On the assessment of the ability of measurements, nowcasts, and forecasts to track changes

Jonas Rieger[1*], Bolin Liu[1], Bernd Saugel[2], Oliver Grothe[1]

[1]Institute for Operations Research, Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, Karlsruhe, 76131, Germany.

[2]Department of Anesthesiology, Center of Anesthesiology and Intensive Care Medicine, University Medical Center Hamburg-Eppendorf, Martinistrasse 52, Hamburg, 20246, Germany.

*Corresponding author(s). E-mail(s): jonas.rieger@kit.edu;

Abstract

Background: Measurements, nowcasts, or forecasts ideally should correctly reflect changes in the values of interest. In this article, we focus on how to assess the ability of measurements, nowcasts, or forecasts to correctly predict the direction of changes in values – which we refer to as the ability to track changes.
Methods: We reviewed and developed visual techniques and quantitative measures to assess ATC. Extensions for noisy data and estimation uncertainty were implemented using bootstrap confidence intervals and exclusion areas.
Results: We exemplarily illustrate the proposed methods to assess the ability to track changes for nowcasting during the COVID-19 pandemic, patient admissions to an emergency department, and non-invasive blood pressure measurements. The proposed methods effectively evaluated ATC across different applications.
Conclusions: The developed ATC assessment methods offer a comprehensive toolkit for evaluating the ATC of measurements, nowcasts, and forecasts. These techniques provide valuable insights into model performance, complementing traditional accuracy measures and enabling more informed decision-making in various fields, including public health, healthcare management, and medical diagnostics.

1 Introduction

Measurements, nowcasts, or forecasts ideally should correctly reflect changes in the values of interest. It is thus important to meticulously assess the ability of measurements, nowcasts, or forecasts to correctly predict the direction of changes in values – which we refer to as the ability to track changes (ATC). Although measurements, nowcasts, and forecasts fundamentally differ as they either measure or predict a value, similar methods can be used to assess their ATC.

Forecasting methods predict the future based on historical data, patterns, or exogenous factors. A forecast is computed based on the current value of interest and an estimate of its future development. In medicine and healthcare, forecasting – for example – is used to predict patient volumes in emergency departments [28, 36] or the demand of emergency medical services [22].

Methodologically evolved from forecasting [6], nowcasting methods focus on predictions for the present, the immediate future, and the recent past [1, 48]. Nowcasting methods use high-frequency indicators or preliminary measurements related to the value of interest and focus on updating predictions using currently available information [7]. Nowcasting, for example, can assess the current situation during an ongoing epidemic, considering the main pathogenic, epidemiological, clinical, and socio-behavioral factors [49] or provide daily numbers of COVID-19 cases for events that have occurred but have not yet been reported [20, 47].

Measurements aim to obtain accurate and precise values of a measurable quantity (measurand; [40, 41]). Repeated measurements can be used to track changes in a value over time. When introducing new measurement methods, they are evaluated against current reference methods, often called the “gold standard”, by simultaneously measuring the same quantity with the new method and the reference method – often in various individuals or different clinical settings.

In forecasting and nowcasting the evaluation of performance or goodness is usually based on statistical methods quantifying the accuracy such as the root mean square error (RMSE), probabilistic scoring rules, and calibration measures [16, 20, 47], see also Steyerberg et al. [42] for an overview in the medical context together with less known measures. However, although techniques like ROC curves for binary direction models or more general concordance measures [see, e.g., 18] are used, there is no general measure directly framed to assess the method’s ATC. In a forecasting competition on armed conflicts, the assessment of the ATC recently gained attention as Vesco et al. [45] proposed the novel targeted absolute deviation with direction augmentation (TADDA) score with an additive tracking-changes-component for evaluation. However, the score poses an unintuitive incentive to forecasters and is thus theoretically problematic [4]. When evaluating the performance or goodness of a measurement method like in classical method comparison studies, comparative statistics such as Bland-Altman analysis [3] and the percentage error [12] are commonly used. In this strand of literature, the best way to assess the ATC of measurement methods is a field of active ongoing research [10, 11, 38, 37, 24].

In this article, we focus on how to assess the ability of measurements, nowcasts, or forecasts to track changes. We formalize the concept of ATC and present visual techniques and quantitative measures to assess it – considering both noiseless data

Application	Predicted change computation
Measurement	$(x_t - x_{t-l})_{t=l}^T$
Nowcasting	$\mathbf{x}^{\Delta,l} = \begin{cases} (x_{t\vert t} - x_{t-l\vert t})_{t=l}^T, & \text{if } y_{t-l} \text{ is not known at time } t, \\ (x_{t\vert t} - y_{t-l})_{t=l}^T, & \text{otherwise.} \end{cases}$
Forecasting	$\mathbf{x}^{\Delta,l} = (x_{t\vert t-l} - y_{t-l})_{t=l}^T$

Table 1: Computation of the predicted change in the different applications. For nowcasting and forecasting, $x_{t|\tau}$ refer to values issued at τ with a target time t. For measurement, x_t denotes the test device measurement at time t.

and data with noise and small non-informative changes. We introduce the conditional ATC plot, a new graphical method for assessing the local ability, and review bootstrap methods for calculating confidence intervals. We extend the concept of assessment to probabilistic predictions. We exemplarily illustrate the proposed methods to assess the ATC for nowcasting during the COVID-19 pandemic, patient admissions to an emergency department, and non-invasive blood pressure measurements – and thus provide blueprints for future assessments. We discuss practical implementation and interpretation of the measures, thus providing the basis for communicating model limitations to forecasters, nowcasters, and public health officials. Ready-to-use code is available on https://github.com/jo-rie/aatc.

2 Assessment of the ability to track changes (ATC)

2.1 Computing changes and notation

We base the assessment of ATC on the measured/observed/true and the predicted changes in a value of interest over a time horizon l. The *true* change is straightforward to compute for all types of measurement, nowcast, or forecast. Let $\mathbf{y} = (y_t)_{t=0}^T$ denote the actual values for nowcasting or forecasting, or gold standard measurements up to time T. The sequence of changes is then given by the differences of values in $\mathbf{y}$ with horizon l, that is,

$$\mathbf{y}_t^{\Delta,l} = (y_t - y_{t-l}) \quad \text{for } t = l, \dots, T. \tag{1}$$

The definition of the *predicted* change depends on the context; Table 1 summarizes the notation for measurements, nowcasts, or forecasts and the computation of the predicted change. While the computation for measurements is straightforward and well-established, we develop the framework for nowcasts and forecasts in the following sections. For nowcasting, let $x_{t|\tau}$ denote the nowcast for time t computed with the knowledge of time τ. We call t the *target time* and τ the *issue time*. The predicted change is computed by

$$\mathbf{x}^{\Delta,l} = \begin{cases} (x_{t|t} - x_{t-l|t})_{t=l}^T & \text{if } y_{t-l} \text{ is not known at time } t, \\ (x_{t|t} - y_{t-l})_{t=l}^T & \text{otherwise.} \end{cases} \tag{2}$$

When computing the predicted change of a nowcast for a time t, we use the best knowledge available at that time t, and the true value might not be known yet. If the true value y_{t-l} is known at time t, the predicted change is computed by the difference between the nowcast and the true value, as y_{t-l} is also known by the nowcaster and incorporated into the nowcast. Through the computation in Equation (2), the predicted change can be computed with the knowledge of the nowcaster at time t.

The notation is similar for forecasting: Let $x_{t|\tau}$ denote the forecast for target time t and issue time τ. The predicted change is computed by

$$\mathbf{x}^{\Delta,l} = (x_{t|t-l} - y_{t-l})_{t=l}^{T} \tag{3}$$

with the same structure as in the nowcasting case and consistent indices with $\mathbf{y}^{\Delta,l}$. If the true value y_{t-l} is not known at time $t-l$, a similar modification can be made as in Equation (2).

The distinction between forecast and issue time is unnecessary in measurement analysis, as the measurement is typically available with a very short time lag. Thus, x_t denotes the test method measurement for time t. The computation

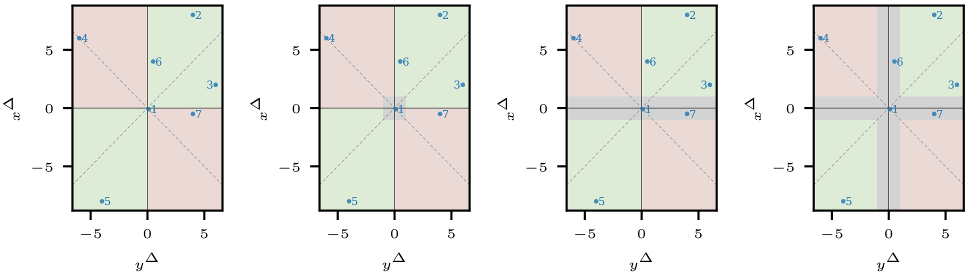
$$(x_t - x_{t-l})_{t=l}^{T} \tag{4}$$

yields the change by the test method. It is computed purely by the test method without the gold standard y_t to analyze whether the gold standard and test method changes are consistent. Accordingly, y_{t-l} is not used in the computation even if known at time t in contrast to forecasting and nowcasting.

In applications, data are often not available for all time steps, for example, due to technical problems or delays in data transfer (see the examples in Sections 3.1 and 3.2). We refer to time steps for which either measurement, nowcast, or forecast or true values are unavailable as missing values. Systematical missing values could lead to a biased assessment, and missing data should be inspected for any underlying patterns. If the missing values are not systematic, random, and occur scarcely, data pairs with missing values can be excluded from the data to calculate the measures [see 44, Section 1.3]. Note that in the case of measurement data, one missing value in the time series leads to two undefined differences in the change series; that is if x_t is missing, x_t^{Δ} and x_{t+l}^{Δ} are undefined; if an observation y_t is missing, y_t^{Δ} and y_{t+l}^{Δ} are undefined. The data pair is excluded even if the corresponding nowcast or forecast is available.

2.2 The four-quadrant plot

Formally, the assessment of ATC is the same for measurements, nowcasts, and forecasts, given the notation for the respective application of Section 2.1. In the following, we omit the horizon l for ease of notation; $\mathbf{x}^{\Delta}$ and $\mathbf{y}^{\Delta}$ refer to $\mathbf{x}^{\Delta,l}$ and $\mathbf{y}^{\Delta,l}$ for a common horizon l. The ATC is maximal if all predicted change directions are correct; that is, the sign of all elements of $\mathbf{x}^{\Delta}$ and $\mathbf{y}^{\Delta}$ coincide. Consequently, when assessing the ATC, we examine the statistical consistency of $\mathrm{sign}(\mathbf{x}^{\Delta})$ and $\mathrm{sign}(\mathbf{y}^{\Delta})$. A simple



(a) Basic four-quadrant plot.

(b) Four-quadrant plot with rectangular exclusion area. Point 1 is excluded.

(c) Four-quadrant plot with horizontal exclusion area. Points 1 and 7 are excluded.

(d) Four-quadrant plot with cross-shaped exclusion area. Points 1, 6 and 7 are excluded.

Figure 1: Illustrations of the four-quadrant plot with sample points and with and without exclusion areas. The rectangular exclusion area in Figure 1b excludes only points where both components are likely to be noise-driven, while the exclusion areas in Figures 1c and 1d exclude points where at least one component is noise-driven.

yet insightful method is the four-quadrant plot, which is well-established in measurement analysis and can be extended to nowcasts and forecasts [see, e.g., 32, 38]. In a four-quadrant plot, the occured changes and the predicted changes are plotted together, that is, (y_t^Δ, x_t^Δ) for $t = l, \ldots, T$. Thus, the x-axis of a four-quadrant plot shows the true value differences, whereas the y-axis displays the prediction data differences. Points in the green upper right and lower left quadrants reflect a correct change direction for the respective time step, whereas points in the remaining red quadrants show incorrectly predicted changes. Figure 1a displays a basic four-quadrant plot, and Figure 2a shows a four-quadrant graph for simulated data with $T = 1461$, for example, four years of daily data (for the data generation, see Appendix A.1).

The four-quadrant plot can be extended by including information on the time index in the point color to reveal effects over time. In Figure 2b, the point colors turn from blue to green for higher time indices t, that is, more recent values; (y_l^Δ, x_l^Δ) is blue and turns green until (y_T^Δ, x_T^Δ). However, four-quadrant plots become crowded for larger datasets, and sequential information on the differences is complex to assess thoroughly.

The four-quadrant plot is intuitive to interpret, and the magnitude and direction of change are shown simultaneously. Other visualization techniques, such as polar plots, lack the four-quadrant plot's clarity and intuition without adding more information on the ATC [38].

