FSA project T01051
Interpretation of Margins of Exposure for Genotoxic Carcinogens

Final Report for Objective 02

Theoretical basis for how assessment factors should be combined in risk assessments

November 13, 2013

Contents

1 Introduction **2**

2 Foundational basis for the use of assessment factors **3**

3 Trajectories **5**

3.1 Multiple assessment factors 6

3.2 Multiplying assessment factors 7

3.2.1 Cooke's criticism 7

3.2.2 Other criticisms 7

3.2.3 Analysis from a probabilistic perspective 8

3.3 Being overcautious 8

4 Modelling scaling and variability — the "random chemicals" approach **9**

4.1 Uncertainty 11

4.2 Cooke on independence 12

5 Incorporating uncertainty **12**

6 A numerical exploration **13**

7 Case Study **14**

7.1 Introduction 14

7.2 Source of data, models and uncertainties 16

7.3 Results 17

8 Conclusions **20**

9 Future work **21**

References **21**

The goal of this document is work towards a more coherent rational approach to assessment factors, one which should have the potential to be practically helpful in relation to carcinogens and hopefully also to other risk assessment contexts.

At a foundational level, this report is strongly influenced by the experience of members of the project team, by interactions with Professor Michael Goldstein of Durham University, by the

writing of De Finetti (the founder of subjective Bayesianism) and by three main threads of literature on probabilistic approaches to assessment factors (also known as uncertainty or safety factors):

- Cooke (2010): article in Risk Analysis and associated discussion, giving a probabilistic perspective on a number of apparent contradictions or paradoxes arising from ways in which people have in the past combined uncertainty factors;
- articles by Kodell and Gaylor (henceforth referred to as "K&G") on combining assessment factors for non-carcinogens: references Kodell and Gaylor (1999), Gaylor and Kodell (2000), Gaylor and Kodell (2002).
- articles by Slob and various co-authors (henceforth referred to as "Slob+al"): selected references Slob and Pieters (1998), Vermeire et al (1999), Bosgra et al (2005).

K&G and Cooke draw on a lot of earlier literature [including Slob+al, Renwick and co-authors, etc].

At a more practical level, this report is influenced again by the experience of members of the project team but also by many others with whom members of the team have collaborated in a variety of areas of risk assessment.

1 Introduction

Although the motivation for undertaking this work comes from the context of carcinogens and the "margin of exposure" idea, the majority of what follows is intended to be more general. Occasional links will be made to carcinogens and other potential applications.

In many areas of risk assessment, one or more assessment factors are applied to a measurement (the point of departure) of some quantity in order to provide a number (the point of arrival) which is felt to be sufficiently conservative relative to some important (target) quantity for which more direct measurement is difficult or impossible[1].

Example 1: In aquatic ecotoxicology, the point of departure might be a laboratory measurement of the 24 hour LC_{50} for a particular test species exposed to a chemical substance and the target quantity might be the environmental concentration which would cause no meaningful disruption to real ecological communities; the point of arrival is obtained by dividing the LC_{50} by a number of assessment factors to account for inter-species variation, differences between acute and chronic tolerances, differences between laboratory and field outcomes, etc. The intention is that the point of arrival can be used as a threshold to determine if an exposure scenario is acceptable.

Example 2: For carcinogens, the point of departure might be a $BMDL_{10}$ for an experimental animal population while the target quantity might be the dose which corresponds to a human population excess lifetime cancer risk of 1 in 100,000. The point of arrival might be obtained by dividing the $BMDL_{10}$ by a total assessment factor of 10000; the result is a dose which might be considered to be acceptable in some jurisdictions, although not in the UK.

Example 3: The context explored by K&G is the derivation of safe human exposure levels for non-carcinogenic toxicants starting from an animal sub-chronic LOAEL or BMD_{10} (the point of departure) and applying 4 assessment factors: AF_H to account for uncertainty about variation in sensitivity between humans; AF_A to account for uncertainty about difference in sensitivity between humans and animals; AF_S to account for uncertainty in extrapolating from sub-chronic data to chronic exposure; and AF_L to account for uncertainty in extrapolation from a "low-risk" animal measurement (the LOAEL or BMD_{10}) to the target level of negligible risk to humans. They use the term uncertainty in each case although they subsequently use empirical information about various kinds of variability to derive values for the assessment factors. They

[1] The difficulty of measurement of the target may arise for a variety of technical, financial or ethical reasons.

use the notation U_H etc rather than AF_H; the change in notation here is to make it possible to distinguish the assessment factor itself from the uncertainty/variability it addresses.

The title of this part of the "margin of exposure" project is "Examine the theoretical basis for how assessment factors should be combined[2] in risk assessment". In reality, as indicated in the project proposal, it will also be necessary to explore the theoretical basis underlying the use of a single assessment factor in order to understand what constraints should apply to the process of combining assessment factors.

In what follows, the notation PoD will be used for the point of departure, T for the target, PoA for the point of arrival and AF for a single assessment factor. Where multiple assessment factors are involved, they will be subscripted as AF_1, AF_2 and so on.

2 Foundational basis for the use of assessment factors

The starting point of the need for assessment factors is uncertainty about the target T. It will be assumed that the uncertainty is not about the definition of the target but about its numerical value. Thus, in the context of carcinogens, it will be assumed that doubts have been resolved about such issues as whether the relevant level of excess risk is 1 in 100,000 people or 1 in 1,000,000 people or which human population should be considered; these are policy issues. The remaining uncertainty to be addressed is about the dose giving rise to the specified level of excess risk in the given population.

The target T is by definition not going to be measured directly. Instead a different number, the point of departure PoD, will be obtained or has already been obtained. The definition of the PoD should also be unambiguous. The reasons for obtaining the PoD are (a) that it is possible to do so; (b) that it is conceptually related to the target T; and (c) that experts feel able to make some quantitative judgements about the difference between PoD and T. Those quantitative judgements may be empirically founded or be based on expert synthesis of evidence or be a combination of both.

Supposing that a single assessment factor AF is being applied so that

$$\mathrm{PoA} = \mathrm{PoD}/\mathrm{AF},$$

several questions immediately arise if we want to rationalise the value chosen for AF:

1. What does AF represent here?
2. What are we trying to achieve: what relationship do the experts want between PoA and T?
3. Why are we dividing by a constant AF rather than (say) subtracting some constant?
4. How should we choose a value for AF? What considerations should we taken into account and how should we go about quantification?

In reality, the answers to these questions are not completely independent. However, taking them one by one:

3. Implicitly, these are contexts where T is positive and where 0 represents complete safety, although it may be impossible or impractical or uneconomic to achieve it. Moreover, the contexts generally deal with doses or concentrations and it tends to be natural to think about the effect of halving a dose or concentration rather than reducing it by a fixed

[2] In practice, for carcinogens, this is not to derive a "safe" dose, but rather to help interpret the level of concern associated with the margin of exposure.

amount; in particular, at lower doses and concentrations, subtracting a fixed amount would lead to the impossible situation of having negative doses and concentrations.

There is a basic scientific intuition that ratios of doses (or concentrations as appropriate) are more likely to be stable between chemicals than other possible ways of comparing two doses, whether the doses correspond to different effect levels in the same species or to the same effect level in different species or to various other differences which are covered by assessment factors.

Moreover, in the carcinogen context (and others), the linear dose-response relationship lurks in the background as a default possibility. For the linear dose-response, dividing by a fixed factor has the same relative size of effect on the response at all doses.

1. The AF represents both our ability to extrapolate from the point of departure PoD to the target T and our uncertainty about that extrapolation.

   By extrapolate, we mean the possibility that experts judge that PoD and T will have similar orders of magnitude, or that T is likely to be a couple of orders of magnitude smaller than S, without being able to make a precise prediction. In other words, the experts have some idea about the size of the ratio T/PoD but are aware that it varies from assessment to assessment and hopefully are able to say something quantitative about the variation and about their uncertainty.

2. Implicitly, we are in situations where lower values of PoA involve less harm.

   Thus, the goal is to have sufficiently high confidence that the point of arrival PoA is lower than the unknown numerical value for the conceptually well-defined target T. By confidence, at this point, it is not meant necessarily to imply quantification in terms of frequency or probability although it seems difficult to avoid some degree of such a quantification if one wants to construct assessment factors rationally.

   A key here is that the goal is one-sided in two ways. One is not looking for a precise level of confidence nor to hit the target precisely; instead one wants some minimum degree of confidence that PoA is lower than T. However, it is also the case that there is a penalty (economic or otherwise) for obtaining too low a value for the PoA; otherwise the value 0 would surely suffice from the safety viewpoint.

