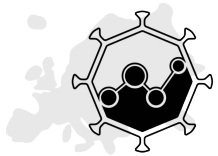
Improving Short-Term Forecasts by Anticipating Data Revisions

Valeria Marras, ISI Foundation
2025/03/03

ISI Foundation



RESPICAST
ECDC RESPIRATORY DISEASES FORECASTING HUB

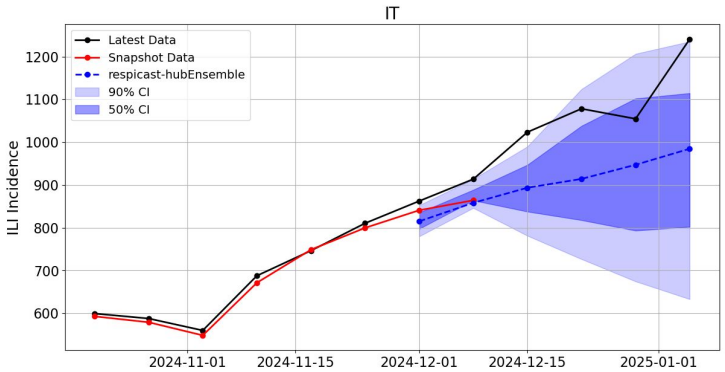
Problem statement and goals

- The quality of epidemiological data is crucial for accurate surveillance and forecasting
- Data collected in surveillance systems are often **retrospectively revised**, leading to discrepancies between initially reported and final values

Why data are revised?

- Surveillance systems rely on sentinel physicians' reports
- In certain periods of the year the reports are delayed

Problem:
Forecast trained on **initially reported data** and compared to **later revised** values.



Problem statement and goals

Analysis of retrospective corrections in **ERVISS surveillance data**[1]

- **Data Revision**: Process of updating previously published surveillance data
- We considered the following **targets**
  - ILI incidence
  - ARI incidence

Goals:

- Measure the problem and characterise it:
  - Estimate the magnitude of data revisions
  - Finding the factors influencing the review process
- Develop correction models to estimate revisions for newly reported data
- Improve the quality of forecasting models by understanding data revision patterns

[1] **https://erviss.org/**

Outline

1. Data Revision magnitude overview
2. Influence of time-dependent patterns in the revision process
3. Data revision estimation for forecasting improvement
4. Models & Methodology
5. Results & next steps

Outline

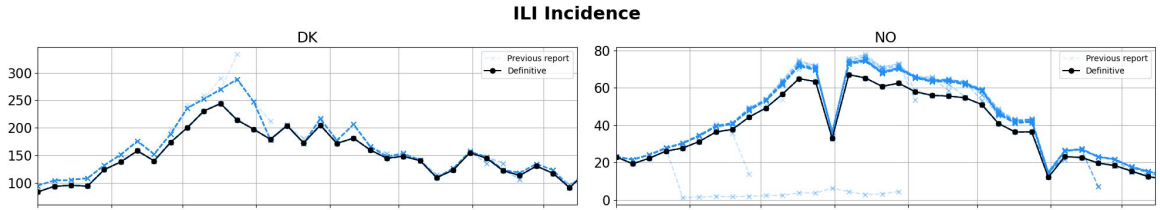
1. Data Revision magnitude overview
2. Influence of time-dependent patterns in the revision process
3. Data revision estimation for forecasting improvement
4. Models & Methodology
5. Results & next steps

Data Revision Overview

Weekly incidence values from the **latest update** and all **previous reports** for the same week

Looking at the curves:

- Revision process:
  - **Some countries** carry out revisions

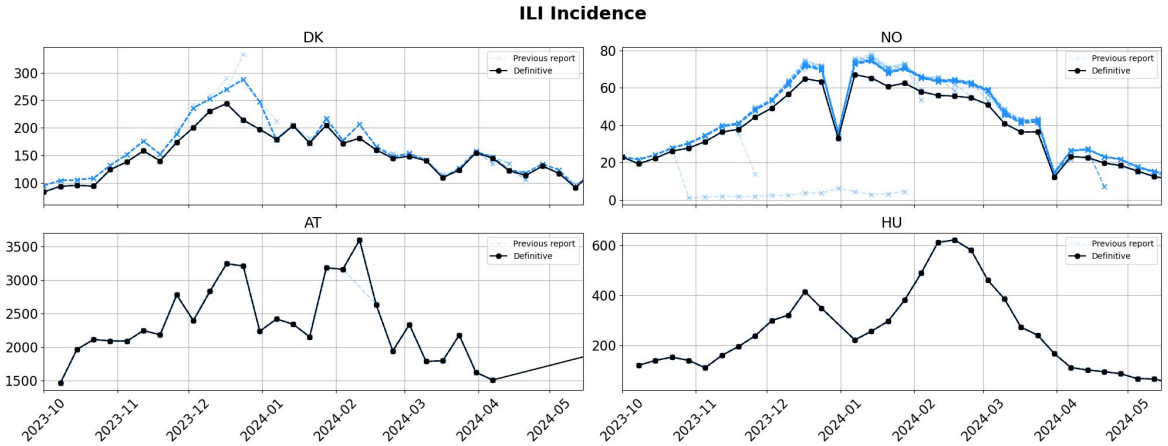


Data Revision Overview

Weekly incidence values from the **latest update** and all **previous reports** for the same week

