Flusion: Integrating multiple data sources for accurate influenza predictions
Supplemental materials

Evan L. Ray, Yijin Wang, Russ Wolfinger, Nicholas G. Reich

July 11, 2024

1 Introduction

This document has supplemental materials.

2 Reporting adjustments for FluSurv-NET data

In this section, we describe the adjustments we made to the reported FluSurv-NET data. The measure of influenza activity reported by FluSurv-NET is the rate of positive influenza cases per 100,000 population in the catchment area of a reporting healthcare facility or group of facilities. For the purposes of FluSurv-NET, "a case is defined as a person who is a resident in a defined FluSurv-NET catchment area and tests positive for influenza by a laboratory test ordered by a health care professional within 14 days prior to or during hospitalization" [2]. This measure of influenza activity may be impacted by underdetection of influenza cases either if patients with influenza are not tested or if they are tested but the test generates a false negative result.

The US Centers for Disease Control and Prevention (CDC) produces annual estimates of influenza disease burden at the national level that adjust for testing rates and test sensitivity, including point estimates of total nationwide influenza hospitalizations in each season along with 95% uncertainty intervals [1, 3]. We used these to estimate a season-specific scale up factor α that was used to adjust the FluSurv-NET data. This factor was obtained by solving the following equation for α based on the total hospitalization rate over the course of the season that was reported across the entire FluSurv-NET network, the point estimate of national hospital burden due to influenza from CDC, and the US population in units of 100,000 people as reported by the US Census Bureau for the first year of the influenza season:

$$\alpha \cdot (\text{cumulative reported hospitalization rate, FluSurv-NET}) = \frac{\text{National burden estimate}}{\text{100k US population}}$$

Table 1 summarizes these terms and the resulting estimated scale-up factors for each season with FluSurv-NET data in our training set. Note that the scale-up factors are larger in earlier seasons than later seasons, indicating that data from FluSurv-NET undercounted influenza activity more in earlier seasons.

Season	Cum. rate	US population	Est. burden (count)	Est. burden (rate)	α
2010/11	21.7	309,321,666	290,000	93.8	4.3
2011/12	8.6	311,556,874	140,000	44.9	5.2
2012/13	44.0	313,830,990	570,000	181.6	4.1
2013/14	35.2	315,993,715	350,000	110.8	3.1
2014/15	64.0	318,301,008	590,000	185.4	2.9
2015/16	31.5	320,635,163	280,000	87.3	2.8
2016/17	62.0	322,941,311	500,000	154.8	2.5
2017/18	102.7	324,985,539	710,000	218.5	2.1
2018/19	63.5	326,687,501	380,000	116.3	1.8
2019/20	65.7	328,239,523	390,000	118.8	1.8
2022/23	62.4	333,287,557	475,000	142.5	2.3

Table 1: Reported data, intermediate calculations, and final estimates for FluSurv-NET burden adjustments in each training season where we used FluSurv-NET data. The 'Cum. rate' column shows the cumulative reported hospitalization rate over the course of the season for the entire FluSurv-NET network. The US populaton column shows an estimate of the US population size from the US Census Bureau in the first year of the season (e.g., the value shown for the 2010/11 season is the population estimate for 2010). The 'Est. burden (count)' column shows the point estimate of influenza hospitalization burden produced by CDC for each season, and the 'Est. burden (rate)' column expresses these burden estimates as a rate per 100,000 population in the US by dividing the estimated burden count by the US population in units of 100,000 people. The scale-up factor α is the ratio of the values in the 'Est. burden (rate)' and 'Cum. rate' columns.

3 Features measuring local level, slope, and curvature of the surveillance signal

As was described in section 5 of the main text, the GBQR models used features based on rolling means and the coefficients of Taylor polynomials fit to rolling windows of the data. These features are designed to estimate the local level, slope, and curvature of the surveillance signal at each point in time, and we describe their calculation here. Recall the notation $\tilde{z}_{l,s,t}$ representing the value of the signal for location l and data source s at time t, after some initial standardizing transformations as described in section 5.1 of the main text.