   A simple example of the benefit of the one-sidedness is to think about being asked by a visitor when the next bus will go past; without detailed knowledge of the timetable, most of us would find it hard to specify a precise time with any confidence but in many places could say with considerable confidence that the waiting time would be less than an hour. Moreover, one would probably be more comfortable expressing one's certainty in the prediction by giving a lower bound for one's probability than if one had to specify a precise probability for the event happening[3]

4. It seems natural that the most basic quantitative goal in determining an AF to use should be to require that

   $$\mathrm{Prob}(\mathrm{PoA} > \mathrm{T}) < \gamma \tag{1}$$

   for some probability γ which is likely to be small. Note that $\mathrm{PoA} > \mathrm{T}$ if and only if $\mathrm{PoD}/\mathrm{AF} > \mathrm{T}$ and that $\mathrm{PoD}/\mathrm{AF} > \mathrm{T}$ if and only if $\mathrm{PoD}/\mathrm{T} > \mathrm{AF}$. Consequently, $\mathrm{Prob}(\mathrm{PoA} > \mathrm{T}) = \mathrm{Prob}(\mathrm{PoD}/\mathrm{T} > \mathrm{AF})$ and so one way to understand the assessment factor AF is as a value which PoD/T is unlikely to exceed. However, this does beg many issues.

   - What kind of probabilities are these? We incline towards the subjective motivation/interpretation of probability despite some difficulties: objections on principle

[3] Some statisticians see this as a failure to adequately engage with specification of probablities but it certainly matches the instinctive attitude of many scientists and others.

from some quarters; the individual as opposed to group nature of the theory; and difficulty of elicitation/specification especially for one-off binary outcomes.

However, there may be situations in which the probability could be partially or completely frequency based which would be more comfortable for many people.

- Is γ explicit or just implicitly small? In practice, γ has rarely been explicitly specified in risk assessments using assessment factors. However, we shall see that combining assessment factors rationally almost certainly requires an explicit value for γ along with many other probabilities.
- One of the reasons for use of assessment factors is the difficulty of quantifying probabilistic arguments in these kinds of risk assessments. Moreover, the assessment factor is usually specified not for a single risk assessment but to cover many risk assessments.

Thus it seems that the assessment factor has two possible goals: one is frequency based which is to ensure that PoA $\leq$ T for most applications of the assessment factor and the second is to achieve a high probability that PoA $\leq$ T in each individual application.

Now these are in fact related goals: the latter can imply the former in the sense that if one judges that equation (1) holds for each assessment, then one also judges that PoA will only exceed T in at most proportion γ of applications. The proviso is that the assessments must be independent in a certain sense; if there is a shared component of uncertainty, the law of large numbers need not apply[4]. On the other hand, in the absence of other information specific to a particular application of the assessment factor, the frequency property leads many to conclude that (1) holds as a judgement for that application. Again, the difficulty is that one can only mathematically derive a frequency property for a statistical procedure if certain independence assumptions hold.

So what does this all boil down to? Fundamentally, we are suggesting that experts make limited probability judgements about

$$\frac{\text{PoA}}{\text{T}} = \frac{\text{PoD}/\text{AF}}{\text{T}} = \frac{\text{PoD}}{\text{T}} \Big/ \text{AF} = \text{U}/\text{AF} \tag{2}$$

where $U = \text{PoD}/\text{T}$. To be precise, the experts need low probability that the ratio in (2) exceeds 1. Since PoA $>$ T if and only if PoD$/$T $>$ AF, (1) becomes

$$\text{Prob}(\text{PoD}/\text{T} > \text{AF}) = \text{Prob}(\text{U} > \text{AF}) < \gamma \tag{3}$$

It seems likely that, informally, this is pretty much exactly what is going on. The experts are of course thinking about how large the ratio PoD$/$T might be and are choosing a larger value for AF to compensate. In fact this is obvious. So why labour the point in this way? Because rational combination of uncertainty factors requires sophisticated reasoning about uncertainty from the experts and experience from a variety of contexts shows that that in general this is hard and may well not be done to a high standard.

3 Trajectories

Now let us start to think about situations where more than one assessment factor ends up being applied.

We shall assume that, conceptually, the link between PoD and T is a "trajectory" from PoD to T via a number of "way-points" $\text{W}_0, \ldots, \text{W}_m$:

$$\text{PoD} = \text{W}_0 \leftrightarrow \text{W}_1 \leftrightarrow \text{W}_2 \leftrightarrow \cdots \leftrightarrow \text{W}_m = \text{T} \tag{4}$$

[4]Both mathematically and in practice, the issue of what constitutes sufficient independence is important and needs more consideration at some point in the future.

and where the advantage of introducing the way-points is that (i) the conceptual links are stronger between adjacent pairs than across longer spans and/or (ii) it's easier to think about quantifying each W_{i-1}/W_i than directly quantifying PoD/T. Note that m is the number of steps in the trajectory and that m will vary from context to context.

We already have a trajectory in our carcinogen "example" by taking PoD $=$ W_0 to be the $BMDL_{10}$ and W_1 to be the BMD_{10} for the experimental animal population. This leads to two areas for potential quantification: (i) the animal experiment which links the measurement PoD $=$ $W_0 = BMDL_{10}$ and population benchmark dose $W_1 = BMD_{10}$ for animals; (ii) the relationship between W_1, the animal BMD_{10}, and W_2 which is T, the target dose for humans. The former has been well-studied and the procedure for quantification is essentially determined by the statistical model and the experimental design.

However, we could further insert W_2 to be the BMD_{10} for the human population and have W_3 be the usual target T. If this helps, what was previously the second area of quantification would be subdivided into quantifying the difference between human and animal BMD_{10}s and the difference for humans between BMD_{10} and the target dose. We do not wish to imply here that this is a sensible approach to the carcinogen context; the purpose is only to illustrate the concept of "way-point". This approach was considered in more detail during the expert elicitation exercise conducted as part of Sub-task 03/02 of the project.

In the K&G example, we have $m = 4$ steps in the trajectory. Depending on the order in which the assessment factors are applied, one might define the way-points differently; we will consider just one possible version where we apply them in the order AF_S, AF_A, AF_H, AF_L. The PoD $= W_0$ is the animal sub-chronic LOAEL or BMD_{10}; W_1 is the animal chronic LOAEL or BMD_{10}; W_2 is the typical human chronic LOAEL or BMD_{10}; W_3 is the sensitive human chronic LOAEL or BMD_{10}; and T $= W_4$ is the sensitive human "negligible-risk" dose.

The purpose of the diagram in (4) is to emphasise the idea that each step along the "trajectory" has to be quantified in some way. Please note that it is not being suggested that successive steps are necessarily independent although experts might make such a judgement in any particular context. The purpose of the way-points is really to break the journey from point of departure to target into manageable pieces.

3.1 Multiple assessment factors

Once we have trajectories, we have the possibility of introducing more than one assessment factor where each assessment factor covers one or more steps in the trajectory. The idea is that the whole trajectory is covered by assessment factors and that assessment factors shouldn't overlap.

So when faced with a problem with two intermediate way-points:

$$\text{PoD} = W_0 \leftrightarrow W_1 \leftrightarrow W_2 \leftrightarrow W_3 = \text{T}$$

we could have

- a single assessment factor taking us from PoD to T; or
- an assessment factor taking us from PoD to W_1 and a second from W_1 to T; or
- an assessment factor taking us from PoD to W_2 and a second from W_2 to T; or
- an assessment factor taking us from PoD to W_1, and a second from W_2 to W_2 and a third from W_2 to T

The appropriate way to approach this depends on our understanding/knowledge of the steps and how they relate.

An issue which might arise is that there could exist more than one plausible trajectory: for the carcinogen "example", we suggested the possibility of taking W_2 to be the human BMD_{10}; an alternative would be to take W_2 to be the dose causing an excess 1 in 100,000 lifetime risk in the animal population. We are not asserting that either proposal makes any sense nor that one is better than the other; the intention is simply to illustrate possibilities.

3.2 Multiplying assessment factors

The natural reaction to a situation with more than one step in the trajectory and where assessment factors have been provided for individual steps is to multiply the assessment factors. This is obviously motivated by

$$\mathrm{U} = \frac{\mathrm{PoD}}{\mathrm{T}} = \frac{\mathrm{W}_0}{\mathrm{W}_m} = \frac{\mathrm{W}_0}{\mathrm{W}_1}\frac{\mathrm{W}_1}{\mathrm{W}_2}\cdots\frac{\mathrm{W}_{m-1}}{\mathrm{W}_m} = \mathrm{U}_1 \times \mathrm{U}_2 \times \cdots \times \mathrm{U}_m$$

where we define $\mathrm{U}_i = \mathrm{W}_{i-1}/\mathrm{W}_i$. If each AF_i is large relative to the corresponding ratio U_i it seems intuitively obvious that the product $\mathrm{AF}_1 \times \cdots \times \mathrm{AF}_m$ would be large compared to the product of the individual ratios $\mathrm{U} = \mathrm{PoD}/\mathrm{T}$. Consequently it would seem sensible to take $\mathrm{AF} = \mathrm{AF}_1 \times \cdots \times \mathrm{AF}_m$ in calculating $\mathrm{PoA} = \mathrm{PoD}/\mathrm{AF}$.

