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

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

July 8, 2024

## 1 Introduction

This document has supplemental materials.

## 2 Reporting adjustments for FluSurv-NET data

## 3 Features derived from Taylor polynomial coefficients

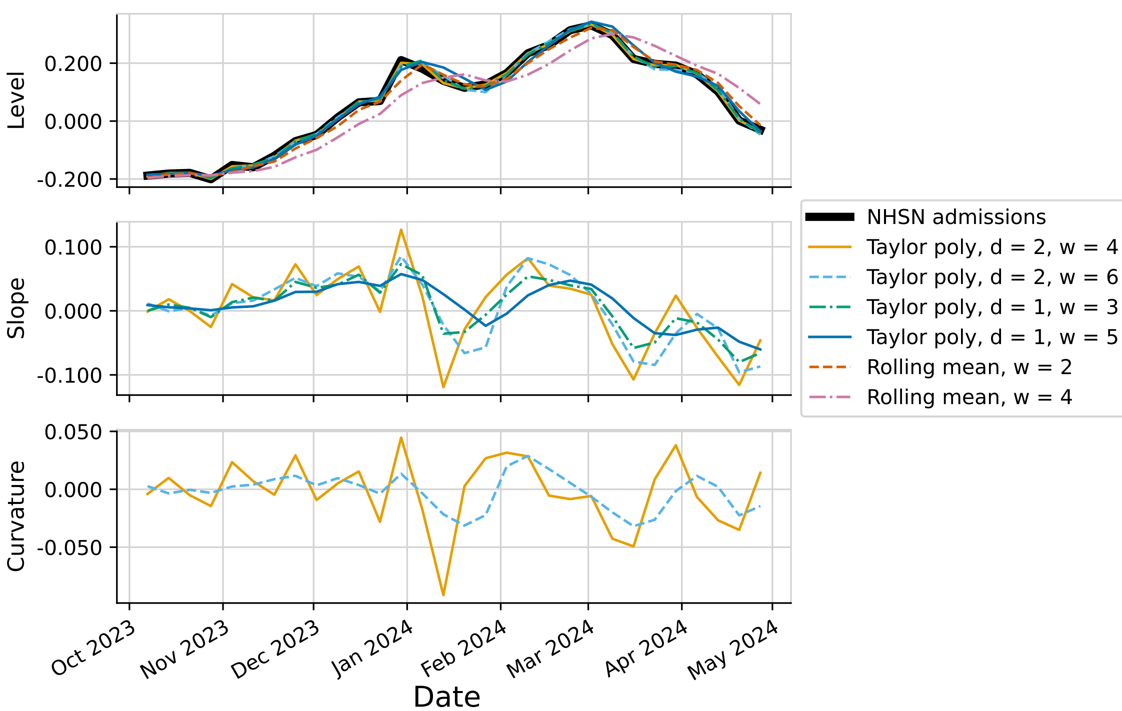


Figure 1: Caption about features.

# 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 1 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.

# 6 Experimental results: sensitivity analysis for data revisions

| 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 1: 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.

