AstraZeneca

Winter is coming; forecasting the next season's wave using ML

Federico Francone

Clinical Data Science, Infectious Disease

02 June 2026

Company restricted

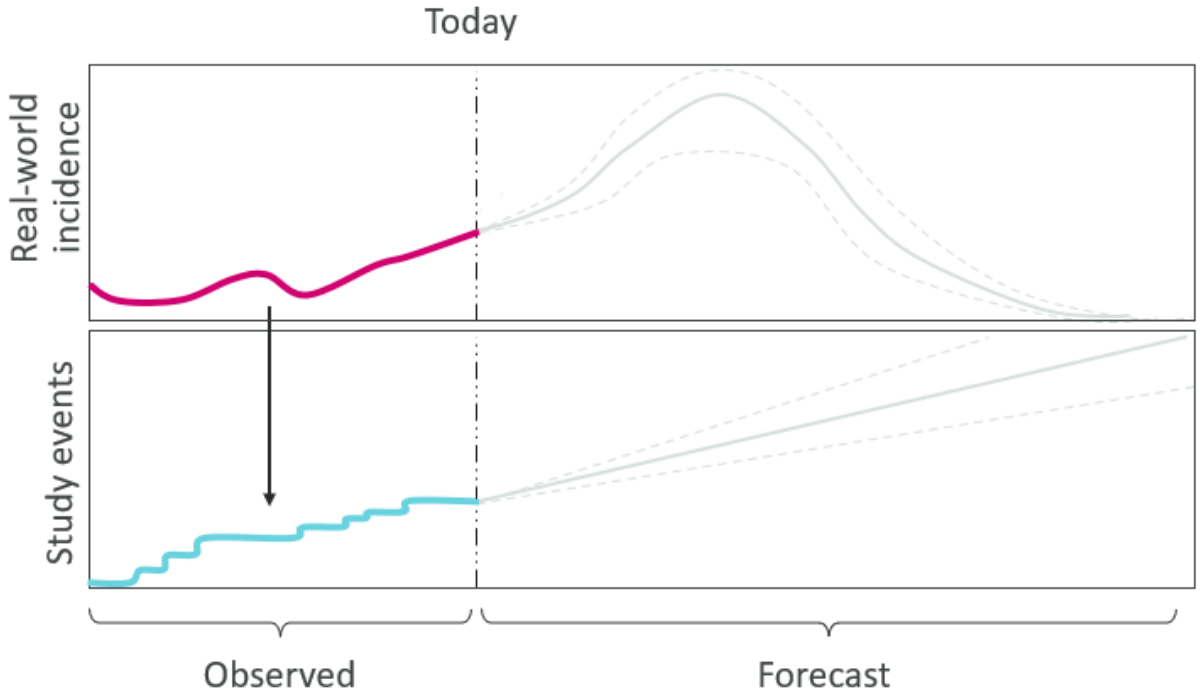
Agenda

- Why we want to forecast disease incidence
- Forecast performance on the latest US RSV Season
- Overview of the Forecasting Methodology:
  - Selection of key time-shifted signals
  - Detection of optimal model stack
  - Model Interpretation
- Conclusions and Next Steps
- Q&A

Accurate study events estimation enables key operational decisions

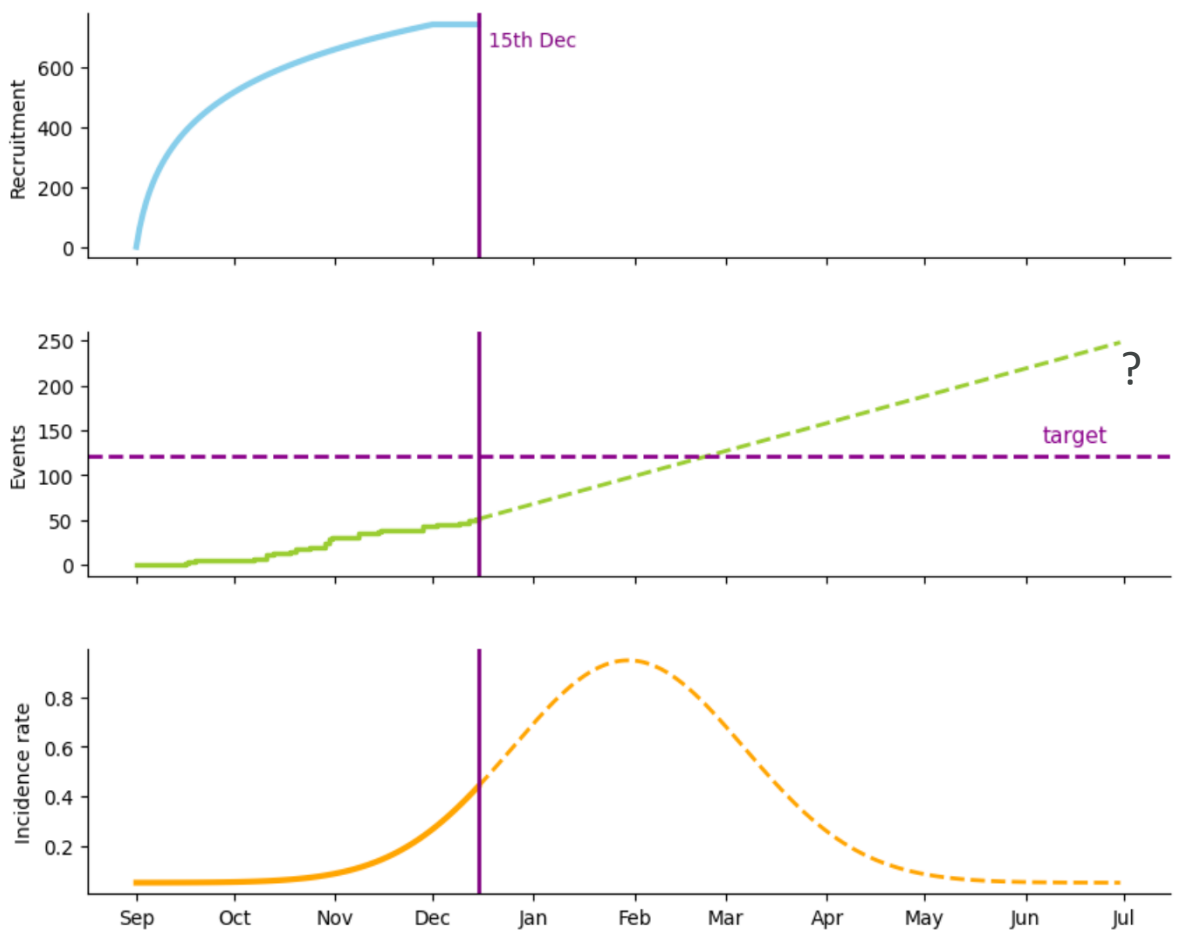
- A key challenge in vaccine efficacy studies is anticipating if and when enough events will accrue to achieve sufficient statistical power for the primary analysis.

- Monitoring ongoing studies and forecast their progress enables timely operational decisions on study design, such as:
  - anticipate the date of primary and interim analyses
  - adjust recruitment to increase probability of technical success.



Seasonal diseases are complex for study design

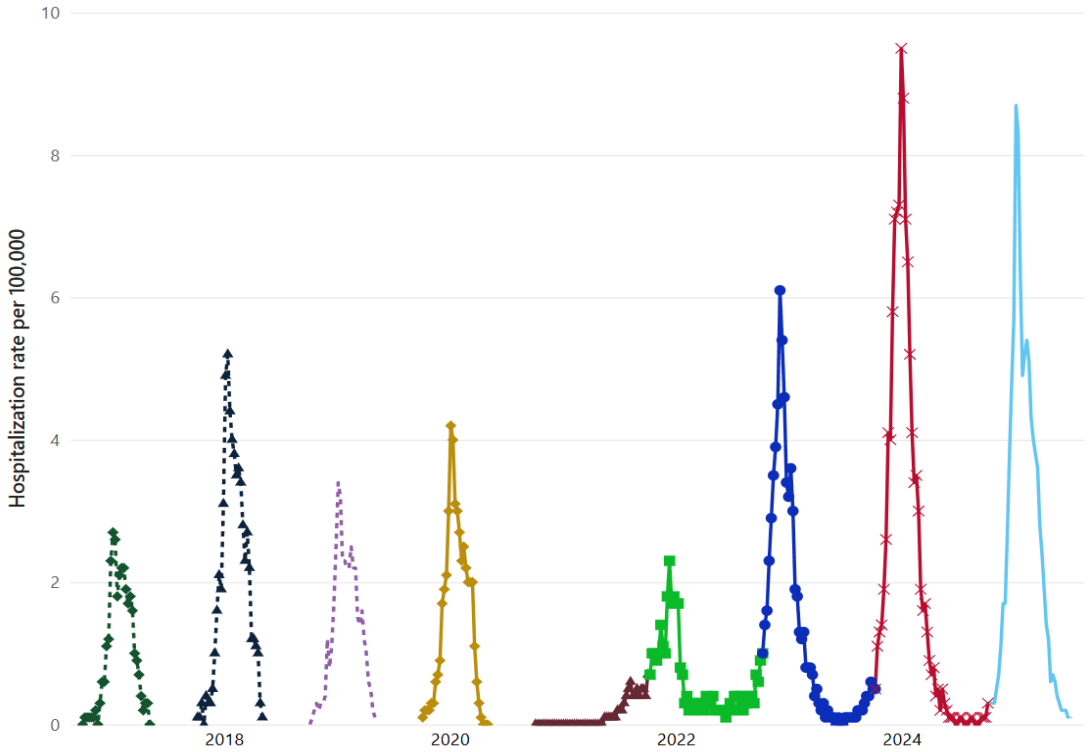
- Infectious diseases bring additional complexity compared to other therapy areas, due to inconsistent seasonal patterns and therefore attack rates.
- We need to forecast incidence rate in advance to plan number of patients to enrol to meet the required endpoints.
- In preparation for RSV efficacy studies - which involve low attack rate and high sample size, we developed a forecasting framework which accurately predicted the season more than six months in advance.
- The same methodology could be leveraged to forecast any seasonal disease incidence rate.