Although it is intuitive to multiply assessment factors in this way, the detailed consequences are unclear. Over the last 15 years, several authors have argued that a more detailed rationale is needed and some have provided alternative calculations which do, however, require additional information to be obtained/provided.

3.2.1 Cooke's criticism

Cooke (2010) argues that, in any situation where there is more than one assessment factor being applied and where the values of the assessment factors are fixed, very strong constraints are being imposed in relation to the steps in the trajectory. For example, if using three fixed size assessment factors in the hypothetical two intermediate way-point situation considered, he argues that there is an implied judgement that $U_1 = \mathrm{W}_0/\mathrm{W}_1$, $U_2 = \mathrm{W}_1/\mathrm{W}_2$ and $U_3 = \mathrm{W}_2/\mathrm{W}_3$ are judged to be completely dependent whereas most of us would have the instinctive feeling that the reason for introducing three steps would be a judgement of independence or near-independence of the three ratios.

Cooke's is a quite sophisticated probabilistic argument; it is mathematically valid given a number of assumptions. Slob (2011) has strongly questioned those assumptions in his comment on Cooke's article but it is not clear that Slob's objections are entirely well-founded. However, Cooke's argument ignores two key features of the use of assessment factors. The first is the one-sidedness discussed earlier which means that we are often trying just to get a "safe enough" outcome rather than to get a "just safe enough" outcome. The second (related) feature is that assessment factors are usually round numbers such as 100, 10, 5, 3 and 2 rather than being specified to high precision. Both features act to limit the strength of the conclusions one can draw about experts' judgements on the basis of how they combine uncertainty factors. Nevertheless, Cooke's argument shows considerable insight and indicates the potential for serious difficulties with the practice of multiplying assessment factors.

3.2.2 Other criticisms

Both K&G and Slob+al[5] have also criticised the multiplication of standard assessment factors as being insufficiently based on evidence and likely to lead to excessively conservative combined assessment factors. Further discussion of their proposed solutions is deferred to the later "random chemicals" section.

[5]There are probably other papers in the literature

3.2.3 Analysis from a probabilistic perspective

The major limitation is that the kind of quantitative judgement expressed in (1) may not be strong enough to be useful when combining assessment factors.

Suppose we are in a situation with a single way-point and two assessment factors AF_1 and AF_2 which have been arrived at separately and which we now wish to combine. The equivalents of (1) to be satisfied by the individual assessment factors are

$$\text{Prob}(\text{PoD}/W_1 > AF_1) \equiv \text{Prob}(U_1 > AF_1) < \gamma_1 \tag{5}$$

and

$$\text{Prob}(W_1/T > AF_2) \equiv \text{Prob}(U_2 > AF_2) < \gamma_2 \tag{6}$$

Mathematically the only conclusion we can now draw involving the product of the assessment factors is that

$$\text{Prob}(\text{PoD}/T > AF_1AF_2) \equiv \text{Prob}(U_1U_2 > AF_1AF_2) < \gamma_1 + \gamma_2 \tag{7}$$

From a precautionary perspective, this is of course fine if the probabilities γ_1 and γ_2 used to derive the individual assessment factors were small enough that $\gamma = \gamma_1 + \gamma_2$ is small enough for our overall goal in (3). Then it is reasonable to take overall assessment factor AF in (3) to be $AF = AF_1AF_2$. However, we shall see in the next section that there are grounds for considering it to be excessively conservative in many situations.

Why is (7) the only conclusion we can draw from (5) and (6)? At one extreme, we could have U_2 just at or below AF_2 whenever $U_1 > AF_1$ and vice versa; this leads to $\text{PoD}/T > AF_1AF_2$ whenever either of the individual ratios exceeds its corresponding assessment factor and yet the individual ratios cannot simultaneously exceed their assessment factors. At the other extreme, it could be that U_2 is small whenever $U_1 > AF_1$ and vice versa; then the probability in (7) would actually be zero.

Note that in both extreme situations, U_1 and U_2 are far from independent. However, adding an assumption of independence of the ratios does not help by itself; we need more information about the ratios than independence if we are to make more precise probability statements about their product.

3.3 Being overcautious

The analysis in section 3.2.3 conceals a more serious issue which is that the procedure obtained by multiplying assessment factors may actually be much more cautious than the individual assessment factors appear to be. Moreover, this is not just a possible issue but is one which is likely often to arise.

Suppose that instead of simply choosing the individual assessment factors to be certainly large enough to satisfy (5) and (6), the individual factors were actually chosen to be calibrated in the sense that

$$\text{Prob}(U_1 > AF_1) \preceq \gamma_1 \tag{8}$$

and

$$\text{Prob}(U_2 > AF_2) \preceq \gamma_2 \tag{9}$$

where by $\preceq$ we mean that each probability is less than or equal to γ_i but not very much smaller than γ_i. This can be seen as retaining the key one-sidedness property but informally restricting the degree of one-sidedness. The intention is to exclude situations where an assessment factor is being set very much larger than would actually be needed for most chemicals. Such assessment factors might be described as "approximately calibrated".

Then about the only thing we can deduce mathematically is that

$$\text{Prob}(\text{PoD}/\text{T} > \text{AF}_1\text{AF}_2) = \text{Prob}(\text{U}_1\text{U}_2 > \text{AF}_1\text{AF}_2) \leq \gamma_1 + \gamma_2 \tag{10}$$

Note that the symbol in (10) is $\leq$ not $\preceq$. In other words, we retain the one-sidedness but we lose the restriction on one-sidedness that protects us from being excessively conservative; the probability might actually be zero or very nearly so. This could be a problem if our reason for approximately calibrating the original assessment factors was to avoid a harm that comes from being over-cautious in the risk assessment.

Interestingly, it does not help very much to assume/believe/know that U_1 and U_2 are independent; it simply replaces $\gamma_1 + \gamma_2$ by $\gamma_1 + \gamma_2 - \gamma_1\gamma_2$ which is likely to be very similar to $\gamma_1 + \gamma_2$ since each γ_i will be near 0. The lower bound is not precise due to the $\preceq$ for the individual assessment factors but is of order of magnitude $\gamma_1\gamma_2$. The reason independence doesn't help is that the probability in (10) still depends on the detail of the probability distributions for U_1 and U_2. Any fixed positive value of U_2 can lead to $\text{U}_1\text{U}_2 > \text{AF}_1\text{AF}_2$ but the chance of this happening depends on the distribution of U_1 for every value of U_2 and then the overall chance also depends on the distribution of U_2.

If we actually know the distributions of U_1 and U_2 and that U_1 and U_2 are independent, we can then compute $\text{Prob}(\text{U}_1\text{U}_2 > \text{AF}_1\text{AF}_2)$; however, in that situation we can do better by simply calculating the distribution of $\text{U} = \text{U}_1\text{U}_2 = \text{PoD}/\text{T}$ and then finding a single assessment factor AF satisfying

$$\text{Prob}(\text{U} > \text{AF}) = \gamma \tag{11}$$

for our chosen value of γ. This is discussed further in the next section.

At this point, we have two extreme possible ways of proceeding: (i) to multiply assessment factors which is guaranteed to be semi-conservative and may well be over-conservative; or (ii) to specify independent probability distributions for the ratios U_1 and U_2 in order to obtain an overall assessment factor offering the desired level of protection. Is there any possibility in between? One limited possibility is to build in correlations when using log-normal distributions for individual ratios. There is certainly no immediately available more general technology; it might be possible to build one based on imprecise probability, where exact distributions would not need to be specified, but it is not clear at this time that the resulting methods would be any easier to apply or would yield more reliable results.

4 Modelling scaling and variability — the "random chemicals" approach

Cooke uses the term "random chemicals" for what he describes as the methodological approach advocated by K&G, Slob+al and many others. In reality, these authors' approaches differ in quite fundamental ways. K&G use distributions based on data analysis to describe only variability whereas Slob+al are happy for distributions also to represent uncertainty, whether it be formally statistical or a representation of expert judgements or even just a "what-if" representation which combines both variability and uncertainty. This section is confined to consideration of distributions representing variability.

The core of the K&G method is that there may be existing data on W_{i-1} and W_i for chemicals considered relevant to the chemical(s) for which the assessment factor is being developed. This data may then be used to infer a specific distribution for U_i, often some log-normal distribution. The term "random chemicals" arises because the chemical(s) for which the assessment factor is being developed are then assumed to be sampled from a population of chemicals and the chemicals for which there was data are assumed to have been sampled from the same population. The sampling need not be (and almost certainly isn't) random; a more realistic requirement which yields the same mathematics is that the experts judge potential toxicity measurements for the new chemicals to be exchangeable with the ones for which data has been obtained.

