRPS; (33)) and generalizes the absolute error to probabilistic forecasts. Its values can be interpreted on the natural scale of the data and measures how far the observed value y is from the predictive distribution (lower values are thus better). For deterministic one-point forecasts the WIS reduces to the absolute error. A useful property of the WIS is that it inherits the decomposition of the interval score into forecast spread, overprediction and underprediction, which makes average scores more interpretable. As secondary measures of forecast quality we use the absolute error to assess the central tendency of forecasts and interval coverage rates of 50% and 95% prediction intervals to assess calibration.

As specified in our study protocol, whenever forecasts from a model were missing for a given week, we imputed the score with the worst (largest) value achieved by any other model for the respective week and target. However, almost all teams provided complete sets of forecasts and very few scores needed imputation.

4.3 Individual models

During the evaluation period, forecasts from fifteen different models run by fourteen independent teams of researchers were collected. Thirteen of these were already available during the first part of our preregistered study, see Table 3 and Supplementary Note 3 of (12) for detailed descriptions. Table 3 provides a slightly extended summary of model properties, including the two new models, `itwm-dSEIR` and `Karlen-pypm`; a more detailed description of the latter can be found in Supplement A.

During the evaluation period, only the `ICM-agentModel` explicitly accounted for vaccinations (given the low realized vaccination coverage by the end of the study period this aspect likely had limited impact). Only four models (`ICM-agentModel`, `Karlen-pypm`, `LeipzigIMISE-SECIR` and `MOCOS-agent1`, all only for certain weeks) explicitly accounted for the presence of multiple variants.

To put the results achieved by the submitted models into perspective, the Forecast Hub team generated forecasts from three simple reference models: a last-observation-carried-forward model (`KIT-baseline`), a multiplicative extrapolation model which continues exponential growth or decline based on the last three observations (`KIT-extrapolation_baseline`) and an exponential smoothing time series baseline (`KIT-time_series_baseline`)

which has been taken from (34). Detailed descriptions can be found in (12), Supplementary Note 2. As a further external comparison we added publicly available death forecasts by the Institute for Health Metrics and Evaluation (IHME, University of Washington (35); available under the CC BY-NC 4.0 license). Here, we always used the most recent prediction available on a given forecast date.

4.4 Forecast ensembles

The Forecast Hub team used the submitted forecasts to generate three different ensemble forecasts:

`KITCOVIDhub-median_ensemble` The α-quantile of the ensemble forecast is obtained as the median of the α-quantiles of the member forecasts.

`KITCOVIDhub-mean_ensemble` The α-quantile of the ensemble forecast is obtained as the mean of the α-quantiles of the member forecasts.

`KITCOVIDhub-inverse_wis_ensemble` The α-quantile of the ensemble forecast is a convex combination of the α-quantiles of the member forecasts. The weights are chosen inversely to the mean WIS value obtained by the member models over the last six evaluated forecasts (last three one-week-ahead, last two two-week-ahead, last three-week-ahead; missing scores are imputed by the worst score achieved by any model for the respective target). This is done separately for each time series to be predicted.

Inverse score weighting has recently also been employed by (36) who found it to perform well in a re-analysis of forecasts from the US COVID-19 Forecast Hub. In the study protocol, the median ensemble was defined as our primary ensemble approach (10), which is why we displayed this version in all figures and focused our discussion on it. We have previously discussed advantages and disadvantages of the different ensemble approaches in (12).

There were no formal inclusion criteria other than completeness of the submitted set of 23 quantiles. The Forecast Hub team did, however, occasionally exclude forecasts with highly implausible central tendency or degree of dispersion. These exclusions have been documented in the Forecast Hub platform.

Data availability

The forecast data generated in this study have been deposited in a GitHub repository (https://github.com/KITmetricslab/covid19-forecast-hub-de), with a stable Zenodo release available under accession code 5608390 https://zenodo.org/record/5608390#.YYFxdJso9H4. This repository also contains all truth data used for evaluation. Forecasts can be visualised interactively at https://kitmetricslab.github.io/forecasthub/.

Code availability

Codes to reproduce figures and tables are available at https://github.com/KITmetricslab/analyses_de_pl2, with a stable version at https://zenodo.org/record/5639514#.YYF1aZso9H4. The results presented in this paper have been generated using the release *preprint2* of the repository https://github.com/KITmetricslab/covid19-forecast-hub-de, see above for the link to the stable Zenodo release.

Competing interests

The authors declare no competing interests.

References

[1] Ray, E. L. *et al.* Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. *medRxiv* (2020). URL https://www.medrxiv.org/content/early/2020/08/22/2020.08.19.20177493.

[2] Cramer, E. *et al.* Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *medRxiv* (2021). URL https://www.medrxiv.org/content/early/2021/02/05/2021.02.03.21250974.

[3] Borchering, R. K. *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* **70**, 719–724 (2021).

[4] European Center for Disease Prevention and Control. Forecasting COVID-19 cases and deaths in Europe – new hub will support European pandemic planning. Press release, published online 22 April 2021, https://www.ecdc.europa.eu/en/news-events/forecasting-covid-19-cases-and-deaths-europe-new-hub.

[5] McGowan, C. J. *et al.* Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Scientific Reports* **9**, 683 (2019).

[6] Del Valle, S. *et al.* Summary results of the 2014–2015 DARPA Chikungunya Challenge. *BMC Infectious Diseases* **18**, 245 (2018).

[7] Johansson, M. A. *et al.* An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences* **116**, 24268–24274 (2019).

[8] Nature Publishing Group. Editorial: Developing infectious disease surveillance systems. *Nature Communications* **11**, 4962 (2020).

[9] Dirnagl, U. Politikberatung, bis der Elefant mit dem Rüssel wackelt! *Laborjournal* **5/2021**, 22–24 (2021).

[10] Bracher, J., the German and Polish COVID-19 Forecast Hub Team & Participants. Study protocol: Comparison and combination of real-time COVID19 forecasts in Germany and Poland. Deposited 8 October 2020, Registry of the Open Science Foundation, https://osf.io/k8d39 (2020).

[11] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20**, 533–534 (2020).

[12] Bracher, J. *et al.* A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nature Communications* **12**, 5173 (2021).

[13] Robert Koch Institute. Bericht zu Virusvarianten von SARS-CoV-2 in Deutschland, insbesondere zur Variant of Concern (VOC) B.1.1.7, 31 March 2021. Available at https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/Bericht_VOC_2021-03-31.pdf (2021).

[14] MI2 Data Lab, Warsaw University of Technology. Monitor of SARS-CoV-2 variants, version 2021-05-05 (2021). Available online at https://monitor.crs19.pl/2021-05-05/poland/?lang=en.

[15] GISAID Initiative. Enabling rapid and open access to epidemic and pandemic virus data – tracking of variants (2021). Available online, https://www.gisaid.org/hcov19-variants/.

[16] Hale, T. *et al.* A global panel database of pandemic policies (Oxford COVID-19 government response tracker). *Nature Human Behaviour* **5**, 529–538 (2021).

[17] Bracher, J., Ray, E. L., Gneiting, T. & Reich, N. G. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology* **17**, e1008618 (2021).

[18] Davies, N. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372** (2021). URL https://science.sciencemag.org/content/372/6538/eabg3055. https://science.sciencemag.org/content/372/6538/eabg3055.full.pdf.

[19] Berndt, C., Endt, C. & Müller-Hansen, S. Die unsichtbare Welle. *Süddeutsche Zeitung* (2021). Published online, 5 February 2021, https://www.sueddeutsche.de/wissen/coronavirus-mutante-b117-daten-1.5197700.

[20] Fischer-Fels, J. Erste Hochrechnung zur Verbreitung der Coronamutationen. *Ärzteblatt* (2021). 3 February 2021, available at https://www.aerzteblatt.de/nachrichten/120768/Erste-Hochrechnung-zur-Verbreitung-der-Corona-Mutationen.

[21] Landesgesundheitsamt Baden Württemberg. Tagesbericht COVID-19, Montag 8.2.2021 (2021). Available online at https://www.gesundheitsamt-bw.de/fileadmin/LGA/_DocumentLibraries/SiteCollectionDocuments/05_Service/LageberichtCOVID19/COVID_Lagebericht_LGA_210208.pdf.

[22] Robert Koch Institute. Bericht zu Virusvarianten von SARS-CoV-2 in Deutschland, insbesondere zur Variant of Concern (VOC) B.1.1.7, update 10 February 2021. Available at https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/Bericht_VOC_2021-02-10.pdf (2021).

[23] Reich, N., Tibshirani, R., Ray, E. & Rosenfeld, R. On the predictability of COVID-19. Blog post, International Institute of Forecasters, https://forecasters.org/blog/2021/09/28/on-the-predictability-of-covid-19/ (2021).

[24] Keyel, A. C. & Kilpatrick, A. M. Probabilistic evaluation of null models for West Nile Virus in the United States. *bioRxiv* (2021). URL https://www.biorxiv.org/content/early/2021/07/26/2021.07.26.453866. https://www.biorxiv.org/content/early/2021/07/26/2021.07.26.453866.full.pdf.

[25] Funk, S. *et al.* Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. *medRxiv* (2020). URL https://www.medrxiv.org/content/early/2020/11/13/2020.11.11.20220962.

[26] Ray, E. L. *et al.* Challenges in training ensembles to forecast COVID-19 cases and deaths in the United States. Blog post, International Institute of Forecasters, https://forecasters.org/blog/2021/04/09/challenges-in-training-ensembles-to-forecast-covid-19-cases-and-deaths-in-the-united-states/ (2021).

[27] Castro, M., Ares, S., Cuesta, J. & Manrubia, S. The turning point and end of an expanding epidemic cannot be precisely forecast. *Proceedings of the National Academy of Sciences* **117**, 26190–26196 (2020). URL https://www.pnas.org/content/117/42/26190. https://www.pnas.org/content/117/42/26190.full.pdf.

[28] Wilke, C. O. & Bergstrom, C. T. Predicting an epidemic trajectory is difficult. *Proceedings of the National Academy of Sciences* **117**, 28549–28551 (2020). URL https://www.pnas.org/content/117/46/28549. https://www.pnas.org/content/117/46/28549.full.pdf.

[29] Harvey, A. Time series modeling of epidemics: Leading indicators, control groups and policy assessment. Tech. Rep. 2114, Cambridge Working Papers in Economics (2021). Available online at https://www.repository.cam.ac.uk/handle/1810/318300.

[30] McDonald, D. J. *et al.* Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction? *medRxiv* (2021). URL https://www.medrxiv.org/content/early/2021/06/25/2021.06.22.21259346.

[31] Berndt, C., Hametner, M., Kruse, B., Müller-Hansen, S. & Witzenberger, B. Ist die dritte Welle überstanden? *Süddeutsche Zeitung* (2021). Published online, 4 May 2020, https://www.sueddeutsche.de/gesundheit/corona-infektionen-trendwende-modellierungen-1.5284545.

[32] Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. & Gneiting, T. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science* **32**, 106–127 (2017).

[33] Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378 (2007).

[34] Petropoulos, F. & Makridakis, S. Forecasting the novel coronavirus COVID-19. *PLOS ONE* **15**, e0231236 (2020).

[35] IHME COVID-19 Forecasting Team. Modeling COVID-19 scenarios for the United States. *Nature Medicine* **27**, 94–105 (2021).

