

Swab statistics sharpen sewer sequencing sensitivity

Vladimir Sedlacek*¹, Lennart Justen (mentor)^{2,3}
¹Pivotal Research ²Massachusetts Institute of Technology ³Broad Institute



Week by week and nose by nose:
P2RA sharpness rose.

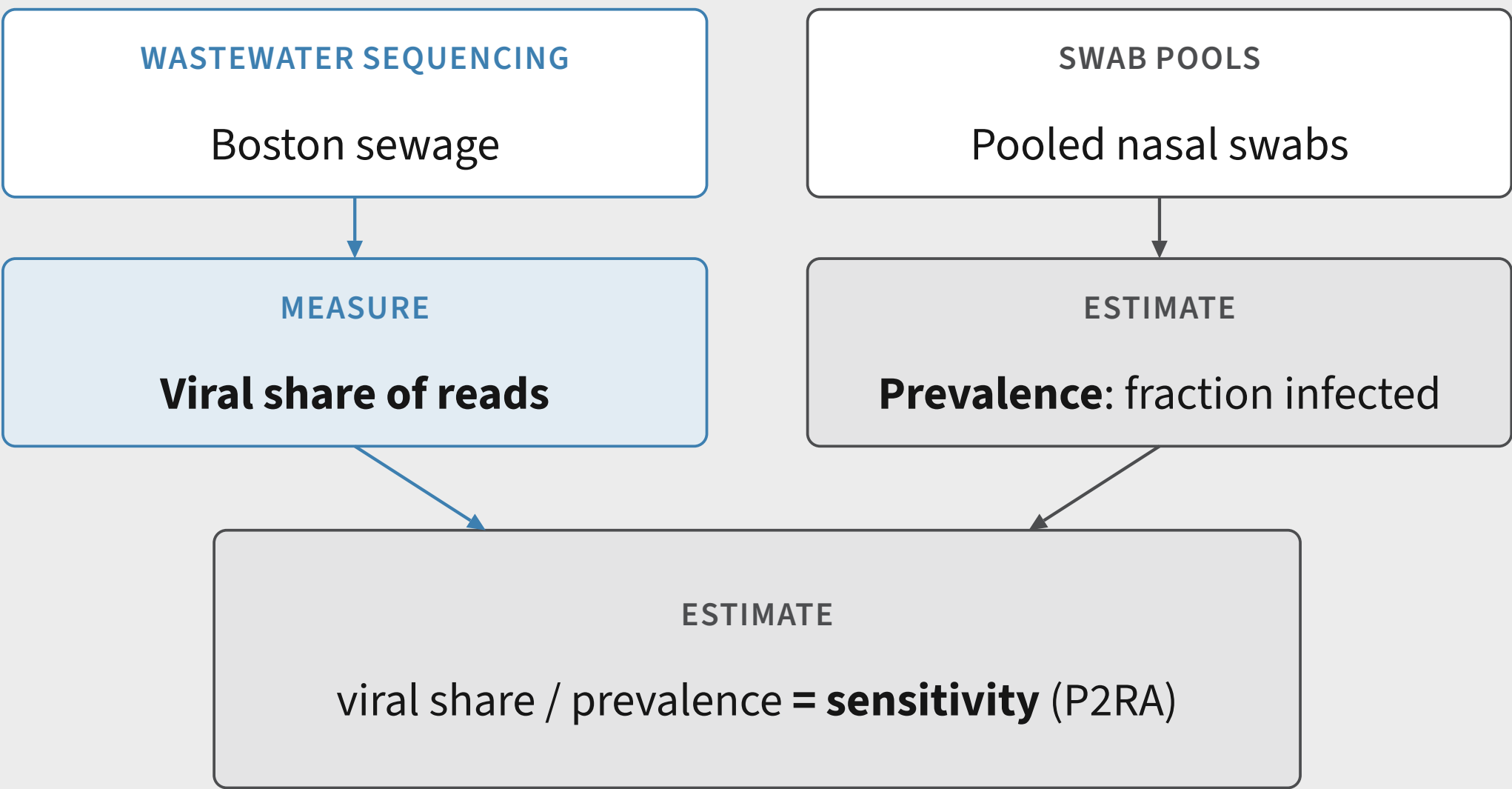
Why this matters

Untargeted sequencing of a city’s sewage can detect a range of viruses (even novel ones), providing early warning for pandemics. Its value depends on its **sensitivity (P2RA)**: how strong is a virus’s sewage signal at a given prevalence level (e.g., 1%)? The answer will inform the ideal sequencing depth and help budget allocators decide where to spend resources. We aimed to improve SecureBio’s P2RA estimates for respiratory viruses with more swab data.

Methods

- **Sequenced inputs:** We used weekly wastewater data from the Boston Deer Island plant (part of CASPER), and pooled nasal swabs (about 60 per pool) from Zephyr program volunteers. We kept a subset with a single lab protocol for fair comparisons, leaving 65 pools and 24 sewage samples.
- **Bayesian statistics:** The model starts with initial assumptions about the unknowns (viral prevalence, average yield, variance), and updates them based on the data, so every estimate includes a plausible range.
- **Signal:** To estimate prevalence, we count pools with at least one viral read.
- **Seasonality:** We updated the original SecureBio model to allow a different viral prevalence each week, instead of a constant rate independent of time.

