Bayesian Analysis (2013) **8**, Number 1, pp. 169–186

A Bayes Linear Approach to Weight-of-Evidence Risk Assessment for Skin Allergy

John Paul Gosling *, Andy Hart †, Helen Owen ‡,
Michael Davies §, Jin Li ¶, and Cameron MacKay ‖

Abstract. We introduce a strategy for quantifying and synthesising uncertainty about elements of a risk assessment using Bayes linear methods. We view the population of subjective belief structures and the use of Bayes linear adjustments as a flexible and transparent tool for risk assessors who want to quantify their uncertainty about hazard based on disparate sources of information. For motivation, we use an application of the strategy to human skin sensitisation risk assessment where there are many competing sources of information available.

Keywords: Bayes linear kinematics, Bayes linear methods, risk assessment, skin sensitisation, subjective judgement, weight-of-evidence

1 Introduction

We present a flexible method for combining lines of evidence of varying quality in a risk assessment. The method is based upon the theory of Bayes linear statistics (Goldstein and Wooff 2007) and its extension by Bayes linear kinematics (as described in Goldstein and Shaw 2004). The approach is amenable to weight-of-evidence assessments because it allows us to model the influence of many disparate sources of information and the computations for a new assessment can be done almost instantaneously.

The concept is to build a belief structure that relates the sources of information to the quantitative endpoint of the risk assessment. Within the Bayes linear framework, the risk assessors only need to specify first- and second-order moments for the quantities of interest. We use formal expert elicitation techniques to capture their knowledge (elicitation is reviewed in O'Hagan et al. 2006). Due to the flexibility of Bayes linear methods, we can accommodate incomplete data and informal evidence provided that beliefs about the link between the evidence and the quantity of interest have been specified.

We motivate the method by considering an application in the risk assessment of new chemicals. Specifically, we will be considering the potency that causes human

*School of Mathematics, University of Leeds, Leeds, UK, j.p.gosling@leeds.ac.uk
†Food and Environment Research Agency, York, UK, andy.hart@fera.gsi.gov.uk
‡Food and Environment Research Agency, York, UK, helen.owen@fera.gsi.gov.uk
§Safety & Environmental Assurance Centre, Unilever, Colworth, UK, michael.davies@unilever.com
¶Safety & Environmental Assurance Centre, Unilever, Colworth, UK, jin.li@unilever.com
‖Safety & Environmental Assurance Centre, Unilever, UK, cameron.mackay@unilever.com

 DOI:10.1214/13-BA807

sensitisation when a chemical is applied to the skin. Skin sensitisation and the resulting allergic contact dermatitis is an undesired immune response caused by a chemical coming into contact with the skin that presents clinically as a rash, skin lesion, papules or blistering at the site of exposure. The minimum amount of chemical (in $\mu g/\text{cm}^2$ skin) required to cause such a response is known as the potency of the chemical and more generally as the inherent hazard of the chemical. Risk assessors working in this area often need to weigh-up several lines of evidence from *in vivo* and, increasingly, *in vitro* experiments when characterising the potency for a new chemical in order to determine a safe dose for exposed individuals. Current quantitative risk assessment of skin sensitising chemicals often consists of estimating single values for potency and exposure levels, on some common scale, and comparing them as a ratio. This approach is deterministic, and any uncertainties in the estimates are accounted for through the use of safety factors. There have been recent attempts to assess parts of the risk problem probabilistically (see Jaworska et al. 2010, 2011; Safford 2008) and to model competing data sources formally (see Ellison et al. 2010). We take this approach further by explicitly considering the link between the available test data and actual human skin sensitisation potency. We use a Bayes linear framework to model the assessors' expectations and uncertainties and to update those beliefs in the light of the competing data sources. Such an approach to synthesising multiple lines of evidence and estimating hazard provides a transparent mechanism to defend and communicate risk management decisions.

The aim of the present paper is to demonstrate a methodology that captures experts' quantitative beliefs about hazard for use in a risk assessment. We do not wish to impose a rigid model structure for performing calculations: the aim of this paper is to communicate one plausible and defensible strategy for completing weight-of-evidence risk assessments using Bayes linear methods. In the subsequent sections, we describe methods that could be adapted to many different quantitative assessments where the data sources are varied and their relative influence is complex. In the discussion section at the end of the paper, we highlight the benefits of our approach over probabilistic Bayesian modelling for this type of application.

2 Bayes linear models and kinematics

Bayes linear methods provide a framework for modelling beliefs about the relationships between variables of interest. In particular, they are methods for statistical modelling and inference within a subjectivist framework. Traditional subjective Bayesian analysis is based upon fully-specified probability distributions, which can be difficult to obtain at the necessary level of detail. Bayes linear methods attempt to solve this problem by developing updating rules using partially specified beliefs: expectations and covariances are used to model beliefs rather than full probability distributions. Some people view Bayes linear analyses as approximations to traditional Bayesian analyses. Nevertheless, Bayes linear methods follow the same principle: prior beliefs are specified and then updated in the light of new data. Bayes linear methodology provides a tractable framework for updating beliefs when full specification of joint probability distributions is too costly. Practical examples of the application of Bayes linear methods are given

in O'Hagan et al. (1992); Craig et al. (1997); Farrow et al. (1997). A comprehensive introduction to the theory and methods can be found in Goldstein and Wooff (2007).

