Supplementary methods: Statistical analysis

Contents

1 Comparing swabs

1.1 Experimental design

1.1.1 Swab efficacy

In this experiment, virus was diluted 10-fold in plain DMEM from 5 x 10^5 pfu/ml down to 5 x 10^{-4} pfu/ml. For each dilution a swab (as identified Materials and Methods) was dipped in the virus dilution and transferred to a different vial containing 0.5 ml of plain DMEM. A sample (140 µl) of this resuspension was used to isolate RNA using the QIAmp Viral RNA mini kit, as per manufacturer's instructions. An RT-qPCR assay was then run as described in the Materials and Methods. This was repeated twice, two individuals performed the entire process independently.

1.1.2 Swab sampling volume

In this experiment, we aimed to measure the average volume collected by the different swabs being evaluated. A tube was filled with 1 ml of plain DMEM and weighed. A swab was used to collect DMEM the same way as it was done above. The tube was weighed again and the difference in weight was used to estimate the volume of DMEM collected by the swab. Comparing the weights of the empty tube with that of the tube + 1 ml of DMEM allowed an estimation of the density to calculate the volume sampled from the difference in weights. This was done with 5 different units for each kind of swab.

1.2 Statistical model

Over a wide-enough range of concentrations, biological assays often produce sigmoidal responses. Simple plots of the data revealed that this was the case here.

In our experiment, the swabs are expected to essentially perform a dilution of the original sample. The swabs themselves are not expected to, for example, release reverse-transcription or PCR inhibitors into the resuspension medium. In that context, the parameters of the sigmoidal curve that depend on the assay are

expected to be the same across the swabs tested. One difficulty with the data set is that the bottom plateau is never reached; however, there are constraints on the possible values for this parameter (e.g. it cannot be negative; it cannot be too close to 0 [e.g. 1 or 2] or the system will not call a C_q; it must be low enough to allow all the existing data to happen).

1.2.1 Fitting the swab comparison data in GraphPad Prism 8

Initially, we attempted to fit the 4-parameter curve using GraphPad Prism 8. Since the C_q values decrease with increasing virus concentration, we fitted a "log(inhibitor) vs response - variable slope" model. GraphPad Prism defines this model as:

$$\text{Response} = \text{Bottom} + \frac{(\text{Top} - \text{Bottom})}{1 + 10^{((\log_{10} \text{IC}_{50} - X) * \text{HillSlope})}}$$

Where the HillSlope is restrained to be negative, to ensure that the response decreases with increasing concentration.

As stated above, the swab experiment can be thought of as a dilution of the original virus solutions. In that context, the values of Bottom, Top, and HillSlope were constrained to be equal across all conditions. This leaves the midpoint (IC_{50} above) as the only parameter that changes across conditions.

Running the regressions with those constraints leads to an estimate of the Bottom parameter that is negative. Since such a result would be impossible, we attempted to add a constraint setting a minimum value for the Bottom parameter. Constraints were compared setting minimum values from 2 to 9 C_qs. This confirmed that the midpoint estimate is very sensitive to the location of the Bottom, as would be expected from the equation. In those attempts, the calculations would always hit the constraint, preventing calculation of the uncertainty of the midpoint.

1.2.2 Modelling the swab comparison data using a bayesian model

Bayesian models allow the estimation of the most likely values of parameters by combining 4 different inputs: 1) the equation(s) relating the parameters and the experimental variables (both fixed [e.g. virus concentration] and measured [C_q]); 2) the likelihood, representing the distribution of errors between expected and measured values of measured variables; 3) the data; and 4) the information encoded in the prior distributions of the parameters. Bayesian models are not frequently used for most applications due to two main factors: they can be computationally intensive (which made them all but impossible to use until modern powerful desktop computers became widespread) and the use of prior distributions, which can possibly be used to force prespecified outcomes (unconsciously or consciously). However, such models allow us to average over unknown quantities much more easily than traditional (least-squares-fitted) models.

The model of the C_q vs virus concentration (ln-transformed) was specified as:

$$\text{Response} = \text{Bottom} + \frac{\text{Span}}{1 + e^{((X - \ln \text{EC}_{50}) * \text{HillSlope})}}$$