[36] Taylor, J. W. & Taylor, K. S. Combining probabilistic forecasts of COVID-19 mortality in the United States. *European Journal of Operational Research* (2021). URL https://www.sciencedirect.com/science/article/pii/S0377221721005609.

[37] Kheifetz, Y., Kirsten, H. & Scholz, M. On the parametrization of epidemiologic models – lessons from modelling covid-19 epidemic (2021). 2109.11916.

[38] Karlen, D. Characterizing the spread of CoViD-19. https://arxiv.org/abs/2007.07156 (2020).

Acknowledgements

J. Bracher, M. Schienle and T. Gneiting acknowledge support from the Helmholtz Foundation via the SIMCARD Information and Data Science Pilot Project. T. Gneiting and D. Wolffram are grateful for support by the Klaus Tschira Foundation. D. Wolffram's contribution was moreover supported by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health'.' N.I. Bosse was supported by the Health Protection Research Unit (grant code NIHR200908). S. Funk and S. Abbott were supperted by the Wellcome Trust (210758/Z/18/Z). The itwm-dSEIR forecasting team (J. Fiedler, N. Leithäuser, J. Mohring) was supported by the Ministry of Health and Science of Rhineland Palatinate and the Fraunhofer Anti-Corona Program. S. Bhatia acknowledges funding from the Wellcome Trust (219415). Work on the ICM UW epidemiological model (J.M Nowosielski, M. Radwan, F. Rakowski) was supported by the Polish Minister of Science and Higher Education grant 51/WFSN/2020 given to the University of Warsaw. Development of the IMISE-SECIR model (Y. Kheifetz, H. Kirsten, S. Scholz) was funded in the framework of the project SaxoCOV (Saxonian COVID-19 Research Consortium). SaxoCOV was co-financed with tax funds on the basis of the budget passed by the Saxon state parliament. Model presentation was funded by the NFDI4Health Task Force COVID-19 (www.nfdi4health.de/task-force-covid-19-2) within DFG project LO-342/17-1.

We thank Dean Karlen for contributing forecasts and the Institute for Health Metrics and Evaluation, University of Washington, for making forecasts publicly available under a free license. We are moreover grateful for support and advice from the organizing team of the US COVID-19 Forecast Hub.

The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the institutions they are affiliated with.

Author contributions

JB, DW, TG and MSe conceived the study with advice from AU. JB, DW, JD, KG and JK put in place and maintained the forecast submission and processing system. AU coordinated the creation of an interactive visualization tool. JB performed the evaluation analyses with inputs from DW, TG, MSe and members of various teams. SA, MVB, DB, SB, MB, NIB, JPB, LC, GF, JFr, JFn, SF, AG, KG, SH, TH, YK, HK, TK, EK, NL, MLL, JHM, BM, IJM, JMe, JMg, PN, JMN, TO, MR, FR, MS, SS, and AS contributed forecasts (see list of contributors by team). JB, TG and MSe wrote the manuscript. All teams and members of the coordinating team provided feedback on the manuscript and descriptions of the respective models.

List of contributors by team

`CovidAnalytics-DELPHI`: Michael Lingzhi Li (Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA), Dimitris Bertsimas, Saksham Soni (both Sloan School of Management, Massachusetts Institute of Technology, Cambridge, USA)

`epiforecasts-EpiExpert` and `epiforecasts-EpiNow2`: Sam Abbott, Nikos I. Bosse, Sebastian Funk (all London School of Hygiene and Tropical Medicine, London, UK)

`FIAS_FZJ-Epi1Ger`: Maria Vittoria Barbarossa (Frankfurt Institute for Advanced Studies, Frankfurt, Germany), Jan Fuhrmann (Institute of Applied Mathematics, University of Heidelberg, Heidelberg, Germany), Jan H. Meinke (Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany)

`SDSC_ISG-TrendModel`: Antoine Flahault, Elisa Manetti, Kristen Namigai (all Institute of Global Health, Faculty of Medicine, University of Geneva, Geneva, Switzerland), Christine Choirat, Benjamin Bejar Haro, Ekaterina Krymova, Gavin Lee, Guillaume Obozinski, Tao Sun (all Swiss Data Science Center, ETH Zurich and EPFL Lausanne, Switzerland), Dorina Thanou (Center for Intelligent Systems, EPFL, Lausanne Switzerland)

`ICM-agentModel`: Filip Dreger, Łukasz Górski, Magdalena Gruziel-Słomka, Artur Kaczorek, Antoni Moszyński, Karol Niedzielewski, Jedrzej Nowosielski, Maciej Radwan, Franciszek Rakowski, Marcin Semeniuk, Jakub Zieliński (all Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland), Rafał Bartczuk (Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw and Institute of Psychology, John Paul II Catholic University of Lublin, Lublin, Poland), Jan Kisielewski (Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw and Faculty of Physics, University of Białystok)

`Imperial-ensemble2`: Sangeeta Bhatia (MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics (J-IDEA), Imperial College, London, UK), Pierre Nouvellet (School of Life Sciences, University of Sussex, Brighton, UK)

`itwm-dSEIR`: Michael Burger, Robert Feßler, Jochen Fiedler, Michael Helmling, Karl-Heinz Küfer, Neele Leithäuser, Jan Mohring, Johanna Schneider, Anita Schöbel, Michael Speckert, Raimund Wegener, Jarosław Wlazło (all Fraunhofer Institute for Industrial Mathematics, Kaiserslautern, Germany)

`ITWW-county_repro`: Przemyslaw Biecek (Warsaw University of Technology, Warsaw, Poland), Viktor Bezborodov, Marcin Bodych, Tyll Krueger (all Wroclaw University of Science and Technology, Poland), Jan Pablo Burgard (Economic and Social Statistics Department, University of Trier, Germany), Stefan Heyder, Thomas Hotz (both Institute of Mathematics, Technische Universität Ilmenau, Ilmenau, Germany)

`LeipzigIMISE-SECIR`: Yuri Kheifetz, Holger Kirsten, Markus Scholz (all Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany)

`MIMUW-StochSEIR`: Anna Gambin, Krzysztof Gogolewski, Błażej Miasojedow, Ewa Szczurek (all Institute of Informatics, University of Warsaw, Warsaw, Poland), Daniel Rabczenko, Magdalena Rosińska (Polish National Institute of Public Health – National Institute of Hygiene)

`MOCOS-agent1`: Marek Bawiec, Viktor Bezborodov, Marcin Bodych, Radosław Idzikowski, Tyll Krueger, Tomasz Ożański, Barbara Pabjan, Ewaryst Rafajłlowicz, Ewa Skubalska-Rafajłowicz, Wojciech Rafajłlowicz (all Wroclaw University of Science and Technology, Poland), Przemyslaw Biecek (Warsaw University of Technology), Agata Migalska (Wroclaw University of Science and Technology, Poland and Nokia Solutions and Networks, Wroclaw, Poland), Ewa Szczurek (University of Warsaw)

`USC-SIkJalpha` Ajitesh Srivastava, Frost Tianjian Xu (both University of Southern California, Los Angeles, USA)

Supplementary Materials for Bracher et al (2021): National and subnational short-term forecasting of COVID-19 in Germany and Poland, early 2021

A Detailed description of new models

We only provide detailed descriptions of models which were added to our project for the second evaluation period. Descriptions for the other models can be found in Supplementary Note 3 of (12). A more detailed documentation of the `LeipzigIMISE-SECIR` model which had not been available at the appearance of (12) can be found in (37).

itwm-dSEIR Fraunhofer-ITWM's predictions are based on a cohort model that groups people according to four age groups and according to the status infected, detected and since 19 April successfully vaccinated (i.e., this extension was added after the evaluation period). The dynamics of the epidemic are described by integral equations, assuming an infectious period with fixed onset, end and infectivity. The most important parameters are contact rates between age groups, detection rates and times, and death rates and times, which are adjusted to the historical data of the RKI. For forecasts, the simulation is continued with the parameters determined for the last week. In principle, the forecast quality could be improved by anticipating the effects of events such as the end of public holidays on contact and detection rates. However, this is not yet done in the automatic submissions. All calculations use automatic differentiation. This speeds up parameter adjustment and allows for error estimates. The latter are determined by comparing counted and simulated cases and by matching the empirical standard deviations with the standard deviations predicted by the calculated sensitivities. The model is described in detail in https://www.itwm.fraunhofer.de/de/presse-publikationen/presseinformationen/2021/2021-06-22_Dritte_Welle_Starker-Effekt-von-Schnelltests-an-Schulen.html.

Karlen-pypm The python Population Modeller (pyPM, (38)), is a mechanistic modeling framework to describe viral spread via discrete-time difference equations. In a pyPM model, different population objects are connected by a list of directional connector objects. The adjustable parameters of the model are stored in parameter objects. The core of the model consists of a model of the infection cycle involving the susceptible, infected (but not yet contagious) and contagious parts of the population. The contagious population is modelled in more detail by introducing symptomatic, test-positive, hospitalized (normal ward and ICU) and deceased populations. The model takes time series of cases, deaths and intensive care occupancy as data inputs. Forecasts are generated at the regional level (German states) first and subsequently aggregated to the national level. Starting from 1 March 2021, the model was stratified into spread of the wild type of the virus and the B.1.1.7 variant, and integrated genetic sequencing data on their respective importance.

