In Vitro Phototoxicity Testing: Development and Validation of a New Concentration Response Analysis Software and Biostatistical Analyses Related to the Use of Various Prediction Models

Björn Peters and Hermann-Georg Holzhütter

Humboldt-Universität Berlin, Medizinische Fakultät (Charité), Institut für Biochemie, Monbijoustrasse 2, 10117 Berlin, Germany

Summary — As demonstrated in several validation studies, the dermal phototoxic potential of chemicals in humans can be effectively assessed by *in vitro* methods. The core of these methods is to monitor dose–response curves of a chemical in the absence and presence of light, to quantify the difference between these two curves by appropriate measures (either the photo-irritancy factor [PIF], or the mean photo effect [MPE]), and to use these measures as predictors of *in vivo* phototoxicity. We present new concentration–response analysis software for *in vitro* phototoxicity testing, which runs on current personal computers, and takes into account all the limitations identified when using a former program. We also demonstrate the validity and robustness of this new software by applying it retrospectively to all data available from two phases of the EU/COLIPA validation trial for the 3T3 neutral red update *in vitro* phototoxicity test. Some frequently raised questions pertaining to the use of prediction models in phototoxicity testing are addressed, including: the necessity of using prediction models based on a cut-off; whether it is justifiable to use sharp prediction cut-off values; whether there is a biostatistical justification for the highest concentration of the test chemical; and whether repeated testing of a chemical is required.

Key words: *bootstrapping, computer program, dose–response curve, phototoxicity, statistics.*

Introduction: the Historical Background

The dermal phototoxicity of a chemical is defined as a toxic response that is elicited after exposure of skin to the chemical or systemic administration of the chemical, and subsequent exposure to light. As demonstrated in several validation studies (1–3), the phototoxic potential of chemicals can be effectively assessed by *in vitro* methods. In 1996, an OECD workshop recommended an *in vitro* tier-testing approach for phototoxicity assessment (4). In 2000, the Commission of the European Communities put into force *Directive 2000/33/EC*, which introduces the *in vitro* 3T3 neutral red uptake (NRU) phototoxicity test as a validated replacement for testing methods involving the use of laboratory animals. The essence of the *in vitro* 3T3 NRU phototoxicity test is to compare the cytotoxicity of a chemical when tested in the presence and absence of exposure to a non-cytotoxic dose of UVA/visible light. Cytotoxicity is expressed as the concentration-dependent reduction of the uptake of the vital dye neutral red (5), 24 hours after treatment with the chemical.

Two prediction models have been proposed, which differ in the definitions of the measure used to quantify the difference between the concentration–response curves recorded in the presence (+UV) and absence (–UV) of light.

Prediction model 1

The *photo-irritancy factor* (PIF) relates the half-effective concentration value EC50(–UV)[1] of the curve for darkness, to the half-effective concentration value EC50(+UV) of the curve in the presence of light, by means of the following formula:

$$\mathrm{PIF} = \frac{\mathrm{EC50(-UV)}}{\mathrm{EC50(+UV)}} \qquad \text{(Equation 1)}$$

Depending on whether the PIF value is larger or smaller than a properly chosen cut-off value ($\mathrm{PIF_c}$), the chemical is classified as phototoxic or non-phototoxic. A shortcoming of the measure in Equation 1 is that additional *ad hoc* definitions are required

[1]In practical applications, no distinction is normally made between the half-effective concentration value EC50, representing the bio-available concentration of the chemical actually sensed by the biological target, and the half-inhibition concentration IC50, which is the added concentration of the chemical at which the response amounts to 50% of the original value.

to cope with situations where no half-effective concentration values can be derived from the corresponding concentration–response curve: a) if no EC50 value can be derived from one of the two curves, the corresponding EC50 value in Equation 1 is replaced by the highest concentration tested, and the chemical is classified as phototoxic if this modified PIF value is larger than unity; b) if no EC50 value exists for both curves, the chemical is considered non-phototoxic.

Prediction model 2

A second measure of the difference between the dark curve and light curve, the so-called mean photo effect (MPE), was proposed by Holzhütter (6). It aims to overcome the obvious limitations in the application of the PIF, by comparing the two curves at arbitrary doses. The MPE is defined as a weighted average across a set of individual photoeffect values.

$$\mathrm{MPE} = \frac{\sum_{i=1}^{n} w_i \mathrm{PE}_{Ci}}{\sum_{i=1}^{n} w_i} \qquad \text{(Equation 2)}$$

In Equation 2, the *photo effect* (PE_C) at an arbitrary concentration C is defined as the product of the *response effect* (RE_C) and the *dose effect* (DE_C), i.e. $PE_C = RE_C \times DE_C$. The definition is illustrated in Figure 1.

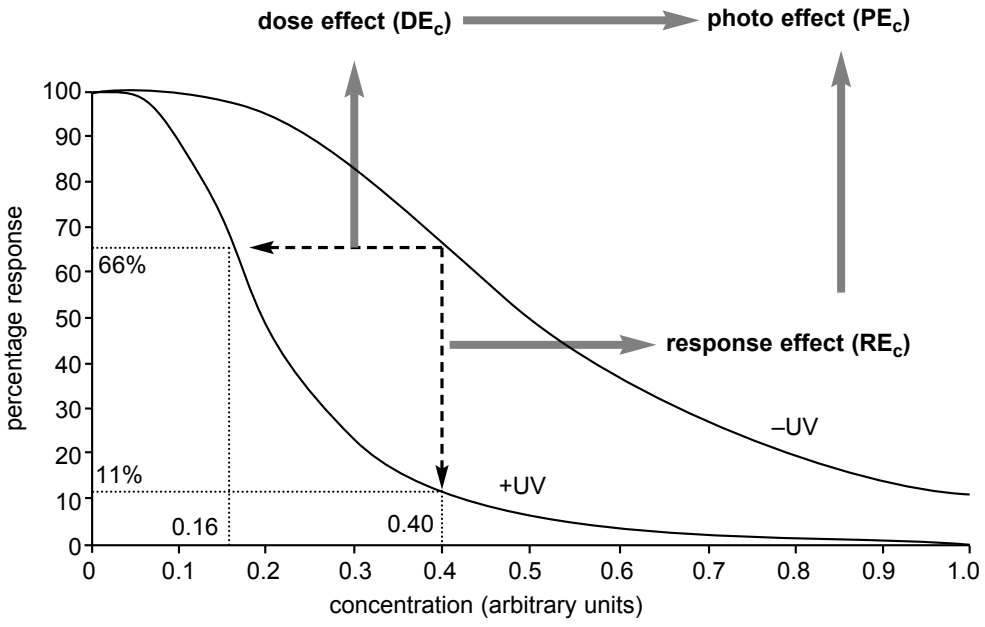
The response effect is the difference between the responses observed in the absence and presence o f light, i.e. $RE_C = R_C(-UV) - R_C(+UV)$. The dose effect is given by the formula:

$$DE_C = \left| \frac{C/C^* - 1}{C/C^* + 1} \right|$$

where C* represents the equivalence concentration to C, i.e. the concentration at which the +UV response equals the –UV response at concentration C. If C* cannot be determined because the response values of the +UV curve are systematically higher or lower than RE_C, the dose effect is set to 1. The weighting factors w_i are given by the highest response value, i.e. $w_i = \mathrm{Max}\{R_i(-UV), R_i(+UV)\}$.

The discrete concentration values C_i, used for the calculation of MPE according to Equation 2, are distributed such that the same number of data points fall into the concentration intervals defined by the concentration values used in the experiment. The calculation of the MPE is restricted to the maximum concentration value at which at least one of the two curves still exhibits a response value of at least 10%. If this maximum concentration is higher than the highest concentration used in the +UV

Figure 1: Illustration for the photo effect calculation



Calculation of the photo effect at the concentration 0.4. Applying the equations given in the text gives: response effect $RE_{0.4}$ = (66% – 11%)/100% = 0.55, dose effect $DE_{0.4}$ =(0.4/0.16 – 1)/(0.4/0.16 + 1) = 0.43, and photo effect $PE_{0.4}$ = 0.24. The mean photo effect is obtained by averaging over the values for the photo effect at various concentrations.

experiment, the residual part of the +UV curve is set to the response value 0. The MPE-based prediction model of *in vivo* phototoxicity classifies a chemical as "phototoxic" if the MPE value is larger than a cut-off value, otherwise the chemical is classified as "non-phototoxic".