2.1 Bayes linear adjustments

Let $\mathbf{B}$ denote a collection of quantities about which we wish to make inferences. Allow $\mathbf{D}$ to be a subset of $\mathbf{B}$ for which we will learn values. Before we learn $\mathbf{D}$, we specify our prior beliefs over $\mathbf{B}$ (this encompasses our beliefs about $\mathbf{D}$): the expectation of $\mathbf{B}$, $\mathrm{E}(\mathbf{B})$, and the corresponding variance-covariance matrix, $\mathrm{Var}(\mathbf{B})$. We can update our beliefs about $\mathbf{B}$ given an observed $\mathbf{D}$ using the following formulae:

$$\begin{aligned} \mathrm{E}_{\mathbf{D}}(\mathbf{B}) &= \mathrm{E}(\mathbf{B}) + \mathrm{Cov}(\mathbf{B},\mathbf{D})\mathrm{Var}^{-1}(\mathbf{D})\left[\mathbf{D} - \mathrm{E}(\mathbf{D})\right], \\ \mathrm{Var}_{\mathbf{D}}(\mathbf{B}) &= \mathrm{Var}(\mathbf{B}) - \mathrm{Cov}(\mathbf{B},\mathbf{D})\mathrm{Var}^{-1}(\mathbf{D})\mathrm{Cov}(\mathbf{D},\mathbf{B}). \end{aligned}$$

These formulae are reached by choosing the linear combination of the elements of $\mathbf{D}$ that minimise the squared difference between that linear combination and the elements of $\mathbf{B}$; full details and justifications for these updating rules are given in Goldstein and Wooff (2007). In practice, as in a probabilistic Bayesian analysis, the updates can be done sequentially. When setting up an environment within which these updates take place, it can be useful to decompose the update into scalar-update steps because it is then possible to gauge the impact of each data source.

In our particular application, the information that is brought to bear when weighing the evidence is not in such a clear-cut form. For some pieces of information, we may be able to rule out portions of the variable-space or we might have some anecdotal evidence that would make us want to shift our prior beliefs. We could extend our Bayes linear model to accommodate such information without changing the updating rules, but this would result in an increase in the number of variables over which we need to specify beliefs. An alternative approach is to use Bayes linear kinematics.

2.2 Bayes linear kinematics

The Bayes linear adjustments can be modified when, rather than obtaining a single value for $\mathbf{D}$, we receive information that changes our beliefs about $\mathbf{D}$ through the use of kinematics. The theory underpinning these methods is given in Goldstein and Shaw (2004) and the general concept of kinematics is introduced in Jeffrey (1983).

Again, we have an initial prior specification over $\mathbf{B}$: $\mathrm{E}(\mathbf{B})$ and $\mathrm{Var}(\mathbf{B})$. We now change our beliefs about a subset of $\mathbf{B}$; that is, we change our beliefs from $\mathrm{E}(\mathbf{D})$ and $\mathrm{Var}(\mathbf{D})$ to $\mathrm{E}_n(\mathbf{D})$ and $\mathrm{Var}_n(\mathbf{D})$ respectively where n denotes new beliefs. We can update our beliefs about $\mathbf{B}$ using Bayes linear kinematics updating formulae:

$$\begin{aligned} \mathrm{E}_n(B) &= \mathrm{E}(\mathbf{B}) + \mathrm{Cov}(\mathbf{B},\mathbf{D})\mathrm{Var}^{-1}(\mathbf{D})\left[\mathrm{E}_n(\mathbf{D}) - \mathrm{E}(\mathbf{D})\right], \\ \mathrm{Var}_n(\mathbf{B}) &= \mathrm{Var}(\mathbf{B}) - \mathrm{Cov}(\mathbf{B},\mathbf{D})\mathrm{Var}^{-1}(\mathbf{D})\mathrm{Cov}(\mathbf{D},\mathbf{B}) \\ &\quad + \mathrm{Cov}(\mathbf{B},\mathbf{D})\mathrm{Var}^{-1}(\mathbf{D})\mathrm{Var}_n(\mathbf{D})\mathrm{Var}^{-1}(\mathbf{D})\mathrm{Cov}(\mathbf{D},\mathbf{B}). \end{aligned}$$

These are an extension of the usual Bayes linear updating rules: if the value of $\mathbf{D}$ is learnt with certainty, then $\mathrm{E}_n(\mathbf{D}) = \mathbf{D}$ and $\mathrm{Var}_n(\mathbf{D}) = \mathbf{0}$.

These updating rules result from satisfying a Bayes linear sufficiency condition. The condition is that, if we were to learn the exact value of $\mathbf{D}$, then $\mathrm{E}_{\mathbf{D}}(\mathbf{B}) = \mathrm{E}_{n,\mathbf{D}}(\mathbf{B})$ and $\mathrm{Var}_{\mathbf{D}}(\mathbf{B}) = \mathrm{Var}_{n,\mathbf{D}}(\mathbf{B})$ regardless of the specified $\mathrm{E}_n(\mathbf{D})$ and $\mathrm{Var}_n(\mathbf{D})$. Adherence to this condition implies that care needs to be taken when using Bayes linear kinematics to update a belief specification sequentially. If we apply new sets of beliefs to data nodes sequentially, then we will get different results if we change the order of the updates. In practice, if there is only one non-zero element in $\mathrm{Var}_n(\mathbf{D})$ (and the experts are not available to judge the impact of the other sequential adjustments on $\mathrm{E}_n(\mathbf{D})$ and $\mathrm{Var}_n(\mathbf{D})$), the kinematic update should be performed first. This can be followed by the single Bayes linear updates in any order. If there are several variables being updated using the kinematic procedure, there are two main ways that we can deal with the problem of consistency. The most useful would be to check the sensitivity of the analyses to the ordering. Alternatively, we could complete the kinematic update for all the variables that are going to be treated in that way before updating sequentially using the standard Bayes linear update.

3 Assessment of skin allergy risk

Skin sensitisation refers to a human health risk (manifesting as allergic contact dermatitis) that can be caused by skin contact with a wide range of chemicals, including those used in personal care products. In terms of biological mechanism, the key step is regarded to be the chemical reaction between the sensitising chemical and proteins in the skin (Basketter et al. 1995). This insight has motivated development of a number of *in vitro* tests for assessing protein reactivity of skin sensitisers (see Aleksic et al. 2009; Gerberick et al. 2004, for example). Further explanation of this health risk and current risk assessment procedures is given in Basketter (2008).