If one additionally assumes that $\mathrm{U}_1, \mathrm{U}_2, \ldots$ are independent, one can then calculate the distribution of their product. This is particularly straightforward in the case where each U_i has a log-normal distribution, i.e. $\log \mathrm{U}_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$. Then $\log \mathrm{U} \sim \mathrm{N}(\sum \mu_i, \sum \sigma_i^2)$. However, if the individual ratios have specific distributions other than log-normal, there is no great difficulty in numerically computing the distribution of U and software support exists for this operation, for example in the `distr` add-on package (Ruckdeschel et al , 2006) for the free statistical software R (R Development Core Team , 2011). Departures from independence are more difficult to handle and, as suggested earlier, the most straightforward approach would be to assume multivariate log-normality as is the case for some examples in the numerical exploration which comes later.

In the log-normal case, we can give clear and distinct meanings to the parameters μ_i and σ_i of the distribution for U_i. μ_i is both the median and the mean of the distribution for $\log \mathrm{U}_i = \log \mathrm{W}_{i-1} - \log \mathrm{W}_i$ and therefore $\tilde{\mu}_u = \exp(\mu_i)$ is both the median and the geometric mean of U_i. The quantity $\tilde{\mu}_i$ will be referred to as the "scaling"[6] involved in the step along the trajectory from W_{i-1} to W_i; it is the part of the step which is predictable. The term "scaling" will continue to be used, along with the notation $\tilde{\mu}_i$ to refer to the geometric mean of U_i, when other distributions are used.

σ_i then represents the amount of variability in the same step in the trajectory. This separation of scaling from the amount of variability generalises to other distributions by defining variability as the distribution of $\tilde{\mathrm{U}}_i = \mathrm{U}_i / \tilde{\mu}_i$. The separation is important as we shall see in the later numerical exploration that relative contributions of scaling and variability to the assessment factor seem to have a particularly large effect on what assessment factor we derive when we combine the distributions of individual ratios in order to determine the distribution of $\mathrm{U} = \mathrm{U}_1 \times \cdots \times \mathrm{U}_m$. Note that

$$\mathrm{U} = \mathrm{U}_1 \ldots \mathrm{U}_m = (\tilde{\mu}_1 \tilde{\mathrm{U}}_1) \ldots (\tilde{\mu}_m \tilde{\mathrm{U}}_m) = (\tilde{\mu}_1 \ldots \tilde{\mu}_m)(\tilde{\mathrm{U}}_1 \ldots \tilde{\mathrm{U}}_m) = \tilde{\mu}\tilde{\mathrm{U}}$$

which separates the overall ratio U into a single combined scaling $\tilde{\mu}$, which is the geometric mean of U, and a single combined variability component $\tilde{\mathrm{U}}$.

As indicated earlier, having computed the distribution of U, the assessment factor is then obtained by replacing the inequality in (3) by equality to obtain

$$\mathrm{Prob}(\mathrm{U} > \mathrm{AF}) = \gamma \tag{12}$$

and then solving to find the value of AF given a value of γ to be specified by the experts. If plenty of uncontroversial good quality relevant data are available, much of the effort in specifying assessment factors is reduced to specification of an appropriate value of γ. However, in practice, there is a paucity of such data.

Another way of understanding (12) is that AF is the $100(1-\gamma)$th percentile of the distribution of U. In general this means that AF is the product of $\tilde{\mu}$ and the $100(1-\gamma)$th percentile of the distribution of $\tilde{U}$. In the case where each U_i is log-normal, U is log-normal as described earlier and we compute the overall assessment factor as

$$\mathrm{AF} = \tilde{\mu} \exp[\sigma \Phi^{-1}(1-\gamma)]$$

where $\tilde{\mu} = \tilde{\mu}_1 \ldots \tilde{\mu}_m$, $\sigma = \sqrt{\sigma_1^2 + \cdots + \sigma_m^2}$ and $\Phi(.)$ is the standard normal cumulative distribution function.

Kodell and Gaylor (1999) apply this idea to the example described in the introduction. They use earlier studies to infer a distribution for each of U_H, U_S, U_L and U_A and then compare the standard products of powers of 10 to the assessment factors which would be needed to

[6] This term is not ideal as it does not immediately suggest the capacity to predict that one way-point is smaller than the other, although that is what it really means. However, a better alternative has not been found.

achieve various particular value of γ in (12). They also consider individual steps in the trajectory, replacing the tight inequalities in (8) and (9) by equalities so that we require

$$\text{Prob}(\text{U}_i > \text{AF}_i) = \gamma_i$$

and compare the individual powers of 10 to the individual assessment factors which would be required for various values of γ_i. They conclude that the individual assessment factors are probably less protective than intended but that the result of multiplying them is conservatism, possibly rather extremely so. Gaylor and Kodell (2000) take the matter further and compute the value of γ corresponding to each of the standard powers of 10 and also make comparison to earlier Monte Carlo studies reported by Baird et al and Swartout et al.

K&G's assumption of log-normality greatly simplifies the numerical calculation of the overall assessment factor relative to the use of other distributions but this is not crucial. The crucial components of the process are the specification of γ and the determination, without uncertainty, of independent distributions for the ratios. K&G derive those distributions reasonably explicitly from data but some or all could simply be specified by experts, an approach taken by Slob+al. Experts may however find it difficult to provide precise values for μ_i and σ_i; this is clearly a more demanding task than the informal specification of an individual assessment factor to satisfy some version of $\text{Prob}(\text{U}_i > \text{AF}_i) < \gamma_i$.

The consequences of unjustified log-normality and independence assumptions may be to compute overall assessment factors in which one has little trust. This point is implicit in the final sentence of Gaylor and Kodell (2002) which calls for an investigation of sensitivity to the assumption of log-normality and to values of parameters used.

4.1 Uncertainty

Kodell and Gaylor (1999) briefly discuss the issue of uncertainty about the parameters of the log-normal distribution. They consider the possibility that some allowance should be made for uncertainty but do not see a good way to proceed taking into account uncertainty for more than one ratio. They suggest instead that a smaller value of γ should be used in determining the assessment factor but provide no rationale for how to choose the smaller value. They consider neither the possibility that the log-normal model is inadequate nor consequences.

Various authors have found log-normal distributions to be an adequate approximation to the distribution of various ratios and have provided parameter estimates and derived individual assessment factors for those ratios. But consideration of the consequences of model uncertainty and parameter uncertainty has been neglected. One difficulty in relation to these studies is that the log-normal fit has been assessed using data-sets of moderate size, large enough to make some assessment of normality in the centre of the distribution but not large enough to guarantee normality in the tails; of course, it isn't possible to guarantee normality in the extreme tails but we should at least try to assess sensitivity of the assessment factor calculation to the choice of distribution family.

Assessment of dependence between different ratios is more difficult as there is less data available. Calabrese and Gilbert (1993) demonstrate a dependence between two ratios but this is actually deduced from the fact that the two ratios share a segment of the trajectory and is not obtained from analysis of data. There may be empirical studies of this issue but we are not presently aware of them.

In the absence of such studies, independence is an appealing starting point but should not be taken as given. Consequently, establishing the sensitivity of AF to dependence assumptions is important.

4.2 Cooke on independence

One of the assertions made by Cooke is that independence is unlikely to be a reasonable assumption; the denominator of U_i is the numerator of U_{i-1} and he argues that this alone makes it unlikely that U_i and U_{i-1} can be independent. This argument is not totally convincing. There are two different aspects to the problem: one is whether independence is conceptually incorrect and the second is whether it is incorrect when data are involved.

The conceptual issue really depends on fundamental judgements about the relationships between things. If W_i is a way-point between W_i and W_{i+1} but one would judge that W_{i-1} provides information about W_{i+1} which is additional to that provided by W_i, then independence of U_i and U_{i+1} is not an appropriate assumption. If one judges that W_{i-1} does not provide additional information, independence of U_i and U_{i+1} may well be reasonable. Note that the language here is deliberately of judgements rather than of inference based on data as it is relatively unlikely that there will be many chemicals for which the three way-points involved will all have been measured.

The issue about independence when data are involved is that the data are often not directly about U_i (or U_{i+1}) but actually involve the empirical relationship between some measurement of W_i and (some measurement of) W_{i-1}. Then the same measurement(s) of W_i may be used in assessing the distributions of U_i and U_{i+1}; if so, the measurement variability is common to the information about U_i and U_{i+1} and should induce a negative correlation between them although this correlation may be small. Essentially, in the trajectory, the conceptual quantity W_i has been replaced by a measurement $\hat{W}_i$ of it and this introduces some overlap between the step from W_{i-1} to $\hat{W}_i$ and the step from $\hat{W}_i$ to W_{i+1}.