At time t, the rolling mean over the trailing window of length w is computed as

$$\frac{1}{w} \sum_{u=t-w+1}^{t} \tilde{z}_{l,s,u}. \tag{1}$$

The coefficients of a degree d Taylor polynomial based on the trailing window of length w relative to the anchor point t are obtained by fitting the following model to the observations $\{\tilde{z}_{l,s,u} : u = t-w+1, \ldots, t\}$:

$$\begin{aligned} \tilde{z}_{l,s,u} &= \sum_{c=0}^{d} \frac{1}{c!} \beta_c (u-t)^c + \varepsilon_u \\ \varepsilon_u &\sim \text{Normal}(0, \sigma^2) \end{aligned} \tag{2}$$

For example, with $d = 2$ we fit the quadratic model

$$\tilde{z}_{l,s,u} = \beta_0 + \beta_1(u-t) + \frac{1}{2}\beta_2(u-t)^2 + \varepsilon_u$$
$$\varepsilon_u \sim \text{Normal}(0, \sigma^2)$$

To motivate this, suppose that the underlying signal follows a mean trend over time given by the smooth function $g(u)$, with observation noise due to, e.g., the reporting process. The function g can be written in terms of its derivatives $g^{(c)}$ using the Taylor expansion about the point t:

$$g(u) = \sum_{c=0}^{\infty} \frac{g^{(c)}(t)}{c!}(u-t)^c.$$

Truncating to the first $d+1$ terms yields an approximation to g in the neighborhood of t, and the coefficient estimates β_c from the linear model (2) can be regarded as estimates of the corresponding derivatives $g^{(c)}(t)$. We refer to estimates of β_0, β_1, and β_2 as estimates of the local level, trend, and curvature of the signal respectively. The highest degree we used in any of our feature computations was $d = 2$. Note that the rolling mean of Equation (1) could also be obtained from this process using a Taylor polynomial of degree $d = 0$, though in practice we used a more direct implementation.

Figure 2 illustrates the values of these features for the NHSN admission signal in the state of Michigan in the 2023/24 season. As expected, the features calculated based on longer window sizes w and lower polynomial degrees d vary more smoothly over time than features calculated based on shorter windows or higher polynomial degrees. Nevertheless, the features generally agree in terms of when the slope and curvature are positive or negative.

Note that at the end of the signal, only observations on or before the last time point are available. This motivates the use of a trailing window for feature calculation: with this choice, the features computed at both the end of the time series and at earlier time points can be expected to have similar characteristics as measures of local derivatives of the signal's trend. In contrast, if a centered window were used, estimates at earlier time points (when all observations within the centered window are available) would be more reliable than estimates at the end of the series.

Importantly, we do not account for the history of data revisions when we calculate these features. For example, for model fitting on reference date t, training examples are assembled for past times $u < t$ that include features measuring the local level, slope, and curvature at those times u. Those features are calculated based on the latest available data at time t, not based on the data that would have been available at time u. This means that our model implicitly estimates the relationships between these features and the target when the features are calculated on finalized, fully reported data. However, when predictions are generated extending from the reference date t, those features are calculated at the end of the time series when reported values more likely to be subsequently revised, leading to a mismatch between the data used for model fitting and the data used for prediction. This is a challenging problem to address in a setting like ours where the target data system has only a short reporting history and the characteristics of its revision process are not well known.

