FECAL SHEDDING OF SARS COV-2 RNA

BST 228

Speaker notes

No notes on this slide.

TODAY

- Case study.
- Project work.

Speaker notes

- This lecture demonstrates how the different topics you've learned about fit together.

Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **72**, 330–345
https://doi.org/10.1093/jrsssc/qlad011
Advance access publication 17 March 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Faecal shedding models for SARS-CoV-2 RNA among hospitalised patients and implications for wastewater-based epidemiology

Till Hoffmann[1] and **Justin Alsing**[2]

[1]Department of Mathematics, Imperial College London, London, UK
[2]Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, Stockholm, Sweden

Address for correspondence: Till Hoffmann, Sir Ernst Chain Building, Imperial College Rd, London SW7 5NH, United Kingdom. Email: t.hoffmann@imperial.ac.uk

Abstract

The concentration of SARS-CoV-2 RNA in faeces is not well characterised, posing challenges for quantitative wastewater-based epidemiology (WBE). We developed hierarchical models for faecal RNA shedding and fitted them to data from six studies. A mean concentration of 1.9×10^6 mL^{-1} (2.3×10^5–2.0×10^8 95% credible interval) was found among unvaccinated inpatients, not considering differences in shedding between viral variants. Limits of quantification could account for negative samples based on Bayesian model comparison. Inpatients represented the tail of the shedding profile with a half-life of 34 hours (28–43 95% credible interval), suggesting that WBE can be a leading indicator for clinical presentation. Shedding among inpatients could not explain the high RNA concentrations found in wastewater, consistent with more abundant shedding during the early infection course.

Keywords: hierarchical modelling, viral load, wastewater-based epidemiology

Speaker notes

- The discussion is based on a recent publication on shedding of SARS-CoV-2 RNA in feces, which is important for interpreting data from wastewater-based epidemiology.

Speaker notes

- What is wastewater-based epidemiology (WBE)? Collecting and analyzing biomarkers wastewater samples to learn about the state of public health.
- Samples are typically collected at the pipe entering wastewater treatment works, but samples can also be collected upstream (e.g., manholes) for more granular spatial analysis or downstream (e.g., sludge from sedimentary tanks during the treatment process).
- Analytes include markers of infectious diseases (e.g., viral RNA), metabolites of prescription and illicit drugs, and more.

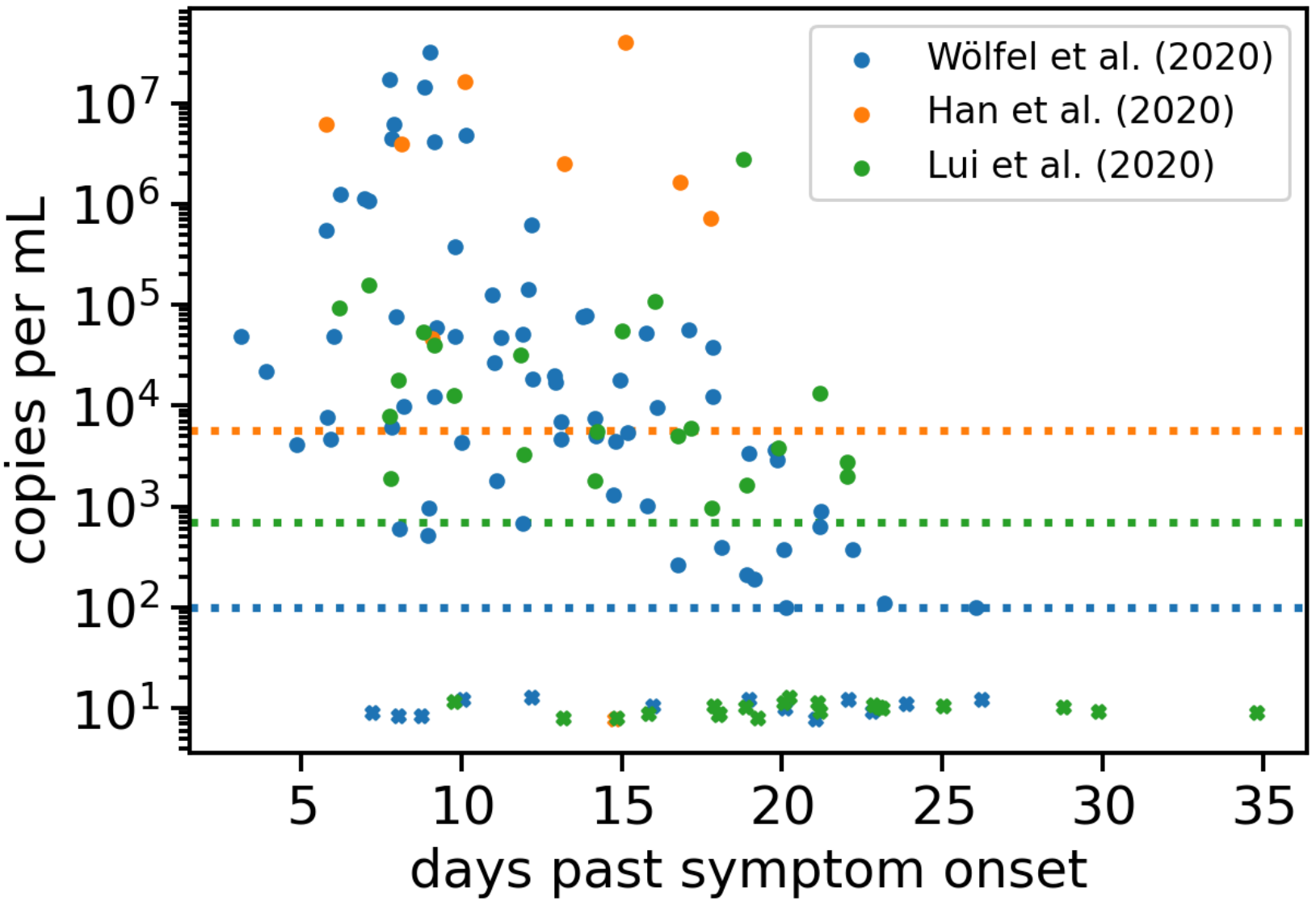
QUESTIONS FOR WASTEWATER-BASED EPIDEMIOLOGY

In the context of SARS-CoV-2, we might want to answer the following questions.