Looking at the curves:

- Revision process:
  - **Some countries** carry out revisions
  - Others **do not**

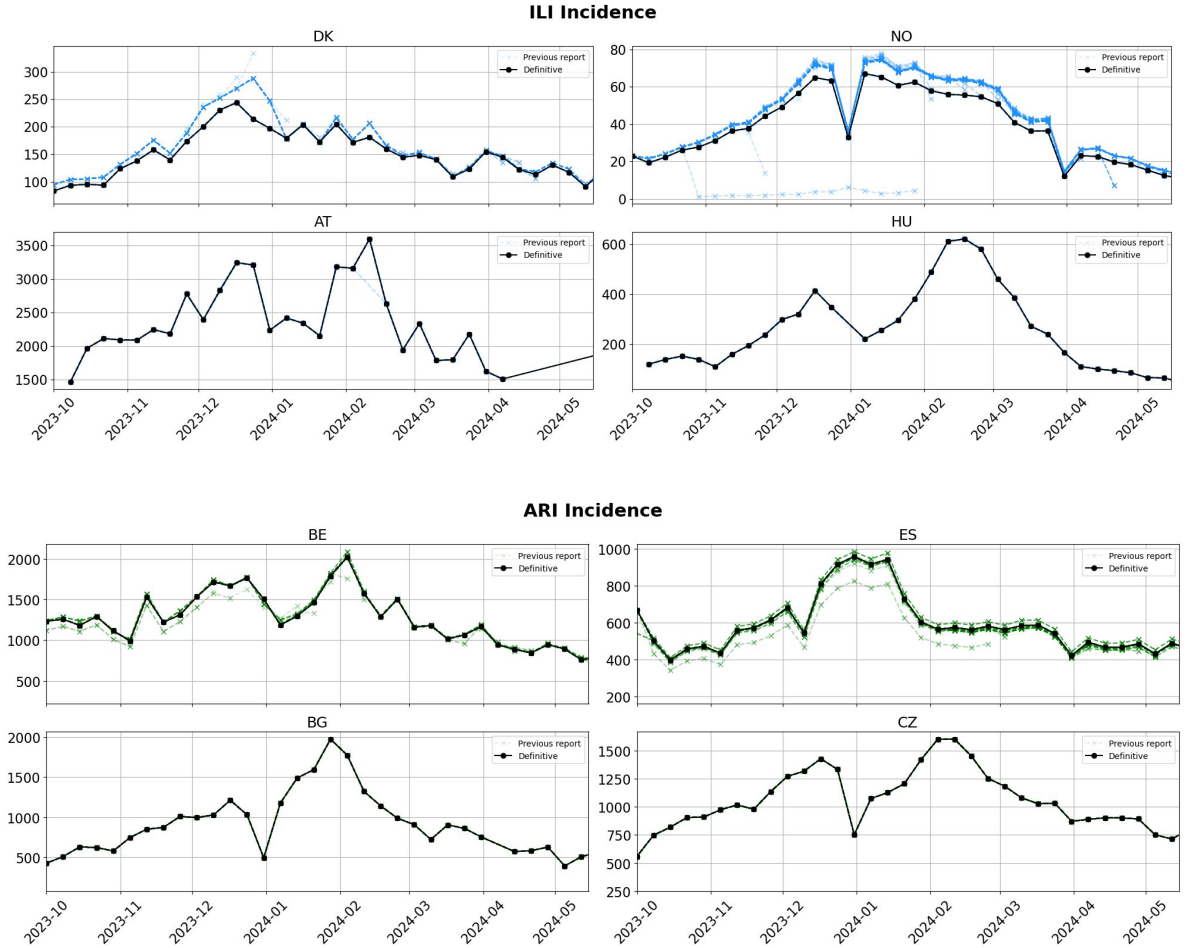


Data Revision Overview

Weekly incidence values from the **latest update** and all **previous reports** for the same week

Looking at the curves:

- Revision process:
  - **Some countries** carry out revisions
  - Others **do not**