Both measures of curve difference, PIF and MPE, represent statistical estimates that have to be derived from the observed concentration–response relations. To facilitate this work, and to harmonise the process of data analysis among various laboratories, in 1996 we developed a computer program, *NRU-PIT2*. It calculates the statistical distribution of PIF and MPE values for given pairs of +UV and –UV concentration–response data. Based on these distributions, and at given cut-off values, the program calculates the probability (p-value) that a test chemical is phototoxic. The *NRU-PIT2* software is a hybrid program, composed of a 16-bit Visual Basic™ dialogue shell, and a fitting module written in Turbo Pascal™, which we adopted from our program *SIMFIT* (7). Its use is restricted to the now-obsolete Windows 3™ group of operating systems. Moreover, intensive use of this program during several international validation studies, as well as in in-house applications, has revealed some weak spots, both in the handling of the program and in the implementation of the MPE-based prediction model. Therefore, we have developed a new computer program, *PHOTOTOX*, which was written in C++, and runs on Windows 95™ or higher operating systems. The major functions of the program are described below, the substantial changes made in comparison with *NRU-PIT2* are discussed, and the overall performance of the new program is assessed. To this end, we have applied the new software to all data compiled for the 3T3 NRU assay during two international studies (2, 3), as well as for the keratinocyte NRU assay (8), which was initially used to determine the cut-off value for the MPE-based prediction model (6). The main objective of this re-analysis was to make sure that the modifications made in the new software do not compromise the results obtained with the old program, *NRU-PIT2*.

We have also addressed some problems that specifically arose from comments on OECD draft Test Guideline *ENV/JM/TG(2001)7: Status Report on Proposals for Test Guidelines on the* In Vitro *Skin Corrosion Test, and the* In Vitro *3T3 NRU Phototoxicity Test* (private communication). The following questions were asked: a) is the use of cut-off based prediction models needed, or is it sufficient to check for a statistically significant difference between the two dose–response curves by using a simple t-test; b) given that we cannot renounce cut-off based prediction models, what are the "optimal" cut-off values for the PIF and the MPE in the light of all available data; c) are there biostatistical arguments for limiting the highest concentration used in the 3T3 NRU assay; and d) what is the benefit of performing at least two independent repeat experiments?

Materials and Methods: Software Design

The main functions of *PHOTOTOX* are depicted in Figure 2. The major changes made in comparison with the old program *NRU-PIT2* are shown in grey boxes.

Data input

The program accepts three modes of data input: a) import of absorbance values from a 96-well plate for arbitrary plate layouts (i.e. arrangement of blanks, controls and wells containing the test chemical), and transformation of these values into dose–response data; b) direct entry of dose–response data as a 2-column data sheet; and c) import of *.ddd-files generated by the old software *NRU-PIT2*.

Bootstrapping

This program module performs a bootstrap resampling of the original concentration–response data (9, 10), which results in a set of new computer-generated concentration–response data, which can be considered as equally probable realisations of the "true" concentration–response data hidden in the experimental observations.

Curve fitting

The central module of the software is a conjugated-gradient minimiser that performs fitting of a continuous model function to the original or bootstrap-resampled sets of discrete concentration–response data. The mathematical concentration–response model used belongs to the large class of polynomials. Nevertheless it has a semi-empirical background, in that it refers to a multi-state compartment system, which upon addition of the test chemical is thought to be driven from its native state (with a response value of 100%) through a series of non-native states with altered response values (11). If the number of compartments is chosen correctly, the model is flexible enough to cope with complicated curve shapes, exhibiting, for example, plateaus or extreme points. A typical shortcoming of polynomial models, however, is their general tendency to exhibit artificial oscillations. Therefore, model fitting to the data is combined with a damping procedure (constraint minimisation). To avoid damping out of real

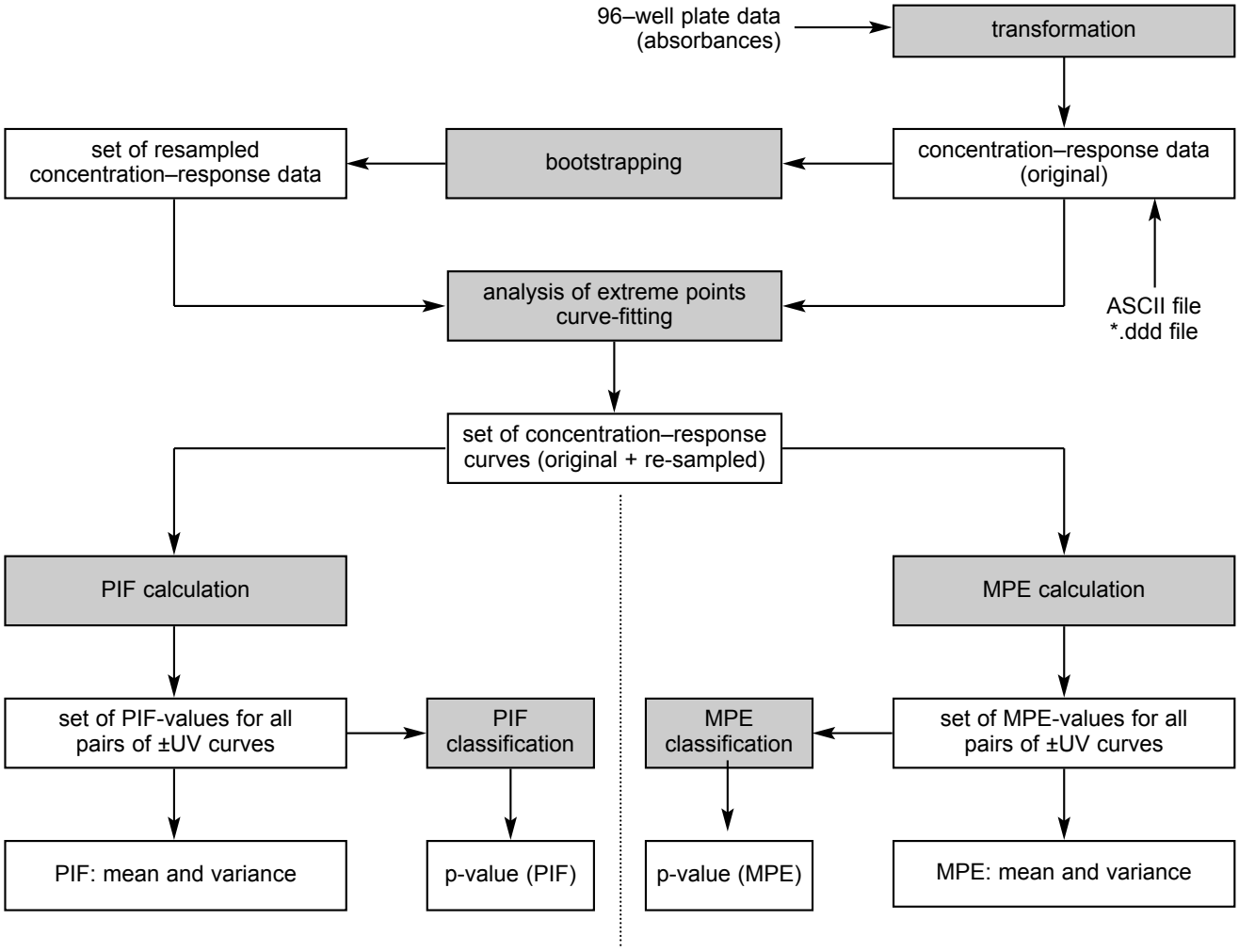
extreme points, curve fitting is preceded by an analysis of extreme points.

In *NRU-PIT2*, an extreme point was defined by a set of three consecutive response values R_{i-1}, R_i, R_{i+1} obeying the condition that the middle value is significantly larger (smaller) than the two neighbouring values. In mathematical terms R_i is a maximum if $R_{i-1} < R_i$ and $R_{i+1} < Ri$, and Ri is a minimum if $R_{i-1} > R_i$ and $R_{i+1} > R_i$. When this definition was used, the program failed to detect smooth extreme points in flat curves. Therefore, we now define an extreme point by a set of consecutive response values R_i, R_{i+1}, … , R_{i+k} ($k \geq 0$), which meet the following two conditions: a) the differences between consecutive points are statistically not significant, i.e. $R_i \approx R_{i+1} \approx \ldots \approx R_{i+k}$; and b) the first and the last response value of this set, R_i and R_{i+k} , are both significantly smaller (maximum) or larger (minimum) than their left and right neighbours, i.e.:

$R_i \; ›/‹ \; R_{i-1}$ and $R_{i+k} \; ›/‹ \; R_{i+k+1}$

Based on the fitted model, the software offers the possibility of calculating the mean value and variance of the effective dose ECx at which the response reaches x% ($0 < x < 100$) of the initial value. The default value of the residual response is x = 50% yielding the common EC50 value used in the calculation of a PIF. This feature makes the program a valuable tool for dose–response analysis beyond phototoxicity testing.