B Additional forecast visualizations

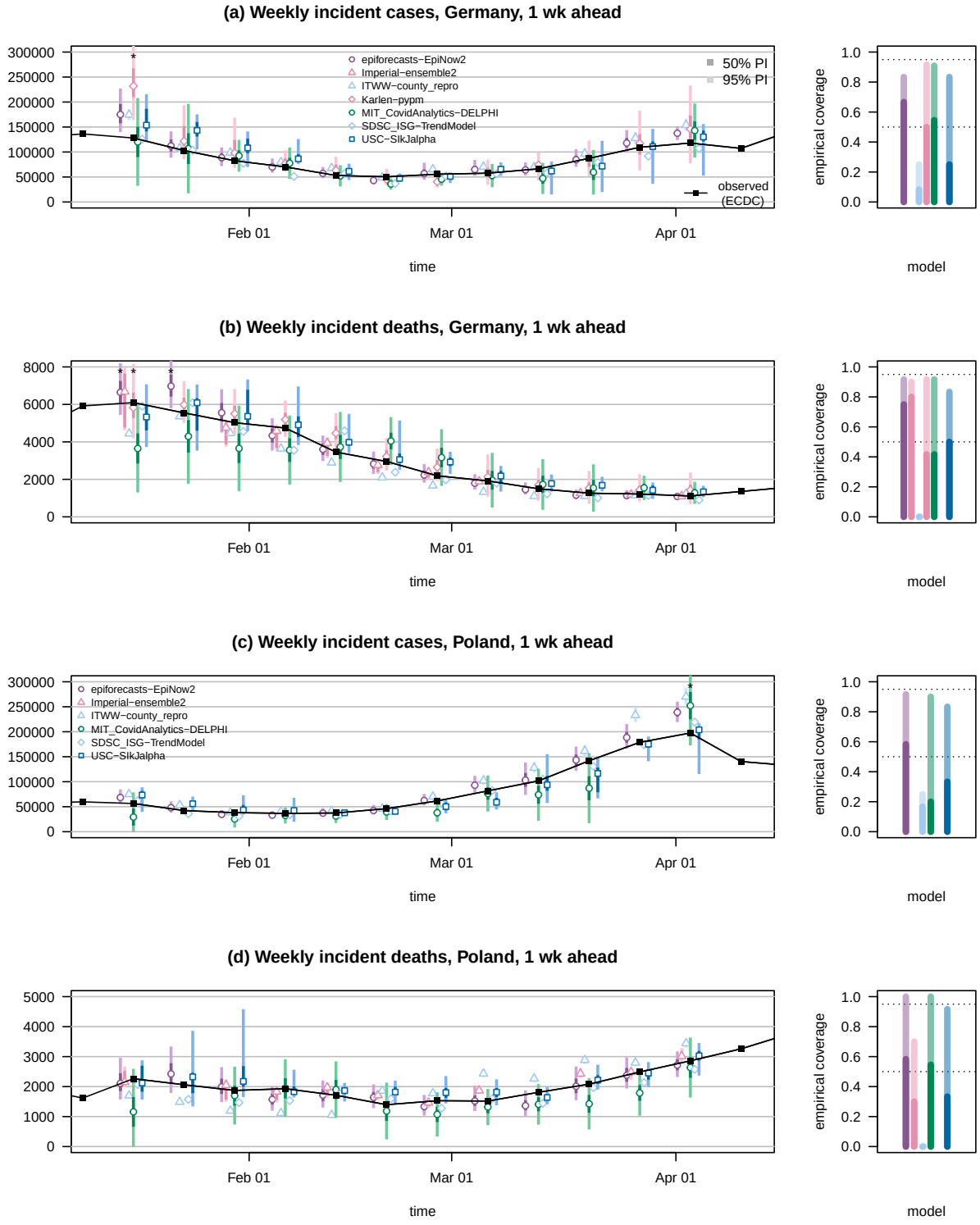

Figure 8: One-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland. Asterisks mark prediction intervals exceeding the upper plot limit. The figure shows forecasts from models not displayed in Figure 2

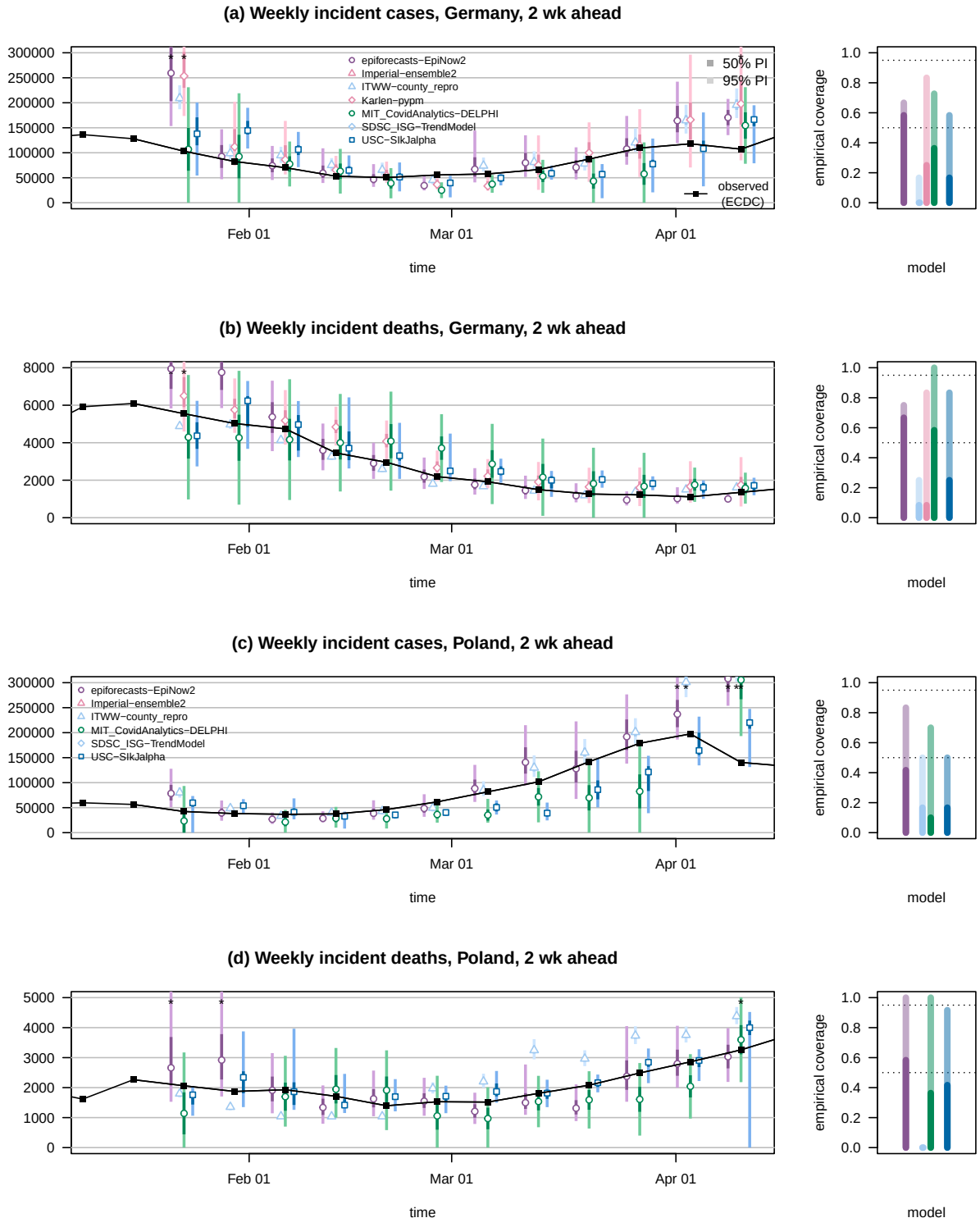

Figure 9: Two-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 8. Asterisks mark prediction intervals exceeding the upper plot limit.

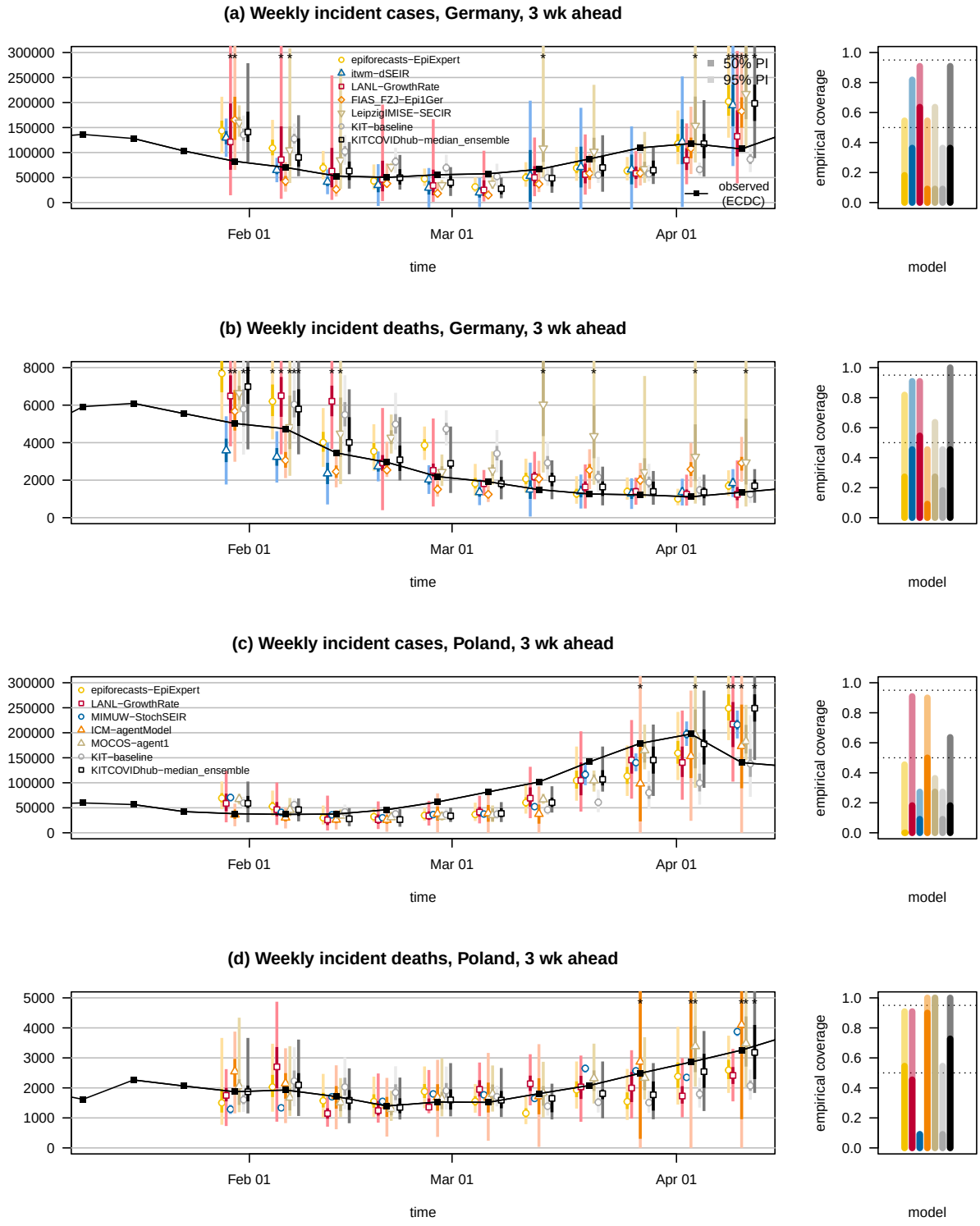

Figure 10: Three-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 2. Asterisks mark prediction intervals exceeding the upper plot limit.

