



RESEARCH ARTICLE

REVISED

Baseline nowcasting methods for handling delays in epidemiological data

[version 2; peer review: 1 approved, 3 approved with reservations]

Kaitlyn E. Johnson ¹, Maria L. Tang², Emily Tyszka ³, Laura Jones ³, Barbora Nemcova ⁴, Daniel Wolfram⁴, Rosa Ergas³, Nicholas G. Reich⁵, Sebastian Funk ¹, Jonathon Mellor², Johannes Bracher⁴, Sam Abbott ¹¹Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine Faculty of Epidemiology and Population Health, London, England, NW5 1JG, UK²Chief Data Officer Group, United Kingdom Health Security Agency, London, UK³Massachusetts Department of Public Health, Boston, MA, USA⁴Institute of Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany⁵University of Massachusetts Amherst School of Public Health and Health Sciences, Amherst, Massachusetts, USA**V2** First published: 05 Nov 2025, 10:614
<https://doi.org/10.12688/wellcomeopenres.25027.1>Latest published: 14 Feb 2026, 10:614
<https://doi.org/10.12688/wellcomeopenres.25027.2>

Abstract

Background

Up-to-date real-time disease surveillance data can provide critical public health insights, however reporting delays can create downward bias in the latest data. Nowcasting methods designed to correct for this bias remain underused in public health practice due to their complexity, lack of tailored documentation, or technical barriers. Methodological advances in nowcasting are also hampered by the absence of standardised benchmarks for evaluating new methods.

Methods

To address these needs, we developed a family of nowcasting methods and an accompanying R package, *baselinenowcast*. We validated our method against the baseline method that was used in the German COVID-19 Nowcast Hub and on which our approach was based. Using this data, we conducted an analysis to compare different specifications of our method which were designed to address common issues in epidemiology such as weekday patterns in reporting and the ability to share estimates across different strata. We used our approach on norovirus surveillance data from the United Kingdom Health Security Agency (UKHSA) and compared the

Open Peer Review

Approval Status

	1	2	3	4
version 2 (revision) 14 Feb 2026		 view		 view
version 1 05 Nov 2025	 view	 view	 view	

- Felix Günther** , Robert Koch Institute, Berlin, Germany
- Giorgio Guzzetta**, Fondazione Bruno Kessler, Trento, Italy
- Sandra Flores Alvarado** , Universidad de Chile Escuela de Salud Publica (Ringgold ID: 193077), Independencia, Chile
- Fatima Dakroub** , American University of Beirut, Beirut, Lebanon

Any reports and responses or comments on the article can be found at the end of the article.

performance of three of our method specifications against three methods evaluated in a previous study.

Results

Our baseline method improved estimates compared to unadjusted data across all case studies. We found that the optimal choice of baseline method specification depends on context but that our default method specification performed well in a range of settings. Applied to UKHSA norovirus data, our method helped us understand the performance of the model currently used in public health practice.

Conclusions

Our method and software can be used both as a straightforward nowcasting method and provides a benchmark for nowcasting model development.

Plain Language Summary

Reporting delays in public health surveillance systems can create a misleading impression of declining trends in recent data. While a number of “nowcasting” methods have been developed to correct for this bias, widespread adoption in public health practice has yet to be realized. Currently, there is no simple method to perform nowcasting that both meets the needs of public health practice and can be used as a benchmark for further methodological advancement. To address these two gaps, we developed a family of baseline nowcasting methods and an accompanying software package. Using data from COVID-19 in Germany and norovirus in England, we evaluated the performance of our default method against other methods used in previous studies and assessed the performance of our different method specifications designed specifically for common problems in epidemiology such as weekday patterns in reporting and sharing estimates across different strata. Our findings indicate that our baseline methods improve performance over unadjusted data or more ad-hoc baseline nowcasting approaches, provide an interpretable and accessible nowcasting solution for public health practice, and are useful as a standard benchmark for enhanced understanding of more advanced methods.

Keywords

nowcasting, right-truncation, benchmarking, infectious disease surveillance

Corresponding author: Kaitlyn E. Johnson (kaitlyn.johnson@lshtm.ac.uk)

Author roles: **E. Johnson K:** Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **L. Tang M:** Data Curation, Formal Analysis, Validation, Writing – Review & Editing; **Tyszka E:** Visualization, Writing – Review & Editing; **Jones L:** Visualization, Writing – Review & Editing; **Nemcova B:** Writing – Review & Editing; **Wolfram D:** Data Curation; **Ergas R:** Visualization, Writing – Review & Editing; **G. Reich N:** Funding Acquisition, Writing – Review & Editing; **Funk S:** Conceptualization, Funding Acquisition, Methodology, Resources, Supervision, Writing – Review & Editing; **Mellor J:** Data Curation, Formal Analysis, Validation, Writing – Review & Editing; **Bracher J:** Data Curation, Methodology, Writing – Review & Editing; **Abbott S:** Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome [210758]; Centers for Disease Control and Prevention [NU38FT00008]; Deutsche Forschungsgemeinschaft [512483310]; Helmholtz Association [HIDSS4Health].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 E. Johnson K *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: E. Johnson K, L. Tang M, Tyszka E *et al.* **Baseline nowcasting methods for handling delays in epidemiological data [version 2; peer review: 1 approved, 3 approved with reservations]** Wellcome Open Research 2026, **10**:614 <https://doi.org/10.12688/wellcomeopenres.25027.2>

First published: 05 Nov 2025, **10**:614 <https://doi.org/10.12688/wellcomeopenres.25027.1>

REVISED Amendments from Version 1

In this updated version, the main changes include:

- A reorganisation of the methods so that we describe the overall method, and then have a separate section for the defaults rather than having the default method described.
- A more in-depth description of the way that the retrospective nowcasting for uncertainty estimation works to make clear that the set of reference times used for this overlaps with those use for delay estimation
- An additional section was added to the supplement to describe edge cases in real-world data and options for handling them
- Figure 2 was edited to improve the overall flow of the figure and clean up some messiness on my part.
- Figure 5 was edited to change the colour scheme of the weekdays
- The first paragraph of the discussion was edited to make it a bit more succinct and clear in the aims of the project (as a practical tool and a benchmark)

We also made some minor changes to the text throughout to reflect typos and have added some additional references.

Any further responses from the reviewers can be found at the end of the article

Introduction

Delays in epidemiological data present a challenge in public health surveillance, as they can create a misleading impression of declining trends in recent data. This bias is due to “right-truncation” of primary events (e.g. symptom onset or hospital admission), where recent events are systematically under-represented because reporting delays cause some portion of these events to not yet be present in the real-time surveillance data. This affects numerous surveillance systems across pathogens, including COVID-19, mpox, and norovirus.^{1–7} Recently, there have been several applications of nowcasting methods aimed at correcting for this downward bias. Examples include public facing data dashboards such as at Massachusetts Department of Public Health,⁸ Dutch National Institute for Public Health and the Environment (RIVM),⁹ and the U.S. Centers for Disease Control and Prevention (U.S. CDC),¹⁰ and internally at UK Health Security Agency (UKHSA) for norovirus, flu, and RSV nowcasting, Robert Koch Institute (RKI) for acute respiratory incidence nowcasting, and the U.S. CDC for estimation of the effective reproductive number for flu and COVID-19, on publicly available data,^{3,11,12} and as part of outbreak responses.^{13,14} However, these applications still remain the exception, with the majority of potential use-cases handled by showing uncorrected data or shortening time series to remove partially complete data, which has the effect of reducing the amount of information available on the latest trends.^{2,15}

Several challenges remain in addressing systematic downward bias in surveillance systems due to right-truncation of primary epidemiological events in public health practice. Frontline practitioners require methods that are statistically rigorous, straightforward to implement, and can be easily communicated to stakeholders. However, no currently available tools meet these needs fully. The chain ladder method,^{16,17} Lawless is an empirically-driven approach that could meet the needs of practitioners as it is close to the kinds of methods often implemented in real world settings. However, current implementations of the method either lack domain-specific features and documentation (e.g. the *chainladder* R package¹⁸ designed for nowcasting insurance claims) or have been implemented in an ad-hoc manner rather than intended for general applicability.^{3,8} Other methods designed specifically for nowcasting epidemiological data are often complex due to the flexibility and feature set that supporting a range of epidemiological settings requires,^{19–23} leading to a high barrier to entry. This means practitioners may have to choose between delaying insights (by not making full use of recent data), risk misleading interpretations from unadjusted data, developing methods that may lack reusability across different contexts or not be fully validated, or use methods not appropriate for their context.

For methodological research, the absence of standardised benchmarks makes understanding the value of novel methods more challenging. In other fields and other settings, baseline methods, which are relatively straightforward methods to implement and serve as a reference point for evaluating other methods,^{24,25} have been suggested as a potential solution for this issue.²⁵ Without consistent baselines for evaluation, researchers and public health practitioners may struggle to quantify incremental improvements or understand trade-offs between model complexity and performance across different epidemiological contexts.²⁴

The chain ladder approach has been used on an ad-hoc basis to meet this need in 3, 8, 26. For wider usage, a standardised method is needed.

We propose a family of baseline nowcasting methods and an accompanying R package, *baselinenowcast*, designed to address these dual needs in infectious disease surveillance, building on the empirically-driven chain ladder implementation described by Wolfram *et al.*^{3,16,18} In this work, we validate our approach by comparing the Wolfram *et al.*³ method against our baseline method with a matching specification using German COVID-19 data. We then compare the performance of our different method specifications applied to both the German COVID-19 data and UKHSA norovirus surveillance data² alongside existing nowcasting methods.

Methods

Input data requirements: reporting triangle

Our method requires the number of incident cases indexed by the reference time (e.g. date of symptom onset, specimen collection, or hospital admission) and report time (e.g. the date a positive test is reported into a surveillance system).

Data in this format allows us to reconstruct reporting delays (i.e. the time from the reference time to the report time). Combining these reporting delays results in a reporting delay distribution (i.e. the distribution of times from the reference time to the time that a case is reported) for each reference time. In wide format, where each reference time is a row and each reporting delay is a column, this data is known as the “reporting triangle”²⁷ because the observations form a triangle with the bottom right elements not yet observed as of the most recent reference time. The goal of nowcasting is to estimate these unobserved events. If we do not do so, and instead sum only the observed columns, then our case counts will be systematically downward biased. The fully observed version of these data, the “complete reporting matrix”, can be used to estimate the reporting delay distribution as each row can be normalised by the sum of that row to give the proportion reported at a given reporting delay. It is data of this format, or reconstructions thereof, that we use in our nowcasting method.

Figure 1 depicts a visual representation of the reporting triangle on the left and the corresponding case counts aggregated by reference time on the right. The solid shading indicates what has been observed as of the most recent reference time, and the lighter shading indicates what will eventually be observed. If we were to aggregate cases by reference time as of the most recent reference time, we would be missing cases from delays 1, 2, and 3 for the most recent reference time, delays 2 and 3 from the second most recent reference time, etc., resulting in systematic downward bias in the recent trend, as indicated by the gray line in the bar chart. Eventually, the counts would be revised upwards as the elements of the reporting triangle are observed at later delays, resulting in the trend in the final counts being corrected (black line).

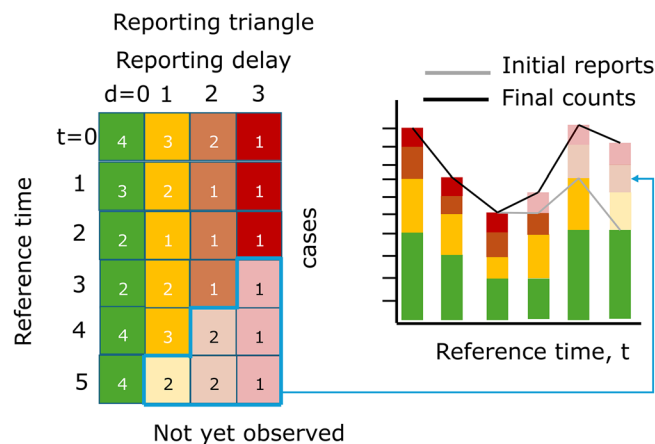


Figure 1. Illustration of the reporting triangle and corresponding case counts aggregated by reference time.

Rows of the reporting triangle represent reference times and columns represent reporting delays with numbers indicating the incident case count at that reference time and reporting delay. Colours indicate the delay, with solid shading indicating observed case counts and transparent shading indicating not yet observed cases, as of the most recent reference time. The bar chart shows the cases summed by reference time, with gray lines indicating the total number of cases reported at each reference time as of the most recent reference time, and black lines indicating the eventually observed number of cases.

