Combining evidence on multiple endpoints in dose-response assessments: multivariate models

Wout Slob
RIVM
The Netherlands

We normally observe different PoDs (NOAELs or BMDLs) for different endpoints

- within the same study
- among different studies

"The endpoint with the lowest POD is the most sensitive endpoint "

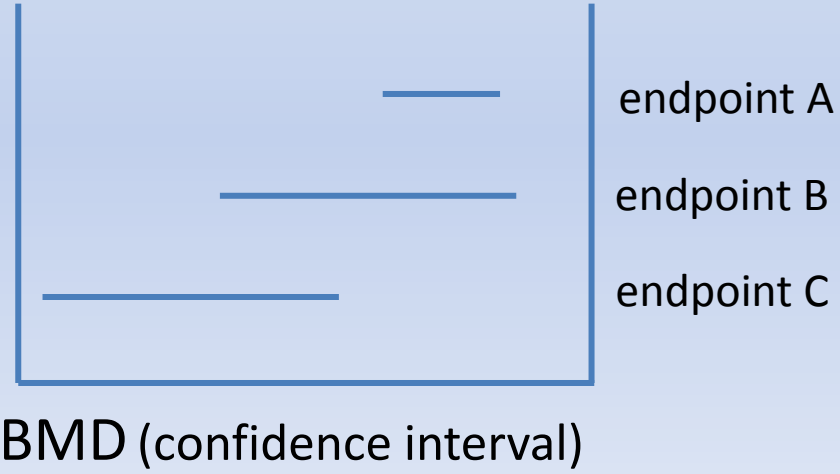
First thing to keep in mind: uncertainties in the PoD itself

- NOAELs are imprecise

  which could be the reason of NOAELs being different

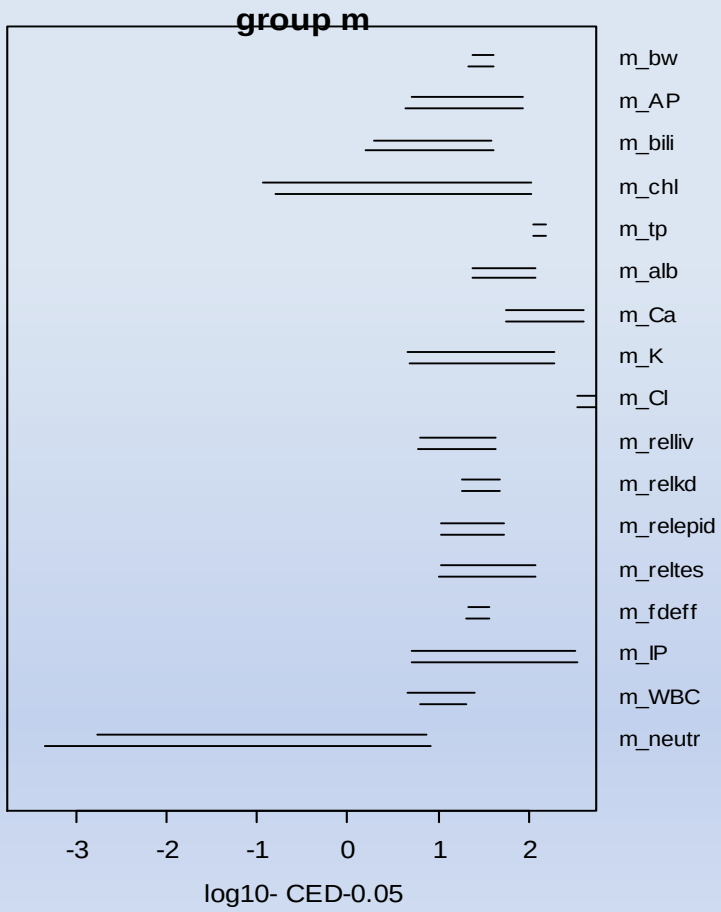
  (statistical sensitivity $\neq$ biological sensitivity)

- BMDLs may differ due to larger uncertainty in one endpoint over another



Example 1: Subchronic study (anonymous)

BMD CIs for all endpoints with a significant trend



BMR = 5% for all endpoints

Two CIs for each endpoint, relating to the exponential and Hill model

This can be done by one automated run (PROAST)

The crucial question in deriving BMDs for continuous endpoints is:

What value for the BMR should be used for each endpoint?