Note that the subtraction was reversed, this allows us to make all parameters strictly positive, and that we use the Span rather than the (Top - Bottom) formulation. The full specification of the model is as follows:

$$\begin{aligned}
C_q &\sim \text{Normal}(\mu_i,\ \sigma) \\
\mu_i &= \text{Bottom} + \frac{\text{Span}}{1 + e^{((X - \ln \text{EC}_{50\,i})\ *\ \text{HillSlope})}} \\
\sigma &\sim \text{Normal}(0,\ 2)T[0, \infty] \\
\text{Span} &\sim \text{Normal}(35,\ 10)
\end{aligned}$$

$$\ln(\text{HillSlope}) \sim \text{Normal}(0,\ 2)$$

$$\ln \text{EC}_{50_i} \sim \text{FlatNormal}\left(-10,\ 16,\ 1,\ 1\right)$$

$$\text{Bottom} \sim \text{FlatNormal}(2,\ 9.5,\ 0.5,\ 0.5)$$

The index i iterates over the different conditions, the index relating the C_q to its concentration X within each condition was omitted from the notation. The priors for the midpoints ($\ln \text{EC}_{50[\text{i}]}$) and the Bottom are specially designed to provide a flat region and "softly" exclude values outside of those regions (the prior probability outside the flat regions are not 0, therefore not impossible). The prior for the $\ln \text{EC}_{50[\text{i}]}$ (it is the same for all conditions) looks like this:

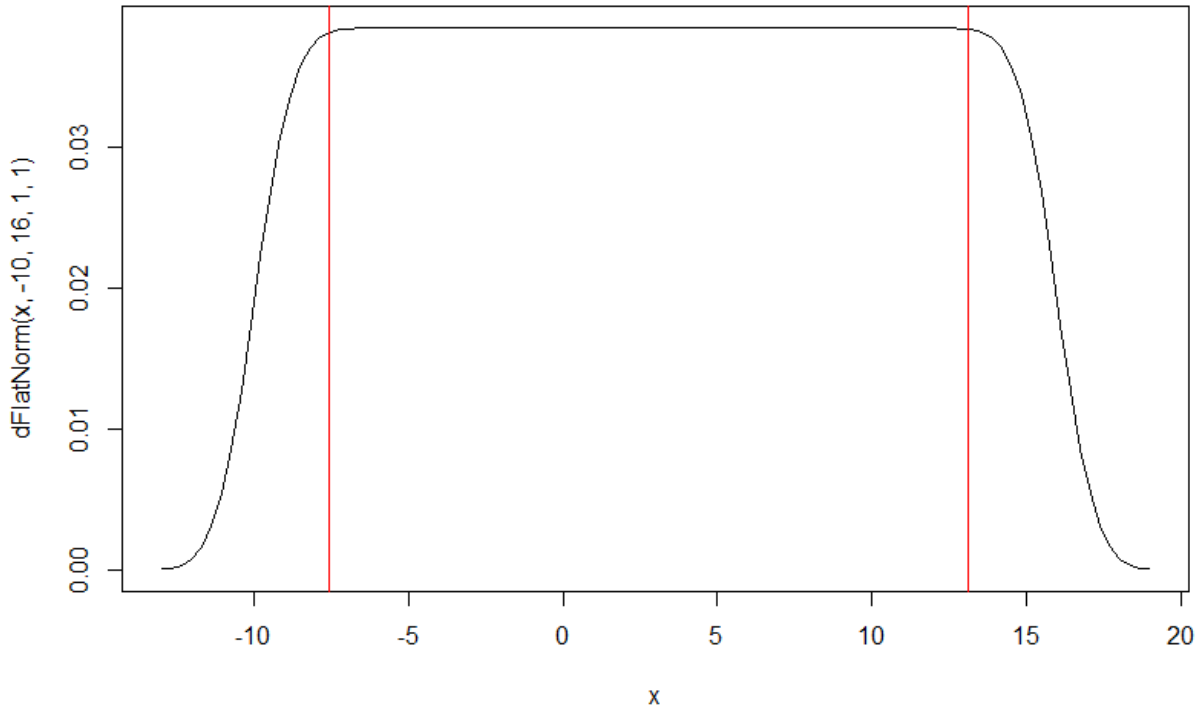


Figure 1: Prior distribution of the ln(EC50) on the ln(PFU/ml) scale. The red vertical lines represent the highest and lowest concentrations of virus tested.

It allows any value inside the range of the data. Essentially assuming that we would not run the model if the midpoint was not inside the tested range of concentrations.

The prior for the Bottom looks like this:

The bounds for the prior on the Bottom were chosen so that it is unlikely to be at the very early cycles, as those are always used to calculate the background signal that is subtracted from the fluorescence data. It also allows a small chance that the Bottom plateau is technically above the lowest C_q, since we expect measurement noise and the curve is meant to predict the average of the data.