Finally, we highlight that although features such as the rolling mean or the intercept of a Taylor polynomial only directly measure the local level of the signal, when their lags are also included as features they can provide information about trend as well. For example, if we see that the rolling mean at time t is larger than the rolling mean at time $t-1$ we may infer that the value of the signal

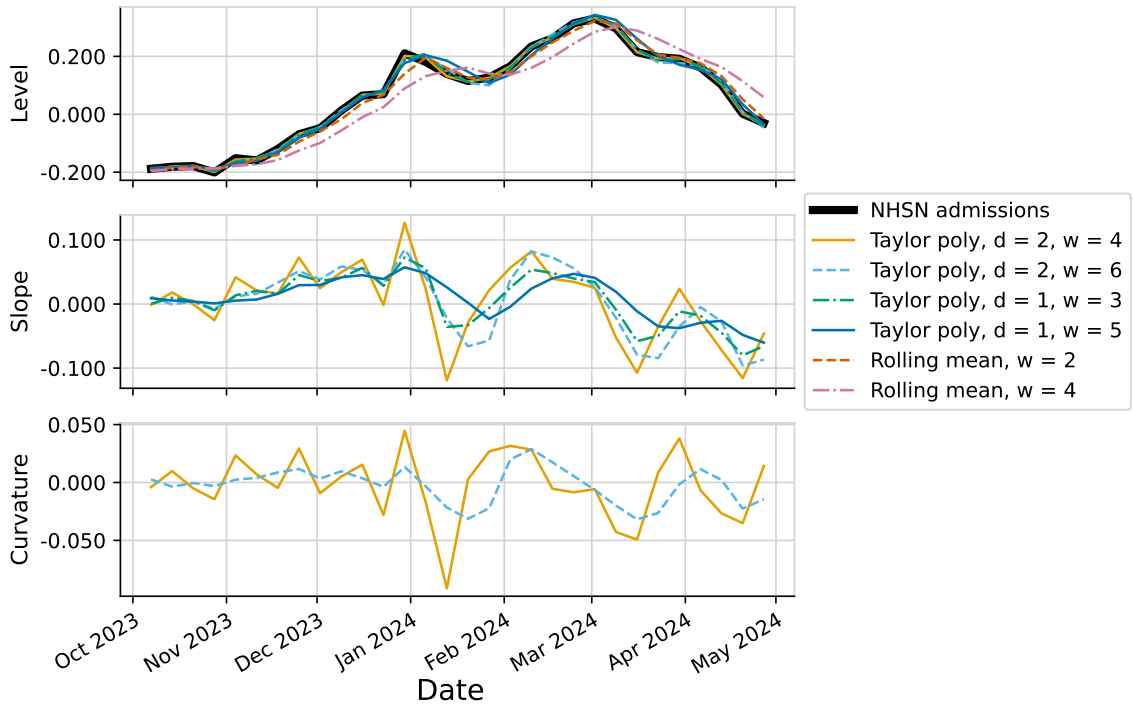

Figure 1: Example of features measuring the local level, trend, and curvature of the standardized NHSN admissions signal for the state of Michigan in the 2023/24 season (shown in black in the top panel for reference). At each time on the horizontal axis, a vertical line will intersect features calculated based on a trailing window ending on that date. For example, on Christmas week (just before Jan 2024), features based on a Taylor polynomial of degree $d = 2$ fit to a trailing window of size $w = 4$ produced a local level estimate that closely matched the Christmas peak observed in the data, a positive slope just over 0.1 on the scale of the standardized data, and a positive curvature just under 0.05 indicating that the trend was increasing over that four week period.

is rising.

4 Feature importance

TODO clean up names of features in figure, describe how importance is calculated

5 FluSight results: sensitivity analysis for data revisions

Table 2 contains MAE, MWIS, and PI coverage rates for real-time FluSight predictions, omitting predictions made on combinations of location and reference date for which the most recent available data at the time the prediction was generated were subsequently revised by 10 or more admissions. This represents a generous sensitivity analysis, omitting 265 out of 1590 combinations of location and reference date for which predictions were submitted. Figure 3 displays information about the magnitudes of these revisions.

Comparing with Table 1 in the primary manuscript, we note that the main results discussed there still hold: Flusion has the best MAE and MWIS values by a substantial margin, while the marginal coverage rates of its central prediction intervals are too conservative.