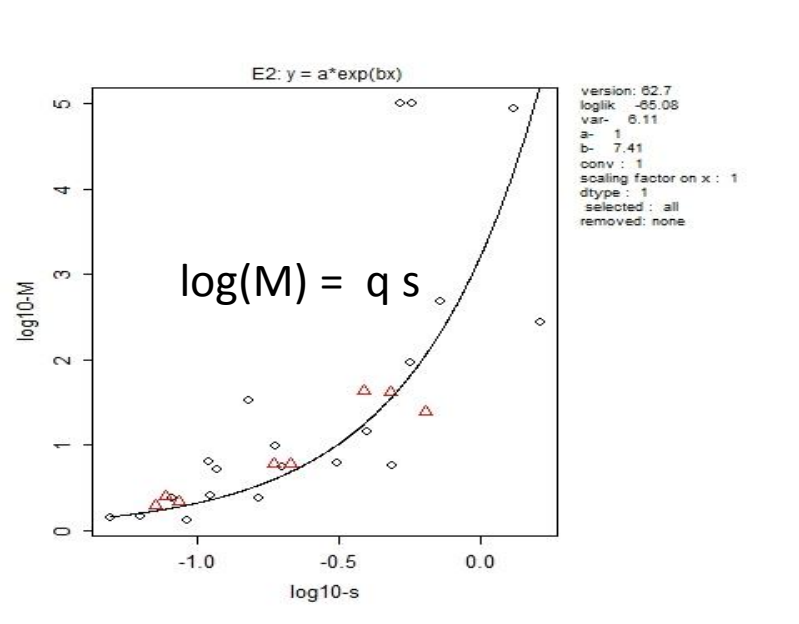
Slob (2017) presents a theory that may provide an answer:

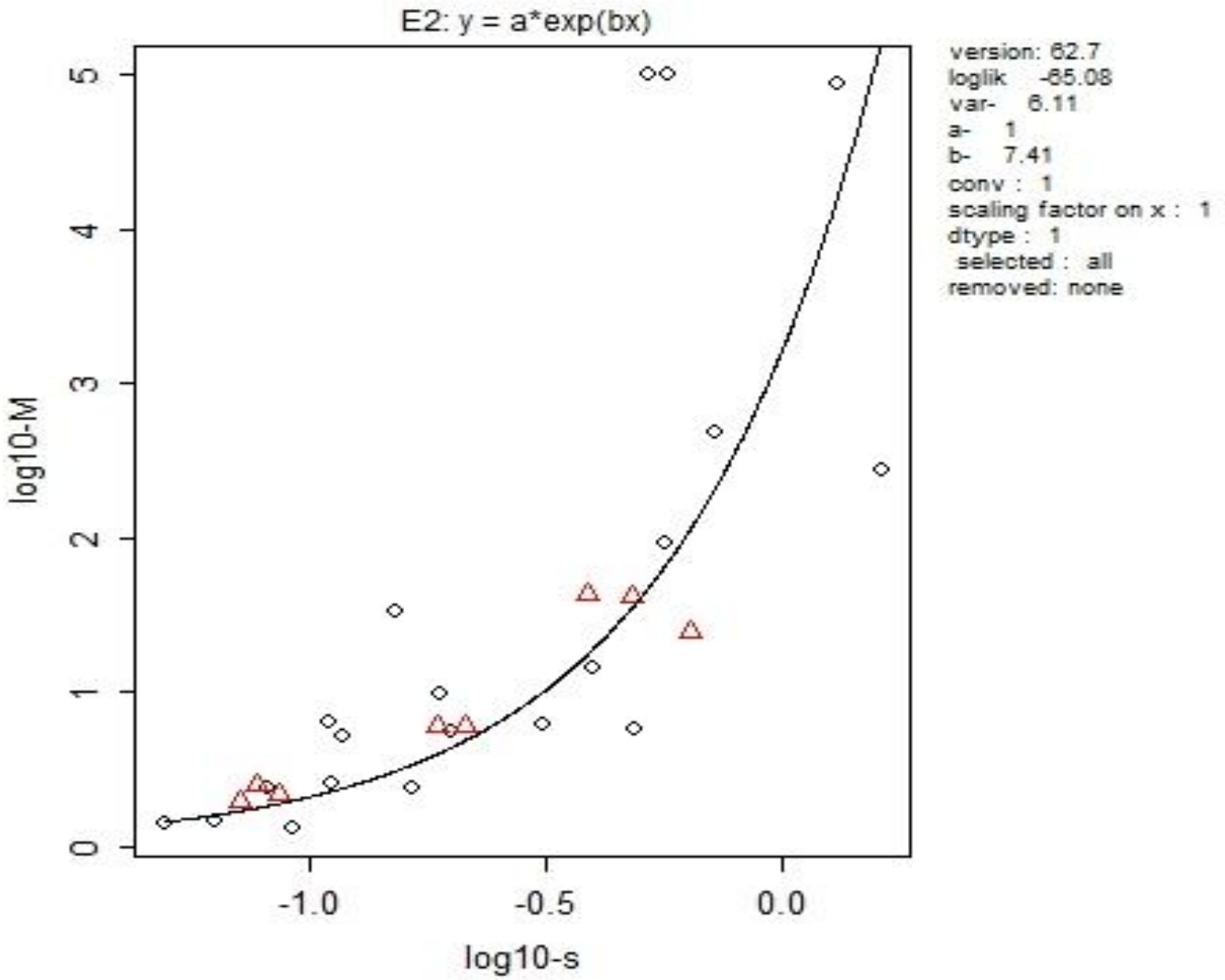
Scale the BMR to the maximum response: log(M)