We can use the same sampling algorithm we will use for the complete model to get an idea of what the prior looks like compared with the data:

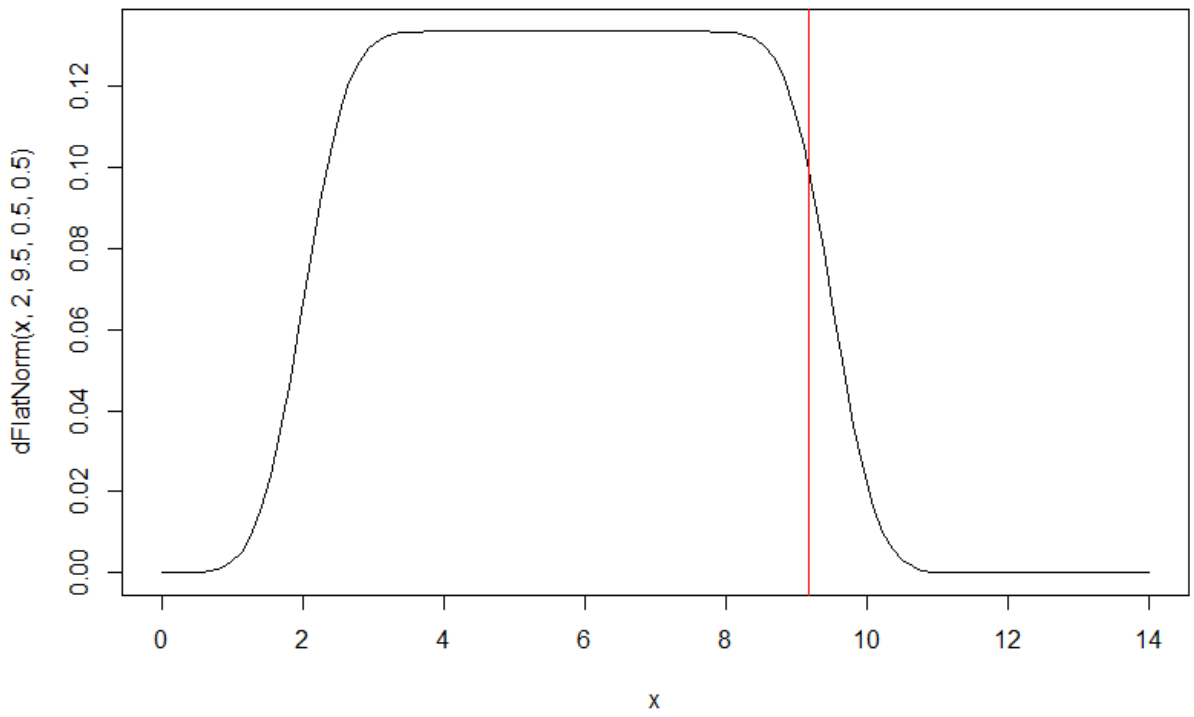
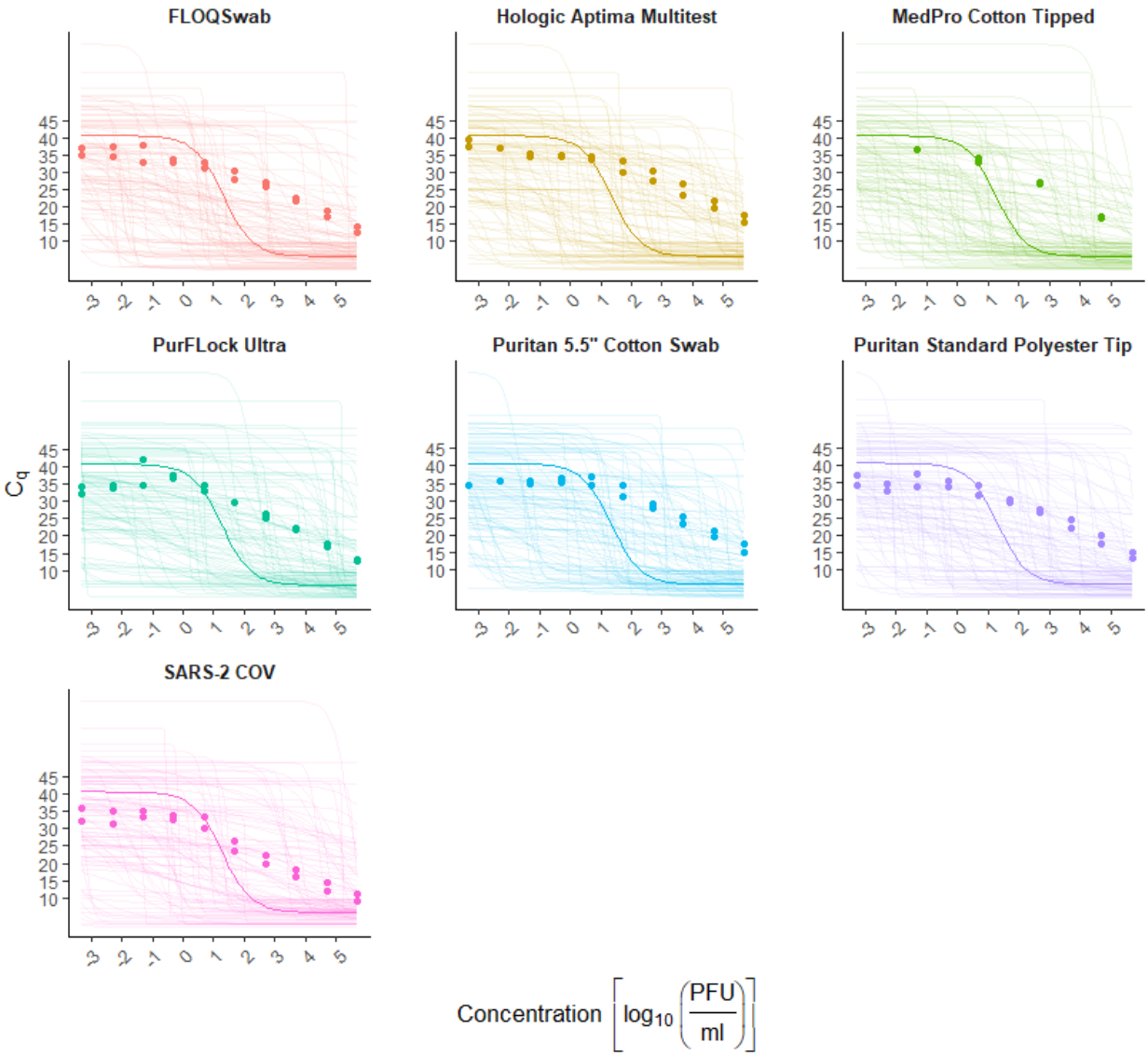