1. How many SARS-CoV-2 RNA copies are shed on average?
2. Does everyone shed RNA?
3. How does shedding vary over time?
4. …

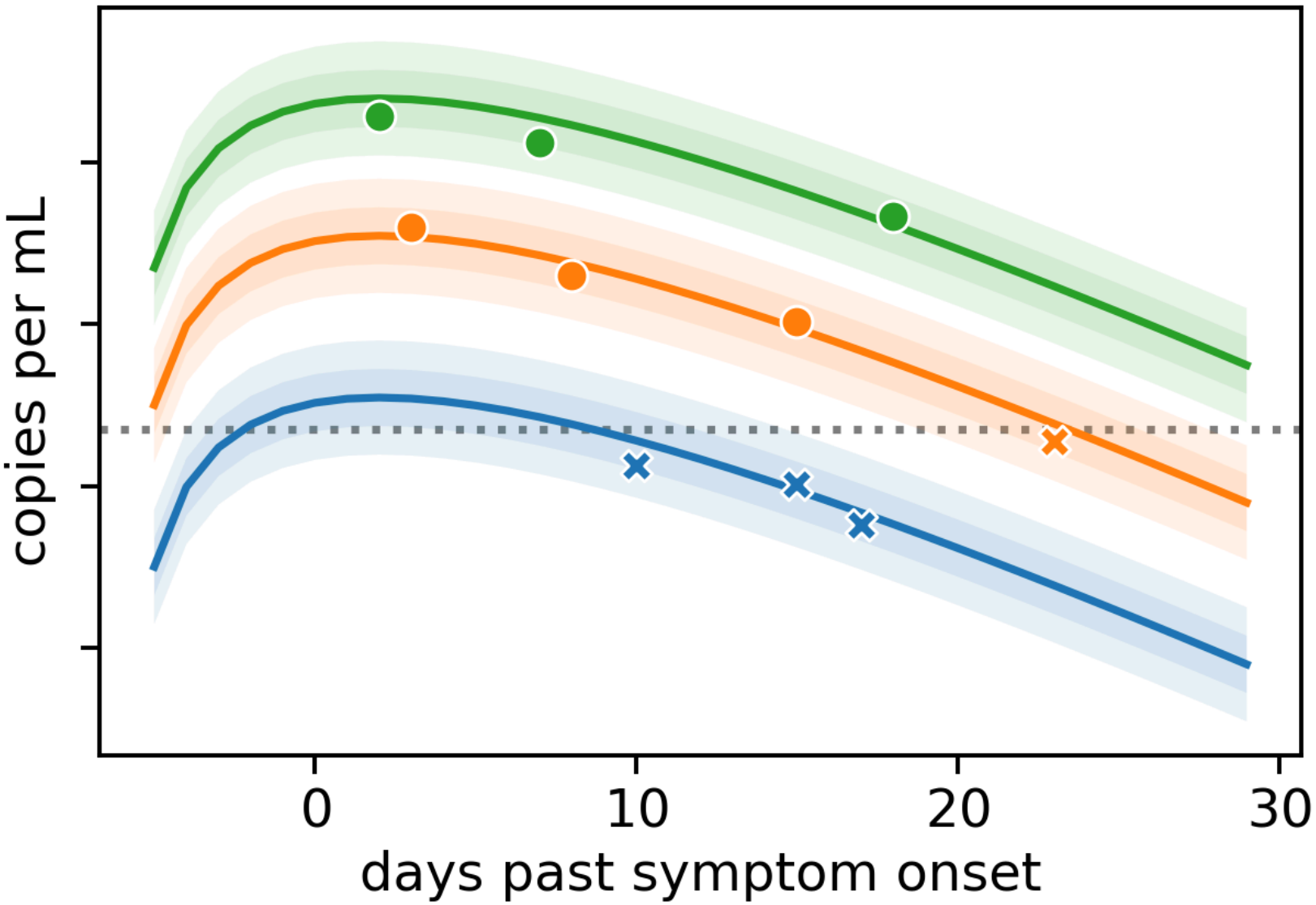
Speaker notes

1. Knowing how many copies are shed is fundamental for trying to estimate disease prevalence.
2. If not everyone sheds viral RNA, we need to correct prevalence estimates. A bimodal distribution of shedding behavior can also inform disease characteristics. E.g., do only some people have enteric infections?
3. If most shedding happens during the early infection course, WBE data can be leading indicator.



Speaker notes

- The figure shows viral RNA concentrations against days past symptom onset in individual samples collected from different patients, i.e., we have longitudinal data. Data are extracted from three early studies indicated by different colors.
- Horizontal lines indicate the limit of quantification which vary across studies.
- Concentration appears to decline over time, but the data are noisy. Early shedding is poorly constrained because data from hospitalized patients.
- We cannot use simple summaries. E.g., if there are many samples per patient, the patient will dominate the summaries.



Speaker notes

- The figure illustrates a potential model. Each color represents a different patient. We use a shared temporal profile common to all patients and a common within-patient variance between samples. Overall shedding levels may vary between patients.
- The “green patient” has all positive samples over the study period. “Orange” has initial positive and later negative samples. “Blue” has all negative samples because they were admitted late in the infection course.
- Cf. lecture 21 on missing data and censored data here. What is the missing data mechanism?

SHEDDING MODEL

$$\mu_i \mid M, S, Q \sim \text{GeneralizedGamma}\left(M, S, Q\right)$$

Speaker notes

- We have random effects for each patient i from a generalized gamma distribution (cf. lectures 14-16 on hierarchical models).
- We could use any distribution with positive support, e.g., log-normal, gamma, Weibull (cf. early lectures on choosing the right likelihood).
- We use the generalized gamma distribution because it includes the other distributions as special cases and allows us to control different tail shapes. We will get back to the importance of this choice.

SHEDDING MODEL

$$\mu_i \mid M, S, Q \sim \text{GeneralizedGamma}\left(M, S, Q\right)$$

$$p\left(x_{ij} \mid \mu_i, \sigma, q, \theta\right) = \begin{cases} f\left(x_{ij} \mid \mu_i \times g\left(t_{ij}\right), \sigma, q\right) & \text{if } x_{ij} > \theta \end{cases}$$

Speaker notes

- We have j^{th} observation x_{ij} for patient i above the limit of quantification θ.
- f is the density of a generalized gamma distribution but with different scale σ and shape q than the population-level distribution.
- The location parameter is $\mu_i \times g(t_{ij})$ with temporal shedding profile g, where t_{ij} is the time at which sample j was collected from patient i.