Due to the correlation between M and s, scaling to s could be used as a proxy

s = within-group SD on log-scale

Question (hypothesis):
Are endpoints equally sensitive?
(when using the scaled BMR)

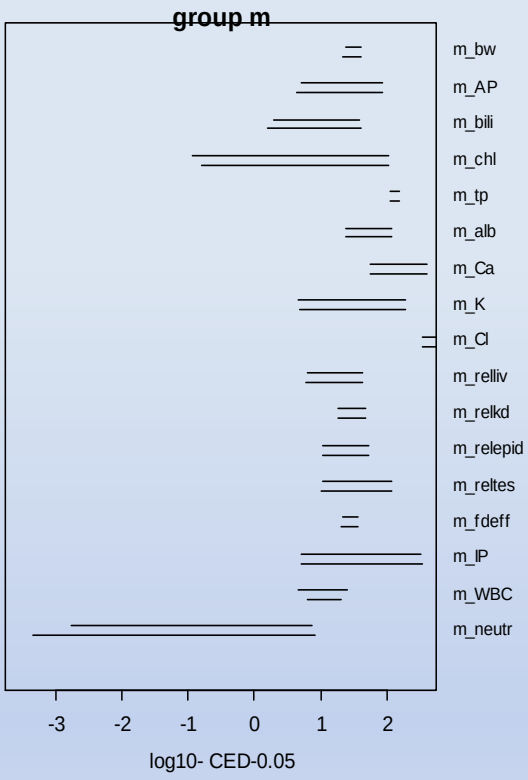




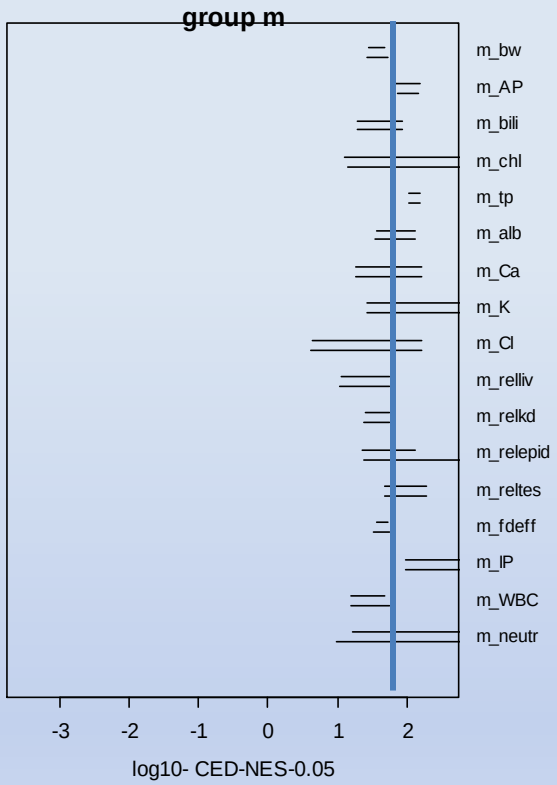
red triangles: estimates of M and s are based on multiple studies

Example 1: Subchronic study (anonymous)