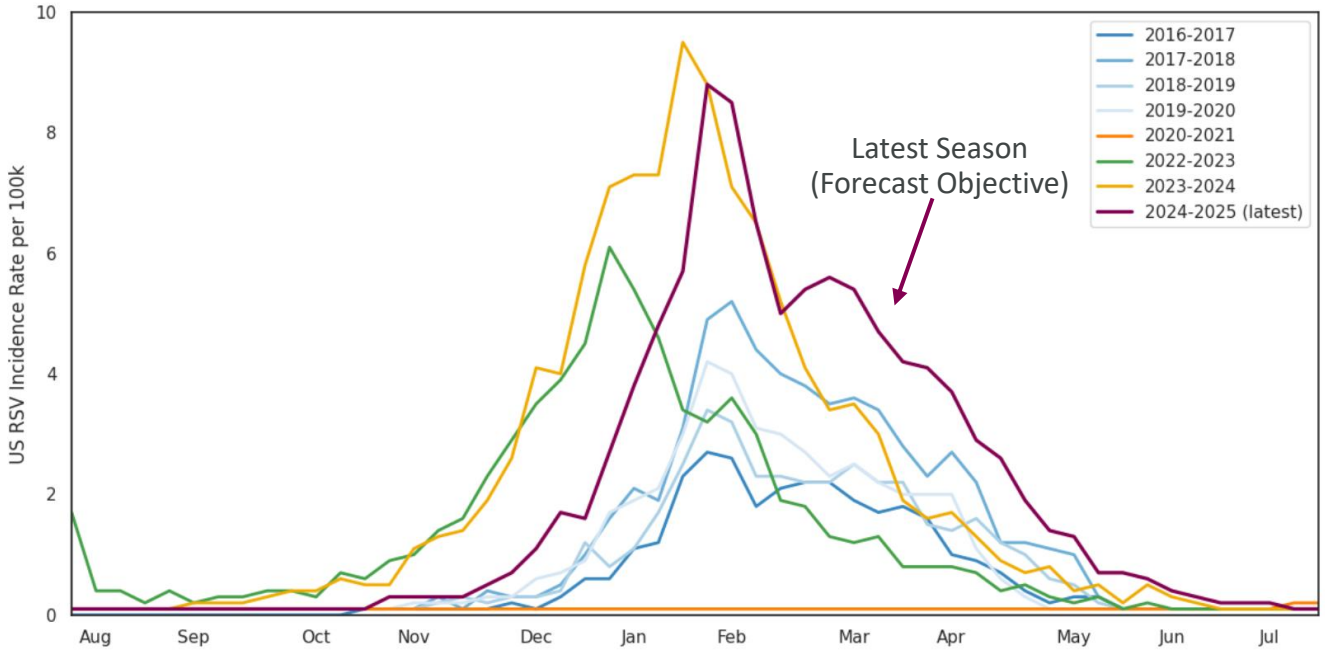


Forecasting enables decision making on study design and monitoring

US RSV seasons show significant changes over time

- Big differences in start time, duration, peak timing and magnitude
- Post-COVID seasons have been larger than before

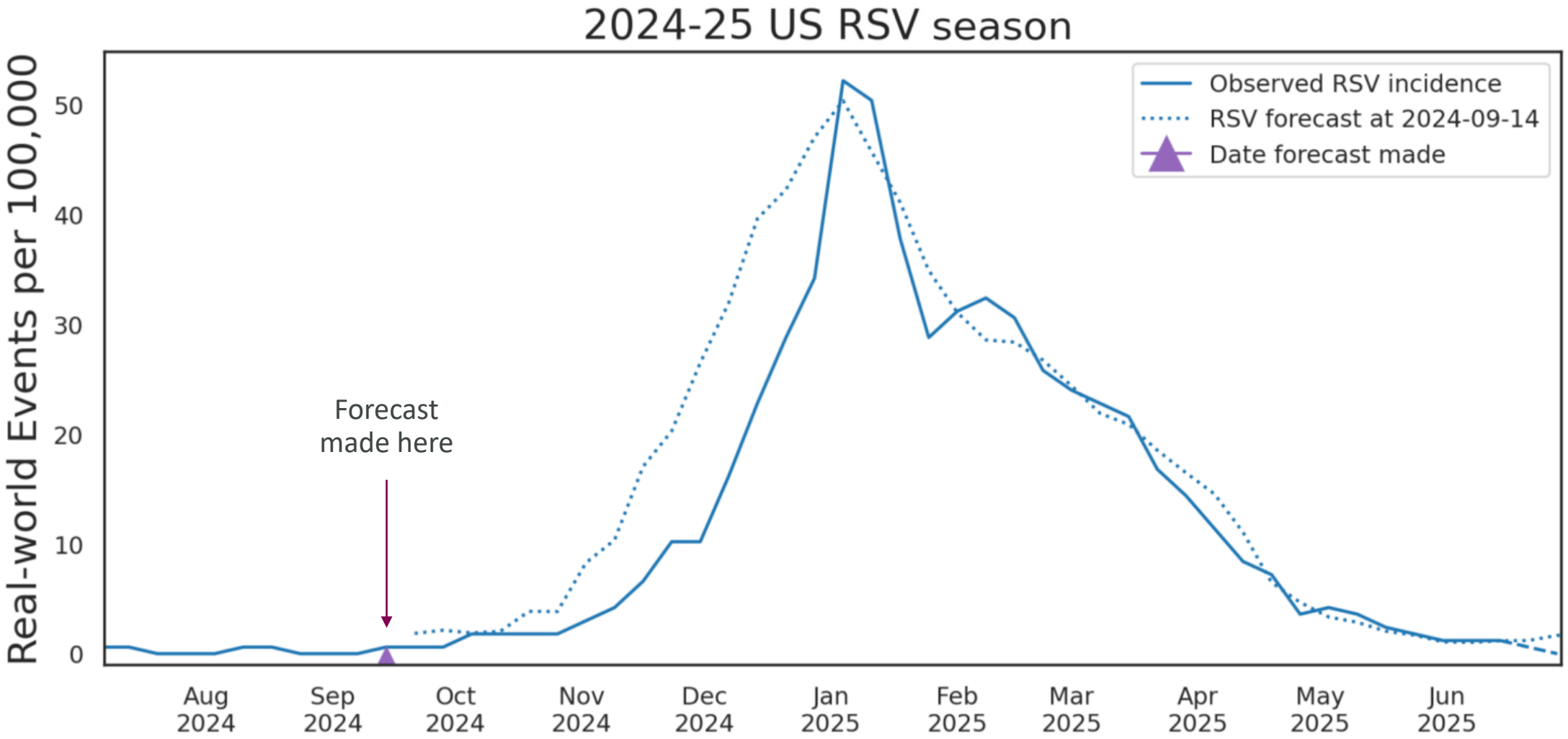




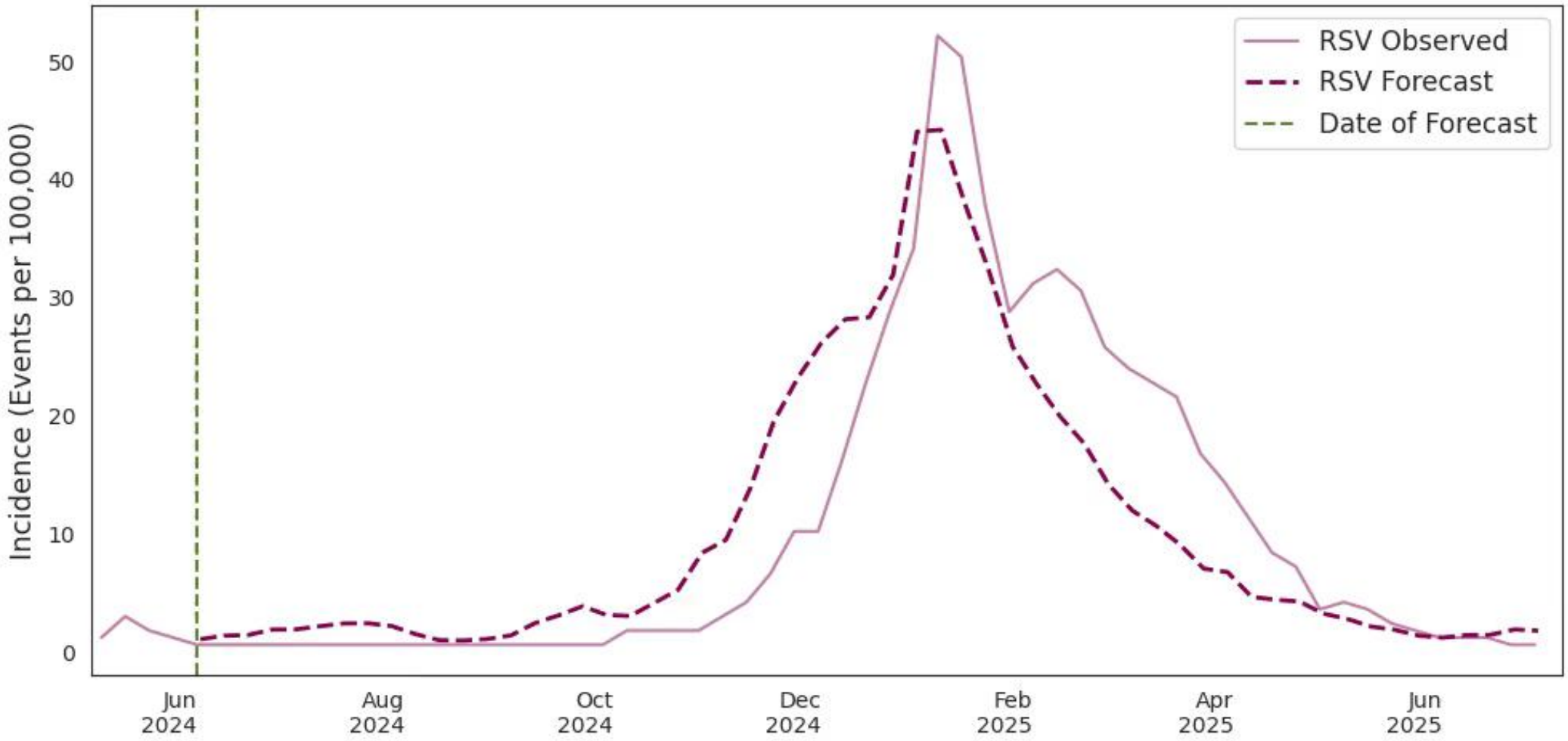
Source: https://www.cdc.gov/rsv/php/surveillance/rsv-net.html

US RSV Incidence Rate for 65+ years population

We accurately predicted the last season before it started



Our forecast refreshes as the season evolves



Accurate predictions require both informative and leading input signals

To find predictive signals for US RSV 65+, we have:

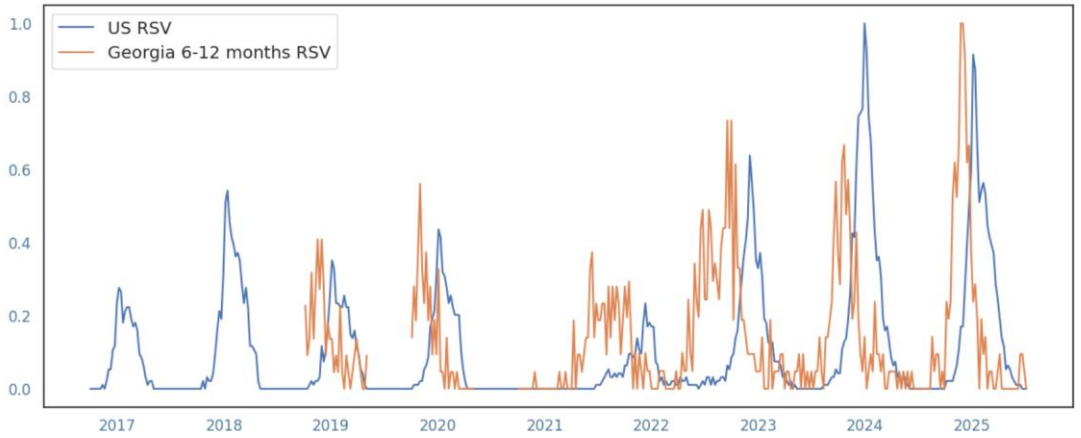
1. **Conducted literature and DB search on relevant data**
   Weather, RSV, Covid and Flu incidence in other countries, RSV incidence in US regions/demographic strata, etc.