Figure 2: Data processing with *PHOTOTOX*



MPE = mean photo effect; PIF = photo-irritancy factor.

ECx mean values and variances of all dose–response curves contained in the project are tabulated in an extra report sheet. They are automatically updated if the user changes the residual response level x. The default value is x = 50%.

PIF/MPE calculation

Curve fits to the original and bootstrapped data sets provide a bundle of concentration-response curves for a single (+UV or –UV) experiment (Figure 3a). By pairing each curve of the –UV experiment with each curve of the +UV experiment, one arrives at an ensemble of different values for the PIF and the MPE. The statistical distribution of these values (Figure 3b) serves as a basis for the calculation of p-values (see below).

A

The choice of the concentration grid used in the calculation of the MPE according to Equation 2 was changed in the new program. In *NRU-PIT2*, 20 different concentration values were distributed equidistantly along the whole dose interval defined through the highest concentration value common to both data sets under comparison. This choice of the concentration grid has two drawbacks: a) it gives a higher weight to the high-concentration parts of the two curves, if the test concentrations are increased in a geometric series (as in most experiments); and b) it restricts the calculation of the MPE to the concentration interval shared by both experiments. In the new program, the calculation of the MPE is performed across a concentration interval that is defined through the highest concentration value (C_{max}) up to which at least one curve exhibits a response value of 10% or higher. If C_{max} is not reached in one of the two paired experiments (usually in the +UV experiment, because the response values have already dropped to zero at smaller concentrations), the missing responses up to C_{max} are considered insignificant, and are therefore set to 0.

B

The concentration grid is now chosen so that the same numbers of points fall into each concentration interval (defined by the concentration values used in the experiment). In *PHOTOTOX*, the total number of grid points is 50. For example, in Figure 3, the concentration values applied were [0, 1.5625, 3.125, 6.25, 12.5, 25, 50, 100, 200] for the +UV experiment, and [0, 7.8125, 15.625, 31.25, 62.5, 125, 250, 500, 1000] for the –UV experiment. For both experiments, a 500-point dose grid is constructed, containing the same number of equidistant points between each pair of doses used in the experiment. For the +UV experiment, this is: [0, 0.0252, 0.0504, …, 1.5373, 1.5625, 1.5877, …, 3.010, 3.125, 3.175, …, 196.8, 198.4, 200]. The grids for both experiments are merged to a give a 1000-point grid. From the merged grid, 50 points are selected, ranging from 0 to C_{max}. The resulting evaluation grid with C_{max} = 173 is [0.000, 0.328, 0.680, …, 26.5, 29.0, 31.8, …, 143, 158, 173].

C

In *NRU-PIT2*, the weighting factors w_i in Equation 2 were taken as the sum of the response values of the two curves, $w_i = R_i(+UV) + R_i(–UV)$, to reduce the influence of data points with low responses. Because of the changes in the choice of the concentration grid (see *B*), the weighting factors w_i are now defined through the highest response value, $w_i = MAX\{R_i(+UV), R_i(–UV)\}$. This enables a sharper separation of MPE values between phototoxic and non-phototoxic chemicals, as higher weighting is given to that part of the concentration grid where one curve has already dropped while the other still comprises high response values.

PIF/MPE-based classification of phototoxicity

The probability (p-value) of a chemical exhibiting a phototoxic potential in a single ±UV experiment is defined by:

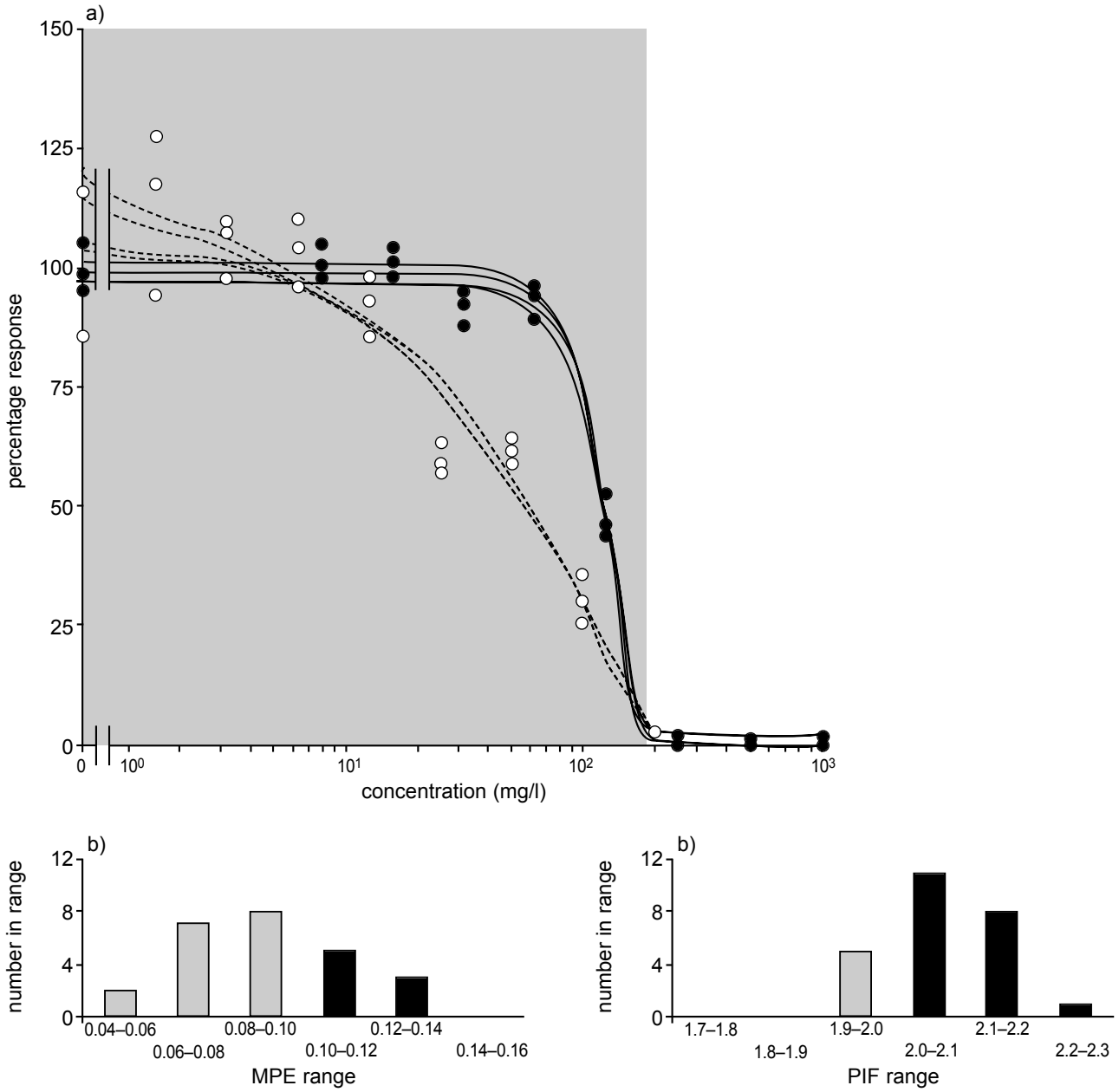
$$p_{tox} = \frac{n_{>} - n_{<}}{n_{>} + n_0 + n_{<}} \qquad \text{(Equation 3)}$$

where $n_{>}$, $n_{<}$ and n_0 denote, respectively, the number of PIF/MPE values within the bootstrap ensemble which are either larger than the cut-off value indicative for phototoxicity (i.e. the toxicity of the chemical is significantly increased in the presence of light), smaller than the cut-off value indicative for a photo-protective effect (i.e. the toxicity of the chemical is significantly decreased in the presence of light), or do not meet the previous two conditions.

