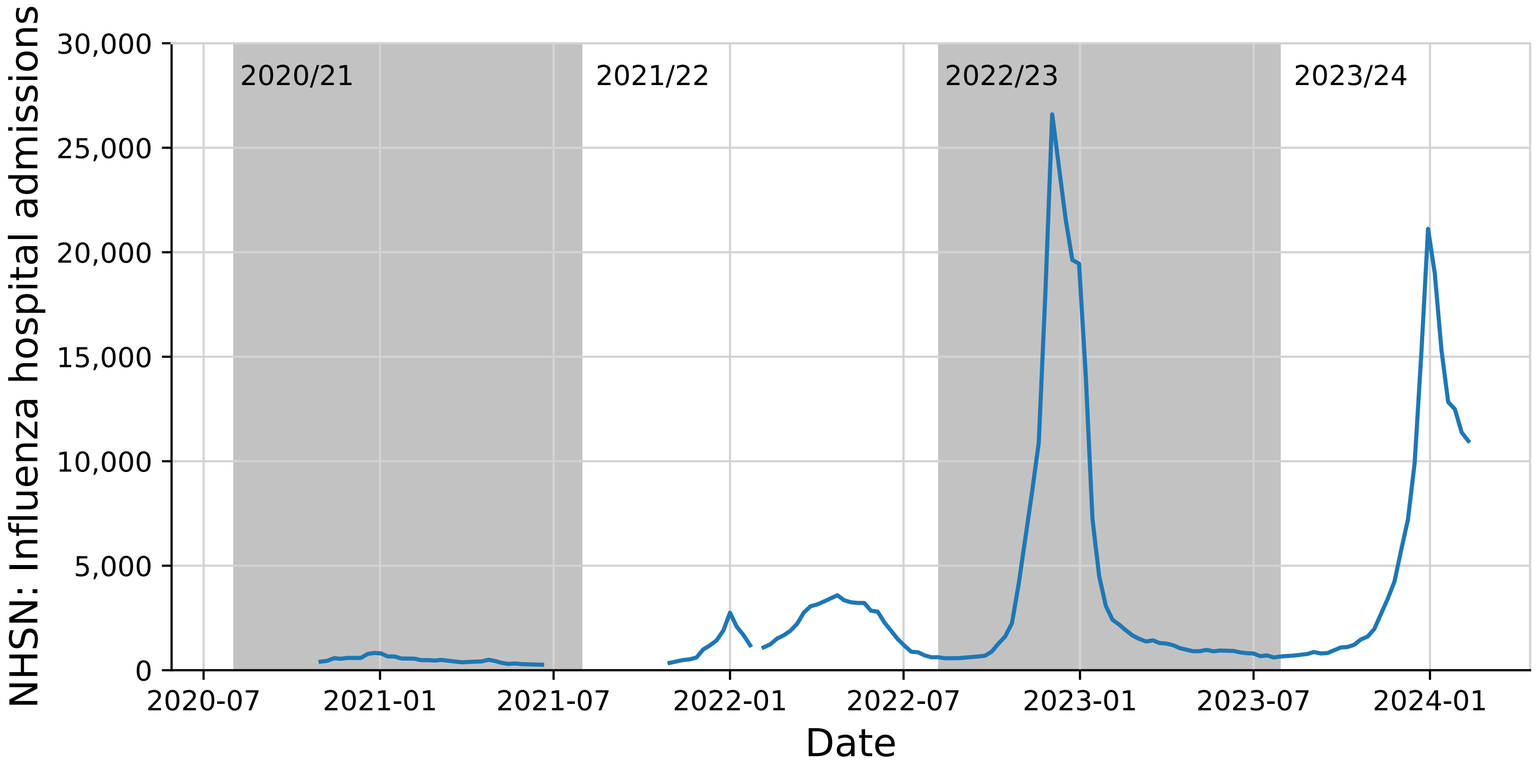
Data sources for flu forecasting

Evan L. Ray
University of Massachusetts, Amherst
MIDAS/CDC Infectious Disease Forecasting
2024-02-16