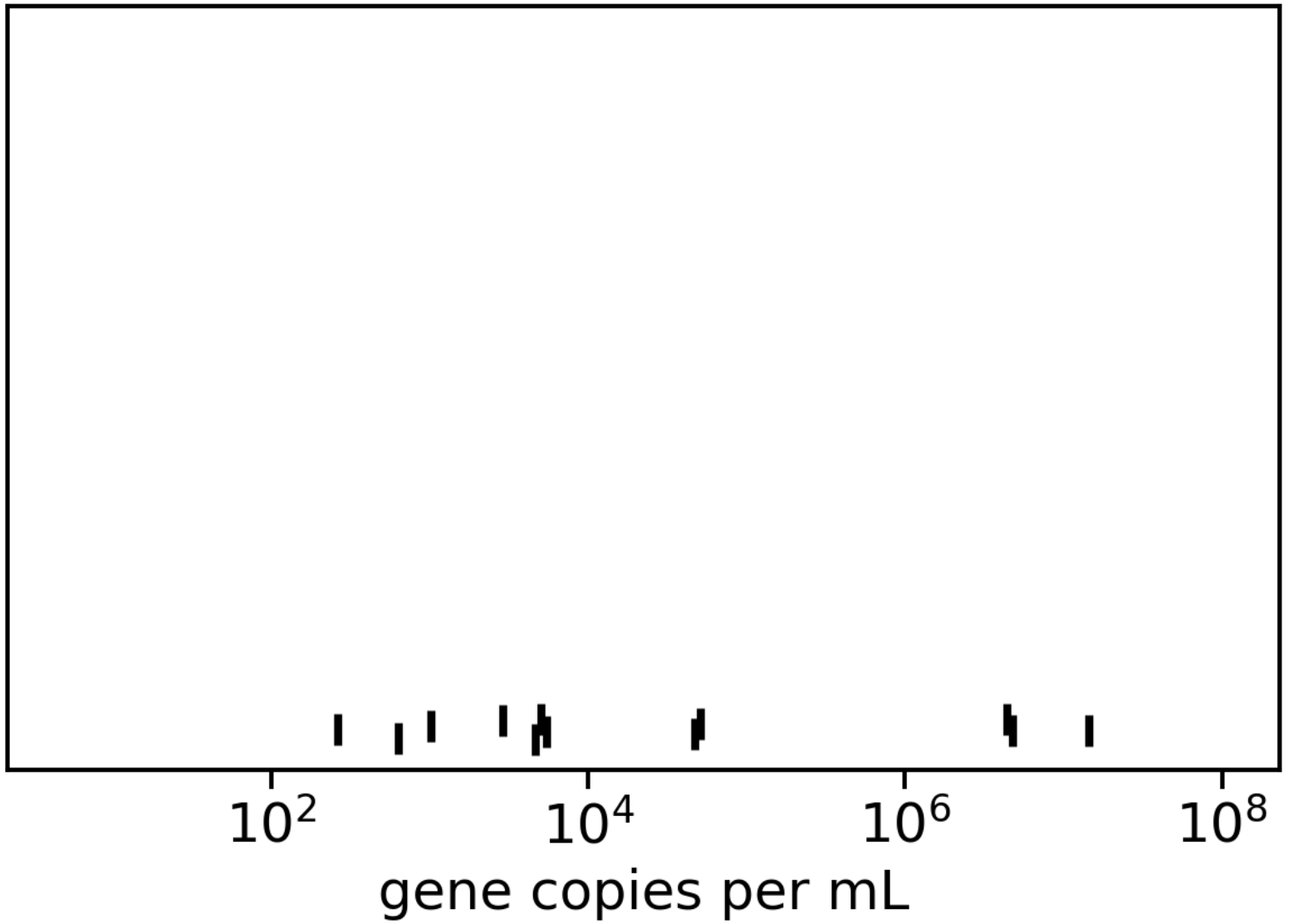
SHEDDING MODEL

$$\mu_i \mid M, S, Q \sim \text{GeneralizedGamma}\left(M, S, Q\right)$$

$$p\left(x_{ij} \mid \mu_i, \sigma, q, \theta\right) = \begin{cases} f\left(x_{ij} \mid \mu_i \times g\left(t_{ij}\right), \sigma, q\right) & \text{if } x_{ij} > \theta \\ F\left(\theta \mid \mu_i \times g\left(t_{ij}\right), \sigma, q\right) & \text{otherwise} \end{cases}$$