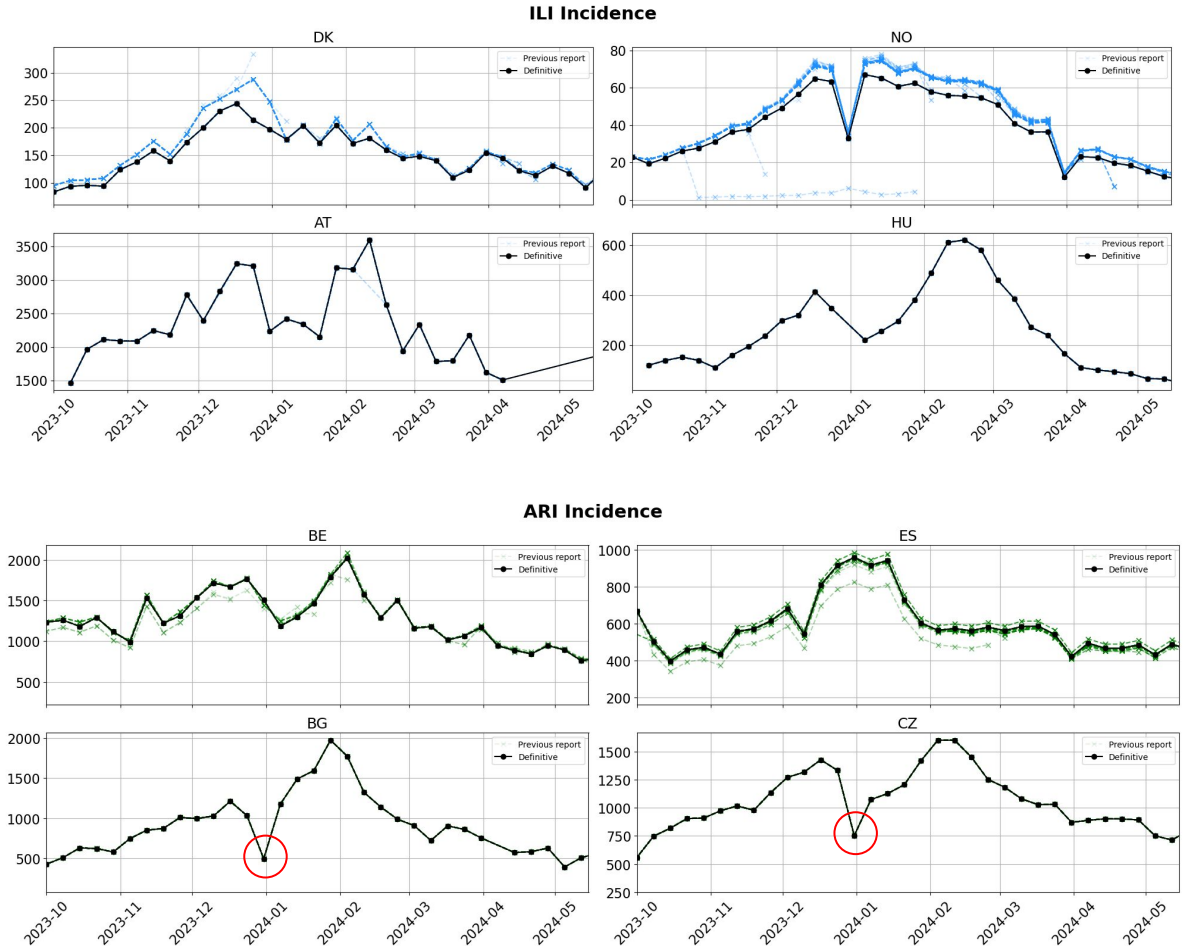
Data Revision Overview

Weekly incidence values from the **latest update** and all **previous reports** for the same week

Looking at the curves:

- Revision process:
  - Some countries carry out revisions
  - Others do not

- Challenging behaviour
  - New Year: critical week
  - Possible **under reporting**
  - Never revised in subsequent reports