BMD CIs per endpoint



BMD CIs per endpoint

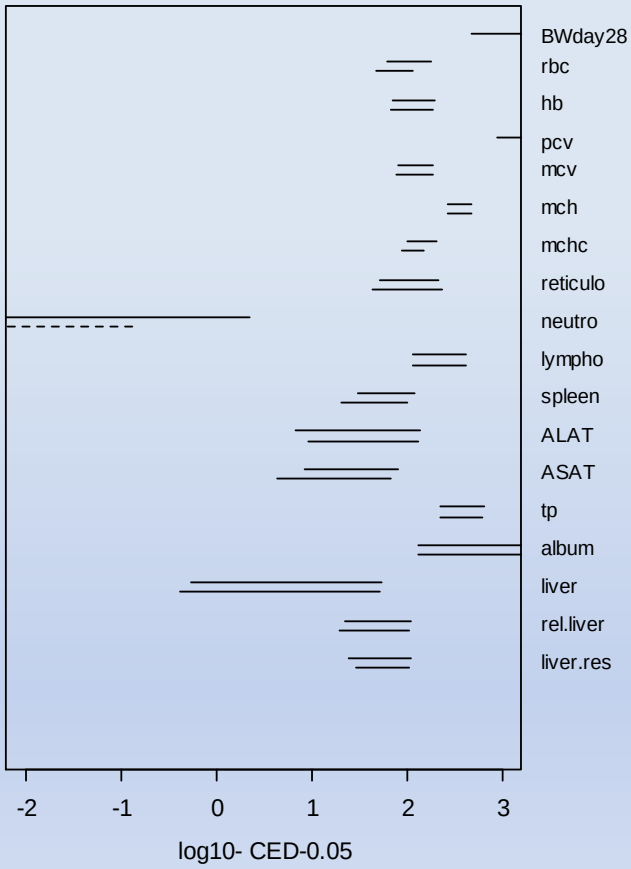


BMR = 5% for all endpoints

endpoint-specific value for BMR (scaled to within group s)

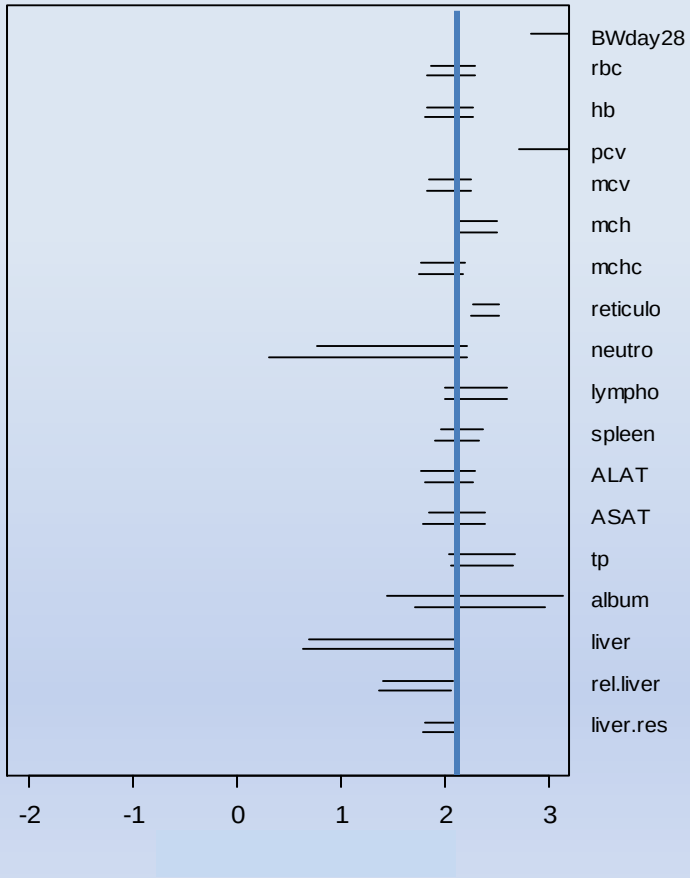
Example 2: 28-Day study with Rhodorsil Silane

BMD CIs per endpoint



BMR = 5% for all endpoints

BMD CIs per endpoint



endpoint-specific BMR

The scaling of the BMR is based on the “incidental” s in the study itself, rather than on the average of a large number of studies.

may explain the remaining (small) differences

I recently developed a model that reflects the ES-theory by substituting the maximum response parameter by q s, leaving just one parameter for the CED for all endpoints

within-group s_{endp}	dependent on endpoint
background response a_{endp}	dependent on endpoint
q in $\log(M_{endp}) = q\ s_{endp}$	q: common
steepness (d)	d: common
BMD	BMD: common