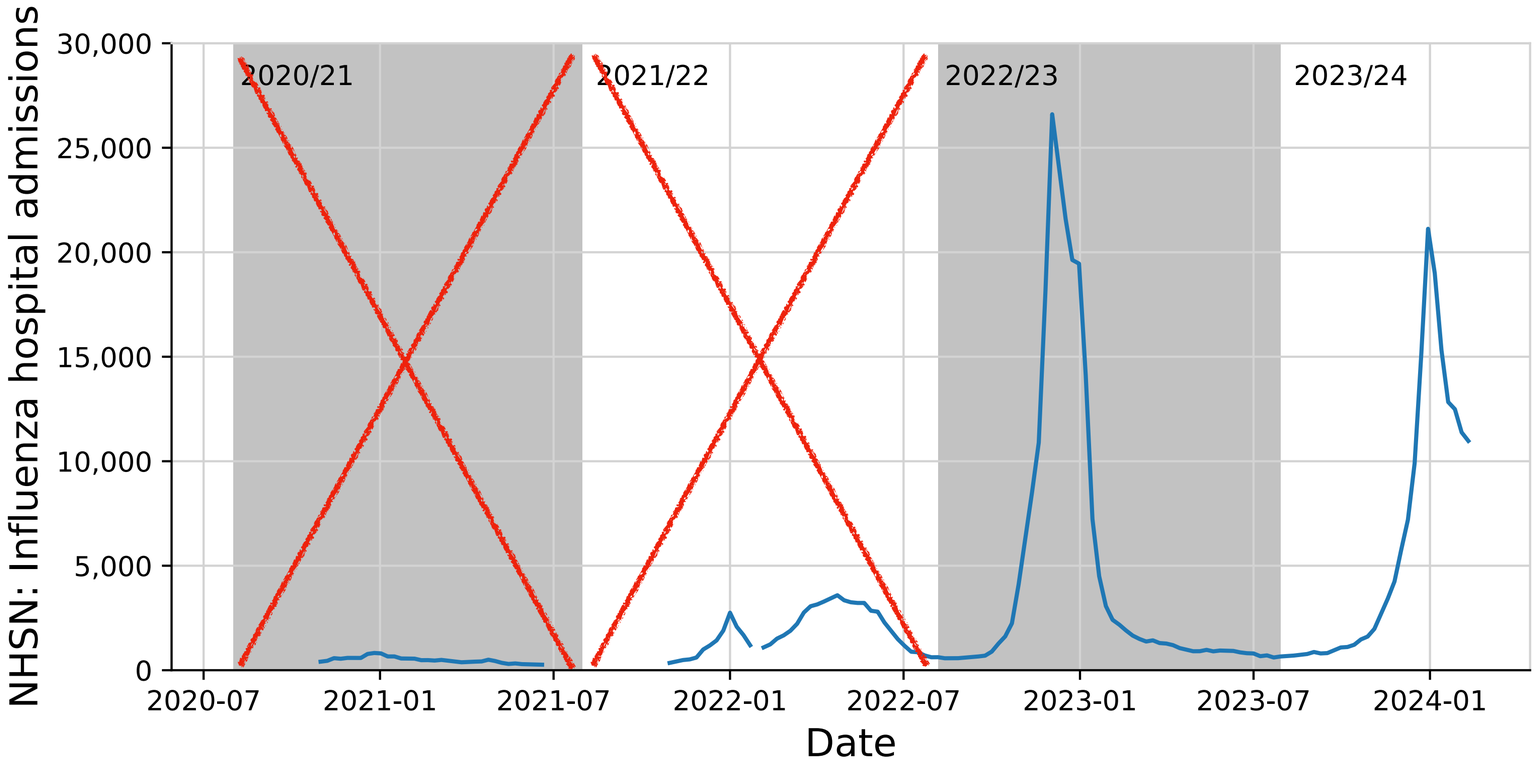
Motivation

- The target for the FluSight forecasting exercise is hospital admissions as reported by NHSN



Motivation

- The target for the FluSight forecasting exercise is hospital admissions as reported by NHSN



- Challenge: at the start of this season, this source had only 1 useful past season of data
  - this data stream had only been active for ~3 seasons
  - 2020/21 had very limited flu activity
  - 2021/22 had limited flu activity and inconsistent reporting

3 signals of influenza activity

The UMass-Flusion model is trained on 3 signals of influenza activity:

1. Hospital admissions as reported by NHSN
   - This is the prediction target for FluSight
   - We drop data prior to the 2022/23 season; **1 season of history**
2. FluSurv-NET, adjusted for detection rates
   - Laboratory confirmed influenza hospitalizations at sites in ~14 states
   - We drop data prior to the 2010/11 season; **11 seasons of history**
3. ILI+
   - An estimate of influenza activity among outpatient doctor visits
   - We have **26 seasons of data at regional level, 11 at state level**

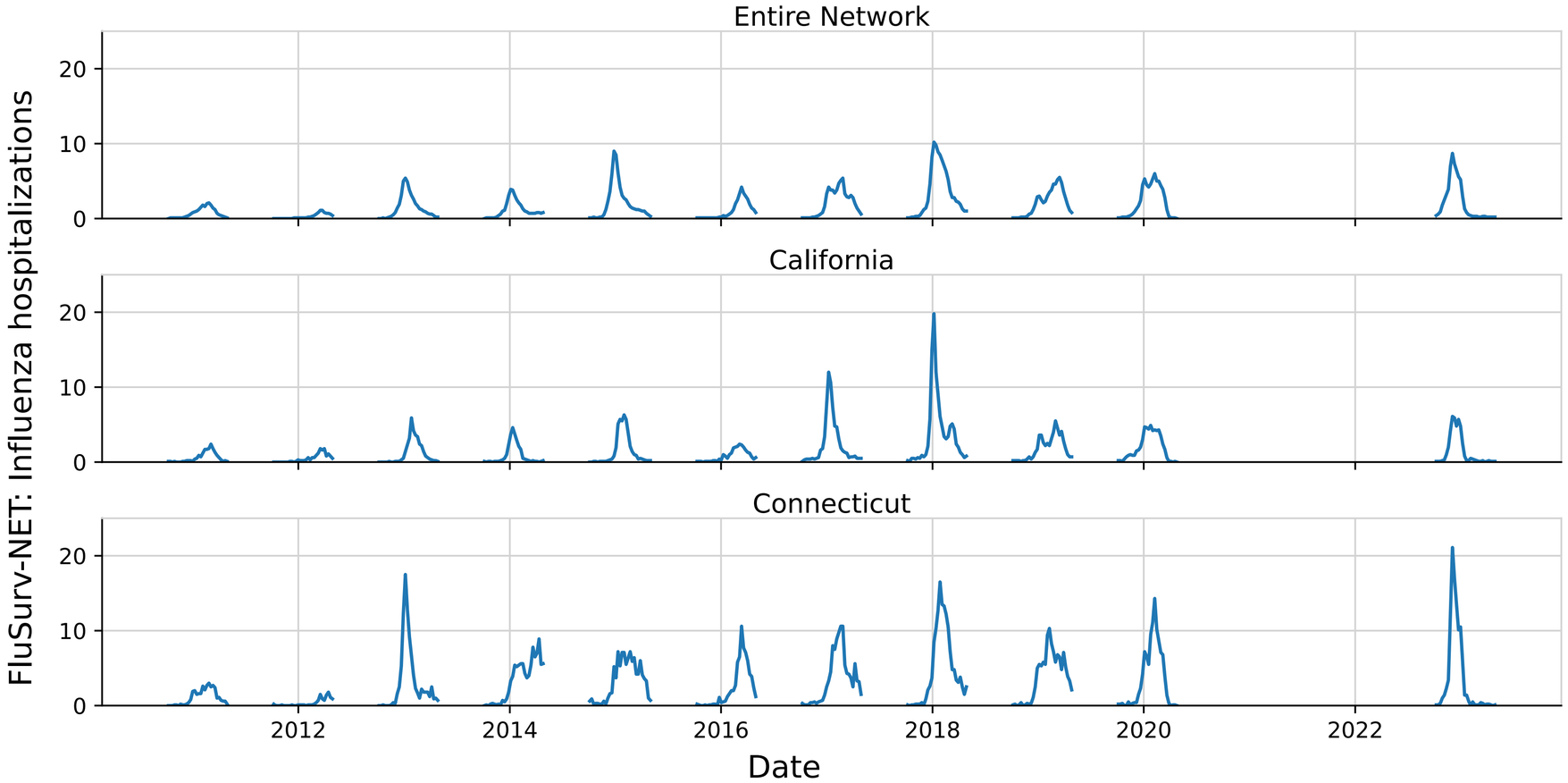
For all data sources, we drop seasons impacted by pandemics (2008/09, 2009/10, 2020/21, 2021/22)