There are a number of ways in which dealing with real-world data can create edge cases and complexities, such as for example when the “reference date” can refer to different primary events, or when reporting delays are registered as being negative. We discuss these and ways to handle them in this framework in the *SI Text. Dealing with edge cases in real world data.*

Statistical framework

Overview

Our approach, based on the chain ladder method in the version described by Wolfram *et al.*,³ broadly consists of generating a point nowcast and estimating uncertainty in the point nowcast using past nowcast errors. We have broken this down further into five key steps, each of which are separate modular operations that can be modified (Figure 2). In the following, we give a high-level overview of our method. For the full details see the *SI Text. Mathematical model.*

Step 1: Delay estimation

- The empirical delay distribution can be estimated directly from a complete reporting matrix with N rows (reference times) by dividing the sum of each column (delays) by the total count. However, if we want to use the most recent data points for delay estimation, we have to impute the missing observations (bottom right of the reporting triangle, Figure 2).
- The expected values in the partial reporting triangle are imputed by assuming that the delay distribution is constant over time and:
 - a. First compute the ratio between the first delay and the second delay for all reference times when both delays are fully observed (e.g. the ratio of the sum of the first 5 green vs yellow rows in Figure 1 $(4+2+2+3+4)/(3+2+1+2+3) = 1.36$)
 - b. Divide the partially observed row of the first delay column by the ratio to impute the missing component of the second delay (e.g. the case count at reference time = 5 and delay = 0, 4, from Figure 1, is used to impute the case count at reference time = 5, delay = 1 as $4/1.36 = 2.93$)

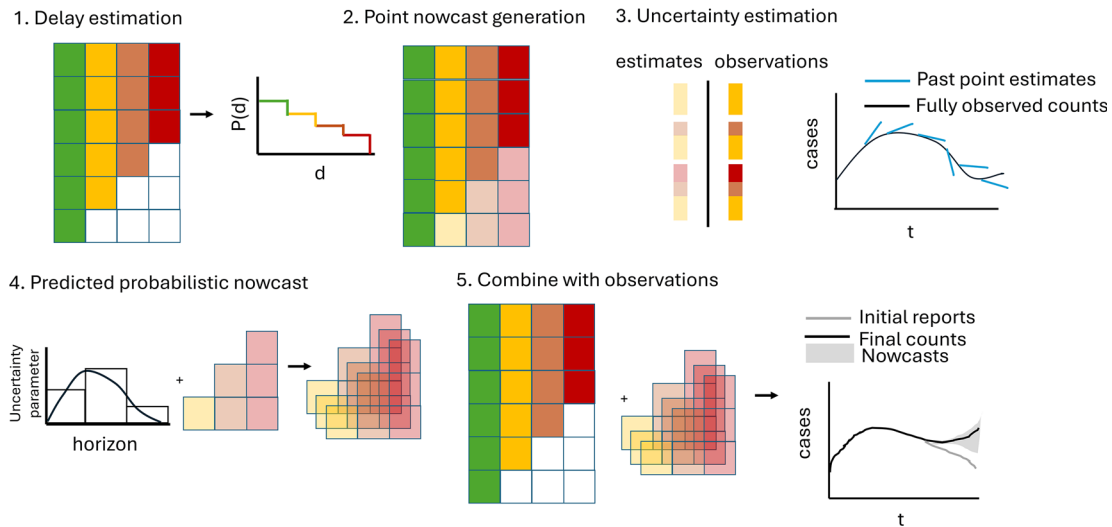


Figure 2. Schematic illustration of the baselinenowcast method. The reporting triangle contains the number of cases by reference time (rows) and reporting delay (columns), where colours represent the delays, and with white indicating observations that have not yet been observed and shading here indicating whether values have been observed (solid) or imputed (shaded). The delay distribution is estimated empirically from the reporting triangle. Point nowcasts are generated using the delay distribution and the partial observations at each reference time to impute the missing values in the bottom right of the triangle. Uncertainty is quantified by repeating this procedure for a number of past retrospective nowcast times and comparing the estimates to the observations. The point nowcast predictions and the uncertainty parameters are used to generate probabilistic predictions, and these are combined with the observations to generate probabilistic nowcasts.

- c. Repeat for the missing elements of each column filling in from left to right.

Step 2: Point nowcast generation

- Given a delay distribution and a reporting triangle, we estimate the missing observations to generate a point nowcast matrix: a complete reporting matrix where the missing observations have been filled in with point estimates (indicated by the shaded bottom right of the triangle in [Figure 2](#), step 2). The rows sum to a point estimate of the final counts at each reference time.
- We use a similar approach to that used to impute the missing elements of the reporting triangle in step 1 but now we:
 - a. Estimate the final case counts by dividing the current total reported counts by the proportion estimated to be reported by this time (e.g. if there are 4 case counts observed so far at a reference time of 5 and a delay of 0 from [Figure 1](#), which is estimated to represent 40% of eventual cases from the delay distribution in step 1 of [Figure 2](#), we expect $4/0.40 = 10$ total final cases)
 - b. For each unobserved delay, we multiply the estimated final case counts by the proportion of cases observed on that delay (e.g. if an estimated 10% of cases are reported with a delay of 3 from the delay distribution in step 1 of [Figure 2](#) and we expect 10 total cases, we expect $10 * 0.1 = 1$ to be reported with a delay of 3)
- We implement a slightly more complex version of this to account for zero reported cases observed so far.

Step 3: Uncertainty estimation

- We use past nowcast errors compared to observed values as of the current nowcast date to quantify uncertainty.
 - a. We generate retrospective point nowcasts (light shaded bars and blue lines in [Figure 2](#), step 3) by repeating steps 1 and 2 for a number of past, retrospective nowcast times, M , using N reference times for delay estimation at each retrospective nowcast time. This requires $N + M$ total reference times, which means that the total reference times used for uncertainty estimation overlaps with those used for delay estimation, as each retrospective nowcast requires re-estimating a delay using N reference times.
 - b. An observation model is fit to the predicted (shaded bars, blue lines [Figure 2](#), step 3) and observed (solid bars, black lines [Figure 2](#), step 3) components for each nowcast horizon.

Step 4: Predicted probabilistic nowcast

- We generate probabilistic predictions by sampling draws (indicated by the stacked shaded bottom right reporting triangles in [Figure 2](#), step 4) from the observation model at each horizon with a mean given by the point nowcast and the horizon specific uncertainty parameter (bar chart with distributional overlay in [Figure 2](#), step 4) estimated in step 3.
- By default, the target we compare to is the sum of the not-yet observed components of the point nowcast for each reference time.

Step 5: Combine with observations

- We then add what has already been observed (reporting triangle [Figure 2](#), step 5) to each draw (shaded bottom right of reporting triangle [Figure 2](#), step 5) of the predicted probabilistic nowcast for what has yet to be observed to create a complete probabilistic nowcast of the final counts for each reference time point (indicated by gray shading in line plot [Figure 2](#), step 5).

Default methodological choices

To guide users, we set default behaviours that are based on the maximum delay observed in the data making them more robust to different datasets. The maximum delay is set by the user when creating the reporting triangle based on when the data is expected to stabilise, with shorter maximum delays being more computationally efficient.

1. **Overall data use:** We use three times the maximum delay number of reference times (in line with Wolfram *et al.*³).
2. **Delay distribution estimation:**
 - a. The delay is estimated from the most recent reference times, using the (partial) reporting triangle being nowcasted.
 - b. The number of reference times used for delay estimation, N, is 50% of the total training data with a minimum requirement of at least one more than the number of reference times with partial observations.
- 3. **Uncertainty Estimation:**⁵
 - a. The additional cases at each reference time (the sum of the not-yet observed components of the point nowcast) are assumed to be negative binomially distributed.
 - b. The number of reference times used for producing retrospective nowcasts, M, is 50% of the total training data. If there is insufficient data, the method uses the oldest retrospective nowcast time with at least N reference times for training and requires that there are at least two retrospective nowcasts containing sufficient data.

Available options for model specification

As our method is modular with each step being independent, there are a large number of potential variations, designed with common scenarios in infectious disease surveillance in mind. We have implemented the following:

Input data options:

- Use of a complete reporting matrix or a reporting triangle.
- Changing the total number of reference time points used to estimate the delay distribution and uncertainty parameters, balancing precision (from more data) against capturing recent variability (from more current data).
- Reference and report dates of different temporal granularities, e.g. daily reporting of weekly reference dates, weekly reporting of daily reference dates, and weekly reference and report dates.

Delay estimation approaches:

- Estimate delays from the same data set/stratum being nowcasted or “borrow” estimates from another reporting triangle. This may be useful for sparse data sets, such as in small subpopulations or locations or for rare pathogens and emerging outbreaks.
- Changing the proportion of data used to estimate the delay with the same tradeoffs as for changing the overall amount of data input.
- Adjusting the maximum delay, but this should be based on empirically observed delay rather than being viewed as a modelling option.

Uncertainty quantification variations:

- Observation models can be fit using different datasets than those used for delay estimation or for the nowcast.
- Changing the proportion of data used to estimate the observation model with the same tradeoffs as in the delay estimation case.
- The parametric form of the observation model can be changed.
- The observation model can be fit jointly for any grouping of observations. The target choice for the final nowcast can be specified such that the uncertainty estimation occurs on this target (i.e. 7-day rolling sum).

Software implementation

Our baseline nowcasting methods are implemented in the R package *baselinenowcast*, which provides a modular pipe-friendly interface (allowing functions to be chained together using pipe operators) for nowcasting right-truncated surveillance. The package has minimal dependencies, primarily relying on base R functionality.²⁸ Worked examples applied to common datasets, as well as documentation of package functions and arguments, are provided as part of the package documentation at <https://baselinenowcast.epinowcast.org/index.html>. A primary workflow function *baselinenowcast()* implements the default method, whilst individual component functions enable customisation of specific steps through a consistent interface. The package accepts input data in a standard data frame format with columns for reference date, report date, and count. An additional feature is its compatibility with the *epinowcast* R package objects,²⁰ allowing users to leverage that package's preprocessing and data conversion functions and facilitating comparison to more *epinowcast* nowcasting models. Low-level functions use a matrix input for efficiency. We followed software best practices, including unit testing, continuous integration, code review via pull requests, and documentation driven development throughout the package's development (<https://github.com/epinowcast/baselinenowcast>).^{29,30}

Evaluation analysis

To evaluate nowcast performance from quantile-based outputs, we use the weighted interval score (WIS). WIS is a widely used proper scoring rule used to assess forecast quality of probabilistic quantile-based forecasts^{24,31,32} and is the probabilistic equivalent of the absolute error of point forecasts. We decompose WIS into overprediction, underprediction, and dispersion components. As WIS is an absolute measure of error, we compute the relative WIS throughout, with the default specification of our baseline method used as a comparison unless otherwise noted. We assess forecast calibration using coverage metrics, where we compute the proportion of all observations within a specified prediction interval. For example, a well-calibrated forecast would have 50% of the observations fall within the 50% prediction intervals (between the 25th and 75th quantiles). See the following sections for details on the quantiles used to compute the WIS scores and coverage metrics.

German COVID-19 Nowcast Hub validation

To verify that our *baselinenowcast* implementation worked as expected and to characterize the performance of our method overall and across age groups and nowcast dates, we produced nowcasts using our method's default specification and compared them to the nowcasts produced using the baseline nowcasting method described in Wolfram *et al.*³ (referred to in that paper as KIT simple nowcast) which our method and default model settings were based upon. Both methods were applied to the COVID-19 hospitalization data from the German COVID-19 Nowcast Hub (the Hub).³³ The dataset represents 7-day rolling sums of hospitalisation incidences due to COVID-19 in Germany, which were used to guide public health policy. German hospitalization counts were aggregated by the date of positive test result rather than admission (see Wolfram *et al.*³ for details). They were hence particularly affected by delays, as these not only comprised the time from hospitalization to report, but also from positive test to hospitalization. Data then sometimes took several weeks or months to fully stabilize. We retrained the model and produced nowcasts each day from 22 November 2021 to 29 April 2022, with preliminary data available from 1 July 2021 used for training the earlier models. Because we retrained the model each time we produced a nowcast, the delay estimate is always based on the latest data available at the time the nowcast is made. We used the pre-processed dataset available in the Hub's GitHub repository,³³ which contains a cleaned version of the reporting triangle that teams submitting to the Hub were instructed to use.

Because the 7-day rolling sum was the epidemiologically relevant surveillance statistic and the one used by the German Nowcast Hub, we generated nowcasts of the 7-day rolling sum of COVID-19 hospitalisation incidence in Germany each day for national-level aggregated data and age-stratified data. We applied our baseline nowcast with a maximum delay of 40 days, consistent with the specification of the KIT simple nowcast model used in 3, and with a training volume of

120 reference dates used by each nowcast (this means that 60 reference dates were used for delay estimation and 60 retrospective nowcast datasets were created for uncertainty estimation, with the latter using all 120 reference dates). For consistency with the KIT simple nowcast model, we estimated uncertainty in the predicted components of the final “target”, which in this case was the 7-day rolling sum of the partial expectation of hospitalisation incidence. This made our baseline approach theoretically equivalent to the KIT simple nowcast model described in 3, except that they did not include a correction for zero counts in the partial observations for a given reference time.

However, when verifying our method against the KIT simple nowcast codebase, we identified an issue impacting the dispersion estimate (see SI Text. *Description of KIT simple nowcast implementation issue and revised nowcasts* for more details). We fixed the issue in our fork of the repository and retrospectively generated new nowcasts, which we labelled as the “revised KIT simple nowcast”. We use this as a point of comparison in the main text results, with a comparison to the original implementation detailed in the supplement and referred to as the “original KIT simple nowcast”.

We visualised summary statistics derived from the data. We plotted the mean reporting delay over time for national-level aggregations and individual age groups to identify any temporal patterns. We also visualised the full distribution of delays nationally and by age group to characterise differences in reporting processes across strata.

We then visualised the nowcasts against the eventually observed values to qualitatively verify equivalence for a subset of nowcast dates. We then quantitatively evaluated both sets of nowcasts using a rolling evaluation dataset of 40 days after each nowcast date, restricted to a maximum delay of 40 days for each reference date, based on findings from the analysis performed in section 3.7 of Wolfram *et al.*³ and consistent with other nowcast evaluations^{4,7,22}. This means that for each nowcast date, we used a different evaluation dataset corresponding to the data available 40 days after that nowcast date, which ensures that recent and older nowcasts are evaluated against a dataset with the same lag. This reduces any bias that could be introduced from very long delays. Which would systematically bias towards poorer performance in nowcasts made further in the past. We evaluated nowcasts for horizon days 0 to - 28 days. We used the same quantiles as in 3 (0.025, 0.1, 0.25, 0.5, 0.75, 0.9, 0.975) to compute the WIS. We computed WIS scores for our baseline method, the revised KIT simple nowcast, and the original KIT simple nowcast. We also calculated the empirical coverage at the 50% and 95% prediction intervals. We calculated the relative WIS across multiple dimensions: by nowcast horizon (from 0 to - 28 days), by nowcast date, and by age group.

Performance of different baselinenowcast method specifications applied to German COVID-19 data

To assess whether other possible *baselinenowcast* method specifications designed specifically for the epidemiological context could further improve performance, we performed a sensitivity analysis of the following pairwise method comparisons: borrowing delay and uncertainty estimates from national-level data when nowcasting for specific age groups, using only complete reporting matrices versus reporting triangles, using shorter (50%) and longer (200%) training periods than those used in the default specification, and estimating delays independently for each weekday and combining.

We generated nowcasts of the 7-day rolling sum of COVID-19 hospitalisation incidence in Germany each day for the same period as in the validation study for all age groups with each method specification. We visualised predictions against the 40-day ahead rolling evaluation dataset values, focusing on same-day nowcasts, and again used the weighted interval score (WIS) and coverage metrics for quantitative assessment, using the same quantiles and prediction intervals as described in the validation section. We again stratified our analysis by nowcast horizon (from 0 to - 28 days), and by age group. We calculated WIS for all method specifications and relative WIS compared to the default method specification.