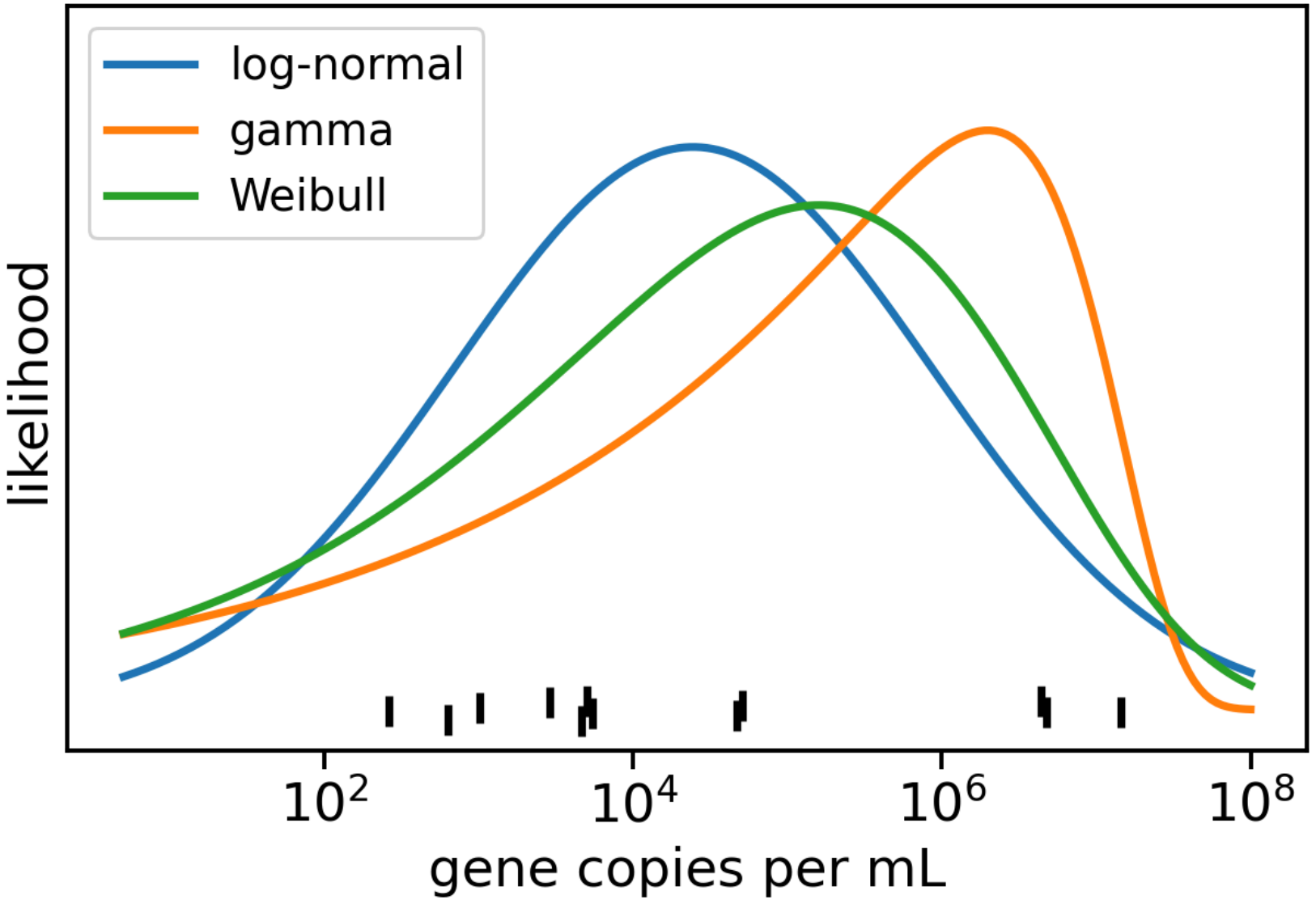
Speaker notes

- For samples below the limit of quantification (LOQ) θ, we evaluate the cumulative distribution function F at the LOQ: We only know that samples are less than the LOQ.



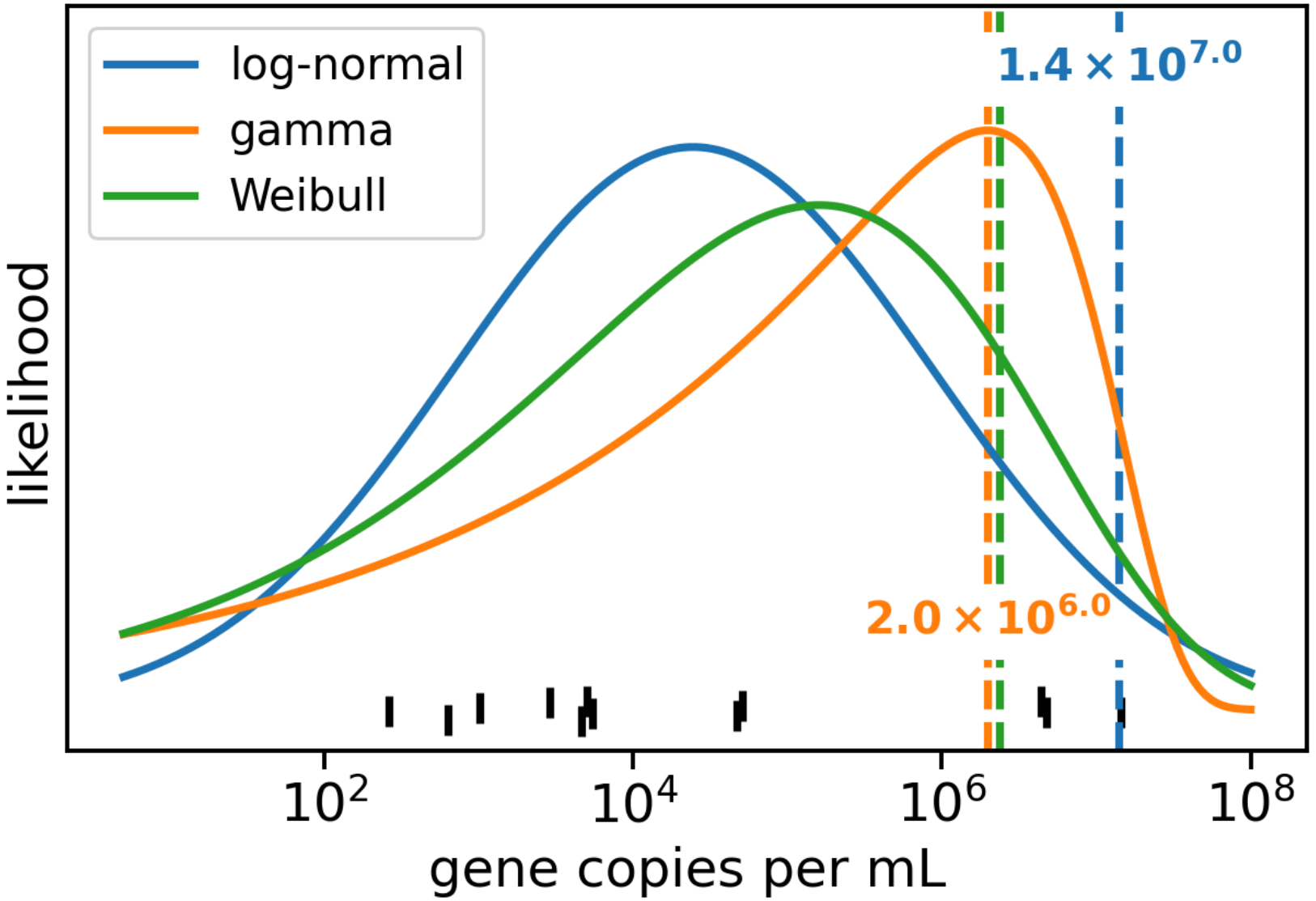
Speaker notes

- Why do we use the more complex generalized gamma distribution instead of a simpler distribution like the log-normal?
- The figure shows 12 samples from the first patient of Wölfel et al. (2020) as a rug plot with jitter to distinguish the samples.