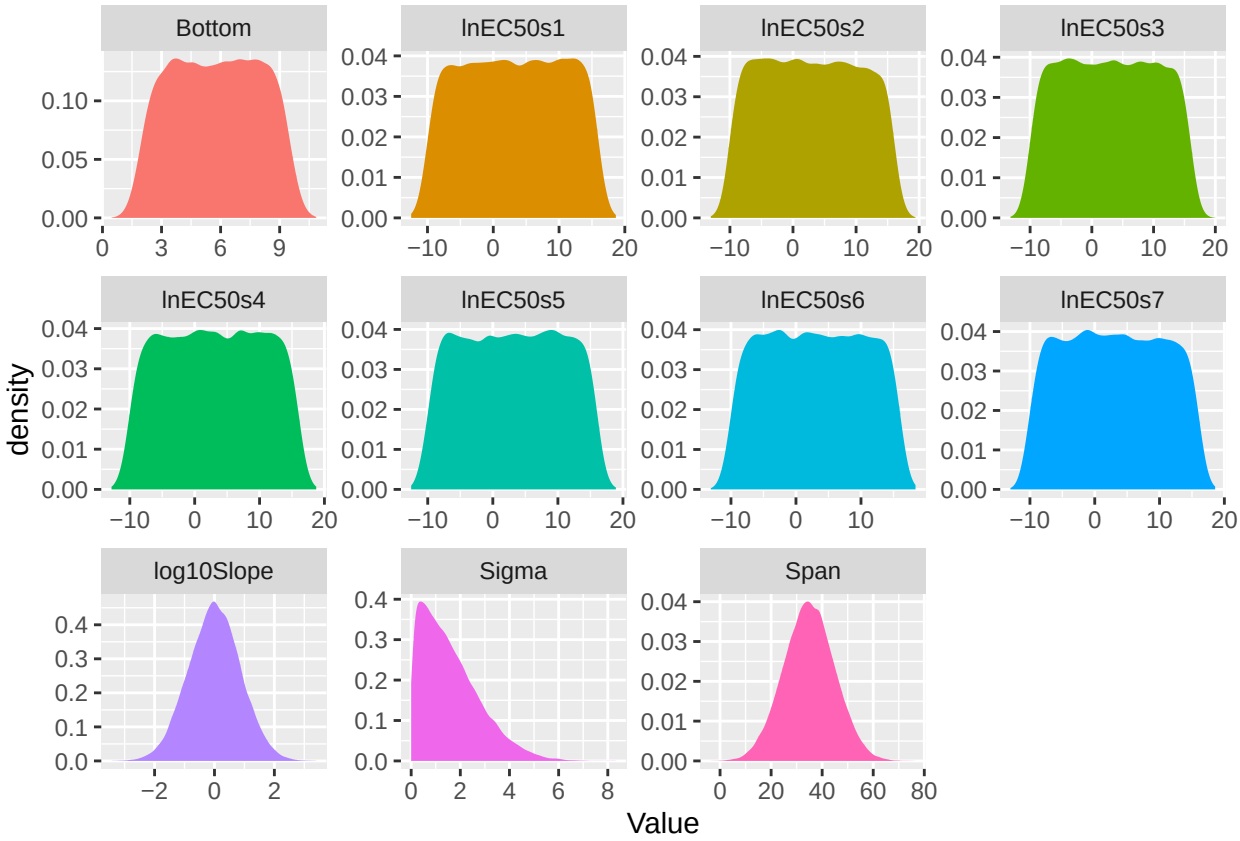
Figure 2: Prior distribution of the Bottom on the Cq scale. The red vertical line represents the lowest Cq measured.