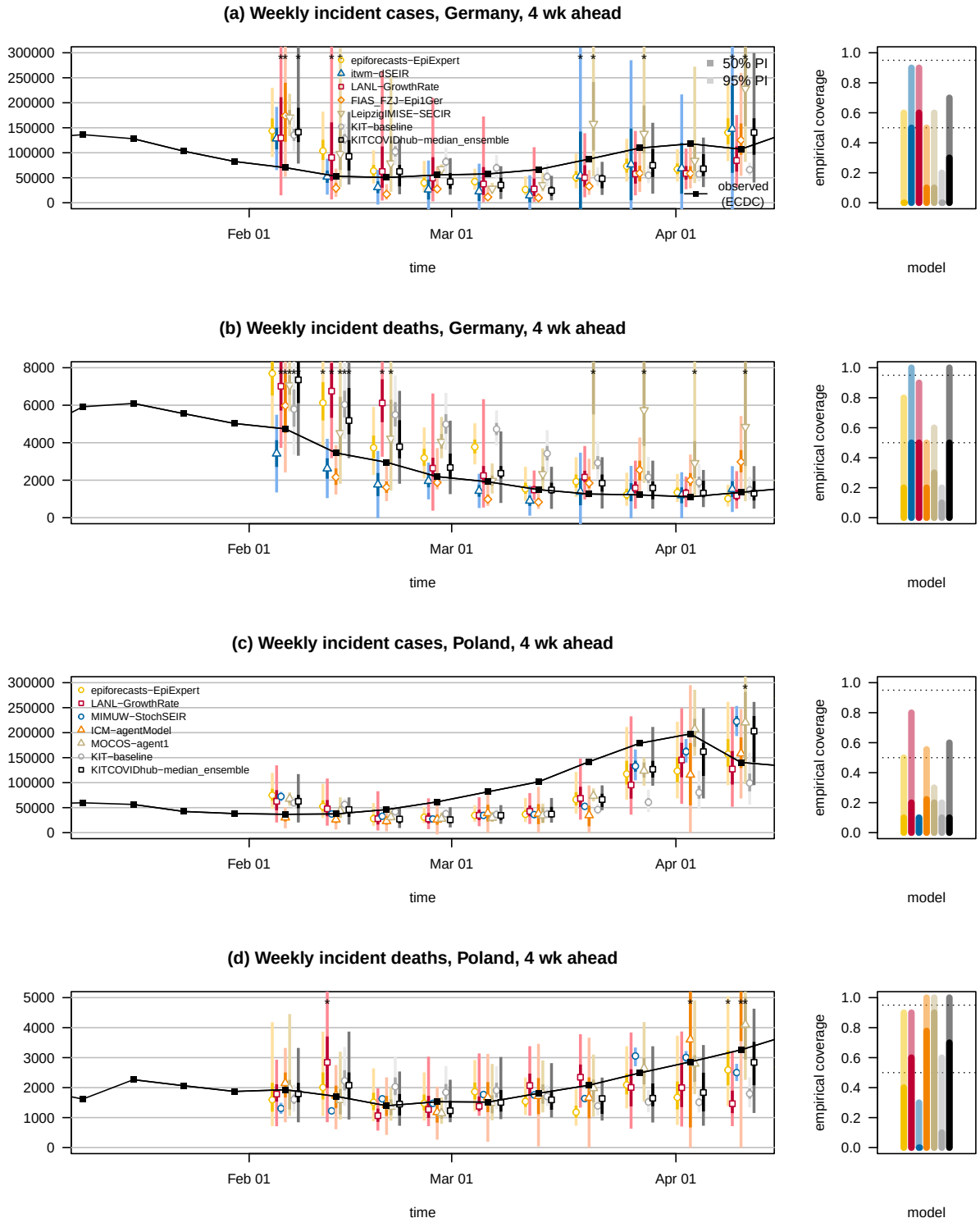

Figure 11: Four-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 2. Asterisks mark prediction intervals exceeding the upper plot limit.

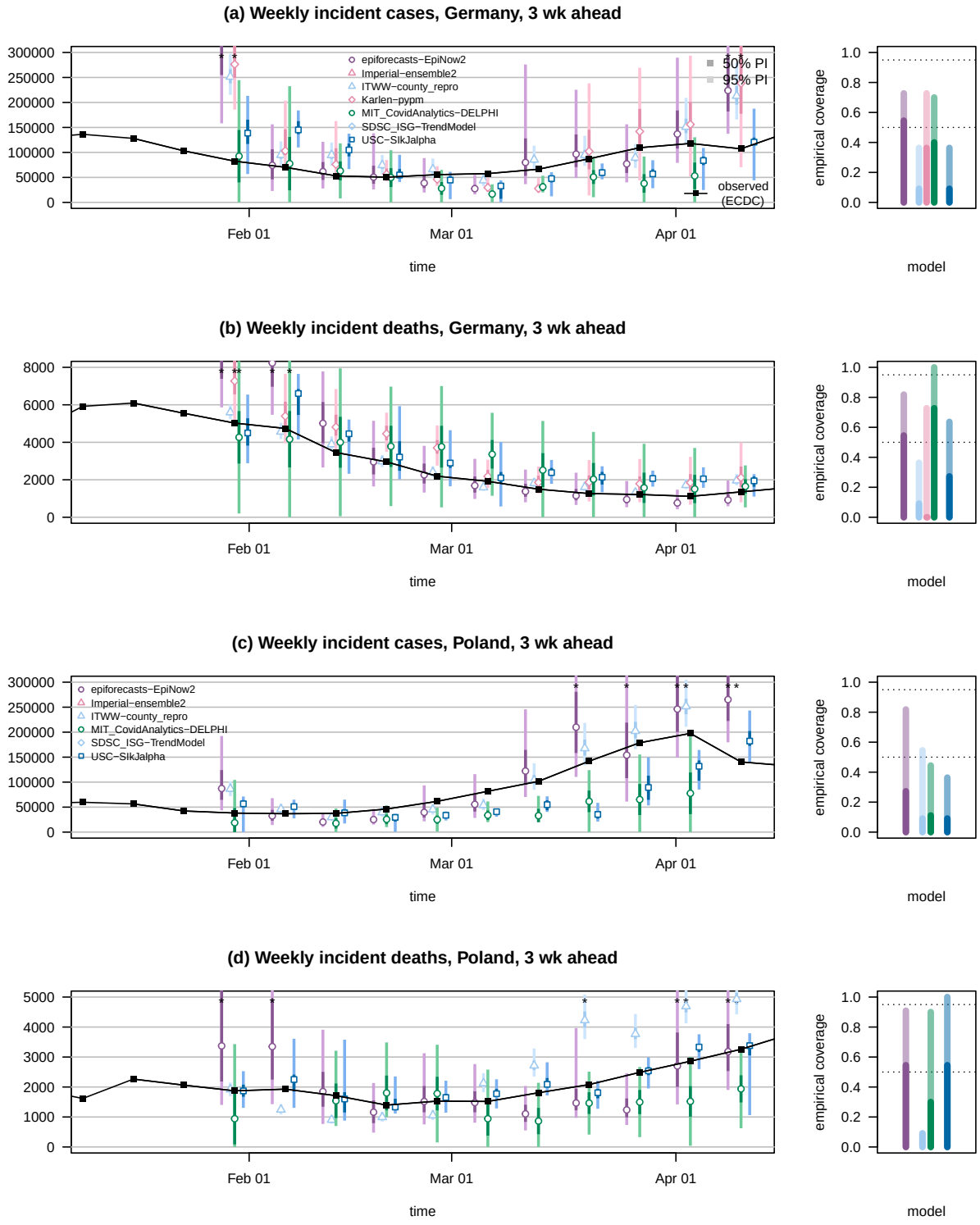

Figure 12: Three-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 8. Asterisks mark prediction intervals exceeding the upper plot limit.

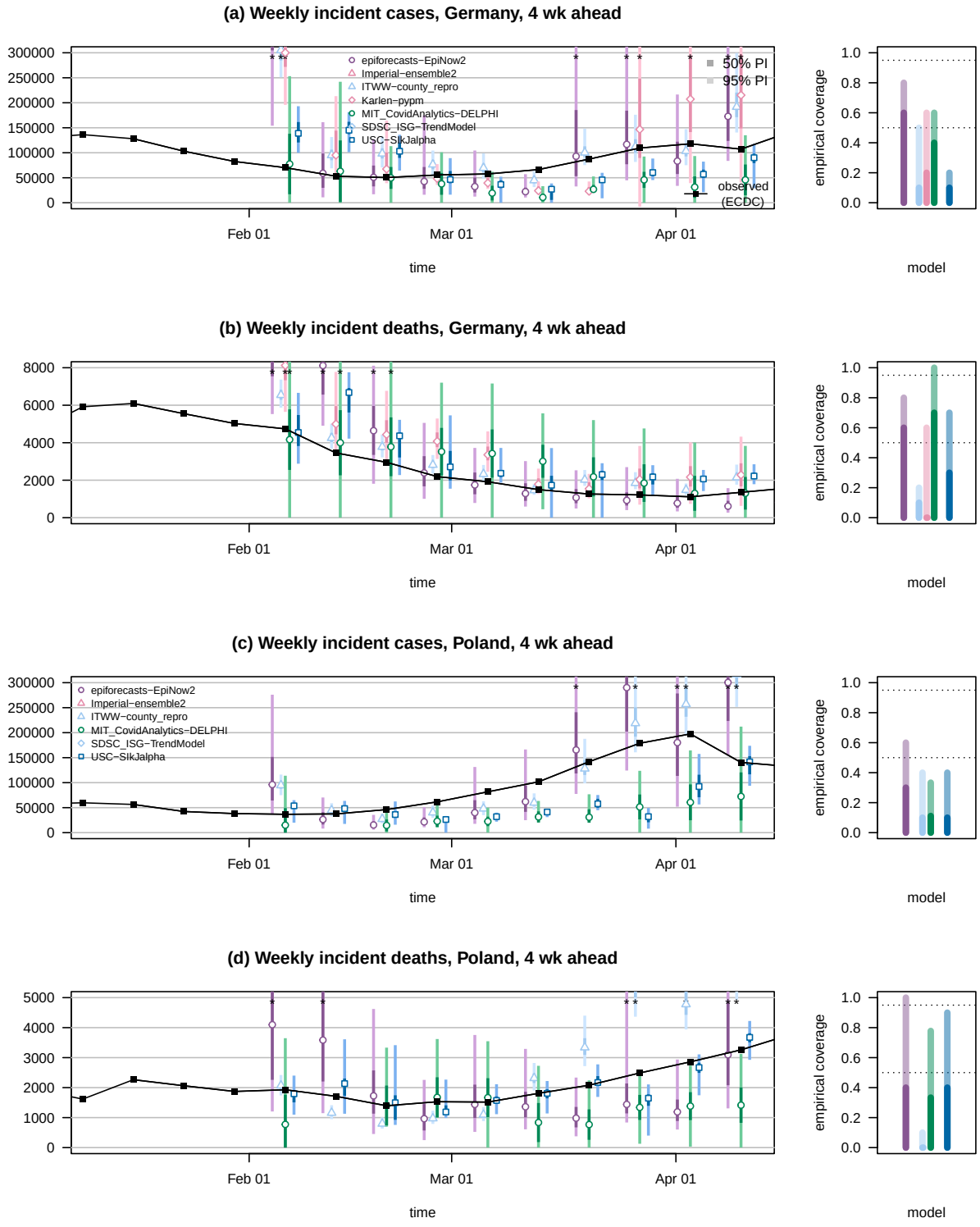

Figure 13: Four-week-ahead forecasts of confirmed cases and deaths from COVID-19 in Germany and Poland, same models as displayed in Figure 8. Asterisks mark prediction intervals exceeding the upper plot limit.

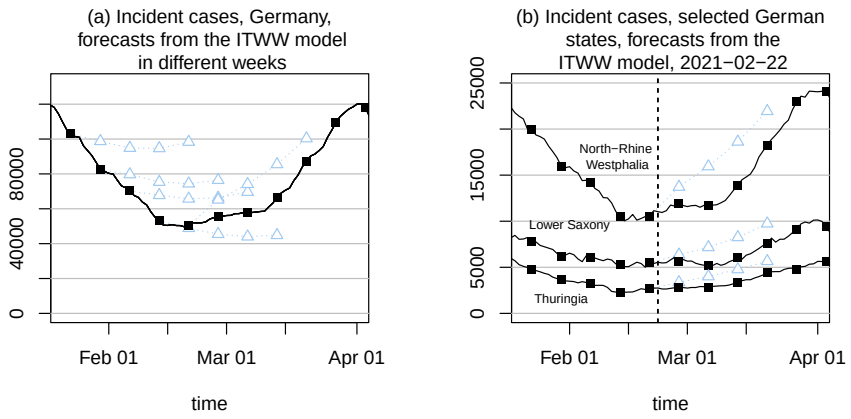

Figure 14: **a** Forecasts of cases in Germany by the `ITWW-county_repro` model, 25 January to 22 February 2021. **b** Forecasts for cases in selected German states by the `ITWW-county_repro` model, 22 February 2021.