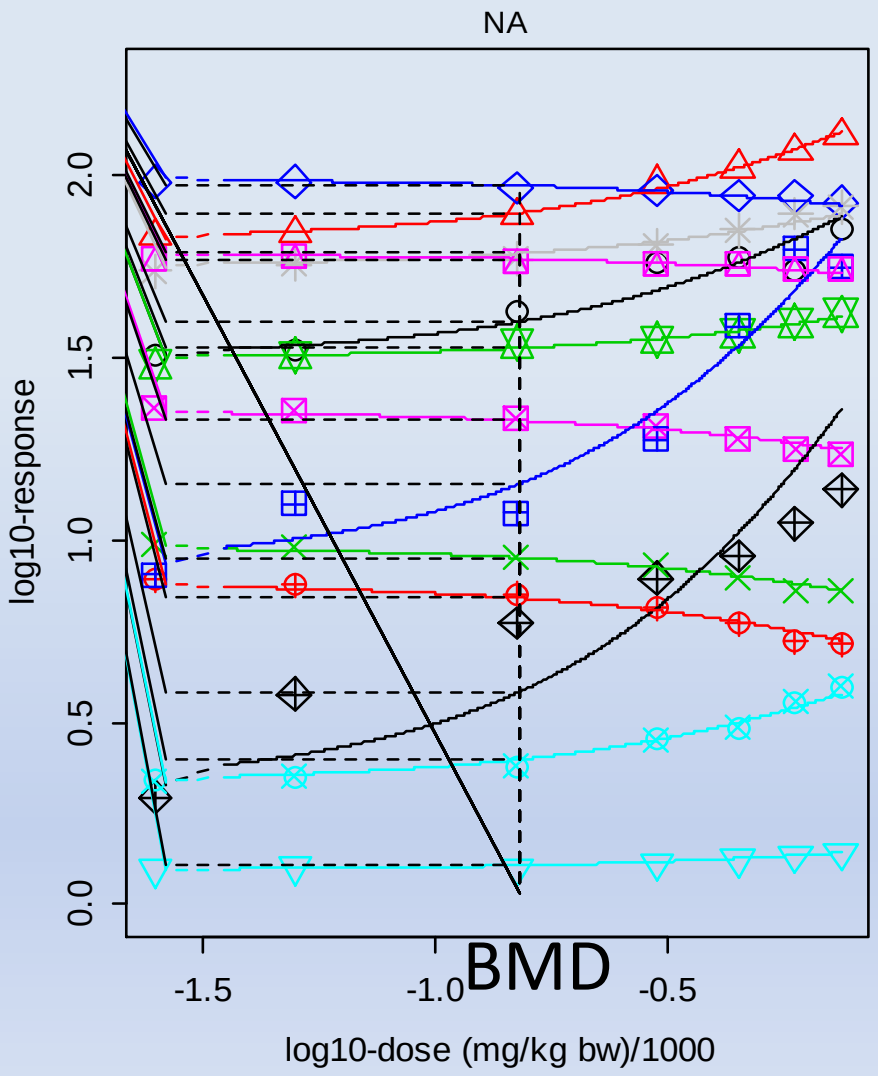
Model with endpoint-dependent background and s,
two shape parameters (d and q),
and ONE parameter for the BMD

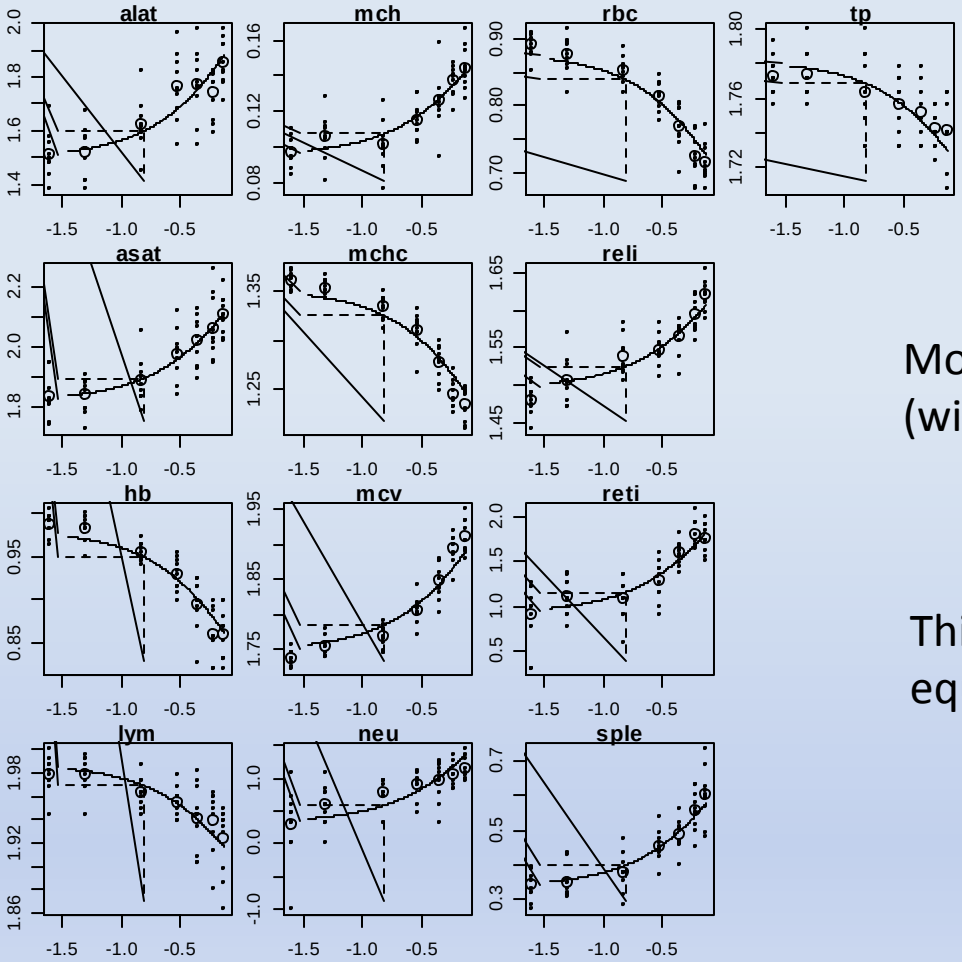
statistical challenge:

How to establish the confidence interval for that single BMD?