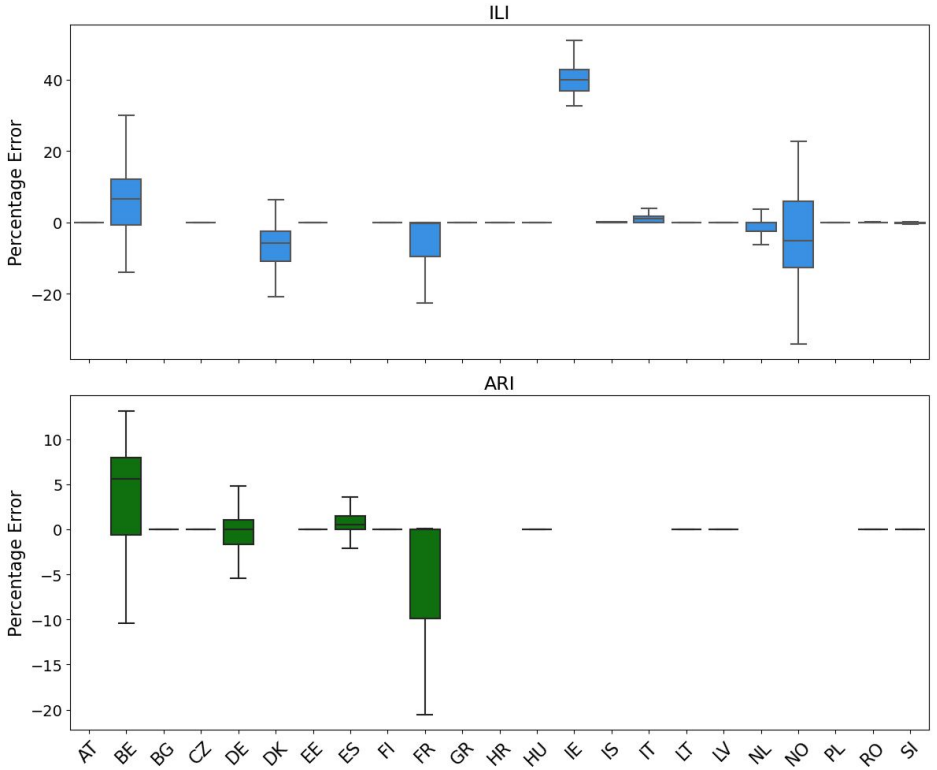


Data revision magnitude

Boxplot of overall magnitude percentage error:

- Difference between first reported value and the final corrected value for a given week (Season 2023-24)

- Countries presenting most significant revisions:
  - **ILI**: BE, DK, FR, NL, NO, IT
  - **ARI:** BE, FR, DE, ES

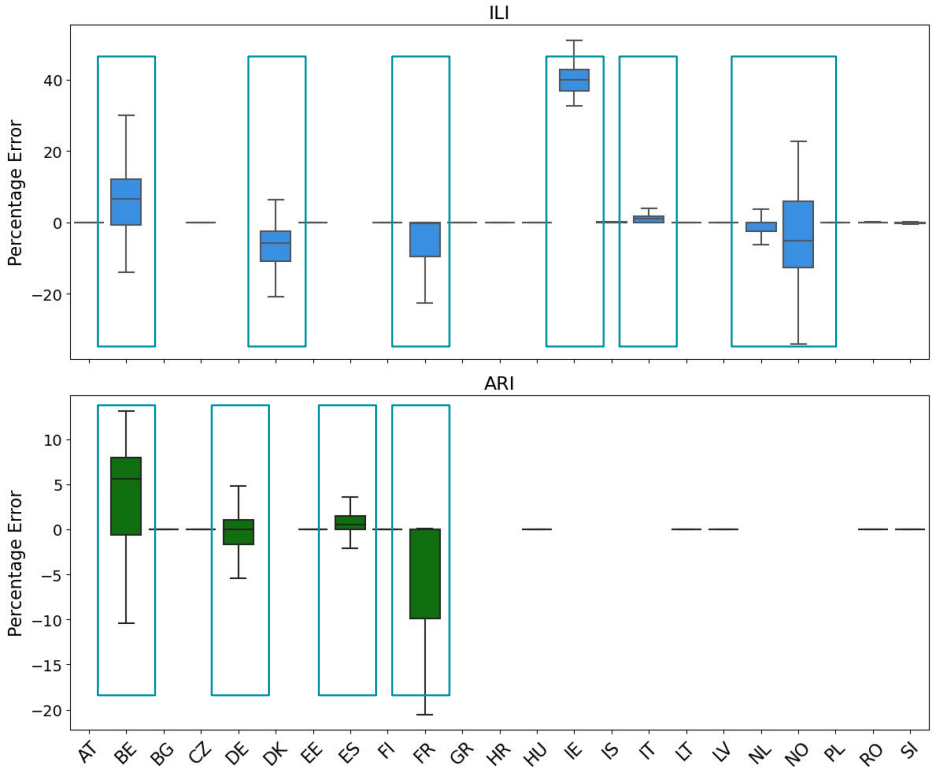


Data revision magnitude

Boxplot of overall magnitude percentage error:

- Difference between first reported value and the final corrected value for a given week (Season 2023-24)

- Countries presenting most significant revisions:
  - **ILI**: BE, DK, FR, NL, NO, IT
  - **ARI:** BE, FR, DE, ES