2. **Identified candidate predictors** - informative signals which are suitable **leading indicators**.

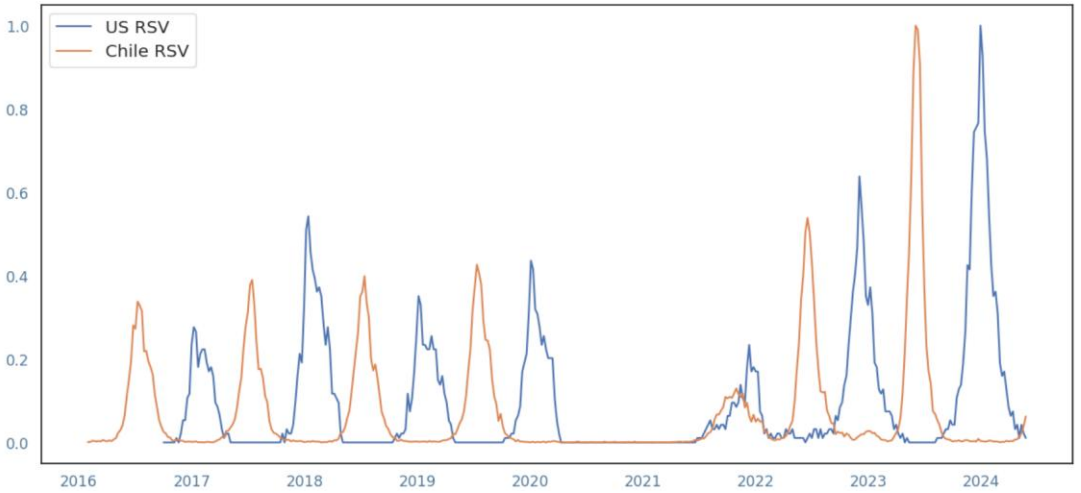
3. **Selected top time-shifted signals** leveraging statistical and ML techniques.

We narrowed down 5000+ potential signals to ~35 high quality, non-redundant time-shifted variables, used as inputs to our models to make predictions.

Georgia 6-12 months RSV



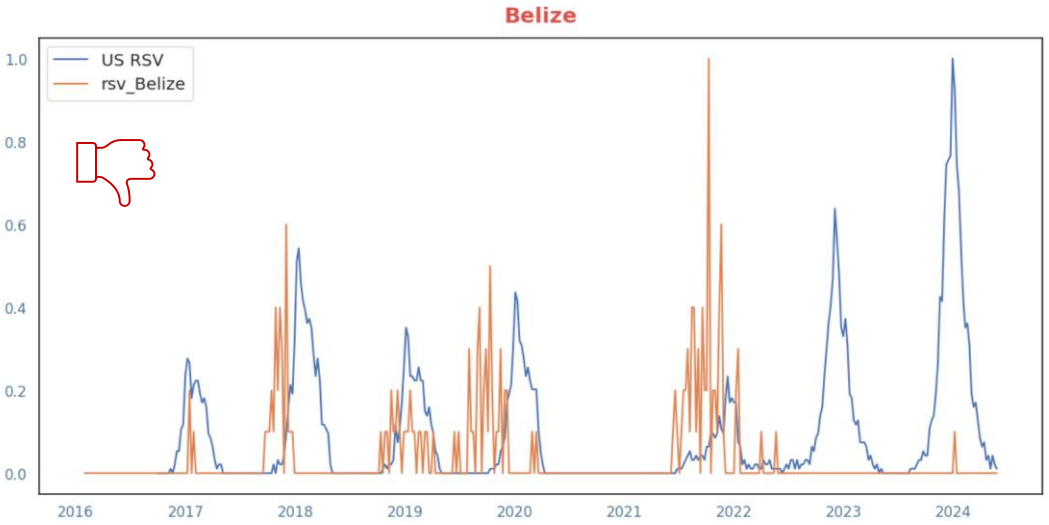
Chile RSV

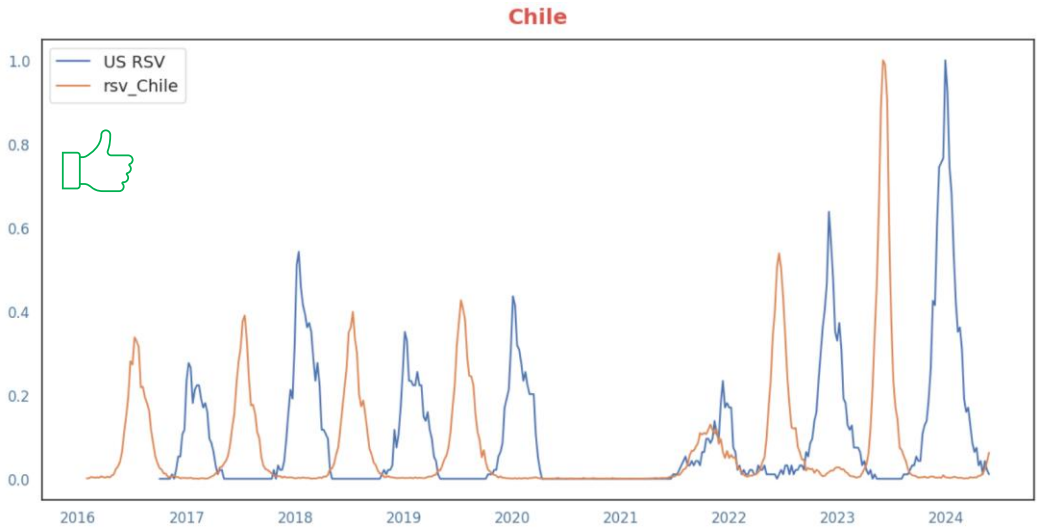


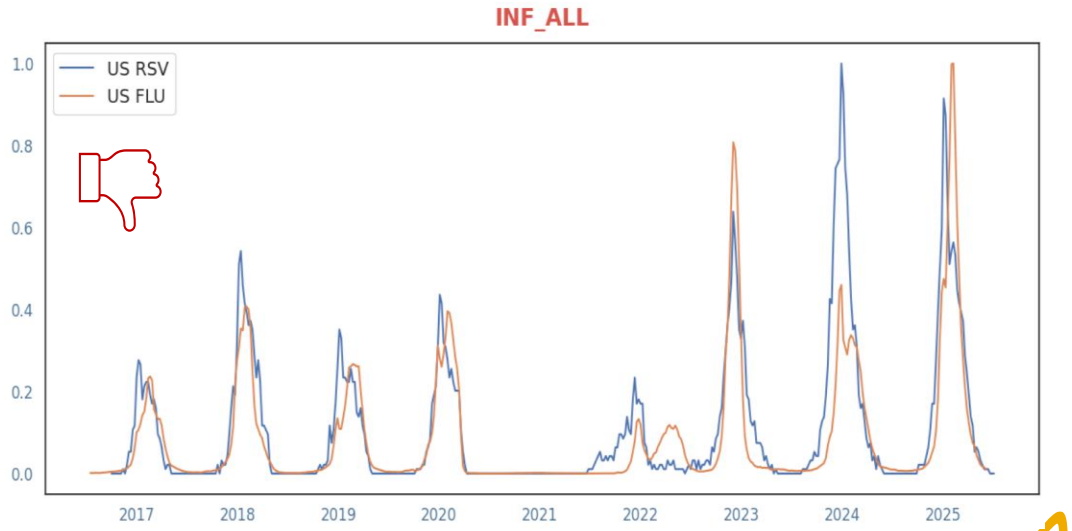
We detect best signals leveraging heuristics and ML methods

A multi-step feature selection approach detects the optimal set of time-lagged predictors to our models:

1. **Data quality** and availability, granularity and timespan
2. **Lead detection & optimal time-shifting**
   - *Custom framework based on Mutual Information*
3. **Correlation** analysis to remove redundant variables
4. **Statistical and ML methods** to detect most informative lagged predictors (*Lasso, XGB)*