2.3 The ATC ratio and other measures

Analyzing the number of points in the green versus red quadrants is a standard approach in the ATC assessment of measurement data [9, 38], which we extend here to forecasts and nowcasts. With that, we estimate the probability of a correctly predicted change direction, $P(X^{\Delta}Y^{\Delta} > 0)$, where Y^{Δ} and X^{Δ} denote random variables for future incremental changes. Since $z_1 z_2 > 0$ imposes the same condition as $\text{sign}(z_1) = \text{sign}(z_2)$ $(z_1, z_2 \in \mathbb{R}\backslash\{0\})$, the standard estimator for $P(X^{\Delta}Y^{\Delta} > 0)$ is

$$\mu(\mathbf{x}^{\Delta}, \mathbf{y}^{\Delta}) := \frac{\sum_{t \in \mathcal{T}} \mathbb{1}\{x_t^{\Delta} y_t^{\Delta} > 0\}}{|T|}. \tag{5}$$

Here, the numerator counts the number of same-sign-changes, while the denominator is the number of considered pairs $(y_t^{\Delta}, x_t^{\Delta})$. Thus, μ is the proportion of concordant changes on all changes. We refer to this estimator as the *ATC ratio* of the prediction and set $\mathcal{T} = \{l, \ldots, T\}$. Visually, the measure computes the fraction of points in the upper right or lower left quadrant. Similar evaluations are used in other scientific areas, for example, with contingency tables in dichotomous forecasting or with confusion matrices in classification analysis (see, e.g., the introductions in [25], Ch. 4, and [27], Ch. 3). Many other measures can be adapted from those fields to deepen the analysis. Two simple measures that focus on a positive or negative predicted change are the positive and negative ATC ratios μ^+ and μ^-, respectively. They are defined as

$$\mu^+(\mathbf{x}^{\Delta}, \mathbf{y}^{\Delta}) := \frac{\sum_{t \in \mathcal{T}} \mathbb{1}\{x_t^{\Delta} y_t^{\Delta} > 0\}\mathbb{1}\{x_t^{\Delta} > 0\}}{\sum_{t \in \mathcal{T}} \mathbb{1}\{x_t^{\Delta} > 0\}}, \text{ and} \tag{6}$$

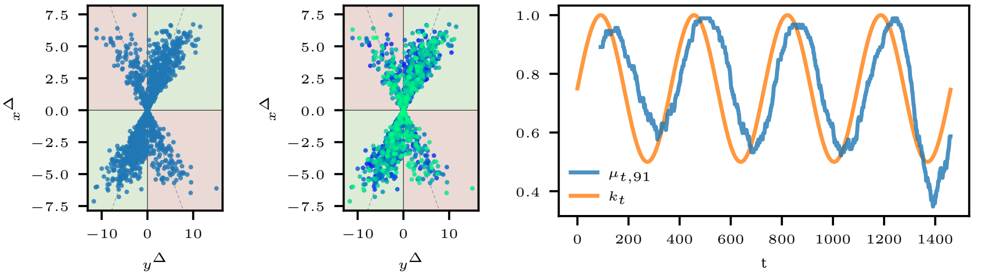
$$\mu^-(\mathbf{x}^{\Delta}, \mathbf{y}^{\Delta}) := \frac{\sum_{t \in \mathcal{T}} \mathbb{1}\{x_t^{\Delta} y_t^{\Delta} > 0\}\mathbb{1}\{x_t^{\Delta} < 0\}}{\sum_{t \in \mathcal{T}} \mathbb{1}\{x_t^{\Delta} < 0\}}. \tag{7}$$

They estimate the probability of a correct prediction of the direction of change, given that the predicted direction is positive or negative, that is, $P(X^{\Delta}Y^{\Delta} > 0 | X^{\Delta} > 0)$ and $P(X^{\Delta}Y^{\Delta} > 0 | X^{\Delta} < 0)$.

Rolling estimates of the above measures detect changes in performance over time and can give a sharper estimate of the current ATC. For the ATC ratio, a rolling estimate with a backward-looking window of length w at time t is given by

$$\mu_{t;w}(\mathbf{x}^{\Delta}, \mathbf{y}^{\Delta}) := \frac{\sum_{t^{\star}=t-w+1}^{t} \mathbb{1}\{x_{t^{\star}}^{\Delta} y_{t^{\star}}^{\Delta} > 0\}}{w}.$$

Backward-looking windows estimate the ATC ratio at a time t considering the w time steps before time t. The window length w controls the smoothing of the estimate; a larger w gives smoother results, while a small w focuses on local variations. Plotting the rolling estimates for $t = w-1, \ldots, T$ yields an estimate of the ATC ratio over time. Figure 2c depicts a rolling window estimate of the ATC ratio for the simulated data of Figures 2a and 2b. While colored four-quadrant plots, as in Figure 2b, illustrate ongoing overall drifts in the ATC, seasonal aspects are only revealed in rolling window estimates.



(a) Four-quadrant plot with simulated data.

(b) The same data is colored according to the time index t, the greener, the later.

(c) Rolling estimate of the ATC ratio over time with window length 91.

Figure 2: Visualizations for data with a time-varying ATC ratio. We defer information on the data generation process to the appendix (see A.1). The ATC ratio for the entire data set is $\mu = 0.7577$. The strong seasonality of the ATC ratio becomes visible in Figure 2c. The green curve k_t shows the theoretical probability that x_t^Δ has the same sign as y_t^Δ for each time step. The ATC ratio has a pronounced sinus-shaped seasonality with a peak after a quarter of a year and a low point after three quarters. The rolling estimates detect the yearly course of the ATC ration. Naturally, they are shifted to the right compared to k_t as the windows look backward.

2.4 Accounting for noise and non-informative small changes and bootstrapping confidence intervals

The above measures can be extended to account for information on the point's location within the quadrant. For example, points close to the zero point may have less explanatory power or may be less reliable than points far away from zero on one of the diagonals. Suppose noise or non-systematic effects are present in the true values or predictions. In that case, noise can drive a point's assignment to a quadrant instead of a systematic ATC. This is more likely for points with at least one small coordinate.

Using an exclusion area around the zero point, as further defined below, is a straightforward and highly interpretable extension of the measures of Section 2.3 accounting for such effects [see, e.g., 38, 9]. Points within that area are omitted in the calculation of the measures. In particular, the measurement, nowcast, or forecast is likely to have a noise component; thus, $\mathbf{x}^\Delta$ should be subject to an exclusion area. The measures of Equations (5), (6) and (7) without points in the exclusion area E are

$$\mu_e(\mathbf{x}^\Delta, \mathbf{y}^\Delta, E) := \frac{\sum_{t\in\mathcal{T}} \mathbb{1}\{\mathbf{x}^\Delta \mathbf{y}^\Delta > 0\}\mathbb{1}\{(y_t^\Delta, x_t^\Delta) \notin E\}}{\sum_{t\in\mathcal{T}} \mathbb{1}\{(y_t^\Delta, x_t^\Delta) \notin E\}}, \tag{8}$$

$$\mu_e^+(\mathbf{x}^\Delta, \mathbf{y}^\Delta, E) := \frac{\sum_{t\in\mathcal{T}} \mathbb{1}\{x_t^\Delta y_t^\Delta > 0\}\mathbb{1}\{x_t^\Delta > 0, (y_t^\Delta, x_t^\Delta) \notin E\}}{\sum_{t\in\mathcal{T}} \mathbb{1}\{x_t^\Delta > 0, (y_t^\Delta, x_t^\Delta) \notin E\}}, \text{ and} \tag{9}$$

$$\mu_e^-(\mathbf{x}^\Delta, \mathbf{y}^\Delta, E) := \frac{\sum_{t\in\mathcal{T}} \mathbb{1}\{x_t^\Delta y_t^\Delta > 0\}\mathbb{1}\{x_t^\Delta < 0, (y_t^\Delta, x_t^\Delta) \notin E\}}{\sum_{t\in\mathcal{T}} \mathbb{1}\{x_t^\Delta < 0, (y_t^\Delta, x_t^\Delta) \notin E\}}. \tag{10}$$

The measures are then estimators for the probability of predicting the correct direction, given that the point's location is not driven by noise or non-informative changes.

The estimators accept various shapes of the exclusion area (see Figure 1). A rectangular exclusion area, $E = \{(y, x) \in \mathbb{R}^2 : (-\varepsilon_x \leqslant x \leqslant \varepsilon_x) \wedge (-\varepsilon_y \leqslant y \leqslant \varepsilon_y)\}$ for $\varepsilon_x, \varepsilon_y > 0$, leaves out points that are small in both components. An exclusion area along one axis, for example, $E = \{(y, x) \in \mathbb{R}^2 : (-\varepsilon_x \leqslant x \leqslant \varepsilon_x)\}$ for $\varepsilon_x > 0$, removes points in which one of the components could change sign by a small amount of noise. A cross-shaped exclusion area, $E = \{(y, x) \in \mathbb{R}^2 : (-\varepsilon_x \leqslant x \leqslant \varepsilon_x) \vee (-\varepsilon_y \leqslant y \leqslant \varepsilon_y)\}$ for $\varepsilon_x, \varepsilon_y > 0$, along both axes accounts for the sign reversal in both components.

In most applications, the shape and size of the exclusion area can be chosen based on domain knowledge or expert opinions. The size determination can also be based on a proportion of the total variance or the total range of the data; for example, the 10% smallest absolute values in each component determine the exclusion area size. A third approach is to visualize the ATC ratio for different sizes of E and thus inspect the effects of the exclusion area size on the estimates. For examples of such plots, see Section 3.1.

Confidence intervals can account for the estimation uncertainty of the measures above, an approach not yet applied in the literature. Bootstrap confidence intervals are based on resampling and not on parametric assumptions as classical confidence intervals are [for introductions see 23, 2]. Many new samples are drawn with replacement from the dataset, and the statistic of interest is computed for each sample, yielding an estimate for the distribution of the statistic of interest. Based on the derived "new" samples of the statistic, the confidence intervals can be derived through different bootstrapping methods. We use the bias-corrected and accelerated (BCa) approach for bootstrapping in the following, as it holds the confidence level for small and large samples and has a moderate computation time (see the simulation study in Appendix A.2).

2.5 The conditional ATC plot

The estimators described above provide information on the probabilities $P(X^\Delta Y^\Delta > 0 | X^\Delta Y^\Delta \notin E)$, $P(X^\Delta Y^\Delta > 0 | X^\Delta > 0, X^\Delta Y^\Delta \notin E)$, and $P(X^\Delta Y^\Delta > 0 | X^\Delta < 0, X^\Delta Y^\Delta \notin E)$. Notably, the visual analysis of these probabilities has not yet been addressed in the literature. A still finer analysis might be gained by considering the conditional distribution $P(X^\Delta Y^\Delta > 0 | X^\Delta = \chi)$ to assess the ATC of a prediction for a specific change $X^\Delta = \chi$ of the measurement, nowcast, or forecast. Thereby, $P(X^\Delta Y^\Delta > 0 | X^\Delta = \chi)$ denotes the probability of a correct direction given a predicted change of χ. Thus, if a change of χ is observed in practice, one can directly assess its credibility regarding the direction. Multivariate kernel density estimation (KDE) facilitates the continuous estimation of $P(X^\Delta Y^\Delta > 0 | X^\Delta = \chi)$ by estimating the components $f_{X^\triangle,Y^\triangle}$ and $f_{X^\triangle}$ of

$$P(X^\triangle Y^\triangle > 0 | X^\triangle = \chi) = \begin{cases} \int_{-\infty}^{0} \frac{f_{X^\triangle,Y^\triangle}(\chi,y)}{f_{X^\triangle}(\chi)}\,\mathrm{d}\,y & \text{if } \chi < 0, \\ \int_{0}^{\infty} \frac{f_{X^\triangle,Y^\triangle}(\chi,y)}{f_{X^\triangle}(\chi)}\,\mathrm{d}\,y & \text{if } \chi > 0, \end{cases}$$

for $\chi \neq 0$ through a KDE. Gramacki [19] provides a comprehensive introduction to multivariate KDE, and implementations are available in many programming languages, for example, in the `statsmodels` in Python [39]. The KDE yields estimates for $P(X^\triangle Y^\triangle > 0 | X^\triangle = \chi)$ for all values of $\chi \in \mathbb{R}$. Multivariate KDE takes a kernel and bandwidth selector as modeling parameters. We advise using a Gaussian kernel and the cross-validation maximum likelihood as bandwidth selector (see Appendix A.3).

Assessing $P(X^\triangle Y^\triangle > 0 | X^\triangle = \chi)$ graphically by drawing $P(X^\triangle Y^\triangle > 0 | X^\triangle = \chi)$ against χ eases the simultaneous evaluation of various χ. Furthermore, the graph facilitates the comparison of various methods in a single graph, and asymmetries of $P(X^\triangle Y^\triangle > 0 | \mathbf{x}^\triangle = \chi)$ with respect to χ in the ATC can be detected. We refer to the plot as a *conditional ATC plot*.

2.6 Probabilistic evaluation

In nowcasting and forecasting, probabilistic predictions have become more prevalent in recent years (see Sections 3.1 and 3.2). In this section, we develop ATC assessments for probabilistic measurements and nowcasts, an approach not yet explored in the literature. Probabilistic predictions issue a probability distribution for the quantity of interest based on their available information and, thus, include a point estimate and information on the prediction uncertainty and quantiles simultaneously. Probabilistic predictions thus also contain a probability of a positive or negative change. For ATC assessment, we compare the predicted probability of positive change, denoted by p_t, with the occurrence of positive changes.

Probabilistic predictions can be a cumulative distribution function (CDF), probability density function (PDF), or quantiles. The CDF is the most general and can be used to derive the others, given that they exist. Let us first assume that the prediction is a CDF, and that y_{t-l} is known at time t (see Table 1). Appendix A.4 extends the analysis to quantile predictions or unknown true values.

Let for a forecast $F_{t|t-l}(x)$ denote the predicted CDF for target time t and issue time $t-l$, where the index is analogous to the point notation of Section 2.1. The CDF $F_{t|t-l}(x)$ specifies the forecasted probability that the quantity of interest is at most x. A positive change occurs for any value at t larger than the true value y_{t-l} and the CDF $F_{t|t-l}(y_{t-l})$ yields the predicted probability of any value at most y_{t-l}, and, thus, a negative change. Accordingly, the forecasted probability of a positive change is

$$p_t = 1 - F_{t|t-l}(y_{t-l}) \quad t = l, \ldots, T.$$

The computation differs slightly for nowcasts, that is,

$$p_t = 1 - F_{t|t}(y_{t-l}) \quad t = l, \ldots, T,$$

with analogous derivations as above. Let z_t denote the indicator that the observed change at time t is positive, that is,

$$z_t = \mathbb{1}\{y_t^{\Delta} > 0\} \quad t = l, \dots, T.$$

The predictive power of $\mathbf{p} = (p_t)_{t=l}^{T}$ for $\mathbf{z} = (z_t)_{t=l}^{T}$ can be assessed using probabilistic dichotomous forecast evaluation methods. Dichotomous forecasts predict a binary outcome, such as a positive or negative change, and are evaluated numerically using scoring rules or visually through reliability diagrams.

The Brier score (BS) is a widely used scoring rule for dichotomous probabilistic forecasts [5]. In our context, it is

$$BS(\mathbf{p}, \mathbf{z}) = \frac{1}{T - l + 1} \sum_{t=l}^{T} (p_t - z_t)^2.$$

Lower values indicate the considered method's higher probabilistic ATC. The BS assesses the calibration and sharpness of the forecast and the observation simultaneously [33, 30]. Calibration refers to the statistical consistency of forecasts and observations; that is, the event occurs with the issued probability and is considered the more fundamental quality [16]. Sharpness refers to the spread of the forecast; probabilities close to zero and one are preferable as they convey a higher certainty.

Graphical methods are a standard tool for evaluating the calibration of probabilistic forecasts in detail. In dichotomous forecasting, the reliability diagram is frequently used [33]. The reliability diagram plots the observed frequency of the positive outcome against the (binned) predicted probability. For example, it shows the proportion of observed increases, given that the predicted probability of increase was approximately 0.7. Ideally, the predicted probability equals the observed frequency, and the reliability diagram is a 45-degree line. Local deviations from the 45-degree line indicate a miscalibration for specific forecast probabilities. Thus, the reliability visualizes the local and overall calibration simultaneously. For an example of a reliability diagram, see Section 3.2.