Model	% Submitted	MWIS	RWIS	MAE	RAE	50% Cov.	95% Cov.
Flusion	100.0	**21.1**	**0.575**	**32.3**	**0.626**	0.597	0.971
Other Model #1	100.0	26.2	0.714	39.3	0.762	0.565	0.939
FluSight-ensemble	100.0	26.1	0.715	40.3	0.784	0.521	0.930
Other Model #2	89.3	28.8	0.732	42.9	0.773	0.495	0.912
Other Model #3	97.5	28.4	0.760	42.4	0.809	0.364	0.795
Other Model #4	100.0	28.3	0.773	43.3	0.844	**0.498**	0.882
Other Model #5	100.0	29.1	0.796	43.1	0.837	0.491	0.881
Other Model #6	85.1	32.8	0.811	47.1	0.826	0.421	0.827
Other Model #7	100.0	30.3	0.828	47.0	0.913	0.474	0.902
Other Model #8	100.0	30.3	0.829	43.9	0.854	0.447	0.830
Other Model #9	98.7	31.8	0.855	48.1	0.924	0.473	**0.944**
Other Model #10	95.3	33.1	0.885	48.7	0.927	0.575	0.881
Baseline-trend	100.0	32.5	0.890	49.5	0.967	0.639	0.929
Other Model #11	87.3	32.1	0.902	49.5	0.988	0.443	0.923
Other Model #12	96.9	31.7	0.925	47.3	0.984	0.429	0.892
Other Model #13	92.8	34.7	0.937	49.8	0.955	0.464	0.829
Other Model #14	95.9	35.7	0.945	46.8	0.881	0.242	0.778
Other Model #15	98.1	36.3	0.976	51.2	0.981	0.393	0.772
Other Model #16	68.3	43.1	0.982	63.7	1.030	0.416	0.829
Other Model #17	99.2	35.2	0.982	42.9	0.850	0.580	0.789
Baseline-flat	100.0	36.4	1.000	51.2	1.000	0.308	0.903
Other Model #18	74.0	43.4	1.020	62.1	1.030	0.304	0.739
Other Model #19	88.5	41.5	1.070	60.7	1.110	0.383	0.814
Other Model #20	85.5	42.5	1.150	54.1	1.050	0.327	0.632
Other Model #21	85.3	40.9	1.200	54.8	1.150	0.379	0.770
Other Model #22	72.4	33.8	1.320	46.5	1.300	0.398	0.783
Other Model #23	85.8	50.8	1.350	64.8	1.230	0.226	0.508
Other Model #24	91.5	52.6	1.360	72.1	1.320	0.404	0.825
Other Model #25	92.5	89.5	2.390	110.0	2.080	0.184	0.443

Table 2: Overall evaluation results for forecasts submitted to the FluSight Forecast Hub, omitting forecasts made on combinations of reference date and location for which the latest available NHSN data at the time of the forecast were subsequently revised by 10 or more admissions. Model names other than Flusion, FluSight-ensemble, Baseline-flat, and Baseline-trend are anonymized. The percent of all combinations of location, reference date, and horizon for which the given model submitted forecasts is shown in the "% Submitted" column; only models submitting at least 2/3 of forecasts were included. Results for the model with the best MWIS, RWIS, MAE, and RAE are highlighted. Results for the models where empirical PI coverage rates are closest to the nominal levels are highlighted.

6 Experimental results: sensitivity analysis for data revisions

Table 3 contains results from the post hoc experiments described in section 7 of the main text, omitting forecasts produced for combinations of location and reference date where the latest available NHSN data as of the reference date were subsequently revised up or down by at least 10 admissions. Comparing with table 2 of the main text, we see that the qualitative modeling results discussed there still hold in this sensitivity analysis.

References

[1] Centers for Disease Control and Prevention. *How CDC Estimates the Burden of Seasonal Influenza in the U.S.* Accessed: 2024-07-11. 2024. URL: `https://www.cdc.gov/flu/about/burden/how-cdc-estimates.htm`.

[2] Centers for Disease Control and Prevention. *Influenza Hospitalization Surveillance Network (FluSurv-NET)*. Accessed: 2024-07-03. 2023. URL: https://www.cdc.gov/flu/weekly/influenza-hospitalization-surveillance.htm.