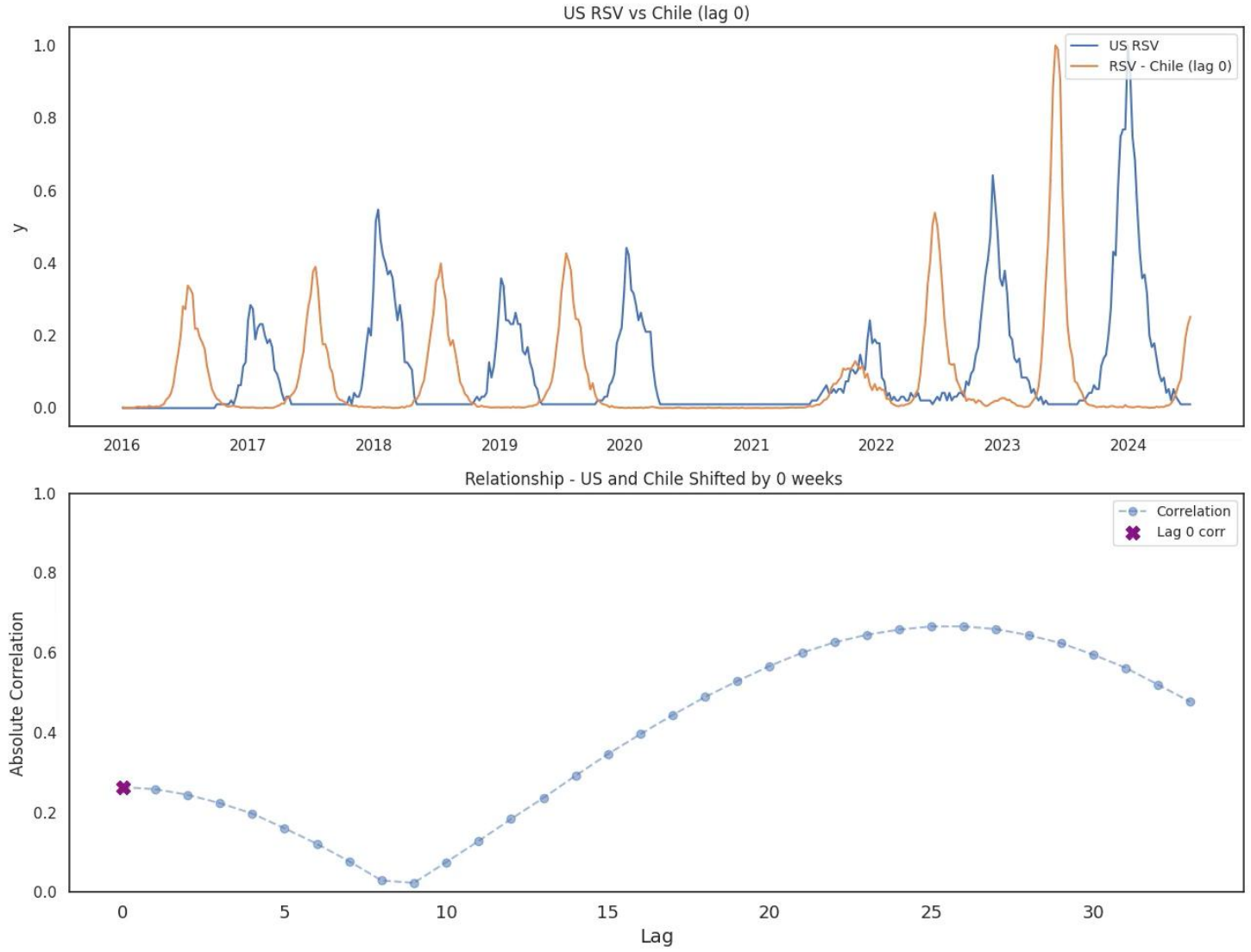






We search for optimal shifts in our predictor signals to improve prediction

- Chile RSV is an example of a candidate leading signal for US RSV. Correlation is maximised when applying a lag of 26 weeks
- This implies that Chile RSV **(predictive signal)** is anticipating US RSV **(target signal)** 26 weeks in advance
- Or, in other words, to forecast US RSV incidence today, we can look at Chile RSV incidence 26 weeks ago
- We consider multiple lags for each predictor, to account for non-constant relationships with target signals over time.



This translates into input data suitable to fit ML models

- Input data structure from previous slide is shown below. On this, we can fit different type of predictive models.
  - Here a simplified version with a few signals and one individual time series.
- Shifted signals enable forecasting into the future: we can use today's Chile RSV to predict 26 weeks ahead

RSV in Chile 26 weeks ago

target

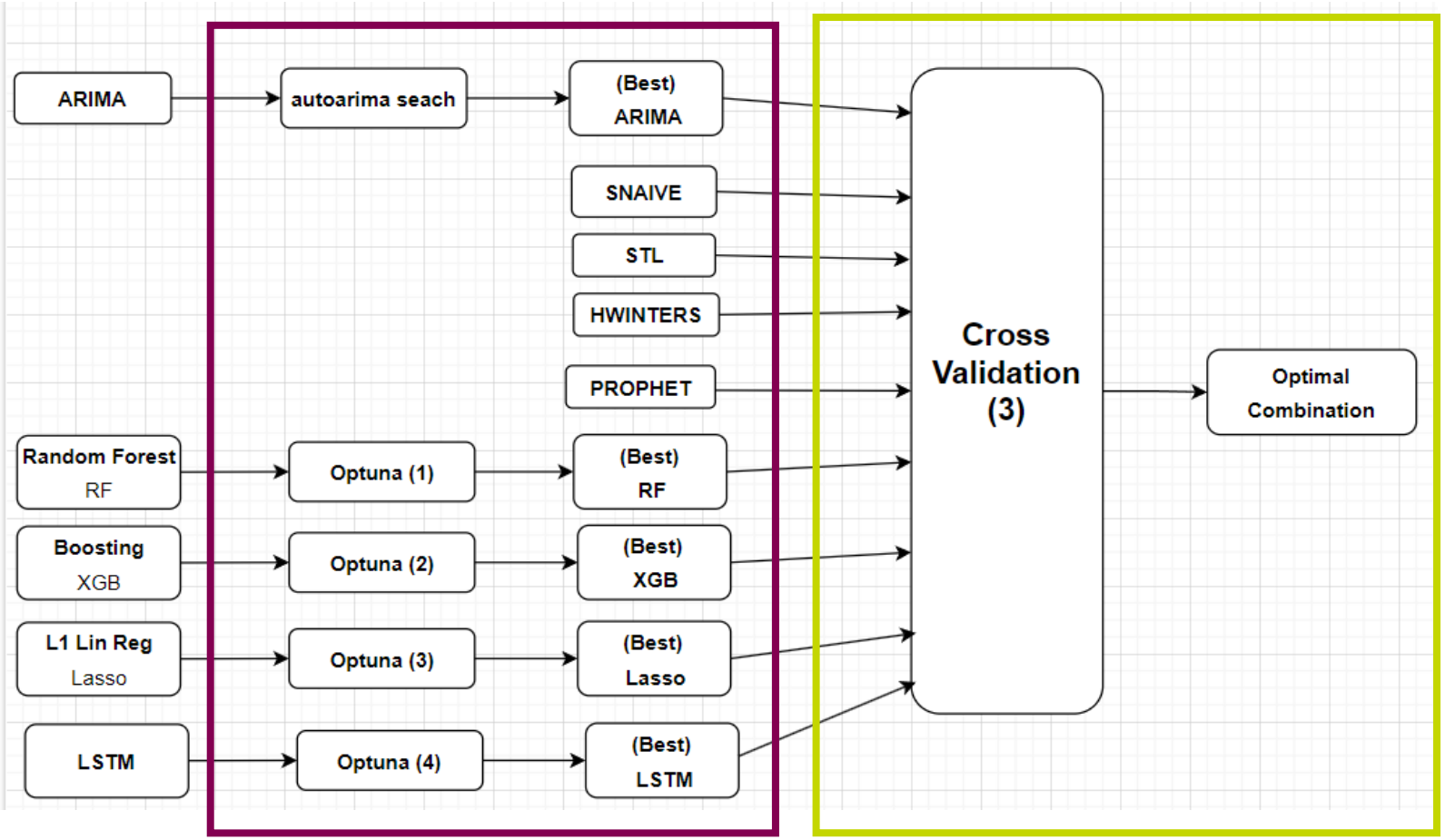
ds	rsv_Chile_lag26	rsv_Japan_lag22	rsv_Nicaragua_lag18	rsv_Uruguay_lag22	rsv_Uruguay_lag28	flu_Togo_lag11	Georgia_6mo12_months_All_All_lag12	y
2017-05-13	0.013020	0.382716	0.018868	0.000000	0.0	0.000000	0.0	0.1
2017-05-20	0.012018	0.296296	0.000000	0.000000	0.0	0.018519	0.0	0.1
2017-05-27	0.005008	0.308642	0.000000	0.015385	0.0	0.000000	0.0	0.1
2017-06-03	0.004006	0.148148	0.000000	0.000000	0.0	0.000000	0.0	0.1
2017-06-10	0.005508	0.148148	0.000000	0.000000	0.0	0.000000	0.0	0.1
...	...	...	...	...	...	...	...	...
2025-08-02	0.007511	0.271605	0.000000	0.000000	0.0	0.000000	0.0	0.1
2025-08-09	0.001502	0.481481	0.009434	0.000000	0.0	0.000000	0.0	0.1
2025-08-16	0.004507	0.469136	0.018868	0.000000	0.0	0.018519	0.0	0.1
2025-08-23	0.005008	0.271605	0.018868	0.000000	0.0	0.000000	0.0	0.1
2025-08-30	0.005008	0.197531	0.028302	0.000000	0.0	0.000000	0.0	0.1

We stack multiple models to make forecast more robust

- Besides local forecasting models (e.g., *ARIMA*), global models are trained on stacked series, leveraging more data to learn shared patterns and improve robustness.

- First, hyperparameter tuning is performed for each individual model.

- Next, all model combinations are compared by averaging their predictions on unseen data.
  The **optimal model stack** is selected based on highest Cross-Validation score.