A

Note that, in the definition of the phototoxicity probability used in *NRU-PIT2*, the num-

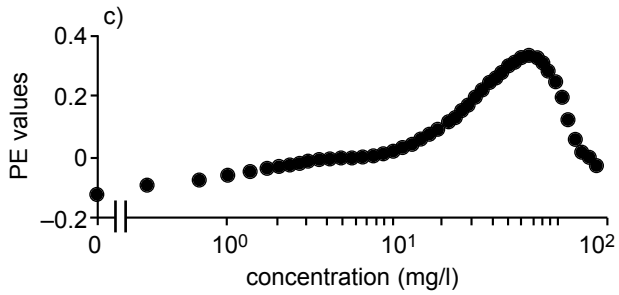
Figure 3: The main functions of *PHOTOTOX*



a) Experimental and theoretical concentration-response data. Five bootstrap concentration-response curves were constructed for both the –UV (continuous lines) and +UV (dotted lines) experiments, by fitting the concentration–response model to novel "synthetic" data sets derived from the original data (solid points) by randomly selecting response values with replacement from the original group of six replicates at each concentration value. The calculation of mean photo effect (MPE) was restricted to the maximal concentration C_{max} = 173 concentration units (shaded area) at which the –UV curve falls below 10% of the maximal response. To include the control values (concentration = 0) on an otherwise logarithmic axis, the concentration axis consists of two parts. The scaling between 0 value and the first concentration value (here, 1.56) is linear, whereas the rest of the axis is scaled logarithmically. The length of the linear part of the axis is chosen to be equal to the distance between the first and the second concentration value in the logarithmic part. Thus, a geometric concentration series, including a control concentration of 0, will be displayed as equidistant. The shaded box indicates the range considered for curve comparison.

b) Statistical distribution of photo-irritancy factor (PIF) and MPE values. Values of the PIF and the MPE were calculated for all 25 possible pairs of the five –UV and five +UV bootstrap curves shown in Figure 3a. Calculation of p-values according to Equation 3 gives p_{MPE} = 8/25 = 0.32. and p_{PIF} = 2/25 = 0.80, i.e. the MPE classifies the chemical as non-phototoxic ($p < 0.5$), whereas the PIF classifies it as phototoxic ($p > 0.5$).

Figure 3: continued



c) *Concentration-dependent photo effect. The plot shows how the photo effect changes over the range of concentrations used in the experiment shown in Figure 3a. At each concentration value, the depicted photo effect is the average of all 25 possible pairs that can be combined from the 5 –UV and +UV bootstrap curves. A significant photo effect (PE values > MPE_c = 0.1) occurs at concentrations between 20 and 120 concentration units. The concentration axis is scaled by using a modified log axis (compare legend to Figure 3a).*

ber of cases indicating a photo-protective effect ($n_<$) has not been considered in the numerator of expression (Equation 3), i.e. the p-value was computed as the relative number of bootstrap pairs yielding a PIF or MPE value larger than a given cut-off value. Such a classification scheme results in a systematic bias toward higher probability values, because all curve pairs indicating the presence of a phototoxic potential enlarge the p-value, whereas all curve pairs indicating the presence of a photo-protective effect, do not reduce the p-value. At the extreme, positive p-values could be produced by chance, even if the mean dose–response data for both experiments were identical, but large errors in the measured response data gave rise to single values of PE or PIF (derived from single curve pairs) exceeding the cut-off value.

B

Because of the low calculation speed of *NRU-PIT2*, each bootstrap curve of the +UV experiment was paired with only a single (arbitrarily chosen) curve of the –UV experiment. In contrast, the new program provides more-reliable statistical estimates, by including all possible pairs of bootstrap curves. For example, if ten bootstrap curves are generated for each experiment, 100 curve pairs are included in the computation of classification probabilities (see below).

The cut-off values for photo-protection are chosen as $1/PIF_c$ and $–MPE_c$, where PIF_c and MPE_c are the respective cut-off values indicative of phototoxicity. Averaging across the p-values obtained in several independent experiments (runs) provides a mean p-value for both prediction models, which can be employed for decision making.

Concentration-dependent photo effect

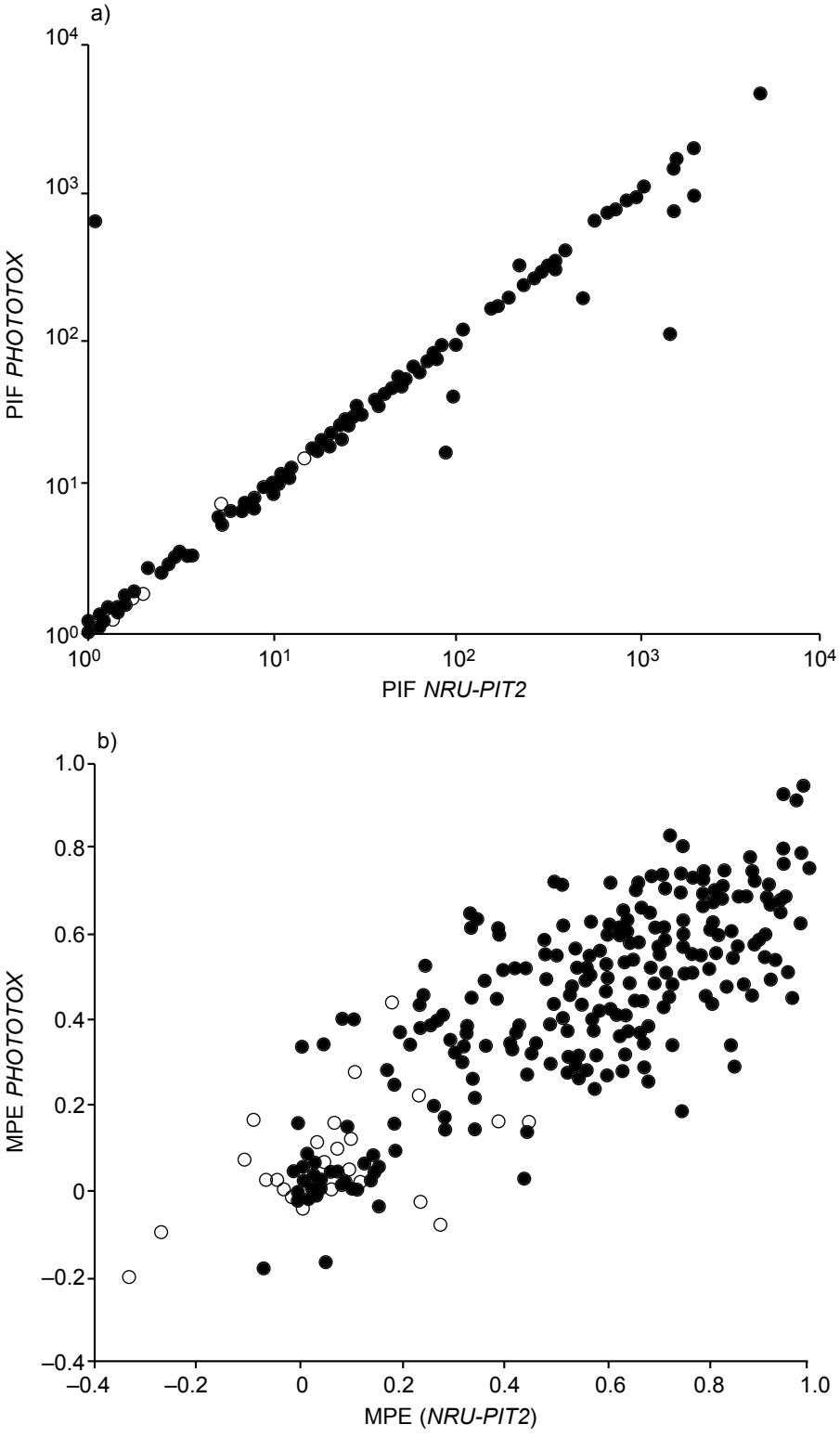
The value of the MPE represents a weighted average across individual photo effect values calculated at various concentrations of the test chemical. Both MPE and PIF measure the *difference* between the +UV and –UV curves without making explicit reference to the concentration range over which the light-induced difference between the curves is relevant. To visualise the concentration range for which significant photo effects can be expected, the program provides a concentration-dependent plot of the photo effect, as well as of the two effects (response effect and dose effect) contributing to it (Figure 3c).

