Fortnightly meeting 4th September 2019

Mark Nelms

4/9/2019

Contents

Questions from last time

After our last meeting we had questions to answer:

- Where a chemical has multiple NOECs, how often is the subchronic the lowest NOEC?
- Can we calculate the 5th %ile without fitting a distribution?
- Why are our recalculated 5th %ile and TTC values for Cramer class I so different?
- Can we identify enriched chemotypes for chemicals where the local NOEC was driving the general NOEC in Escher et al?

1 Inhalation TTC

Filtering criteria for inhalation studies:

- Test species: Rat, mouse, other rodents, or rabbit
- Route of exposure: Inhalation
- Study duration: subchronic, chronic, reproductive, developmental, and multigeneration
- POD type: NO(A)EL, NOEL, NO(A)EC, NOEC

1.1 Before we begin

I went back and had a look at the code for calculating the 5th percentile and TTC values - I did keep all of the chemicals even if we didn't have the SMILES for the chemical. The only chemicals that were removed were those where the NOEC value was either `NA` or `0`. I had to do this, otherwise the `fitdist` function wouldn't allow me to calculate either value.

1.2 How often is the subchronic the lowest NOEC?

There are a total of 478 chemicals from the ToxVal Inhalation TTC dataset that fall into either Cramer class I, II, or III and have at least one non-NA toxicity value. The NA values are due to either the toxicity value being given in mg/kg-day or being unspecified in ToxVal (and so I couldn't calculate in terms of ppm). These 478 chemicals have 2,596 associated studies.

NB: I have removed NAs from the calculations but I haven't removed outliers (i.e. x1.5 inter-quartile range) from any of the calculations: outliers are identified by the grey "X" in Figures 1 and 2.

Of these 2,596 studies, only 105 are chronic studies, whilst there are 1,590 subchronic studies.

However, the above numbers include 132 chemicals that only have one study associated with them. This leaves a total of 346 chemicals that have multiple toxicity values.
Below, I've plotted a box and scatter plot of all of the associated toxicity values for each of the chemicals with multiple studies.