We can use these graphs to confirm that the prior is not concentrating the curves near the data, but it does seem to allow curves that should fit the data.

The full summary of the parameters is in the table and graph below.

Parameter	Mean	Median	L_95	L_50	U_50	U_95
Bottom	5.764416	5.7813467	1.9566409	3.2918123	7.0423776	9.509660
lnEC50s1	3.077579	3.0801254	-9.5609272	0.2752041	13.1562141	15.480870
lnEC50s2	2.887631	2.8092241	-9.6385612	-8.3773899	4.4632847	15.529656
lnEC50s3	2.950776	2.9457050	-9.3831427	-7.3381930	5.5467621	15.699695
lnEC50s4	3.024900	3.0186032	-9.5614149	-0.5231116	12.2879487	15.595898
lnEC50s5	3.025150	3.0786316	-9.5760752	-1.3920843	11.5009883	15.558278
lnEC50s6	2.990950	2.9682224	-9.5391906	-7.1579909	5.7037748	15.524478
lnEC50s7	2.919221	2.8661083	-9.5513825	-7.3249664	5.4841994	15.600170
log10Slope	-0.000794	0.0038821	-1.6780381	-0.5225493	0.6442449	1.725365
Sigma	1.599579	1.3570270	0.0001519	0.0004805	1.3572703	3.926282
Span	34.883566	34.8644996	14.9798574	28.6431827	42.0631798	54.025081



1.2.3 Modelling the swab volumes

Since the volume collected by each swab was measured on different experimental units than the ones used for virus sampling, we will use the average volume. Also, the curve parameters that we estimate provide estimates of the average C_q, so it makes sense to estimate the average recovery efficiency at the average swab transfer volume. Note that any volume can technically be put through the calculations.

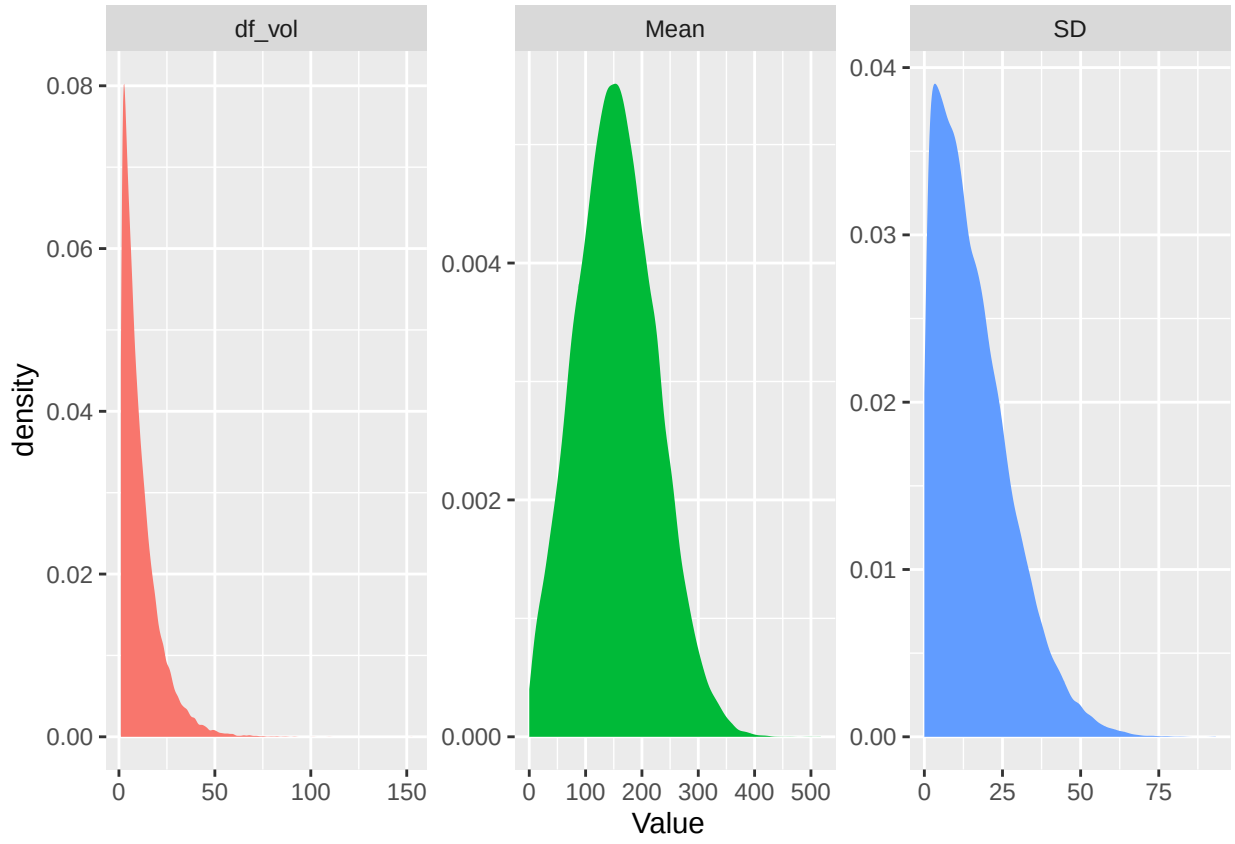
In order to have a more robust (i.e. less sensitive to extreme values) estimate of the mean, we will model the volumes as Student t-distributed measurements with estimated degrees of freedom. The full model, for each measurement m for swab i is described below:

$$\text{Volume}_{i,m} \sim \text{Student t}\,(\eta, \mu_i, \sigma_i)$$
$$\eta \sim \text{exponential}\,(1/10)\,T\,[1, \infty]$$
$$\mu_i \sim \text{Normal}\,(150, 75)\,T\,[0, \infty]$$
$$\sigma_i \sim \text{Normal}\,(0, 40)\,T\,[0, \infty]$$

By using the Student t distribution we can allow the data to generate more outliers than when assuming that the volumes are normally distributed. In the case where the data is fairly normal, the degrees of freedom should be large. The priors on the means and standard deviations are left intentionally vague to cover swabs with low and high volumes. Both are truncated to be positive as standard deviations and volumes cannot be negative. The prior on the degrees of freedom implies that the data is likely to be not very normal, but is relatively weak (the exponential prior does not normalize as strongly as the normal prior; especially if the data suggest that large values are more likely).

Here is the summary for the prior on the swab volumes (all the means and SDs are the same, so only one of each is shown here):

Parameter	Mean	Median	L_95	L_50	U_50	U_95
df_vol	11.03770	8.020096	1.0010037	1.0010037	8.020338	30.75749
Mean	154.83997	152.905590	15.0903595	98.7454038	196.522974	283.81156
SD	15.93321	13.419494	0.0019508	0.0034784	13.422635	39.20245



1.2.4 Fitting the models

Both the curve fitting and the modeling of the swab volumes are combined in a single model. This creates a joint posterior, making it easier to carry the uncertainties through all downstream calculations. It also increases the uncertainty of individual parameter estimates based the total number of parameters in the model (this is a consequence of the calculations, not an adjustable aspect of the model).

The model is sampled from using Stan version 2.19.3. The Stan code for the model is:

```
functions{
  //The prior for the Bottom and ln(EC50)
  real Flat_Normal_lpdf(real y, real mu1, real mu2, real sigma1, real sigma2){
    real first_term;
    real second_term;

    first_term = Phi_approx((y - mu1) / sigma1);
```

```
    second_term = Phi_approx((y - mu2) / sigma2);

    return(log(first_term - second_term) - log(mu2 - mu1));
  }
}
data {
  //Data for the swab comparison
  int<lower=0> N;
  vector[N] Cq;
  int<lower = 1> N_swabs_curve;
  int<lower = 1, upper = N_swabs_curve> swab_curve[N];
  vector[N] ln_conc;

  //Data for the volume
  int N_vols;
  int N_swabs;
  int<lower = 1, upper = N_swabs> swab[N_vols];
  vector[N_vols] volumes;

  //A simple switch, if fit = 0, then we sample the prior, if fit = 1, then we sample the posterior
  int<lower = 0, upper = 1> fit;
}
parameters {
  real Span;
  real Bottom;
  real ln_Slope;
  vector[N_swabs_curve] lnEC50;
  real<lower = 0> sigma;

  vector<lower = 0>[N_swabs] mean_vol;
  vector<lower = 0>[N_swabs] sd_vol;
  real<lower = 1> df_vol;
}

model {
  Span ~ normal(35, 10);
  Bottom ~ Flat_Normal(2, 9.5, 0.5, 0.5);
  ln_Slope ~ normal(0, 2);
  for (i in 1:N_swabs_curve)
    lnEC50[i] ~ Flat_Normal(-10, 16, 1, 1);
  sigma ~ normal(0, 2);


  df_vol ~ exponential(1.0 / 10.0);
  sd_vol ~ normal(0, 20);
  mean_vol ~ normal(150, 75);

  if(fit){
    real Slope = exp(ln_Slope);
    vector[N] Cq_hat = Bottom + exp(log(Span) - log1p(exp(Slope * (ln_conc - lnEC50[swab_curve]))));

    Cq ~ normal(Cq_hat, sigma);
    volumes ~ student_t(df_vol, mean_vol[swab], sd_vol[swab]);
  }
```