3 Application to medical/healthcare nowcasting, forecasting, and measurement data

3.1 Nowcasting during the COVID-19 pandemic

In Germany, the seven-day hospitalization rate was established as a central steering measure in November 2021 during the COVID-19 pandemic, and the imposition of severe public restrictions was based on it [34]. However, the publication of the definite hospitalization rate was substantially delayed and partially flawed for two main reasons. First, the reporting process was delayed because – among other reasons – different authorities were involved in passing the data to the RKI [35]. Second, the seven-day hospitalization rate allocated all COVID-19-related hospitalizations to the

Model	RMSE	MAE	Count
ILM	648	504	153
RKI	810	670	156
RIVM	820	674	159
ENS-MED	832	675	158
ENS-MEAN	841	666	158
LMU	979	810	159
SZ	1,048	834	159
SU	1,127	899	159
KIT	1,161	912	159
EPI	1,513	1,006	159

Table 2: Point evaluation measures for the issued mean of the different models in COVID-19 nowcasting. “RMSE” and “MAE” are accuracy measures, while “Count” lists the number of non-missing values. The RMSE orders the models. The evaluation period comprises 159 days, and only a few nowcasts are missing [for explanations of the missing values, see 47, Tables A2, A3, and A4]. Note that the high values for the EPI model could be driven by an exceptionally far-off value at the end of the evaluation period (see Figure 12).

date of the first positive test [for a detailed description, see 47]. The COVID-19-Nowcasting-Hub [14] collected various nowcasts in a predefined setup, including the mean, median and other quantiles of the predicted seven-day hospitalization rate (for further information see [47] and Table 9 for the abbreviations used). In addition to those nowcasts, Wolffram et al. [47] construct two ensemble methods using the ensembles’ mean or median. We denote them by ENS-MEAN and ENS-MED. In line with the initial study design, we consider the period from November 22, 2021, to April 29, 2022, as the evaluation period. We use the data from February 8, 2024, for the true values and focus on nowcasts for all inhabitants of Germany. Figure 12 in Appendix B displays the true and nowcast data for the evaluation period. The time comprises the fourth wave’s end in December 2021 and nearly the entire fifth wave of the pandemic in Germany, lasting until May 28, 2022 [43].

Traditional methods often focus on point or distributional accuracy of hospitalization rates, but understanding the direction and reliability of changes is crucial for effective decision-making. The ATC assessment offers easy-to-interpret insights and helps determine not only if rates are rising or falling, but also how confidently we can make this determination. For instance, if hospitalization rates are rising, public health measures may need to be tightened. Conversely, falling rates might justify loosening restrictions. The ATC assessment’s ability to reveal asymmetries is especially valuable, as it can show whether certain models are more adept at recognizing decreases than increases, or vice versa. This information, which is not readily apparent from traditional methods, can significantly impact the interpretation of nowcasts and subsequent policy decisions for public health officials.

	μ^7	$\mu^{+,7}$	$\mu^{-,7}$	$\mu^7_{q_{0.1}}$	$\mu^{+,7}_{q_{0.1}}$	$\mu^{-,7}_{q_{0.1}}$
EPI	0.77 (0.71, 0.82)	0.67 (0.58, 0.75)	0.87 (0.79, 0.92)	0.78 (0.72, 0.83)	0.68 (0.59, 0.77)	0.88 (0.81, 0.93)
ILM	0.85 (0.80, 0.89)	0.73 (0.64, 0.80)	0.99 (0.94, 1.00)	0.85 (0.80, 0.90)	0.74 (0.65, 0.81)	0.99 (0.94, 1.00)
KIT	0.74 (0.69, 0.79)	0.64 (0.55, 0.72)	0.87 (0.80, 0.93)	0.75 (0.69, 0.80)	0.64 (0.55, 0.72)	0.88 (0.81, 0.94)
LMU	0.80 (0.74, 0.85)	0.70 (0.62, 0.79)	0.91 (0.84, 0.95)	0.81 (0.75, 0.86)	0.72 (0.63, 0.79)	0.92 (0.85, 0.96)
ENS-MEAN	0.82 (0.76, 0.86)	0.71 (0.63, 0.79)	0.94 (0.89, 0.99)	0.82 (0.76, 0.87)	0.71 (0.63, 0.78)	0.96 (0.90, 0.99)
ENS-MED	0.82 (0.76, 0.87)	0.70 (0.62, 0.78)	0.96 (0.90, 0.99)	0.83 (0.77, 0.87)	0.72 (0.63, 0.79)	0.96 (0.90, 0.99)
RIVM	0.83 (0.77, 0.87)	0.74 (0.65, 0.81)	0.92 (0.86, 0.96)	0.83 (0.78, 0.88)	0.74 (0.65, 0.81)	0.93 (0.87, 0.97)
RKI	0.72 (0.65, 0.77)	0.60 (0.51, 0.67)	0.98 (0.92, 1.00)	0.73 (0.67, 0.78)	0.61 (0.52, 0.68)	0.98 (0.92, 1.00)
SU	0.81 (0.75, 0.86)	0.71 (0.62, 0.78)	0.92 (0.85, 0.96)	0.81 (0.75, 0.85)	0.71 (0.63, 0.79)	0.92 (0.85, 0.96)
SZ	0.78 (0.72, 0.83)	0.67 (0.58, 0.75)	0.91 (0.84, 0.96)	0.78 (0.72, 0.83)	0.67 (0.58, 0.75)	0.92 (0.85, 0.97)

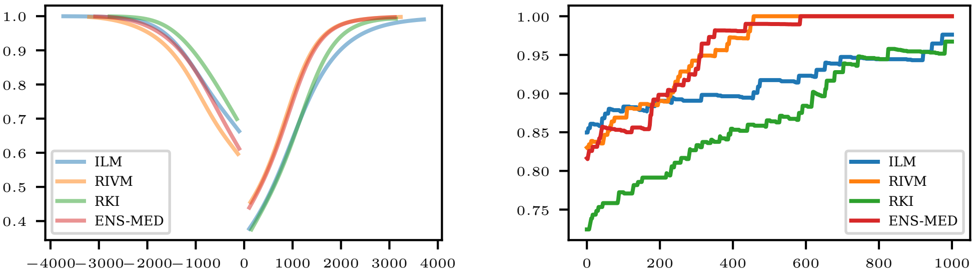
Table 3: The ATC ratio μ^7, positive ATC ratio $\mu^{+,7}$, and negative ATC ratio $\mu^{-,7}$ for the models without and with exclusion areas for the horizon seven days in COVID-19 nowcasting. The exclusion areas are rectangles centered on the zero points with a width and height of twice the 10%-quantile of the absolute values of nowcast and true values. The subscript $q_{0.1}$ denotes the measures with exclusion area. There are 66 positive and 93 negative actual changes in the considered time period.

Results

Table 2 summarizes the non-ATC-aware point evaluation measures for the issued mean of the different models. The best-performing models in terms of RMSE and MAE are the ILM and RKI models. The ensemble methods ENS-MED and ENS-MEAN perform worse than the best models regarding the mean location. The performance of the models is diverse, with more than twice as high RMSE values for the worst models compared to the best models.

In the following, we apply the ATC assessment for the short-term horizons one and medium-term horizons seven and 14 days. The horizons seven and 14 reflect a typical period until new policy changes are taken. We start by providing background information on the marginal distributions of the actual value and nowcast changes for the different horizons in Table 10 in Appendix B such as standard deviation and quantiles of the nowcasts and true values. The variability and general level of changes grow with the horizon: The standard deviation increases from roughly 300 for horizon one to 1,200 for horizon seven and 2,000 for horizon 14 days. Similarly, the 10%-quantile of changes, the basis for the exclusion area size, increases. The exclusion area is rectangular; a point falls within it if both $\mathbf{y}^\Delta$ and $\mathbf{x}^\Delta$ are below the respective 10%-quantile of the absolute changes. Thus, points are still included in the ATC assessment if they are large in one dimension but not in the other, thus ensuring that substantial changes in, for example, $\mathbf{y}^\Delta$ are to be recognized by the nowcast and vice versa.

Table 3 lists the ATC ratios for all models without and with exclusion areas for the horizon of seven days. The ATC ratios without exclusion area range from 0.72 to 0.85 for the horizon of seven days. The negative ATC ratios are higher than the positive ATC ratios for all models. The confidence intervals for the positive and negative



(a) Conditional ATC plot.

(b) ATC ratio over exclusion area size in $\mathbf{x}^{\Delta}$.

Figure 3: Conditional ATC plot and ATC ratio over exclusion area for the nowcasts of the seven-day hospitalization rate ILM, RKI, RIVM, and ENS-MED for the horizon seven days in COVID-19 nowcasting.

ATC ratios do not overlap for all models, indicating that the ATC ratios are indeed different. The 10%-quantile exclusion areas have, at most, an influence of 0.03 on the ratios. The model with the highest ATC ratio is the ILM model, and the model with the lowest is the RKI model. The confidence intervals between all models without and with exclusion areas overlap. The positive ATC ratio implies a similar ranking of the models than the overall ATC, while the negative ratio provides a different ranking, for example, for the RKI model. For the horizons of one and 14 days, we refer to Table 11 in Appendix B.

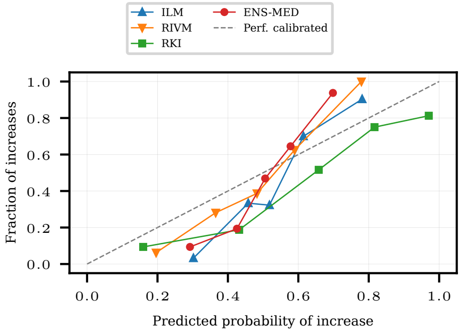
Figure 3 shows the conditional ATC plots and the ATC ratio over the exclusion area for the horizon seven days; the respective plots for the horizons one day and 14 days are shown in Figure 14. Here, only the best models in point evaluation measures, ILM, RKI, RIVM, and ENS-MED, are shown to keep the plots easily readable. If RKI or ILM issues a fall in the hospitalization rate, the probability of a fall is higher than if RIVM or ENS-MED issues a fall. The opposite is the case for a nowcasted hospitalization rate increase, and the difference between the models' performance is more prominent than for a fall. Similar observations can be made for the horizon of 14 days in Figure 14b. For a horizon of one day, the models' conditional ATC difference is less pronounced (see Figure 14a). The RKI model is still less conclusive when issuing an increase in the hospitalization rate, while RIVM is most informative in that case. The curves cross for an issued fall, with ENS-MED being on top for issued falls above 250.

The ATC ratios for various exclusion areas are shown in Figure 3b. The ATC ratio generally increases with larger exclusion areas. While the RIVM and ENS-MED ATC ratios evolve similarly, the RKI and ILM ATC ratios get closer. For the horizon of one day, the RKI ATC ratio decreases with increasing exclusion area size while the other models rise (see Figure 14c). For the horizon of 14 days, all ATC ratio curves increase with the exclusion area size (see Figure 14d).

Figure 4 shows the Brier score (BS) and reliability diagrams for the same subset of models, the ILM, RKI, RIVM, and ENS-MED. The probabilities of increase for the different models are computed using the nowcast quantiles. For each horizon l, $10,000$

	1 d	7 d	14 d
ILM	0.1783	0.1269	0.1119
RIVM	0.1606	0.1136	0.1274
RKI	0.1893	0.2200	0.1712
ENS-MED	0.1812	0.1096	0.1066

(a) BS for the different models and horizons.



(b) Reliability diagram for horizon one day.

Figure 4: Brier scores and reliability diagrams for the COVID-19 nowcasting models ILM, RKI, RIVM, and ENS-MED. The reliability diagram bins are chosen according to the empirical quantiles of the predicted probabilities. In the computation of BS and reliability diagram, missing values are excluded. The reliability diagram for the horizons seven and 14 days is in the appendix (see Figure 15).

samples of the forecast date t and the forecast date $t - l$ based on the nowcasts of issue date t are generated, and the proportion of positive changes is computed (see Appendix A.4). Remember that a low BS and a reliability diagram along the diagonal are signs of a high ATC. The BS is the lowest for the RIVM model for a one-day-horizon, while the ENS-MED model has the best BS for the horizon of seven and 14 days. The RKI model yields the highest BS for all horizons. Note that the BS is 0.25 for random guessing; thus, all models perform better than random guessing. The reliability diagrams show that the models are not well calibrated for the horizon of one day. While for predicted probabilities below 0.5, the observed ratio of increases is smaller, it is higher than predicted for probabilities above 0.5. Figure 15c shows the observed predicted increase probabilities for all horizons. For the horizons of seven and 14 days, the nowcasters issue only a few moderate probabilities, and most probabilities are near zero and one.

Discussion

For all horizons, the influence of the exclusion area on the 10%-quantile level is negligible. For example, the ATC ratio changes at most by 0.03 for the EPI model with $\mu^{-,14}$. The exclusion areas are thus not crucial for the ATC assessment in the case of the nowcasts of the seven-day hospitalization rate. The lower bound of confidence intervals is at least 0.68 for all models, indicating that they perform better than random guessing the trend.

ATC assessment evaluates the models differently from point evaluation measures. RKI is among the best in point evaluation measures but performs worse in ATC assessment. The assessment of asymmetry in the conditional ATC plots is crucial for interpreting the ATC ratios, with the RKI model being the most prominent example.

Figure 3b shows that larger exclusion areas increase the ATC ratio, indicating that the predicted direction is more accurate for large predicted changes.

The probabilistic ATC assessment shows that the models are better than random guessing. The reliability diagram cannot provide information if specific probabilities are issued scarcely. Thus, the reliability diagrams for the horizons of seven and 14 days do not contain information on moderate probabilities. The BS values, however, work well for those examples and provide a good measure for the ATC of the models.

A more extensive data size would be beneficial for assessing the models' performance. For the evaluation period of 159 days, the ATC ratio confidence intervals overlap; thus, no conclusions can be drawn from the ATC assessment comparing the models.

3.2 Forecasting patient admissions to an emergency department

In a second example, we consider forecasting patient admissions to an emergency department per hour with data and models by Rostami-Tabar, Browell, and Svetunkov [36]. Every 12 hours, the models issue hourly forecasts for the next 48 hours.