The quantity of interest in our weight-of-evidence assessment is the mean threshold for skin sensitisation (or sensitisation potency) for the population of consumers. For our purposes, sensitisation potency is measured in terms of chemical mass per unit area of skin ($\log_{10} \mu g \text{ cm}^{-2}$). Before considering the sensitising potency of the chemical, the risk assessors consider whether the chemical has the potential to be a skin sensitiser in humans. A method to approach this, which handles the disparate lines of evidence within a probabilistic framework, has been presented in Owen et al. (2012). Here, we consider the quantification of uncertainty about the sensitising potency, and we assume that the chemical under consideration is a skin sensitiser.

The overall aim of this case study is to capture the risk assessors' current views on the relative influence of different tests on beliefs about sensitisation potency for humans. We used formal expert elicitation techniques to capture knowledge about the relationships between the different experiments and skin sensitising potency for humans. We split the problem into two parts: building a conceptual model for the links between the experiments and the quantity of interest, and populating the model

with beliefs about the first- and second-order moments of the modelled quantities. We facilitated nine elicitation workshops at Unilever[1] between July 2010 and March 2011. The experts were risk assessors and chemists, nominated by Unilever, who had a wide range of experience covering both the application and analysis of *in vivo* and *in vitro* experiments. The facilitators of the exercise were the authors who have experience in facilitating group elicitation sessions, expertise in statistical modelling and extensive experience of risk analysis. The facilitators led the elicitation process and carried out calculations with the elicited judgements. Although we have an understanding of the application area, it was important for us to remain impartial and not influence the judgements. The strategy for eliciting the group's opinions is similar to the method of Gosling et al. (2012) where the focus of the facilitated meetings was on the capturing of a group consensus. This is not the only method we could have employed. A recent review of the group elicitation problem is given in French (2011).

3.1 Influence diagrams and their population

Undirected graphs provide a simple way of representing beliefs about dependencies between a collection of variables. Figure 1 illustrates a portion of the graph from the skin sensitisation risk model. In Figure 1, three variables of interest are displayed as nodes on a graph. If there is a direct link between nodes and we learn about one of the linked nodes, then beliefs about the other nodes will change. If there is no direct link between the nodes (for example, between "Guinea pig maximisation test" and "True human potency"), learning about one of the nodes will inform us about the intermediate nodes, and they, in turn, will alter the beliefs about the other nodes. Hence, we use these graphs to build-up a model for the conditional independence for the variables.

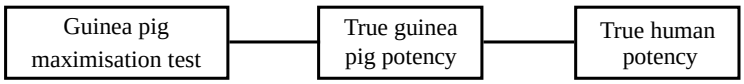

Figure 1: An undirected influence diagram.

Uncertainties about the true skin sensitising potency for humans stem from the inadequacies of experimentation (impurities in the test chemical for example) and the fact that chemical reactivity and laboratory animals are not perfect models for the average human. In our modelling, we make the distinction between the experimental result and the true chemical sensitisation potency and chemical reactivity[2]. This allows us to separate the uncertainty about the adequacy of the experiments from the uncertainty about whether laboratory animals and reactivity are suitable predictors for human sensitisation. This approach of separating true quantities from experimental results is similar in principle to the work of Turner et al. (2009), who separated different sources of bias using idealised studies, and of Goldstein and Rougier (2009), who introduced the concept

[1] Colworth Science Park, Bedford, UK.

[2] Here chemical reactivity is a measure of the rate of binding of the chemical to a suitable protein (glutathione at 20 °C and pH 7.4, measured in $\log_{10}$ per second per molar, $\log_{10} s^{-1}M^{-1}$) and is the focus of the *in vitro* tests.

of an idealised computer model.

If the experts are able to identify conditional independence between variables, the following formula for belief separation can be applied. If we have three collections of random variables, denoted by X, Y and Z, and we believe that, if we know the value of Y, then learning Z will tell us nothing more about X, then

$$\mathrm{Cov}(X, Z) = \mathrm{Cov}(X, Y)\mathrm{Var}(Y)^{-1}\mathrm{Cov}(Y, Z)$$

will reduce the number of judgements that the experts need to make. We are making a distinction between the true values for the quantities of interest and corresponding experimental observation. This gives us an opportunity to build conditional independence into our model because, if the assessors were to know the true values of the animal potencies, then experimental data on one animal would not tell us anything extra about an experiment on another species.

Figure 2 shows one version of the influence diagram that was proposed by the risk assessors, which was constructed by identifying plausible conditional independencies. The shaded boxes represent experimental results that we might obtain for a new chemical and the unshaded boxes represent true quantities for the chemical. The three shaded boxes on the top right represent *in vivo* experimental results and the bottom two shaded boxes represent *in vitro* experimental results. As mentioned previously, there are no direct links between any of the experiments because, if we knew the true animal potency, the corresponding *in vivo* experiment would give us no extra information about the other true quantities (or the experiments that aim to measure them). In Table 1, we list details of the variables given in Figure 2 along with the abbreviations we will use hereafter.

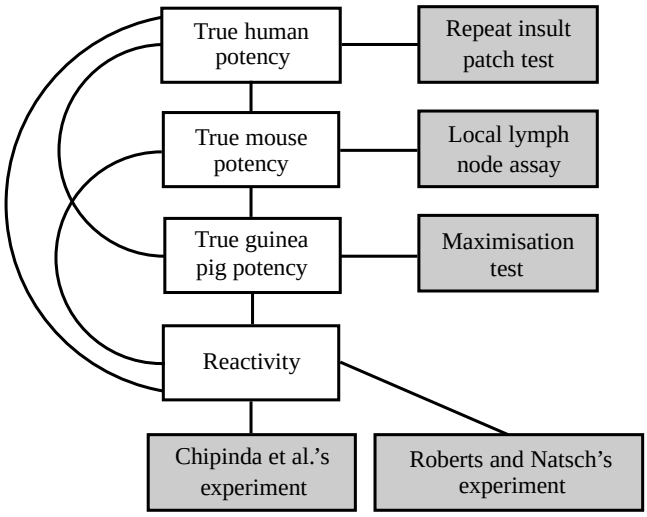

Figure 2: The complete influence diagram for the skin sensitisation application.