}

The parameter of interest is the ability of the swab to recover, in the resuspension, as much of the virus that was present in the volume of sample it transferred. If a swab is considered as a way to collect a mostly liquid sample from a surface and dilute it into a resuspension medium we can calculate, from the shift in the EC_{50} between the original and diluted solutions, the efficiency with which the virus was released in the resuspension medium. This is likely to be a simplified model of swabbing, neglecting the roles of molecular interactions between the virus and the swab.

We do this by using the estimated average swab volume to calculate the expected dilution factor that should be produced by the swab. Next, we calculate the expected EC_{50} using the expected dilution factor and the EC_{50} of the original solution. The ratio of the expected EC_{50} to the estimated EC_{50} (which should be larger, as recovery should not be perfect) provides an estimate of the recovery from the swab.

Code for the calculations and graphing is provided in supplementary materials.

2 Comparing the resuspension buffers

2.1 Experimental context

The MedPro Cotton Tipped swab was used to sample a virus solution at different concentrations and was resuspended in various media (PBS, normal saline, 100% ethanol, Virus Transport Medium (VTM), and DMEM). The transferred samples were used to isolate RNA at various time points (0, 1, 2, and 3 days post resuspension) and the C_q was measured.

2.2 Model

Here we simply model the change in C_q over time as a linear trend. The long-term change in RNA is likely non-linear, but the experimental design did not extend far enough to make the non-linearity apparent. The change should also be monotonic, so a straight line should still provide some information about the overall rate of change over time.

Since at least one medium has samples where virus could not be detected by RT-qPCR, it is important to have an estimate of the probability of detecting virus at all. To this end, the linear model is built into a hurdle model, which estimates the probability, for every dose for every medium, of detecting virus (this probability is called θ).

For medium m and sample i we have:

$$p(C_{q_{m,i}} \mid \theta, \mu_m, \sigma) = \begin{cases} (1-\theta) & \text{if} \quad \text{Undetected} \\ \theta * \text{Normal}\left(\mu_m, \sigma\right) \ T[-\infty, 45] & \text{Otherwise} \end{cases}$$

$$\theta \sim \text{Beta}(1,\ 1)$$

$$\mu_m = \text{Intercept}_m + \text{Slope}_m * \text{Day}_i$$

$$\text{Intercept}_m \sim \text{Normal}\left(25, 10\right)$$

$$\text{Slope}_m \sim \text{Normal}\left(0, 1\right)$$

$$\sigma \sim \text{Normal}\left(0, 1\right) \ T[0, \infty]$$

The swab sampling volumes are also included, using the same model as before. The prior on the probability of detection is left uninformative, the Beta(1, 1) assigns an equal probability to all values between 0 and 1, inclusively. This prior is also slightly normalizing, so that even if all samples are detected, the posterior mean and median will not be at 1 (although the posterior mode might). The prior on the intercepts softly restricts