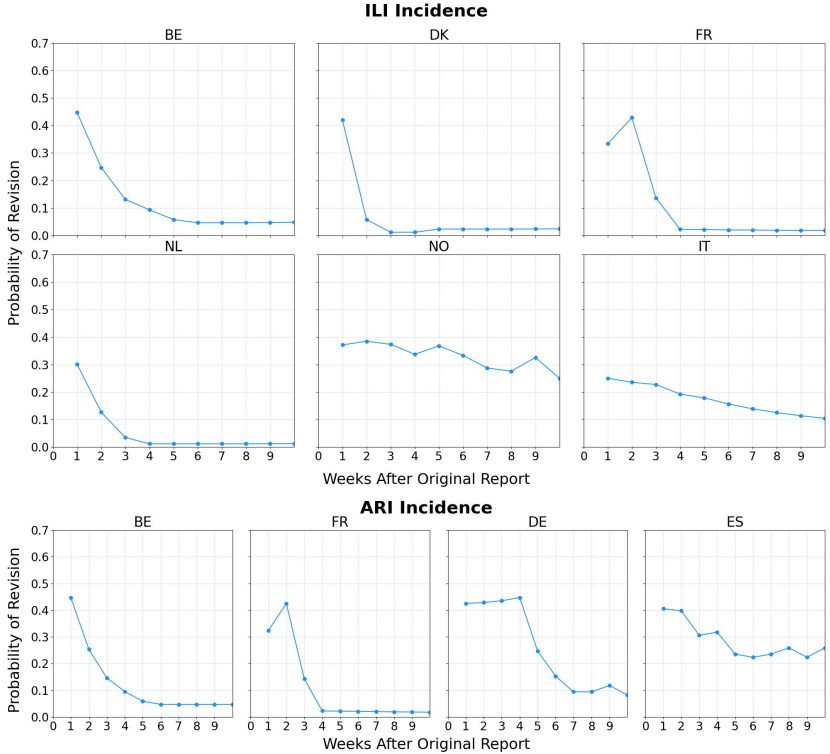
Outline

1. Data Revision magnitude overview
2. Influence of time-dependent patterns in the revision process
3. Data revision estimation for forecasting improvement
4. Models & Methodology
5. Results & next steps

Influence of time-dependent patterns in the revision process

Probability of Revision

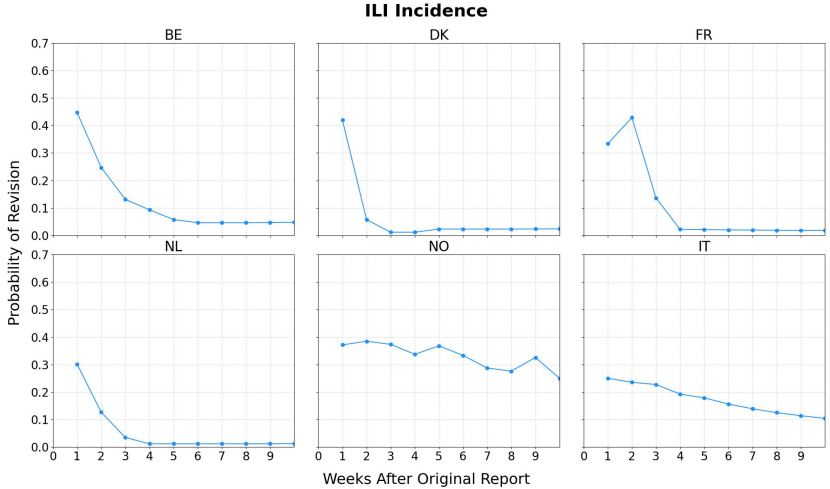
Probability that a given week's data will be revised in future reports.