Table 1: Descriptions of the variables in Figure 2

Node label	Abbr.	Description
True human potency	H_A	Mean human potency (threshold for sensitisation) under occlusion for the population of consumers (measured in $\log_{10} \mu g \text{ cm}^{-2}$).
True mouse potency	M_A	Mean mouse potency (threshold for sensitisation) under occlusion for the population of CBA-strain mice used in experiments (measured in $\log_{10} \mu g \text{ cm}^{-2}$).
True guinea pig (GP) potency	G_A	Mean GP potency (threshold for sensitisation) under occlusion for the population of GPs used in experiments (measured in $\log_{10} \mu g \text{ cm}^{-2}$). It is assumed that the GP has been shaved.
Repeat insult patch test (HRIPT)	H_D	The no-observed-effect levels (NOELs) from a human test (measured in $\log_{10} \mu g \text{ cm}^{-2}$). This could be largest NOEL from several reported experiments. The experiment follows the protocol of Stotts (1980).
Local lymph node assay (LLNA)	M_D	This is the experimental effective concentration that sees a three times increase in lymph node stimulation over control mice (EC3) (measured in $\log_{10} \mu g \text{ cm}^{-2}$). The desired vehicle is a 4:1 mix of acetone and olive oil. The test is on an inbred strain of mice. The experiment follows the protocol of Gerberick et al. (2000).
Maximisation test	G_D	This is a Gerberick classification taken from the following set once a maximisation test has been performed on a number of GPs: "very weak", "very weak/weak", "weak", "moderate", "moderate/strong", "strong", "potent" and "negative". It is assumed that the GPs have been shaved. The experiment follows the protocol described in Andersen and Maibach (1985).
Reactivity	R_A	This is the rate constant ($\log_{10} k$) for a reaction in glutathione at 20 °C and pH 7.4 (measured in $\log_{10} s^{-1} M^{-1}$).
Chipinda et al.'s reactivity experiment	R_{D1}	This is a rate constant ($\log_{10} k$) measured by kinetic profiling for a reaction in nitrobenzenethiol at 25 °C and pH 7.5 (in $\log_{10} s^{-1} M^{-1}$). This should be corrected to 20 °C and pH 7.4 before use in the model. The experiment follows the protocol described in Chipinda et al. (2010).
Roberts and Natsch's reactivity experiment	R_{D2}	This is a rate constant ($\log_{10} k$) measured by kinetic profiling for a reaction in the peptide Ac-RFAACAA at 25 °C and pH 7.5 (in $\log_{10} s^{-1} M^{-1}$). This should be corrected to 20 °C and pH 7.4 prior to entry into the model. The experiment follows the protocol described in Roberts and Natsch (2009).

For each of the variables shown in Figure 2, we elicited information from the experts about the mean and variance when considering all possible chemicals they might assess in the future. We also asked them to restrict their focus to chemicals that fall within the Michael acceptor domain because the assessors expected that the different chemical domains have different relationships between the experiments (see Tokoroyama 2010, for further details on the Michael acceptor class). Our strategy for this part of the elicitation was

1. revisit the precise definition of the quantity,
2. consider situations where the value could be relatively very small or large,
3. elicit summaries of a probability distribution,
4. fit an appropriate distribution to the judgements (using the technique of O'Hagan 1998),
5. feed back the fitted distribution to the experts (showing plots of the density and reporting different distributional summaries),
6. allow the experts to revise their original judgements until they are satisfied with the representative distribution,
7. use the fitted distribution to derive an expectation and variance for the quantity of interest.

The most important part of the modelling is the characterisation of the links between the nodes. These links represent correlations that need to be elicited from the assessors. To do this, we used two strategies. When the quantities linked together are on the same scale and of the same type (human to guinea pig potency, for example), we asked the assessors about the difference between the two quantities. Then using the earlier judgements about the marginal quantities, it is trivial to calculate the corresponding correlation. When the quantities were not of the same type (the relationship between reactivity and mouse potency, for example), we set hypothetical values for the first variable and elicited beliefs about the second conditional on the first. By repeating this process, we were able to recover the implied correlation between the variables. These methods are described in more detail in Clemen et al. (2000) and, from a Bayes linear perspective, in Revie et al. (2010).

We needed to elicit information about nine expectations, nine variances and eleven correlations to complete the prior belief specification based on the influence diagram of Figure 2. In practice, we elicited more information than this to check the consistency of the risk assessors' judgements and the suitability of the model. Throughout the elicitation sessions, we were wary of questioning fatigue in the risk assessors. We cut short the planned elicitation sessions if the experts were beginning just to repeat the same judgements, and we spent a lot of time feeding back results and consequences of judgements to break up the repetitive process of defining variables and making judgements on them (steps 1 to 3 in the strategy given above).

Of course, all of the judgements are based on the experience of the risk assessors and are strongly influenced by the many datasets they have seen. This is intentional: we are trying to capture the assessors' process, and we want to understand how they use the disparate data sources in an assessment given their knowledge and experience of skin sensitising potency and the chemicals they assess. Given the inherent subjectivity of the approach, we dedicated whole day sessions to exploring the consequences of judgements with the experts particularly highlighting the judgements that greatly affected the adjusted beliefs for H_A and considering the impact of changing the model structure shown in Figure 2.

As mentioned in Section 2.1, Bayes linear updating proceeds by altering the expectation and variances for the quantities of interest using observed data from the modelled experiments. For M_D, R_{D1} and R_{D2}, the experiment data come in the required, single-value form. However, for the other experiments, the results are not so clear cut. For H_D, the result will typically be a no-observed-effect level (NOEL), which can be thought of as a lower bound on a result from an experiment measuring the mean sensitisation potency for humans. For G_D, the result of the experiment is a classification into a potency category. The risk assessors view this as the true guinea pig potency falling within a range of possible potencies that typically span one order of magnitude. We can treat this as information that alters the risk assessors' beliefs about the result of a guinea pig experiment that directly measures sensitisation potency. For both of these sources of information, we must invoke the kinematic adjustment rules as described in Section 2.2. When updating the model sequentially, we first adjusted the model using G_D to ensure consistency with the judgements the experts made about the guinea pig classifications. Again, it was important for us to check the consistency of the adjustment mechanism with the risk assessors' presumptions for the model's behavior. We did this by first testing the model on a number of chemicals with known skin sensitisation properties (one such example is given in the next section), and we investigated the robustness of our results by perturbing the judgements over sensible ranges and by changing the sequence of the Bayes linear adjustments.