Rostami-Tabar, Browell, and Svetunkov [36] publish means and probabilistic quantile forecasts for various models and input data. We use the published mean as a point forecast for the ATC assessment and evaluate the probabilistic ATC based on the quantile forecasts subsequently. Considering only the forecasts of at least 36 hours ahead, we restrict the evaluation period to March 2, 2018, at noon, to February 28, 2019, at 23:00, comprising 8,724 hours.

While traditional methods provide point estimates, ATC assessment offers a simple, intuitive way to evaluate model performance and directional accuracy. This approach is particularly valuable for management, as it facilitates easy comparisons between expected workload and recent shifts. For instance, if staff was near capacity during the last shift, ATC assessment can clearly indicate whether an increase in patient admissions is likely, allowing management to proactively adjust resources. This directional insight, often hidden in conventional methods, enables more informed and timely decision-making.

The number of patient admissions has a strong weekly and daily pattern. Thus, we consider the horizons of 72 hours, the last already observed shift of the same hour of day, and seven days, the previous shift of the same hour and day, in ATC assessment.

Results

Table 4 lists the point evaluation measures and the count of available forecasts. The best-performing models regarding RMSE and MAE are the NBI-2 and Poisson-2 models. More than 8,600 forecasts are available for all models, with changes in the number due to missing values on four afternoons in 2018.

We start by analyzing the marginal distributions for the predicted and observed changes for the three- and seven-day horizons in Table 5, again. The computed difference aligns with Section 2.1, that is, the difference between the forecasted mean and true value of three and seven days before, as the actual value is available when

Model	RMSE	MAE	Count
NBI-2	8.883	3.200	8688
Poisson-2	8.884	3.200	8688
Poisson-1	9.164	3.238	8688
Benchmark-2	9.246	3.236	8688
Ttr-2	9.394	3.266	8688
NOtr-1	9.413	3.276	8688
NOtr-2	9.413	3.276	8688
Poisson-2-I	9.458	3.276	8688
Benchmark-1	10.065	3.331	8688
GBM-2	11.663	3.542	8688
tbats	12.905	3.912	8724
Prophet	13.078	3.877	8724
qreg-1	13.337	3.758	8688
Regression-Poisson	21.162	4.818	8724
ADAM-iETSX	28.000	5.561	8724
ETS	29.358	5.742	8724

Table 4: Point evaluation measures for the forecasting models for patient admissions to an emergency department. The smaller count for some models stems from missing forecasts scattered throughout the evaluation period. Note that the reported values for the RMSE differ from those in Rostami-Tabar, Browell, and Svetunkov [36] due to differences in the evaluation period.

issuing the forecast. The positive change fraction varies between 0.39 and 0.63 for the horizon of three days and between 0.37 and 0.63 for the horizon of seven days. The variability of changes decreases for the larger horizon for most models; only for the ETS model does it increase. The 10%-quantile of the changes is between zero and one for all models and horizons. Thus, we use an exclusion area of size 1. The resulting fraction of included values in the computation is also listed in Table 5 and is at least 79% of the values.

Table 6 lists the ATC ratios for all models for three and seven-day horizons. The ATC ratios range from 0.68 to 0.84 for a horizon of three days and from 0.68 to 0.82 for seven days. The negative and positive ATC ratios differ for all models and horizons. For some models, for example, the GBM-2 model, the positive ATC ratio is higher than the negative ATC ratio, and for some models, for example, the tbats model, vice versa. The confidence interval width is at most 0.02 for the ATC ratios and at most 0.03 for the positive and negative ATC ratios. The models GBM-2, qreg-1, and Benchmark-1 have the highest positive ATC ratio for the three and seven-day horizons, while Poisson-2 and NBI-2 have the highest negative ATC ratio.

Figure 5 shows the conditional ATC plots for the models Benchmark-1, GBM-2, NBI-2, Poisson-2, and qreg-1 for the horizons three and seven days and thus inspects the local ATC of the models with highest positive and negative ATC ratio. The conditional ATC plots show similar courses for the two horizons, though the curves are shifted downwards for the horizon of seven days. The model's relative ATC evolves

	(1), l=3	$\sigma_{x^{\Delta,3}}$	$q_{0.1}(x^{\Delta,3})$	(2), l=3	(1), l=7	$\sigma_{x^{\Delta,7}}$	$q_{0.1}(x^{\Delta,7})$	(2), l=7
ADAM-iETSX	0.57	7.76	0.83	0.88	0.57	7.49	0.78	0.87
Benchmark-1	0.45	5.05	0.50	0.80	0.44	4.43	0.47	0.78
Benchmark-2	0.51	5.11	0.52	0.80	0.50	4.29	0.45	0.78
ETS	0.58	7.49	0.78	0.87	0.58	7.68	0.84	0.88
GBM-2	0.39	4.93	0.51	0.80	0.37	4.61	0.49	0.79
NBI-2	0.53	5.04	0.52	0.81	0.53	4.41	0.48	0.79
NOtr-1	0.52	5.03	0.51	0.81	0.51	4.41	0.49	0.79
NOtr-2	0.52	5.03	0.51	0.81	0.51	4.41	0.49	0.79
Poisson-1	0.53	5.04	0.51	0.81	0.52	4.38	0.48	0.79
Poisson-2	0.53	5.05	0.52	0.80	0.53	4.42	0.48	0.78
Poisson-2-I	0.51	5.03	0.51	0.81	0.50	4.42	0.49	0.79
Prophet	0.62	5.27	1.00	0.91	0.62	5.15	1.00	0.91
Regression-Poisson	0.51	6.65	0.67	0.85	0.51	6.49	0.67	0.85
Ttr-2	0.51	5.03	0.50	0.81	0.50	4.41	0.49	0.79
qreg-1	0.39	5.01	0.49	0.81	0.39	4.84	0.51	0.80
tbats	0.63	5.35	1.00	0.92	0.63	5.04	1.00	0.92
True	0.54	6.61	1.00	0.93	0.55	5.90	1.00	0.92

Table 5: Marginal analysis of the forecast and true changes in patient admissions to an emergency department. The column (1) shows the fraction of values greater than zero for horizon l, $\sigma_{x^{\Delta,l}}$ the standard deviation, and $q_{0.1}(x^{\Delta,l})$ the 10% quantile of the changes' absolute values. Column (2) shows the fraction of values not in the exclusion area of size one.

consistently for the two horizons, with the NBI-2 and Poisson-2 models being indistinguishable. The GBM-2 model outperforms the qreg-1 model for all predicted changes. The models NBI-2 and Poisson-2 have the highest ATC for all negative predicted changes and the lowest for all positive predicted changes. Benchmark-1 lies between the other models for all predicted changes.

Figure 6 visualizes the probabilistic ATC assessment for the same subset of models. The Brier scores (BSs) are shown in Figure 6a, and the reliability diagrams for the horizons three and seven days in Figures 6b and 6c. The BSs are smallest for NBI-2 and Poisson-2 for both horizons, while the BSs for the other models are larger and differ more. The qreg-1 model has both horizons' highest BS. The reliability diagrams of GBM-2 and NBI-2 are also close and show a too-small fraction of increases for the predicted probability overall. For the other models, the reliability diagrams show a fraction of increases that are too large for the corresponding predicted probability.

Discussion

The ATC is consistent for the two horizons, with the models' relative ATC evolving similarly for the two horizons. The models' ATC is generally higher for the smaller horizon, but the changes are minor, and confidence intervals overlap.

The positive and negative ATC ratios differ for all models. While some models, such as GBM-2 and qreg-1, have the highest positive ATC ratio, others, such as Poisson-2 and NBI-2, have the highest negative ATC ratio. Thus, the uncertainty of the model's predicted change has to be assessed differently based on the direction.

The probabilistic ATC assessment results endorse the point ATC assessment and assign the best scores to NBI-2 and Poisson-2. The reliability diagrams show that they underestimate the fraction of increases slightly.

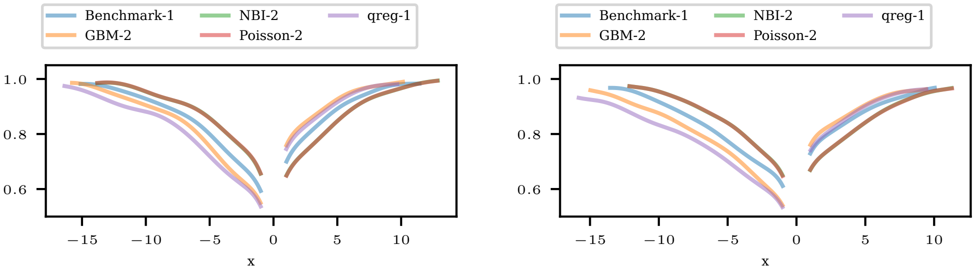
	μ^{3}_{+}	$\mu^{+,3}_{+}$	$\mu^{-,3}_{+}$	μ^{7}_{+}	$\mu^{+,7}_{+}$	$\mu^{-,7}_{+}$
ADAM-iETSX	0.70 (0.69, 0.71)	0.68 (0.67, 0.69)	0.72 (0.71, 0.73)	0.68 (0.67, 0.69)	0.67 (0.66, 0.69)	0.69 (0.67, 0.70)
Benchmark-1	0.83 (0.82, 0.84)	0.86 (0.85, 0.87)	0.81 (0.79, 0.82)	0.81 (0.80, 0.82)	0.86 (0.85, 0.87)	0.78 (0.76, 0.79)
Benchmark-2	0.84 (0.83, 0.84)	0.83 (0.82, 0.85)	0.84 (0.83, 0.85)	0.82 (0.81, 0.83)	0.83 (0.82, 0.84)	0.80 (0.79, 0.82)
ETS	0.68 (0.67, 0.69)	0.66 (0.65, 0.67)	0.70 (0.69, 0.72)	0.67 (0.66, 0.68)	0.66 (0.64, 0.67)	0.68 (0.66, 0.69)
GBM-2	0.82 (0.81, 0.82)	0.90 (0.89, 0.91)	0.77 (0.76, 0.78)	0.78 (0.77, 0.79)	0.88 (0.87, 0.90)	0.73 (0.72, 0.74)
NBI-2	0.84 (0.83, 0.85)	0.83 (0.82, 0.84)	0.85 (0.84, 0.86)	0.82 (0.81, 0.83)	0.82 (0.81, 0.83)	0.82 (0.81, 0.83)
NOtr-1	0.83 (0.83, 0.84)	0.83 (0.82, 0.84)	0.84 (0.82, 0.85)	0.81 (0.80, 0.82)	0.82 (0.81, 0.83)	0.80 (0.79, 0.81)
NOtr-2	0.83 (0.83, 0.84)	0.83 (0.82, 0.84)	0.84 (0.82, 0.85)	0.81 (0.80, 0.82)	0.82 (0.81, 0.83)	0.80 (0.79, 0.81)
Poisson-1	0.84 (0.83, 0.84)	0.82 (0.81, 0.83)	0.85 (0.84, 0.86)	0.82 (0.81, 0.82)	0.82 (0.81, 0.83)	0.81 (0.80, 0.82)
Poisson-2	0.84 (0.83, 0.85)	0.83 (0.82, 0.84)	0.85 (0.84, 0.86)	0.82 (0.81, 0.82)	0.82 (0.81, 0.83)	0.82 (0.80, 0.83)
Poisson-2-I	0.83 (0.83, 0.84)	0.84 (0.83, 0.85)	0.83 (0.82, 0.84)	0.81 (0.80, 0.82)	0.83 (0.81, 0.84)	0.80 (0.79, 0.81)
Prophet	0.75 (0.74, 0.76)	0.72 (0.71, 0.73)	0.79 (0.77, 0.80)	0.74 (0.73, 0.74)	0.72 (0.70, 0.73)	0.76 (0.75, 0.77)
Regression-Poisson	0.72 (0.71, 0.73)	0.73 (0.71, 0.74)	0.72 (0.70, 0.73)	0.70 (0.69, 0.71)	0.71 (0.70, 0.73)	0.69 (0.67, 0.70)
Ttr-2	0.84 (0.83, 0.84)	0.84 (0.83, 0.85)	0.83 (0.82, 0.85)	0.81 (0.80, 0.82)	0.83 (0.82, 0.84)	0.80 (0.79, 0.81)
qreg-1	0.80 (0.79, 0.80)	0.88 (0.87, 0.89)	0.75 (0.74, 0.76)	0.77 (0.76, 0.78)	0.86 (0.85, 0.88)	0.71 (0.70, 0.72)
tbats	0.75 (0.74, 0.76)	0.72 (0.71, 0.73)	0.80 (0.78, 0.81)	0.73 (0.72, 0.74)	0.71 (0.69, 0.72)	0.76 (0.74, 0.77)

Table 6: ATC ratio μ_{+}, positive ATC ratio μ^{+}_{+}, and negative ATC ratio μ^{-}_{+} for the models for the horizons 72 hours and seven days in the forecasting of patient admissions to an emergency department with an exclusion zone equal to the axis indicated by the subscript +. There are 4030 positive changes and 4051 negative changes for the horizon of 72 hours and 4042 and 3961 for the horizon of 7 days.

Overall, the example provides performance assessments that are different from standard point evaluation measures and thus provide further insights into the strengths and weaknesses of the models. While the models with the lowest RMSE, NBI-2, and Poisson-2, also have a high ATC, three models with below-average point evaluation measures, Benchmark-1, GBM-2, and qreg-1, have a high positive ATC.

3.3 Non-invasive blood pressure monitoring

We here consider the ATC of non-invasive blood pressure measurements from the MIMIC-III database that includes data of critically ill patients treated in intensive care units of the Beth Israel Deaconess Medical Center in Boston (Massachusetts, USA, [26], [31]; available through [17]). We focus on invasive arterial blood pressure (ABP) and non-invasive blood pressure (NBP) measurements and thus limit our analysis to datasets containing simultaneous measurements of ABP and NBP simultaneously. 2,548 datasets include at least one measurement of systolic ABP and NBP, and 1,327 include at least one pair of simultaneously measured systolic ABP and NBP; for the mean ABP and NBP, the numbers are 2,605 and 1,516, respectively. We assess the ATC of non-invasive blood pressure measurements (test method) compared to intraarterial blood pressure measurements (reference method, gold standard). We consider



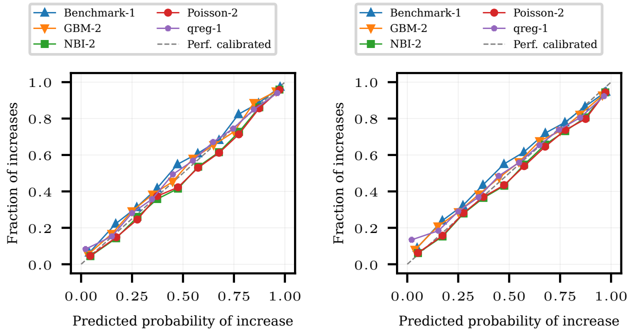
(a) Horizon three days.