Model workflow



Limitations

- While the pilot and the refit use the same city and wastewater lab, they differ in season, sites and swab protocol.
- The swabs come from adult volunteers at transit sites, while the sewer includes viruses from children and the housebound.
- The data cannot determine the false negative detection rate: how often does a swab from an infected person show no viral reads?

Narrower violins = more certainty

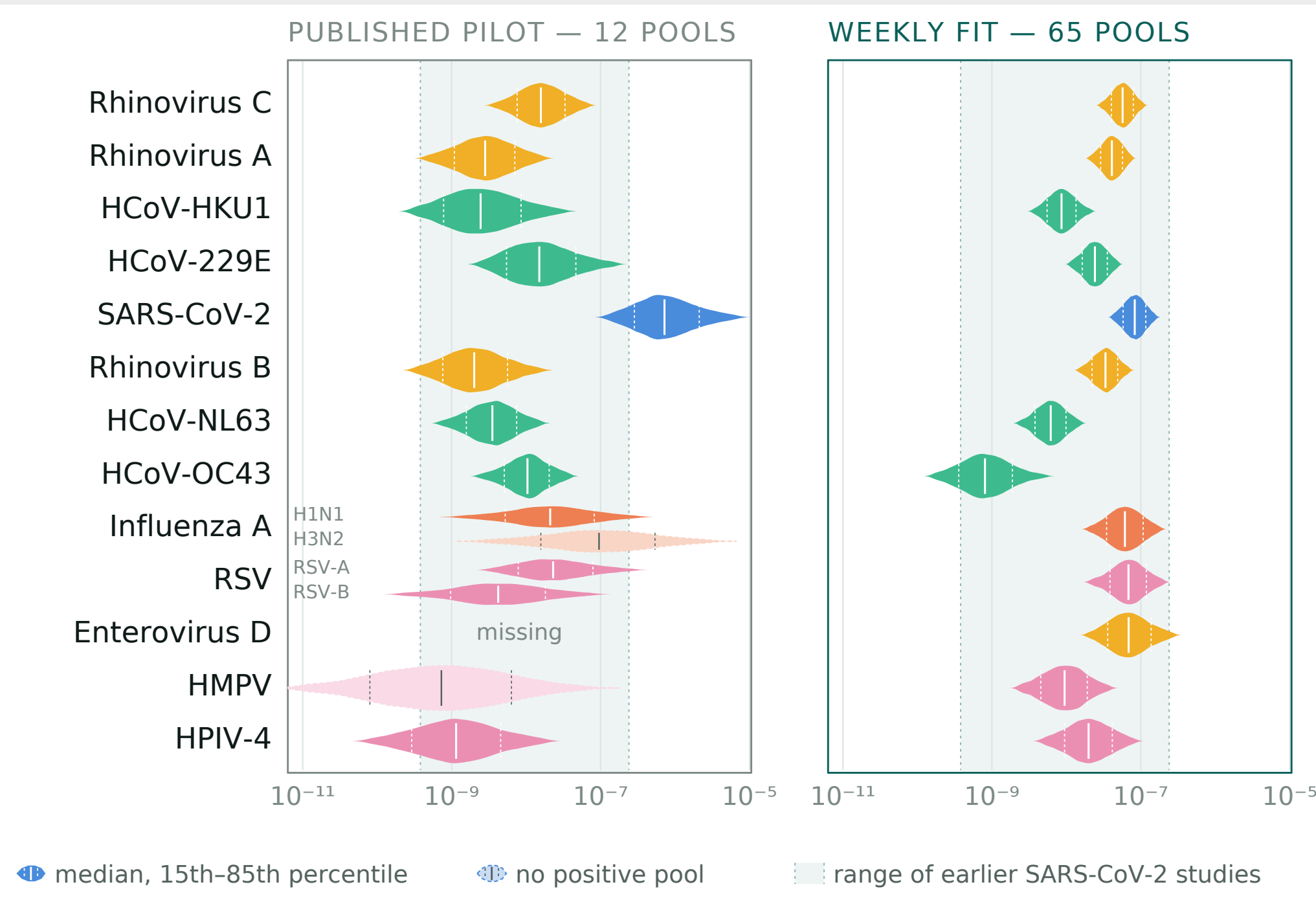


Figure 2: P2RA(1%) estimates for 13 viruses, ranked by positive pool counts: the published pilot (Jan–Feb 2025) beside our weekly-prevalence fit (Nov 2025 – Feb 2026).

Takeaways

- Prevalence can vary wildly with time. Influenza A was found in half the pools during its January peak week, but only in 9% of the pools over the whole winter.
- For the seven most observed viruses, the range of plausible P2RA values shrank from a 4–11-fold spread to about 2-fold in our weekly fit.
- The model is sensitive to the swab data, as there is comparatively less of it.
- P2RA depends not just on the virus, but also on the city, sewer system, lab and protocol. Similar published SARS-CoV-2 estimates from different cities and labs differ by up to 630-fold.

Next steps

- We plan to find a way to compare swab protocols and rerun the data through the latest sequencing pipeline, in order to extend the observed time period and stop discarding useful information.
- We want to use a more realistic viral shedding distribution that could handle many weeks without positive pools, as well as a few weeks with all pools positive.
- We will consider clinical and literature estimates to choose more accurate prevalence priors and the false negative detection rate.
- We would like to utilize metadata to account for the swab quality.