↓

multivariate methods





Model with the same CED for all endpoints (with endpoint-specific CES)

This model reflects that all endpoints are equally sensitive

Hypothesis not rejected

Correlations among endpoints (after correcting for the dose-response)

model 47		rhodorsil															
	alat	album	asat	hb	liver	lympho	mch	mchc	mcv	neutro	pcv	rbc	relaliver	reticulo	spleen	termbw	tp
alat	1	-0.064	0.631	-0.087	-0.185	-0.035	-0.102	0.096	-0.113	0.086	-0.16	-0.033	-0.049	-0.2	0.057	-0.118	-0.161
album	-0.064	1	0.072	-0.21	-0.108	-0.164	0.1	-0.328	0.334	0.092	0.344	-0.256	0.319	0.046	0.261	0.145	0.695
asat	**0.63**	0.072	1	-0.206	-0.068	-0.019	-0.007	-0.106	0.093	0.007	-0.021	-0.18	-0.052	0.027	-0.019	0.127	-0.109
hb	-0.087	-0.21	-0.206	1	0.255	0.334	-0.185	0.689	-0.64	-0.119	0.144	**0.92**	-0.236	-0.337	-0.413	-0.109	0.004
liver	-0.185	-0.108	-0.068	0.255	1	0.227	-0.04	0.112	-0.106	-0.066	0.135	0.225	0.407	-0.127	-0.242	0.583	0.131
lympho	-0.035	-0.164	-0.019	0.334	0.227	1	0.065	0.126	-0.056	-0.688	0.294	0.232	-0.135	0.068	0.008	0.362	-0.136
mch	-0.102	0.1	-0.007	-0.185	-0.04	0.065	1	-0.223	0.611	-0.164	0.101	-0.541	0.043	0.186	0.195	0.098	-0.165
mchc	0.096	-0.328	-0.106	0.689	0.112	0.126	-0.223	1	-0.905	0.011	-0.551	**0.71**	-0.369	-0.272	-0.419	-0.324	-0.084
mcv	-0.113	0.334	0.093	-0.64	-0.106	-0.056	0.611	-0.905	1	-0.093	0.515	**-0.81**	0.33	0.301	0.427	0.343	0.012
neutro	0.086	0.092	0.007	-0.119	-0.066	-0.688	-0.164	0.011	-0.093	1	-0.222	-0.018	0.158	-0.131	-0.034	-0.287	0.044
pcv	-0.16	0.344	-0.021	0.144	0.135	0.294	0.101	-0.551	0.515	-0.222	1	0.008	0.299	0.004	0.129	0.527	0.198
rbc	-0.033	-0.256	-0.18	0.922	0.225	0.232	-0.541	0.708	-0.814	-0.018	0.008	1	-0.246	-0.36	-0.438	-0.189	0.043
relaliver	-0.049	0.319	-0.052	-0.236	0.407	-0.135	0.043	-0.369	0.33	0.158	0.299	-0.246	1	-0.102	0.196	0.115	0.191
reticulo	-0.2	0.046	0.027	-0.337	-0.127	0.068	0.186	-0.272	0.301	-0.131	0.004	-0.36	-0.102	1	0.204	0.138	0.022
spleen	0.057	0.261	-0.019	-0.413	-0.242	0.008	0.195	-0.419	0.427	-0.034	0.129	-0.438	0.196	0.204	1	-0.02	0.148
termbw	-0.118	0.145	0.127	-0.109	0.583	0.362	0.098	-0.324	0.343	-0.287	0.527	-0.189	0.115	0.138	-0.02	1	0.188
tp	-0.161	0.695	-0.109	0.004	0.131	-0.136	-0.165	-0.084	0.012	0.044	0.198	0.043	0.191	0.022	0.148	0.188	1

Conclusions

In various studies examined so far, endpoints seem to be equally sensitive, or at least close to that

If so, a single BMD (with CI) could be derived from a study, covering all endpoints

← multivariate methods

Suppose we have a compound with four studies, resulting in the following NOAELs:

	rat	mouse
Subchr (liver effect)	100	30
Developm (foetal BW)	20	10

Conclusions :

- foetal BW more sensitive endpoint than liver effects
- mice more sensitive than rats

Correct ?