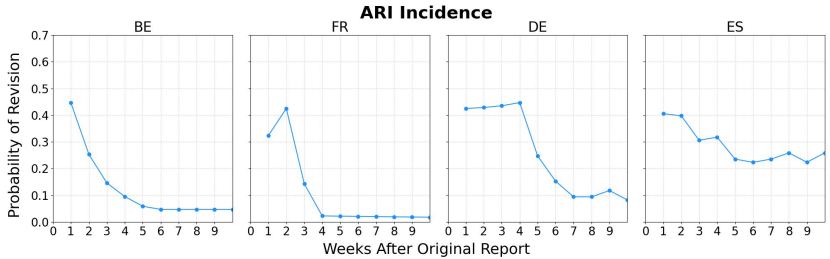


Influence of time-dependent patterns in the revision process

Probability of Revision

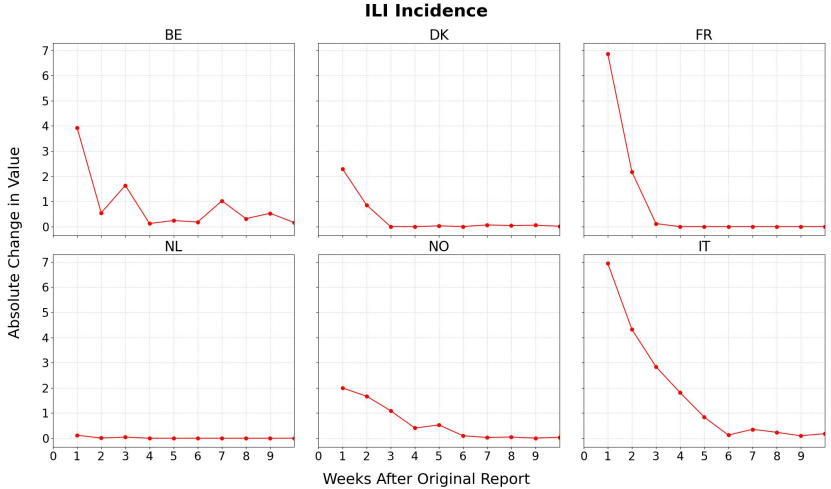
Probability that a given week's data will be revised in future reports.

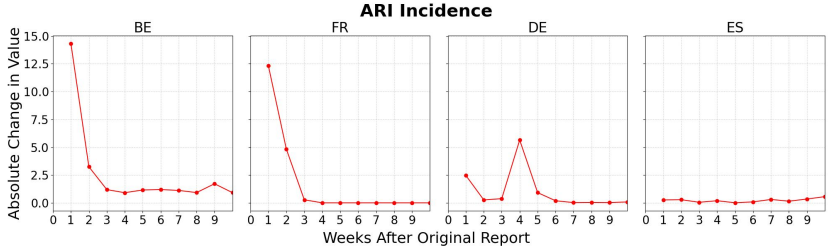




Magnitude of Revisions

Average absolute change in reported values

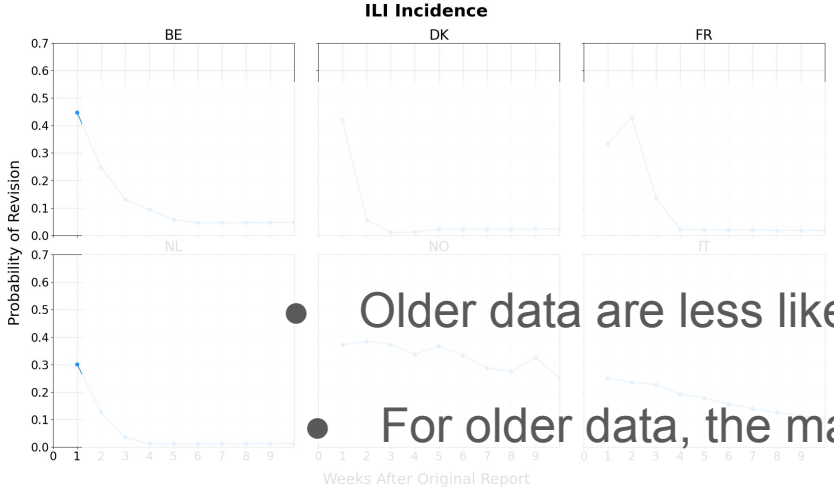




Influence of time-dependent patterns in the revision process

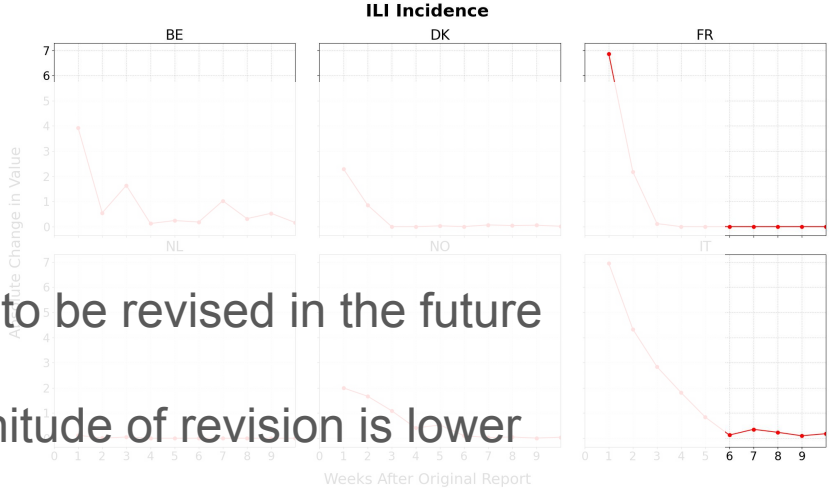
Probability of Revision

Probability that a given week's data will be revised in future reports.