Case study: UKHSA norovirus surveillance

To further validate our approach and explore its utility as a baseline method we produced nowcasts using three specifications of our baseline method for UKHSA’s norovirus surveillance data, and compared the performance of those nowcasts to the performance of three models in Mellor *et al.*² including an alternative baseline approach. These data consist of laboratory-confirmed norovirus cases in England from winter 2023/2024. Each case record includes a specimen date (when the sample was collected) and a report date (when the test result was ingested into the surveillance system). The difference between these dates forms the reporting delay, which typically peaks at 1–2 days with most reports occurring within 7 days. However, approximately 20% of specimens obtained in a given week were not reported by the end of that week, creating a substantial right-truncation effect in recent data. Because of the variable results across days of the week observed in the original study, we assumed that the dataset exhibited strong day-of-week-periodicity, with reporting patterns varying systematically across weekdays.

We implemented our baseline nowcasting method using three different method specifications. All method specifications used a maximum delay of 14 days to match the approach in the original study for the Generalized Additive Model (GAM) model. The first method was our default baseline method, except that we fit to the most recent 56 days of data rather than the default of 42 days (3 times the maximum delay of 14 days). This was motivated by following the settings used for the GAM model so that the methods used equivalent amounts of training data and so were more directly comparable. We refer to this method specification as the base specification. Based on our hypothesis that day of the week variations in the delays were present we used two further methods. The second method independently estimates nowcasts for each day of the week, using the most recent 11 weeks of data (11 data points per day of week), to use a similar amount of data to the base specification (56 days), though in practice we needed more than 56 days of data due to data sparsity in some weekdays. The third model had the same structure as the second model but with more training data, instead using the 56 latest data points from each particular day of the week (spanning approximately 56 weeks of historical data) or as much as was available, to use the same amount of training data as the base specification for each weekday's delay estimation. We refer to these method specifications as the *baselinenowcast* weekday filter small/large training volume specifications respectively.

Following Mellor *et al.*,² we nowcast each Sunday from November 5th 2023, to March 10 2024 and evaluated performance on horizons 0 to -7 days using rolling 50-day ahead of the nowcast date evaluation datasets (similar to the rolling evaluation applied to the German COVID-19 data). For consistency with the evaluation in Mellor *et al.*,² we did not restrict the evaluation dataset to only reports within the maximum delay. We then compared nowcast performance from three of the models considered in Mellor *et al.*: a GAM, an implementation of epinowcast, and a naive baseline method. The GAM was developed by the authors of the original study² and uses a latent model of the number of cases as a function of reference time and delay, accounting for weekday effects in the reference time and the reporting delay. The epinowcast implementation uses the R package *epinowcast*²⁰ to estimate a Bayesian hierarchical model of the number of observed cases by reference time and delay, using a latent variable approach to model the underlying trend in the epidemic growth rate accounting for weekday effects in the underlying infections. The Mellor *et al.*² baseline implementation assumes that the cases predicted each day of this week will be equal to the observed count from the previous week (essentially propagates the previous week forward), producing only a median nowcast.

We first visualised the nowcasts against the eventually observed values to qualitatively assess performance. To align with Mellor *et al.*,² we calculated WIS using only the 50% and 90% prediction intervals. Coverage metrics were also calculated using the 50% and 90% prediction intervals. We otherwise followed the evaluation approach of our earlier case studies, using our base specification as the baseline for the relative WIS calculation. We stratified by day of week in addition to week of nowcast, and stratified empirical delays by day of week as well as over time. Unlike the original paper, which examined both daily and weekly aggregated scores, we focused on daily scores.

Evaluation analysis implementation

Our analysis pipeline was implemented using *targets*³⁴ for reproducible workflow management, with *scoringutils*^{32,35} used for nowcast evaluation. All analyses were performed in R²⁸ with key dependencies including *dplyr*³⁶ for data manipulation and *ggplot2*³⁷ for visualisation. The code and analyses are available at <https://github.com/epinowcast/baselinenowcast-paper>.^{38,39}

Results

German COVID-19 Nowcast Hub validation

Because our baseline method's implementation is a generalization of the KIT simple nowcast used by Wolfram *et al.*,³ we initially verified our method by comparing our baseline method's results against retrospective nowcasts generated directly from the KIT simple nowcast codebase using the same input data. In this process, we identified a difference in the outputs despite following identical model specifications and having nearly identical mathematical methods, which led us to further investigate the method implemented in the original KIT simple nowcast codebase. We then identified a bug in the original implementation (see SI Text. *Description of KIT simple nowcast implementation issue and revised nowcasts* for more details). We consulted with the authors, whom are co-authors of this work, fixed the bug in our fork of the repository, and regenerated retrospective nowcasts which we have labelled as the revised KIT simple nowcast, with comparisons to the original implementation in the Supplement labelled as the original KIT simple nowcast (SI Figure 1).

Our method produced equivalent nowcasts to the revised KIT simple nowcast (Figure 3A, 3B, 3C). Compared to the original KIT simple nowcast, our baseline method performed better, with an overall relative WIS of 0.88 nationally and 0.91 by age groups (SI Figure 1A, B), indicating that the bug fix improved nowcast performance. Exploratory data analysis indicated that the delay distribution (Figure 3E) differed between age groups, with the 00–04 age group having

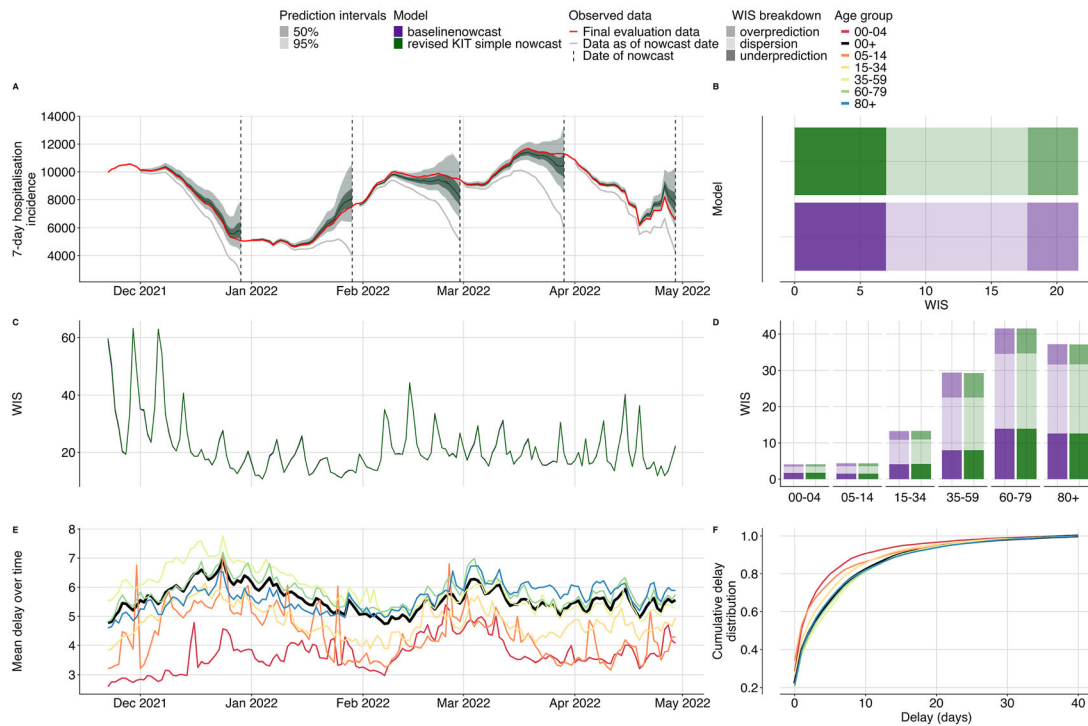


Figure 3. Validation of our baseline nowcasting method using German COVID-19 data compared to the revised KIT simple nowcast model. **A.** Nowcasts from our baseline model and the revised KIT simple nowcast model for all age groups plotted against eventually observed values aggregated across all age groups, coloured by model. Because both sets of nowcasts were nearly identical, colouring and shading is not distinguishable between the two. Observed data are shown as solid lines with gray indicating data available at the nowcast date and red indicating the final evaluation data. Dashed vertical lines indicate the date of the nowcasts. Shading indicates prediction intervals of each nowcast. Nowcasts are shown from horizons 0 to -28 days. **B.** Decomposed overall WIS aggregated across age strata. Shading indicates WIS component (dispersion, overprediction, underprediction), colour indicates model. **C.** Daily mean WIS summarised across age strata and 28 horizon days, coloured by model. Identical WIS over time renders the two lines indistinguishable. **D.** Decomposed WIS across age groups. Shading indicates WIS component, colour indicates model. **E.** Mean reporting delay over time. Colour indicates age group, with the national average highlighted in black. Mean reporting delays were calculated using the last 40 reference dates, corresponding to the maximum delay, for each nowcast date and age group. **F.** Cumulative distribution functions (CDFs) of the mean delay distribution across the entire study period. Colour indicates age group with the national average highlighted in black.

the shortest mean delay of 3.71 days whilst the 35–59 age group experienced the longest delays, averaging 5.88 days. The mean reporting delay showed changes over time roughly on the time scale of changes in case counts (Figure 3E, 3A).

Our baseline method was consistently less biased than no adjustment (gray line Figure 3A). Underprediction, overprediction, and dispersion contributed approximately 32%, 18%, and 50% respectively of the total score across age groups. Our method had 55% coverage at 50% prediction intervals and 96% coverage at 95% prediction intervals, indicating that a greater proportion of observed data points fell within the prediction intervals than would be theoretically expected in a perfectly well-calibrated model (SI Figure 6,7). There was moderate variation across nowcast dates that appeared to be correlated with changes in reporting delays and peaks in case counts (Figure 3A, 3C, 3E), as is to be expected as higher case counts will be associated with scores of a higher magnitude. Across age groups, WIS ranged from 4.05 for the 00–04 age group to 45.76 for the 60–79 age-group, again reflecting the magnitude in case counts with 00–04 contributing the fewest hospital admissions and 60–79 year olds contributing the most admissions (Figure 3D).

Performance of different *baselinenowcast* model specifications applied to German COVID-19 data

We found varying patterns of relative performance compared to the default configuration, depending on age group and method specification (Figure 3). Borrowing delay and uncertainty estimates for each age group from aggregate data overall reduced nowcast performance (relative WIS (rWIS) of 1.14, Figure 4B). Using only the complete reporting matrices, rather than the default of including the most recent partial reports, resulted in a slight reduction in performance

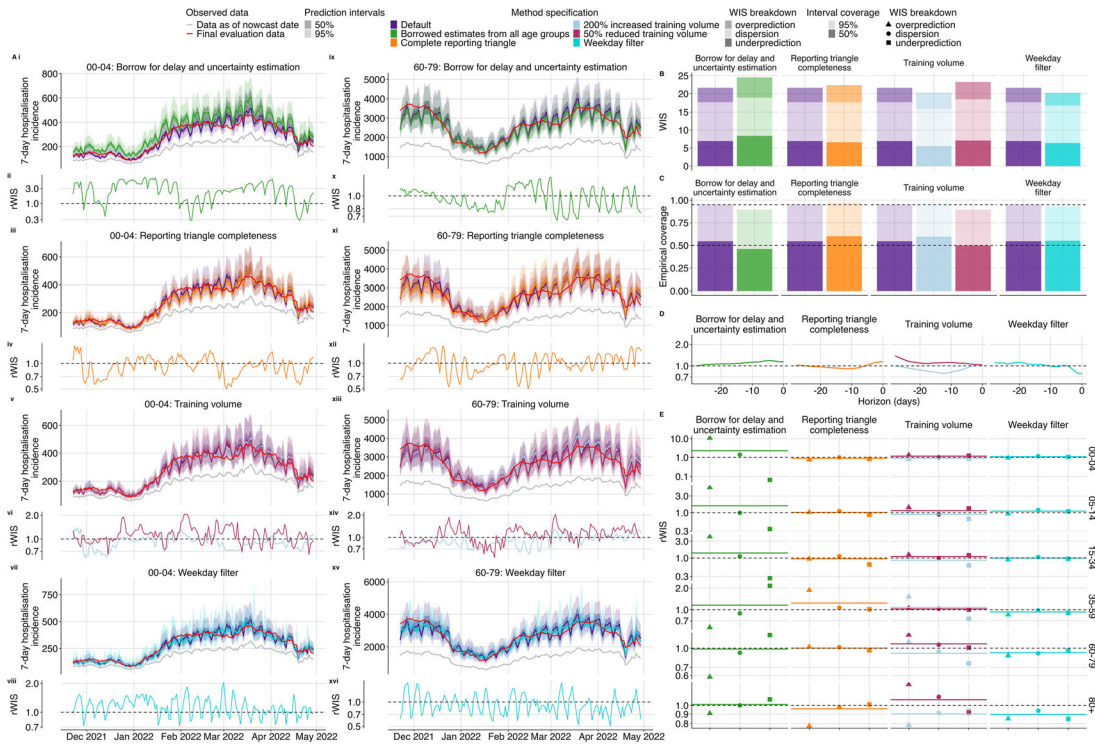


Figure 4. Performance comparison of different *baselinenowcast* method specifications compared to the default method. **A.** Upper plots show the daily same-day (0- horizon) nowcasts from the 00–04 (left) and 60–79 (right) age group for each method specification grouping showing the default validation model alongside variants from a specific permutation category: first row shows borrowing strategy variations (i, ix), second row shows reporting triangle completeness variations (iii, xi), third row shows training data volume variations (v, xiii), and the fourth row shows weekday filter variations (vii, xv). Observed data are shown as solid lines with gray indicating data available at the nowcast date and red indicating the final evaluation data. Colours indicate the method specification throughout, with the default approach in purple. Lower subplots show the relative WIS over time from the 00–04 (left) and 60–79 (right) age group relative to our default model with first row showing borrowing strategy variations (ii, x), second row showing reporting triangle completeness variations (iv, xii), third row showing training data volume variations (vi, xiv), and the fourth row showing weekday filter variations (viii, xvi). The horizontal reference line at 1.0 indicates performance parity with the validation default, values below 1.0 indicate improved performance of the method specification being considered. **B.** Overall performance comparison between all model configurations across all age groups, with decomposed absolute WIS (dispersion, overprediction, underprediction). Colour indicates method specification, with the default approach in purple. **C.** Empirical coverage for age group stratified nowcasts at 50% and 95% prediction intervals by method specification. Horizontal reference lines indicate the nominal coverage levels (50% and 95%). **D.** Relative WIS by nowcast horizon summarised across age groups. The horizontal reference line at 1.0 indicates performance parity with the default with values lower than 1.0 indicating improved performance of the method specification being considered. **E.** Relative decomposed WIS by age group, faceted by method specification grouping. Shape indicates relative WIS component (underprediction, dispersion, and overprediction).

overall (rWIS 1.04, Figure 4B). Doubling the training volume used for delay and uncertainty estimation improved performance (rWIS 0.94, Figure 4B), while reducing it by half decreased performance (rWIS 1.08, Figure 4B). Applying the weekday filter method improved performance by the greatest amount out of all the method specifications we explored (rWIS 0.93, Figure 4B).