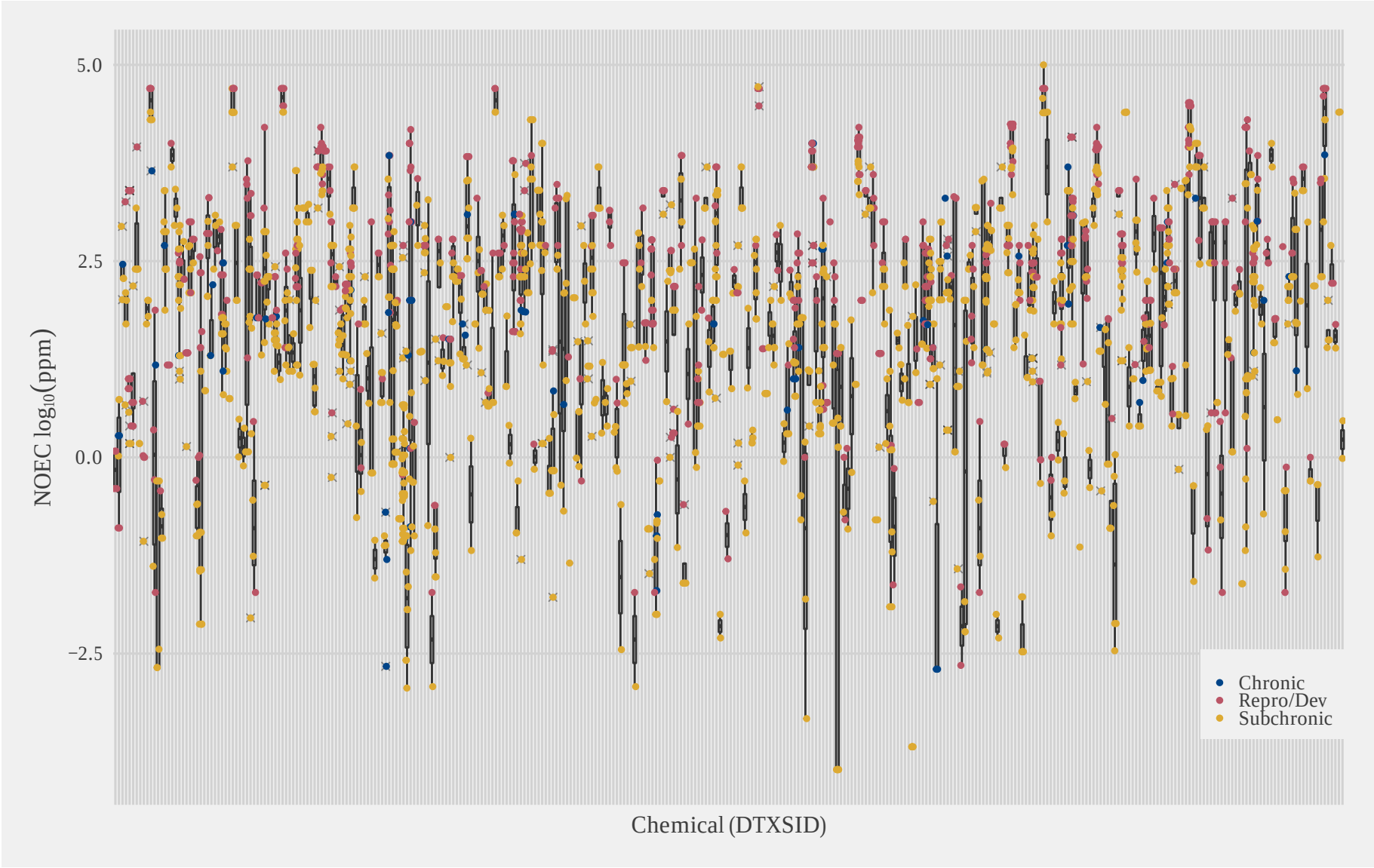

Figure 1: Scatterplot of NOECs for all chemicals with multiple studies

Because of the number of chemicals with multiple NOECs, Figure 1 only serves to back up the fact that the vast majority of the studies in our Inhalation TTC dataset are subchronic studies (red points), we can't really decipher much more than this.

Next, I filtered the 346 chemicals with multiple studies to retain only the ones that have both chronic **AND** subchronic data, this whittled the number of chemicals down to 53.