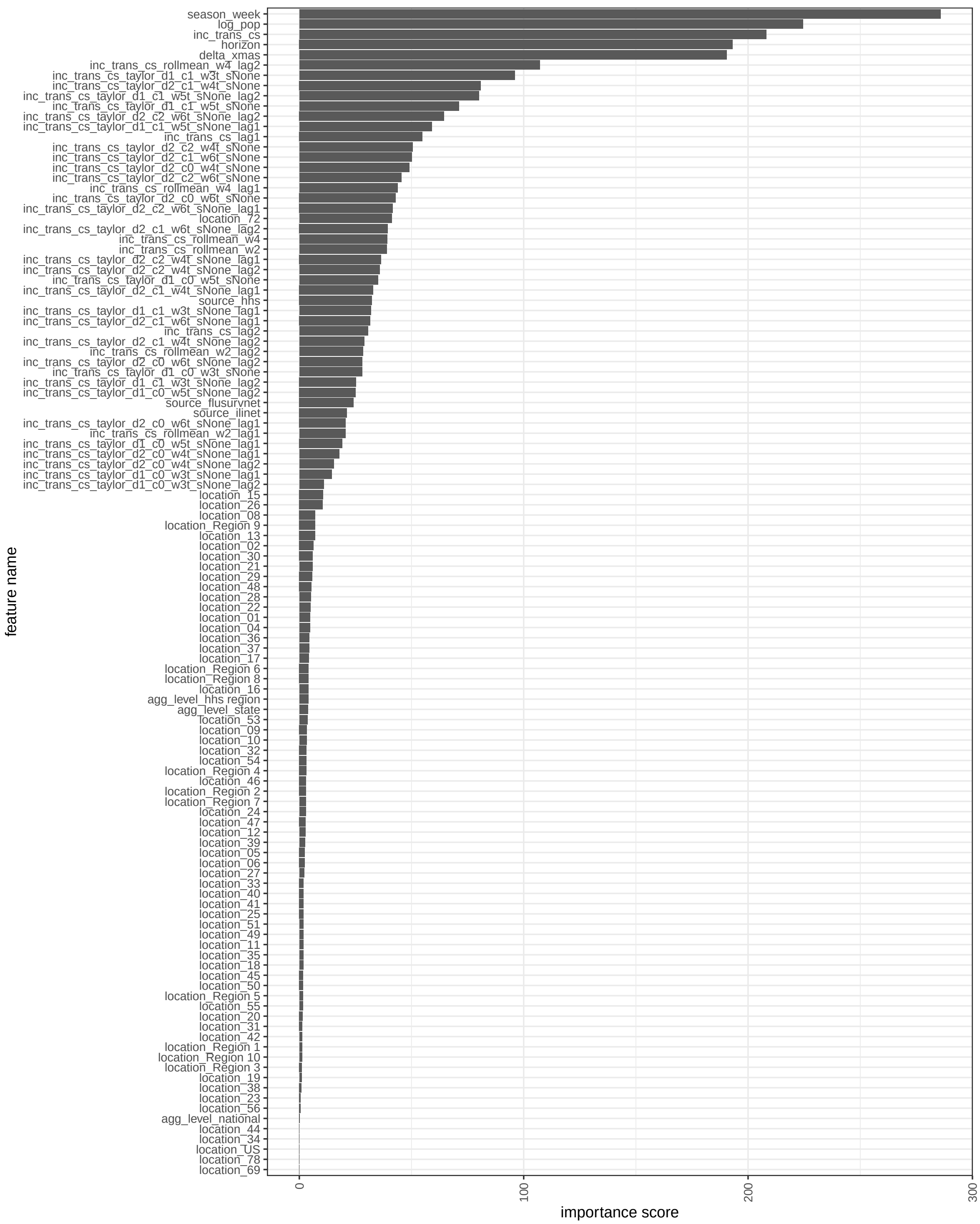


Figure 2: Caption about feature importance.

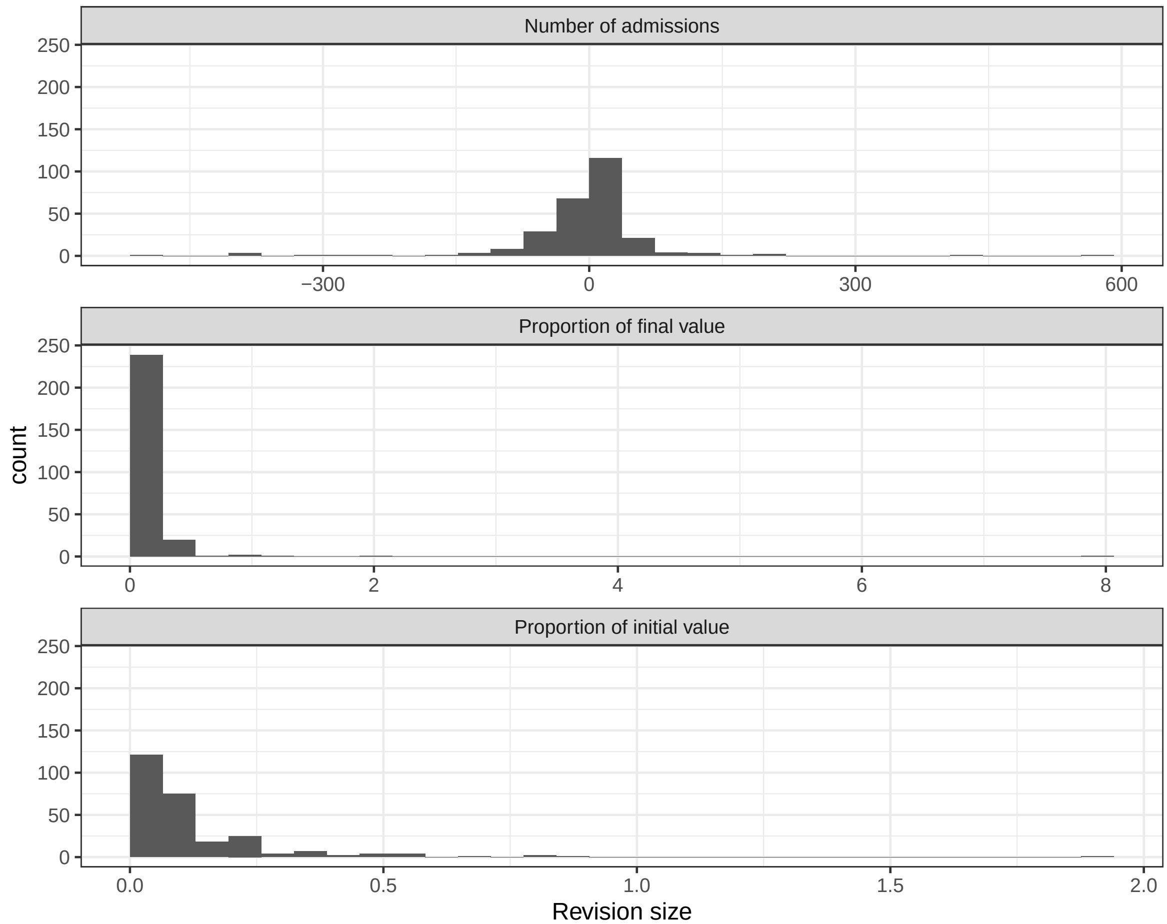


Figure 3: Measures of the size of reporting revisions for combinations of location and reference date that were omitted in the sensitivity analysis. For legibility, only those revisions that were dropped (i.e., where the revision amount was at least 10 admissions up or down from the initial reported 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 2: 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.