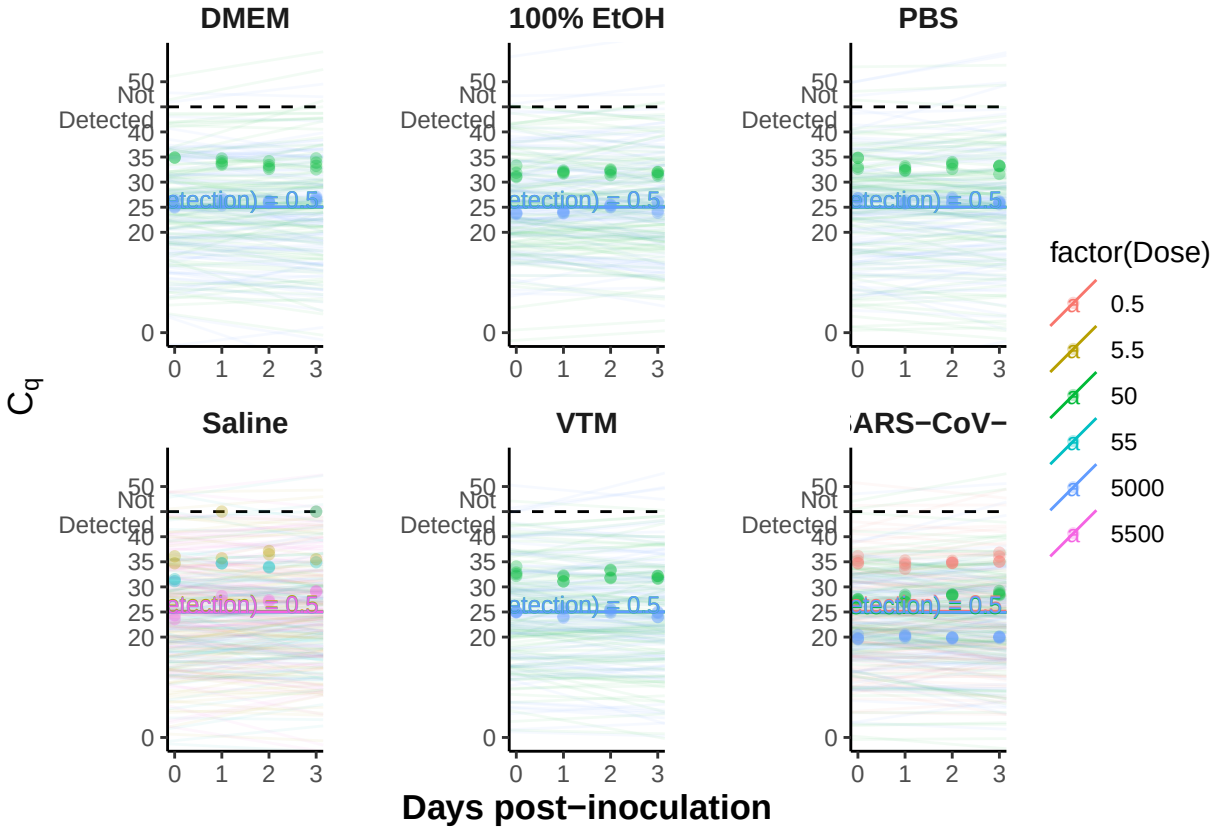
the it to a reasonable range (the C_q at time 0 should be measureable). The prior on the slope was chosen both for normalizing properties and from our experience with viral suspensions. Viral suspensions do not generally loose many C_qs over a day. This is not necessarily true when such suspensions are prepared from patient samples (such as oro-pharyngeal swabs), which will contain many contaminating agents, including bacteria, proteases, and nucleases.

The full code for the model (implemented in Stan) is below:

```
data{
  int N;
  int N_media;
  int<lower = 1, upper = N_media> media[N];
  vector[N] dpi;
  vector[N] Ct;

  int N_vols;
  int N_swabs;
  int<lower = 1, upper = N_swabs> swab[N_vols];
  vector[N_vols] volumes;

  int<lower = 0, upper = 1> fit;
}
parameters{
  vector[N_media] intercepts;
  vector[N_media] slopes;
  real<lower = 0> sigma;
  vector<lower = 0, upper = 1>[N_media] thetas;

  vector<lower = 0>[N_swabs] mean_vol;
  vector<lower = 0>[N_swabs] sd_vol;
  real<lower = 1> df_vol;
}
model{
  vector[N] means = intercepts[media] + slopes[media] .* dpi;

  intercepts ~ normal(25, 10);
  slopes ~ normal(0, 1);

  sigma ~ normal(0, 1);

  df_vol ~ exponential(1.0 / 10.0);
  sd_vol ~ normal(0, 20);
  mean_vol ~ normal(150, 75);

  thetas ~ beta(1, 1);

  if(fit){
    for (i in 1:N){
      if(Ct[i] >= 45){
        0 ~ bernoulli(thetas[media[i]]);
      } else {
        1 ~ bernoulli(thetas[media[i]]);
        Ct[i] ~ normal(means[i], sigma) T[, 45];
      }
    }
```

```
    volumes ~ student_t(df_vol, mean_vol[swab], sd_vol[swab]);
  }
}
```

The priors produce the following "fit":



As with the EC_{50}s in the previous section, the intercepts can be used to evaluate the early efficiency with which the various media can release viruses from the MedPro Cotton swab. The calculation is similar to that above, where the volume transferred by the MedPro Cotton swab is used to estimate an expected C_q and the difference between the expected C_q and the estimated intercept provides an estimate of the recovery efficiency.

The code for fitting the model and performing graphing/calculations is in Supplementary Materials.