Borrowing delay and uncertainty estimates from national data for each age group had the most pronounced reduction in performance in the sparsest age group (00–04) (rWIS 2.28), with the degradation in performance compared to the default specification driven by overprediction (Figure 4Ai, 4Aix, 4E, SI Figure 10). This reduction in performance when borrowing delay and uncertainty from all age groups is possibly due to the shorter delay in the 00–04 age group compared to the overall delay across all age groups (Figure 3E, 3F). The greatest improvement amongst the age strata from borrowing delay and uncertainty estimates was modest, occurring in the 60–79 age group (rWIS 0.99) (Figure 4E, SI Figure 10). Borrowing strategies showed reduced prediction interval coverage compared to the default (Figure 4C). Borrowing strategies also resulted in reduced performance at recent horizons (Figure 4D) with the least available data,

likely reflecting the reduced performance for the 00–04 age group at a horizon of 0 (Figure 4A). Averaging across age strata, performance was reduced due to borrowing for nearly all nowcast dates, except for at the beginning of February when we saw a slight improvement from borrowing (SI Figure 12), which might be due to the rapid fluctuation in the mean delay in the 05–14 age group a few weeks prior (Figure 4E).

Compared to the default, which uses the most recent partial observations in the reporting triangle, visual inspection of nowcasts using only complete reporting matrices showed slightly improved performance in the 00–04 age group and degraded performance in the 60–79 age group (Figure 4Aiii, 4Aiv, 4Axi, 4Axii). The greatest improvement from using complete reporting matrices was in the 00–04 age group (rWIS 0.90, Figure 4Aiii, 4Aiv) and the greatest reduction in performance was in the 35–59 age group (rWIS 1.24) (Figure 4E, SI Figure 10). Across all age strata, complete reporting matrices demonstrated improved performance at the end of December 2021 and February 2022, with reduced performance elsewhere. This suggests improved performance from using complete reporting matrices during times of stable reporting delays (SI Figure 12). Reporting triangle completeness resulted in improvements near intermediate horizons -8 to -15 (Figure 4D) and increased coverage (Figure 4C).

Varying training data volume resulted in the most substantial differences in performance in the 00–04 age group which had the lowest counts. For this age group, doubling training volume showed improved performance (rWIS 0.89), while reducing the training volume by 50% reduced performance (rWIS 1.16) (Figure 4E, SI Figure 10). Visual inspection of the nowcast performance over time in the 00–04 and 60–79 year old age groups shows that trends in performance from increased and decreased training volume tend to oppose one another (Figure 4Av, 4Avi, 4Axi, 4Axiv), indicating that there might not be a single optimal setting due to tradeoffs between overfitting to noise in reporting delays versus failing to incorporate the most recent temporal changes in reporting delay (SI Figure 12). Doubling the training volume resulted in improved performance at horizons between -8 and -15, whereas a 50% reduction in the training volume had considerably worse performance at distant horizons (-20 to -28) (Figure 4D). Increasing training data volume resulted in higher coverage (98% coverage), while reducing training volume reduced coverage (90% coverage), compared to the default (96% coverage) at 95% prediction intervals (Figure 4C).

Performing independent nowcasts of each weekday and combining them improved nowcast performance by reducing the large day-of-week variation observed in the default method (Figure 4Avii, 4Aviii, 4Axv, 4Axvi). Accounting for weekday effects in reporting had the greatest performance improvement in the most recent 7 horizon days (Figure 4D, SI Figure 9). Across age groups, the greatest performance improvement was observed in the 80+ age group (rWIS of 0.89), followed by the 60–79 age group (rWIS 0.92), with reductions in performance observed in the 00–04 and 05–14 age groups (rWIS of 1.09 and 1.11 respectively), which may be due to less of a consistent weekday effect in the sparser age groups compared to the larger age groups (Figure 3E).

Case study: UKHSA norovirus surveillance

Visual inspection indicates that nowcasts from our base specification showed better alignment with eventual observed data compared to the naive baseline approach used in Mellor *et al.*² which was based on projecting the previous week's data forward (Figure 5A, 5B, SI Figure 13). Comparing the external model implementations to our base specification, the epinowcast model had slightly better performance than our base specification (Figure 5E, 5F, SI Figure 13), with similarly high levels of uncertainty, while the GAM model showed reduced uncertainty and better overall performance compared to our base specification (Figure 5C, 5D, SI Figure 13).

Overall, compared to our base specification, the Mellor *et al.*² baseline performed worse (rWIS 2.29), and both the the GAM model (rWIS 0.65) and the epinowcast implementation performed better (rWIS 0.89) (Figure 5I). The GAM model performed best relative to our base specification in December 2023 and late January to early February 2024 (Figure 5D), whereas the epinowcast implementation performed best relative to our base specification early December (Figure 5F). The absolute WIS of our base specification and the Mellor *et al.* models was consistent over time except for on January 14, 2024 (SI Figure 13), where all models performed poorly due to what appears to be initial underreporting in the most recent two days resulting in underpredictions by all models. Relatively speaking the GAM and epinowcast models both performed better than our base specification at this point (Figure 5D, 5F, SI Figure 13).

Our base specification and the GAM model were similarly well-calibrated (SI Figure 15) with slightly different coverage patterns. All models had 50% prediction intervals that were too wide with 69% (base specification), 59% (GAM), and 58% (epinowcast) coverage. At 90% prediction intervals, all models were well covered at 90% (base specification), 91% (GAM) and 88% (epinowcast) coverage (SI Figure 15). The Mellor *et al.*² baseline predictably undercovered for all prediction intervals and weekdays since it contained only a median estimate (SI Figure 15,16).

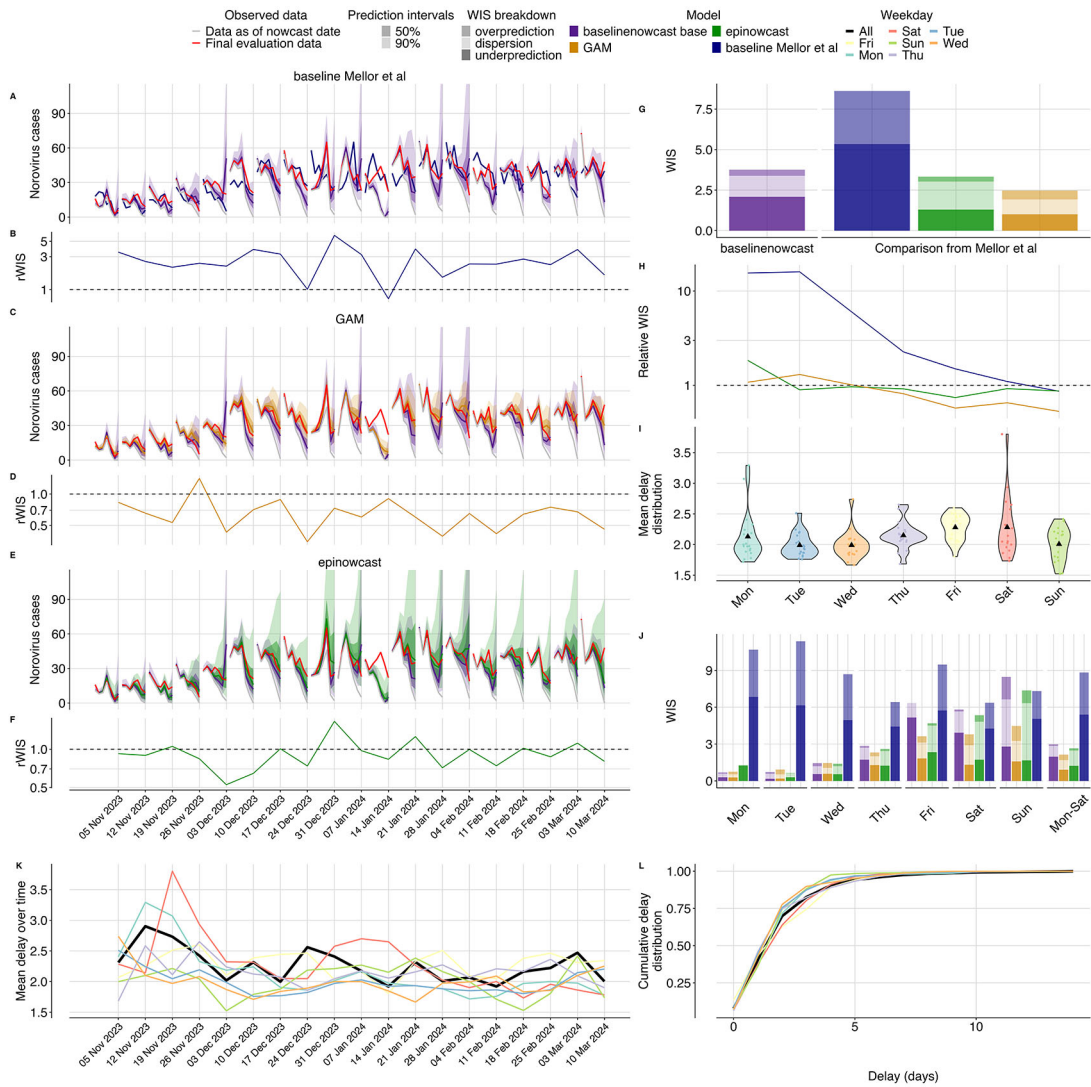


Figure 5. Nowcasting performance applied to UKHSA norovirus surveillance data. A - C - E - G. Weekly nowcasts compared to our base configuration, for three models used in Mellor *et al.* and compared to our two day-of-week filtered configurations. Observed data are shown as solid lines with gray indicating data available at the nowcast date and red indicating evaluation data. Nowcasts are shown out to 7 days as in Mellor *et al.* Shading indicates prediction intervals and colour indicates model, with the base specification in purple. **B - D - F - H.** Relative WIS over time (by week) compared to our base configuration. colours indicate models. Horizontal dashed line at 1.0 indicates performance parity with our base configuration, with value less than 1.0 indicating performance improvement of the method being considered. **I.** Overall performance comparison between all models, with decomposed absolute WIS (dispersion, overprediction, underprediction) displayed as stacked bar charts for direct comparison across our three variants and the three reference models from Mellor *et al.* Colour indicates model. **J.** Relative WIS by day of week for each model compared to our base specification with colour indicating model. Horizontal dashed line at 1.0 indicates performance parity with our base configuration. **K.** Distribution of mean reporting delays by weekday shown as violin plots with colours indicating the weekday. Black triangle indicates the mean delay for that weekday across nowcast dates. **L.** Mean reporting delays were calculated using the last 14 days of data (corresponding to the maximum delay). **M.** Mean reporting delay over time, with colours indicating the day of week and black indicating the mean reporting delay averaged across all weekdays. Mean reporting delays were calculated using the last 14 days of data (corresponding to the maximum delay). **N.** Cumulative distribution functions (CDFs) of the mean delay distribution. Colours indicate day of the week with black indicating the mean reporting delay averaged across all week days.

Performance varied substantially by day of week in the Mellor *et al.* models compared to our base specification (Figure 5J, 5L). As nowcasts were only conducted on Sundays, weekdays effectively reflect performance by nowcast horizon, though weekday effects in both specimen collection date and reporting also play a role. As expected, the most

recent day with the least amount of data (Sunday, 0 horizon) showed the worst performance across models, whilst days further in the past showed progressively improved performance (Figure 5L). When comparing the GAM method and our base configuration, the most pronounced relative reduction in performance of our base specification was on Sunday, largely driven by reduced dispersion in the GAM compared to our base specification (Figure 5J, 5L), which had high dispersion on Sundays. Both the GAM and our method performed similarly for horizons -7 to -4 (weekdays Mon-Thurs, Figure 5J, 5L). Comparing the epinowcast implementation to our base specification revealed the largest relative improvement in performance of the epinowcast implementation was on Friday, when our base specification tended to underpredict (Figure 5L). Our base specification performed better than the epinowcast implementation for horizons -7 to -4 (weekdays Mon-Thur, Figure 5L).

Comparing our model variants to the base configuration, both weekday filter small and large training volume configurations reduced performance (rWIS 1.51 and 1.26 respectively) (Figure 5G, 5H, 5I). The visual differences were most notable in December 31, 2023 and January 7, 2024, where the weekday-filter variants appeared to exhibit more dispersion and overprediction than our base specification (Figure 5G, 5H, SI Figure 13). Overall, our base specification underpredicted whereas the weekday filter variations had higher dispersion and over prediction (Figure 5I). At 50% prediction intervals the weekday filter small and large training volume configurations had 74% and 76% coverage respectively compared to 58% coverage in our base specification (SI Figure 15). Our weekday-filter configurations had less stable performance than our base model specification over time, with particularly poor relative performance of the weekday-filter with small training volume on November 26, 2023, and both configurations on December 31, 2023 and January 7, 2024 (Figure 5G, 5H, SI Figure 13). Similar to the GAM and epinowcast models, the relative performance differences were most pronounced on Sunday. Both weekday-filter variants performed worse relatively on Sunday compared to our base specification (Figure 5J), but generally performed similarly on other days of the week (Figure 5J,L). The reduction in performance on Sunday appeared to be due in large part to dispersion for both weekday-filter variants, and in particular overprediction for the weekday-filter variant with a small training volume (Figure 5L).

Discussion

In this work, we demonstrate that our baseline nowcasting method is both a practical tool for public health practice and can be used as a benchmark for further model development. We demonstrated that the method improved estimates of recent disease incidence compares to unadjusted data across all datasets evaluated. As the original implementation of the KIT simple nowcast, despite the implementation bug, was competitive with more complex methods in German Nowcast Hub,³ ours is likely to have been even more competitive. Analysis of the performance of additional epidemiology-specific features such as weekday-specific delay estimation and strata sharing applied to the COVID-19 data showed how exploratory data analysis can help determine optimal baseline method specifications catered to specific tasks. Our base specification substantially improved performance over the baseline used in the Mellor *et al.*² study, and helped to identify the horizons, weekdays, and nowcast dates in which the more complex models differed in performance compared to our baseline, highlighting areas for future model development.