PoDs may differ due to:

- different endpoints
- different species
- different exposure durations
- different routes
- .....

- different strains, labs, diets, study conditions, etc

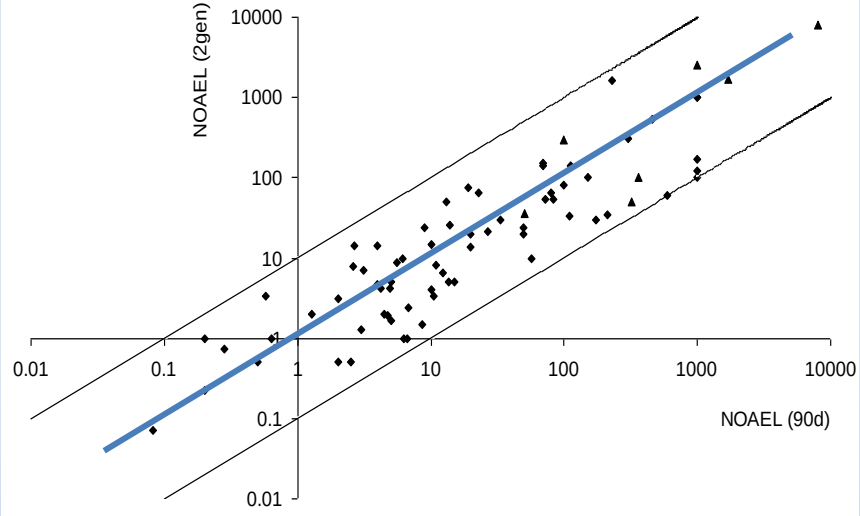
- uncertainty in the POD itself

study
replication
error

Comparing endpoints in distinct studies

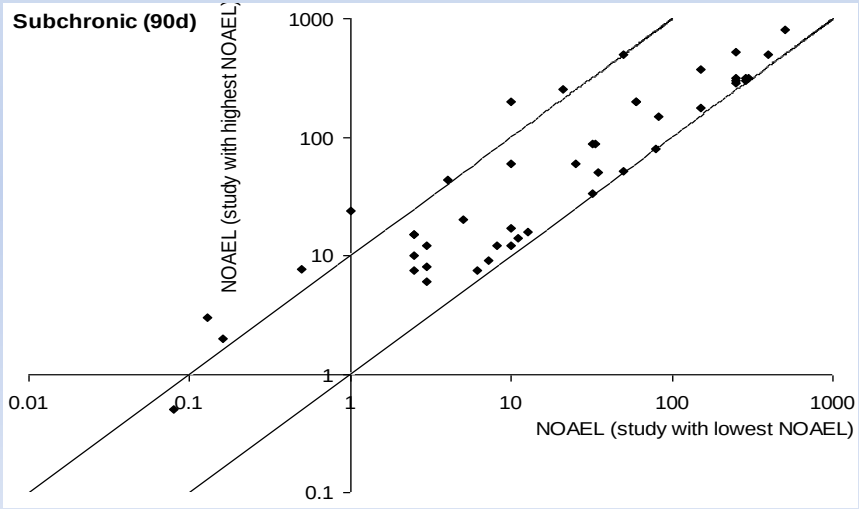
(Janer et al. 2007) compared NOAELs in 2-gen vs. subchronic studies

2-gen NOAEL vs. subchronic NOAEL



highest NOAEL vs. lowest NOAEL in replicated subchronic studies

replicated studies show a similar scatter



Conclusion

study replication error might explain the observed differences between 2-gen NOAELs and subchronic NOAELs

Do endpoints in distinct studies show similar sensitivity as well?