3.2 Results

Case study: cinnamic aldehyde

We begin this section by looking at one application of the model. Cinnamic aldehyde is used to give products a cinnamon aroma, and it is a known skin sensitiser (Schorr 1975). A risk assessment for cinnamic aldehyde was considered because almost all of the experiments in the model have been performed for this chemical (for some experiments, the tests have been carried out many times). The different sources of information that are relevant to the model for this chemical are listed in Table 2. Note that, the lower the sensitisation concentration, the more potent the chemical is said to be.

Given the experts' belief specifications and if we treat cinnamic aldehyde as a new chemical, we can use the risk assessment model of the previous section. The risk assessors begin with the prior expectation of 2 and a variance of 2 (on the $\log_{10}$-scale

Table 2: The experimental data used to update the model.

	Source of information[3]	Value used
H_D:	We have the following three NOELs: 97, 388 and 591 (in μg cm^{-2}).	591 μg cm^{-2}
M_D:	We have that a concentration of 1.56% gives a stimulation index of 3 in the mouse (this can be converted to μg cm^{-2}).	390 μg cm^{-2}
G_D:	There are several maximisation test results, but they show "strong" sensitising potential.	Strong
R_{D2}:	A value for log(k) of -1.66 has been measured using the method of Roberts and Natsch (2009).	-1.66 $\log_{10}\left(s^{-1}M^{-1}\right)$

for H_A). This implies that the risk assessors have little idea where the true sensitising concentration for humans lies over seven orders of magnitude. After updating the model with the information in Table 2, the beliefs are adjusted to an expectation of 2.96 with variance of 0.11. This could be interpreted as being quite sure that the true potency is within half an order of magnitude of 1000 μg cm^{-2}. It is clear that this value is some way from the experimental results reported in Table 2. This is because the data points are thought to be no-expected-sensitisation levels whereas the model focusses on the mean threshold for skin sensitisation for the population of consumers, which is likely to be higher.

Earlier, when introducing Bayes linear methods, we mentioned the benefits of avoiding making full probabilistic specifications. However, by using this approach, we lose the ability to make statements about the probability of the potency being less than some threshold of interest. Bayes linear approaches are a pragmatic step to getting a quantitative indication of the uncertainties in the quantities of interest and their relative influence on human potency. That said, there are ways to construct bounds for probabilities from expectations and variances using Chebyshev's or the Vysochanskij-Petunin inequality (Vysochanskij and Petunin 1980). Using these, we could say something about the probability of exceeding (or falling below) a potency threshold, which could be useful to the risk assessors.

Consequences of the model

We can also use the model to explore some of the features of the risk assessor's beliefs. For instance, we can calculate which sources of information reduce the expert's uncertainty about the quantity of interest the most.

[3]The HRIPT NOELs, the LLNA data and the guinea pig classification (as defined in Table 1) are taken from an unpublished data set.

The resolution of a variable X induced by observing variables D is defined as

$$\text{Res}_D(X) = 1 - \frac{\text{Var}_D(X)}{\text{Var}(X)}.$$

The resolution is related to the R^2 value that is often calculated when performing standard linear regression because it measures how much of the uncertainty has been explained by the information that we have added into the model. Resolution is reported on a scale of zero to one: a value of zero means that we have learnt nothing about the variable from observing D and a value of one means that we have nothing more to learn about the variable.

In Table 3, we list the resolutions of H_A, M_A, G_A and R_A obtained when updating the risk assessors' beliefs using the data sources shown. If we accept the Bayes linear model as an adequate representation of the risk assessors' beliefs, then the resolutions in Table 3 show that the *in vivo* experiments tell the assessors more about the animal potencies than the *in vitro* experiments. We also have that the *in vitro* experiments tell us more about the chemical reactivity. It is also clear that the assessors currently believe that a result from the mouse experiment (M_D) produces the largest reduction in the uncertainty about human potency (H_A) despite the presence of human experimental data (H_D). This is due to the human data being included in the model as a bound on the potency rather than a single value for a threshold like the mouse data. It should also be noted that the resolutions for M_D, R_{D1} and R_{D2} are not affected by the actual experimental result (this is obvious from the variance adjustment formula of Section 2.1). The resolutions are affected significantly when changing the value of H_D because we are using the kinematic adjustment formulae and, as we increase the NOEL value, we rule out more potency values.

	Resolution for			
Data used	H_A	M_A	G_A	R_A
$\text{H}_D = 0 \log_{10} \mu g \text{ cm}^{-2}$	0.24	0.22	0.22	0.02
$\text{H}_D = 2 \log_{10} \mu g \text{ cm}^{-2}$	0.59	0.55	0.55	0.06
$\text{H}_D = 4 \log_{10} \mu g \text{ cm}^{-2}$	0.78	0.73	0.73	0.08
$\text{M}_D = 0 \log_{10} \mu g \text{ cm}^{-2}$	0.88	0.94	0.88	0.10
$\text{M}_D = 2 \log_{10} \mu g \text{ cm}^{-2}$	0.88	0.94	0.88	0.10
$\text{M}_D = 4 \log_{10} \mu g \text{ cm}^{-2}$	0.88	0.94	0.88	0.10
G_D = "Very weak"	0.86	0.86	0.92	0.09
G_D = "Moderate"	0.83	0.83	0.89	0.09
$\text{R}_{D1} = \text{-2} \log_{10} s^{-1} M^{-1}$	0.10	0.10	0.10	0.99
$\text{R}_{D1} = \ \ 0 \log_{10} s^{-1} M^{-1}$	0.10	0.10	0.10	0.99
$\text{R}_{D2} = \text{-2} \log_{10} s^{-1} M^{-1}$	0.10	0.10	0.10	0.99
$\text{R}_{D2} = \ \ 0 \log_{10} s^{-1} M^{-1}$	0.10	0.10	0.10	0.99

Table 3: Resolutions induced in H_A, M_A, G_A and R_A given the single results shown.