[3] Centers for Disease Control and Prevention. *Past Seasons Estimated Influenza Disease Burden*. Accessed: 2024-07-03. 2024. URL: https://www.cdc.gov/flu/about/burden/past-seasons.html.

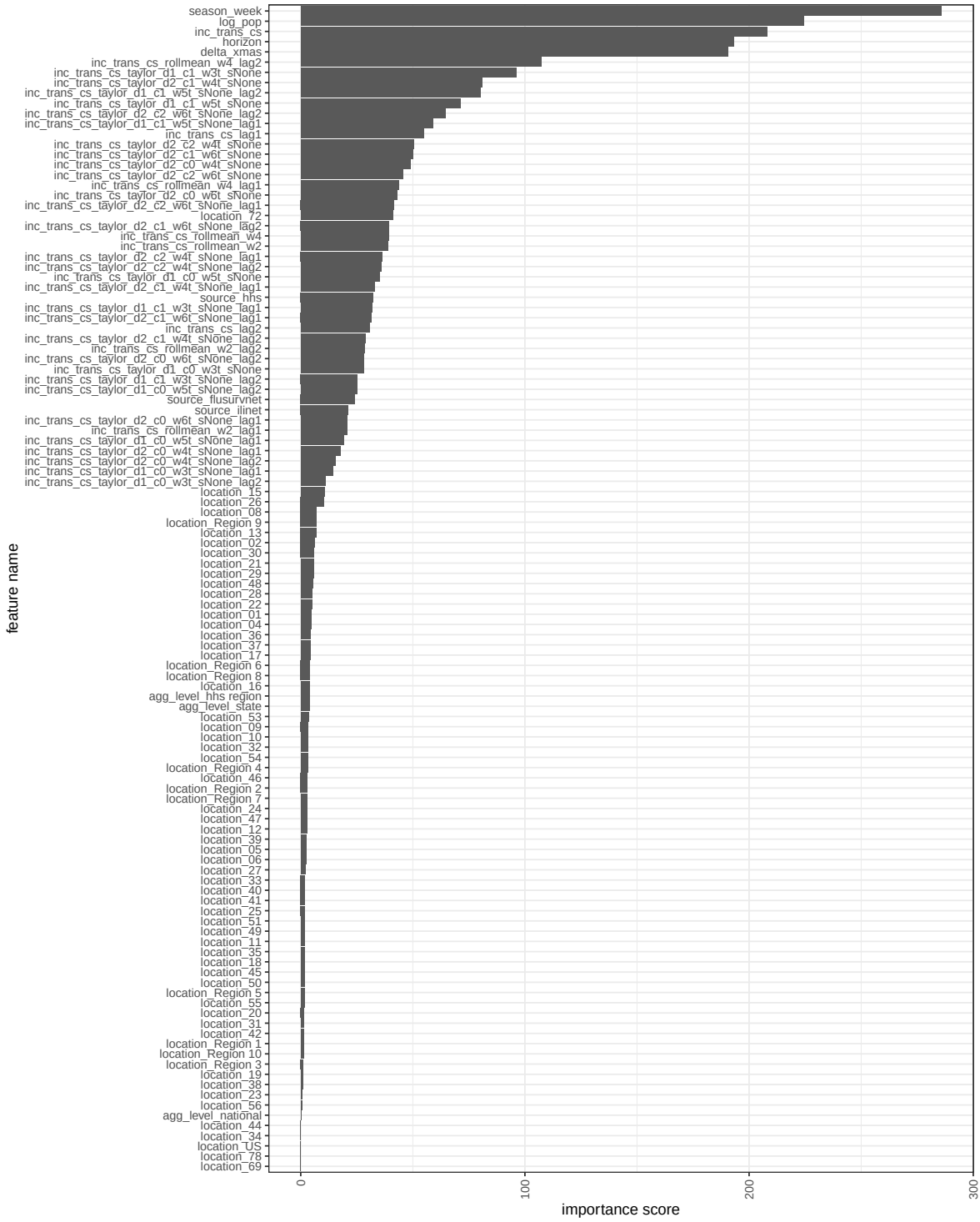

Figure 2: Caption about feature importance.