C Additional summary tables on forecast evaluation

Table 4: Forecast evaluation for Germany and Poland, 3 and 4 weeks ahead (incidence scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	29,111	20,140	3/12	7/12	37,879	24,429	1/12	8/12	712	463	3/12	10/12	925	614	3/12	10/12
epiforecasts-EpiNow2	51,095	36,966	6/12	9/12	76,218	54,216	6/12	8/12	940	653	6/12	9/12	1,234	834	6/12	9/12
FIAS_FZJ-Epi1Ger	39,102	25,939	1/12	7/12	52,516	35,472	1/12	7/12	920	625	1/12	6/12	1,071	688	2/12	6/12
IHME-CurveFit									839				967			
Imperial-ensemble2																
itwm-dSEIR	31,177	20,772	4/12	10/12	46,895	29,388	5/12	11/12	547	348	5/12	11/12	588	366	5/12	12/12
ITWW-county_repro	47,796	38,477	1/12	4/12	66,520	54,070	1/12	5/12	350	266	1/12	4/12	753	599	1/12	2/12
Karlen-pypm	56,242	40,195	4/12	9/12	89,745	63,976	2/12	8/12	933	618	0/12	9/12	1,309	905	0/12	8/12
LANL-GrowthRate	23,760	17,577	8/12	11/12	29,959	21,224	8/12	11/12	715	450	6/12	11/12	1,037	651	6/12	11/12
LeipzigIMISE-SECIR	49,422	33,952	1/12	8/12	79,741	54,914	1/12	8/12	1,570	1,107	3/12	8/12	2,493	1,773	3/12	8/12
MIT_CovidAnalytics-DELPHI	*28,404	*22,268	5/11	8/11	*37,661	*29,934	5/11	7/11	713	459	9/12	12/12	688	486	9/12	12/12
SDSC_ISG-TrendModel																
USC-SIkJalpha	35,434	26,133	1/12	5/12	46,612	36,819	2/12	3/12	762	492	3/12	8/12	947	617	3/12	8/12
KIT-baseline	34,871	26,554	1/12	5/12	44,270	35,228	0/12	3/12	1,168	834	2/12	6/12	1,494	1,132	1/12	4/12
KIT-extrapolation_baseline	36,817	23,333	2/12	10/12	50,965	34,906	2/12	6/12	1,331	865	3/12	9/12	1,893	1,273	2/12	9/12
KIT-time_series_baseline	47,164	33,930	2/12	7/12	60,577	46,342	4/12	7/12	1,299	983	6/12	9/12	1,791	1,432	5/12	10/12
KITCOVIDhub-inverse_wis_ensemble	34,038	21,456	4/12	11/12	48,943	30,201	5/12	10/12	494	320	9/12	12/12	678	450	7/12	12/12
KITCOVIDhub-mean_ensemble	33,754	21,288	3/12	10/12	48,688	29,683	4/12	10/12	518	318	6/12	12/12	712	435	6/12	12/12
KITCOVIDhub-median_ensemble	30,154	19,767	4/12	11/12	45,928	27,397	3/12	9/12	543	332	5/12	12/12	690	456	5/12	12/12

Poland

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	46,220	33,301	0/12	5/12	63,055	46,882	1/12	5/12	350	220	7/12	11/12	491	295	5/12	11/12
epiforecasts-EpiNow2	55,368	39,353	3/12	9/12	82,709	59,002	3/12	6/12	552	391	7/12	11/12	810	576	6/12	12/12
ICM-agentModel	*33,401	*20,103	6/11	10/11	* 40,524	*26,560	4/11	7/11	* 592	*536	10/11	11/11	* 643	* 546	9/11	11/11
IHME-CurveFit									637				505			
Imperial-ensemble2																
ITWW-county_repro	63,327	54,555	1/12	6/12	108,188	93,286	1/12	4/12	1,047	915	1/12	1/12	1,466	1,241	0/12	1/12
LANL-GrowthRate	42,864	24,958	2/12	10/12	55,486	35,013	2/12	9/12	463	281	5/12	11/12	635	388	7/12	10/12
MIMUW-StochSEIR	33,532	30,118	1/12	3/12	58,203	52,730	1/12	1/12	365	331	1/12	1/12	408	327	1/12	4/12
MIT_CovidAnalytics-DELPHI	*74,004	*52,292	1/10	4/10	* 94,153	*70,385	1/10	3/10	* 717	*435	4/11	10/11	*1,019	* 665	3/11	8/11
MOCOS-agent1	22,746	19,350	4/12	5/12	39,711	29,751	3/12	5/12	235	166	11/12	12/12	243	202	11/12	12/12
SDSC_ISG-TrendModel																
USC-SIkJalpha	48,043	40,067	1/12	4/12	63,352	56,280	1/12	4/12	256	166	6/12	12/12	371	252	5/12	11/12
KIT-baseline	41,804	34,355	1/12	4/12	54,127	45,833	2/12	3/12	618	451	1/12	6/12	791	611	1/12	6/12
KIT-extrapolation_baseline	55,011	45,278	4/12	5/12	98,777	79,912	2/12	4/12	530	399	7/12	8/12	852	624	5/12	7/12
KIT-time_series_baseline	53,619	41,830	5/12	8/12	94,873	72,907	4/12	8/12	644	539	5/12	7/12	1,014	824	4/12	6/12
KITCOVIDhub-inverse_wis_ensemble	38,590	27,016	3/12	8/12	59,449	42,678	3/12	7/12	242	178	9/12	12/12	408	265	8/12	12/12
KITCOVIDhub-mean_ensemble	41,216	27,591	3/12	8/12	62,918	43,800	3/12	6/12	169	185	10/12	12/12	353	263	9/12	12/12
KITCOVIDhub-median_ensemble	41,979	28,852	2/12	7/12	61,177	43,683	1/12	7/12	196	173	9/12	12/12	354	256	9/12	12/12

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

Table 5: Forecast evaluation at the regional level, Germany and Poland, 3 and 4 weeks ahead (incidence scale, based on RKI/MZ data). Results are averaged over the different regions (states in Germany, voivodeships in Poland). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

Model	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiNow2	2,697	1,803	63/192	141/192	3,879	2,630	55/192	135/192	98	66	85/192	154/192	136	95	81/192	149/192
FIAS_FZJ-Epi1Ger	2,840	1,929	36/192	112/192	4,192	2,862	34/192	99/192	75	54	31/192	86/192	96	65	25/192	98/192
IHME-CurveFit									64				68			
ITWW-county_repro	3,123	2,100	75/192	147/192	4,394	2,970	85/192	149/192	34	23	66/192	151/192	52	34	60/192	148/192
Karlen-pypm	3,634	2,261	83/192	162/192	5,740	3,578	58/192	144/192	64	36	79/192	177/192	84	49	63/192	164/192
LeipzigIMISE-SECIR			2/12	6/12			2/12	8/12			1/12	4/12			2/12	5/12
USC-SIkJalpha	2,190	1,584	44/192	124/192	2,830	2,094	36/192	104/192	53	36	67/192	158/192	65	45	57/192	156/192
KIT-baseline	2,194	1,636	32/192	96/192	2,774	2,168	21/192	76/192	79	54	48/192	133/192	96	69	38/192	109/192
KIT-extrapolation_baseline	2,496	1,629	41/192	132/192	3,392	2,315	30/192	116/192	114	76	69/192	150/192	157	106	68/192	144/192
KIT-time_series_baseline	2,806	1,989	38/192	117/192	3,571	2,748	41/192	106/192	132	90	72/192	149/192	178	130	61/192	140/192
KITCOVIDhub-inverse_wis_ensemble	2,622	1,675	73/192	158/192	3,679	2,375	68/192	152/192	55	35	79/192	178/192	77	51	74/192	171/192
KITCOVIDhub-mean_ensemble	2,503	1,582	71/192	165/192	3,496	2,254	71/192	154/192	61	39	84/192	179/192	86	58	85/192	171/192
KITCOVIDhub-median_ensemble	2,589	1,607	62/192	163/192	3,669	2,323	51/192	146/192	51	32	92/192	182/192	70	45	77/192	176/192

Poland

Model	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiNow2	3,849	2,745	69/192	133/192	5,889	4,209	51/192	117/192	77	63	86/192	159/192	149	194	79/192	160/192
ITWW-county_repro	4,319	3,222	71/192	140/192	7,133	5,353	55/192	126/192	71	53	39/192	115/192	98	70	49/192	132/192
USC-SIkJalpha	2,543	1,910	46/192	104/192	3,438	2,665	41/192	95/192	28	18	93/192	174/192	35	23	85/192	168/192
KIT-baseline	2,743	2,195	30/192	76/192	3,519	2,932	20/192	66/192	47	32	57/192	135/192	58	43	51/192	118/192
KIT-extrapolation_baseline	4,002	3,073	50/192	110/192	7,041	5,491	38/192	92/192	52	34	84/192	166/192	70	48	74/192	159/192
KIT-time_series_baseline	3,905	2,855	65/192	127/192	6,563	5,056	53/192	115/192	58	43	60/192	114/192	75	57	45/192	111/192
KITCOVIDhub-inverse_wis_ensemble	3,327	2,253	62/192	140/192	5,558	3,892	44/192	124/192	32	22	104/192	180/192	48	46	98/192	179/192
KITCOVIDhub-mean_ensemble	3,107	2,051	63/192	154/192	5,205	3,560	44/192	129/192	34	25	106/192	181/192	59	61	99/192	179/192
KITCOVIDhub-median_ensemble	3,301	2,095	68/192	145/192	5,412	3,441	47/192	119/192	30	20	103/192	177/192	40	28	98/192	171/192

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

Table 6: Forecast evaluation for Germany and Poland, 1 and 2 weeks ahead (cumulative scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

Model	1 wk ahead cumul case				2 wk ahead cumul case				1 wk ahead cumul death				2 wk ahead cumul death			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	8,921	5,605	4/12	12/12	25,991	16,113	4/12	9/12	291	202	7/12	11/12	744	493	5/12	11/12
epiforecasts-EpiNow2	14,097	11,023	8/12	9/12	42,806	32,126	7/12	9/12	588	457	8/12	10/12	1,130	804	7/12	10/12
FIAS_FZJ-Epi1Ger	10,859	6,690	6/12	11/12	35,351	21,697	2/12	10/12	578	458	0/12	3/12	976	739	3/12	6/12
Imperial-ensemble2									* 193	* 136	8/10	9/10				
itwm-dSEIR	7,517	5,379	5/12	8/12	26,189	19,526	4/12	8/12	726	548	3/12	6/12	1,149	972	4/12	6/12
ITWW-county_repro	15,223	12,418	1/12	3/12	46,473	37,889	1/12	4/12	564	527	0/12	0/12	813	742	1/12	2/12
Karlen-pypm	18,532	13,629	6/12	11/12	53,323	38,243	4/12	11/12	380	232	5/12	11/12	963	596	5/12	11/12
LANL-GrowthRate	12,623	10,542	9/12	12/12	24,393	23,423	9/12	12/12	338	222	6/12	12/12	705	450	8/12	12/12
LeipzigIMISE-SECIR	17,708	12,470	0/12	5/12	41,912	29,244	3/12	6/12	1,474	1,335	6/12	6/12	3,142	2,544	6/12	7/12
MIT_CovidAnalytics-DELPHI									911	560	4/12	11/12	1,593	987	1/12	8/12
SDSC_ISG-TrendModel	10,394								384							
USC-SIkJalpha	16,854	11,177	4/12	10/12	40,016	30,407	2/12	7/12	467	314	4/12	8/12	1,035	686	2/12	5/12
KIT-baseline	12,756	7,953	5/12	11/12	35,996	26,510	2/12	5/12	411	277	7/12	11/12	1,164	811	3/12	9/12
KIT-extrapolation_baseline	8,823	5,715	6/12	12/12	31,598	19,803	5/12	9/12	456	269	4/12	12/12	1,219	748	4/12	12/12
KIT-time_series_baseline	15,583	10,281	3/12	9/12	47,712	33,183	3/12	8/12	406	263	8/12	12/12	1,247	849	7/12	12/12
KITCOVIDhub-inverse_wis_ensemble	10,649	6,614	7/12	11/12	32,451	20,251	4/12	10/12	303	183	6/12	12/12	463	319	8/12	12/12
KITCOVIDhub-mean_ensemble	9,715	6,000	8/12	12/12	31,185	19,363	6/12	12/12	280	178	5/12	12/12	451	325	6/12	12/12
KITCOVIDhub-median_ensemble	8,118	5,344	6/12	12/12	27,596	18,028	5/12	11/12	283	161	6/12	12/12	544	342	6/12	12/12