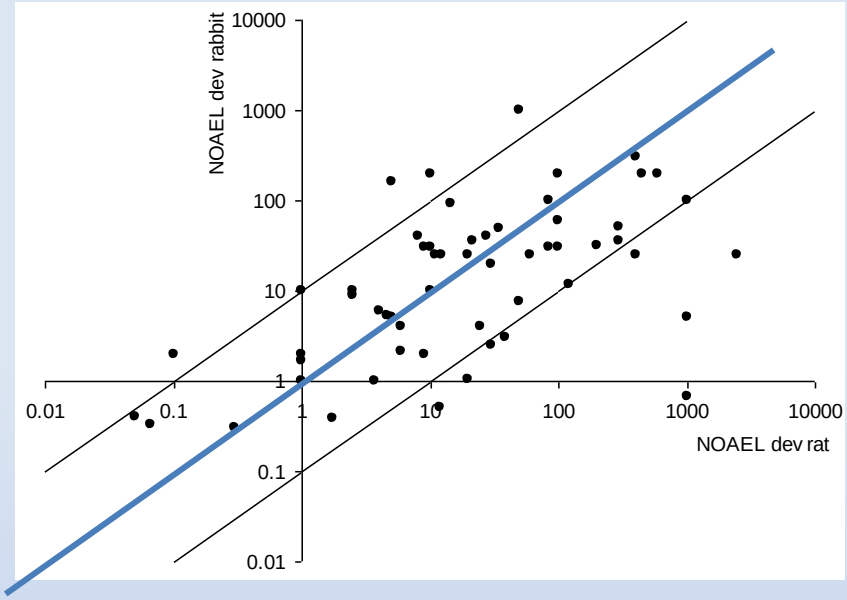
Impact of species

Various studies have shown that species are, on average over compounds, equally sensitive

Janer et al. 2008:	rat vs. rabbit in developm. studies, 54 compounds (NOAELs)
Bokkers and Slob, 2007:	rats vs. mice in 958 NTP datasets (NOAELs and BMDs) (after allometric scaling)
Braakhuis et al. (in prep) :	rat vs. rabbit in 1273 developm. studies (LOAELs) (after allometric scaling)
Bokkers (2009):	mouse, rat, rabbit, monkey, dog, human (kinetics parameters) (after allometric scaling)

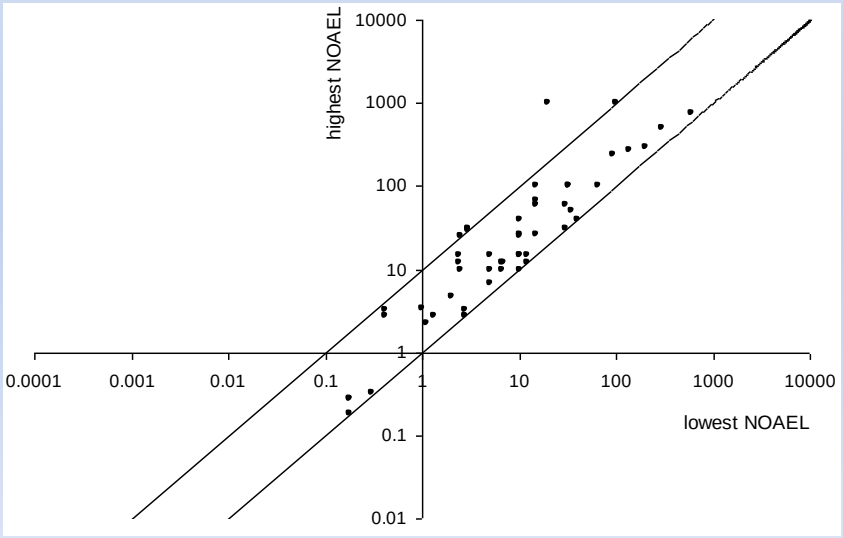
For example, developmental NOAELs in rabbit vs. rat (Janer et al. 2008):

rabbit NOAEL vs. rat NOAEL



highest NOAEL against lowest NOAEL for the same species (and co mpound)

replicated studies show a similar scatter



Using a larger database (1273 studies) Braakhuis et al. (in prep) confirmed that rat and rabbit are equally sensitive for the *individual* compounds

So, interspecies differences might not be as large as we always thought, even for individual compounds

More research on species-compound interaction is needed for other study types/effects

Before addressing the question:

What to do with multiple DR datasets ?

we must know where the differences in PODs from different studies come from

endpoints?
species?
routes?
labs?
data errors?
others?

Hypothesis 1:

All endpoints (within a study) are (more or less) equally sensitive,
and can be used for estimating one single BMD (confidence interval).

Hypothesis 2:

Interspecies differences in sensitivity are minor, and studies using different (wildtype) species can be used for estimating an average BMD.

Hypothesis 3:

Exposure duration has an impact on the BMD, but the impact is more or less the same for all chemicals. So, the ratio of BMDs for two exposure durations is a constant.

and, similarly, other hypotheses may be investigated

If these hypotheses are (approximately) true, we can handle multiple studies by simply taking the (geometric) mean of the study BMDs, and calculate a confidence interval for that mean

(taking BMD CIs into account by taking weighted mean)

(BMDs are corrected by a constant for exposure duration)

By selecting the lower bound of that confidence interval, more studies is “rewarded” by a higher value for the lower confidence bound

Simple numerical example

Four studies, with PODs: 20, 50, 200, 500 mg/kg

geometric mean: 100 mg/kg

lower confidence bound (95% confidence) : 19 mg/kg

POD for this compound

With more studies, the lower confidence bound will tend to be higher (and with fewer it will tend to be lower)