Results: Software Validation

Database

As outlined above, the new program *PHOTOTOX* comprises a number of modifications of the calculation procedure for MPE and of the curve-fitting to the data, in comparison with the old program *NRU-PIT2*. To investigate the consequences of these modifications for numerical PIF and MPE values and the related *in vivo* classifications of phototoxicity, we have re-analysed the experimental data gathered for the 3T3 NRU phototoxicity test during two international studies (2, 3) and for the keratinocyte NRU phototoxicity test (8; see Table 1). The latter database was used to define the cut-off value for the MPE-based prediction model in our previous work (6). Compiling the original measurements gathered in the first international study (2), we encountered technical difficulties in extracting data from compressed files from one laboratory. Therefore, these data could not be used in the re-analysis. Moreover, we did not include the data for one chemical tested in two studies (2) and (8), for which no *in vivo* toxicity was reported. The whole database used in the software validation comprised 635 different pairs of dose–response curves. Generally, each laboratory recorded two independent pairs of +UV/–UV curves per chemical (replicates), but sometimes only one, while sometimes, three pairs were recorded. By averaging the results for the replicates, 322 classification results are obtained. If no other database is explicitly men-

Figure 4: Comparison of PIF and MPE values obtained by application of the two computer programs *NRU-PIT2* and *PHOTOTOX*



● = in vivo *phototoxic chemical;* ○ = in vivo *non-phototoxic chemical*

a) Photo-irritancy factor (PIF). Only those PIF values obtained from curve pairs with two existing EC50 values are plotted. Measure of determination: r^2 *=0.992. To understand how the striking differences between some PIF values (for example. PIF* NRU-PIT2 ≈ *1, PIF* PHOTOTOX ≈ *1000) could be generated, we examined the original data. These do not support the PIF values published for* NRU-PIT2*, so we have to conclude that errors were made when manually entering the concentration values for those curves. b) Mean photo effect (MPE).*

Measure of determination: r^2 *=0.831.*

Table 1: Database used for the validation of the new software and for the computations related to statistical issues with prediction models

	EU/COLIPA: phase I (2)	EU/COLIPA: phase II (8)	Study on UV filter chemicals (3)
Number of test chemicals	29	29	20
Number of laboratories	8	1	4
Number of chemicals/laboratory for which at least one curve pair was available	29, 29, 28, 29, 16, 29, 26, 29	29	20, 19, 20, 19

EU/COLIPA = international EU/COLIPA in vitro *phototoxicity validation study.*

tioned below, the results will refer to the database shown in Table 1.

Comparing the old and the new software

We have checked whether the changes in the software give rise to significantly different *in vivo* classifications of phototoxicity compared with those obtained with the old software *NRU-PIT2*. As can be seen in Table 2, this is not the case. For 302 out of 322 MPE-based classifications, and for 313 out of 322 PIF-based classifications, the two versions of the software provided identical results. The total misclassification rates obtained by means of the new software were 30/322 = 9.3% for PIF and 26/322 = 8.1% for MPE: these are marginally smaller than the total misclassification rates of 31/322 = 9.6% (PIF) and 28/322 = 8.7% (MPE) obtained when using the old software. Instead of comparing classification results, a more-sensitive detection of differences between the performance of the two software versions is achieved by comparing directly "old" and "new" PIF and MPE values (Figure 5). The concordance between old and new PIF values is higher ($r^2 = 0.992$) than for the MPE values ($r^2 = 0.831$), because the estimate of EC50 values, and hence the estimate of the PIF value, is less sensitive to slight changes in the shapes of the fitted dose–response curves. The few striking differences between old and new PIF values turned out to be caused by incorrect handling of the corresponding raw data in the previous analysis made with *NRU-PIT2*.

Table 2: Comparison of classification results obtained by the old software (*NRU-PIT2)* and the new software (*PHOTOTOX)*

	Classification by *PHOTOTOX*					
Classification	**MPE**	true	false	**PIF**	true	false
by	true	**285**	9	true	**287**	4
NRU-PIT2	false	11	**17**	false	5	**26**

MPE = mean photo effect; PIF = photo-irritancy factor.

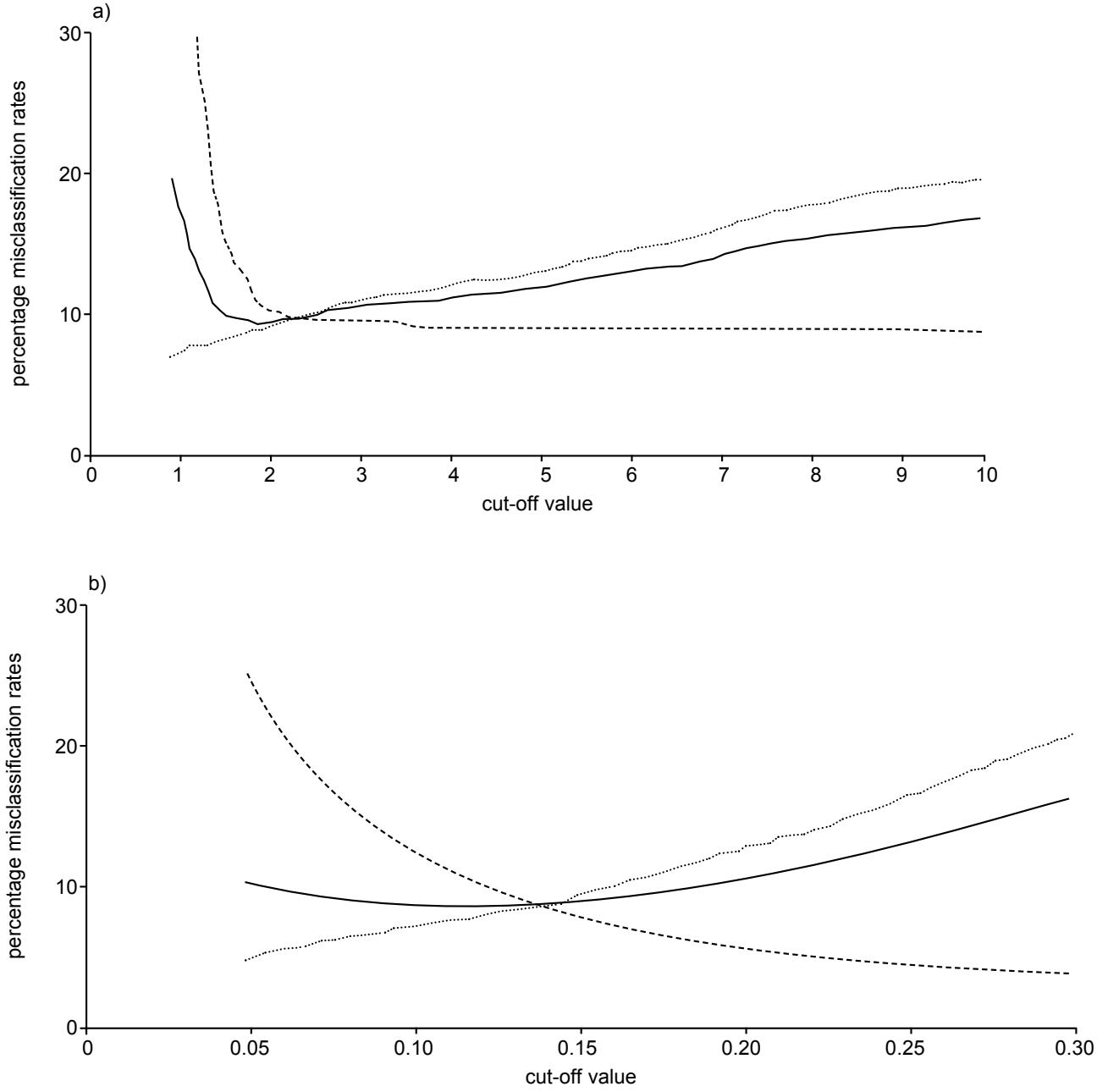
Results: Biostatistical Issues Related to the Use of Prediction Models

The following results were obtained in attempts to respond to typical questions raised by users of the program, and in helping them to minimise the number of misclassifications in future applications of the software. Moreover, we have responded to some critical remarks and suggestions made by international experts in their comments on the OECD draft Test Guideline *ENV/JM/TG(2001)7 Status Report on Proposals for Test Guidelines on the* In Vitro *Skin Corrosion Test, and the* In Vitro *3T3 NRU Phototoxicity Test* (private communication).