In the German COVID-19 data used in the Hub, we found modest improvements from applying a weekday filter to account for day of week effects in reporting and from increasing the amount of training volume used for delay and uncertainty estimation. Had these features been implemented in the German Nowcast Hub, this model would have been even more competitive, approaching the performance of the top performing model. This highlights the importance of exploratory data analysis to identify patterns, such as weekday effects and the time scale of changing reporting delays, to help determine optimal method specifications. We found reduced performance when deploying a strategy to borrow delay and uncertainty estimates from across all age groups. This approach reduced performance in the age groups with a greater difference in their average delay compared to the national average, as using a misspecified delay generally led to overprediction. However, at times of rapidly fluctuating reporting delays and in settings with sparse data, we would expect borrowing to be more beneficial as it will act to reduce noise in the delay estimate due to within-strata outliers. These findings highlight that the distribution of delay estimates across strata, weekdays, and nowcast dates are key considerations when choosing model specifications, and can likely be used as a heuristic to choose which model specifications are most appropriate in different epidemiological contexts.

In the UKHSA norovirus dataset, the use of our baseline nowcasting method allowed us to distinguish between areas of improved and reduced performance of two of the more competitive nowcasting models, epinowcast and the GAM model. Previously, the baseline model used was so poor that it couldn't effectively illuminate horizons and nowcast dates with performance variability across models. In contrast to the German COVID-19 dataset, in the UKHSA norovirus dataset we did not see improved performance compared to the default configuration from fitting weekday specific delays, likely because this methods required that we use a longer historical time period to have sufficient historical data for delay and uncertainty estimation. In this case, it appears that the reporting delay shifted enough over time to favour a more recent

delay estimate, with less drastic weekday effects in delays than we had initially anticipated. These findings highlight that the distribution of delay estimates across strata, weekdays, and nowcast dates are key considerations when choosing model specifications, and can likely be used as a heuristic⁴⁰ to choose which model specifications are most appropriate in different epidemiological contexts.

Whilst we evaluated our approach across several datasets and through a range of method specifications, we did not consider a range of other possible method specifications. For example, users might also wish to use a different parametric distribution to model nowcast uncertainty. In this analysis we only considered negative binomial observation models applied to the target dataset, which in this case were always count data. Other methods for uncertainty estimation could have been explored, such as using bootstrapping or computing percentiles directly from past nowcast errors in a non-parametric form, or exploring other parametric observation models. The modular design of our package enables exploration of other valid error models. Additionally, users might wish to use our existing functionality in different combinations. For example, in this paper, we have only considered sharing information across all groups for both delay estimates and uncertainty, compared to independent estimates of both. A user might instead want to pool information for delay estimation but estimate uncertainty separately for each strata, or vice versa. We also only considered increasing training volume for both delay and uncertainty estimation together, and a user may wish to modify the combination of training data used for each task. A strength of our framework is that it supports these alternatives through its modular design.

For consistency and comparability with the original studies, we used the same evaluation and target metrics as those studies. This approach facilitated direct comparisons but meant we did not explore potentially more appropriate metrics, such as evaluating on log-transformed scales,⁴¹ which can more effectively capture relative differences when counts span multiple orders of magnitude, such as in the different age strata in the German Nowcast Hub COVID-19 data. Similarly, whilst WIS is a valid choice, we could have used Continuous Ranked Probability Score (CRPS), which scores probabilistic forecasts based on a full distribution of samples from the predictive distribution, avoiding information loss from using a limited set of quantiles. This was a particular issue in the Mellor *et al.*² comparison, where we followed their approach of using just two intervals (50th and 90th) when calculating WIS, where additional intervals may have better reflected the distribution of predictions. While this reduces the breadth of evaluation of the full predictive distribution, it has the benefit of focusing on evaluating only the prediction intervals that drive decision-making, namely those that are being visualised and presented to decision-makers. Another issue is that the use of retrospective datasets may not fully capture performance in truly prospective settings where evolving data availability and reporting changes create additional challenges. This is particularly true of the German COVID-19 data, where we used a final pre-processed reporting triangle for all analyses as vintages of that reporting triangle available each day were not systematically archived. However, we mitigated this issue by using case studies where data on reference and report times was available, as much as possible, as it would have been in real-time, and not tuning our models based on performance against these case studies. Additionally, for the German COVID-19 data we used the 7-day rolling sums as the nowcast target because this was what was epidemiologically relevant and used in the German Nowcast Hub, however, the use of this target meant that the daily noise and weekday effects were removed, preventing us from evaluating the methods ability to capture these effects. Our focus on proper scoring rules provides a standardised assessment but might not capture all dimensions of nowcast quality relevant to decision-making.⁴² Finally, the datasets selected, while diverse, do not represent the full spectrum of delay structures and reporting processes encountered in public health surveillance.

Multiple methods have been proposed as nowcasting approaches in infectious disease surveillance. Most of these methods jointly estimate reporting delays and the final number of cases,^{19,22,23,27,43} though other methods for nowcasting are similar to ours in that they require specifying an independently estimated delay distribution.⁴⁴ The joint approach has the advantage that the full set of training data is used to estimate both delays and uncertainty, whereas our baseline method requires use of a proportion of the training data for delay estimation, so that we have sufficient data to estimate uncertainty using the same amount of training data applied at different retrospective nowcast times. Other methods differ in how they model the time evolution of the final number of cases, the delay distribution, and the observation model. For example, some assume a parametric form for the reporting delay distribution, while others assume a non-parametric delay. A few methods enable estimation of time-varying delays.^{7,43} Commonly, hazard-based formulations are used.^{19,22,45} Methods also take varying approaches to smoothing, including the use of geometric random walks^{22,27,43} and B-splines.²³ Individual methods have also been extended to support modelling day of the week variation, covariate inclusion, public holiday anomalous reporting, and transmission processes.^{4,6,7,20,21} Bayesian methods^{6,7,19,20,27,45,46} can be particularly useful in settings with sparse data, where the ability to specify priors and expected reporting patterns is critical, for example, in novel outbreaks or situations with very low case counts. However, many of these methods are relatively complex which can make them difficult to inspect, validate, or extend, and may make it unclear when these models are inappropriate and prone to failure due to reasons such as unidentifiability or non-convergence. While other similar multiplicative methods

exist⁴⁷ and have been evaluated,⁴⁸ these methods are point-based rather than providing estimates with uncertainty, and do not contain the features developed here such as handling of initially observed zeros, strata sharing, and accounting for weekday effects. While other domains have implementations of a similar multiplicative approach to ours, such as the *chainladder* R^{16,18,17} package designed for nowcasting insurance claims, we have not seen widespread adoption of these methods in public health practice, likely due to the lack of examples in epidemiological contexts or the addition of features specific to the infectious disease context.

In contrast, our approach focuses on a minimal set of well-defined, modular components that are simple and easy to explain. The method has a modular, pipe-friendly design enabling users to customise specific components. By breaking the nowcasting method into distinct components, users can adapt, inspect and validate each part of the model. This simplicity makes implementation and model fitting straightforward without requiring specialist knowledge, making it suitable for implementation in public health agencies with moderate statistical expertise. A key advantage of our method over similar methods, such as the *chainladder* R package,¹⁸ is simply that it is tailored to common epidemiological contexts; with its modular design enabling the generation of nowcasts for different temporal granularities in reference and report dates, ease and demonstration of how to “borrow” delays and uncertainty across different strata, and the ability to independently fit each weekday to account for weekday effects in reporting, all of which are common scenarios encountered in epidemiology and which *baselinenowcast* was designed to handle. Additionally, based on retrospective performance, the uncertainty quantification approach provides relatively well-calibrated nowcasts. We note this was not the case for many of the models submitted to the original German Nowcast Hub³ and we consider this to be a key advantage of this baseline approach over other baseline methods such as the method used by Mellor *et al.*² which propagated the previous week’s data forward with no uncertainty. Lastly, by providing a standardised R package with comprehensive documentation (see <https://baselinenowcast.epinowcast.org/index.html> for tutorials and function documentation), our method aims to address challenges in reproducing and re-implementing different baseline methods, ensuring that analyses applied in one context can be readily extended to other pathogens/surveillance systems with relative ease and standardized approaches.

However, important limitations in our method exist. The method follows a pipeline approach where delay distributions are estimated independently from case occurrence processes. As demonstrated by 5, such approaches can propagate errors and may not fully account for interdependencies between these processes. The method also lacks flexibility for incorporating time-varying delay distributions or partially pooled estimates across strata, which may be crucial during periods of changing surveillance intensity or for rare diseases where flexibly borrowing information is essential.

Our approach to handling day-of-week effects highlights an important trade-off. By filtering data to use only observations from the same day of the week, we provide a simple solution that avoids complex modelling whilst effectively capturing weekly patterns. However, this simplicity comes at the cost of discarding potentially valuable data on recent trends in delay estimation, which proved to be problematic when performing independent estimates of the delay across a large time frame in the norovirus case study. More sophisticated approaches, such as those incorporating day-of-week terms in statistical models,²⁰ as was used by the GAM implementation in 2, can utilise all available data but require greater statistical complexity. This illustrates a broader tension in nowcasting between accessible methods that may sacrifice efficiency and complex methods that maximise statistical power but may be less accessible to practitioners.

More fully featured nowcasting frameworks, such as the *epinowcast* R package (a semi-mechanistic Bayesian approach,²⁰ which was used in the Mellor *et al.* study²), and generalised additive model-based approaches,⁴ encompass the functionality of other nowcasting approaches whilst offering greater flexibility, often via formula interfaces, and standardised tooling. This increased flexibility allows users to customise models for their specific problem reflecting their familiarity with details of the data and how they are collected, but has the downside of increasing model, software, and interface complexity. The evidence supporting the benefits of this increased flexibility is currently limited, as little work has been done to evaluate framework-based approaches.

Future research should focus on five key areas. First, developing standardised benchmarks and datasets would enable rigorous evaluation of nowcasting methods, quantifying trade-offs between simplicity and performance across different surveillance contexts. Second, research addressing practitioners’ needs through decision support tools could guide method selection based on data characteristics and available resources. Relatedly, co-creation of documentation with public health practitioners, as well as evaluation in the real-world public health context will be critical to help address implementation barriers. Thirdly, research is needed to implement a wider array of error models for improved uncertainty quantification. Fourthly, reporting processes can be complex, for example they can include downward revisions or spatial drop out in reporting. Both the type of empirically based methods we discuss in this work and more complex methods need to be adapted for these settings. Finally, advancements in model composition could enable seamless transitions between nowcasting and forecasting and address error propagation issues inherent in pipeline methods.^{5,49}

In conclusion, our baseline nowcasting method addresses gaps in the epidemiological toolbox by providing both a practical nowcasting solution for infectious disease surveillance and a standard benchmark for further model development. Its performance across datasets demonstrates that straightforward approaches can provide substantial improvements over unadjusted data whilst remaining accessible to practitioners. Our discovery of the implementation issue in the KIT simple nowcast method highlights the value of implementing reproducible, tested, and validated software. The performance of different method specifications highlighted the importance of exploratory data analysis to identify patterns in the delay distributions across time and strata. The method's pipe-friendly, modular design enables simple yet effective adaptations, such as the day-of-week stratification, without requiring complex statistical frameworks. Our method, developed with public health practitioners, is robust, interpretable, and relatively straightforward, evidence by our public health co-authors who were able to implement the package in their systems within a few days unsupported. By establishing this benchmark, we aim to reduce the barriers to implementing nowcasting solutions in public health practice and provide a resource for improving nowcasting methods, ultimately contributing to more timely and effective public health responses.

AI disclosure

Generative AI tools were used to assist with preparation of the manuscript text and as part of code review for both the *baselinenowcast* R package and the analysis applying it to the two case studies presented here. Coderabbitai <https://github.com/apps/coderabbitai> was used for automated code review. Claude Pro version 4 <https://claude.ai/> was used for assistance with writing code and for revisions of text in the manuscript. Claude Sonnet 4 <https://www.anthropic.com/claude/sonnet> was used for writing and revising code. Opus <https://www.anthropic.com/claude/opus> was used for assistance with writing.

Data availability statement

Extended data

Supplementary Information to this manuscript is available at:

Repository name: baselinenowcast-paper. <https://github.com/epinowcast/baselinenowcast-paper/blob/main/docs/Supplement.docx>

All code to reproduce results and figures are available publicly at:

Repository name: baselinenowcast-paper. <https://github.com/epinowcast/baselinenowcast-paper>

Version DOI at: <https://doi.org/10.5281/zenodo.18430500>

Functions used to generate nowcasts are available in the baselinenowcast R package available publicly at:

Repository name: baselinenowcast. <https://github.com/epinowcast/baselinenowcast>

Underlying data used to generate COVID-19 nowcasts from Germany is available publicly at:

Repository name: hospitalization-nowcast-hub. <https://github.com/KITmetricslab/hospitalization-nowcast-hub>.

Underlying data used to generate nowcasts of norovirus cases in England anonymized and with statistical noise added is available publicly at:

Repository name: norovirus-nowcast-baselinenowcast. <https://github.com/jonathonmellor/norovirus-nowcast-baselinenowcast>.

A pre-print⁵⁰ of this work is available at: <https://www.medrxiv.org/content/10.1101/2025.08.14.25333653v2>.