We can also consider the results of exposing the model to conflicting experimental results. Conflicting experimental results do occur for some chemicals; the model can cope

with conflicts because we have explicitly modelled differences between species and between experimental results and true values for the chemical potency. For example, if we have conflicting evidence from M_D and G_D ($M_D = 20,000 \mu g$ cm^{-2} and $G_D =$"Strong"), then the Bayes linear adjustment gives us an expectation of $4,700 \mu g$ cm^{-2}, which is a compromise between the two experimental results. When such conflicts exist in the data, it would be good practice for the risk assessors to consider why such differences occur and whether the original belief specifications were appropriate. It is possible to detect such outliers using standardised versions of the quantities of interest as described in Goldstein and Wooff (2007).

The resolution calculations can be performed on the model without much computational cost. This means that the consequences of the experts' judgement can be fed back quickly and thorough sensitivity analyses can be performed to establish where more effort in the expert elicitation process might benefit the accuracy of the model. For instance, in Figure 3, we have produced a plot of the resolution of the assessors' uncertainty about the true human potency given *in vitro* data (either R_{D1} or R_{D2}) for different values of correlation between reactivity and animal potency. It is clear that the value specified for this correlation has a large impact on inferences from the model when *in vitro* data are available. Given this information, a large portion of the elicitation sessions was focussed on this parameter (this was not just due to this sensitivity: the correlation was also difficult for the assessors to express beliefs about due to relative lack of familiarity with new *in vitro* methods compared to established *in vivo* methods).

A by-product of the model is an assessment of the true mouse potency given results from the *in vitro* experiments alone. This is beneficial because, prior to obtaining M_D, the model can tell us what value of M_D to expect. The expected true mean mouse threshold for sensitisation derived from the model could help to set the test concentrations, and, therefore, help avoid testing at unrealistically low and high concentrations. Of course, the same argument could be extended to the other experiments. Indeed, the model itself could easily be extended to include other *in vitro* experiments such as the peptide depletion experiments of Aleksic et al. (2009) and Gerberick et al. (2004). A further benefit of the model is that it gives an understanding of the confidence provided by experimental data for making risk assessment decisions. In addition to experimental data, understanding human exposure is essential to the risk assessment process and limiting exposure is often the risk manager's main tool in being able to control risk. Including exposure in the model would allow risk assessors to have a better understanding of how it determines the risk with varying degrees of experimental data. Such an understanding of uncertainty in both exposure and experimental data would be of great utility in deciding whether new *in vitro* experiments provide sufficient confidence for making risk assessment decisions and thus potentially obviating the need for *in vivo* testing.

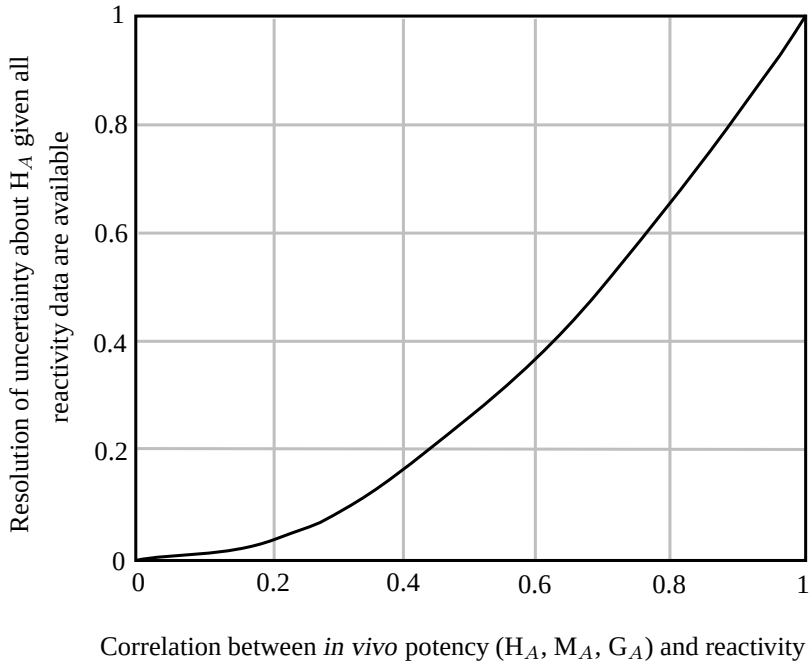

Figure 3: The effect of changing the value of correlation on the resolution of uncertainty about human potency.

4 Discussion

4.1 Application and extensions of the skin sensitisation model

Skin sensitisation is an important safety consideration in risk assessment of consumer products for which topical exposure is intended. Risk assessment of potential skin sensitising ingredients proceeds by ensuring the predicted consumer exposure to the product ingredient (μg cm^{-2} of skin) is sufficiently far from the expected sensitisation threshold in the consumer population (μg cm^{-2} of skin). The sensitisation threshold is chemical specific and is determined through an expert weight-of-evidence interpretation of the available data. There are many alternative strategies for completing weight-of-evidence assessments: many are reviewed in Weed (2005) and Suter and Cormiera (2011). Our proposed method gives a simple framework within which many competing sources of information can be handled, the assessors' beliefs about the collected information and the quantities of interest are explicit, and the model can be extended and refined as knowledge increases and new experimental approaches are developed. These qualities should make the approach valuable to risk assessors in many different fields.