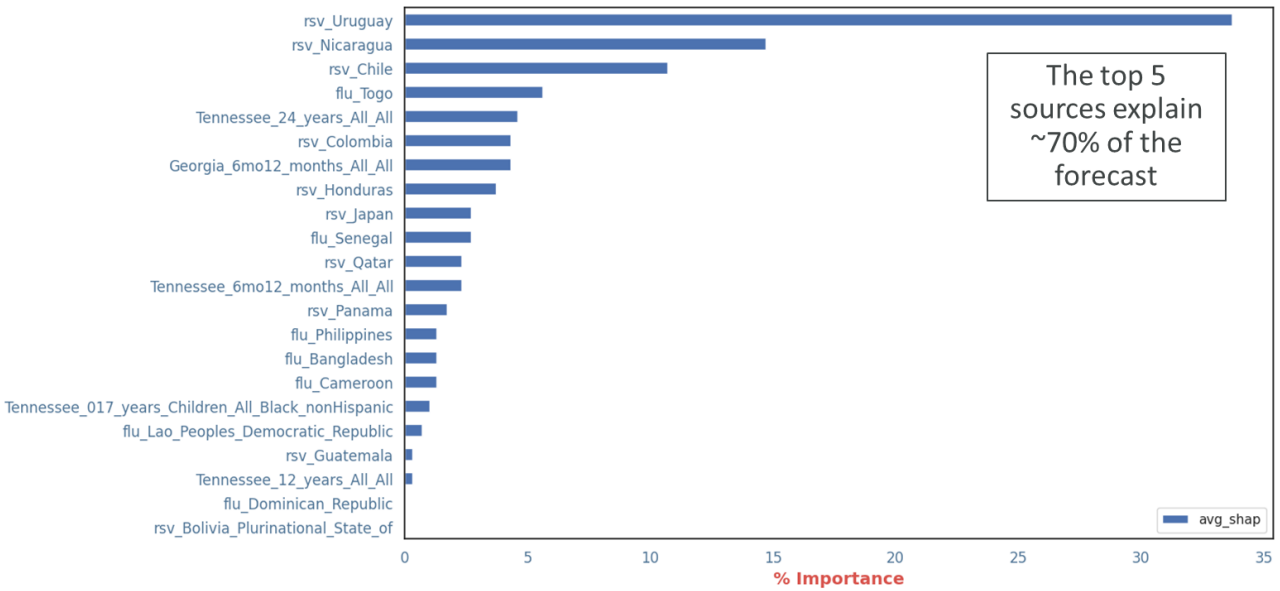


Stacking Model Approach

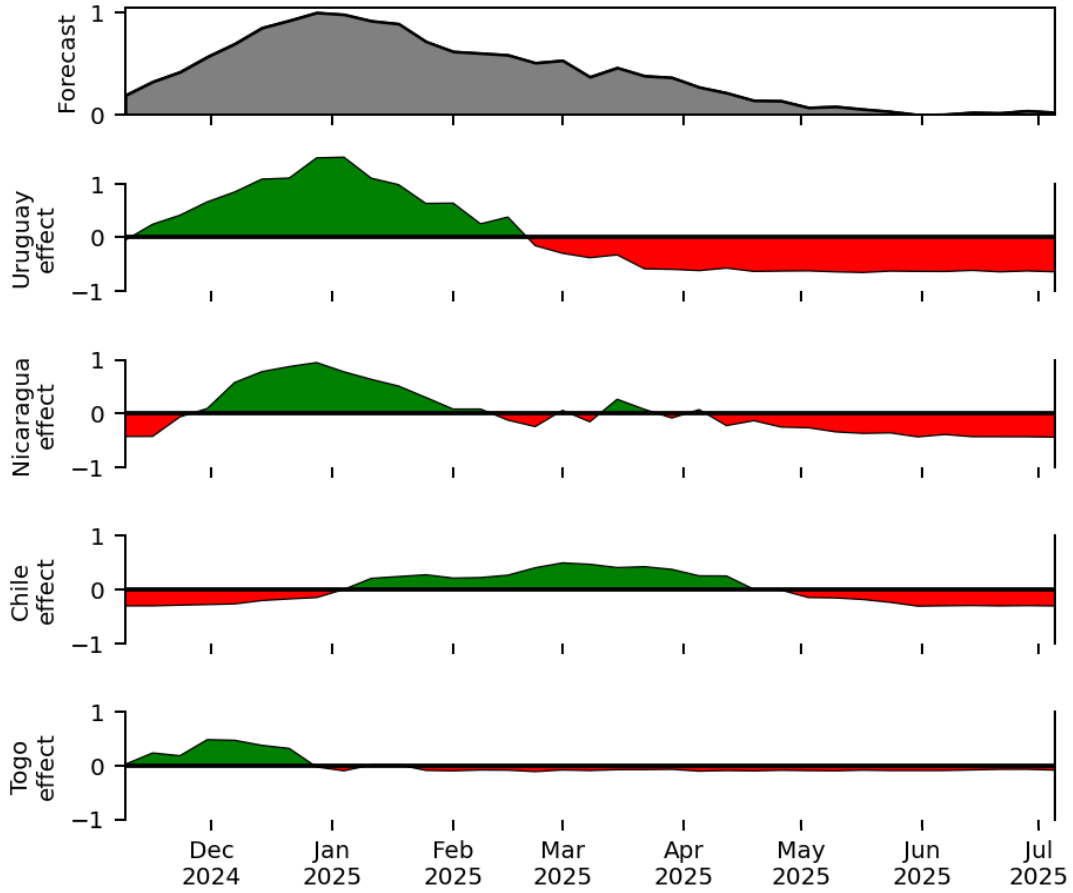
We explore combinations of 20+ models, from traditional Time Series methods, Machine Learning, Deep Learning to recently released TS foundation models (e.g., *TimeGPT*)

Which countries are most important to our forecast?

RSV incidence in south America ~6 months ago is the best predictor of US incidence today



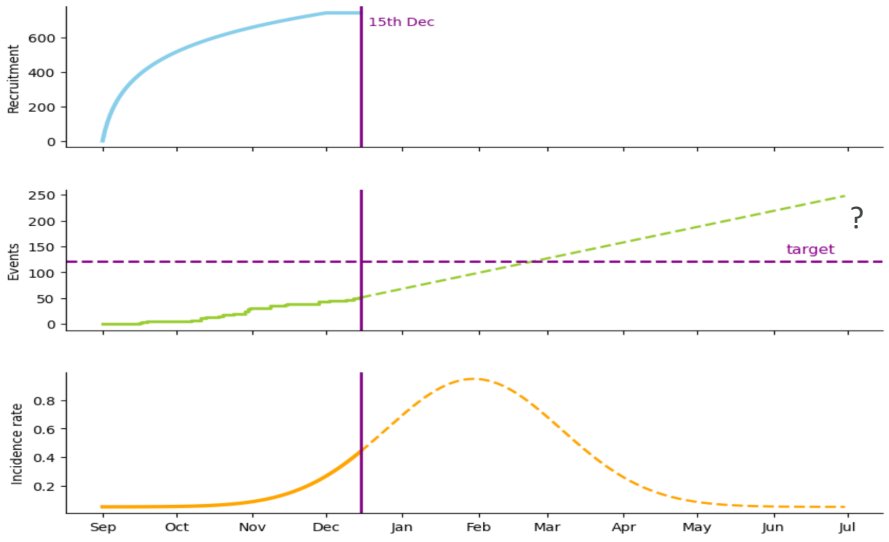
We can show the **most important countries** to our forecasting model



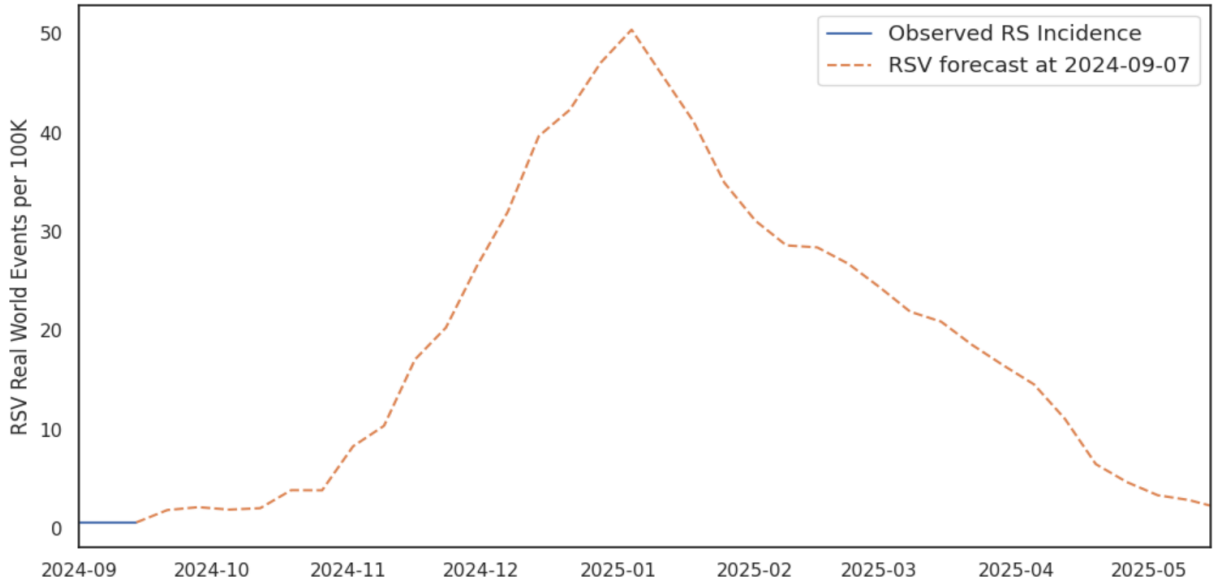
We can show *how each* ***predictive signal affects our forecast over time***

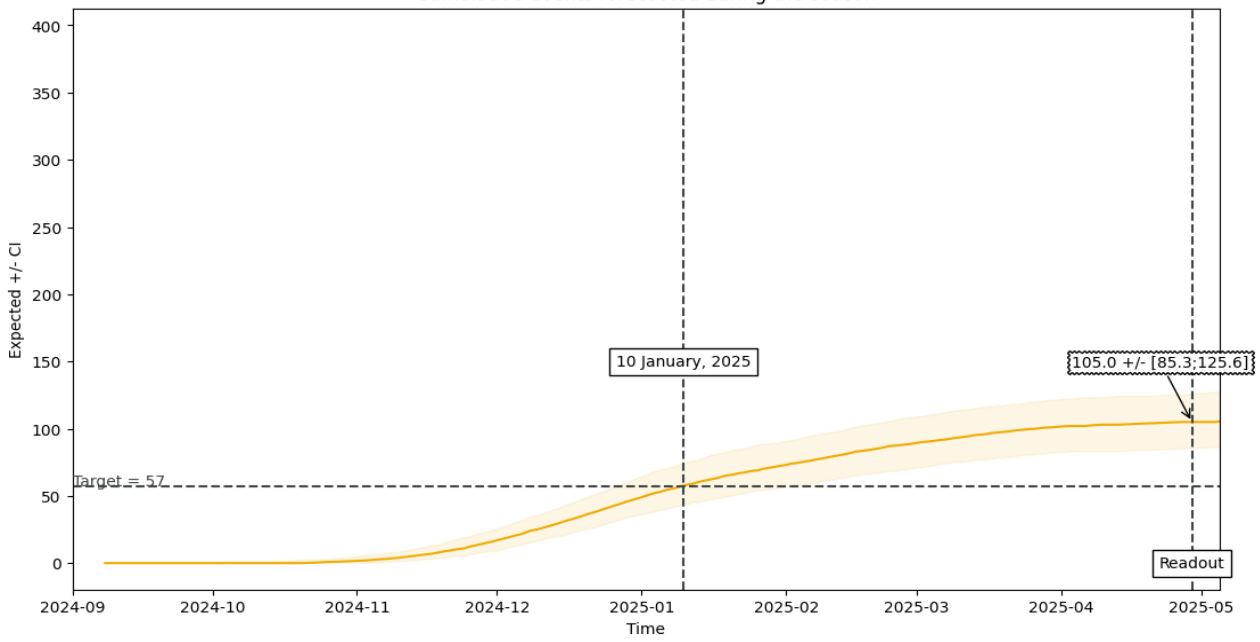
Forecast is combined with recruitment plan to simulate expected study events

- The forecasted real-world incidence powers a probabilistic simulation to estimate the number of events