References

1. Ward T, Overton CE, Paton RS, *et al.*: **Understanding the infection severity and epidemiological characteristics of mpxv in the UK.** *Nat Commun.* 2024; **15**(1): 2199.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Mellor J, Tang ML, Finch E, *et al.*: **An application of nowcasting methods: cases of norovirus during the winter 2023/2024 in England.** *PLoS Comput Biol.* 2025; **21**(2): e1012849.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Wolfram D, Abbott S, An der Heiden M, *et al.*: **Collaborative nowcasting of COVID-19 hospitalization incidences in Germany.** *PLoS Comput Biol.* 2023; **19**(8): e1011394.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. van de Kassteelle J, Eilers PHC, Wallinga J: **Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing.** *Epidemiology.* 2019; **30**(5): 737–745.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Lison A, Abbott S, Huisman J, *et al.*: **Generative Bayesian modeling to nowcast the effective reproduction number from line list data with missing symptom onset dates.** *PLoS Comput Biol.* 2024; **20**(4): e1012021.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Bergström F, Günther F, Höhle M, *et al.*: **Bayesian nowcasting with leading indicators applied to COVID-19 fatalities in Sweden.** *PLoS Comput Biol.* 2022; **18**(12): e1010767.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Günther F, Bender A, Katz K, *et al.*: **Nowcasting the COVID-19 pandemic in Bavaria.** *Biom J.* 2021; **63**(3): 490–502.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Massachusetts syndromic surveillance data and moving epidemic method. In:
[Reference Source](#) | [Reference Source](#)
9. Monitoring mortality figures in the Netherlands. In: rivm.nl, [cited 15 May 2025].
[Reference Source](#)
10. Respiratory virus hospitalization surveillance network (RESP-NET). In: CDC RESP-NET, 10 Oct 2025.
[Reference Source](#)
11. Scholler S, Wuensche H, Fesser A: **COVID-19 hospitalisierungen in Deutschland.** In: *GitHub*, 15 May 2025; [cited 15 May 2025].
[Reference Source](#)
12. Wolfram D, Bracher J: **The RESPINOW Study Group, et al.: Integrating nowcasts into an ensemble of data-driven forecasting models for SARI hospitalizations in Germany.** *medRxiv.* 2025; 2025.02.21.25322655.
[Publisher Full Text](#)
13. Overton CE, Abbott S, Christie R, *et al.*: **Nowcasting the 2022 mpxv outbreak in England.** *PLoS Comput Biol.* 2023; **19**(9): e1011463.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Tang ML, McFarlane IS, Overton CE, *et al.*: **Nowcasting cases and trends during the measles 2023/24 outbreak in England.** *medRxiv.* 2025; 2025.04.08.25325363.
[Publisher Full Text](#)
15. Measles outbreak – July 15, 2025. [cited 16 Jul 2025].
[Reference Source](#)
16. Gesmann M, Tsanakas A: **Conference report: R in Insurance.** *R J.* 2014.
[Reference Source](#)
17. Lawless JF: **Adjustments for reporting delays and the prediction of occurred but not reported events.** *Can. J Stat.* 2009; **22**(1): 15–31.
[Publisher Full Text](#)
18. ChainLadder: *Claims reserving with R.* [cited 25 Jul 2025].
[Reference Source](#)
19. Meyer S, Held L, Höhle M: **Spatio-temporal analysis of epidemic Phenomena using the R package surveillance.** *J Stat Softw.* 2017; **77**(11): 1–55.
[Publisher Full Text](#)
20. Abbott S, Lison A, Funk S, *et al.*: **Epinowcast: flexible hierarchical nowcasting.**
[Publisher Full Text](#)
21. Lopes R, Bastos L: *Nowcaster: Nowcaster.* 2025;
[Reference Source](#)
22. McGough SF, Johansson MA, Lipsitch M, *et al.*: **Nowcasting by Bayesian Smoothing: a flexible, generalizable model for real-time epidemic tracking.** *PLoS Comput Biol.* 2020; **16**(4): e1007735.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Gressani O: *EpiLPS: a fast and flexible Bayesian tool for estimating epidemiological parameters.* [Computer Software]. 2021.
[Reference Source](#)
24. Cramer EY, Ray EL, Lopez VK, *et al.*: **Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States.** *Proc Natl Acad Sci U S A.* 2022; **119**(15): e2113561119.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Morasso C, Dolci G, Galazzo IB, *et al.*: **Guidelines for the choice of the baseline in XAI attribution methods.** *arXiv [cs.AI].* 2025;
[Reference Source](#)
26. European CDC **respiratory diseases forecasting hub.** *Respicast.* [cited 8 Aug 2025].
[Reference Source](#)
27. Höhle M, Ma d H: **Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011.** *Biometrics.* 2014; **70**(4): 993–1002.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. The R project for statistical computing. [cited 15 Apr 2025].
[Reference Source](#)
29. Johnson K, Abbott S, Bracher J, *et al.*: **baselinenowcast: Baseline nowcast.**
[Reference Source](#)
30. Johnson K, Abbott S, Funk S, *et al.*: **baselinenowcast: baseline nowcast from right-truncated epidemiological data.** *Zenodo.* 2025;
[Publisher Full Text](#)
31. Bracher J, Ray EL, Gneiting T, *et al.*: **Evaluating epidemic forecasts in an interval format.** *PLoS Comput Biol.* 2021; **17**(2): e1008618.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Bosse NI, Gruson H, Cori A, *et al.*: **Evaluating forecasts with scoringutils in R.** *arXiv [stat.ME].* 2022.
[Reference Source](#)
33. Lab KM, Bracher J, Wolfram D: **hospitalization-nowcast-hub: collecting nowcasts of the 7-day hospitalization incidence in Germany.** *GitHub.*
[Reference Source](#)
34. Landau W: *The {targets} R package user manual.* [cited 15 Apr 2025].
[Reference Source](#)
35. Bosse N, Abbott S, Gruson H, *et al.*: **scoringutils: utilities for scoring and assessing predictions.** 2025.
[Reference Source](#)
36. Wickham H, François R, Henry L, *et al.*: **dplyr: a grammar of data manipulation. CRAN: Contributed Packages.** *The R Foundation* 2014;
[Publisher Full Text](#)
37. Wickham H: **ggplot2: elegant graphics for data analysis.** New York: Springer-Verlag; 2016.
[Reference Source](#)
38. Johnson K, Abbott S: *baselinenowcastpaper.* 2025.
[Reference Source](#)
39. Johnson K, Abbott S, Bracher J: **epinowcast/baselinenowcast-paper: initial submission.** *Zenodo.* 2025.
[Publisher Full Text](#)
40. Gelman A, Vehtari A, Simpson D, *et al.*: **Bayesian Workflow.** *arXiv [stat.ME].* 2020.
[Reference Source](#)
41. Bosse NI, Abbott S, Cori A, *et al.*: **Scoring epidemiological forecasts on transformed scales.** *PLoS Comput Biol.* 2023; **19**(8): e1011393.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. Gerding A, Reich NG, Rogers B, *et al.*: **Evaluating infectious disease forecasts with allocation scoring rules.** *J R Stat Soc Ser A Stat Soc.* 2024; **188**(4): 1299–1325.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
43. Hinch R, Panovska-Griffiths J, Ward T, *et al.*: **Quantification of the time-varying epidemic growth rate and of the delays between symptom onset and presenting to healthcare for the mpxv epidemic in the UK in 2022.** *Sci Rep.* 2024; **14**(1): 19755.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
44. Abbott S, Hellewell J, Sherratt K, *et al.*: *EpiNow2: estimate real-time case counts and time-varying epidemiological parameters.* 2025.
[Reference Source](#)
45. Salmon M, Schumacher D, Höhle M: **Monitoring count time series in R: aberration detection in public health surveillance.** *J Stat Softw.* 2016; **70**(10): 1–35.
[Publisher Full Text](#)
46. Gressani O, Wallinga J, Althaus CL, *et al.*: **EpiLPS: a fast and flexible Bayesian tool for estimation of the time-varying reproduction number.** *PLoS Comput Biol.* 2022; **18**(10): e1010618.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

47. Tariq A, Yan P, Blichrodt A, *et al.*: **Nowcast-It: A practical toolbox for real-time adjustment of reporting delays in epidemic surveillance.** *Viruses*. 2025; **17**: 1598.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
48. Bizzotto A, Guzzetta G, Marziano V, *et al.*: **Increasing situational awareness through nowcasting of the reproduction number.** *Front. Public Health*. 2024; **12**: 1430920.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
49. Nicholson G, Blangiardo M, Briers M, *et al.*: **Interoperability of statistical models in pandemic preparedness: principles and reality.** *Stat Sci*. 2022; **37**(2): 183–206.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
50. Johnson KE, Tang ML, Tyszka EJ, *et al.*: **Baseline nowcasting methods for handling delays in epidemiological data.** *medRxiv*. 2025.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:    

Version 2

Reviewer Report 20 March 2026

<https://doi.org/10.21956/wellcomeopenres.28633.r147768>

© 2026 Guzzetta G. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Giorgio Guzzetta**

Fondazione Bruno Kessler, Trento, Italy

The manuscript has significantly improved in terms of clarity and precision of communication, and I now do not see major issues with the proposed methods and analyses.

I still think that the claims of the manuscript are overstated, especially in the opening paragraph of the discussion:

- the nowcasting method is not an innovation proposed by the authors (only the R package is).
- the practical use of the baseline nowcasting method for public health would require testing under simulated real-time conditions, whereas the authors admittedly used consolidated data for model training.
- the usability of the package for public health professionals needs to be demonstrated; the claim that "public health co-authors were able to implement the package in their systems" only constitutes anecdotal evidence, not a proper demonstration.

Other minor points below:

- In my opinion, Figure 3 and related results are just a sanity check, not a scientific result, and should be moved to the appendix.
- I still think Figure 4 is overloaded with information (26 panels!) and could benefit from moving non-essential information to the Supplementary Materials. The y label is missing in Figure 4D.
- A brief explanation of how the overdispersion parameter of the negative binomial observation model is chosen would be helpful.
- P. 7 "We implement a slightly more complex version of this to account for zero reported cases observed so far." -> A short explanation of how this is implemented in the main text is needed, at least in terms of general principles, along with a reference to the supplementary materials.
- Possible typos:
 - P. 4 "The chain ladder methods [refs], Lawless"
 - P. 17 "[...] random walks [refs], d B-splines"

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease modelling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 20 March 2026

<https://doi.org/10.21956/wellcomeopenres.28633.r150108>

© 2026 Dakroub F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Fatima Dakroub 

American University of Beirut, Beirut, Lebanon

The study presents a nowcasting approach and accompanying R package to address reporting delays in real-time disease surveillance data. The topic is relevant and valuable for public health practice. The study is generally well-structured, and the methods are clearly written. However, it may benefit from some refinement to improve readability. Below are some minor comments:

1. Consider replacing the numerical labels for each panel in Figure 2 with letters and referring to these in the legend to improve clarity and make it easier for the reader to follow.

2. References: The formatting of references is inconsistent, with some citations presented in superscript and others not. Please ensure a uniform citation style is used throughout the manuscript.

Also consider replacing “three models in Mellor et al.” with “three models reported by Mellor et al.” Regarding this sentence, please add the author’s name et al.: “As demonstrated by 5, such approaches can propagate errors and may not fully account for interdependencies between these processes.”

“GAM implementation in 2,”: you can delete “in”

3. “Each case record includes a specimen date (when the sample was collected) and a report date (when the test result was ingested into the surveillance system).”

The term “ingested” may be unclear. Consider replacing it with a more standard phrasing, such as “when the test result was recorded in the surveillance system” to improve clarity for the reader.

4. “The difference between these dates forms the reporting delay, which typically peaks at 1–2 days with most reports occurring within 7 days.”

The description of the reporting delay may be unclear, as stating that it “peaks at 1–2 days” while “most reports occur within 7 days” could be misinterpreted as contradictory. Consider rephrasing by first stating that the majority of cases are reported within 7 days, followed by clarifying that the most common delay is 1–2 days.

Also replace 7 with “seven” (numbers below 10 should be written in text form)

5. Discussion:

- “Lastly, by providing a standardised R package with comprehensive documentation (refer to reference no. 1)”.

There is no need to add the link, as it was already referenced in the materials and methods

- “Relatedly, co-creation of documentation with public health practitioners”

The phrase is not clear; if the authors mean that public health practitioners should be involved in the documentation process alongside researchers, please re-write the phrase more clearly

References

1. EpiNowcast Project Team: Baseline Nowcasting Tutorials and Function Documentation. Available at: <https://baselinenowcast.epinowcast.org/index.html>. –.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Public health, active surveillance, infectious diseases

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 21 January 2026

<https://doi.org/10.21956/wellcomeopenres.27585.r142120>

© 2026 Alvarado S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Sandra Flores Alvarado 

¹ Programa de Bioestadística, Universidad de Chile Escuela de Salud Publica (Ringgold ID: 193077), Independencia, Región Metropolitana, Chile

² Programa de Bioestadística, Universidad de Chile Escuela de Salud Publica (Ringgold ID: 193077), Independencia, Región Metropolitana, Chile

The article introduces `baselinenowcast`, an R package designed to provide a standardized, low-complexity baseline for nowcasting epidemiological count data. This tool is highly relevant for public health surveillance, particularly in emergency contexts where rapid implementation and transparency are paramount to inform policy decisions. The package successfully implements a version of the chain-ladder method and provides uncertainty quantification via a negative binomial model. While the software represents a valuable contribution to the field, there are certain methodological and structural aspects that could be further clarified. The following suggestions are intended to help strengthen the manuscript and its practical application for public health users.

A central point for clarification is the deterministic nature of the initial estimation step. The current approach uses the chain-ladder method to calculate a point estimate of the delay distribution, which is then used as a fixed baseline for the negative binomial model. This assumes that reporting delay proportions remain relatively constant, which may not fully capture the high variability inherent in real-time reporting during an active outbreak. Integrating this stochasticity more directly into the delay estimation process would likely provide a more realistic representation of uncertainty.

The validation using German COVID-19 data relies on 7-day moving averages rather than raw daily counts. The use of smoothed data removes the daily noise and weekend effects that a nowcasting tool is specifically intended to resolve. Because the data has been pre-processed, the estimated overdispersion parameters and the resulting prediction intervals may reflect the effects of the smoothing rather than the model's inherent ability to handle raw surveillance data. Testing the package with raw daily reports would provide a more robust evaluation of its performance in real-world settings.

Regarding the comparative analysis, the manuscript uses a revised version of the KIT simple nowcast instead of the official gold standard. Given that several co-authors are also linked to the original KIT model, it is important to maintain a clear distinction between these two projects. The KIT model's implementation issues should ideally be addressed within its own framework, and this paper should focus on an independent validation of `baselinenowcast` against established, official benchmarks to ensure the objectivity and reproducibility of the results.

As a software tool article, the manuscript would be more effective if it focused less on the narrative history of the project and more on its technical implementation. Currently, the text lacks specific examples of the package's architecture, such as function definitions, primary arguments, or code snippets. The claims regarding usability, low computational costs, and minimal expertise would be better supported if the main text provided technical evidence or a clear walkthrough of the functions necessary for a user to implement the tool during an emergency.

Finally, the visual presentation of Figure 2 could be improved to better guide the reader through

the methodology, as the current layout and curve rendering appear unpolished and lack a clear reading order. Additionally, it would be helpful to define how the model handles specific data edge cases, such as instances where a report occurs before the recorded symptom onset date (e.g., asymptomatic cases). Clarifying whether to use the earliest available date or another criterion, along with refining the graphical rendering, would significantly strengthen the manuscript for indexing.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biostatistics, epidemiological surveillance, epidemiology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 02 Feb 2026

Kaitlyn Johnson

Thank you for your helpful feedback. Please find our responses to each point below. A central point for clarification is the deterministic nature of the initial estimation step. The current approach uses the chain-ladder method to calculate a point estimate of the delay distribution, which is then used as a fixed baseline for the negative binomial model. This assumes that reporting delay proportions remain relatively constant, which may not fully capture the high variability inherent in real-time reporting during an active outbreak. Integrating this stochasticity more directly into the delay estimation process would likely provide a more realistic representation of uncertainty.