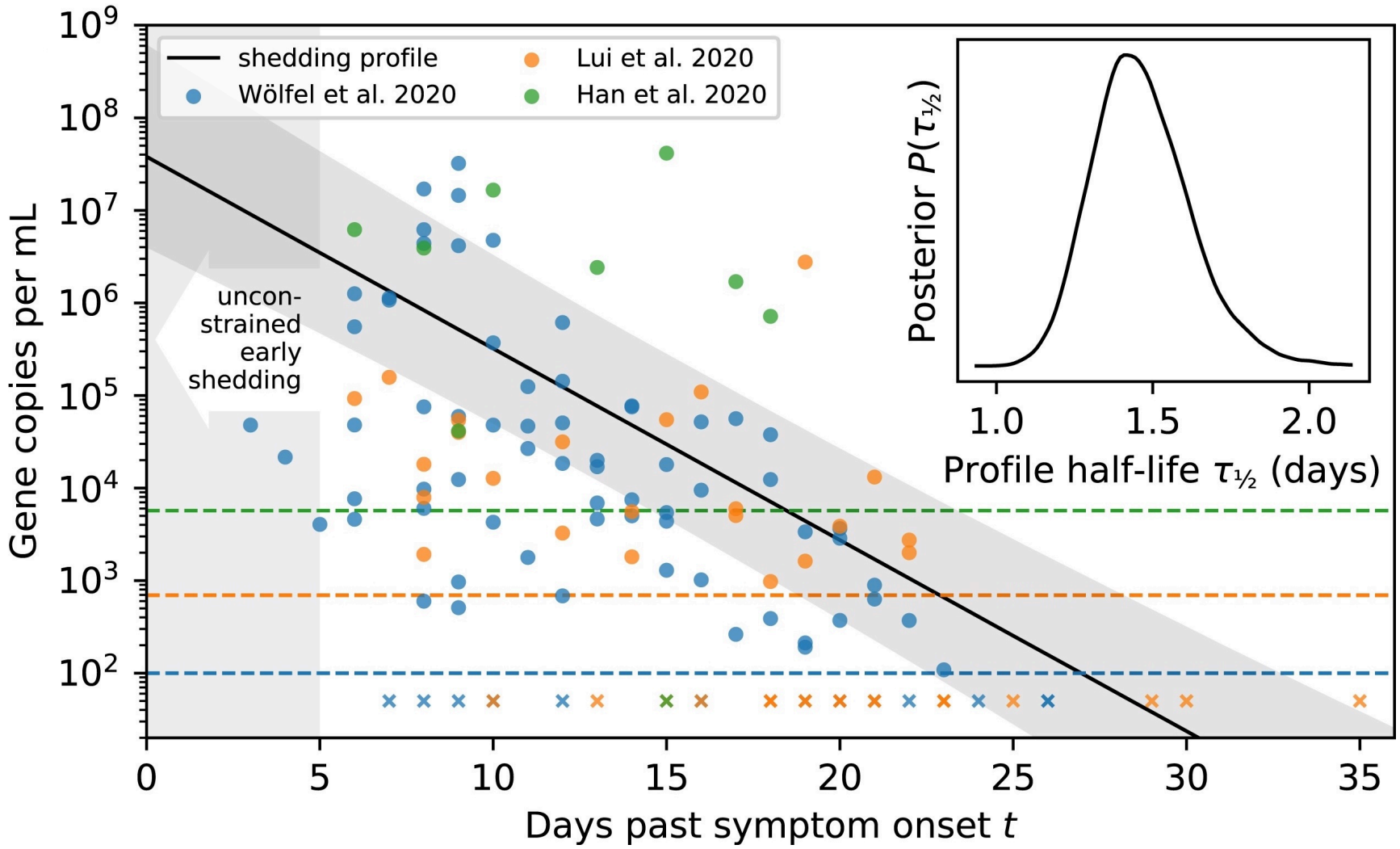
Speaker notes

- After fitting three distributions using maximum likelihood estimation, we observe very different densities because the shape is fully determined by the functional form.
- Log-normal has more mass at low values but also heavier tail; gamma is most concentrated and has light tails.
- Cf. early lectures where we discussed that the support of the data does not uniquely determine the “correct” likelihood.



Speaker notes

- What are the predicted means?
- Predicted mean of log-normal is almost 10x larger compared with gamma.
- Predicted mean is almost as large as the largest observed value. For some datasets, the predicted mean is larger than the maximum of the dataset.
- Ironically, adding a measurement at 10 gc/mL would increase the predicted mean to 4.1×10^7 because log-normal scale increases.
- Means are all about the tails. But we *need* the mean for WBE applications because the collected data is a mixture of fecal matter from many people.
- Choosing a distribution is a commitment to the tail shape which we cannot know a priori.
- By using a generalized gamma distribution, we are still committing to a functional form here, but it is more flexible.



Speaker notes

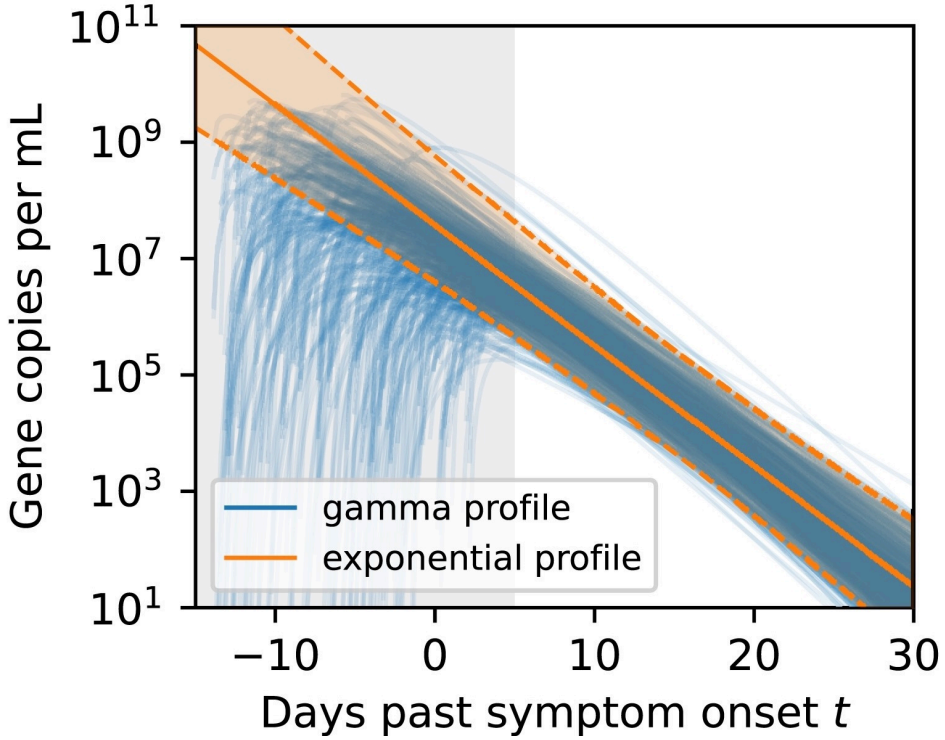
- The figure shows a model fit using Monte Carlo samples in three replicates for different seeds (cf. lecture 8 on MCMC and 10 on diagnostics).
- Stan was not suitable because the posterior had bad geometry (cf. lecture 12 regression case study). We used a different but slow sampler which requires running at scale (cf. Dr Schwartz's guest lecture on distributed computing).
- We did not use variational inference (cf. lecture 27) because we really care about tails: super-shedders.
- Here, we used an exponential shedding profile and find a find a half-life of 1.5 days (that's half life of the profile not half life of RNA degradation).