Signal 2: FluSurv-NET

- Reported value is:

$$\text{Hospitalization rate} = \frac{\text{\# hospitalized with a positive test}}{\text{(\# of residents in catchment area)/100k}}$$

- Number of participating sites varies over time, but generally in the teens



FluSurv-NET: handling underreporting

- Reported value is:

$$\text{Hospitalization rate} = \frac{\text{\# hospitalized with a positive test}}{\text{(\# of residents in catchment area)/100k}}$$

- This is generally an underestimate, depending on testing rates and test sensitivity
- CDC produces yearly estimates of total hospital burden that adjust for these factors
  - https://www.cdc.gov/flu/about/burden/past-seasons.html
- We scale up by estimating a multiplicative factor a such that

$$a \times (\text{cumulative reported hospitalization rate, entire network}) \times (\text{100k US population})$$

$$= \text{National burden estimate}$$

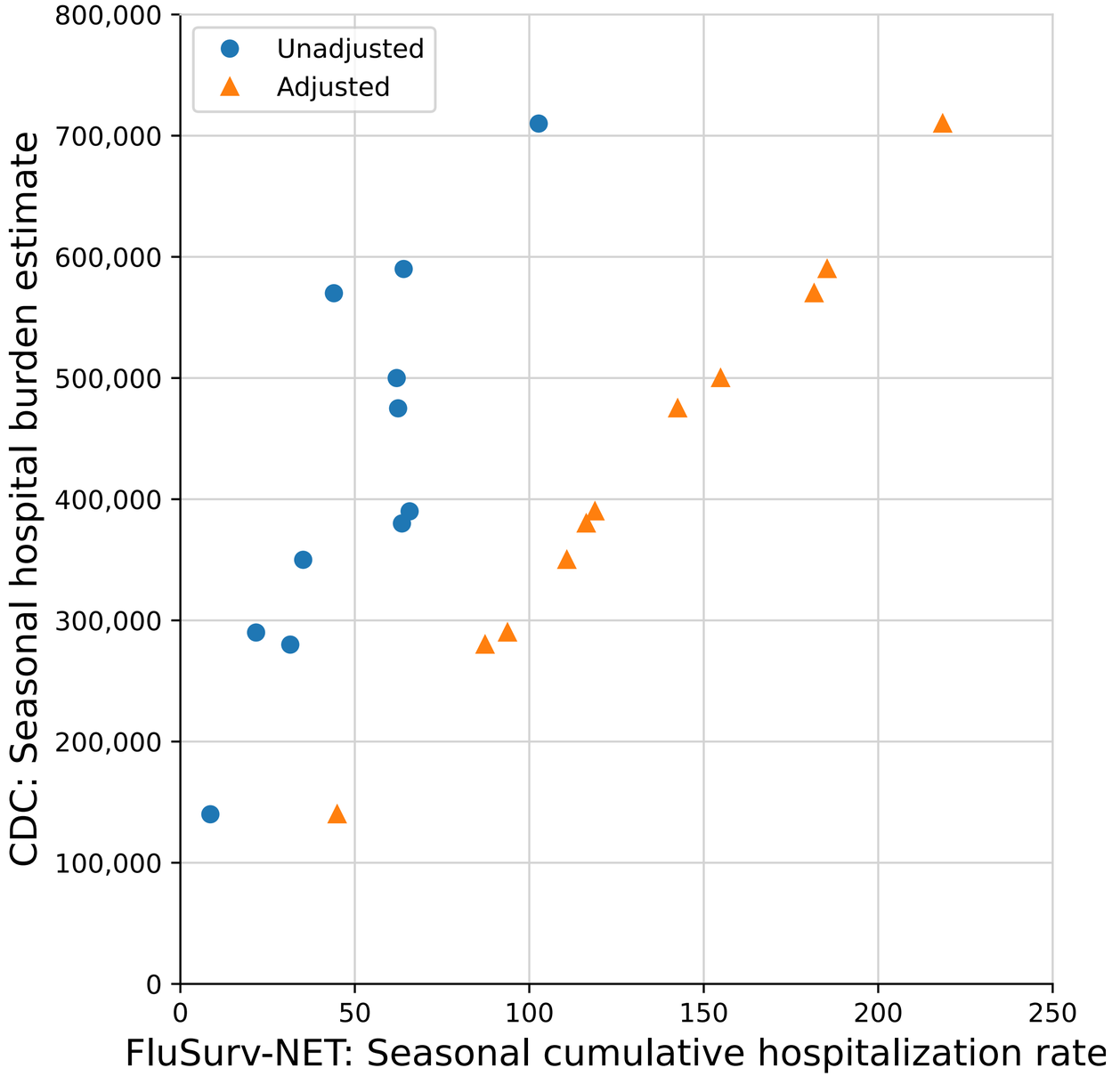
FluSurv-NET: handling underreporting

Season	FluSurv-NET Cumulative Rate	US Population	Hosp. Burden (from FluSurv-NET)	Hosp. burden (CDC)	Adjustment factor
2010/11	21.7	309321666	67122.8	290000.0	4.320439
2011/12	8.6	311556874	26793.8	140000.0	5.225072
2012/13	44.0	313830990	138085.6	570000.0	4.127873
2013/14	35.2	315993715	111229.7	350000.0	3.146639
2014/15	64.0	318301008	203712.6	590000.0	2.896237
2015/16	31.5	320635163	101000.1	280000.0	2.772275
2016/17	62.0	322941311	200223.6	500000.0	2.497208
2017/18	102.7	324985539	333760.1	710000.0	2.127276
2018/19	63.5	326687501	207446.6	380000.0	1.831797
2019/20	65.7	328239523	215653.4	390000.0	1.808458
2022/23	62.4	333287557	207971.4	475000.0	2.283968

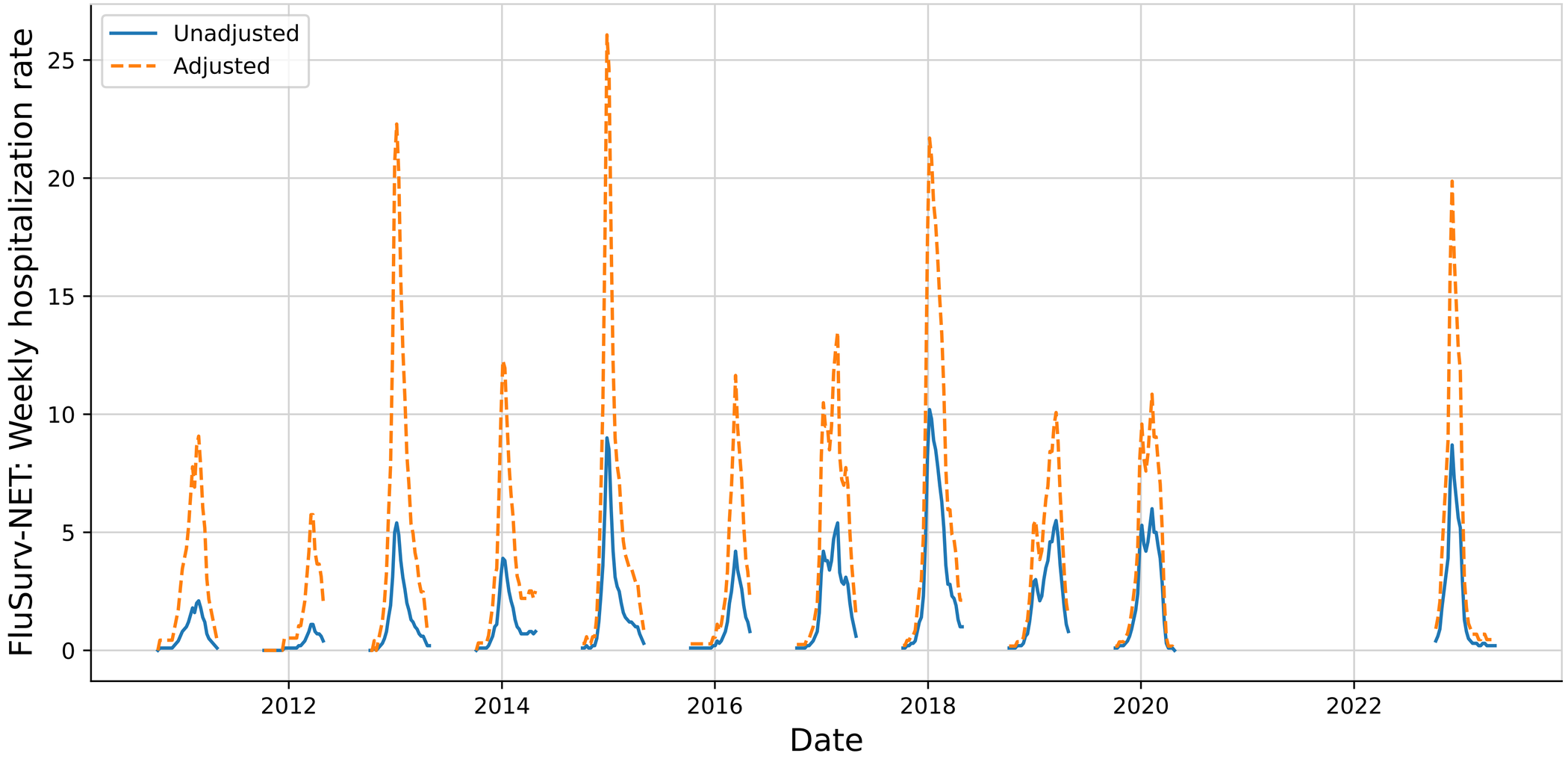
- In recent seasons, the “scale-up factor” has stabilized around 2
- It was larger in past seasons