Optimal cut-off values for PIF and MPE

Figure 5 shows the percentage of total misclassifications, false-negative classifications and false-positive classifications at varying cut-off values of PIF and MPE. Putting together the classifications obtained for the three databases, a minimum of misclassifications is achieved when choosing $PIF_{cut\text{-}off} = 2$ and $MPE_{cut\text{-}off} = 0.12$. Hitherto, for the PIF, a cut-off value of 5 was stipulated in several international studies, and was also recommended in the Standard Operating Procedures (SOPs) of the 3T3 NRU phototoxicity test. This cut-off value was pro-

Figure 5: Misclassification rates at varying cut-off values for photo-irritancy factor (PIF) and mean photo effect (MPE)



a) PIF; b) MPE.

The rate of false-positive classifications is given as the percentage of misclassifications within the group of in vivo *non-phototoxic chemicals. The rate of false-negative classifications is given as the percentage of misclassifications within the group of* in vivo *phototoxic chemicals.*

------ *average probability of a non phototoxic chemical to be classified as phototoxic;*

..... *average probability of a phototoxic chemical to be classified as phototoxic;*

—— *average misclassification probability.*

Table 3: Predictions of *in vivo* phototoxicity based on a prediction model which classifies a chemical as phototoxic if the EC(+UV) value is significantly smaller than EC(–UV) value according to Student's t test

	$\alpha = 0.05$ ($t_c = 1.73$)			$\alpha = 0.001$ ($t_c = 3.61$)			$\alpha = 9.5E\text{-}19$ ($t_c = 3.61$)		
t test results		***in vivo* toxic**	***in vivo* not toxic**		**toxic**	**not toxic**		**toxic**	**not toxic**
Classified	toxic	**436**	104	toxic	**434**	100	toxic	**422**	44
	not toxic	31	**64**	not toxic	33	**68**	not toxic	45	**124**

EC(+UV) = half-effective concentration in the presence of UV/visible radiation; EC(–UV) = half-effective concentration in the absence of UV/visible radiation.

posed on the basis of a prevalidation study carried out in 1994 (1), and has not been submitted to critical revision since then. In the light of all the relevant data now available, a cut-off value of about 2 seems more appropriate. The optimal cut-off value of 0.12 determined for the MPE is very close to the previous cut-off value of 0.1, which was proposed by Holzhütter (6), on the basis of data obtained by a keratinocyte NRU phototoxicity test (8). We conclude that there is no necessity to alter this historical cut-off value for the MPE.

Is a statistically significant difference between EC50(–UV) and EC50(+UV) a reliable indicator of a phototoxic effect?

We analysed whether a statistically significant difference between EC50(–UV) and EC50(+UV) is a reliable indicator for the phototoxic potential of a chemical. To this end, we applied Student's t test, according to which the two EC50 values have to be considered different at the level of confidence $(1 - \alpha)$, if the test quantity:

$$t = \frac{|EC50(-UV) - EC50(+UV)|}{\sqrt{\sigma^2_{(-)} + \sigma^2_{(+)}}} \sqrt{n} \quad \text{(Equation 4)}$$

Table 4: Concordance and discordance of classifications made by the photoirritancy factor (PIF) and the mean photo effect (MPE) for all 635 runs of the database

		PIF	
		true	false
MPE	true	**555**	23
MPE	false	22	**35**

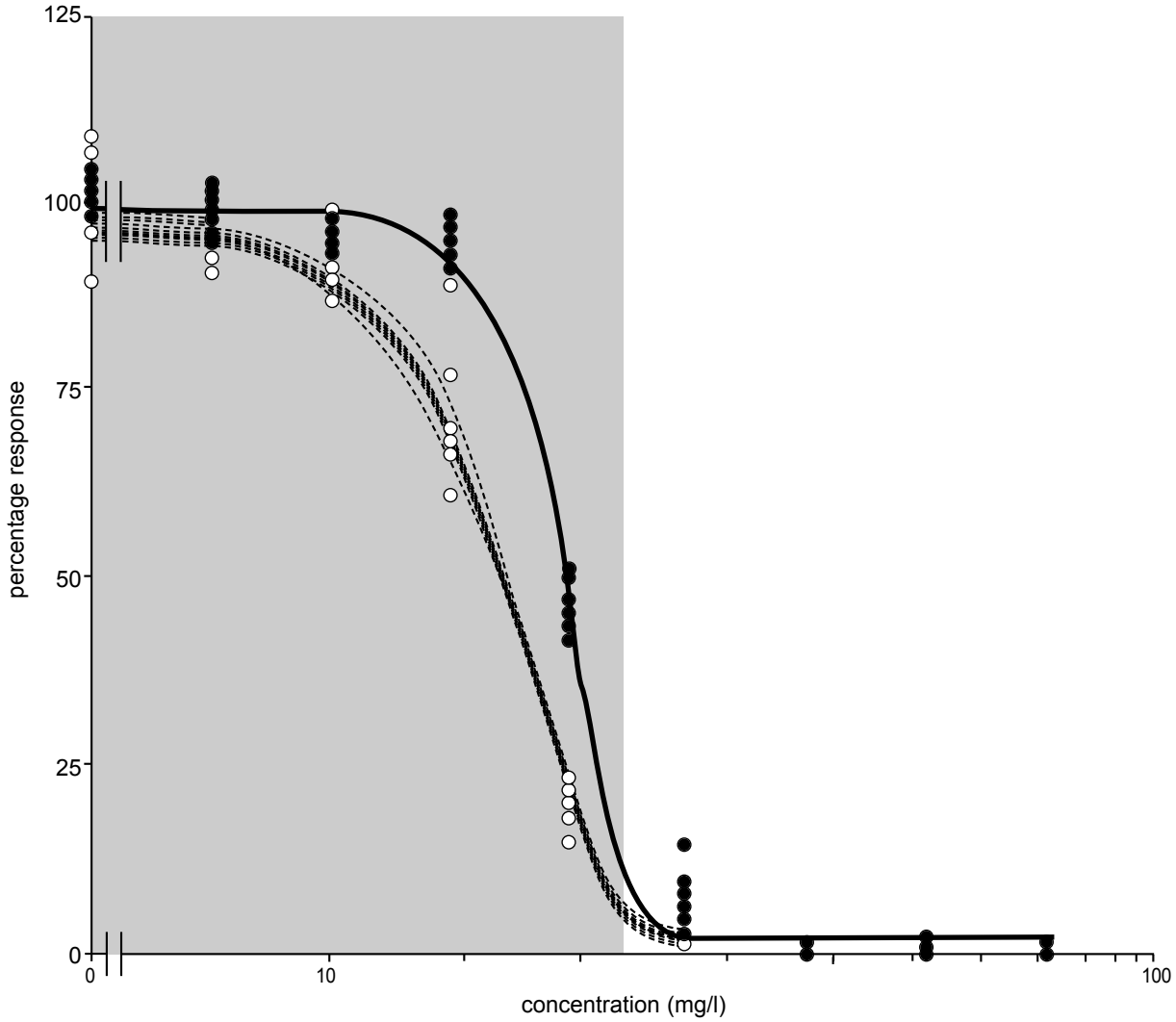
is larger than a critical table value t_c, which depends upon the chosen error probability α. In Equation 4, $\sigma_{(\pm)}$ denotes the variance of the corresponding EC50 value, and n is the number of individual replicates of EC50 used to calculate its mean value and variance. In our calculations, we have used $n = 20$ replicates derived from bootstrap curves. Where one of the EC50 values was missing, the classification scheme for the PIF (see above) was used.

Table 3 depicts the outcome of the t-test for all pairs of ±UV curves. At two common confidence levels of $\alpha = 0.05$ (5% error probability) and $\alpha = 0.001$ (0.1% error probability), most of the related EC50(±UV) values are assessed to be different from each other. This is because the variance of the EC50 value is usually very small, so that a slight difference between EC50(–UV) and EC50(+UV) produces a large t-value. As a consequence, most of the *in vivo* non-toxic chemicals would be misclassified as "phototoxic", if such classification was based solely on a statistically significant difference between EC50(–UV) and EC50(+UV) (see Figure 6 for an example). Because it might be argued that the table statistics t_c for Student's t test at $\alpha = 0.05$ and $\alpha = 0.001$ were still too small for effectively discriminating equal EC50 values from different ones, we have systematically searched for an optimal value of t_c which minimises the misclassification rate. This resulted in an optimal value of $t_c = 37$, which corresponds to an error probability of $\alpha = 9.5 \times 10^{-19}$! Apart from the credibility of such an extreme statistical figure, the minimal misclassification rate is still 14%, i.e. significantly above the misclassification rates reached with the PIF or the MPE. Hence, a statistically significant difference between EC50(–UV) and EC50(+UV) is not sufficient for the identification of a phototoxic potential.