(b) Horizon seven days.

Figure 5: Conditional ATC plots for the horizons three and seven days and the models with the best positive or negative ATC in forecasting the patient admissions to an emergency department. The plots of NBI-2 and Poisson-2 are indistinguishable.

	3 d	7 d
Benchmark-1	0.1586	0.1761
GBM-2	0.1590	0.1759
NBI-2	0.1549	0.1717
Poisson-2	0.1549	0.1714
qreg-1	0.1679	0.1843

(a) BS for the different models and horizons.



(b) Reliability diagram for horizon three days.

(c) Reliability diagram for horizon seven days.

Figure 6: Probabilistic ATC assessment for the models Benchmark-1, GBM-2, NBI-2, Poisson-2, and qreg-1 for the horizons three and seven days in forecasting patient admissions to an emergency department. The Brier score in Figure 6a evaluates the calibration and sharpness of the probabilistic ATC simultaneously, while the two plots on the right assess solely the calibration, that is, whether the predicted probability of increase occurs empirically. Probabilistic ATC by the BS is best for the models NBI-2 and Poisson-2 for both lags.

the horizons of one minute, five minutes, and 15 minutes for the ATC assessment, as those are typical intervals of NBP measurements.

While ATC assessment is well-established in measurement analysis, our new methods extend its application, revealing insights such as asymmetries. This extension enhances the utility of ATC assessment beyond its conventional use, offering a more comprehensive evaluation of model performance.

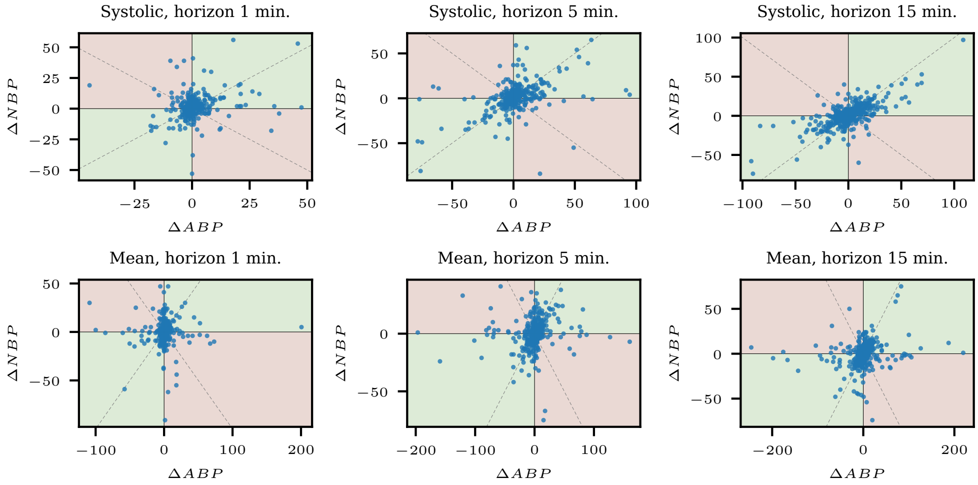


Figure 7: Four-quadrant plots for the different horizons l and the systolic and mean blood pressure measurements. The upper row contains systolic measurements, and the lower row contains mean measurements. The columns contain the horizons one, five, and 15 minutes.

Results

Again, we exclude the smallest 10% of absolute changes in ATC assessment. The resulting four-quadrant plots of the mean and systolic blood pressure measurements for the different horizons are shown in Figure 7. The number of points in the four-quadrant plot is smaller due to the restriction to data records with measurements of mean or systolic ABP and NBP simultaneously for two consecutive times with the specified horizons. Thus, we use the NBP measurements as the test method and the ABP measurements as the gold standard. For the systolic measurements, 290, 332, and 442 points are available for the horizons of one, five, and 15 minutes; for the mean measurements, 406, 430, and 542.

The ATC ratios, including confidence intervals for the different horizons, are listed in Table 7. The confidence intervals have lower bounds of 0.5 or slightly above for the measurements with a horizon of one minute. For larger horizons, the ATC ratio increases. The difference between positive and negative ATC ratios is small for all types and horizons, with overlapping confidence intervals.

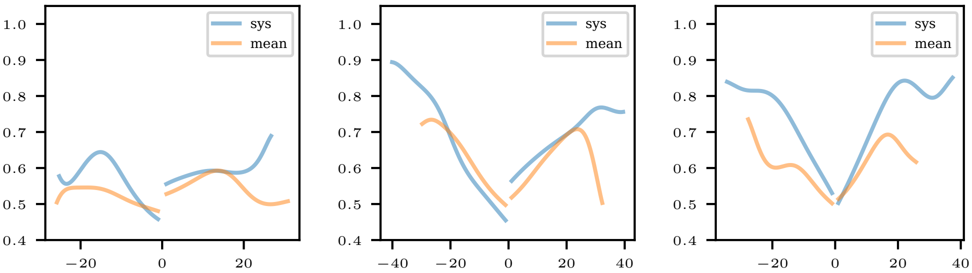
Figure 8 shows the conditional ATC plots for the different horizons and the systolic and mean blood pressure measurements. It becomes apparent that the systolic measurements have a higher ATC than the mean measurements, except for small negative predicted changes.

Discussion

The four-quadrant plots contain a considerable number of extreme points. Whether these points are due to measurement errors or extreme values is not distinguishable.

Type	l	μ^l	$\mu^{+,l}$	$\mu^{-,l}$
Systolic	1	0.55 (0.50, 0.60)	0.59 (0.52, 0.65)	0.58 (0.50, 0.66)
Systolic	5	0.63 (0.59, 0.68)	0.70 (0.64, 0.75)	0.62 (0.56, 0.69)
Systolic	15	0.69 (0.65, 0.73)	0.72 (0.66, 0.76)	0.74 (0.69, 0.79)
Mean	1	0.55 (0.51, 0.59)	0.62 (0.56, 0.68)	0.56 (0.50, 0.62)
Mean	5	0.59 (0.55, 0.64)	0.65 (0.59, 0.71)	0.62 (0.56, 0.68)
Mean	15	0.62 (0.58, 0.65)	0.65 (0.60, 0.70)	0.66 (0.61, 0.71)

Table 7: ATC ratios for the different horizons l and the systolic and mean blood pressure measurements.



(a) Horizon one minute.

(b) Horizon five minutes.

(c) Horizon 15 minutes.

Figure 8: Conditional ATC plot for the systolic and mean blood pressure measurements and the horizons one, five, and 15 minutes.

Some authors argue to exclude the measurements below the 10%-quantile of the absolute changes and the points above the 90%-quantile [see 9]. We do not follow this approach here, as the extreme values are not necessarily measurement errors and could be particularly relevant.

The difference between positive and negative ATC ratios is small in this example. The positive and negative ATC ratios have overlapping confidence intervals, the conditional ATC plots do not contain prominent deviations in the course, and the four-quadrant plots do not display asymmetries.

The bootstrap confidence intervals are wide. The width is around 0.1 for the ATC ratio, while it gets up to 0.16 for the negative ATC ratio for systolic measurement and the horizon of one minute. Thus, the confidence intervals cover 0.5 for systolic measurement and the horizon of one minute, and the equality to random guessing cannot be excluded.

4 Discussion and conclusion

In this paper, we examine various methods to assess the ability to track changes (ATC) for measurements, nowcasts, or forecasts, that is, whether they correctly predict the direction of changes in values. While the computation of predicted change varies between the application areas of measurement, nowcasting, and forecasting, the assessment can be based on the same methods. The ATC assessment can accompany other evaluation techniques, such as measures of deviation or probabilistic scoring rules.

Four-quadrant plots facilitate the visual inspection of the ATC for a measurement, nowcast, or forecast (see Section 2.2). The ATC ratio, the ratio of change directions predicted correctly over the total number of changes, numerically evaluates ATC. Visually, it is the proportion of concordant points in a four-quadrant plot (see Section 2.3). The positive and negative ATC ratios analyze the ATC ratio given whether the predicted change is positive or negative, respectively. Thus, they quantify the credibility of the respective predictions. The applications of Section 3.3 show that models, in general, indeed have different positive and negative ATC and that they add valuable information to the ATC ratio. In the applications, the bootstrap confidence intervals of Section 2.3 are used to quantify the estimation uncertainty of the ATC measures. The width of the confidence intervals is around 0.1 for around 100 samples, while it is around 0.01 for 8000 samples. For models with reasonably high ATC, 100 samples are thus sufficient to differentiate from random guessing or to assess models with high ATC differences.

A conditional ATC plot visualizes the probability of correct change direction prediction over the predicted change of the measurement, nowcast, or forecast (see Section 2.5). It is based on a multivariate kernel density estimation (KDE) of predicted and observed change. In the application, the conditional ATC plot gives reasonable insights into the local effects of the ATC. Section 2.6 adapts measures of probabilistic forecast evaluation to the ATC assessment of probabilistic forecasts and nowcasts. The Brier score (BS) as numerical assessment of probabilistic ATC is introduced, and reliability diagrams are used to visualize the local ATC of probabilistic forecasts.

The methods of ATC assessment are applied to COVID-19-nowcasting, forecasting the patient admissions to an emergency department, and invasive and non-invasive blood pressure measurements in Section 3. While ATC assessment should not be the only aspect, it is a valuable addition to evaluating nowcasts, forecasts, and measurements. Models with highly different accuracies are usually scored similarly in ATC assessment, but ATC assessment can differentiate between models with similar accuracies. As in the application in Section 3.1, models with medial point forecast evaluation measures can have the most meaningful positive ATC.

For public health officials and forecasters analyzing methods, we recommend presenting four-quadrant plots, ATC ratios and the conditional ATC plots as a fundamental approach. This presentation makes a model's ATC easily interpretable and and reveals performance asymmetries. An exclusion area can be incorporated to address for non-relevant changes.

We did not expand on two modeling aspects throughout this paper, which we leave for further research. In the estimation, we did not consider sequential correlation. The computation of differences is a standard procedure in time series analysis to remove sequential dependence, but, in general, some could remain, and the estimators could account for it. Similarly, the bootstrap confidence intervals could be adapted to consider sequential correlation using time-series bootstrap methods [21, 29].

The estimators of Section 2.3 do not account for imbalances in the number of observed positive and negative changes [for theoretical analysis, see 27, Chapter 3]. Significant differences in the number of observed positive and negative changes are unlikely in the ATC setting, as $\mathbf{y}^{\Delta}$ is obtained from differencing time series data and occur, for example, if the true value contains a few high jumps in one direction and many smaller jumps in the other. However, if the number of positive and negative observed changes differs widely, unbalanced-data-aware measures should be considered. There are various adapted measures for unbalanced outcomes, for example, Cohen's κ [8] or those listed in Jolliffe and Stephenson [27, Table 3.3].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the study are available in the corresponding databases or repositories (https://covid19nowcasthub.de/, https://github.com/bahmanrostamitabar/hourly-emergency-care, and https://physionet.org/content/mimic3wdb/1.0/). All code is available in the repository https://github.com/jo-rie/aatc.

Competing interests

The authors declare that they have no competing interests.

Funding

JR gratefully acknowledges funding from the Bischoefliche Studienstiftung Cusanuswerk (Bonn, Germany).

Authors' contributions

JR, BL, and OG developed the theoretical formalism. JR and BL implemented the methods. JR conducted the simulations and analyzed the data. All authors interpreted the data. BS and OG substantively revised the manuscript. JR and BL drafted the manuscript. All authors have approved the submitted version and have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Acknowledgments

Not applicable.

References

[1] Marta Bańbura et al. "Now-Casting and the Real-Time Data Flow". In: *Handbook of Economic Forecasting.* Vol. 2. Elsevier, 2013, pp. 195–237. ISBN: 978-0-444-53683-9. DOI: 10.1016/B978-0-444-53683-9.00004-9.

[2] Felix Bittmann. *Bootstrapping: An Integrated Approach with Python and Stata.* Berlin Boston: De Gruyter Oldenbourg, 2021. 120 pp. ISBN: 978-3-11-069440-6.

[3] Martin J. Bland and Douglas G. Altman. "Statistical methods for assessing agreement between two methods of clinical measurement". In: *The Lancet* 327.8476 (Feb. 1986), pp. 307–310. ISSN: 01406736. DOI: 10.1016/S0140-6736(86)90837-8.

[4] Johannes Bracher et al. "Direction Augmentation in the Evaluation of Armed Conflict Predictions". In: *International Interactions* 49.6 (Nov. 2, 2023), pp. 989–1004. ISSN: 0305-0629, 1547-7444. DOI: 10.1080/03050629.2023.2255923.

[5] Glenn W. Brier. "Verification of forecasts expressed in terms of probability". In: *Monthly Weather Review* 78.1 (Jan. 1950), pp. 1–3. ISSN: 0027-0644, 1520-0493. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[6] K. A. Browning and C. G. Collier. "Nowcasting of Precipitation Systems". In: *Reviews of Geophysics* 27.3 (Aug. 1989), pp. 345–370. ISSN: 8755-1209, 1944-9208. DOI: 10.1029/RG027i003p00345.

[7] Jennifer Castle, David Hendry, and Oleg Kitov. "Forecasting and Nowcasting Macroeconomic Variables: A Methodological Overview". In: *Handbook on Rapid Estimates: 2017 Edition.* Ed. by European Commission. Statistical Office of the European Union. Luxembourg: Publications Office, 2017. URL: https://data.europa.eu/doi/10.2785/488740 (visited on 04/09/2024).

[8] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (Apr. 1960), pp. 37–46. ISSN: 0013-1644, 1552-3888. DOI: 10.1177/001316446002000104.