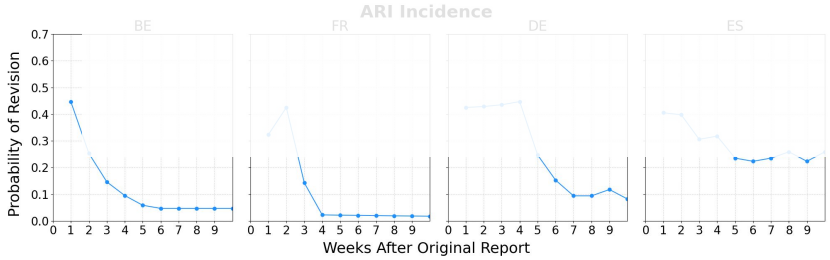


Magnitude of Revisions

Average absolute change in reported values



- Older data are less likely to be revised in the future
- For older data, the magnitude of revision is lower



Data revision estimation for forecasting improvement

- **Fom data analysis**
  - Many countries do not carry out retrospective revisions
  - Both positive and negative revisions
  - Heterogeneity in revising countries
  - Age is an important feature to characterize the problem

- **Goals**
  - Estimate the magnitude of data revision based on the historical dataset
  - Improve forecasting models by incorporating revision patterns and probabilities

Outline

1. Data Revision magnitude overview
2. Influence of time-dependent patterns in the revision process
3. Data revision estimation for forecasting improvement
4. Models & Methodology
5. Results & next steps

Models & Methodology

Given the **heterogeneity** of curves across countries

- Explicit mathematical form is not ideal
- We chose a machine learning approach.

- **Models used to estimate revision magnitude:**
  - Decision Tree Regressor (**DT**):
    Captures non-linear relationships by splitting data into decision nodes.
  - Random Forest Regressor (**RF**):
    An ensemble of decision trees that reduces variance and improves generalization by averaging multiple predictions.

Models & Methodology

- **Features & Target**
  - Features: Age (weeks after original report), Value
  - Target: Magnitude of data revision

- **Approach**
  - Training: Data from 2023/24 season
  - Test: Data from 2024/25 season
  - Countries: Revising countries from previous season

Models & Methodology

We want to estimate **how good** the predictions of the models are compared to baseline models

- **Baseline Model**: uses historical revision distribution by age to randomly select revision values
- **Persistence Model**: assumes no revision occurs

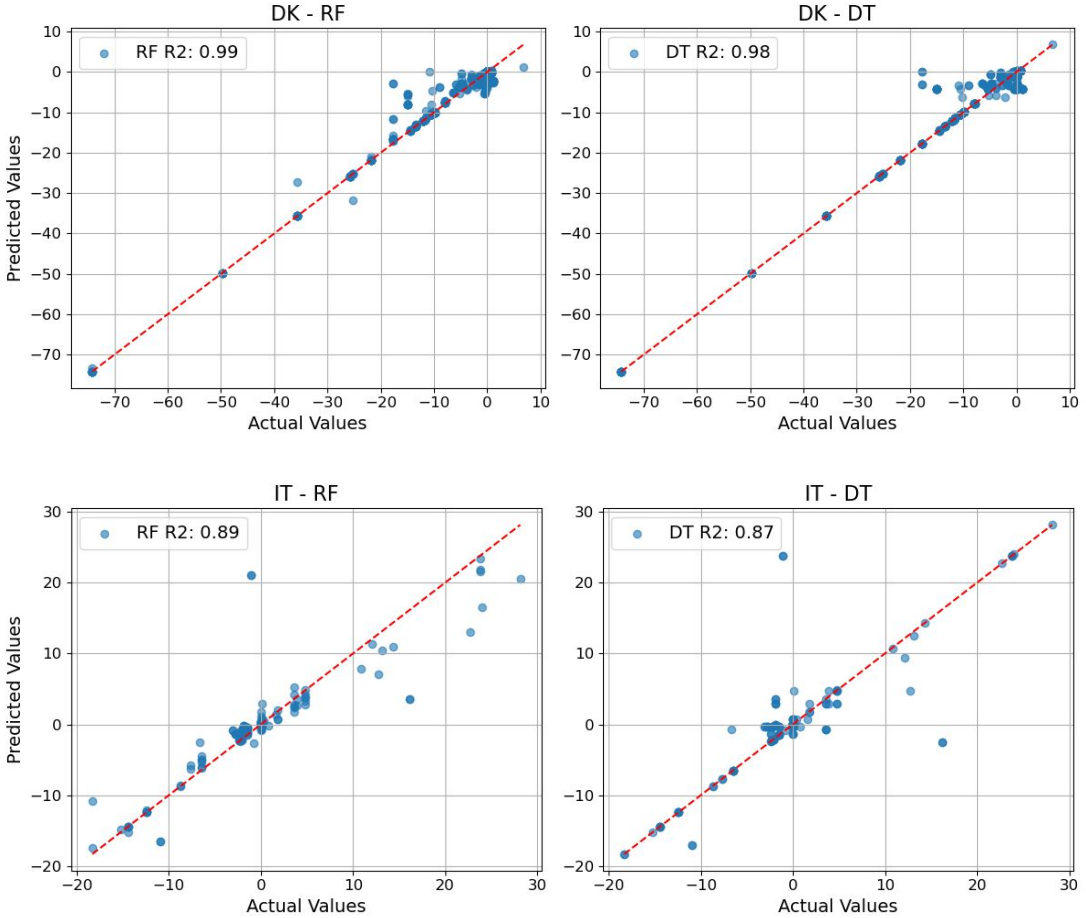
Results (ILI)