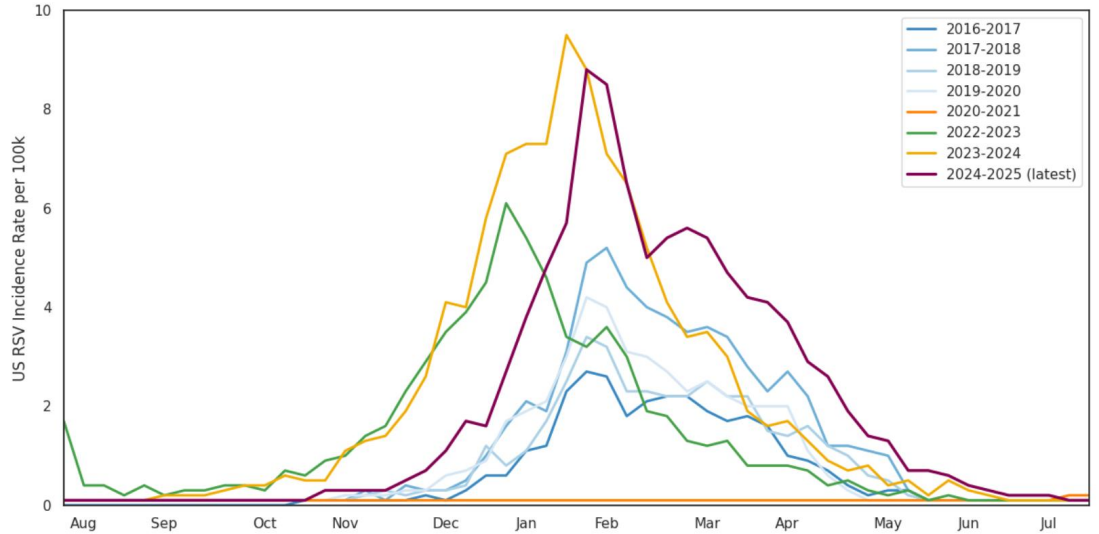
Forecasting enables decision making on study design and monitoring

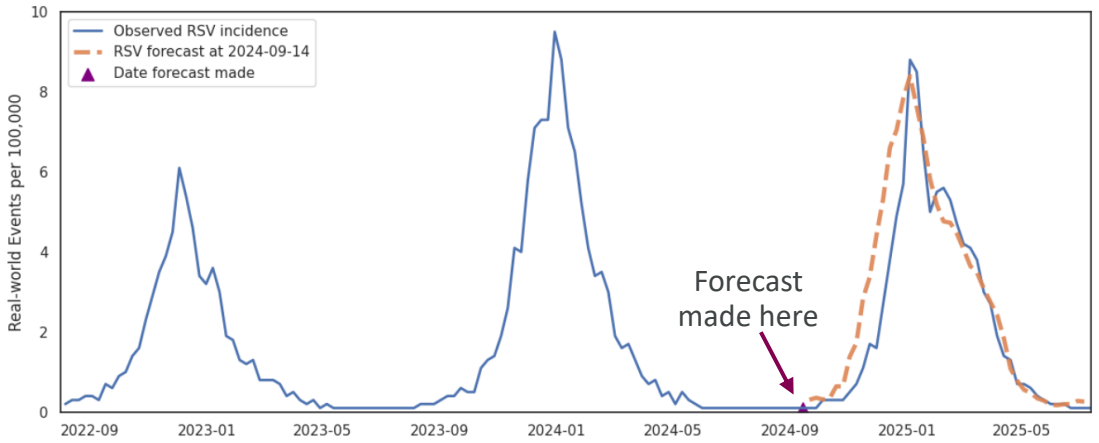




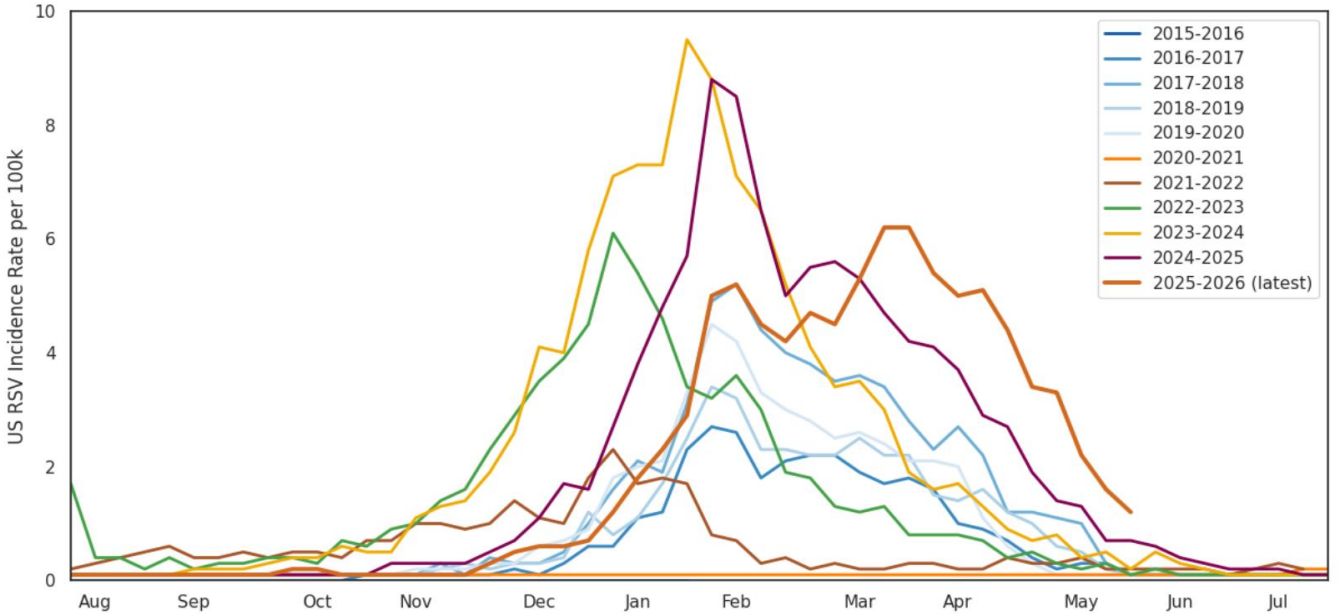
Conclusions and next steps

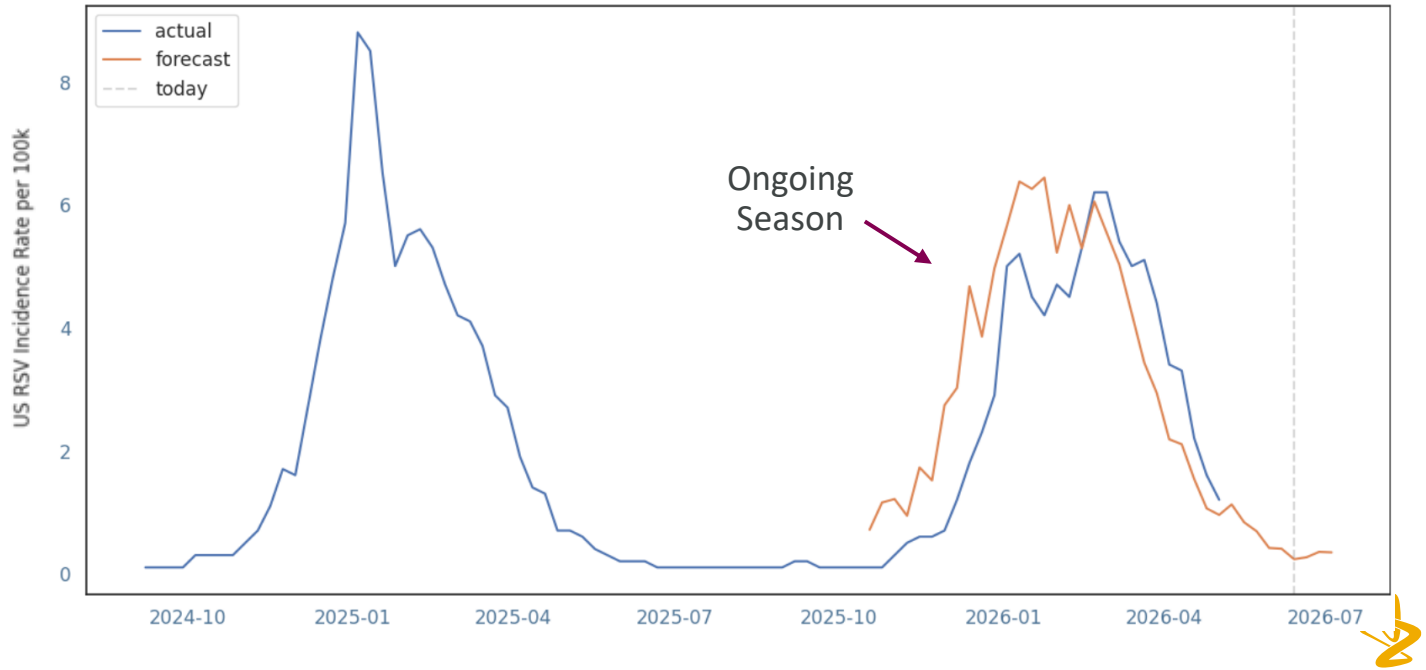
- Accurate forecasts enable decision making on study design and monitoring activities
- This approach delivered an accurate forecast of a US RSV season 6 months in advance, despite shifts in seasonality and season profile post-COVID
- We aim to apply the same methodology to forecast other disease rates in different countries, expanding the set of raw signals to include further disease incidence indicators and vaccination rates





Thank you.