FluSurv-NET: handling underreporting

- Adjustments shown for the “Entire Network” (each point represents 1 season)
- The same adjustments are applied to all sites
- Goal: after adjustment, we have a more meaningful representation of variability across flu seasons
- That said, I haven’t yet actually measured whether this helps forecast skill
  - Experiments planned for near future

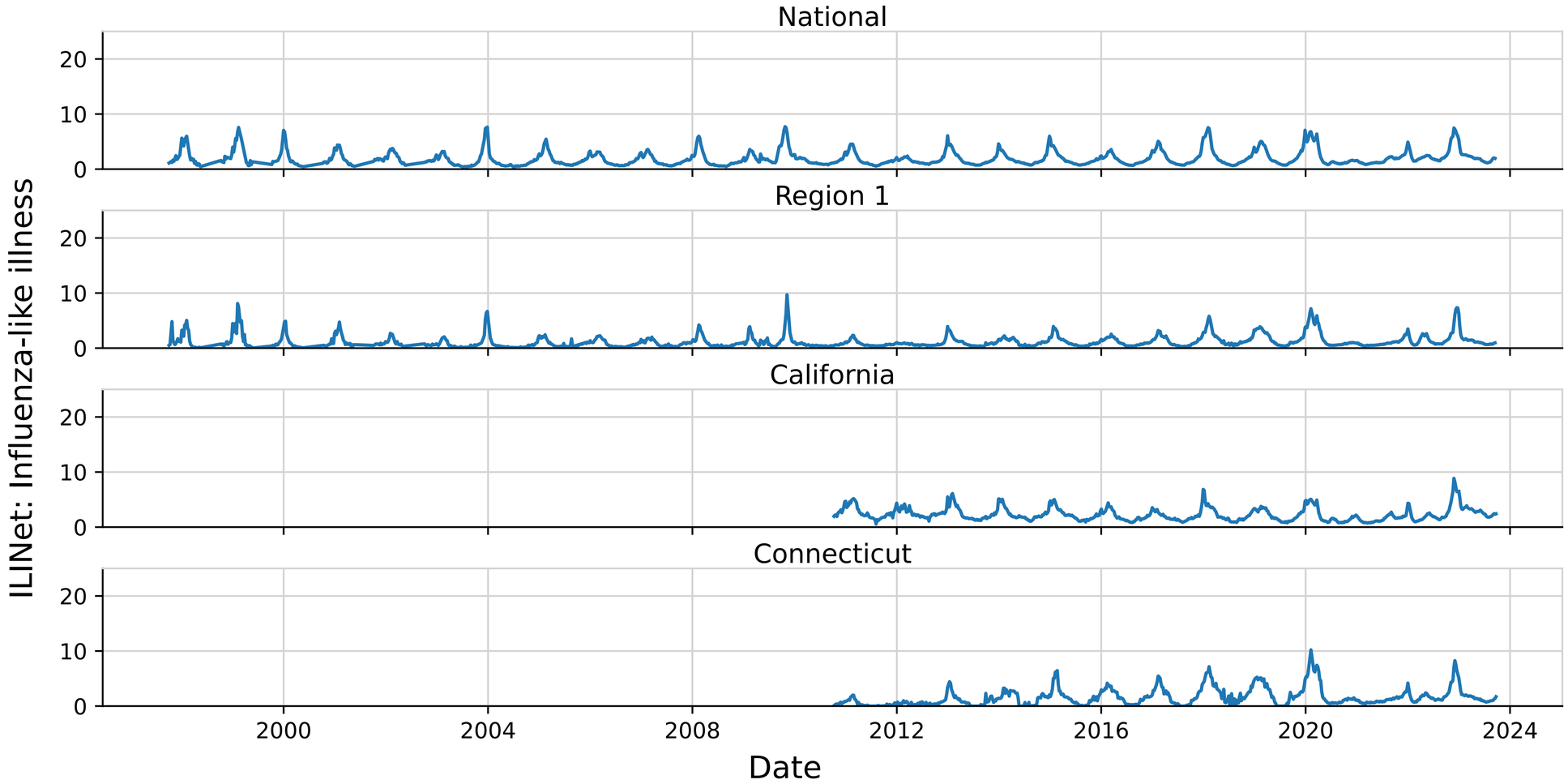


FluSurv-NET: handling underreporting



Signal 3: ILINet ➔ ILI+

- ILINet reports on influenza-like illness among outpatient doctor visits
- Defined symptomatically, includes people who have diseases other than flu
- This was always messy, but it’s particularly messy in recent years