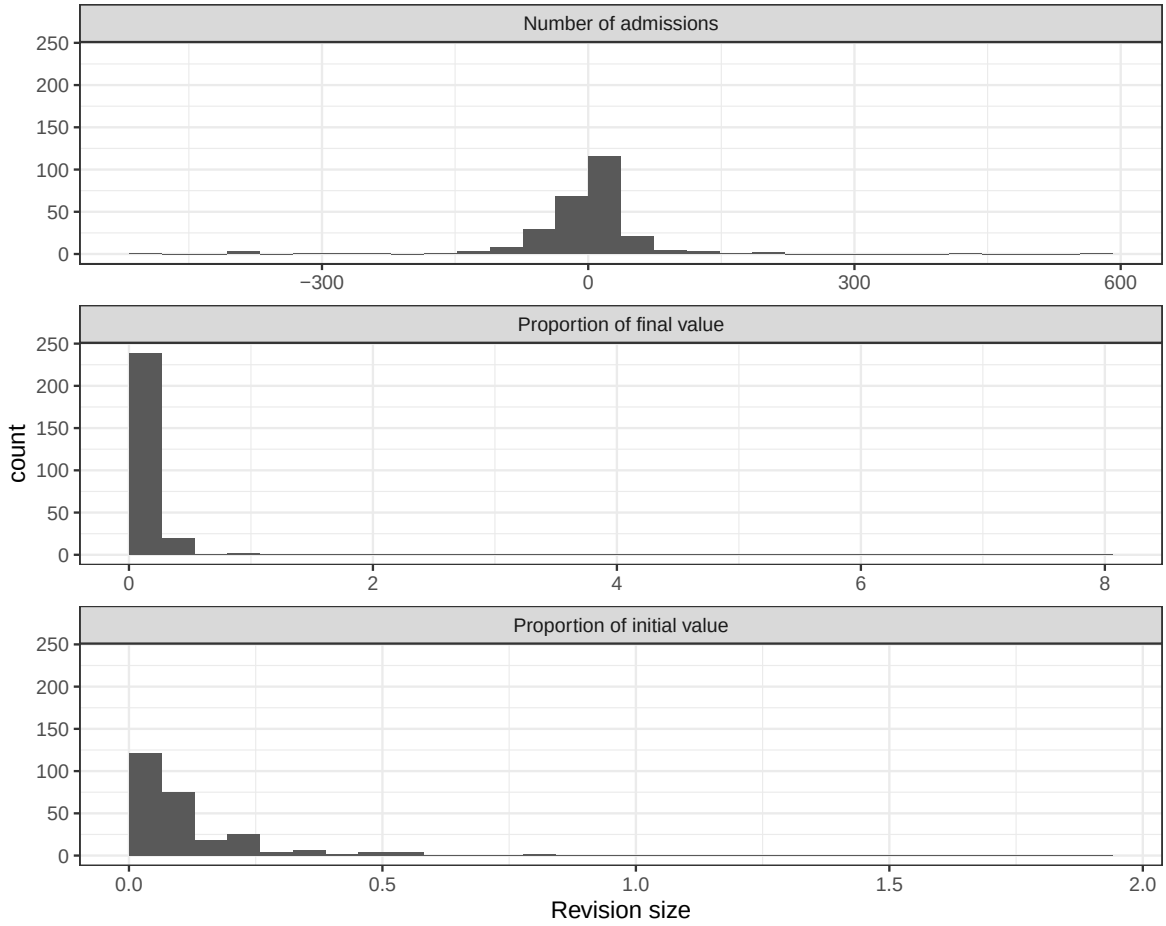

Figure 3: Measures of the size of reporting revisions for combinations of location and reference date that were omitted in the sensitivity analysis. For legibility, only those revisions that were dropped (i.e., where the revision amount was at least 10 admissions up or down from the initial reported value) are displayed; most revisions were small. The top panel shows the size of the revision in units of hospital admissions, where positive numbers indicate an upward revision of the initially reported value. The second panel shows the absolute value of the revision size as a proportion of the final reported value. The third panel shows the absolute value of the revision size as a proportion of the initial reported value. When computing proportions, we add one to the denominator to avoid division by zero. As an example, for October 7, 2023 (the last date for which data were available when producing predictions with a reference date of October 14, 2023), in Washington state the initial reported value was 43, which was subsequently revised down to a final value of 4. The revision amount is -39, which is 7.80 when expressed as a proportion of the final reported value or 0.89 when expressed as a proportion of the initial reported value.

Experiment A: Component model performance

Model	% Submitted	MWIS	RWIS	MAE	RAE	50% Cov.	95% Cov.
GBQR, ARX	100.0	**21.2**	**0.582**	**32.0**	**0.626**	0.589	0.967
GBQR	100.0	21.3	0.583	32.4	0.632	0.544	0.952
Flusion	100.0	21.4	0.588	32.9	0.642	0.574	0.969
GBQR, GBQR-no-level	100.0	21.5	0.589	32.9	0.643	0.555	0.961
GBQR-no-level, ARX	100.0	23.7	0.651	37.0	0.723	0.542	0.964
GBQR-no-level	100.0	24.0	0.657	36.7	0.717	0.525	**0.949**
ARX	100.0	28.1	0.771	43.0	0.841	**0.508**	0.934
Baseline-flat	100.0	36.4	1.000	51.2	1.000	0.308	0.903

Experiment B: Reduced training data

Model	% Submitted	MWIS	RWIS	MAE	RAE	50% Cov.	95% Cov.
GBQR	100.0	**21.3**	**0.583**	**32.4**	**0.632**	**0.544**	**0.952**
GBQR-by-location	100.0	25.5	0.701	39.3	0.768	0.340	0.895