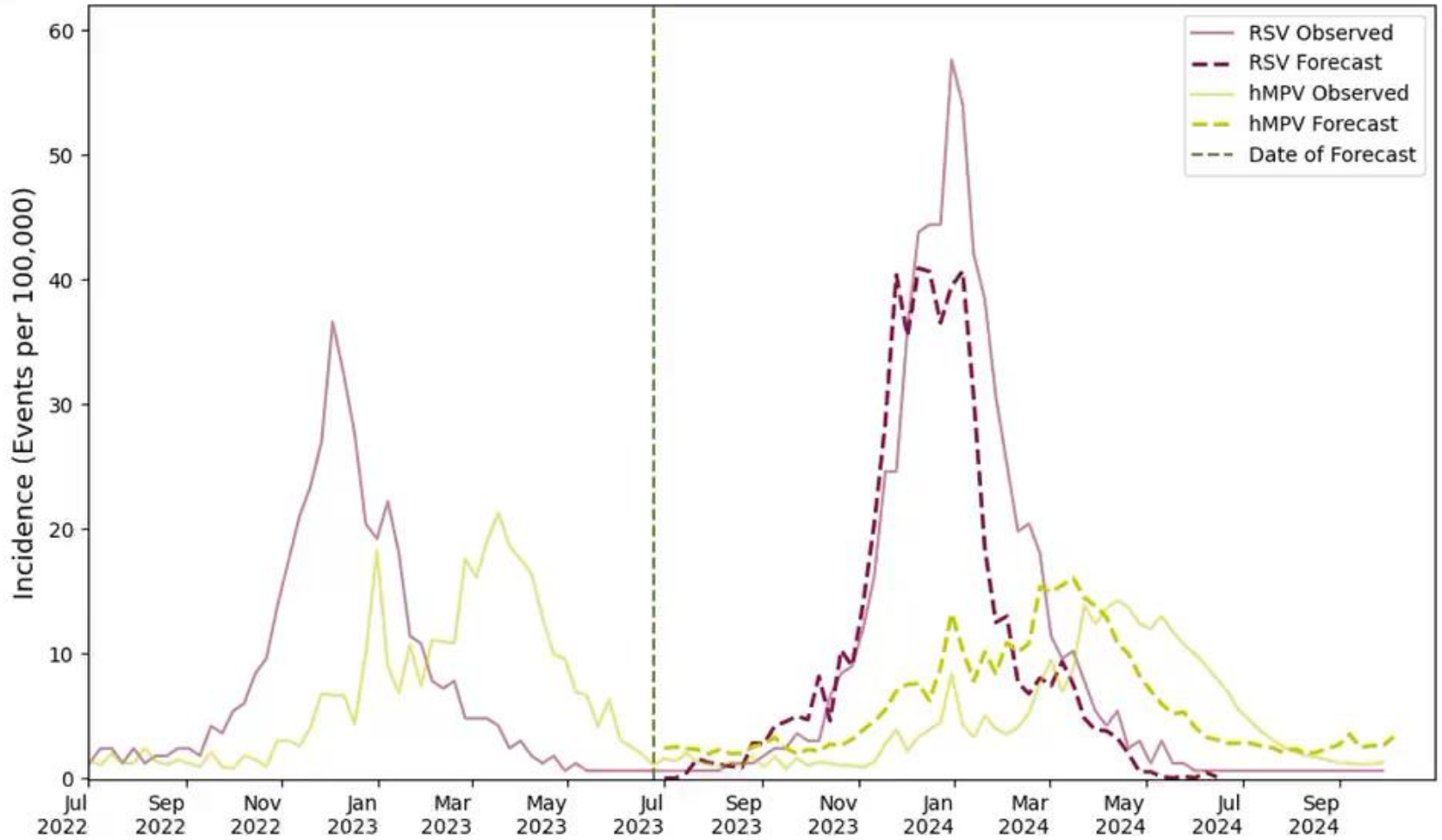

US RSV Incidence Rate per 100k
10
8
6
4
2
0
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
2015-2016
2016-2017
2017-2018
2018-2019
2019-2020
2020-2021
2021-2022
2022-2023
2023-2024
2024-2025
2025-2026 (latest)


actual
forecast
today
US RSV Incidence Rate per 100k
8
6
4
2
0
Ongoing
Season
2024-10
2025-01
2025-04
2025-07
2025-10
2026-01
2026-04
2026-07

5

Appendix

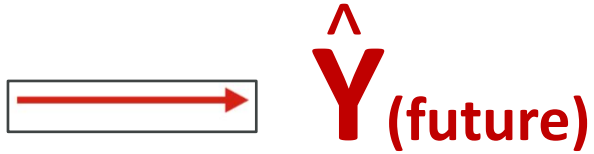
The same model worked well also in the previous seasons



Forecasting into the future

- A simplified version (i.e. few signals) of the input data from the previous slide is shown below. On this, we can fit different type of predictive models.
- To forecast into the future, we'll have same structure of input matrix, with future values in the covariates. The larger the lag, the more historic actual values of the predictor are used.

ds	rsv_Chile_lag30	rsv_Japan_lag22	rsv_Nicaragua_lag18	rsv_Uruguay_lag22	rsv_Uruguay_lag28	flu_Togo_lag11	Georgia_6mo12_months_All_All_lag12
2025-09-06	0.003005	0.148148	0.000000	0.000000	0.000000	0.000000	0.094899
2025-09-13	0.002003	1.012346	0.000000	0.000000	0.000000	0.000000	0.094899
2025-09-20	0.001002	0.901235	0.028302	0.000000	0.000000	0.000000	0.047450
2025-09-27	0.002504	0.691358	0.000000	0.000000	0.000000	0.000000	0.000000
2025-10-04	0.002003	0.604938	0.000000	0.000000	0.000000	0.000000	0.000000
2025-10-11	0.000501	0.320988	0.000000	0.015385	0.000000	0.000000	0.047450
2025-10-18	0.002504	0.456790	0.000000	0.015385	0.000000	0.000000	0.094899
2025-10-25	0.003505	0.469136	0.000000	0.046154	0.000000	0.018519	0.047450
2025-11-01	0.006510	0.234568	0.009434	0.153846	0.000000	0.037037	0.000000
2025-11-08	0.006009	0.382716	0.009434	0.076923	0.000000	0.148148	0.000000
2025-11-15	0.006510	0.432099	0.009434	0.200000	0.000000	0.000000	0.047450



Our predictions uses multiple Forecasting Models

- Our final forecast is the combination of different Statistical and Machine Learning

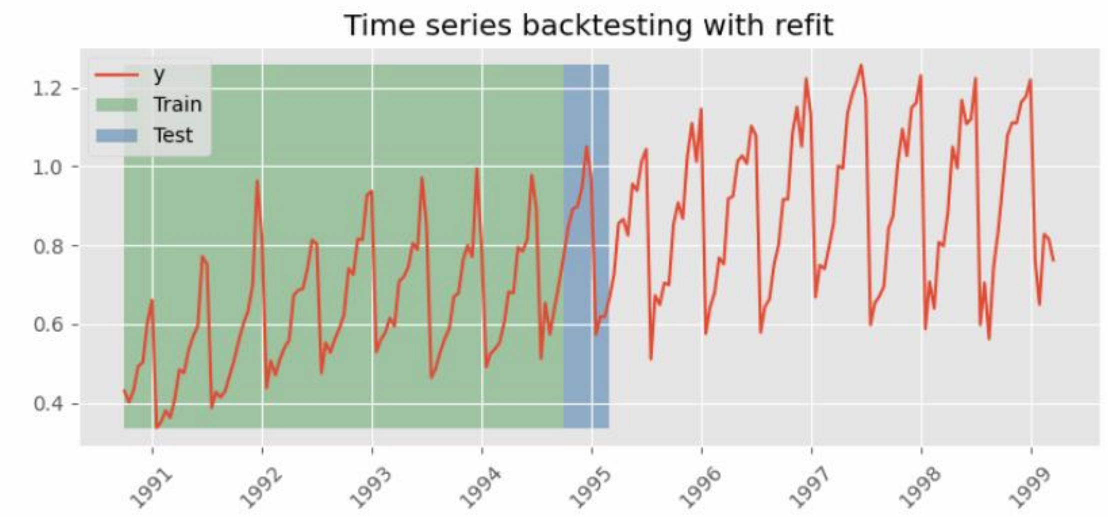
$$\text{Forecast}_{\text{final}} = w_1 \cdot \text{Forecast}_{\text{model 1}} + w_2 \cdot \text{Forecast}_{\text{model 2}} + w_3 \cdot \text{Forecast}_{\text{model 3}}$$

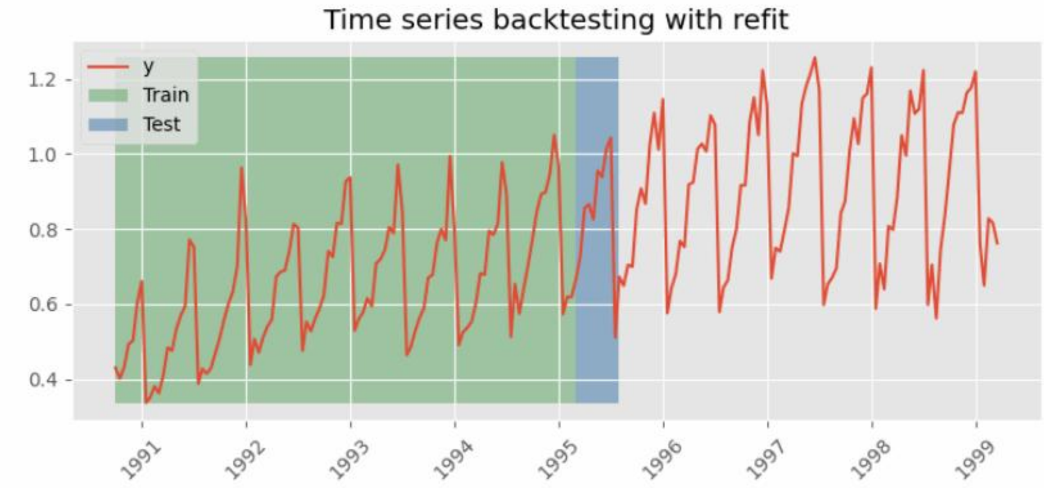
- Each model predicts future US RSV incidence as a function of our 35 top RSV and FLU

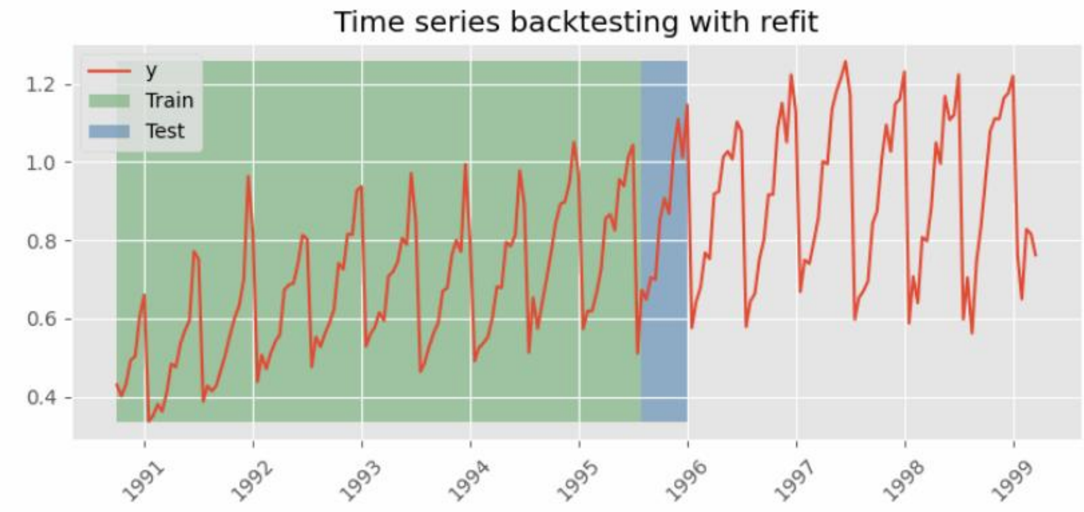
$$\text{Forecast}_{\text{US RSV, next week}} = f\left(\text{RSV}_{\text{Uruguay, lag 28}},\ \text{RSV}_{\text{Uruguay, lag 22}},\ \text{Flu}_{\text{Togo, lag 9}},\ \ldots\right)$$