In the later part of his paper, Cooke essentially argues that the only way forward is some form of multivariate modelling of the target, way-points and point of departure; he suggests possible use of Bayesian belief networks but does not offer a concrete example. This is certainly theoretically correct but may not actually differ from what is described here. Useful BBNs tend to involve substantial numbers of (conditional) independence assumptions in order to simplify learning from data and other computations.

5 Incorporating uncertainty

Although K&G avoid dealing with the issues, there will surely be uncertainty about both the log-normal model and specific parameter values for the model. The approach taken by Slob+al is just to combine scaling and variability of U_i with associated uncertainty to make a single distribution for U_i and then use that distribution in the calculation of the distribution of U. Mathematically, this requires approaching uncertainty from a Bayesian perspective although this is not explicit in their work.

One could of course object to the Slob+al approach on the grounds that it would be better to separate uncertainty from variability so that one would actually deal with uncertainty at the end of the process by arriving at either a distribution for AF given a choice of γ or a distribution for γ given a particular value for the AF; in either case, the distribution would purely represent uncertainty. However, this is only really necessary if we want to separate uncertainty from variability. Provided the ratios $U_1, \ldots, U_m$ have independent variability, and uncertainty about each ratio's variability is also independent of other ratios, mathematically we obtain the same overall distribution for U, combining uncertainty and variability, regardless of whether we incorporate uncertainty into each individual U_i distribution and then combine them or we first combine variability distributions and then incorporate uncertainty.

What we should not do is derive a separate assessment factor for each U_i allowing for uncertainty and then simply multiply them.

In principle, incorporating parameter uncertainty for the log-normal distribution is relatively

		Number of ratios m				
		1	2	3	4	5
Scaling	1	10	26	54	100	170
	2	10	39	130	400	1200
	5	10	67	420	2500	15000
	9	10	94	870	8100	75000

Table 1: Assessment factors (2 significant figures) for independent identically log-normally distributed ratios.

easy if the distribution is inferred from data. Unless the sample size is very small, using the standard Jeffreys independent prior will not be very influential or contentious and the conjugate update means that the predictive distribution for U_i will be a re-located and re-scaled t-distribution.

6 A numerical exploration

Studying the practical implications of all this is scheduled for later in the project. The numerical study here will be confined to doing some illustrative calculations of how the final assessment factor depends on the number of steps m in the trajectory, the distributions being used and the value of γ chosen.

The examples directly explore sensitivity of the assessment factor to distributional choices and γ and how this depends on m and also indirectly explore a little of the uncertainty issue. Six different scenarios are considered for the type of distribution used for the $\log \mathrm{U}_i$: normal, t (10 degrees of freedom), t (5 df), t (2df), skew-normal (skew=0.8) and skew-normal (skew=-0.8). This means, for example, that in the first scenario the distribution of U_i is log-normal.

Each scenario consists of a number of examples varying: (a) the number of ratios m; (b) the value of γ; and (c) the scaling $\tilde{\mu}_i$ used for each U_i. Each example has the same basic structure: within the example, the same distribution is used for each U_i and that distribution is obtained by taking the basic log-ratio distribution and changing its location and scale to: (i) give the required scaling for U_i; and (ii) make the $100(1-\gamma)$th percentile of U_i be 10. The distribution of $\mathrm{U} = \mathrm{U}_1 \times \cdots \times \mathrm{U}_m$ is then computed numerically and used to find the $100(1-\gamma)$th percentile of U which is the required AF assuming that the same γ is of interest.

The purpose of (ii) is that 10 (the most common standard assessment factor for a single source of uncertainty) is then the AF for all single-step ($m = 1$) trajectories in all examples for all scenarios. The purpose of (i) is to enable us to explore how changing the balance between scaling and variability affects the AF which would be required for a multi-step trajectory. The scalings considered are 1, 2, 5 and 9; the first is a situation where there is no predictable scaling and there is just variability while the last is a situation where most of each AF addresses scaling rather than variability.

The t-distribution scenarios provide some exploration of the consequences of both some excess kurtosis in the variability and some uncertainty about parameters for log-normal variability.

Table 1 shows the outcomes (values of AF) for the log-normal scenario. Note that γ does not appear; it is one of the many unusual special features of the normal distribution that the ratio of the pth percentile for a convolution of m standard normals to the pth percentile of a single standard normal does not depend on p. Consequently, the AF does not depend on γ, provided that the same γ is used when determining the distribution used for each ratio. When there is no scaling, we see that the the AF grows much more slowly than the 10, 100, 1000, 10000, 100000 that would be obtained by multiplying individual assessment factors. On the other hand, when most of each single assessment factor addresses scaling, the final AF is not much smaller than the product of the individual assessment factors.

Table 2 shows the outcomes for all 6 distributional scenarios. Entries for $m = 1$ have been omitted as they are all equal to 10. Again the effect of scaling is striking but we can also see substantial differences between distributions, with high kurtosis (5 degrees of freedom t-distribution) and negative skew at opposite extremes. The differences are more noticeable when there is less scaling. The results might seem counter-intuitive[7]. What one must remember is that the distribution for an individual ratio is set up to have 10 as the required percentile. The effect of multiplying ratios is to add log-ratios which means that the log-ratio distribution moves towards normality as m increases. Consequently, the right tail gets longer for the negatively skewed log-ratio distribution, which leads to a higher AF than for the normal log-ratio case, while it shortens for the positively skewed distribution. Similarly, both tails shorten for the excess-kurtosis cases. We can see that the percentile chosen does have an effect and that the nature of that effect depends on the distribution scenario but in general is larger when there is more variability and less scaling.

The effects of adding correlation to the scenario with log-normal ratios are also investigated. The individual ratio distributions are as in the previous study but correlation is introduced between each adjacent pair of log-ratios: $\log \mathrm{U}_{i-1}$ and $\log \mathrm{U}_i$. There is no correlation between $\log \mathrm{U}_i$ and $\log \mathrm{U}_j$ for $|j - i| > 1$. Table 3 shows results for 5 different correlation scenarios ranging from -0.5 (the minimum possible) to 0.5 (the maximum possible). As in the original log-normal scenario, the AF does not depend on the value of γ chosen. One sees that correlation can have striking effects, especially when there is less scaling. When the correlation is -0.5, effectively there are m scaling components but only one component of variability regardless of m.

7 Case Study

7.1 Introduction

We now give closer attention to the non-carcinogen setting discussed by Gaylor and Kodell (2000). They consider moving from the point of departure which is an animal sub-chronic LOAEL to the point of arrival which is a reference dose (RfD). Adapting their notation slightly, their fundamental scheme is

$$\text{RfD} = \frac{\text{LOAEL}}{\text{AF}_A \times \text{AF}_H \times \text{AF}_S \times \text{AF}_L \times \text{AF}_D \times M}$$

where: AF_A, AF_H, AF_S, AF_L are assessment factors addressing uncertainty about corresponding ratios U_A, U_H, U_S, and U_L; AF_D is an assessment factor covering uncertainty due to an inadequate database; and M is a "modifying factor" to cover any additional uncertainty. Gaylor and Kodell (2000) are not completely precise about the definitions of the ratios, stating: AF_A represents a dose reduction factor due to the uncertainty of using animal data for human effects, AF_H represents the variable sensitivities in a human population, AF_S represents the uncertainty of predicting chronic effects from sub-chronic studies and AF_L represents the ratio of the LOAEL to the NOAEL when a LOAEL is used instead of NOAEL. They consider only the first four assessment factors in the rest of their article and we shall do the same.

A possible trajectory from PoD, being the measured animal sub-chronic LOAEL, to our chosen point of arrival, the chronic NOAEL for a sensitive human, is the following sequence of way-points: animal sub-chronic NOAEL, animal chronic NOAEL, chronic NOAEL for typical human, chronic NOAEL for sensitive human. There are clearly many other possible choices but fundamentally we have to make a step from LOAEL to NOAEL, a step from sub-chronic to chronic, a step from animal to human and a step from typical human to sensitive human. For the trajectory given:

(i) U_L is the ratio of the measured sub-chronic LOAEL (the PoD) to the first way-point: the unknown sub-chronic NOAEL for the same chemical in the same animal species;

[7] The behaviour for t (2df) relative to t(5df) really is counter-intuitive; something strange happens for df< 5.