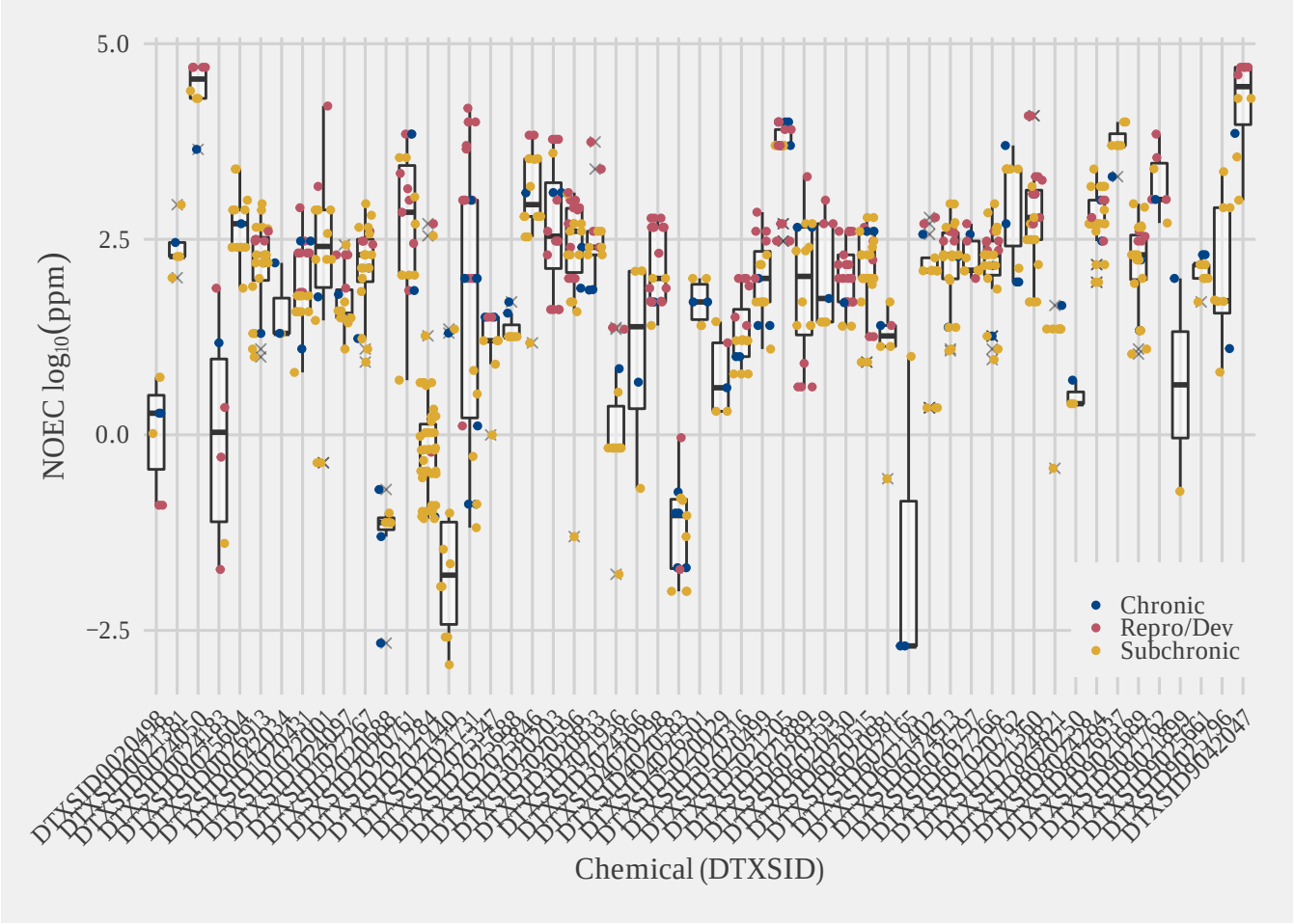

Figure 2: Scatterplot of NOECs for chemicals with both subchronic and chronic studies

As you can see from Figure 2 very few chemicals have a chronic study as their lowest toxicity value; in fact, there are only 6 chemicals. Even then, at least 2 of these would be removed for being outliers using the Tukey fence.

Of the 53 chemicals that have both chronic and subchronic studies, if we also class the reproductive/developmental studies as being chronic studies, then we get a total of 12 chemicals whereby the "chronic" study would be the toxicity value used for that specific chemical.

If we zoom back out to consider the whole set of 478 chemicals with at least one, non-NA, toxicity value from our ToxVal inhalation dataset, then there are a total of 9 chemicals where a chronic study is the lowest toxicity value, and 61 chemicals where either a chronic or repro/dev study is the lowest toxicity value.

Even with the 2-fold adjustment made to the subchronic studies to approximate a chronic study if only **9** chemicals out of a possible **478** chemicals are having their toxicity value driven by a chronic study can we say our inhalation TTC is for chronic exposure? (Even if we considered repro/dev toxicity studies to be chronic we'd only have 61 chemicals driven by a "chronic" exposure)

1.3 Can we calculate the 5th %ile without fitting a distribution?

Good news! It looks like the `quantile()` function enables you to make calculations without needing to fit a distribution to the values. However, as you can see from Table 1, this doesn't exactly make things better.

Table 1: Comparison of Published and Recalculated 5th percentiles

	Type of TTC	Cramer class	5th %ile units	Published 5th %ile	Recreated 5th %ile	
					Using fitdist	No distribution fitted
Escher						
	General	1	ppm	0.21	0.253	0.32
		2	ppm	0.028	0.026	0.04
		3	ppm	0.003	0.002	0.003
	Systemic	1	ppm	0.25	0.557	0.313
		2	ppm	0.52	0.156	0.565
		3	ppm	0.006	0.004	0.01
	Local	1	ppm	0.021	0.13	0.145
		2	ppm	0.028	0.019	0.038
		3	ppm	0.003	0.002	0.003
Carthew						
	Systemic	1	mg/kg-day	0.41	0.44	0.37
		3	mg/kg-day	0.07	0.044	0.067
	Local	1	mg/m3	1.4	1.229	1.355
		3	mg/m3	0.47	0.141	0.471
ToxVal						
	General	1	ppm	NA	0.032	NA
		2	ppm	NA	0.042	NA
		3	ppm	NA	0.014	NA

Not fitting a distribution to provide to `quantile()` does improve the 5th percentile calculation for pretty much all of the Carthew values: especially if you round to the same number of significant figures as are reported. But things are a little bit more complicated for the Escher dataset: the Systemic Cramer class I and II 5th percentiles are more in-line with the reported values, but that's about it.