Poland

Model	1 wk ahead cumul case				2 wk ahead cumul case				1 wk ahead cumul death				2 wk ahead cumul death			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	21,893	16,543	4/12	9/12	36,184	24,110	4/12	7/12	316	202	2/12	9/12	499	301	6/12	12/12
epiforecasts-EpiNow2	13,372	9,851	5/12	10/12	43,134	30,915	5/12	9/12	332	261	7/12	10/12	639	440	7/12	10/12
ICM-agentModel	*14,264	*12,390	7/11	11/11	*40,403	*30,044	7/11	11/11	*279	*279	10/11	11/11	* 624	* 735	11/11	11/11
Imperial-ensemble2									*188	*138	3/10	7/10				
ITWW-county_repro	20,054	17,364	2/12	3/12	56,144	48,119	2/12	5/12	589	551	0/12	0/12	1,374	1,276	0/12	0/12
LANL-GrowthRate	8,129	5,787	10/12	12/12	30,850	19,842	7/12	12/12	229	137	2/12	10/12	573	348	4/12	11/12
MIMUW-StochSEIR	5,933	4,132	4/12	10/12	23,176	19,204	1/12	3/12	251	238	2/12	3/12	524	497	0/12	0/12
MIT_CovidAnalytics-DELPHI									*320	*199	6/11	11/11	* 823	* 513	3/11	9/11
MOCOS-agent1	5,173	4,978	5/12	8/12	19,929	15,373	3/12	9/12	158	132	9/12	12/12	349	275	10/12	12/12
SDSC_ISG-TrendModel	12,372								201							
USC-SIkJalpha	10,405	6,919	4/12	10/12	42,376	32,924	1/12	5/12	206	133	4/12	11/12	441	279	2/12	9/12
KIT-baseline	16,407	9,736	5/12	10/12	44,384	32,767	2/12	5/12	258	167	5/12	11/12	664	457	2/12	6/12
KIT-extrapolation_baseline	9,448	5,992	6/12	11/12	38,410	27,616	3/12	9/12	269	190	7/12	10/12	672	466	6/12	10/12
KIT-time_series_baseline	10,784	7,787	9/12	10/12	41,180	29,143	5/12	9/12	300	232	8/12	8/12	735	593	6/12	8/12
KITCOVIDhub-inverse_wis_ensemble	8,743	5,630	5/12	11/12	28,915	18,765	5/12	10/12	131	105	10/12	12/12	320	231	8/12	12/12
KITCOVIDhub-mean_ensemble	8,448	5,465	8/12	11/12	28,569	18,187	5/12	10/12	125	108	10/12	12/12	274	235	12/12	12/12
KITCOVIDhub-median_ensemble	6,334	4,402	8/12	12/12	28,772	19,818	4/12	10/12	144	97	9/12	12/12	280	207	9/12	12/12

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

Table 7: Forecast evaluation for Germany and Poland, 3 and 4 weeks ahead (cumulative scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

Model	3 wk ahead cumul case				4 wk ahead cumul case				3 wk ahead cumul death				4 wk ahead cumul death			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	53,494	34,345	5/12	8/12	85,643	56,327	4/12	8/12	1,414	929	4/12	11/12	2,330	1,515	3/12	11/12
epiforecasts-EpiNow2	93,015	68,352	7/12	8/12	168,222	121,460	7/12	8/12	2,070	1,401	6/12	8/12	3,244	2,189	5/12	9/12
FIAS_FZJ-Epi1Ger	72,325	45,809	2/12	11/12	122,165	79,579	1/12	7/12	1,792	1,290	1/12	6/12	2,774	1,946	1/12	7/12
Imperial-ensemble2																
itwm-dSEIR	56,386	42,720	4/12	5/12	101,463	75,670	4/12	5/12	1,677	1,459	3/12	5/12	2,265	1,945	2/12	6/12
ITWW-county_repro	94,211	75,775	0/12	5/12	160,665	128,935	0/12	5/12	826	658	0/12	2/12	740	589	5/12	8/12
Karlen-pypm	107,997	76,940	4/12	10/12	195,806	138,671	2/12	9/12	1,896	1,172	1/12	10/12	3,205	2,033	0/12	10/12
LANL-GrowthRate	41,353	39,396	9/12	12/12	68,993	58,428	8/12	12/12	1,413	869	7/12	12/12	2,312	1,463	7/12	12/12
LeipzigIMISE-SECIR	81,273	56,348	4/12	7/12	146,917	102,340	4/12	9/12	4,929	3,727	4/12	8/12	6,882	5,009	1/12	8/12
MIT_CovidAnalytics-DELPHI									2,120	1,389	2/12	8/12	2,541	1,712	2/12	8/12
SDSC_ISG-TrendModel																
USC-SIkJalpha	72,202	62,065	1/12	2/12	115,293	105,336	1/12	3/12	1,796	1,368	1/12	3/12	2,744	2,223	0/12	1/12
KIT-baseline	70,752	56,271	1/12	3/12	114,908	96,625	0/12	2/12	2,352	1,760	2/12	5/12	3,867	3,094	1/12	3/12
KIT-extrapolation_baseline	67,400	42,285	3/12	9/12	118,558	76,508	2/12	9/12	2,572	1,608	2/12	9/12	4,493	2,909	2/12	9/12
KIT-time_series_baseline	94,122	67,918	2/12	7/12	153,758	115,049	4/12	7/12	2,553	1,830	6/12	11/12	4,372	3,271	6/12	10/12
KITCOVIDhub-inverse_wis_ensemble	66,373	41,830	4/12	10/12	111,839	72,513	3/12	8/12	893	571	7/12	12/12	1,535	958	6/12	12/12
KITCOVIDhub-mean_ensemble	63,145	40,346	6/12	10/12	108,994	70,062	5/12	10/12	972	599	5/12	12/12	1,684	997	3/12	12/12
KITCOVIDhub-median_ensemble	58,244	38,895	5/12	9/12	100,388	68,015	5/12	10/12	1,104	641	4/12	12/12	1,854	1,072	4/12	12/12

Poland

Model	3 wk ahead cumul case				4 wk ahead cumul case				3 wk ahead cumul death				4 wk ahead cumul death			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	71,493	46,562	2/12	8/12	124,129	86,643	1/12	6/12	801	495	6/12	12/12	1,188	756	6/12	11/12
epiforecasts-EpiNow2	97,781	68,679	4/12	9/12	180,004	125,246	3/12	7/12	1,189	794	7/12	11/12	1,985	1,354	5/12	11/12
ICM-agentModel	* 67,819	* 49,027	6/11	11/11	*101,414	* 73,220	6/11	10/11	*1,207	*1,262	10/11	11/11	*1,834	*1,789	11/11	11/11
Imperial-ensemble2																
ITWW-county_repro	118,193	101,211	2/12	4/12	224,258	192,827	1/12	4/12	2,410	2,187	0/12	0/12	3,854	3,406	0/12	1/12
LANL-GrowthRate	71,765	42,183	6/12	12/12	127,178	74,690	1/12	10/12	1,027	616	4/12	11/12	1,638	977	4/12	10/12
MIMUW-StochSEIR	53,052	46,717	1/12	3/12	106,465	96,945	0/12	0/12	832	787	0/12	2/12	1,186	1,072	0/12	2/12
MIT_CovidAnalytics-DELPHI									*1,540	*1,075	2/11	6/11	*2,558	*1,978	1/11	4/11
MOCOS-agent1	38,471	32,124	5/12	7/12	73,460	59,498	5/12	6/12	531	422	12/12	12/12	699	589	11/12	12/12
SDSC_ISG-TrendModel																
USC-SIkJalpha	84,104	75,355	1/12	3/12	145,603	136,915	1/12	1/12	625	428	4/12	7/12	992	701	2/12	6/12
KIT-baseline	85,544	69,035	1/12	4/12	137,648	116,676	0/12	3/12	1,286	965	0/12	4/12	2,063	1,658	0/12	4/12
KIT-extrapolation_baseline	93,229	72,727	4/12	6/12	184,714	148,616	4/12	4/12	1,203	859	5/12	8/12	2,021	1,465	5/12	8/12
KIT-time_series_baseline	95,539	70,731	6/12	9/12	186,952	142,431	5/12	8/12	1,336	1,132	6/12	6/12	2,342	1,966	5/12	7/12
KITCOVIDhub-inverse_wis_ensemble	64,825	42,213	4/12	10/12	116,123	80,227	3/12	8/12	563	381	9/12	12/12	979	629	8/12	12/12
KITCOVIDhub-mean_ensemble	66,331	42,082	4/12	10/12	118,378	80,761	4/12	9/12	416	386	12/12	12/12	750	617	9/12	12/12
KITCOVIDhub-median_ensemble	68,709	46,381	5/12	10/12	120,285	85,310	3/12	6/12	382	332	9/12	12/12	667	565	8/12	12/12

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

Table 8: Forecast evaluation for Germany and Poland, 1 and 2 weeks ahead (incidence scale, based on JHU data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