Parallel use of the PIF and the MPE?

Considering the rather complicated mathematical structure of the MPE, it has been frequently asked

Figure 6: Statistical difference between related ±UV curves is not a suitable measure of phototoxicity



The ±UV concentration response curves shown were recorded for a non-phototoxic chemical (sodium lauryl sulphate). The bootstrapping procedure yields (18.98 ± 0.12) and (16.0 ± 0.15) for the mean and variance of the two EC50 values. The corresponding t-value is 68, i.e. from the statistical point of view, the two EC50 values are significantly different at an extremely high confidence level of 1 – α = 0.999999…. Despite this statistical difference, both photo-irritancy factor (= 1.2) and mean photo effect (= 0.066) provide a correct prediction (absence of phototoxic potential) for this chemical. The concentration axis is scaled by using a modified log axis (compare with legend to Figure 3a).

● = *data point –UV experiment*

○ = *data point +UV experiment*

—— = *bootstrap curve –UV experiment*

- - - = *bootstrap curve +UV experiment*

▭ = *range considered for curve comparison*

whether the parallel use of two different measures, the PIF and the MPE, is really required. To this end, we have analysed the extent to which two measures are redundant, i.e. the extent to which they have led to identical classifications (see Table 4). If the classification is made for all 635 runs of the database, the PIF and the MPE provide nearly identical portions of false classifications (9.1% and

9.0%, respectively). However, accepting only those classifications consistently made by the MPE and the PIF, the misclassification rate is considerably smaller: 35/590 = 5.9%. In this case, however, there remain 45 conflicting classifications, which require additional testing before their phototoxic potential can be decided upon. If the conflict of unequal classifications by the PIF and the MPE cannot be resolved by repeated testing in the 3T3 NRU phototoxicity test, other *in vitro* tests (for example, skin tests) and all information available on the physical and chemical properties of the relevant chemicals should be examined. Contradictory classifications occurred typically for two curve constellations (Figure 7):

— Case 1: Both curves do not drop down to response values below 50%, but nevertheless clearly diverge. This is a typical constellation if the highest tested concentration was too low. According to the PIF-based prediction model, this automatically entails the classification "non-phototoxic", whereas the MPE value is larger than the cut-off value, and thus leads to the classification "phototoxic". Here, a repeat experiment could clarify whether the difference between the +UV and –UV curve was reproducible.

— Case 2: Only the +UV curve drops below 50%, and the estimated EC50(+UV) value is close to the maximum dose used in the –UV experiment. In this case, the PIF-based prediction model inevitably classifies the substance as "phototoxic", whereas the MPE value is smaller than the cut-off value, because the discrepancy between the curves occurs only over a rather narrow concentration range. Here, the dose range for the repeat experiment should be carefully chosen. If the initial dose range is found to be definitely too narrow, it should be extended to increase the chance of getting an EC50(–UV) estimate. If, however, the EC50(+UV) value is found to be in a dose range above the non-physiological threshold of 1000μg/ml, it would be advisable to reduce the highest dose tested (see also Case 1).

How many independent experiments (runs)?

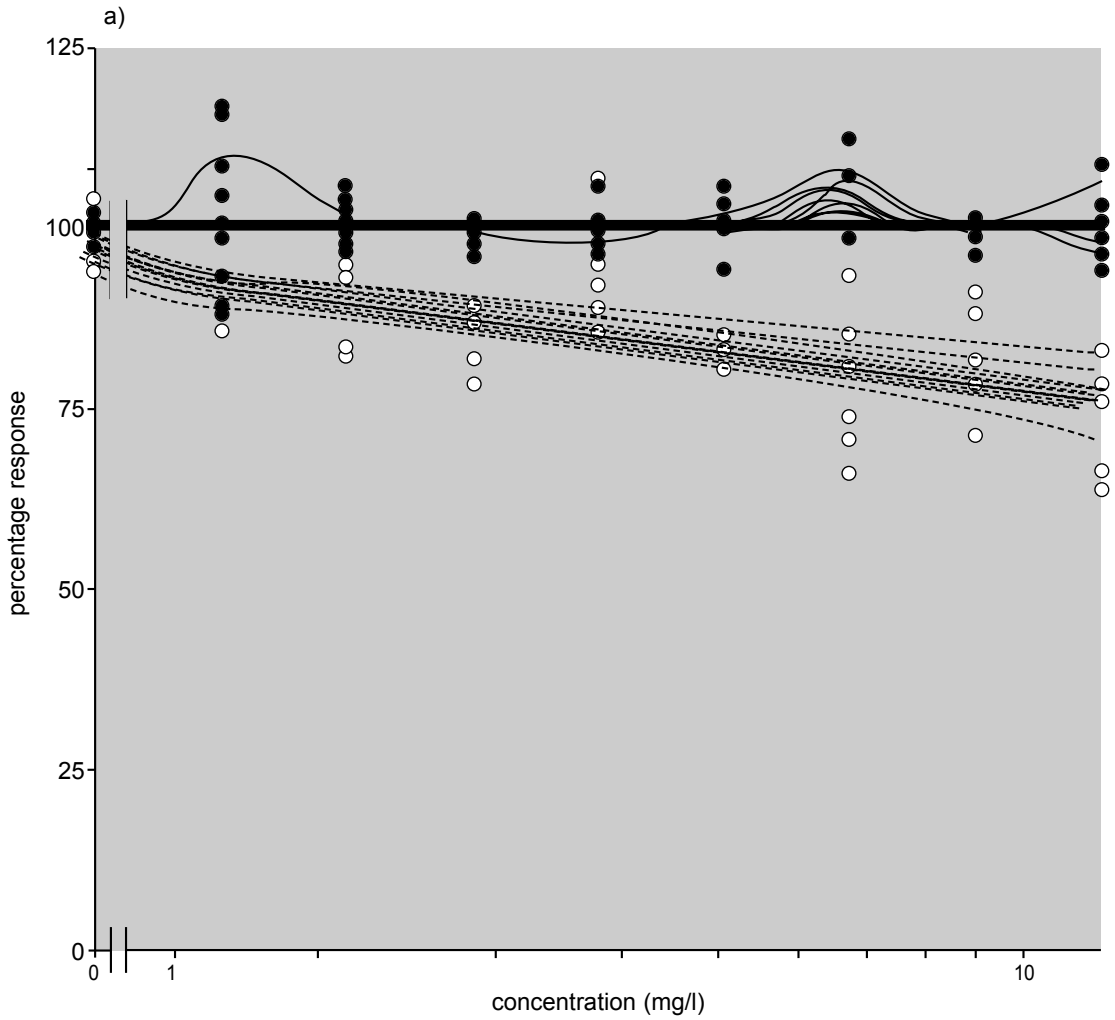
In previous validation studies, the final classification was achieved by taking the mean PIF and MPE values of two independent experiments (runs) and comparing them with the corresponding cut-off values. The question arises whether the classifications based on two runs are significantly better than those based on only a single run. Table 5 shows the number of cases where the classification of the *in vivo* phototoxic potential was either correct in all runs, in conflict, or incorrect in all runs. There are 15/322 = 4.7% cases of conflicting runs for the PIF, and 20/322 = 6.2% cases of conflicting runs for the MPE, i.e. by using the PIF or the MPE, the likelihood that a second run will confirm the classification result of the first run is about 95%. Intriguingly, the misclassification rates obtained on the basis of individual runs (9.1% PIF and 9.0% MPE, see Table 4) are very close to those obtained by averaging across runs (9.3% and 8.1%, see Table 2). This means that making the final decision by merely averaging across the probability values of two conflicting runs does not improve the quality of the classifications. However, the benefit of performing a second run is to increase the confidence in the classification result.