Number of ratios (m):				2				3				4				5			
Percentile:				90%	95%	99%	99.9%	90%	95%	99%	99.9%	90%	95%	99%	99.9%	90%	95%	99%	99.9%
Scaling	1	Distribution	normal	26	26	26	26	54	54	54	54	100	100	100	100	170	170	170	170
			t (10 df)	28	27	24	20	60	56	47	36	120	110	83	58	200	180	140	89
			t (5 df)	30	28	23	18	69	61	44	29	140	120	76	42	260	210	120	58
			t (2 df)	41	37	30	26	130	100	70	53	350	260	140	93	880	570	270	160
			skew +	28	26	24	23	65	55	48	44	140	110	87	77	280	200	150	130
			skew -	26	29	33	34	47	61	80	91	73	110	160	200	100	170	300	410
	2	Distribution	normal	39	39	39	39	130	130	130	130	400	400	400	400	1200	1200	1200	1200
			t (10 df)	41	40	37	33	140	130	120	98	440	410	350	270	1300	1200	1000	740
			t (5 df)	43	41	36	31	150	140	110	84	510	450	330	220	1600	1300	920	550
			t (2 df)	54	50	43	39	240	210	160	130	970	770	510	380	3700	2700	1600	1100
			skew +	41	39	37	36	150	130	120	110	500	430	360	330	1600	1300	1100	950
			skew -	39	42	46	48	120	140	170	190	320	420	560	660	810	1100	1700	2200
	4	Distribution	normal	67	67	67	67	420	420	420	420	2500	2500	2500	2500	15000	15000	15000	15000
			t (10 df)	68	67	65	62	430	420	400	370	2600	2500	2400	2100	15000	15000	14000	12000
			t (5 df)	70	68	64	60	450	430	390	340	2800	2600	2300	1900	17000	16000	13000	11000
			t (2 df)	76	74	69	67	540	510	450	410	3700	3300	2800	2400	24000	21000	17000	14000
			skew +	68	66	65	64	440	420	400	390	2800	2600	2400	2300	17000	15000	14000	13000
			skew -	67	69	72	73	400	430	470	490	2300	2500	2900	3100	13000	15000	17000	19000
	9	Distribution	normal	94	94	94	94	870	870	870	870	8100	8100	8100	8100	75000	75000	75000	75000
			t (10 df)	94	94	94	93	880	880	870	860	8200	8100	8000	7900	75000	75000	74000	73000
			t (5 df)	95	94	94	93	880	880	870	850	8200	8200	8000	7800	76000	75000	74000	71000
			t (2 df)	96	96	95	94	910	900	890	870	8600	8500	8200	8100	81000	79000	76000	74000
			skew +	94	94	94	94	880	880	870	870	8200	8100	8000	8000	76000	75000	74000	74000
			skew -	94	95	95	95	870	880	890	900	8000	8100	8300	8400	73000	75000	77000	78000

Table 2: Assessment factors (2 significant figures) assuming independent identically distributed ratios, for various types of log-ratio distribution. Percentile means $100(1-\gamma)$.

Number of ratios:				2	3	4	5
Scaling	1	Correlation	$\rho = -0.5$	10	10	10	10
			$\rho = -0.2$	18	30	47	70
			$\rho = 0$	26	54	100	170
			$\rho = 0.2$	35	89	190	370
			$\rho = 0.5$	54	170	440	1000
	1	Correlation	$\rho = -0.5$	20	40	80	160
			$\rho = -0.2$	31	87	240	620
			$\rho = 0$	39	130	400	1200
			$\rho = 0.2$	48	180	630	2000
			$\rho = 0.5$	65	290	1100	4000
	5	Correlation	$\rho = -0.5$	50	250	1200	6200
			$\rho = -0.2$	60	350	2000	11000
			$\rho = 0$	67	420	2500	15000
			$\rho = 0.2$	73	480	3000	19000
			$\rho = 0.5$	83	590	3900	25000
	9	Correlation	$\rho = -0.5$	90	810	7300	66000
			$\rho = -0.2$	93	850	7800	72000
			$\rho = 0$	94	870	8100	75000
			$\rho = 0.2$	95	900	8300	77000
			$\rho = 0.5$	97	920	8700	81000

Table 3: Assessment factors (2 significant figures) for correlated log-normally distributed ratios.

(ii) U_S is the ratio of the sub-chronic NOAEL to the chronic NOAEL for the same species;

(iii) U_A is the ratio of the animal chronic NOAEL to the chronic NOAEL for a typical human;

(iv) U_H is the ratio of the chronic NOAEL for a typical human to the chronic NOAEL for a sensitive human.

Gaylor and Kodell (2000) cite Dourson et al (1996) saying that the latter "note from the examination of databases that uncertainty ratios for U_A, U_H, U_S and U_L appear to be approximately log-normally distributed". They go on to choose parameters for each log-normal distribution as follows:

(a) U_L: for 24 chemicals in Abdel-Rahman and Kadry (1995), median of LOAEL/NOAEL is 3.5 with 96% of ratios below 10; from this they deduce that $\ln \mathrm{U}_L$ is normal with mean 1.25 and standard deviation 0.60.

(b) U_S: for 149 chemicals in Pieters et al (1998), they obtained $\ln \mathrm{U}_S$ is normal with mean 0.53 and standard deviation 1.72.

(c) U_H: median is 1 on principle and for 490 chemicals in Dourson and Stara (1983), 10 is 92nd percentile of U_H; from this they deduce that $\ln \mathrm{U}_H$ is normal with mean 0 and standard deviation 1.64.

(d) U_A: median is 1 on principle and (for an unspecified number of chemicals) they interpret Calabrese and Baldwin (1995) to conclude that 26 is 97.5th percentile of U_A; from this they deduce that $\ln \mathrm{U}_A$ is normal with mean 0 and standard deviation 1.66.

7.2 Source of data, models and uncertainties

In all that follows, we follow the random-chemicals approach (as do Gaylor and Kodell (2000)); the four ratios are well-defined (unobserved) quantities for a chemical and, as a core form of

uncertainty about their values for a new chemical, we wish to quantify inter-chemical variability in the ratios and in their product. We also wish to quantify uncertainty about that variability. The calculations made by Gaylor and Kodell (2000) address only inter-chemical variability and make no allowance for uncertainty about either distribution shape or parameter values.

Pieters et al (1998) do justify log-normality for U_S, saying that quantile-quantile plots for log data are consistent with normality. We take the estimates of log-normal parameters provided by Pieters et al (1998) and attach sampling uncertainty to them in the usual way for a random sample from a normally distributed population.

Dourson and Stara (1983) make no such claim for U_H: they do not present the data as numbers but the histogram presented is quite misleading about the shape of the distribution. Careful re-plotting of the approximate data extracted from their histogram suggests that log-normality is not appropriate but that log-log-normality is an adequate fit. U_H is the ratio of typical human to sensitive human (and is related to human dose response slope). Therefore it has median greater than 1 and the data provided by Dourson and Stara (1983) should be seen as data directly on inter-chemical variation of U_H. We fit the log-log-normal distribution carefully to their data and obtain corresponding sampling uncertainty.

For U_L, Abdel-Rahman and Kadry (1995) make no distributional claim and again the data are presented as a histogram. The approximate data were again extracted from the histogram and careful formal statistical testing shows that log-normality is just about acceptable. The data provided by Abdel-Rahman and Kadry (1995) are data directly on inter-chemical (and inter-animal-species) variation of U_L. We fit the log-log-normal distribution carefully to their data and obtain corresponding sampling uncertainty.

For U_A, it is not clear that Gaylor and Kodell (2000) correctly interpreted Calabrese and Baldwin (1995); moreover, a careful reading of Calabrese and Baldwin (1995) and their source Barnthouse et al (1990) suggests that the interpretation made by Calabrese and Baldwin (1995) was not itself sound. For that reason we revisited and made a new interpretation of the data provided by Barnthouse et al (1990). U_A is the ratio of typical animal to typical human. Each line in the relevant part of table 3 of Barnthouse et al (1990) gives us an estimate for a pair of taxonomic orders of the mean difference in LC50s (intercept column) and, indirectly, an estimate of the magnitude of inter-chemical variation in the difference (Prediction interval column). We can turn the latter into the usual estimate of the error variance in a regression by dividing by the 97.5th percentile of the relevant t-distribution and by the usual factor depending on the sample size. We then model the inter-pair-of-orders variation in the mean differences using a log-normal distribution and carefully model the corresponding variation in the underlying true regression error variances by log-normality, in both cases also obtaining sampling uncertainty. Note that for U_A, there are two fundamental sources of uncertainty: firstly about the model for inter-pair-of-orders variability and then that variability is itself a source of uncertainty about inter-chemical variation. These must be carefully propagated when arriving at uncertainty about inter-chemical variation in U_A.

We now have models and sampling uncertainty for inter-chemical variation in each of the ratios. We do not compare our results to those of Gaylor and Kodell (2000) because some of our models are fundamentally different to theirs and we believe our choices to be preferable. We have not considered uncertainty about distribution family for any of these ratios.