The final structure for assessing the chemical potency for skin sensitisation was the result of many hours of discussion. The links between the nodes result from a

consideration of how different sources of information influence the risk assessors' beliefs and of what is possible in terms of quantitative modelling. Due to the flexibility of the Bayes linear approach, the latter set of considerations placed little constraint on the development of the model. We are confident that, in these models, we have correctly captured the experts' beliefs. Time was dedicated to feeding back the consequences of these model choices in the expert workshops and revising the model where appropriate. The feedback and revision elements are important when using expert judgements to build and populate a model: we had several iterations of these stages to capture the experts' beliefs. Of course, models populated by different experts could easily lead to different conclusions with regards to the uncertainty in human sensitisation potency. Due to the simplicity of the model and the transparency of specifying a finite number of model parameters, it would be easy to identify where the differences lie. This sort of information could help experts to focus their research efforts and scientific debate and help to highlight reasons for different opinions to the risk managers.

Risk assessors are interested in a number of questions with regard to this approach for determining sensitisation thresholds: which assays or tests provide the most confidence in specifying the sensitisation threshold; and, do some assays provide equivalent confidence on the sensitisation threshold. In the context of finding alternative approaches to animal testing, the question of which assays provide equivalent or sufficient confidence in the sensitisation threshold is of particular importance. This is reflected in the results presented here where the resolution in human threshold due to animal data (for example, mouse LLNA) can be directly compared with that of non-animal data (for example, the Chipinda reactivity data). However, it is important to emphasise that the sensitisation threshold is only one half of the risk assessment problem and that decisions on safety are always made in the context of consumer exposure. Ideally, the model presented should be extended to make weight-of-evidence judgements on the overall risk of sensitisation for a given exposure. By extending the model in this way, the safe level of exposure supportable for a given test result and resolution could be determined. Such a principle has already been used to set safe exposures using only prior knowledge (Safford 2008; Safford et al. 2011).

4.2 Why is a kinematic approach appropriate?

The model we have presented in this paper gives us a mechanism for updating beliefs about the average human skin sensitisation threshold for a chemical given disparate data sources. Of course, this updating could be done through a probabilistic Bayesian model with some hierarchical structure. If we view our Bayes linear model as a simple approximation to a full Bayes analysis, then a natural extension would be to model the beliefs using probability distributions and handle the updating through appropriate likelihoods. Of course, this would be tremendously challenging in an example where there are several disparate sources of information. However, if we were able to produce a probabilistic model for part of the model (linking one source of data to an animal's average sensitisation threshold say), we can apply a kinematic approach to update part of the belief structure that has undergone a probabilistic Bayesian analysis (see Gold-

stein and Shaw 2004). In this application, we chose to use the Bayes linear framework for several reasons:

1. The risk assessors could potentially have different results for each assay type from different research laboratories and may disagree on the strength of the evidence from each data source. Our model allows them to make changes and view the consequences of their specifications instantaneously. If we had some numerical integration to perform over a multidimensional space in a fully probabilistic treatment, then we could not match this speed of calculation.

2. The simple structure of the belief specification allows us to add new elements to the model relatively easily. In some cases, it could be as easy as specifying just one expectation, variance and correlation. This was important to the risk assessors because many new assays are being developed at the moment.

3. It is not the norm in fully probabilistic models to have data that are uncertain once they have been collected. Although it is possible to model this in a probabilistic way by adding an extra hierarchical level, the Bayes linear kinematic gives a consistent and easily-implementable way of including such data.

4. In specifying the Bayes linear model, the focus is on just expectations, variances and correlations so we need a finite number of judgements. The result of this is that the experts understand the elicitation task, and it is easier to perform a sensitivity analysis to evaluate the robustness of the model to the judgements.

5. Through several trials of the model with the experts, we felt that the resulting adjusted belief structures closely approximated the experts' thought process when assimilating the data.

Overall, we believe that, by following a Bayes linear modelling approach, we have produced a model that is consistent with the risk assessors' thought process, quick to evaluate and relatively simple to modify (either in terms of structure or individual judgements). By following a Bayes linear scheme, we lose the ability to talk directly about our results in terms of chances or probabilities of adverse effects occurring (although we can derive bounds on probabilities as mentioned in the previous section). This omission rules out the application of standard decision theory methods, but there are Bayes linear analogues based around expected utilities (as discussed in Goldstein and Wooff 2007).

If we were to pursue a fully probabilistic modelling approach, we could build a hierarchical model that modelled the data from each source simultaneously with an ordinal variable for the guinea pig maximisation test results. However, we would find that the computation of the posterior would be much more time consuming, and it might be more difficult for the risk assessors to understand the implications of their judgements due to the time needed to perform thorough sensitivity analyses. Ultimately, this latter point could lead to the risk assessors being distrustful of the model and not adopting it in their assessment process.

References

Aleksic, M., Thain, E., Roger, D., Saib, O., Davies, M., Li, J., Aptula, A., and Zazzeroni, R. (2009). "Reactivity Profiling: Covalent Modification of Single Nucleophile Peptides for Skin Sensitization Risk Assessment." *Toxicological Sciences*, 108: 401–411. 172, 180

Andersen, K. and Maibach, H. (1985). *Contact Allergy Predictive Tests in Guinea Pigs, Current Problems in Dermatology*, chapter Guinea pig sensitisation assays: an overview. Chichester: Wiley. 175

Basketter, D. A. (2008). "Skin sensitization: strategies for the assessment and management of risk." *British Journal of Dermatology*, 159: 267–273. 172

Basketter, D. A., Dooms-Goossens, A., Karlberg, A., and Lepoittevin, J. (1995). "The chemistry of contact allergy: why is a molecule allergenic?" *Contact Dermatitis*, 32: 65–73. 172

Chipinda, I., Ajibola, R. O., Morakinyo, M. K., Ruwona, T. B., Simoyi, R. H., and Siegel, P. D. (2010). "Rapid and simple kinetics screening assay for electrophilic dermal sensitizers using nitrobenzenethiol." *Chemical Research in Toxicology*, 23: 918–925. 175

Clemen, R. T., Fischer, G. W., and Winkler, R. L. (2000). "Assessing dependence: Some experimental results." *Management Science*, 46: 1100–1115. 176

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1997). "Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes Linear strategies for large computer experiments." In *Case Studies in Bayesian Statistics III*. New York: Springer. 171