>Thanks for this suggestion, we agree that there is an important distinction to be made between our approach and alternative approaches based on bootstrapping or errors

obtained from combining many different estimated delays. In our methodology, we chose to produce a point nowcast and then estimate a parametric observation model (by default a negative binomial) for each nowcast horizon. However, this is not the only way to do this, for example, some of us at Massachusetts Department of Health developed a similar method which used directly the 95th percentiles on the observed past nowcast errors to compute the current 95% confidence intervals on their nowcast. We have added a point on this in the discussion, stating: "In this analysis we only considered negative binomial observation models applied to the target dataset, which in this case were always count data. Other methods for uncertainty estimation could have been explored, such as using bootstrapping or computing percentiles directly from past nowcast errors in a non-parametric form, or exploring other parametric observation models. The modular design of the package enables exploration of alternative error models. Other valid error models can be readily explored, given the modular design of our baseline package." We also clarified how the uncertainty is estimated in the methods in the uncertainty estimation description. We agree that there still might be additional high variability inherent in real-time reporting during an outbreak. One of the advantages of our approach is its modularity– so while the user can estimate the negative binomial (or any observation model of their choice) using past retrospective nowcast errors as described above, they don't have to – they could instead estimate them from another dataset/outbreak or add additional dispersion to the estimates if they believe them to be not fully representative of the expected variability.

The validation using German COVID-19 data relies on 7-day moving averages rather than raw daily counts. The use of smoothed data removes the daily noise and weekend effects that a nowcasting tool is specifically intended to resolve. Because the data has been pre-processed, the estimated overdispersion parameters and the resulting prediction intervals may reflect the effects of the smoothing rather than the model's inherent ability to handle raw surveillance data. Testing the package with raw daily reports would provide a more robust evaluation of its performance in real-world settings.

>Thanks for bringing this up, we used 7-day sums because this was the epidemiologically relevant surveillance statistic used in Germany and the target for the Nowcast challenge set up and described in Wolfram et al. In our method, we are nowcasting using the individual level data and then summarising that, with the observation model based on the 7-day sum. This does mean that our observations correlated, and in this case, it was shown by Wolfram et al that estimating the observation model on the 7-day sum directly improves performance over estimating for daily data and combining in post-processing. We clarify this choice in the methods:

"Because the 7-day rolling sums was the epidemiologically relevant surveillance statistic and the one used by the German Nowcast Hub, we generated nowcasts of the 7-day rolling sum of COVID-19 hospitalisation incidence each day for national-level aggregated data and age-stratified data." And add to the Discussion a point about the limitations inherent in the choice of using 7-day rolling sums instead of daily level data: "Additionally, for the German COVID-19 data we used the 7-day rolling sums because this was what was epidemiologically relevant and used in the Nowcast Hub, however, the use of this target meant that the daily noise and weekday effects were removed, preventing us from evaluating the methods ability to capture these effects. " Because nowcasts were produced each day and the input, pre-processed data was on a daily scale, we still were able to evaluate the nowcasts ability to handle weekday effects in reporting – and we did in fact find that allowing for weekday-

specific delay distributions improved nowcasts compared to not adjusting for the weekday when computing daily delay distributions (Fig 4B).

Regarding the comparative analysis, the manuscript uses a revised version of the KIT simple nowcast instead of the official gold standard. Given that several co-authors are also linked to the original KIT model, it is important to maintain a clear distinction between these two projects. The KIT model's implementation issues should ideally be addressed within its own framework, and this paper should focus on an independent validation of baselinenowcast against established, official benchmarks to ensure the objectivity and reproducibility of the results.

>Thanks for flagging this, we agree that there is a need for benchmarks to compare against for nowcasting model development, and the lack thereof is a primary motivation of this analysis. The KIT model was found to be a reasonably well-performing baseline in Wolfram et al, which is why we chose that model to develop into a software package which allows tuning for different settings. In an ideal world, we would have performed our validation against only what was submitted to the Hub in real-time. However, because of the implementation bug, which was a result we found as part of this study, we would not have been able to validate that our implementation was working correctly if we had not fixed the bug and made the comparison. To provide a thorough comparison to official results and ensure our method was working as we expected, we compare to the real-time Hub nowcasts in the SI, and discuss the finding of the bug and our resulting validation, in the main text (Fig 3 main text, and against the original version Fig. S1 & S2).

As a software tool article, the manuscript would be more effective if it focused less on the narrative history of the project and more on its technical implementation. Currently, the text lacks specific examples of the package's architecture, such as function definitions, primary arguments, or code snippets.

Thanks for this feedback, we thought that it was better to use this article as an introduction to the project and a validation of its implementation. We think that it is better to demonstrate package code in the package documentation where it can be kept up to date, whereas in a paper which is generally a static document. We agree that more signpost to this is needed and so have added the following signposting to package documentation in the Methods under software implementation: "Worked examples applied to common datasets, as well as documentation of package functions and arguments, are provided as part of the package documentation at <https://baselinenowcast.epinowcast.org/index.html>" And in the Discussion:

"Lastly, by providing a standardised R package with comprehensive documentation (see <https://baselinenowcast.epinowcast.org/index.html> for tutorials and function documentation)," The claims regarding usability, low computational costs, and minimal expertise would be better supported if the main text provided technical evidence or a clear walkthrough of the functions necessary for a user to implement the tool during an emergency. We agree that these claims should be justified. We have edited the concluding paragraph of the discussion to reflect the usability, and have removed any reference to lower computational costs (while we do think this is true, we didn't formally evaluate its speed in this work). "Our method, developed with public health practitioners, is robust, interpretable, and relatively straightforward, evidenced by our public health co-authors who

were able to implement the package in their systems within a few days unsupported. “

Finally, the visual presentation of Figure 2 could be improved to better guide the reader through the methodology, as the current layout and curve rendering appear unpolished and lack a clear reading order. Thanks. We have reorganized Figure 2 to make it more logical and have revised the contents to clean up the presentation. Additionally, it would be helpful to define how the model handles specific data edge cases, such as instances where a report occurs before the recorded symptom onset date (e.g., asymptomatic cases). Clarifying whether to use the earliest available date or another criterion, along with refining the graphical rendering, would significantly strengthen the manuscript for indexing.

>Thanks for this feedback. One of the problems is that there are many such edge cases and in the interest of clarifying for readers it is difficult for us to give a complete overview of what can be achieved with the flexibility of the method. We have added a section on dealing with data edge cases in the Supplement (SI Text. Dealing with edge cases in real-world data) and have added the following reference to it in the main text: “There are a number of ways in which dealing with real-world data can create edge cases and complexities, such as for example when the “reference date” can refer to different primary events, or when reporting delays are registered as being negative. We discuss these and ways to handle them in this framework in the SI Text. Dealing with edge cases in real world data.” This section points to package documentation with further examples on how to deal with some of the edge cases like the one mentioned here.

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 January 2026

<https://doi.org/10.21956/wellcomeopenres.27585.r140119>

© 2026 Guzzetta G. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Giorgio Guzzetta

¹ Fondazione Bruno Kessler, Trento, Italy

² Fondazione Bruno Kessler, Trento, Italy

This manuscript presents an R package implementing a previously proposed multiplicative nowcasting algorithm, explores alternative algorithmic specifications, and evaluates performance using COVID-19 hospitalization data from Germany (late 2021–early 2022) and norovirus surveillance data from the UK (late 2023–early 2024). While the topic is timely and the authors’ intent is commendable, the manuscript in its current form does not fully achieve its two stated aims: providing a practical tool for public health practice and serving as a standardized benchmark for methodological research. In addition, assessment of the methods, results, and overall contribution is hindered by substantial issues related to clarity, excessive verbosity, and limited

synthesis of insights. At several points, the paper reads more like a technical report, leaving much of the burden of identifying the main messages and implications to the reader.

Major issues

1. Clarity of methods

- a. The Methods section is excessively verbose and difficult to follow, as it mixes statistical assumptions, implementation heuristics, software design choices, and sensitivity analyses without a clear hierarchy. For example, much of page 7 falls between a package vignette and a sensitivity analysis description but does not fully succeed in either role. Several expressions (e.g., “data ingestion assumptions,” “pipe-friendly interface”) are not clearly defined. For a paper aiming to promote accessibility and practical usability, the exposition is unexpectedly hard to navigate. A clearer separation between the statistical model, default methodological choices, and software implementation details would substantially improve readability.
 - b. The authors should more clearly explain the rationale behind points (a) and (b) of Step 1 (Delay estimation, page 5). In particular, there is concern that this procedure may propagate errors over longer delays. This issue could be exacerbated when the reporting delay distribution is not monotonically decreasing, as is often the case in practice (especially for delays between symptom onset or hospitalization and reporting). The nowcasting algorithm appears closely related to that proposed by Lawless (1994), yet this reference is not cited.
 - c. It is not clear which data are used to estimate the delay distribution and which are used for nowcasting. For instance, on page 8 the German COVID-19 dataset is described as containing approximately 160 reference dates, with 60 used for delay estimation and 60 for “retrospective nowcasts” (another undefined term). However, Figure 3c appears to report WIS values over the full data domain, creating confusion about the evaluation setup. This applies equally to the norovirus use case and to the age-specific nowcasting applications.
 - d. Figure 3e shows time- and age-varying reporting delays, but it is unclear why (or how) this variability is not incorporated into the nowcasting procedure. Additionally, the absolute WIS values shown in Figure 3c provide limited insight into nowcast quality and do not align clearly with the rWIS values discussed in the text.
 - e. In both applications, nowcasts are validated using X-day-ahead rolling evaluation datasets. It is unclear why the authors did not instead evaluate predictions against consolidated values from the final version of the reporting dataset.
- Overall, the lack of clarity in methodology and investigation pipeline makes it impossible to determine the robustness of the results.

2. Figure quality

Overall figure quality is substandard. Some plots in Figure 2 appear hand-drawn, while Figures 3, 4, and 5 contain an excessive number of panels, with color schemes and legends that are difficult to map to the corresponding results. For example, in Figure 5, it is hard to assess whether any model provides qualitatively accurate nowcasts and which legend applies to panels on the right.

3. Information overload

The manuscript devotes substantial space to issues of limited scientific relevance, such as the identification of implementation bugs and baseline consistency checks (e.g., Figure 3b,d). Similarly, Figures 4 and 5 and their accompanying text present a large volume of detailed results that are difficult to contextualize, especially given the lack of methodological clarity, and to summarize in its epidemiological meaning. Differences across results of model variants appear minimal, yet they are discussed at considerable length.

4. Claims and positioning

In the introduction, the work positions itself as both a practical tool for public health practice and a standardized benchmark for methodological research. However, as also acknowledged in the Discussion, the paper does not succeed in achieving either objective.

In summary, the work appears to be a reasonable validation exercise of an existing and previously validated model using COVID-19 data (conditional on the robustness of the methods, which would need to be clarified more clearly). The extension to norovirus data does not appear competitive relative to other existing models. The main added value of the work seems to lie in providing additional flexibility in algorithmic choices, though these are currently described unclearly, and in wrapping the method into an R package, whose usability and practical value remain to be demonstrated.

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease modelling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 02 Feb 2026

Kaitlyn Johnson

We thank the reviewer for the feedback. We have reorganised the methods section, clarified how training data is used for estimation, and revised the schematic figure to improve its readability. Please find our detailed responses to each review point below.

a. The Methods section is excessively verbose and difficult to follow, as it mixes statistical assumptions, implementation heuristics, software design choices, and sensitivity analyses without a clear hierarchy. For example, much of page 7 falls between a package vignette and a sensitivity analysis description but does not fully succeed in either role. Several expressions (e.g., “data ingestion assumptions,” “pipe-friendly interface”) are not clearly defined. For a paper aiming to promote accessibility and practical usability, the exposition is unexpectedly hard to navigate. A clearer separation between the statistical model, default methodological choices, and software implementation details would substantially improve readability.

> Thanks for this feedback. We agree that the initial methods section interspersed software defaults with methods in a way that wasn't clear. We revised the methods section to clearly delineate between the high-level description of the model and the default method specifications in the package, options for model specifications, and the details of the software implementation. We also removed the term “data ingestion assumptions”, using instead language around the amount of training data used for delay and uncertainty estimation, and defined the term pipe-friendly interface. We think this should help the reader delineate between the core method used in the package, the defaults, and the features that are tunable, with the description of the sensitivity analyses (validation, comparison of method specifications, and comparison to external methods) going within each separate case study.

b. The authors should more clearly explain the rationale behind points (a) and (b) of Step 1 (Delay estimation, page 5). In particular, there is concern that this procedure may propagate errors over longer delays. This issue could be exacerbated when the reporting delay distribution is not monotonically decreasing, as is often the case in practice (especially for delays between symptom onset or hospitalization and reporting).

> We thank the reviewer for raising this concern. We have reworded the description of the delay estimation for clarity (see first bullet point of Step 2). We have tested a range of delays both in the paper and in the package development and did not observe error propagation over longer delays. We agree with the reviewer that there is potential for non-monotonically decreasing CDFs and we highlight this as an area of future work in the Discussion paragraph (second to last paragraph) when we refer to downward revisions. We don't anticipate that the extension to handle downward revisions will propagate errors. We have also added a section to the SI Text. Dealing with edge cases in real world data which discusses additional real-world data issues, since this is not a focus of the main text. The nowcasting algorithm appears closely related to that proposed by Lawless (1994), yet this reference is not cited. We appreciate the reviewers suggestion to cite Lawless (1994) and have added it when we describe the nowcasting algorithm upon which the method is based upon. We have also added a citation to Bizzotto et al to reflect that it is an extension of the Lawless method.

c. It is not clear which data are used to estimate the delay distribution and which are used for nowcasting. For instance, on page 8 the German COVID-19 dataset is described as containing approximately 160 reference dates, with 60 used for delay estimation and 60 for

“retrospective nowcasts” (another undefined term). However, Figure 3c appears to report WIS values over the full data domain, creating confusion about the evaluation setup. This applies equally to the norovirus use case and to the age-specific nowcasting applications.