TEMPORAL SHEDDING PROFILES

- Exponential: $\exp\left(-\lambda\left(t-t_0\right)\right)$
- Gamma: $\left(t-t_0\right)^{a-1}\exp\left(-b\left(t-t_0\right)\right)$
- Teunis et al.: $\left[1-\exp\left(-a\left(t-t_0\right)\right)\right]\exp\left(-b\left(t-t_0\right)\right)$

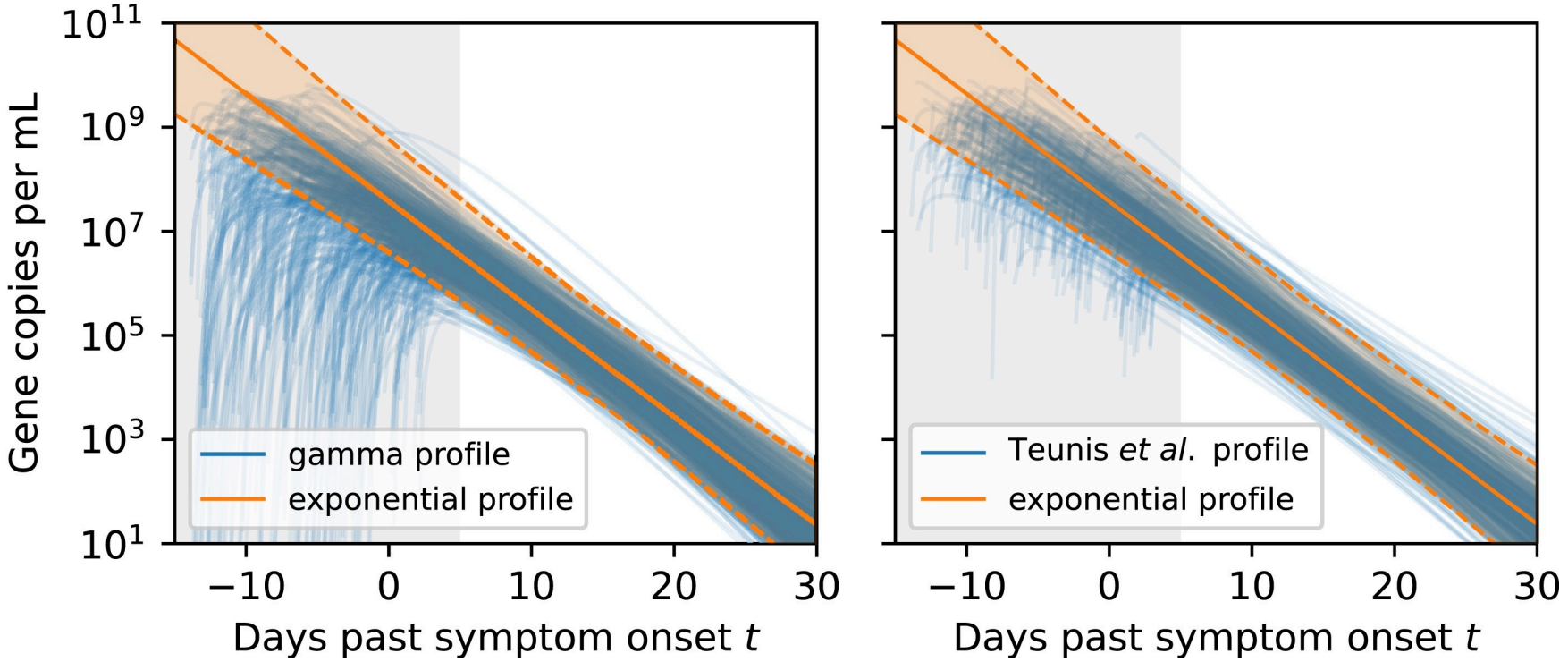
Speaker notes

- The exponential profile is really just a generalized linear model with log link function and generalized gamma likelihood (cf. lecture 13).
- The exponential profile is of course not reasonable because it suggests very large shedding long before symptom onset.
- As a sensitivity analysis, we consider different shedding profiles from the literature, including the shape of a gamma likelihood and an exponential rise and decay profile.
- The two other profiles have a third parameter which we need to infer: The time t_0 at which shedding starts relative to symptom onset.



Speaker notes

- The left panel shows samples of gamma shedding profiles consistent with the data in blue and the exponential profile in orange. The data cannot constrain the early shedding profile (because we do not have any early data). Consequently, t_0 and the time at which shedding peaks cannot be learned from the data we have.



Speaker notes

- We observe the same pattern for the exponential rise and decay profile.
- There are publications that make claims about peak shedding and time of peak shedding. But those claims are entirely assumption based, e.g., $t_0 = 0$ such that shedding starts at symptom onset.
- We cannot constrain the peak without collecting early shedding data, unless we have very good mechanistic models to inform the shedding profile.

SUB-POPULATION MODEL

For patients with all-negative samples,

$$p\left(x_{i\bullet} \mid \mu_i, \sigma, q, \rho\right) = (1-\rho) + \rho \prod_{j=1}^{m_i} F\left(\theta \mid \mu_i \times g\left(t_{ij}\right), \sigma, q\right).$$

Speaker notes

- We consider a sub-population of people who fundamentally do not shed. The probability to shed is ρ.
- For a patient with all negative samples, they either belong to the non-shedding population or each sample is below the LOQ.
- Fitting this model, the highest posterior density interval includes $\rho = 1$, and we find no evidence for a subpopulation of people who do not shed at all.
- The small proportion of positive samples is likely an artifact of collecting data from hospitalized patients late in the infection course.