[9] Lester A. Critchley, Anna Lee, and Anthony M.-H. Ho. "A Critical Review of the Ability of Continuous Cardiac Output Monitors to Measure Trends in Cardiac Output". In: *Anesthesia & Analgesia* 111.5 (Nov. 2010), pp. 1180–1192. ISSN: 0003-2999. DOI: 10.1213/ANE.0b013e3181f08a5b.

[10] Lester A. Critchley, Anna Lee, and Anthony M.-H. Ho. "A critical review of the ability of continuous cardiac output monitors to measure trends in cardiac output". In: *Anesthesia and Analgesia* 111.5 (2010). Cited by: 421; All Open Access, Bronze Open Access, 1180 ? 1192. DOI: 10.1213/ANE.0b013e3181f08a5b.

[11] Lester A. Critchley, Xiao X. Yang, and Anna Lee. "Assessment of Trending Ability of Cardiac Output Monitors by Polar Plot Methodology". In: *Journal of Cardiothoracic and Vascular Anesthesia* 25.3 (2011), pp. 536–546. ISSN: 1053-0770. DOI: https://doi.org/10.1053/j.jvca.2011.01.003. URL: https://www.sciencedirect.com/science/article/pii/S1053077011000048.

[12] Lester A. H. Critchley and Julian A. J. H. Critchley. “A Meta-Analysis of Studies Using Bias and Precision Statistics to Compare Cardiac Output Measurement Techniques”. In: *Journal of Clinical Monitoring and Computing* 15.2 (Feb. 1, 1999), pp. 85–91. ISSN: 1573-2614. DOI: 10.1023/A:1009982611386.

[13] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge ; New York, NY, USA: Cambridge University Press, 1997. 582 pp. ISBN: 978-0-521-57391-7.

[14] Chair of Econometrics and Statistics at Karlsruhe Institute of Technology and Computational Statistics Group at Heidelberg Institute for Theoretical Studies. *Nowcasts Der COVID-19 Hospitalisierungsinzidenz*. Feb. 23, 2024. URL: https://covid19nowcasthub.de/ (visited on 02/23/2024).

[15] Bradley Efron. “Better Bootstrap Confidence Intervals”. In: *Journal of the American Statistical Association* 82.397 (Mar. 1987), pp. 171–185. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1987.10478410.

[16] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. “Probabilistic Forecasts, Calibration and Sharpness”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 69.2 (2007), pp. 243–268. ISSN: 13697412. DOI: 10.1111/j.1467-9868.2007.00587.x.

[17] Ary L. Goldberger et al. “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals”. In: *Circulation* 101.23 (June 13, 2000). ISSN: 0009-7322, 1524-4539. DOI: 10.1161/01.CIR.101.23.e215.

[18] Mithat Gönen and Glenn Heller. “Concordance Probability and Discriminatory Power in Proportional Hazards Regression”. In: *Biometrika* 92.4 (2005), pp. 965–970. ISSN: 00063444. URL: http://www.jstor.org/stable/20441249 (visited on 10/19/2024).

[19] Artur Gramacki. *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Vol. 37. Studies in Big Data. Cham: Springer International Publishing, 2018. ISBN: 978-3-319-71688-6. DOI: 10.1007/978-3-319-71688-6.

[20] Felix Günther et al. “Nowcasting the COVID-19 Pandemic in Bavaria”. In: *Biometrical Journal* 63.3 (Mar. 2021), pp. 490–502. ISSN: 0323-3847, 1521-4036. DOI: 10.1002/bimj.202000112.

[21] Wolfgang Härdle, Joel Horowitz, and Jens-Peter Kreiss. “Bootstrap Methods for Time Series”. In: *International Statistical Review* 71.2 (Aug. 2003), pp. 435–459. ISSN: 0306-7734, 1751-5823. DOI: 10.1111/j.1751-5823.2003.tb00485.x.

[22] Anna Haugsbo Hermansen and Ole Jakob Mengshoel. “Forecasting Ambulance Demand Using Machine Learning: A Case Study from Oslo, Norway”. In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2021 IEEE Symposium Series on Computational Intelligence (SSCI). Orlando, FL, USA: IEEE, Dec. 5, 2021, pp. 01–10. ISBN: 978-1-72819-048-8. DOI: 10.1109/SSCI50451.2021.9659837.

[23] Tim Hesterberg. “Bootstrap”. In: *WIREs Computational Statistics* 3.6 (Nov. 2011), pp. 497–526. ISSN: 1939-5108, 1939-0068. DOI: 10.1002/wics.182.

[24] Mayu Hiraishi, Kensuke Tanioka, and Toshio Shimokawa. “Concordance Rate of a Four-Quadrant Plot for Repeated Measurements”. In: *BMC Medical Research*

Methodology 21.1 (Dec. 2021), p. 270. ISSN: 1471-2288. DOI: 10.1186/s12874-021-01461-0.

[25] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. New York, NY: Springer US, 2021. ISBN: 978-1-07-161418-1. DOI: 10.1007/978-1-0716-1418-1.

[26] Alistair E.W. Johnson et al. "MIMIC-III, a Freely Accessible Critical Care Database". In: *Scientific Data* 3.1 (May 24, 2016), p. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35.

[27] Ian T. Jolliffe and David B. Stephenson, eds. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2. ed. Oxford: Wiley-Blackwell, 2012. 274 pp. ISBN: 978-0-470-66071-3.

[28] Spencer S. Jones et al. "Forecasting Daily Patient Volumes in the Emergency Department". In: *Academic Emergency Medicine* 15.2 (Feb. 2008), pp. 159–170. ISSN: 1069-6563, 1553-2712. DOI: 10.1111/j.1553-2712.2007.00032.x.

[29] Jens-Peter Kreiss and Soumendra Nath Lahiri. "Bootstrap Methods for Time Series". In: *Handbook of Statistics*. Vol. 30. Elsevier, 2012, pp. 3–26. ISBN: 978-0-444-53858-1. DOI: 10.1016/B978-0-444-53858-1.00001-6.

[30] James Mitchell and Kenneth F. Wallis. "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness". In: *Journal of Applied Econometrics* 26.6 (2011), pp. 1023–1040. ISSN: 08837252. DOI: 10.1002/jae.1192.

[31] Benjamin Moody et al. *MIMIC-III Waveform Database*. Version 1.0. PhysioNet, 2017. DOI: 10.13026/C2607M.

[32] Albert C. Perrino, Stephen N. Harris, and Martha A. Luther. "Intraoperative Determination of Cardiac Output Using Multiplane Transesophageal Echocardiography". In: *Anesthesiology* 89.2 (Aug. 1, 1998), pp. 350–357. ISSN: 0003-3022. DOI: 10.1097/00000542-199808000-00010.

[33] Roopesh Ranjan and Tilmann Gneiting. "Combining Probability Forecasts". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 72.1 (2010), pp. 71–91. ISSN: 13697412. DOI: 10.1111/j.1467-9868.2009.00726.x.

[34] Robert Koch Institute. *Aktualisierung Der ControlCOVID-Strategie Zur Vorbereitung Auf Den Herbst/Winter 2021/22*. Sept. 14, 2021. URL: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Downloads/control-covid-2021-09-22.pdf?__blob=publicationFile (visited on 02/23/2024).

[35] Robert Koch Institute. *RKI - Coronavirus SARS-CoV-2 - Antworten Auf Häufig Gestellte Fragen (FAQ) Zu Akuten Atemwegserkrankungen Und COVID-19*. Feb. 21, 2024. URL: https://www.rki.de/SharedDocs/FAQ/COVID-19/FAQ-gesamt.html (visited on 02/23/2024).

[36] Bahman Rostami-Tabar, Jethro Browell, and Ivan Svetunkov. "Probabilistic Forecasting of Hourly Emergency Department Arrivals". In: *Health Systems* (May 1, 2023), pp. 1–17. ISSN: 2047-6965, 2047-6973. DOI: 10.1080/20476965.2023.2200526.

[37] Bernd Saugel, Oliver Grothe, and Julia Y. Nicklas. "Error Grid Analysis for Arterial Pressure Method Comparison Studies". In: *Anesthesia & Analgesia* 126.4 (Apr. 2018), pp. 1177–1185. ISSN: 0003-2999. DOI: 10.1213/ANE.0000000000002585.

[38] Bernd Saugel, Oliver Grothe, and Julia Y. Wagner. “Tracking Changes in Cardiac Output: Statistical Considerations on the 4-Quadrant Plot and the Polar Plot Methodology”. In: *Anesthesia & Analgesia* 121.2 (Aug. 2015), pp. 514–524. ISSN: 0003-2999. DOI: 10.1213/ANE.0000000000000725.

[39] Skipper Seabold and Josef Perktold. “Statsmodels: Econometric and Statistical Modeling with Python”. In: *9th Python in Science Conference*. 2010.

[40] Pierre Squara et al. “Metrology Part 1: Definition of Quality Criteria”. In: *Journal of Clinical Monitoring and Computing* 35.1 (Feb. 2021), pp. 17–25. ISSN: 1387-1307, 1573-2614. DOI: 10.1007/s10877-020-00494-y.

[41] Pierre Squara et al. “Metrology Part 2: Procedures for the Validation of Major Measurement Quality Criteria and Measuring Instrument Properties”. In: *Journal of Clinical Monitoring and Computing* 35.1 (Feb. 2021), pp. 27–37. ISSN: 1387-1307, 1573-2614. DOI: 10.1007/s10877-020-00495-x.

[42] Ewout W. Steyerberg et al. “Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures”. eng. In: *Epidemiology (Cambridge, Mass.)* 21.1 (2010), pp. 128–138. ISSN: 1044-3983.

[43] Kristin Tolksdorf, Anna Loenenbach, and Silke Buda. “Dritte Aktualisierung der „Retrospektiven Phaseneinteilung der COVID-19-Pandemie in Deutschland“”. In: *Epidemiologisches Bulletin* 2022.38 (Sept. 22, 2022), pp. 3–6. DOI: 10.25646/10598.

[44] Stef Van Buuren. *Flexible Imputation of Missing Data, Second Edition*. 2nd ed. Second edition. | Boca Raton, Florida : CRC Press, [2019] |: Chapman and Hall/CRC, July 17, 2018. ISBN: 978-0-429-49225-9. DOI: 10.1201/9780429492259.

[45] Paola Vesco et al. “United They Stand: Findings from an Escalation Prediction Competition”. In: *International Interactions* 48.4 (July 4, 2022), pp. 860–896. ISSN: 0305-0629, 1547-7444. DOI: 10.1080/03050629.2022.2029856.

[46] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[47] Daniel Wolffram et al. “Collaborative Nowcasting of COVID-19 Hospitalization Incidences in Germany”. In: *PLOS Computational Biology* 19.8 (Aug. 11, 2023). Ed. by James M McCaw, e1011394. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011394.

[48] World Meteorological Organization (WMO). *Guidelines for Nowcasting Techniques*. Geneva: WMO, 2017. 82 pp. ISBN: 978-92-63-11198-2. URL: https://library.wmo.int/idurl/4/55666.

[49] Joseph T. Wu et al. “Nowcasting Epidemics of Novel Pathogens: Lessons from COVID-19”. In: *Nature Medicine* 27.3 (Mar. 2021), pp. 388–395. ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-021-01278-w.

A Additional material on Section 2

A.1 Data generation for Section 2

The first dataset is generated by sequentially generating $\mathbf{x}^{\Delta}$ and $\mathbf{y}^{\Delta}$. First, the x_t^{Δ} are sampled as a sum of a standard normal random number and a uniform random number on $(-10, 10)$:

$$x_t^{\Delta} \sim N(0,1) + U(-10,10) \quad t = 1, \ldots, T.$$

Subsequently, the $\mathbf{y}^{\Delta}$ are simulated for a constant ATC ratio k by

$$y_t^{\Delta} = x_t^{\Delta} \cdot n_t \cdot b_t,$$

where n_t is a truncated normal distribution with mean one and standard deviation 0.5, truncated at 0, and b_t is a symmetric Bernoulli random variable with parameter k. For a time-varying ATC ratio, the parameter k is modified to have a wave-shape function over time, that is,

$$k_t = 0.75 + \sin(t/365.25 \cdot 2\pi)/4.$$

For the asymmetric ATC ratio, k is a function of x_t^{Δ},

$$k(x) = 0.5 + \min\left\{\max\left\{\frac{x+5}{10}, 0\right\}, 1\right\}/2.$$

In the second approach, y_t^{Δ} and x_t^{Δ} are modeled to be multivariate normal with mean 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}.$$

Thus, the conditional probability of correct direction prediction can be calculated by a conditional normal distribution to

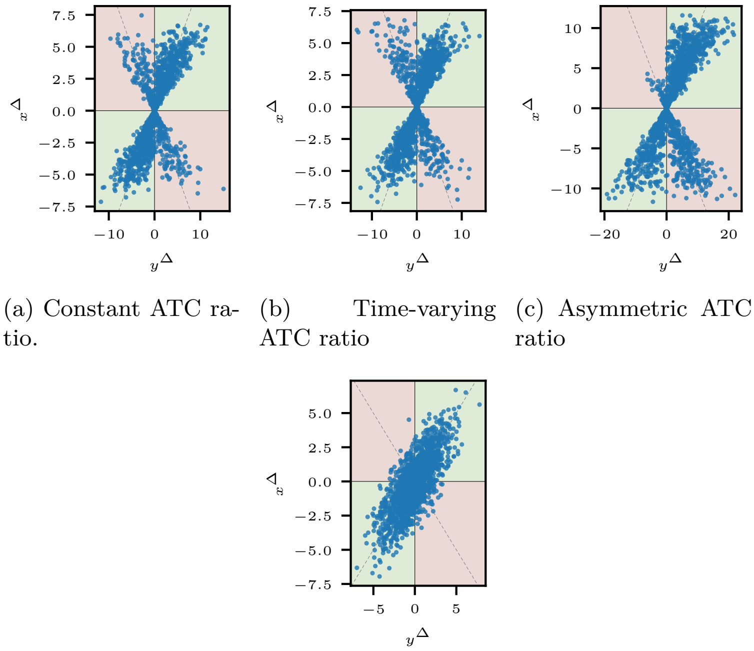
$$P(Y^{\Delta} X^{\Delta} > 0 | X^{\Delta} = x) = \Phi\left(\frac{3}{2\sqrt{7}} x\right),$$

where Φ is a standard normal CDF.

The four-quadrant plots for the sample realizations of the data generation schemes are shown in Figure 9.