However, by not fitting a distribution to the Escher data we don't need to adjust/remove the chemicals with a NOEC of 0.

1.4 Why are our recalculated 5th %ile and TTC values for Cramer class I so different?

As we saw in our last meeting (and also above in Table 1), the original and re-calculated values could be off by quite a bit: especially the local 5th percentiles which were different by ~1 order of magnitude!

To see how inherently variable the 5th percentile NOEC is likely to be, due to the data we have, for each affected organ type, I used bootstrapping to calculate 5000 samples (with replacement) for each affected organ type and used this information to calculate the median 5th percentile, as well as the upper and lower 95% confidence interval. The results of this can be seen below in Figure 3.

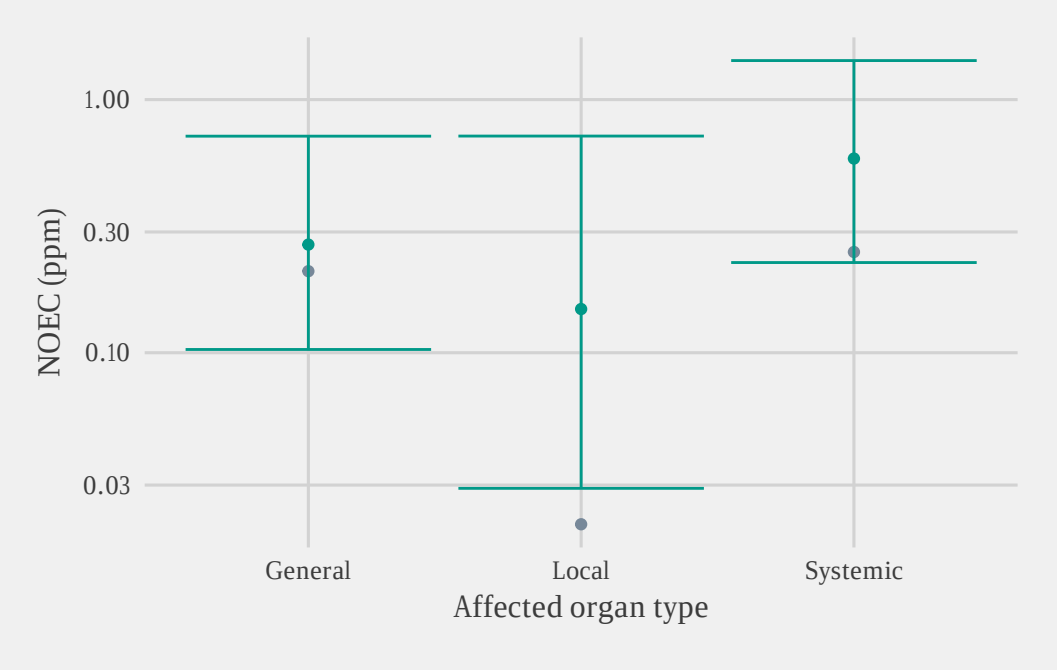

Figure 3: Median and 95% confidence intervals for General, Local, and Systemic 5th percentile NOEC values for Cramer class I from Escher et al. Published values are in grey.

As you can see above, the General 5th percentile is the one we can recalculate with the best accuracy - it's the closest to the published value in Escher et al (grey point). Meanwhile, the re-calculated Local 5th percentile is not accurate at all, with the published 5th percentile value not even being covered by the 95% confidence intervals after 5000 bootstrap samples. The published Systemic 5th percentile is (barely) covered by the 95% confidence interval using the data in the Appendix.