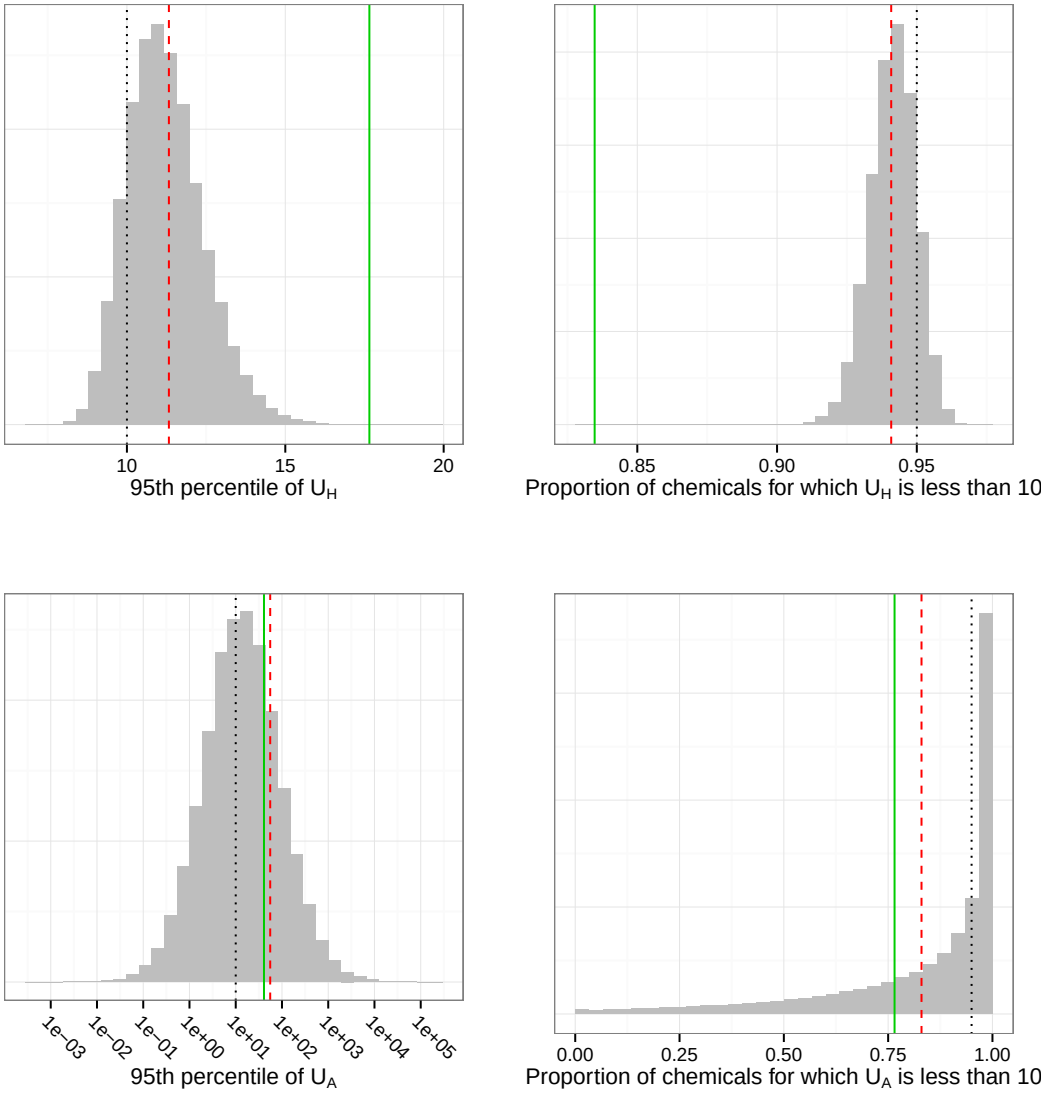
7.3 Results

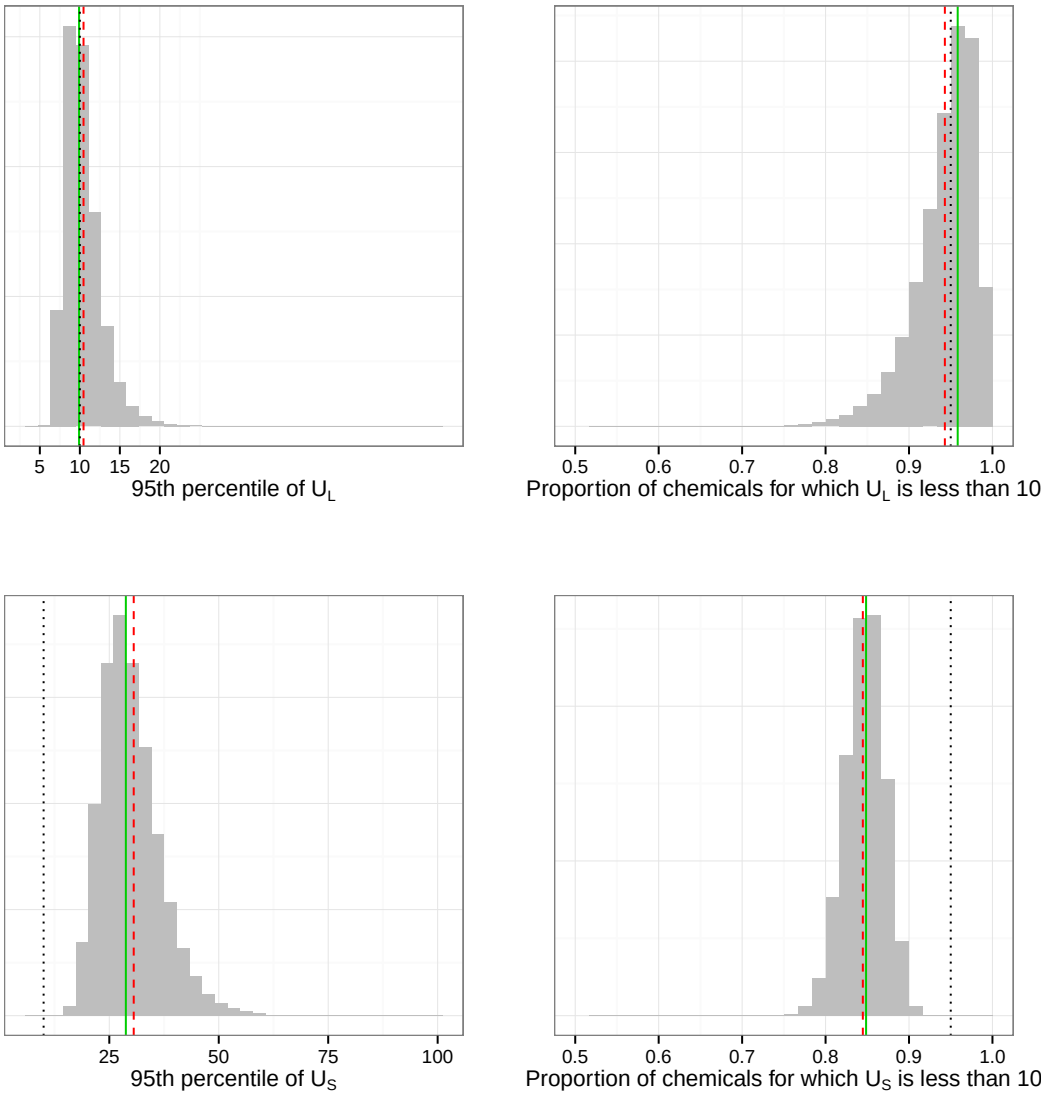
Following the pattern introduced by Gaylor and Kodell (2000), we focus for each ratio on two things: the inter-chemical 95th percentile of the ratio and the proportion of chemicals for which the ratio is less than a standard assessment factor of 10. To be precise, the following panels show for each of the four ratios:

- A histogram showing uncertainty about the inter-chemical 95th percentile of the ratio.

The overall (sampling uncertainty and inter-chemical variability combined) 95th percentile of the ratio for a new random chemical is shown as a dashed red line. The green line is naive estimate of the 95th percentile obtained using simple point estimates of parameters. The dotted black line corresponds to the standard assessment factor of 10.