A.2 Simulation study on bootstrapping confidence intervals

We examine three methods for bootstrapping for computing confidence intervals for the ATC ratio: the intuitive percentile and the more sophisticated basic and BCa method. In the *percentile* approach, the confidence interval for the level α is built directly from the empirical distribution of the bootstrap estimators. The *basic* approach computes the confidence interval based on the non-bootstrap estimate using the bootstrapped quantile deviations [13]. The BCa method modifies the quantiles of the empirical bootstrap distribution by a bias and an acceleration parameter [15]. Typically, the percentile approach requires larger datasets and provides an easy and fast estimate, while the BCa, though computationally expensive, can produce reasonable confidence intervals with smaller datasets. The basic approach balances these



(a) Constant ATC ratio.

(b) Time-varying ATC ratio

(c) Asymmetric ATC ratio

(d) Second approach.

Figure 9: Four-quadrant plots for sample realizations of the data generation schemes of Section A.1. Although the first and second plots differ over time, their difference is not discernible in the plots. The third data set's asymmetry is visible in the plot, but the decrease in the ATC near 0 is not visible.

two objectives. We compare the approaches in a small synthetic data study on their small-dataset behavior and computation time.

We vary the number of time steps T to take typical time-series values, such as 30 for daily data in a month, 52 for weekly data, 168, 365, 720, and 1024. The considered datasets are the first dataset with asymmetric dependence and the second dataset outlined in Appendix A.1. In the calculations, the `scipy` package's implementation of bootstrap confidence intervals is used [46]. The prescribed confidence level is 90 %, and the number of bootstrap samples is $10,000$. The share of confidence intervals covering the true values per method and T are shown in Table 8. The true values of the accuracy are computed based on a dataset of size 10^8, yielding 0.7501 and 0.7700 for the two datasets. The computation times per method and dataset are shown in Figure 10. For the small sample sizes up to $T = 168$, only the BCa method keeps the confidence interval size and yields slightly wider confidence intervals. The method's results are similar for the larger sample sizes. The computation time for the BCa method is slightly larger than for the other methods, but all methods have a moderate computation time. BCa is the only method that maintains the confidence level for small datasets while increasing the computation time only moderately for

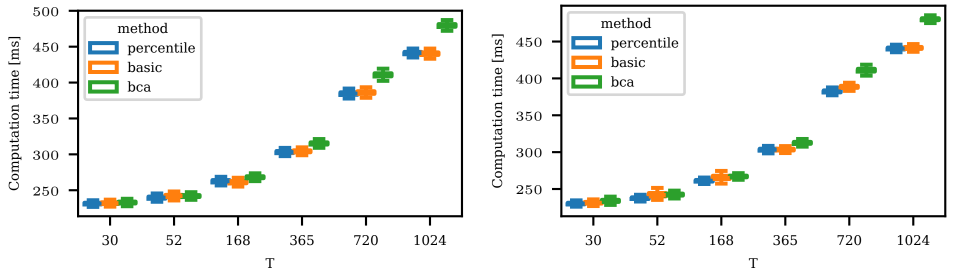
	percentile	basic	BCa
30	0.84 (0.249)	0.86 (0.250)	0.91 (0.257)
52	0.89 (0.194)	0.89 (0.193)	0.89 (0.198)
168	0.91 (0.109)	0.90 (0.109)	0.90 (0.110)
365	0.90 (0.074)	0.90 (0.074)	0.90 (0.074)
720	0.90 (0.053)	0.90 (0.053)	0.90 (0.053)
1024	0.90 (0.044)	0.90 (0.044)	0.89 (0.044)

(a) First dataset.

	percentile	basic	BCa
30	0.87 (0.243)	0.88 (0.242)	0.92 (0.249)
52	0.87 (0.188)	0.89 (0.188)	0.90 (0.192)
168	0.89 (0.106)	0.90 (0.106)	0.90 (0.107)
365	0.90 (0.072)	0.90 (0.072)	0.90 (0.072)
720	0.90 (0.052)	0.90 (0.052)	0.90 (0.052)
1024	0.89 (0.043)	0.90 (0.043)	0.90 (0.043)

(b) Second dataset.

Table 8: Proportion of bootstrap confidence intervals covering the true value of ATC ratio per method and sample size T. The average width of the confidence interval is listed in brackets.



(a) First dataset with asymmetry.

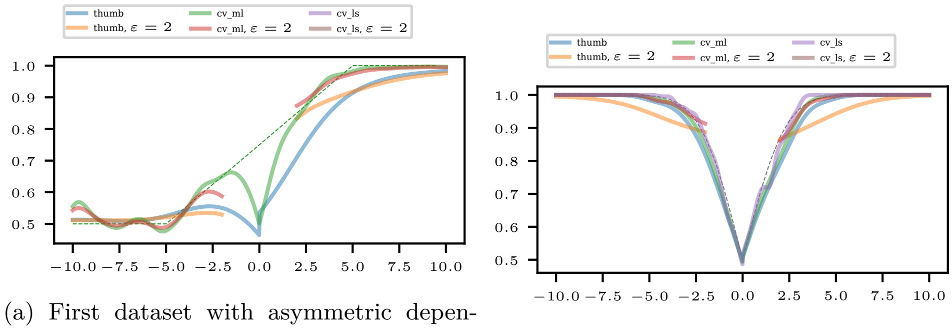
(b) Second dataset.

Figure 10: Boxplot of the computation time for different bootstrapping methods and data set sizes T. The computation time refers to bootstrapping one confidence interval based upon $10,000$ values. Each boxplot reflects $10,000$ samples. The BCa method takes slightly longer than the other two, but the difference is negligible.

larger datasets. Therefore, we use the BCa method for confidence intervals in the applications in Section 3.

A.3 Visualization of different bandwidth selectors in multivariate KDE

We examine the resulting conditional ATC plots for the three well-known KDE bandwidth selectors, rule-of-thumb, cross-validation maximum likelihood, and cross-validation least squares using the `statsmodels` Python package [39] in Figure 11. While the rule-of-thumb is based only on the covariance matrix, the other two numerically optimize the bandwidth with a hold-one-out least squares or likelihood objective function. The dashed line shows the theoretical $P(Y^\Delta X^\Delta > 0|X^\Delta = \chi)$. The second method, cross-validation least squares, requires long computation times while yielding small or no bandwidth results, even for two relatively small datasets. The rule-of-thumb and cross-validation maximum likelihood methods yield reasonable results at moderate computation times.



(a) First dataset with asymmetric dependence.

(b) Second dataset.

Figure 11: Conditional ATC plot for different bandwidth selection processes. Cross-validation least squares takes a considerably larger computation time. It converges neither for the first nor the second data set with an exclusion area and yields a bandwidth too small for the second data set. The rule of thumb is the fastest method but tends to oversmooth. The cross-validation maximum likelihood method yields a more reasonable bandwidth with moderate computation time. ε specifies an exclusion area $E = \{(x, y) \in \mathbb{R}^2 : (-\varepsilon \leqslant x \leqslant \varepsilon)\}$ in $\mathbf{x}^\Delta$-direction.

A.4 Probabilistic ATC evaluation

Section 2.6 outlines the assessment of probabilistic ATC for nowcasts and forecasts and specifies the computation for predictions in terms of a CDF and known true values. Here, we outline the computation for quantile forecasts and yet unknown, probabilistic true values.

If forecasts or nowcasts are given as quantiles, p_t can be determined by interpolations among the quantiles. Let q_p denote the quantiles for target time $t + l$ for even-spaced probabilities $p \in \{1/\hat{p}, \ldots, (p-1)/\hat{p}\}$ $(\hat{p} \in \mathbb{N}\backslash\{1, 2\})$ and y_t the true value at time t. The quantiles q_p generally differ for each time step, but we omit an index here for ease of notation. The probability p_t^c of a *negative* change is between $p^\star$ and $p^\star + 1/\hat{p}$ for

$$p^\star = \max\{p \in \{1/\hat{p}, \ldots, (\hat{p}-1)/\hat{p}\} : q_p \leqslant y_t\}, \quad \text{if } q_{1/\hat{p}} \leqslant y_t \leqslant q_{1-1/\hat{p}}.$$

Quantiles do not determine the location within the interval $[p^\star, p^\star + 1/\hat{p}]$. Under the assumption of a uniform distribution within the quantile interval, the probability of a negative change is

$$p_t^c = \frac{y_t - q_{p^\star}}{\hat{p}(q_{p^\star+1} - q_{p^\star})} + p^\star.$$

The approach does not yet assign probabilities for y_t smaller than the smallest quantile $q_{1/p}$ or greater than the largest quantile. As a simple extension, we assume that the probability mass is uniformly distributed on an interval of the same length as the nearest interval specified by the quantiles. This yields

$$p_t^c = \begin{cases} \max\{\frac{1}{\hat{p}} - \frac{q_{p^\star} - y_t}{\hat{p}(q_{p2/\hat{p}} - q_{1/\hat{p}})}, 0\} & , \text{if } y_t < q_{1/p}, \\ \min\{\frac{1}{\hat{p}} - \frac{y_t - q_{(\hat{p}-1)/\hat{p}}}{\hat{p}(q_{(\hat{p}-1)/\hat{p}} - q_{(\hat{p}-2)/\hat{p}})}, 1\} & , \text{if } y_t > q_{1-1/p}, \\ \frac{y_t - q_{p^\star}}{\hat{p}(q_{p^\star+1} - q_{p^\star})} + p^\star & , \text{otherwise.} \end{cases}$$

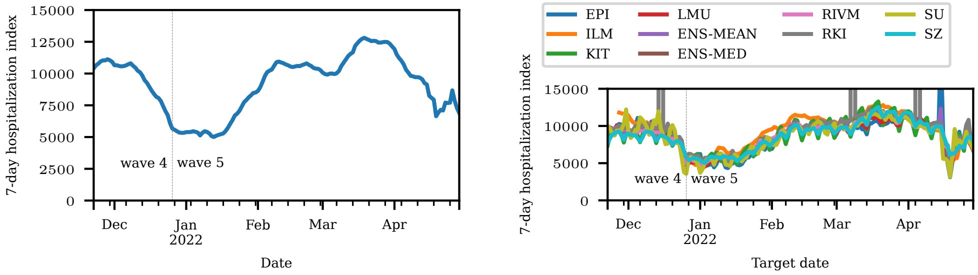
The probability of positive change is $p_t = 1 - p_t^c$.

If the true value is given as a distribution because it is still unknown, the probabilities p_t can be computed by integration. Let for two nowcasts the distributions be given by PDFs $f_{t+l|t+l}$ and $f_{t|t+l}$ with CDFs $F_{t+l|t+l}$ and $F_{t|t+l}$. Then, the probability of a negative change can be computed by

$$\begin{aligned} p_t^c &= \int_{\substack{x_1, x_2 \in \mathbb{R}: \\ x_2 < x_1}} f_{t|t+l}(x_1) f_{t+l|t+l}(x_2) \, \mathrm{d}\,(x_1, x_2) \\ &= \int_{x_1 \in \mathbb{R}} \int_{-\infty}^{x_1} f_{t|t+l}(x_1) f_{t+l|t+l}(x_2) \, \mathrm{d}\,x_2 \, \mathrm{d}\,x_1 \\ &= \int_{x_1 \in \mathbb{R}} f_{t|t+l}(x_1) F_{t+l|t+l}(x_1) \, \mathrm{d}\,x_2 \, \mathrm{d}\,x_1. \end{aligned} \tag{11}$$

Thereby, the distributions are assumed to be independent. If the nowcasts have the form of a multivariate distribution, including the dependence of the two PDFs, $f_{t+l|t+l}(x_2)$ has to be replaced by the PDF conditional on x_1. As a Monte Carlo approximation of Equation (11), the probability can also be calculated by sampling from $f_{t+l|t+l}$ and $f_{t|t+l}$ and calculating the fraction of negative changes. For forecasts, the indexes have to be shifted. If no PDFs are available, they can be estimated from the CDF or quantiles, or the CDF or quantiles can be used to generate samples for the Monte Carlo approximation. This approach is applied in Section 3.1, as the true values are published with a delay of more than 80 days, and the nowcasts are given as quantiles.

B Additional results for Section 3.1



(a) Realisations.

(b) Same-day nowcasts.

Figure 12: True and nowcast data of the seven-day-hospitalization in Germany from November 22, 2021, to April 29, 2022 [14]. The outliers in the RKI model of values above 10^8 are removed before the following analysis.

Abbreviation	Nowcasting hub key
EPI	Epiforecasts-independent
ILM	ILM-prop
KIT	KIT-simple_nowcast
LMU	LMU_StaBLab-GAM_nowcast
RIVM	RIVM-KEW
RKI	RKI-weekly_report
SU	SU-hier_bayes
SZ	SZ-hosp_nowcast
ENS-MEAN	NowcastHub-MeanEnsemble
ENS-MED	NowcastHub-MedianEnsemble

Table 9: Matching the abbreviation to the key in the COVID-19 nowcasting hub. Information on the models and references are listed in Wolffram et al. [47, Table 1].

	(1), l=1	$\sigma_{x^\Delta,1}$	$q_{0.1}(x^{\Delta,1})$	(1), l=7	$\sigma_{x^\Delta,7}$	$q_{0.1}(x^{\Delta,7})$	(1), l=14	$\sigma_{x^\Delta,14}$	$q_{0.1}(x^{\Delta,14})$
EPI	86	520	44	83	1411	78	80	1976	144
ILM	86	281	26	81	1457	102	82	2356	140
KIT	84	354	50	89	1306	171	83	1964	265
LMU	65	285	26	84	1180	124	78	1946	167
ENS-MEAN	85	267	23	86	1213	98	83	1955	235
ENS-MED	88	259	23	88	1206	101	81	1955	186
RIVM	77	241	32	81	1264	123	77	2034	190
RKI	99	362	34	106	1194	145	99	1832	325
SU	91	376	47	85	1390	180	80	2126	263
SZ	92	201	26	89	1154	184	87	1889	241
True	75	262	27	66	1237	126	73	2193	284

Table 10: Marginal analysis of the COVID-19 nowcast and true changes for the horizons one, seven, and 14 days. The column (1), $l = l$ shows the number of values greater than zero for horizon l, $\sigma_{x^\Delta,l}$ the standard deviation, and $q_{0.1}(x^{\Delta,l})$ the 10% quantile of the changes' absolute values.