Ellison, C. M., Madden, J. C., Judson, P., and Cronin, M. T. D. (2010). "Using In Silico Tools in a Weight of Evidence Approach to Aid Toxicological Assessment." *Molecular Informatics*, 29: 97–110. 170

Farrow, M., Goldstein, M., and Spiropoulos, T. (1997). "Developing a Bayes linear decision support system for a brewery." In *The Practice of Bayesian Analysis (eds. S. French and J.Q. Smith)*, 71–106. London: Arnold. 171

French, S. (2011). "Aggregating Expert Judgement." *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales*, 105: 181–206. 173

Gerberick, G., Vassallo, J., Bailey, R., Chaney, J., Morrall, S., and Lepoittevin, J.-P. (2004). "Development of a peptide reactivity assay for screening contact allergens." *Toxicological Sciences*, 81: 332–343. 172, 180

Gerberick, G. F., Ryan, C. A., Kimber, I., Dearman, R. J., Lea, L. J., and Basketter, D. A. (2000). "Local lymph node assay: validation assessment for regulatory purposes." *American Journal of Contact Dermatitis*, 11: 3–18. 175

Goldstein, M. and Rougier, J. C. (2009). "Reified Bayesian modelling and inference for physical systems." *Journal of Statistical Planning and Inference*, 139: 1221–39. 173

Goldstein, M. and Shaw, S. (2004). "Bayes linear kinematics and Bayes linear Bayes graphical models." *Biometrika*, 91: 425–446. 169, 171, 182

Goldstein, M. and Wooff, D. A. (2007). *Bayes Linear Statistics: Theory and Methods*. Chichester: Wiley. 169, 171, 180, 183

Gosling, J. P., Hart, A., Mouat, D., Sabirovic, M., Scanlan, S., and Simmons, A. (2012). "Quantifying experts' uncertainty about the future cost of exotic diseases." *Risk Analysis*, 32: 881–893. 173

Jaworska, J., Gabbert, S., and Aldenberg, T. (2010). "Towards optimization of chemical testing under REACH: a Bayesian network approach to Integrated Testing Strategies." *Regulatory Toxicology and Pharmacology*, 57: 157–167. 170

Jaworska, J., Harol, A., Kern, P. S., and Gerberick, G. F. (2011). "Integrating non-animal test information into an adaptive testing strategy - skin sensitization proof of concept case." *Alternatives to Animal Experimentation*, 28: 211–225. 170

Jeffrey, R. C. (1983). *The Logic of Decision*, 2nd ed.. London: University of Chicago Press. 171

O'Hagan, A. (1998). "Eliciting expert beliefs in substantial practical applications." *The Statistician*, 47: 21–35. 176

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. E., Garthwaite, P. H., Jenkinson, D., Oakley, J. E., and Rakow, T. (2006). *Uncertain judgements: eliciting expert probabilities*. Chichester: Wiley. 169

O'Hagan, A., Glennie, E. B., and Beardsall, R. E. (1992). "Subjective modelling and Bayes linear estimation in the UK water industry." *Applied Statistics*, 41: 563–577. 171

Owen, H., Hart, A., Aleksic, M., Aptula, A., Davies, M., Gilmour, N., Li, J., MacKay, C., Safford, R., and Gosling, J. (2012). "A weight-of-evidence approach to skin sensitisation hazard identification using Bayesian belief networks." Technical report, Food and Environment Research Agency. Submitted to *Regulatory Toxicology and Pharmacology*. 172

Revie, M., Bedford, T., and Walls, L. (2010). "Evaluation of elicitation methods to quantify Bayes linear models." *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 224: 322–332. 176

Roberts, D. W. and Natsch, A. (2009). "High Throughput Kinetic Profiling Approach for Covalent Binding to Peptides: Application to Skin Sensitization Potency of Michael Acceptor Electrophiles." *Chemical Research in Toxicology*, 22: 592–603. 175, 178

Safford, R. (2008). "The Dermal Sensitisation Threshold — A TTC approach for allergic contact dermatitis." *Regulatory Toxicology and Pharmacology*, 51: 195–200. 170, 182

Safford, R., Aptula, A., and Gilmour, N. (2011). "Refinement of the Dermal Sensitisation Threshold (DST) approach using a larger dataset and incorporating mechanistic chemistry domains." *Regulatory Toxicology and Pharmacology*, 60: 218–224. 182

Schorr, W. (1975). "Cinnamic aldehyde allergy." *Contact Dermatitis*, 1: 108–111. 177

Stotts, J. (1980). *Current Concepts in Cutaneous Toxicity*, chapter Planning, conduct and interpretation of human predicitive sensitisation patch tests. Washington: Academic Press. 175

Suter, G. and Cormiera, S. (2011). "Why and how to combine evidence in environmental assessments: Weighing evidence and building cases." *Science of the Total Environment*, 409: 1406–1417. 181

Tokoroyama, T. (2010). "Discovery of the Michael Reaction." *European Journal of Organic Chemistry*, 10: 2009–2016. 176

Turner, R., Spiegelhalter, D., Smith, G., and Thompson, S. (2009). "Bias modelling in evidence synthesis." *Journal of the Royal Statistical Society, Series A*, 172: 23–47. 173

Vysochanskij, D. F. and Petunin, Y. I. (1980). "Justification of the 3σ rule for unimodal distributions." *Theory of Probability and Mathematical Statistics*, 21: 25–36. 178

Weed, D. (2005). "Weight of Evidence: A Review of Concept and Methods." *Risk Analysis*, 25: 1545–1557. 181

Acknowledgments

First, we must thank the experts who invested so much time in the elicitation sessions that led to development of the model in this article: Maja Aleksic, Aynur Aptula, Catherine Clapp, Nicola Gilmour, Robert Safford (Unilever), David Roberts (Liverpool John Moores University), and Terry Schultz (University of Tennessee). This work was wholly funded through Unilever as part of their ongoing efforts to develop novel ways of delivering consumer safety. Finally, we thank the editor and referees for their comments that have greatly improved this article.