>Thanks for pointing out this area lacking clarity. The total reference times used for uncertainty estimation overlaps with those used for delay estimation because the uncertainty is estimated by producing a retrospective nowcast for multiple retrospective nowcast times. At each of those times, the delay is re-estimated, using the same amount of training data specified for delay estimation. The uncertainty estimation step now reads: “We generate retrospective point nowcasts (light shaded bars and blue lines in Figure 2, step 3) by repeating steps 1 and 2 for a number of past, retrospective nowcast times, M , using N reference times for delay estimation at each retrospective nowcast time. This requires $N + M$ total reference times, which means that the total reference times used for uncertainty estimation overlaps with those used for delay estimation, as each retrospective nowcast requires re-estimating a delay using N reference times.” There is a WIS value reported every day in Figure 3C because we are fitting the model and nowcasting for each day i.e time aware cross validation. The value at each time point summarises the score for the nowcast produced on that day. In practice, this means that the delay estimate is “rolling”, as the delay is always based on the latest data available. We have clarified this in the methods section on the German Nowcast Hub validation which now states: “ We retrained the model and produced nowcasts each day from 22 November 2021 to 29 April 2022, with preliminary data available from 1 July 2021 used for training the earlier models. Because we retrained the model each time we produced a nowcast, the delay estimate is always based on the latest data available as of the time the nowcast is made.”

d. Figure 3e shows time- and age-varying reporting delays, but it is unclear why (or how) this variability is not incorporated into the nowcasting procedure. Additionally, the absolute WIS values shown in Figure 3c provide limited insight into nowcast quality and do not align clearly with the rWIS values discussed in the text.

> Thanks for pointing this out. As described above, the delay is estimated at each nowcast date based on the latest data, and we edited the text to clarify this. A larger training volume will lead to less variable (more slowly changing) delay estimates, and a smaller training volume will lead to more variable delay estimates. We appreciate the point about Figure 3C, but we felt that it was important to show the how the absolute WIS changed over time.

e. In both applications, nowcasts are validated using X-day-ahead rolling evaluation datasets. It is unclear why the authors did not instead evaluate predictions against consolidated values from the final version of the reporting dataset.

> Thanks for pointing this out. We followed the recommendations from Wolfram et al to use a X-day ahead rolling evaluation dataset which were based on the analysis they performed in Section 3.7 of their manuscript, which found that using a constant lag between the time the nowcast was made and the evaluation dataset more fairly evaluated nowcast performance across time in the case where very long delays can occur. We have clarified this point in the methods, stating: “We then quantitatively evaluated both sets of nowcasts using a rolling evaluation dataset

of 40 days after each nowcast date, restricted to a maximum delay of 40 days for each reference date, consistent with the analysis performed in section 3.7 of Wolfram et al. [3] and with other nowcast evaluations [4,7,22]. This means that for each nowcast date, we used a different evaluation dataset corresponding to the data available 40 days after that nowcast date, which ensures that recent and older nowcasts are evaluated against a dataset with the same lag, reducing any bias that could be introduced from very long delays, which would systematically bias towards poorer performance in nowcasts made further in the past. " Overall, the lack of clarity in methodology and investigation pipeline makes it impossible to determine the robustness of the results.

> We hope that our changes based on the substantive reviewer comments have made the manuscript easier to follow.

2. Figure quality

Overall figure quality is substandard. Some plots in Figure 2 appear hand-drawn, while Figures 3, 4, and 5 contain an excessive number of panels, with color schemes and legends that are difficult to map to the corresponding results. For example, in Figure 5, it is hard to assess whether any model provides qualitatively accurate nowcasts and which legend applies to panels on the right.

>Thanks for this feedback though we don't agree with the use of "substandard" We agree that Figure 2 could be improved and have reorganized it to more logically guide the reader from steps 1 through 5 and replaced the components of it that were hand-drawn in powerpoint. ... We are concerned that many readers will not also read the SI and so think that the panels in 3, 4, and 5 are important to include in the main text in order to communicate the detail of our evaluation which the reviewer has highlighted is key to communicate. We agree that the legends in Figure 5 are difficult to distinguish, we have altered the color palette of the weekdays to make them more visually distinct from the different model variations.

3. Information overload

The manuscript devotes substantial space to issues of limited scientific relevance, such as the identification of implementation bugs and baseline consistency checks (e.g., Figure 3b,d). Similarly, Figures 4 and 5 and their accompanying text present a large volume of detailed results that are difficult to contextualize, especially given the lack of methodological clarity, and to summarize in its epidemiological meaning. Differences across results of model variants appear minimal, yet they are discussed at considerable length.

> We appreciate this feedback. We hope that other changes based on earlier comments have made the results easier to digest. We agree that our results are discussed at length which we feel is a strength of the manuscript and important for evaluations of this kind in order for users to be able to trust the method. In light of this feedback, we have gone through and reviewed the results and trimmed down instances where we felt the language was overly verbose.

4. Claims and positioning

In the introduction, the work positions itself as both a practical tool for public health practice and a standardized benchmark for methodological research. However, as also

acknowledged in the Discussion, the paper does not succeed in achieving either objective.

> We believe that this work does achieve the stated aims of being both a practical tool for public health practice. We worked with two different public health agencies in both the development of the tool and the analysis presented in this paper. We have revised the discussion to reflect this.

In summary, the work appears to be a reasonable validation exercise of an existing and previously validated model using COVID-19 data (conditional on the robustness of the methods, which would need to be clarified more clearly). The extension to norovirus data does not appear competitive relative to other existing models. The main added value of the work seems to lie in providing additional flexibility in algorithmic choices, though these are currently described unclearly, and in wrapping the method into an R package, whose usability and practical value remain to be demonstrated.

>Thanks for this feedback. Our goal in this work was to first validate that our implementation of the baseline method used in the German Nowcast Hub (Wolffram et al), second assess the performance of the different model specifications applied to German COVID-19 data, and compare to other methods and previously used baselines applied to norovirus data. We agree that the process of improving the usability and practical feature set of the R package will be ongoing, and one which we plan to continue to develop after receiving feedback from users.

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 January 2026

<https://doi.org/10.21956/wellcomeopenres.27585.r139005>

© 2026 Günther F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Felix Günther 

¹ Robert Koch Institute, Berlin, Germany

² Robert Koch Institute, Berlin, Germany

Johnson et al. present a nowcasting model that allows to anticipate future changes to current real-time disease surveillance data, e.g. due to reporting delays or data revision, based on observed patterns of changes in the past. They also provide an accompanying R package, containing an implementation of the method. It allows for easy application of the method and provides flexibility in adapting it to requirements for the respective use case. The model is elegant and relatively simple, it focuses on representing the observed changes of the past in a simple way, combined with an assumption-light, empirical quantification of prediction uncertainty derived from applying the method to past time points.

Besides introducing the method and software, the authors illustrate application of the model based on two use cases, discuss different specifications and modelling choices and demonstrate a solid performance of the method by retrospective evaluation and comparison to alternative, more complex nowcasting methods.

The work is motivated by the observation that nowcasting, i.e., the adjustment of disease surveillance data for reporting delays and corresponding changes in available data over time, is an important task for real-time infectious disease surveillance. In the past, several methods have been proposed for this, but they are not yet in regular use in practice. Furthermore, it is often unclear how the performance of different models differs. The new method should therefore be easy for users to apply, perform reliably, and at the same time provide a basis for comparison with the results of more complex methods.

I enjoyed reading the article and agree with the authors that this is an important piece of work. I see the potential of the proposed method — in conjunction with the published software — to increase the applicability of nowcasting even for users who are not technical experts.

I have two more detailed comments/questions regarding the method and its presentation followed by some minor points and typos:

1) The methods section in the main text of the manuscript states that training data is splitted into two non-overlapping subsets for estimating the delay distribution and uncertainty parameters (50% of the reference times, each). This seems to be appealing as it assures that the uncertainty quantification of the point predictions takes into account an "out of sample" performance/variability when comparing the nowcast point predictions to actual data. I, however, do not understand completely how this is done (I think it's not explicitly mentioned in the supplement again), and I am also wondering, whether it is really necessary. Have you also tried estimating the nowcast uncertainty and delay distribution based on the same set of reference dates and did it reduce the overall performance compared to splitting the training data? This would reduce the amount of historical data/data snapshots required for nowcasting, which could be useful in situations involving a new outbreak or data stream without much historical data. Overall, I would appreciate some more clarification of this aspect of the method.

2) The proposed nowcasting method is described in detail in the supplementary material and explained in an understandable way with the necessary technical depth. However, when I first read the main section on the methods, I had some difficulty understanding the approach, as I felt the conceptual ideas and methodological details of the default settings were mixed up in a somewhat confusing way (e.g., I did not directly understand the necessity for the imputation of missing cells in the reporting triangle, the difference compared to the "point nowcast matrix," and the actual nowcasting based on the currently available cumulative case numbers per reference date). I wonder, whether this section could be made more accessible by first sketching out the general idea of the method (point nowcast + uncertainty quantification based on a comparison of past point nowcasts and final data). This could then be followed by a subsection carefully describing the default choices for each step, and then the already existing subsection on the available options for model specification. I guess, such a description could help to make the basic idea of the method more accessible, also to potential users without a background in the area of nowcasting methods.

I also have a few minor points:

- The term "data ingestion settings" was somewhat unclear to me, might profit from clarification
- Results: German COVID-19: The sentence "Our method had 55% coverage at 50% prediction intervals and 96% coverage at 95% prediction intervals, indicating it had slightly wider prediction intervals than observed in the data" does not exactly makes sense to me, as no prediction intervals are observed in the data. I guess a rephrasing towards the idea that empirical coverage frequencies were higher than desired would be more appropriate.
- Discussion: "for example they can include upwards revisions", should this mean downward revisions?

Typos:

- page 4: "black lines" instead of "black line"
- denominator in Formula 5 in the supplement (first subscript t should probably be removed)
- Supplement: Description of KIT simple nowcast implementation issue and revised nowcasts: "beign" instead of "being"
- Supplement: "less stably" instead of "less stable"

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious disease modelling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 02 Feb 2026

Kaitlyn Johnson

Thank you for your helpful feedback.

We have reorganised the main text methods and added additional explanations of how the training data is used for delay and uncertainty estimation in the supplement. Please find our responses to each point below.

1) The methods section in the main text of the manuscript states that training data is splitted into two non-overlapping subsets for estimating the delay distribution and uncertainty parameters (50% of the reference times, each). This seems to be appealing as it assures that the uncertainty quantification of the point predictions takes into account an "out of sample" performance/variability when comparing the nowcast point predictions to actual data. I, however, do not understand completely how this is done (I think it's not explicitly mentioned in the supplement again), and I am also wondering, whether it is really necessary. Have you also tried estimating the nowcast uncertainty and delay distribution based on the same set of reference dates and did it reduce the overall performance compared to splitting the training data? This would reduce the amount of historical data/data snapshots required for nowcasting, which could be useful in situations involving a new outbreak or data stream without much historical data. Overall, I would appreciate some more clarification of this aspect of the method.

> Thanks for highlighting that this part isn't clear. The training data is not strictly speaking split into two non-overlapping subsets for delay estimation and uncertainty estimation. The total reference times used for uncertainty estimation overlaps with those used for delay estimation because the uncertainty is estimated by producing a retrospective nowcast for multiple retrospective nowcast times. At each of those times, the delay is re-estimated, using the same amount of training data specified for delay estimation. The uncertainty estimation step now reads: "We generate retrospective point nowcasts (light shaded bars and blue lines in Figure 2, step 3) by repeating steps 1 and 2 for a number of past, retrospective nowcast times, M , using N reference times for delay estimation at each retrospective nowcast time. This requires $N + M$ total reference times, which means that the total reference times used for uncertainty estimation overlaps with those used for delay estimation, as each retrospective nowcast requires re-estimating a delay using N reference times. We clarify this in the supplement, revising the section on uncertainty quantification to read:

"To estimate the uncertainty in the nowcasts, we use the nowcast errors from M past nowcasting time points, with each past nowcast produced using N reference times for delay estimation. This means that the total set of reference times used for uncertainty estimation is $N + M$, which partially overlap with the set of reference times N used for delay estimation."

2) The proposed nowcasting method is described in detail in the supplementary material and explained in an understandable way with the necessary technical depth. However, when I first read the main section on the methods, I had some difficulty understanding the approach, as I felt the conceptual ideas and methodological details of the default settings were mixed up in a somewhat confusing way (e.g., I did not directly understand the necessity for the imputation of missing cells in the reporting triangle, the difference compared to the "point nowcast matrix," and the actual nowcasting based on the currently available cumulative case numbers per reference date). I wonder, whether this section could be made more accessible by first sketching out the general idea of the method (point

nowcast + uncertainty quantification based on a comparison of past point nowcasts and final data). This could then be followed by a subsection carefully describing the default choices for each step, and then the already existing subsection on the available options for model specification. I guess, such a description could help to make the basic idea of the method more accessible, also to potential users without a background in the area of nowcasting methods.

> We agree with the reviewer that this was confusing. We have taken the reviewers suggestions and have reorganized the methods so that we provide an overview of the method, a step-by-step description of the method conceptually, a section with the default methodological choices, a section for available options for model specification, and finally a section for the software implementation.

- The term "data ingestion settings" was somewhat unclear to me, might profit from clarification

> The term is meant to refer to the amount of input training data the model uses and for what purposes (i.e. delay and uncertainty estimation). We have removed it, instead describing directly the choices that are made regarding the training data used for delay and uncertainty estimation by default and the available specification options which allow these to be modified.

Results: German COVID-19: The sentence "Our method had 55% coverage at 50% prediction intervals and 96% coverage at 95% prediction intervals, indicating it had slightly wider prediction intervals than observed in the data" does not exactly makes sense to me, as no prediction intervals are observed in the data. I guess a rephrasing towards the idea that empirical coverage frequencies were higher than desired would be more appropriate.

> Thanks for catching this, this was a mistake in our description of interval coverage. We edited the sentence to state that "Our method had 55% coverage at 50% prediction intervals and 96% coverage at 95% prediction intervals, indicating that a greater proportion of observed data points fell within the prediction intervals than would be theoretically expected in a perfectly well-calibrated model (SI Figure 6,7)

Discussion: "for example they can include upwards revisions", should this mean downward revisions?

> Thanks for pointing out this mistake. We have edited this accordingly.

We also addressed each of the typos pointed out by the reviewer and greatly appreciate the thorough attention to detail given in this review.

Competing Interests: No competing interests were disclosed.