ILI+

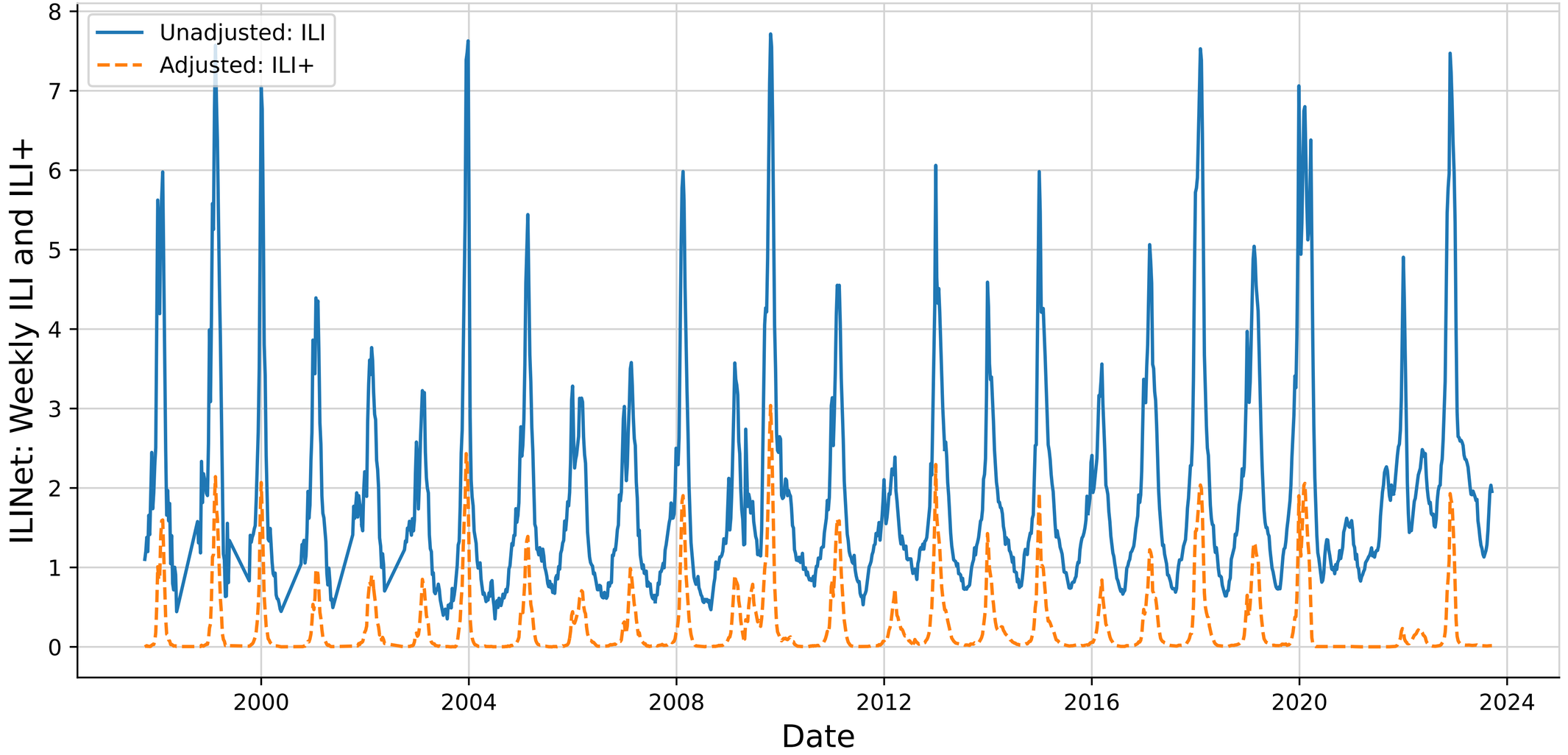
- ILI+ attempts to clean up the mess:

$$\begin{aligned}\text{ILI+} &= (\text{proportion outpatient visits with ILI}) \times (\text{proportion of flu tests with positive result}) \\ &\approx \text{proportion outpatient visits with flu}\end{aligned}$$

- We get virology testing data from WHO/NREVSS
- Again, I haven't done anything to measure benefit to forecasting of using ILI+ instead of ILI
  - Plan to conduct experiments soon