	μ^{1}	$\mu^{+,1}$	$\mu^{-,1}$	$\mu^{1}_{q_{0.1}}$	$\mu^{+,1}_{q_{0.1}}$	$\mu^{-,1}_{q_{0.1}}$
EPI	0.68 (0.62, 0.74)	0.64 (0.55, 0.72)	0.73 (0.63, 0.81)	0.69 (0.63, 0.75)	0.64 (0.56, 0.73)	0.75 (0.65, 0.82)
ILM	0.73 (0.67, 0.79)	0.67 (0.58, 0.76)	0.82 (0.73, 0.89)	0.74 (0.68, 0.79)	0.68 (0.60, 0.76)	0.82 (0.72, 0.89)
KIT	0.62 (0.55, 0.68)	0.58 (0.49, 0.67)	0.65 (0.56, 0.75)	0.62 (0.56, 0.69)	0.59 (0.51, 0.67)	0.66 (0.57, 0.74)
LMU	0.66 (0.60, 0.72)	0.66 (0.57, 0.75)	0.66 (0.57, 0.73)	0.66 (0.59, 0.72)	0.66 (0.55, 0.75)	0.66 (0.57, 0.73)
ENS-MEAN	0.81 (0.75, 0.85)	0.76 (0.68, 0.84)	0.88 (0.81, 0.93)	0.81 (0.75, 0.86)	0.76 (0.68, 0.83)	0.88 (0.81, 0.94)
ENS-MED	0.75 (0.68, 0.80)	0.69 (0.60, 0.77)	0.81 (0.73, 0.89)	0.75 (0.69, 0.80)	0.69 (0.61, 0.77)	0.83 (0.74, 0.90)
RIVM	0.77 (0.72, 0.82)	0.75 (0.66, 0.83)	0.79 (0.71, 0.85)	0.78 (0.72, 0.83)	0.75 (0.66, 0.83)	0.81 (0.73, 0.87)
RKI	0.74 (0.68, 0.80)	0.67 (0.59, 0.75)	0.88 (0.79, 0.93)	0.74 (0.67, 0.79)	0.66 (0.58, 0.73)	0.87 (0.78, 0.93)
SU	0.71 (0.65, 0.77)	0.66 (0.57, 0.74)	0.78 (0.69, 0.85)	0.72 (0.66, 0.78)	0.67 (0.58, 0.75)	0.79 (0.70, 0.87)
SZ	0.74 (0.69, 0.80)	0.68 (0.60, 0.76)	0.82 (0.73, 0.88)	0.74 (0.69, 0.80)	0.68 (0.60, 0.76)	0.82 (0.73, 0.88)

(a) One day.

	μ^{14}	$\mu^{+,14}$	$\mu^{-,14}$	$\mu^{14}_{q_{0.1}}$	$\mu^{+,14}_{q_{0.1}}$	$\mu^{-,14}_{q_{0.1}}$
EPI	0.83 (0.77, 0.87)	0.79 (0.70, 0.85)	0.87 (0.80, 0.92)	0.85 (0.80, 0.90)	0.81 (0.73, 0.87)	0.90 (0.83, 0.95)
ILM	0.86 (0.81, 0.90)	0.78 (0.70, 0.85)	0.96 (0.90, 0.99)	0.87 (0.82, 0.91)	0.80 (0.71, 0.86)	0.96 (0.90, 0.99)
KIT	0.81 (0.75, 0.86)	0.76 (0.67, 0.83)	0.87 (0.79, 0.92)	0.82 (0.76, 0.86)	0.76 (0.68, 0.84)	0.88 (0.81, 0.93)
LMU	0.88 (0.83, 0.92)	0.85 (0.77, 0.91)	0.91 (0.85, 0.95)	0.89 (0.85, 0.93)	0.87 (0.79, 0.92)	0.91 (0.85, 0.95)
ENS-MEAN	0.83 (0.77, 0.87)	0.77 (0.69, 0.84)	0.89 (0.83, 0.95)	0.84 (0.78, 0.88)	0.78 (0.70, 0.85)	0.91 (0.84, 0.95)
ENS-MED	0.84 (0.79, 0.89)	0.79 (0.70, 0.85)	0.90 (0.83, 0.95)	0.85 (0.80, 0.90)	0.80 (0.72, 0.86)	0.91 (0.84, 0.96)
RIVM	0.85 (0.80, 0.89)	0.82 (0.74, 0.88)	0.88 (0.80, 0.93)	0.85 (0.80, 0.90)	0.83 (0.75, 0.89)	0.88 (0.80, 0.93)
RKI	0.81 (0.75, 0.86)	0.71 (0.63, 0.77)	0.98 (0.93, 1.00)	0.81 (0.75, 0.86)	0.71 (0.63, 0.78)	1.00 (nan, nan)
SU	0.88 (0.83, 0.92)	0.84 (0.76, 0.90)	0.92 (0.86, 0.96)	0.89 (0.84, 0.93)	0.85 (0.77, 0.91)	0.94 (0.87, 0.97)
SZ	0.82 (0.77, 0.87)	0.76 (0.68, 0.83)	0.90 (0.83, 0.94)	0.83 (0.78, 0.88)	0.78 (0.69, 0.85)	0.90 (0.83, 0.94)

(b) 14 days.

Table 11: ATC ratio μ, positive ATC ratio μ^{+}, and negative ATC ratio μ^{-} for the models without and with exclusion areas for the horizon one and 14 days in COVID-19 nowcasting. The exclusion areas are rectangles centered on the zero points with a width and height to exclude the 10%-quantile of the absolute values of nowcast or true values. There are 75 positive and 84 negative actual changes in the considered time period for horizon one day and 73 and 86 for the horizon of 14 days.

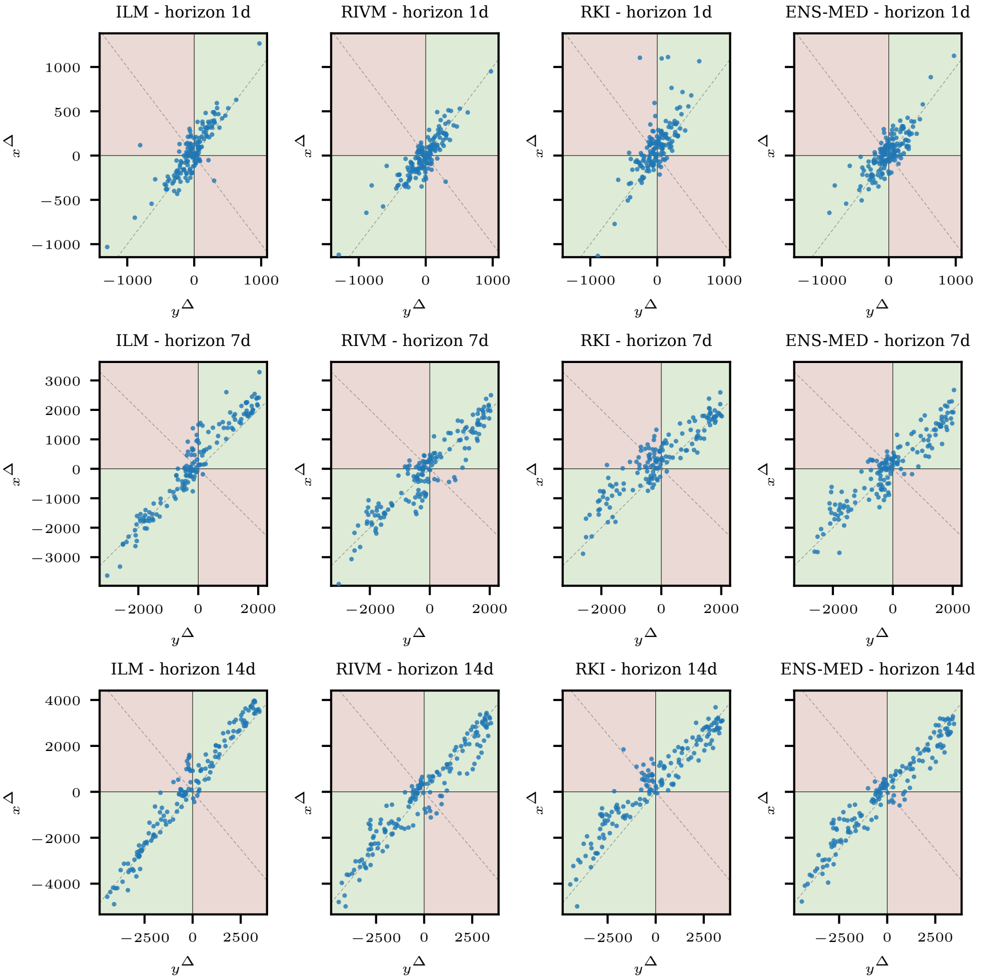
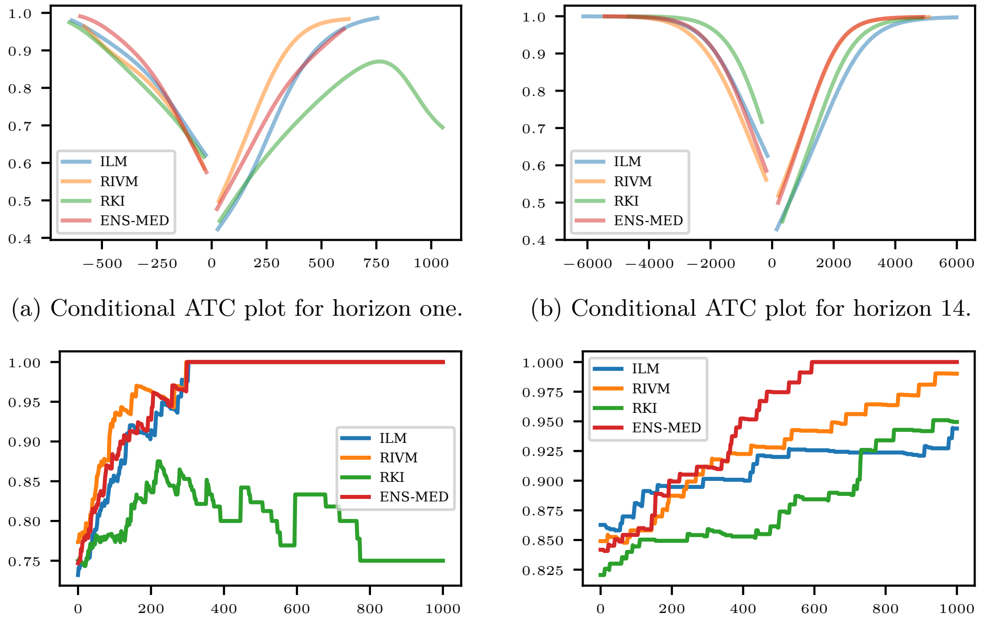


Figure 13: Four-quadrant plots for the COVID-19 nowcast models ILM, RIVM, RKI, and ENS-MEAN and the horizons of one, seven, and 14 days. The spread in both directions increases with the horizon.



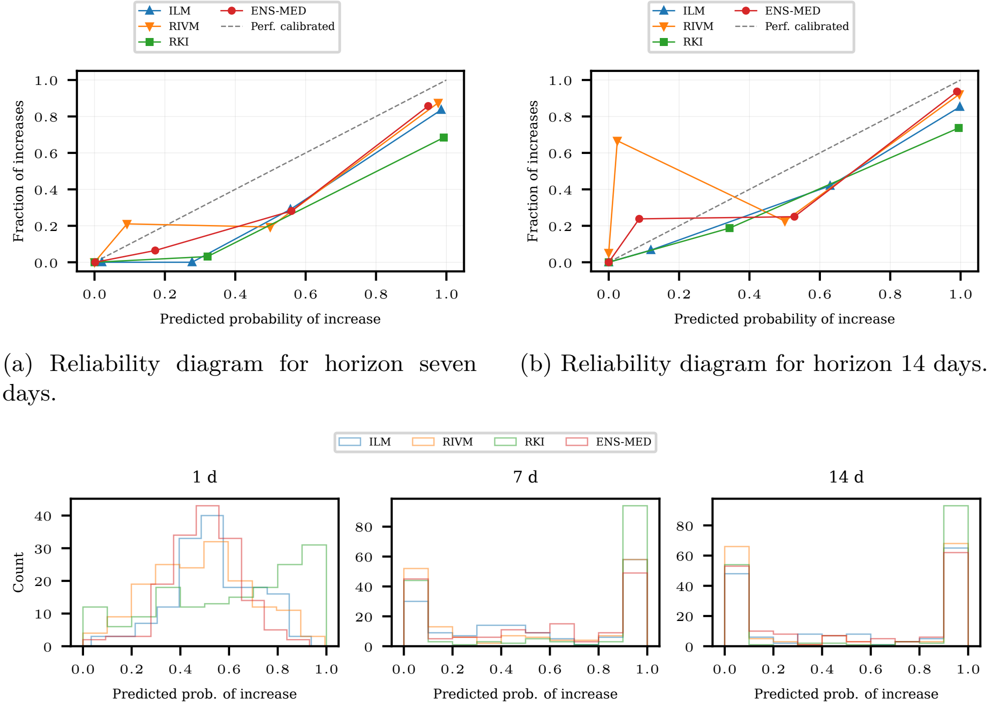
(a) Conditional ATC plot for horizon one.

(b) Conditional ATC plot for horizon 14.

(c) ATC ratio over exclusion area size in $\mathbf{x}^{\Delta}$ for horizon one.

(d) ATC ratio over exclusion area size in $\mathbf{x}^{\Delta}$ for horizon 14.

Figure 14: Conditional ATC plot and ATC ratio over exclusion area for the COVID-19 nowcast models ILM, RKI, RIVM, and ENS-MED for the horizon seven days. Note the different axis scalings of the ATC ratios for the two horizons. While the ENS-MED model has a rather low ATC ratio for small exclusion areas, it is increasing fast for larger exclusion areas. The RKI model is among the worst, except for the small negative predicted changes.



(a) Reliability diagram for horizon seven days.

(b) Reliability diagram for horizon 14 days.

(c) Count histogram of the predicted probabilities for the horizon one, seven, and 14 days.

Figure 15: The reliability diagram for the COVID-19 nowcasting models ILM, RIVM, RKI, and ENS-MED for the horizon seven and 14 days. Additionally, the count of predicted probabilities for the horizons is shown. The reliability diagram bins are chosen according to the empirical quantiles of the predicted probabilities. As the models issue small or large probabilities of increase for the higher horizons, little information on the accuracy of moderate probability predictions is available.