- Test set **RMSE** comparison:
  - RF and DT outperform baseline
  - RF: best performance

Country	DT	RF	Baseline	Persistence
BE	9.42	**7.96**	11.1	10.1
DK	6.3	**6.1**	18	14.9
FR	0.69	**0.63**	3.05	5.04
NL	**0.84**	**0.84**	0.97	0.99
NO	**4.2**	4.3	6.6	7.3
IT	3.8	**3.6**	6.3	4.9

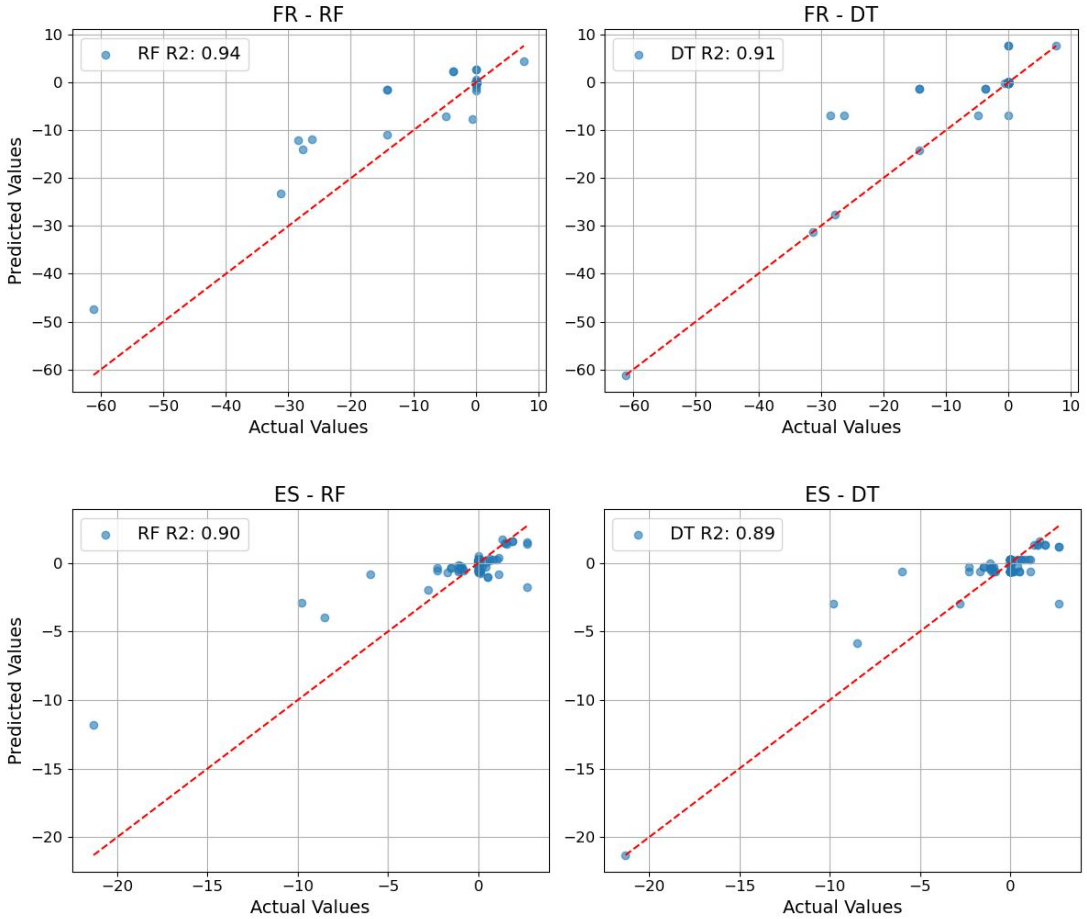
Results (ILI)

- Comparing revision **actual values** with **predicted values** from Decision Tree and Random Forest models
- R^2 confirms better performance of RF over DT



Results (ARI)

- Comparing revision **actual values** with **predicted values** from Decision Tree and Random Forest models
- R^2 confirms better performance of RF over DT



Next Steps

- Use revision magnitude estimation to **improve forecast** model performance
  - Train the forecasting model on data adjusted with estimated revisions and evaluate its performance
  - Compare the results with the same model trained on the original data to assess the impact of revision adjustments

- Extend revision magnitude estimation to other targets:
  - SARI incidence
  - Virological detections (RSV, SARS-CoV-2, . . .)