I'm not sure why we are able to predict the General 5th percentile to a much better degree than either the Local or Systemic 5th percentiles. Maybe it's because all of the chemicals in the dataset have a General NOEC (203), whilst the Local and Systemic affect organ types only have a subset of NOECs: 102 chemicals and 199 chemicals, repsectively. However, that doesn't quite make sense because there are only 4 fewer chemicals with Systemic data than with a General NOEC.

Additionally, the 95% confidence intervals are quite spread out. Both the General and Systemic 95% confidence intervals span almost 1 order of magnitude, whilst the confidence intervals for the Local values is closer to 1.5 orders of magnitude.

Bearing in mind that not fitting a distribution before calculating the 5th percentile doesn't make much of a difference (and in some instances took us further away from the published value, Table 1), I think there has to be something wrong with either the published NOECs or the 5th percentiles. This is especially true for the Local affected organ data for us to be almost 1 order of magnitude away from the published value and for the published value to beyond even the 95% confidence interval!

Either that or the way the log normal distribution is calculated using STATA is different from using `fitdist()`, but without any more information it's all just speculation.

To see how well we were able to recreate the 5th percentiles for the other two Cramer classes I ran the bootstrapping on them too. As you can see in Figure 4, we do much better for the other two Cramer classes (although, I'd take the class II results with a massive pinch of salt given that these values are based on 5/6 chemicals).

I do find it quite strange that we over estimate the 5th percentile for the Cramer class I chemicals and under estimate the 5th percentiles for the class III chemicals - but to a much lesser degree.

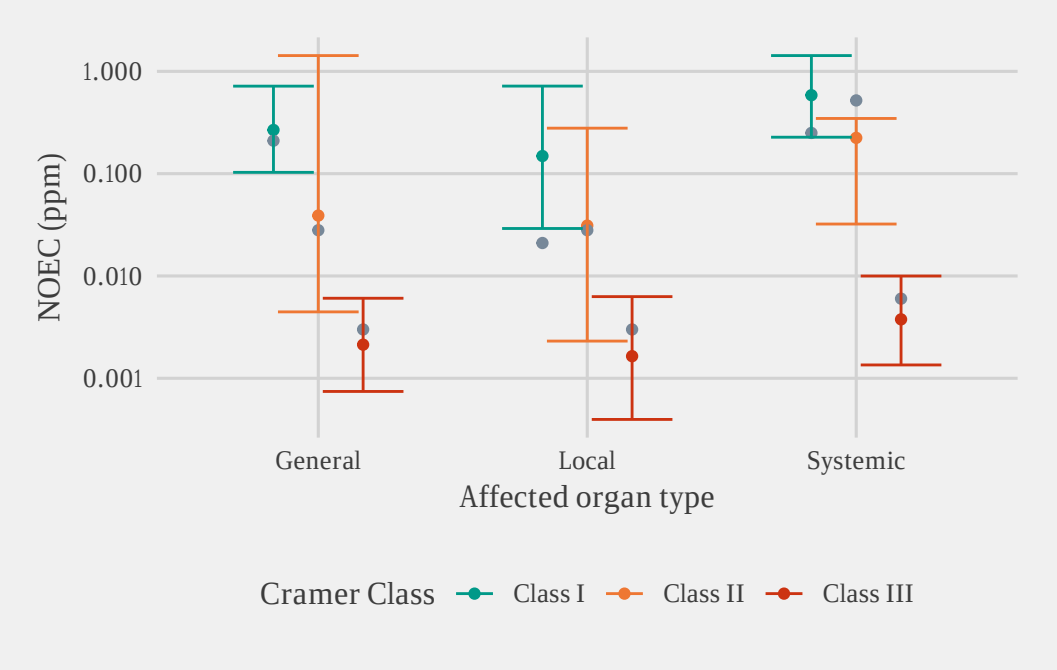

Figure 4: Median and 95% confidence intervals for General, Local, and Systemic 5th percentile NOEC values for Cramer class I from Escher et al. Published values are in grey.

Maybe I should run the chemicals we have structures for through Toxtree to recalculate the Cramer class and see what effect that has? For those chemicals we don't have strucutures for I'd just keep the same Cramer class.

1.5 Can we identify enriched chemotypes for chemicals where the local NOEC was driving the general NOEC in Escher et al?