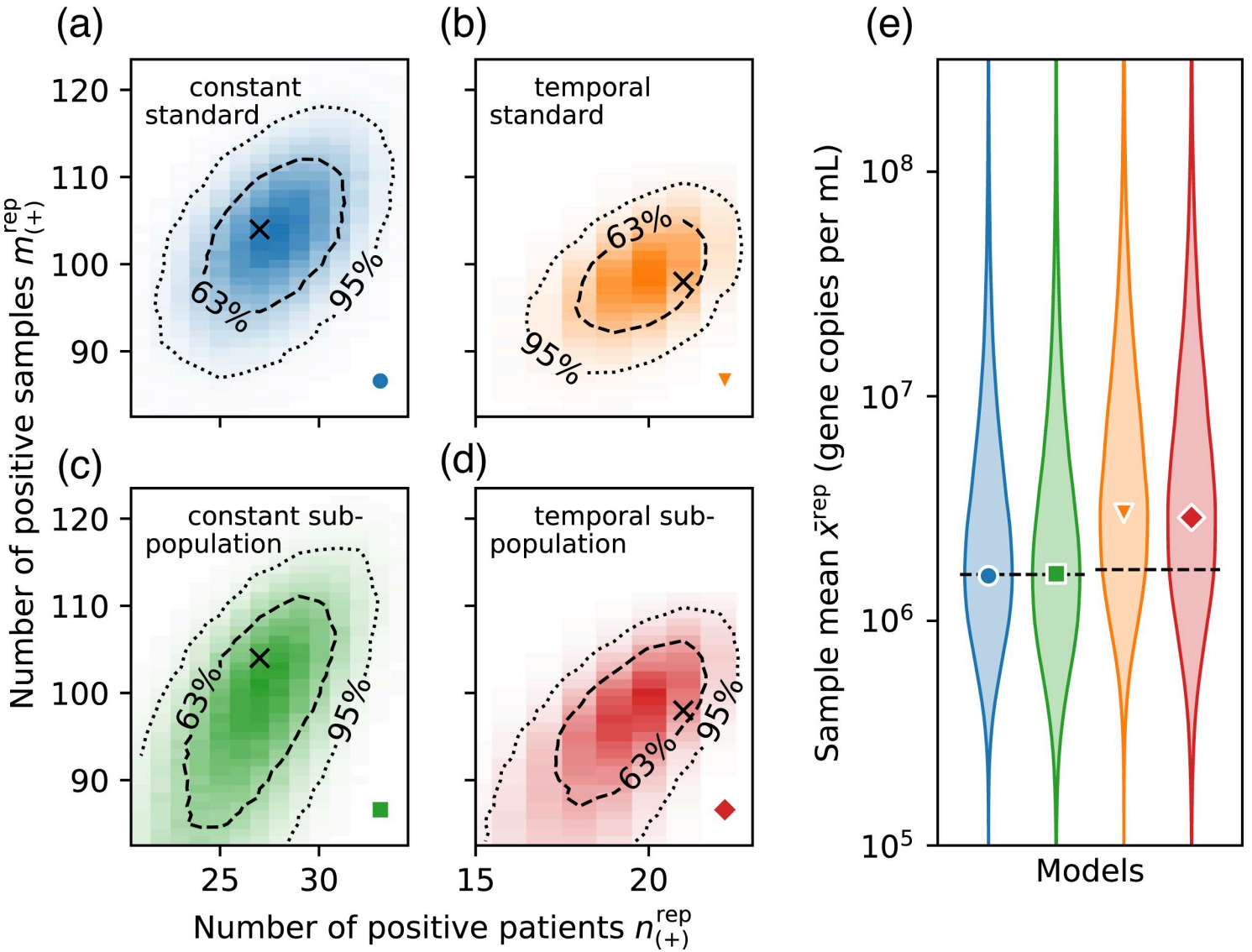
POSTERIOR PREDICTIVE REPLICATION

To check the model, we sample from the posterior predictive distribution to replicate

- $m_{(+)}^{\text{rep}}$, number of positive samples,
- $n_{(+)}^{\text{rep}}$, number of patients with at least one positive sample,
- $\bar{x}^{\text{rep}}$, sample mean of positive samples.

Speaker notes

- We use quantities for replication that are not directly related to parameters of the model. Otherwise, we are likely to replicate the summaries well even if the model is poor (cf. lecture 20 on model checking).
- Remember, replication can only tell us if a model is bad—not if it is good.



Speaker notes

- All models can explain the number of positive samples and number of patients with at least one positive sample.
- Temporal models with exponential profiles have smaller variance because variation is explained by the profile, not marginal variance as is the case for a model with a constant shedding profile.
- Models with a sub-population of non-shedders have larger variance because binary indicators for shedding affects all samples of the patient jointly.
- Sub-population models have higher replicated mean because the mean isn't "pulled down" as far to explain the negative samples.
- Temporal models predict slightly higher mean concentration because of abundant early shedding.

OUT-OF-SAMPLE PREDICTION

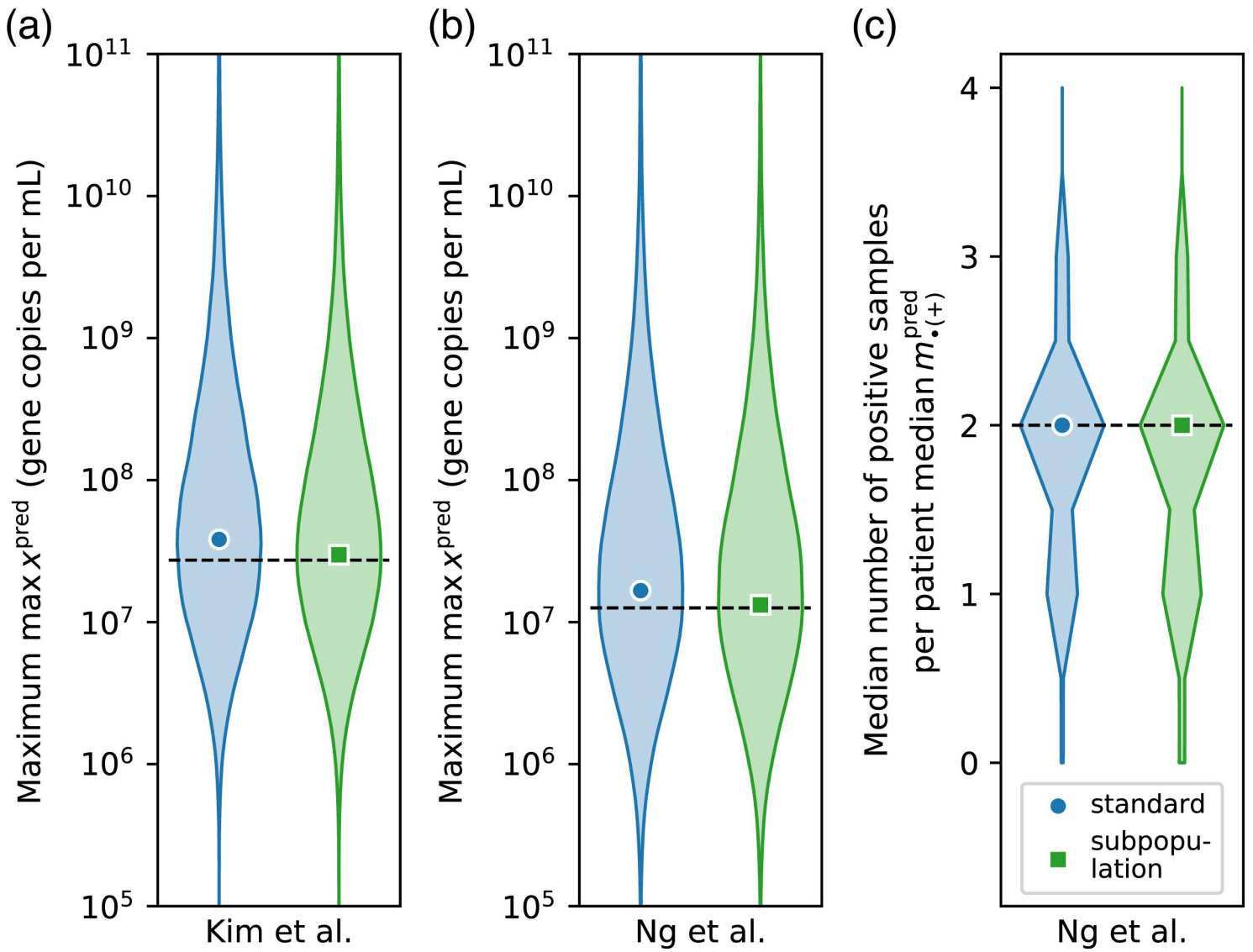
Two studies reported summary statistics without microdata.