GBQR-only-NHSN	100.0	30.1	0.826	46.7	0.912	0.368	0.850
Baseline-flat	100.0	36.4	1.000	51.2	1.000	0.308	0.903

Experiment C: Data preprocessing

Model	% Submitted	MWIS	RWIS	MAE	RAE	50% Cov.	95% Cov.
GBQR-no-reporting-adj	100.0	**20.8**	**0.572**	**31.8**	**0.622**	0.518	0.942
GBQR	100.0	21.3	0.583	32.4	0.632	0.544	**0.952**
GBQR-no-transform	100.0	22.1	0.606	34.0	0.664	**0.496**	**0.948**
Baseline-flat	100.0	36.4	1.000	51.2	1.000	0.308	0.903

Table 3: Evaluation results for post hoc experiments investigating determinants of model performance, omitting forecasts made on combinations of reference date and location for which the latest available NHSN data at the time of the forecast were subsequently revised by 10 or more admissions. Experiment A gives results for individual component models in the Flusion ensemble, ensembles of pairs of components, and the full Flusion ensemble including all three components. Experiment B gives results for the GBQR model, which is trained jointly on data for all locations and data sources, and variations trained separately for each location (GBQR-by-location) and trained only on hospital admissions from NHSN (GBQR-only-NHSN). Experiment C gives results for a variation on the GBQR model that does not incorporate reporting adjustments designed to improve the degree to which ILINet and FluSurvNET data reflect influenza activity (GBQR-no-reporting-adj) and a variation that does not use a fourth-root transform (GBQR-no-transform), along with the original GBQR model which uses the reporting adjustments and the fourth-root transform. The percent of all combinations of location, reference date, and horizon for which the given model submitted forecasts is shown in the "% Submitted" column; in these retrospective experiments, we produced forecasts for all locations and time points. Within each experiment group, results for the model with the best MWIS, RWIS, MAE, and RAE are highlighted. Results for the models where empirical PI coverage rates are closest to the nominal levels are highlighted.