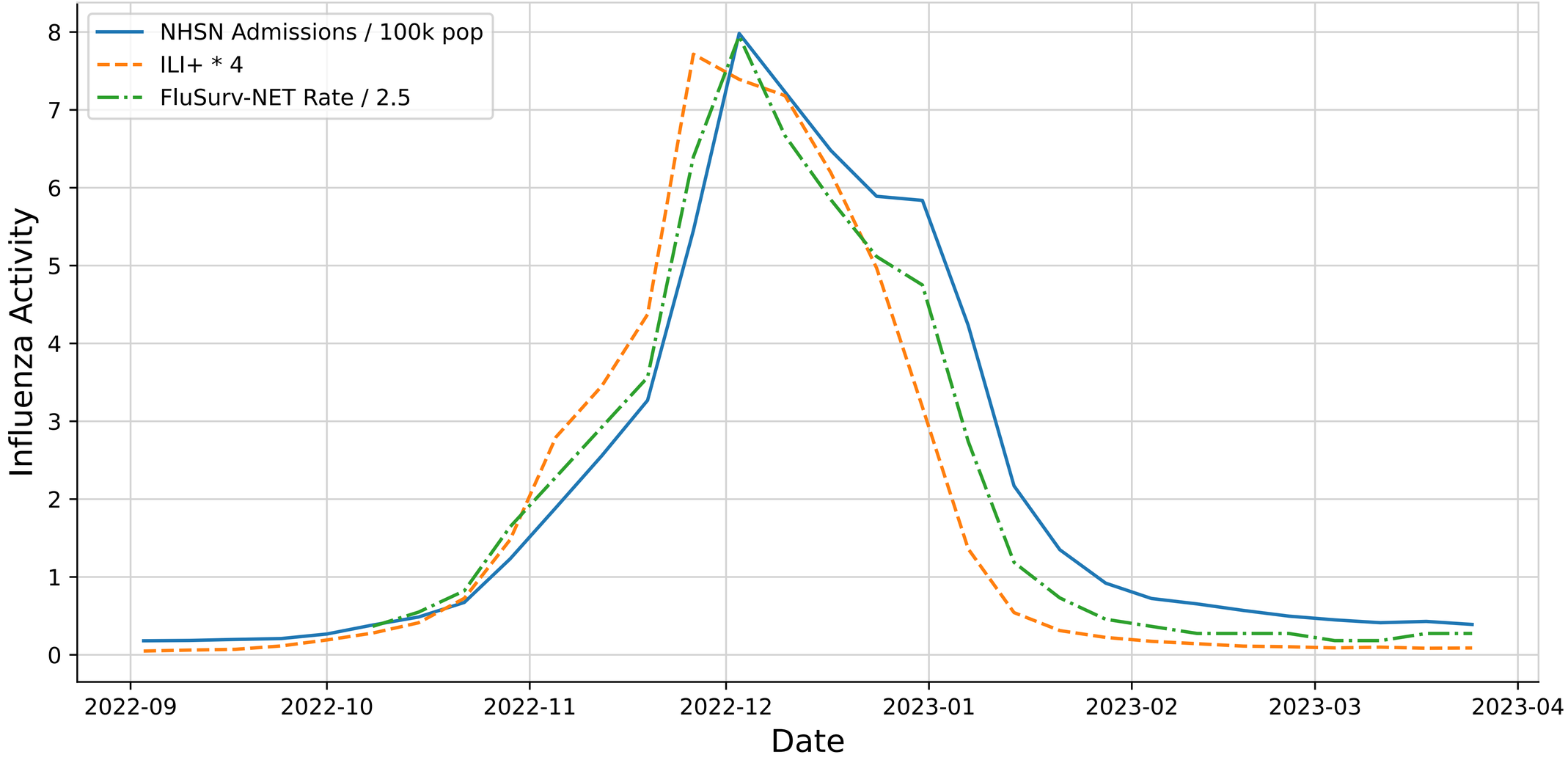
ILI+

- Data shown for US National level
- We get a consistent off-season and a more meaningful measure of flu activity post-COVID



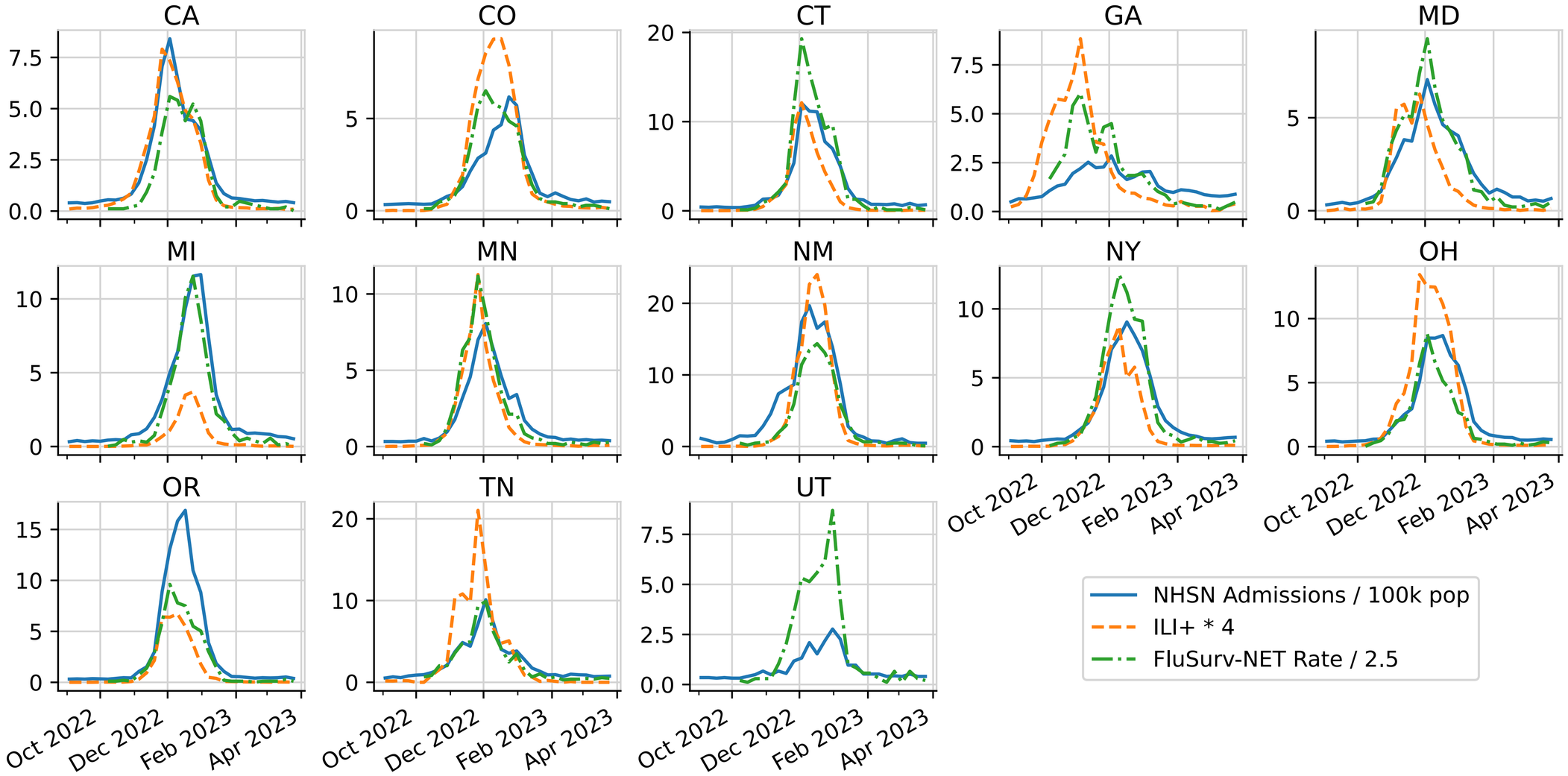
When available, the 3 signals align

- National level, 2022/23 season
- It doesn't always look this good (next slides)



When available, the 3 signals align

- State level (states with FluSurv-NET sites), 2022/23 season
- Timing generally lines up, with fixed scaling factors relative heights vary across states



How to use 3 signals?