- Kim et al. (2020) collected 129 samples from 38 patients; reported maximal concentration.
- Ng et al. (2020) collected 81 samples from 21 patients; reported maximal concentration and median number of positive samples per patient.

We replicated these studies *in silico* to predict summary statistics.

Speaker notes

- We assume one sample from each patient and the remainder of samples are randomly allocated to patients.
- Kim also reported number of positive and negative samples but we were not able to replicate the statistics because they do not report the limit of detection.



Speaker notes

- Studies without sample-level information can be replicated using the model which gives us confidence: The model can make predictions out of sample.
- The maximal observed value for Kim et al. is larger than Ng et al. That is expected: The larger the number of samples, the more likely we are to observe a very large value.
- We can also replicate other summary statistics, such as the slightly unusual median number of positive samples per patient reported by Ng et al.

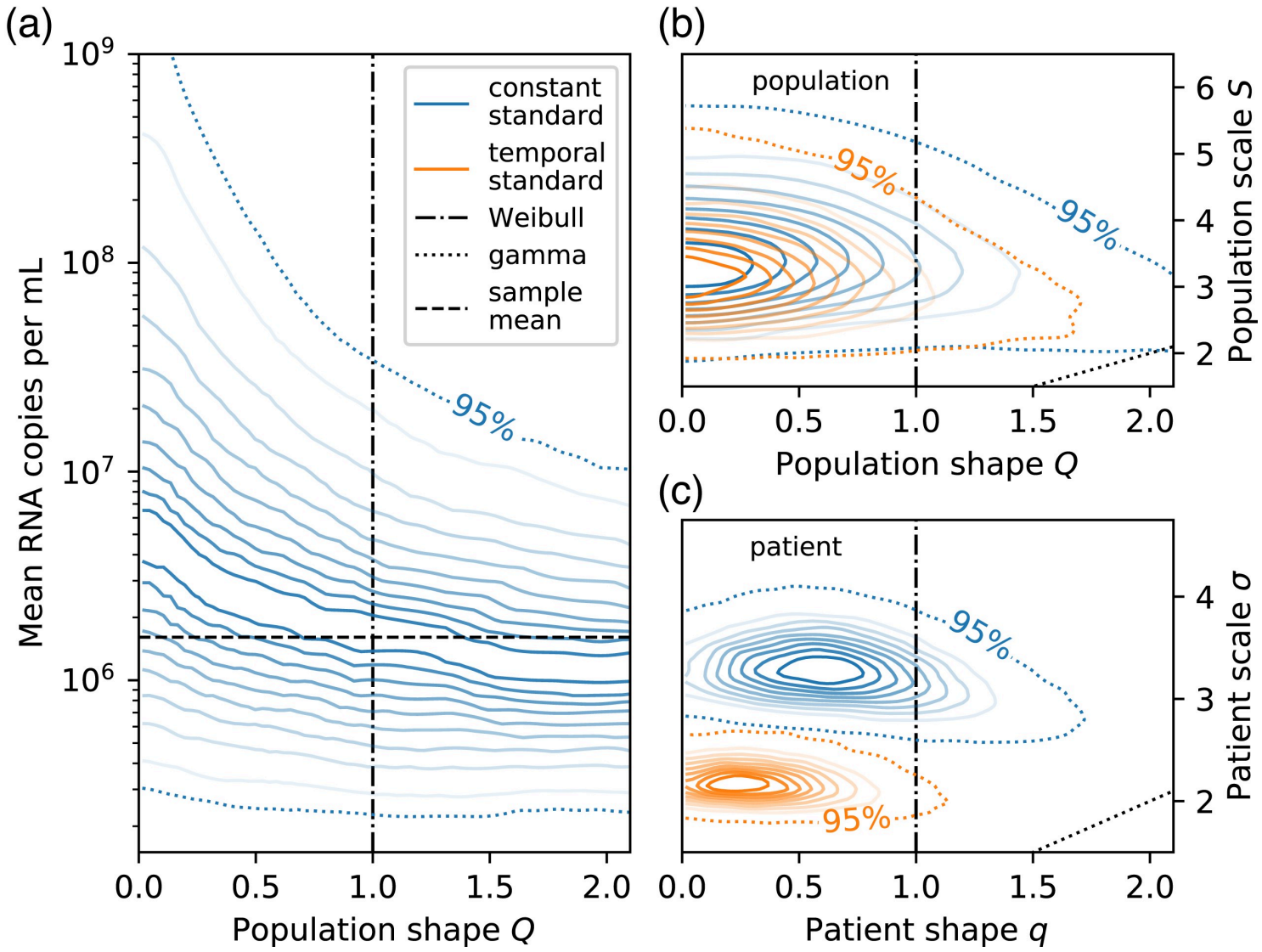
INVESTIGATING THE TAILS

The tail of the generalized gamma distribution is controlled by the shape parameter Q.

- $Q = 0$ recovers the log-normal distribution.
- $Q = 1$ recovers the Weibull distribution.
- $Q = S$ recovers the gamma distribution.

Speaker notes

- This is an example of continuous model expansion, an alternative to discrete model selection and Bayesian model averaging (cf. lecture 19).



Speaker notes

- The patient scale σ is much smaller for models with temporal shedding profile because the profile captures a lot of the variance and the residuals are smaller.
- The predicted mean increases with smaller shape Q because the tails get heavier (up to $Q = 0$ which corresponds the log-normal distribution).

CONCLUSIONS

- Modeling is essential to constrain viral RNA shedding.
- Early shedding behavior can reconcile clinical and wastewater-based data.
- Data are consistent with everyone shedding but to different degrees.

Speaker notes

No notes on this slide.

WHAT'S NEXT?

- Promote and expand the Shedding Hub.
- More flexible non-parametric shedding profiles, e.g., using Gaussian processes.
- Random effects for demographics, studies, gene targets, viral variants, …

Speaker notes

No notes on this slide.

LEARNINGS

- Know your data. Model the most “raw” data you can get your hands on. But not so raw that you can’t make progress.
- Bayesian hierarchical models are a great “language” for building complex models from simple building blocks.
- Really understanding your model matters for both science and computation.
- Model checking is essential!
- Validation on held-out data gives confidence.
- Interpreting parameters can be challenging; interpreting the posterior predictive distribution is often easier and more relevant.

Speaker notes

No notes on this slide.