Model	1 wk ahead case (JHU)				2 wk ahead case (JHU)				1 wk ahead death (JHU)				2 wk ahead death (JHU)			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	13,319	8,491	2/12	9/12	22,391	15,479	3/12	8/12	336	226	4/12	11/12	485	323	4/12	11/12
epiforecasts-EpiNow2	14,041	10,182	4/12	10/12	31,439	22,548	6/12	8/12	284	196	9/12	11/12	601	424	7/12	9/12
FIAS_FZJ-Epi1Ger	14,650	9,740	2/12	8/12	27,022	17,565	1/12	9/12	389	299	2/12	5/12	636	468	2/12	8/12
IHME-CurveFit									461				661			
Imperial-ensemble2									*234	*155	7/10	9/10				
itwm-dSEIR	9,919	7,887	6/12	7/12	20,214	14,003	5/12	9/12	492	319	6/12	11/12	518	336	6/12	10/12
ITWW-county_repro	17,888	15,162	2/12	4/12	33,039	27,132	1/12	3/12	517	483	1/12	2/12	259	204	1/12	2/12
Karlen-pypm	22,754	16,774	4/12	10/12	34,263	25,765	4/12	10/12	393	253	6/12	11/12	656	414	2/12	10/12
LANL-GrowthRate	12,175	10,877	9/12	12/12	17,749	14,958	9/12	11/12	286	193	6/12	12/12	457	286	6/12	12/12
LeipzigIMISE-SECIR	14,283	11,458	2/12	5/12	28,741	20,567	2/12	6/12	406	292	7/12	12/12	912	656	4/12	8/12
MIT_CovidAnalytics-DELPHI	*15,629	*10,544	5/11	10/11	*23,970	*16,970	3/11	8/11	777	468	5/12	12/12	763	452	7/12	12/12
SDSC_ISG-TrendModel	9,981								419							
USC-SIkJalpha	18,702	12,899	2/12	9/12	26,340	18,768	1/12	6/12	393	260	7/12	10/12	573	343	3/12	9/12
KIT-baseline	16,125	10,728	4/12	10/12	26,070	18,921	1/12	6/12	423	291	7/12	11/12	814	544	2/12	8/12
KIT-extrapolation_baseline	13,419	8,526	4/12	10/12	24,159	15,031	3/12	11/12	450	282	5/12	12/12	826	506	4/12	10/12
KIT-time_series_baseline	20,128	13,215	1/12	8/12	34,678	23,736	2/12	7/12	467	295	7/12	12/12	844	610	6/12	11/12
KITCOVIDhub-inverse_wis_ensemble	12,749	8,145	5/12	11/12	24,149	14,899	4/12	11/12	185	148	8/12	12/12	331	213	10/12	12/12
KITCOVIDhub-mean_ensemble	12,822	8,185	6/12	12/12	24,164	14,858	4/12	11/12	202	150	7/12	12/12	360	224	7/12	12/12
KITCOVIDhub-median_ensemble	11,789	7,853	3/12	10/12	21,667	14,075	4/12	10/12	195	147	8/12	12/12	381	231	7/12	12/12

Poland

Model	1 wk ahead case				2 wk ahead case				1 wk ahead death				2 wk ahead death			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	9,348	5,809	4/12	11/12	28,241	19,561	1/12	7/12	221	151	5/12	9/12	316	195	6/12	10/12
epiforecasts-EpiNow2	8,669	6,394	6/12	10/12	30,133	22,536	5/12	10/12	218	138	6/12	11/12	390	245	5/12	12/12
ICM-agentModel	*24,466	*16,579	3/11	10/11	*28,581	*19,023	8/11	11/11	*506	*296	8/11	11/11	*573	*504	9/11	11/11
IHME-CurveFit									352				472			
Imperial-ensemble2									*191	*143	3/10	6/10				
ITWW-county_repro	18,811	16,021	1/12	5/12	36,017	31,212	2/12	6/12	548	510	0/12	0/12	729	656	0/12	0/12
LANL-GrowthRate	10,738	6,547	8/12	12/12	26,194	16,557	5/12	11/12	249	146	1/12	10/12	378	236	3/12	10/12
MIMUW-StochSEIR	8,492	6,067	2/12	7/12	20,002	17,452	1/12	3/12	249	235	1/12	2/12	286	264	0/12	0/12
MIT_CovidAnalytics-DELPHI	*24,435	*14,346	2/10	9/10	*52,595	*35,605	1/10	4/10	*417	*260	5/11	10/11	*536	*314	3/11	11/11
MOCOS-agent1	8,086	6,272	6/12	7/12	17,320	13,263	3/12	6/12	159	134	8/12	12/12	159	141	10/12	12/12
SDSC_ISG-TrendModel	8,333								289							
USC-SIkJalpha	13,247	8,677	3/12	8/12	35,120	26,749	2/12	6/12	210	142	5/12	11/12	252	156	5/12	12/12
KIT-baseline	18,711	11,471	4/12	10/12	34,420	24,900	1/12	5/12	293	188	5/12	9/12	459	315	3/12	8/12
KIT-extrapolation_baseline	10,390	6,698	6/12	11/12	32,416	23,760	3/12	6/12	286	199	6/12	9/12	417	297	5/12	9/12
KIT-time_series_baseline	11,424	8,387	7/12	10/12	31,643	22,688	6/12	9/12	312	246	7/12	8/12	486	388	6/12	7/12
KITCOVIDhub-inverse_wis_ensemble	8,451	5,486	6/12	12/12	25,280	16,907	3/12	10/12	182	119	7/12	12/12	233	153	8/12	12/12
KITCOVIDhub-mean_ensemble	8,817	5,619	8/12	12/12	26,309	17,097	4/12	10/12	169	121	8/12	12/12	174	156	10/12	12/12
KITCOVIDhub-median_ensemble	8,977	5,411	6/12	11/12	25,951	17,939	2/12	9/12	190	114	7/12	12/12	215	142	9/12	12/12

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

Table 9: Forecast evaluation for Germany and Poland, 3 and 4 weeks ahead (incidence scale, based on JHU data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

Model	3 wk ahead case (JHU)				4 wk ahead case (JHU)				3 wk ahead death (JHU)				4 wk ahead death (JHU)			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	30,920	20,680	3/12	7/12	38,187	24,940	2/12	8/12	724	470	3/12	11/12	971	646	2/12	9/12
epiforecasts-EpiNow2	51,165	35,835	5/12	9/12	74,985	54,568	6/12	8/12	946	659	6/12	9/12	1,260	866	6/12	9/12
FIAS_FZJ-Epi1Ger	37,482	25,141	1/12	6/12	52,812	36,746	1/12	7/12	871	592	1/12	7/12	1,071	678	1/12	6/12
IHME-CurveFit									874				1,000			
Imperial-ensemble2																
itwm-dSEIR	31,973	19,974	4/12	10/12	47,231	30,301	5/12	10/12	532	335	5/12	12/12	539	337	5/12	12/12
ITWW-county_repro	48,002	38,724	1/12	5/12	64,365	53,500	3/12	5/12	330	246	2/12	4/12	781	625	1/12	1/12
Karlen-pypm	55,203	39,877	3/12	8/12	91,822	66,228	2/12	8/12	952	628	0/12	10/12	1,342	939	0/12	7/12
LANL-GrowthRate	23,138	17,800	9/12	11/12	32,959	22,202	6/12	12/12	722	457	5/12	11/12	1,076	675	6/12	11/12
LeipzigIMISE-SECIR	48,868	33,487	1/12	8/12	77,302	54,290	3/12	8/12	1,589	1,102	3/12	8/12	2,526	1,798	2/12	7/12
MIT_CovidAnalytics-DELPHI	*29,198	*22,885	5/11	7/11	*39,778	*31,367	5/11	7/11	718	465	9/12	12/12	680	488	9/12	12/12
SDSC_ISG-TrendModel																
USC-SIkJalpha	35,113	26,616	2/12	6/12	49,985	39,839	1/12	2/12	815	519	3/12	8/12	948	619	2/12	7/12
KIT-baseline	37,102	28,457	1/12	4/12	46,124	37,251	0/12	3/12	1,188	858	2/12	6/12	1,529	1,159	1/12	3/12
KIT-extrapolation_baseline	36,004	23,750	2/12	9/12	54,153	37,117	2/12	6/12	1,351	866	2/12	9/12	1,917	1,302	2/12	9/12
KIT-time_series_baseline	46,542	34,947	4/12	7/12	63,585	47,872	3/12	6/12	1,318	981	6/12	10/12	1,815	1,452	5/12	10/12
KITCOVIDhub-inverse_wis_ensemble	35,449	21,286	3/12	11/12	49,251	31,504	5/12	9/12	505	320	9/12	12/12	717	473	8/12	12/12
KITCOVIDhub-mean_ensemble	35,296	21,200	2/12	10/12	48,995	31,027	5/12	10/12	537	319	8/12	12/12	745	454	8/12	12/12
KITCOVIDhub-median_ensemble	31,963	19,924	3/12	10/12	46,236	28,865	3/12	9/12	553	338	4/12	12/12	727	475	5/12	12/12

Poland

Model	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	49,190	36,126	0/12	5/12	66,413	49,981	1/12	5/12	381	248	7/12	11/12	538	316	5/12	11/12
epiforecasts-EpiNow2	57,000	40,754	3/12	7/12	84,239	61,040	3/12	6/12	590	408	7/12	11/12	858	598	5/12	11/12
ICM-agentModel	*36,110	*21,706	5/11	10/11	* 43,759	*28,736	3/11	7/11	* 580	*533	10/11	11/11	* 656	* 544	9/11	11/11
IHME-CurveFit									603				521			
Imperial-ensemble2																
ITWW-county_repro	64,635	56,004	1/12	6/12	111,127	96,150	0/12	3/12	1,020	886	0/12	1/12	1,444	1,219	0/12	2/12
LANL-GrowthRate	45,833	27,090	2/12	9/12	58,322	37,589	2/12	9/12	480	300	6/12	10/12	640	395	7/12	10/12
MIMUW-StochSEIR	37,055	33,387	1/12	3/12	61,665	56,148	1/12	1/12	347	314	2/12	2/12	422	350	2/12	2/12
MIT_CovidAnalytics-DELPHI	*77,200	*55,508	1/10	4/10	* 96,731	*73,059	1/10	3/10	* 770	*476	5/11	10/11	*1,062	* 698	3/11	7/11
MOCOS-agent1	25,951	21,656	3/12	5/12	43,679	32,808	2/12	5/12	210	164	11/12	12/12	220	203	11/12	12/12
SDSC_ISG-TrendModel																
USC-SIkJalpha	50,909	42,933	1/12	3/12	66,710	59,344	1/12	4/12	244	160	7/12	12/12	408	270	5/12	11/12
KIT-baseline	44,230	36,646	2/12	4/12	56,963	48,028	1/12	3/12	647	481	1/12	6/12	812	631	0/12	6/12
KIT-extrapolation_baseline	57,609	47,303	4/12	5/12	102,746	82,981	2/12	4/12	576	431	6/12	8/12	899	667	4/12	7/12
KIT-time_series_baseline	56,074	43,716	5/12	8/12	97,935	75,620	4/12	8/12	701	585	4/12	7/12	1,058	861	2/12	6/12
KITCOVIDhub-inverse_wis_ensemble	41,559	29,313	2/12	8/12	62,808	45,573	3/12	7/12	268	190	9/12	12/12	444	280	6/12	12/12
KITCOVIDhub-mean_ensemble	44,185	30,162	3/12	7/12	66,277	46,910	3/12	6/12	228	197	9/12	12/12	403	283	9/12	12/12
KITCOVIDhub-median_ensemble	44,949	31,671	2/12	6/12	64,535	46,726	1/12	7/12	229	187	9/12	12/12	399	278	7/12	12/12