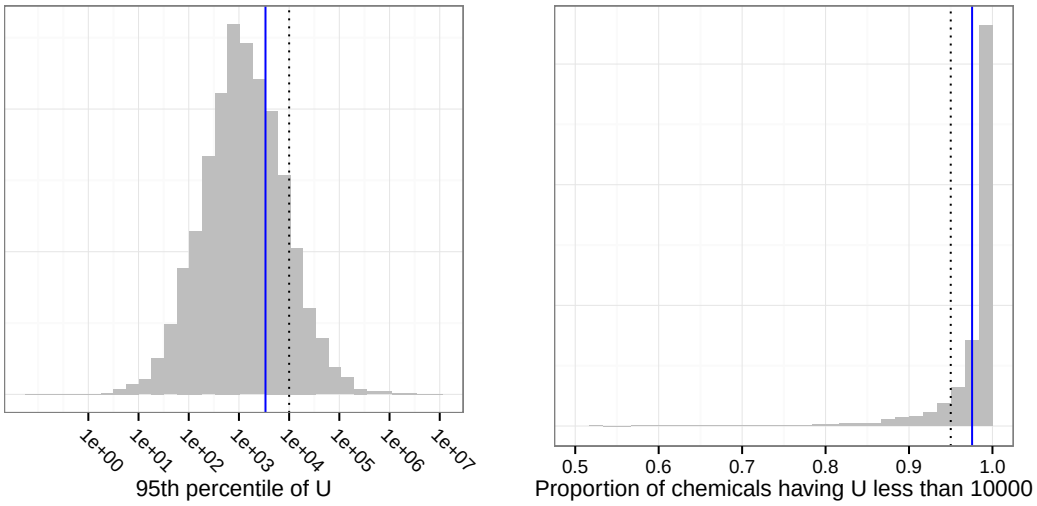
- A histogram showing uncertainty about the proportion of chemicals for which the ratio is less than 10. The dotted black line highlights the situation where the proportion is 95%. The overall (sampling uncertainty and inter-chemical variability combined) probability that the ratio is less than 10 for a new random chemical shown as a dashed red line. The green line is naive estimate of the same probability obtained using simple point estimates of parameters.





We also show results for the overall ratio $\mathrm{U} = \mathrm{U}_A \times \mathrm{U}_H \times \mathrm{U}_L \times \mathrm{U}_S$. The left panel below shows uncertainty about the 95th percentile of the overall ratio and about the proportion of chemicals for which the overall ratio is less than $10 \times 10 \times 10 \times 10 = 10000$:

- A histogram showing uncertainty about the inter-chemical 95th percentile of the overall ratio. The overall (sampling uncertainty and inter-chemical variability combined) 95th percentile of the ratio for a new random chemical is shown as a blue line. The dotted black line corresponds to the standard overall assessment factor of 10000.
- Uncertainty about the proportion of chemicals for which the ratio is less than 10000. The dotted black line highlights the situation where the proportion is 95%. The overall (sampling uncertainty and inter-chemical variability combined) probability that the ratio is less than 10000 for a new random chemical is shown as a blue line.



There are many interesting features to this case study. Some of the more important are:

- We have demonstrated that it is possible to model inter-chemical variability and associated sampling uncertainty for individual ratios. However, it requires careful evaluation and modelling of existing data.
- Individual models can be used to evaluate the effectiveness of individual assessment factors.
- When combining inter-chemical variability and uncertainty to provide uncertainty for random new chemical, although the individual assessment factors of 10 tend not to reach the 95th percentiles of their corresponding ratios, the overall assessment factor of 10000 actually reaches the 97th percentile. However, there is considerable uncertainty about performance if used for many chemicals although it is unlikely that 10000 covers fewer than 90% of chemicals..
- We have made no allowance for uncertainty about the family of distributions to be used for each model. It is unclear whether such uncertainty would make a large further contribution.
- The entire case study lives in the context created by Gaylor and Kodell (2000). We have not tried to re-evaluate the literature in relation to choice of datasets. It is likely that such a re-evaluation would affect the numerical results.

8 Conclusions

The principal conclusion is that we have only two practical choices at this time.

One is to continue to multiply assessment factors, accepting that this process may be excessively conservative when several factors are combined; doing so is undermined by K&G's work suggesting that some standard individual factors are likely to be too small for the ratios they deal with. A further problem is the lack of transparency; the assessment factors used in many areas lack a detailed technical justification. This could of course be rectified but would involve a careful analysis of the relevant data and scientific knowledge; it would also require some formalisation of the degree of confidence required.

The alternative is that we must follow the process proposed by Slob+al and K&G of using data (and biological insight) to determine distributions for individual ratios. However, we must go beyond K&G by addressing the uncertainties involved which include: (i) parameter uncertainty

when fitting distributions to data; (ii) uncertainty about the appropriate family of distributions to use; (iii) uncertainty about dependence between steps in the trajectory. This is quite a strong theme in the discussion of Cooke (2010). In doing so, we are following a fairly standard scientific procedure: we build a (Bayesian) statistical model which relates the uncertain and/or varying quantities which we can measure to the ones about which we wish to make inferences; we use the model as the basis for the decision we wish to make, in this case to determine the AF.

Some will be concerned that considering all the uncertainties will lead to unnecessarily large final assessment factors AF. However, the numerical and case studies provided suggest that the increase in size of the AF due to consideration of uncertainty may be counter-balanced by a reduction in size due to correct combination of distributions for individual steps in the trajectory.

It is expected that expert judgement will have an important role, especially in relation to the treatment of dependence between steps in the trajectory and in bringing understanding of the underlying science to bear on the question of what would be reasonable distributional assumptions.

9 Future work

We need to move further to the practical side and find out whether or not we can apply these ideas in the context of carcinogens and whether or not the framework needs revision.

Applying the methodology specifically to the context of carcinogens would need considerable effort, especially to assess the literature and relevant data. However, it would clearly be well worth doing if sufficient resource was available.

References

Abdel-Rahman M.S. and Kadry A.M. (1995), Studies on the Use of Uncertainty Factors in Deriving RfDs, *Human and Ecological Risk Assessment*, **1**(5), pp 614–624.

Barnthouse L.W., Suter II G.W. and Rosen A.E. (1990), Risk of Toxic Contaminants to Exploited Fish Populations: Influence of Life History, Data Uncertainty and Exploitation Intensity, *Environmental Toxicology*, **9**, pp 297–311.

Bosgra S., Bos P.M.J, Vermeire T.G., Luit R.J. and Slob W. (2005), *Regulatory Toxicology and Pharmacology*, **43**, pp 104–113.

Calabrese E.J. and Baldwin L.A., A Toxicological Basis to Derive Generic Interspecies Uncertainty Factors for Application in Human and Ecological Risk Assessment, *Human and Ecological Risk Assessment*, **1**, pp 555–564.

Calabrese E.J. and Gilbert C.E. (1993), Lack of Total Independence of Uncertainty Factors (UFs): Implications for the Size of the Total Uncertainty Factor, *Regulatory Toxicology and Pharmacology*, **17**, pp 44-51.

Cooke R (2010), Conundrums with Uncertainty Factor (with discussion), *Risk Analysis*, **30**(3), pp 330–338.

Dourson M.L., Felter S.P. and Robinson D. (1996), Evolution of Science-Based Uncertainty Factors in Noncancer Risk Assessment, *Regulatory Toxicology and Pharmacology*, **24**, pp 108-120.

Dourson M.L. and Stara J.F. (1983), Regulatory History and Experimental Support of Uncertainty (Safety) Factors, *Regulatory Toxicology and Pharmacology*, **3**, pp 224–238.

Gaylor D.W. and Kodell R.L. (2000), Percentiles of the Product of Uncertainty Factors for Establishing Probabilistic Reference Doses, *Risk Analysis*, **20**(2), pp 245–250.

Gaylor D.W. and Kodell R.L. (2002), A Procedure for Developing Risk-Based Reference Doses, *Regulatory Toxicology and Pharmacology*, **25**, pp 137–141.

Kodell R.L. and Gaylor D.W. (1999), Combining Uncertainty Factors in Deriving Human Exposure Levels of Noncarcinogenic Toxicants, *Annals of the New York Academy of Sciences*, **895**, pp 188–195.

Pieters M.N., Kramer H.J. and Slob W. (1998), Evaluation of the Uncertainty Factor for Subchronic-to-Chronic Extrapolation: Statistical Analysis of Toxicity Data, *Regulatory Toxicology and Pharmacology*, **27**, pp 108–111.

R Development Core Team (2011), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

Ruckdeschel P., Kohl M., Stabla T. and Camphausen, F. (2006), S4 Classes for Distributions, *R News*, **6**(2), pp 2–6.

Slob W. (2011). Conundrums?, *Risk Analysis*, **31**(1), pp 3–4.

Slob W. and Pieters M.N. (1998). A Probabilistic Approach for Deriving Acceptable Human Intake Limits and Human Health Risks from Toxicological Studies: General Framework, *Risk Analysis*, **18**(6), pp 787–798.

Vermeire T., Stevenson H., Pieters M.N., Rennen M., Slob W. and Hakkert B.C. (1999), Assessment Factors for Human Health Risk Assessment: A Discussion Paper, *Critical Reviews in Toxicology*, **29**(5), pp 439–490.