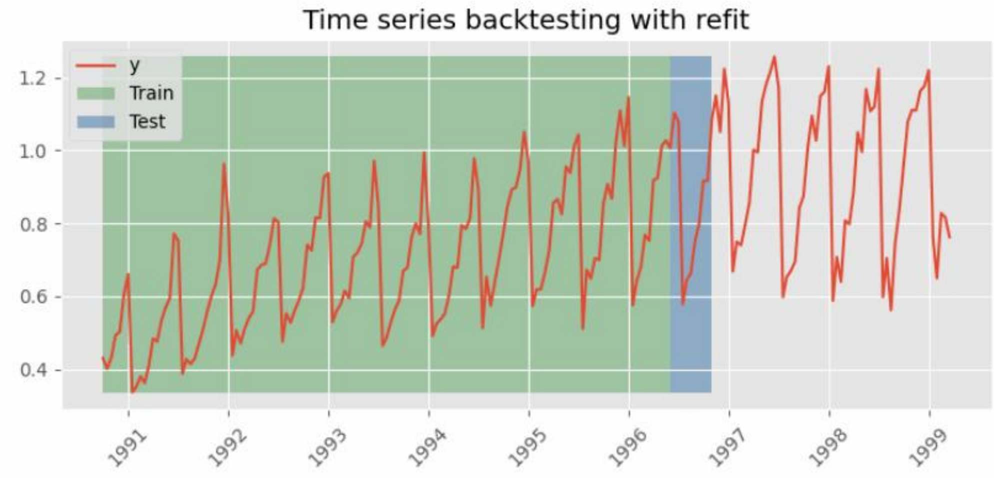
$$\text{Forecast}_{\text{model 1}} = a_1 \cdot \text{RSV}_{\text{Uruguay, lag 28}} + a_2 \cdot \text{RSV}_{\text{Uruguay, lag 22}} + a_3 \cdot \text{RSV}_{\text{Chile, lag 30}} + \ldots + a_{35} \cdot \text{Flu}_{\text{Togo, lag 9}}$$

Time Series Cross Validation









Backtesting with refit and increasing training size (fixed origin).

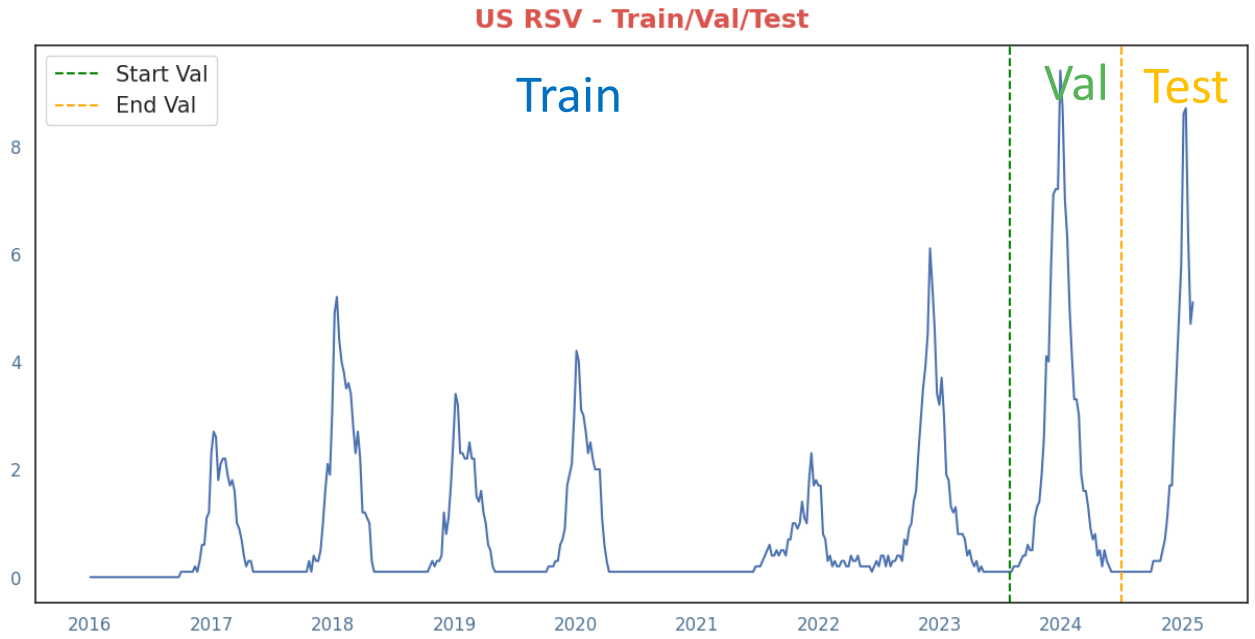
Backtesting with refit and increasing training size (fixed origin).

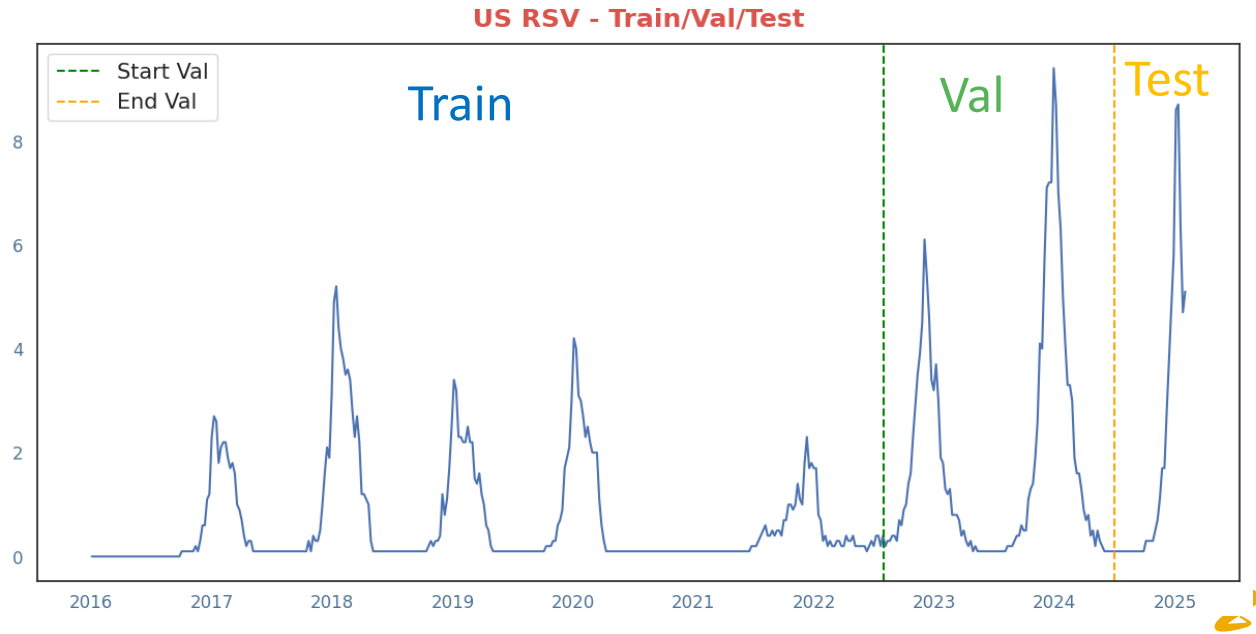
Which one to choose?

- If we had to train a new model to forecast 2025/26 season, which set would you choose to select the Optimal Ensemble?

- There is no an official correct answer, different trade-offs to account for.

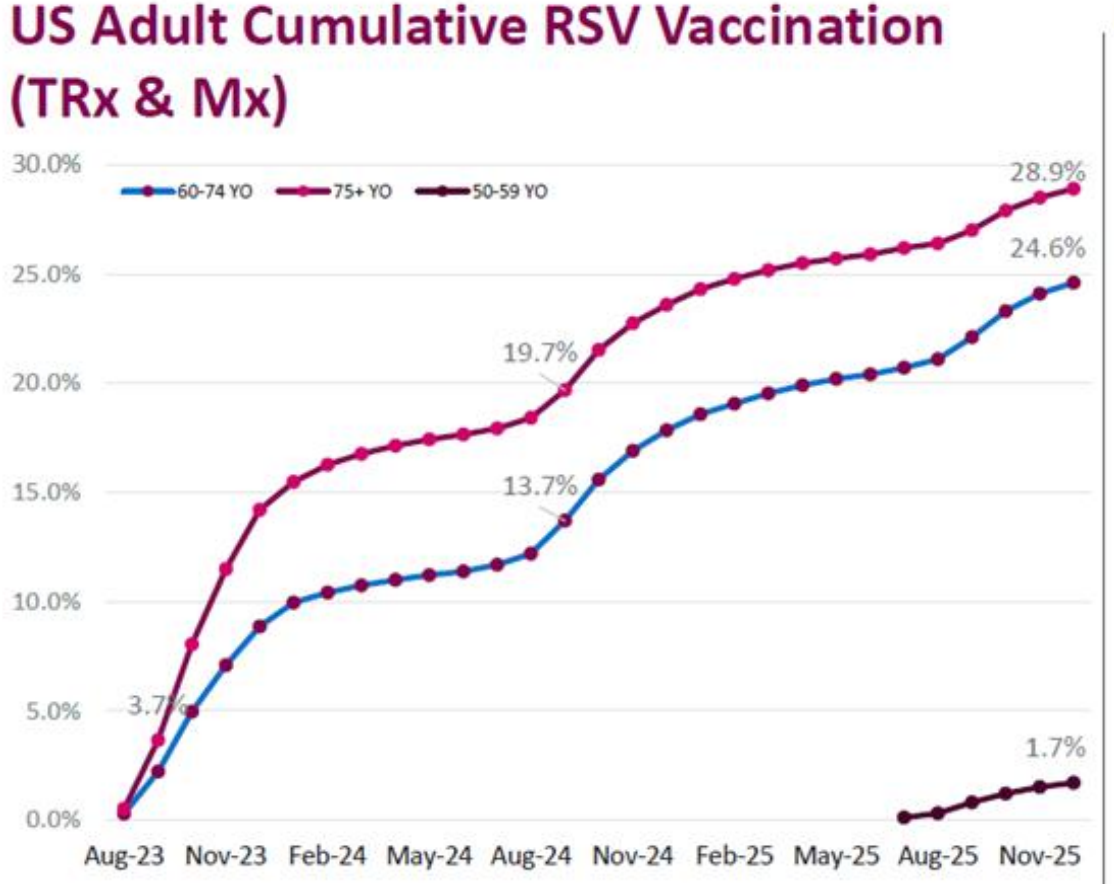
- Key Point: In Regression we can shuffle data points, in TS no. Distributions can significantly change.





US RSV - Vaccination uptake over time

- The latest vaccination data shows that the uptake for the target population is roughly 27% as of Q4 2025
- Uptake has increased roughly by 10% in each RSV season (Aug – July)
- We've implemented a way to discount vaccination impact from our baseline incidence rate forecast



Source: Vaccination Claim data (CDC)