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

Table 10: Forecast evaluation for Germany and Poland, pooled across evaluation periods, 1 and 2 weeks ahead (incidence scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	1 wk ahead case			2 wk ahead case			1 wk ahead death			2 wk ahead death		
Model	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	10,653	6,945	8/22	19/22	24,560	17,271	7/21	9/21	249	171	11/22	18/22	433	285	7/21	17/21
epiforecasts-EpiNow2	10,355	7,229	13/22	17/22	32,772	23,970	11/21	15/21	246	157	14/22	18/22	493	339	11/21	16/21
FIAS_FZJ-Epi1Ger	9,118	6,028	11/22	20/22	27,174	18,523	6/21	14/21	354	285	5/22	7/22	599	457	4/21	10/21
IHME-CurveFit																
Imperial-ensemble2									224*	166*	13/20	14/20				
itwm-dSEIR			5/12	9/12			5/12	9/12			7/12	10/12			6/12	10/12
ITWW-county_repro	23,951	19,913	1/22	5/22	45,782	37,544	0/21	4/21	476	449	1/22	2/22	393	342	2/21	5/21
Karlen-pypm			6/14	13/14			3/13	11/13			5/14	12/14			1/13	10/13
LANL-GrowthRate	22,330*	15,271*	14/19	19/19	36,344*	23,394*	11/18	18/18	285*	187*	9/19	19/19	436*	281*	9/18	17/18
LeipzigIMISE-SECIR	14,097	13,366*	4/17	11/17	37,135	36,634*	2/16	8/16	484	318*	7/17	13/17	828	621*	5/16	10/16
MIT_CovidAnalytics-DELPHI	24,290*	17,004*	7/19	14/19	44,566*	33,677*	6/18	10/18	671*	436*	6/20	14/20	636*	397*	7/19	17/19
SDSC_ISG-TrendModel	9,271								400							
UCLA-SuEIR																
USC-SIkJalpha	16,613		4/13	11/13	27,942		2/12	7/12	430		6/13	10/13	582		3/12	10/12
KIT-baseline	15,355	10,246	10/22	20/22	27,601	20,850	5/21	13/21	442	270	9/22	20/22	804	518	2/21	13/21
KIT-extrapolation_baseline	10,274	7,900	13/22	22/22	28,704	19,614	10/21	16/21	341	208	11/22	21/22	624	385	9/21	18/21
KIT-time_series_baseline	15,492	10,614	8/22	18/22	37,524	24,854	7/21	16/21	330	230	14/22	21/22	754	521	10/21	19/21
KITCOVIDhub-inverse_wis_ensemble	11,055	7,141	12/22	21/22	30,599	19,896	8/21	15/21	200	133	13/22	21/22	285	181	11/21	20/21
KITCOVIDhub-mean_ensemble	12,137	7,732	13/22	20/22	30,563	19,502	7/21	17/21	213	146	10/22	21/22	326	199	11/21	20/21
KITCOVIDhub-median_ensemble	9,249	6,221	13/22	21/22	27,149	17,998	8/21	17/21	217	143	13/22	20/22	358	221	10/21	19/21

Poland

	1 wk ahead case			2 wk ahead case			1 wk ahead death			2 wk ahead death		
Model	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	10,292	6,836	13/22	19/22	30,493	20,653	4/21	13/21	243	155	8/22	21/22	423	264	7/21	18/21
epiforecasts-EpiNow2	9,327	6,422	12/22	18/22	33,681	23,474	7/21	16/21	276	184	10/22	19/22	670	465	9/21	16/21
ICM-agentModel	19,736*	13,900*	5/15	14/15	32,194*	21,406*	8/14	12/14	599*	453*	10/19	14/19	1,101*	791*	9/18	13/18
IHME-CurveFit																
Imperial-ensemble2									293*	188*	6/20	14/20				
ITWW-county_repro	19,188	16,147	4/22	7/22	35,214	29,629	4/21	10/21	560	521	0/22	1/22	748	669	0/21	2/21
LANL-GrowthRate	11,012*	7,151*	13/19	19/19	31,944*	19,233*	8/18	17/18	233*	151*	6/19	17/19	366*	228*	7/18	17/18
MIMUW-StochSEIR	10,072*	6,468*	7/17	15/17	23,492*	19,354*	4/16	6/16	422*	320*	3/17	6/17	771*	557*	1/16	2/16
MIT_CovidAnalytics-DELPHI	27,212*	17,817*	4/19	14/19	54,488*	38,714*	2/18	11/18	447*	282*	8/20	16/20	580*	354*	5/19	17/19
MOCOS-agent1	8,855	6,863	7/22	13/22	22,131	17,207	3/21	11/21	174	139	18/22	22/22	296	202	17/21	20/21
SDSC_ISG-TrendModel	6,918								214							
USC-SIkJalpha	10,353		4/13	11/13	29,100		2/12	6/12	206		4/13	12/13	255		5/12	11/12
KIT-baseline	21,751	13,546	10/22	19/22	41,057	28,022	3/21	11/21	340	216	10/22	21/22	595	384	4/21	15/21
KIT-extrapolation_baseline	13,477	8,685	12/22	21/22	40,533	27,328	6/21	14/21	332	233	13/22	18/22	658	463	10/21	17/21
KIT-time_series_baseline	16,108	10,657	14/22	20/22	43,096	28,569	11/21	18/21	411	281	14/22	18/22	854	574	11/21	15/21
KITCOVIDhub-inverse_wis_ensemble	9,796	6,402	10/22	21/22	28,909	19,458	8/21	16/21	182	130	15/22	22/22	321	216	13/21	20/21
KITCOVIDhub-mean_ensemble	9,646	6,391	12/22	21/22	29,101	19,012	7/21	16/21	191	136	16/22	21/22	349	242	15/21	20/21
KITCOVIDhub-median_ensemble	10,342	6,430	12/22	21/22	30,228	19,832	6/21	16/21	186	124	14/22	22/22	312	202	11/21	21/21

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.

Table 11: Forecast evaluation for Germany and Poland, pooled across evaluation periods, 3 and 4 weeks ahead (incidence scale, based on RKI/MZ data). $C_{0.5}$ and $C_{0.95}$ denote coverage rates of the 50% and 95% prediction intervals; AE and WIS stand for the mean absolute error and mean weighted interval score.

Germany

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	36,332	25,966	5/20	11/20	48,205	32,980	1/19	12/19	673	449	5/20	14/20	936	677	4/19	12/19
epiforecasts-EpiNow2	70,236	51,689	8/20	15/20	134,886	98,969	7/19	12/19	826	574	9/20	15/20	1,336	906	8/19	14/19
FIAS_FZJ-Epi1Ger	53,281	37,314	3/20	13/20	88,229	64,186	3/19	12/19	877	648	1/20	10/20	1,046	737	3/19	10/19
IHME-CurveFit																
Imperial-ensemble2			4/12	10/12			5/12	11/12			5/12	11/12			5/12	12/12
itwm-dSEIR	70,730	58,490	1/20	6/20	96,380	81,756	1/19	6/19	495	405	3/20	7/20	770	625	2/19	4/19
Karlen-pypm			4/12	9/12			2/12	8/12			0/12	9/12			0/12	8/12
LANL-GrowthRate	45,698*	29,603*	9/17	16/17	45,825*	33,757*	9/16	15/16	751*	502*	8/17	15/17	1,044*	711*	8/16	14/16
LeipzigIMISE-SECIR	62,495	72,724*	1/15	8/15	102,043	127,685*	1/14	8/14	1,295	1,108*	3/15	9/15	2,003	1,946*	3/14	8/14
MIT_CovidAnalytics-DELPHI	71,468*	58,057*	5/18	10/18	104,826*	87,522*	5/18	8/18	669*	465*	10/19	16/19	817	636	10/19	14/19
SDSC_ISG-TrendModel																
UCLA-SuEIR	41,039		1/12	5/12	53,915		2/12	3/12	808		3/12	8/12	1,007		3/12	8/12
USC-SIkJalpha																
KIT-baseline	38,805	30,289	5/20	9/20	48,062	37,794	3/19	7/19	1,188	842	2/20	8/20	1,536	1,186	1/19	4/19
KIT-extrapolation_baseline	54,987	36,555	6/20	17/20	93,240	64,624	4/19	11/19	1,099	711	6/20	16/20	1,758	1,169	3/19	15/19
KIT-time_series_baseline	65,037	45,753	6/20	14/20	98,051	78,394	7/19	13/19	1,216	905	8/20	15/20	1,762	1,545	7/19	15/19
KITCOVIDhub-inverse_wis_ensemble	56,618	38,957	5/20	15/20	94,005	65,809	6/19	14/19	473	301	13/20	19/20	687	458	9/19	17/19
KITCOVIDhub-mean_ensemble	53,170	35,696	4/20	14/20	81,675	57,091	6/19	14/19	506	310	9/20	19/20	699	434	8/19	18/19
KITCOVIDhub-median_ensemble	49,788	33,166	6/20	17/20	76,681	51,457	5/19	13/19	523	333	8/20	18/20	657	452	9/19	17/19

Poland

	3 wk ahead case				4 wk ahead case				3 wk ahead death				4 wk ahead death			
Model	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$	AE	WIS	$C_{0.5}$	$C_{0.95}$
epiforecasts-EpiExpert	55,346	40,055	1/20	6/20	81,957	62,080	1/19	5/19	571	369	8/20	16/20	800	520	6/19	14/19
epiforecasts-EpiNow2	73,258	51,543	5/20	13/20	146,975	104,951	5/19	8/19	1,388	970	7/20	15/20	2,471	1,748	6/19	15/19
ICM-agentModel			6/13	10/13			4/12	7/12	1,037*	854*	11/17	13/17	1,020*	846*	10/16	13/16
IHME-CurveFit																
Imperial-ensemble2																
ITWW-county_repro	65,634	56,584	2/20	9/20	122,956	107,847	1/19	5/19	1,073	932	2/20	3/20	1,669	1,446	1/19	2/19
LANL-GrowthRate	57,040*	31,430*	2/17	15/17	72,916*	45,076*	2/16	12/16	530*	329*	7/17	16/17	774*	479*	9/16	14/16
MIMUW-StochSEIR	42,959*	37,974*	2/15	4/15	70,955*	62,821*	1/14	1/14	1,403*	1,004*	1/15	2/15	2,131*	1,468*	1/14	4/14
MIT_CovidAnalytics-DELPHI	90,391*	67,331*	2/17	7/17	125,485*	100,228*	3/16	5/16	875*	589*	5/18	15/18	1,313*	955*	4/17	11/17
MOCOS-agent1	41,792	35,988	5/20	8/20	76,914	65,186	5/19	7/19	517	333	15/20	17/20	847	592	12/19	15/19
SDSC_ISG-TrendModel																
USC-SIkJalpha	42,643		1/12	4/12	51,952		1/12	4/12	344		6/12	12/12	530		5/12	11/12
KIT-baseline	55,156	42,440	2/20	8/20	67,121	52,517	3/19	7/19	854	615	3/20	10/20	1,068	828	2/19	9/19
KIT-extrapolation_baseline	83,345	62,296	7/20	11/20	165,052	127,853	4/19	7/19	1,116	845	12/20	14/20	2,046	1,585	9/19	11/19
KIT-time_series_baseline	81,657	61,117	10/20	16/20	144,083	125,616	9/19	15/19	1,280	918	9/20	14/20	1,777	1,322	7/19	13/19
KITCOVIDhub-inverse_wis_ensemble	54,734	39,215	4/20	12/20	96,497	74,679	5/19	10/19	560	366	11/20	18/20	1,070	701	8/19	17/19
KITCOVIDhub-mean_ensemble	56,238	39,027	5/20	13/20	97,337	73,756	5/19	10/19	567	395	11/20	18/20	1,045	705	9/19	17/19
KITCOVIDhub-median_ensemble	54,928	40,367	4/20	11/20	92,043	70,201	1/19	10/19	476	317	12/20	19/20	913	589	9/19	16/19

*Asterisks mark entries where scores were imputed for at least one week. Weighted interval scores and absolute errors were imputed with the worst (largest) score achieved by any other forecast for the respective target and week. Models marked thus received a pessimistic assessment of their performance. If a model covered less than two thirds of the evaluation period, results are omitted.