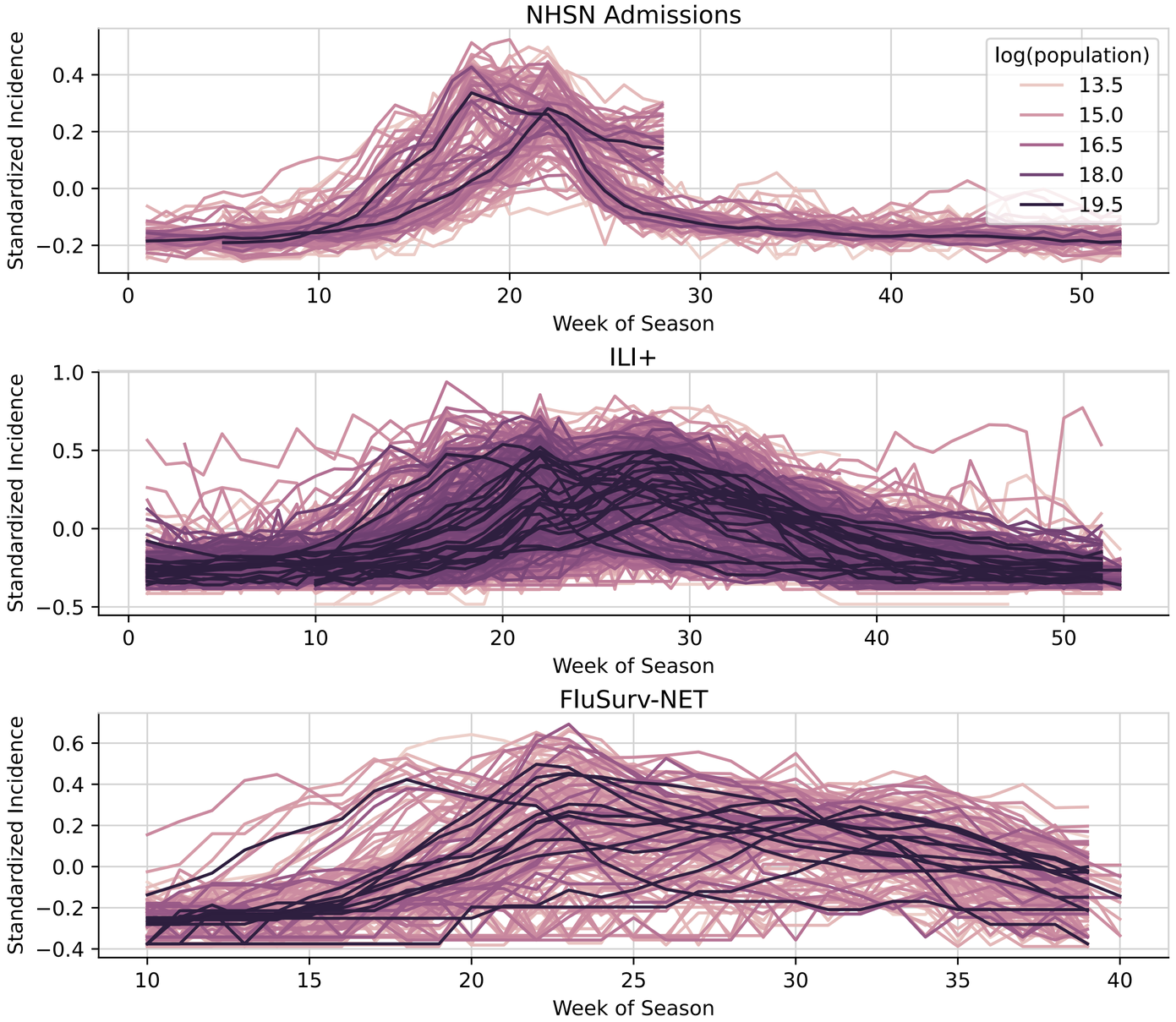
I see two main ways to use these data sources:

1. Parameterize a relationship between the sources, use all three together in real time
    - This seems like where we should be headed, but it's hard:
        - Unstable relationships between the signals across spatial units (and over time?)
        - Different reporting lags and amounts of revisions/backfill for different signals
2. Use historical data from alternate sources to help a model learn what seasonal flu looks like:
    - When does the season tend to peak?
    - What is a typical range for how low/high the peak might be?
    - What do holiday effects look like?

For the current flu season, I'm taking approach 2

What the UMass-flusion model sees

- The UMass-Flusion model is an ensemble of a gradient boosting model and an ARX model
- Both component models see all three data sources at state/site, regional, and national levels after:
  - Fourth root transform
    - Goal: variance stabilization
  - Centering and scaling
    - Separate center/scale parameters for each combination of source and location
    - Goal: make the data more comparable across data sources and locations



Conclusions

- Future work:
  - Measure how much all of this helps the current UMass-flusion model
  - Consider a setup where additional data sources are used to help inform predictions for the current season
- Code for data manipulation is available on GitHub at https://github.com/reichlab/flusion/
  - A hybrid of R (pulling raw data) and python (data manipulation)
  - This is extremely rough/not well documented
  - But I'm in favor of collaboration on a shared code base if others are planning to use these sources. Get in touch: elray@umass.edu
- Acknowledgments to Nick Reich, Rebecca Borchering, Matt Biggerstaff for input and feedback