If only those classifications are accepted in which both the PIF and the MPE provided consistent classifications in two runs (= 271 + 12 cases in Table 5), the misclassification rate is reduced to 4.2%. Again, 39 "unclear" classifications remain, which have to be submitted to further testing. It has to be noted that the chance of arriving at the correct classification in further testing increases considerably, if the external conditions of the experiment (such as treatment of cells, use of solvents, and variation of

Table 5: Comparison of cases where classifications based on photo-irritancy factor (PIF) or mean photo effect (MPE) were consistent for all runs, conflicting, or incorrect for all runs

		PIF		
		consistently classified in all runs	conflicting classifications	misclassified in all runs
MPE	consistently classified in all runs	271	6	6
MPE	conflicting classifications	12	4	4
MPE	misclassified in all runs	2	5	12

Figure 7: Typical curve constellations for which conflicting photo-irritancy factor (PIF), and mean photo effect (MPE), classifications may arise



the dose range) are deliberately varied to a reasonable extent. This is clearly seen from Table 6, which presents the overall classifications obtained by lumping together the results of all the participating laboratories in a study.

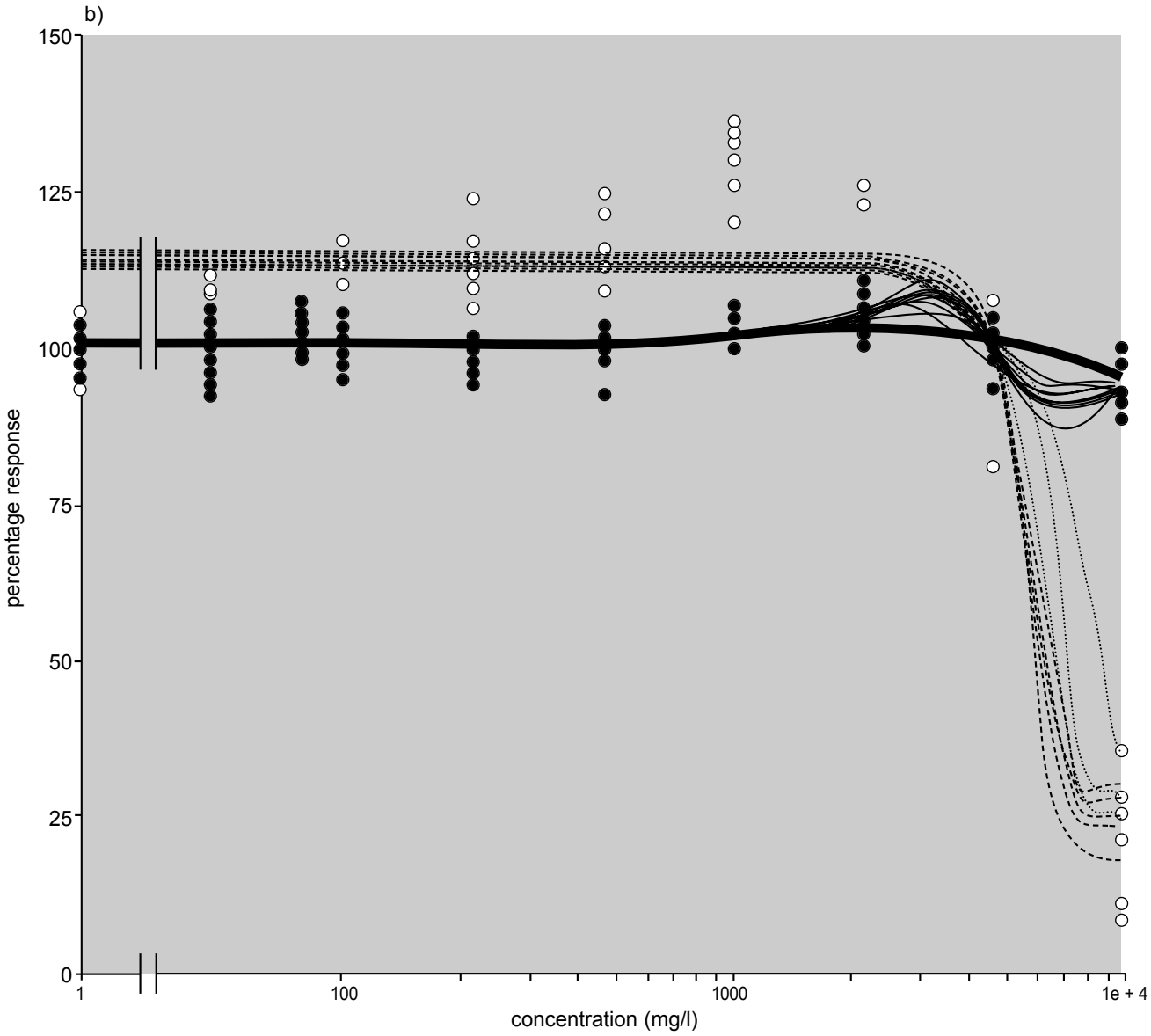
Depending on whether the classification is made by comparing the mean PIF or MPE values with the corresponding cut-off value (as carried out in previous statistical analyses of phase II and III), or in a biostatistically more reliable manner by using the p-value (see "PIF/MPE-based classification of phototoxicity", above) there remain only 3 or 1 (!) misclassifications. Based on mean p-values for the 29 chemicals of the phase II study, an ambiguous classification remains only for the chemical furosemide[1] for both prediction models (p-value = 0.43 for PIF, p-value = 0.54 for MPE). The remaining 28 chemicals are correctly classified. For the 20 chemicals of the phase III study, conflicting p-values remain only for the *in vivo* non-phototoxic chemical terephthalylidene dicamphorsulphonic acid/salts (p-value = 0.63 for PIF, p-value = 0.09 for MPE). Altogether, only one chemical (furosemide) would be misclassified by averaging across the results of independent laboratories

Confining the highest dose tested to an upper limit?

The new software was applied to elucidate the influence of the highest test concentration used on the

[1]It should be noted that expert opinions on the true *in vivo* phototoxic potential of furosemide are divergent, because of contradictory clinical and experimental data.

Figure 7: continued



*Two typical examples for conflicting PIF and MPE classifications. The chemicals used are a) bergamot oil (*in vivo *phototoxic) and b) terephthalidene dicamphor sulfonic acid (*in vivo *non-phototoxic). The concentration axis is scaled using a modified log axis (compare with legend to Figure 3a).*

● = *data point –UV experiment*

○ = *data point +UV experiment*

—— = *bootstrap curve –UV experiment*

- - - = *bootstrap curve +UV experiment*

▭ = *range considered for curve comparison*

Table 6: Classification of the *in vivo* phototoxic potential of 40 chemicals based on average values across all the participating laboratories in a study

			Phase II							Phase III					
			PIF			MPE				PIF			MPE		
Chemical	***In vivo***	**No. runs**	**Mean PIF**	**p-value**	**No. false classified runs**	**Mean MPE**	**p-value × 100%**	**No. false classified runs**	**No. runs**	**Mean PIF**	**p-value**	**No. false classified runs**	**Mean**	**p-value × 100%**	**No. false classified runs**
p-Aminobenzoic acid (PABA)	not pt	16	**2.0**	20%	3	**0.14**	3.2%	6	—	—	—	—	—	—	
Benzophenone	not pt	—	—	—	—	—	—		8	1.5	25%	2	0.05	23%	2
3-Benzylidene camphor	not pt	—	—	—	—	—	—		9	1.1	4%	0	0.01	0%	0
Benzylidene camphor sulphonic acid	not pt	—	—	—	—	—	—		8	1.2	0%	0	0.02	4%	0
Chlorhexidine dihydrochloride	not pt	16	1.2	19%	3	0.04	19%	3	—	—	—	—	—	—	
Hexachlorophene	not pt	16	1.5	7%	1	0.04	7%	1	—	—	—	—	—	—	
L-Histidine-free base	not pt	—	—	—	—	—	—		8	1.0	0%	0	–0.07	4%	0
4-Methylbenzylidene camphor	not pt	—	—	—	—	—	—		8	1.2	0%	0	0.02	0%	0
Octyl methoxycinnamate	not pt	—	—	—	—	—	—		8	1.0	0%	0	0.02	2%	0
Octyl salicylate	not pt	—	—	—	—	—	—		8	1.0	0%	0	0.03	1%	0
Penicillin G	not pt	14	**3.2**	21%	3	0.09	29%	4	—	—	—	—	—	—	
Polyacrylamidomethyl b.c.	not pt	—	—	—	—	—	—		8	1.0	0%	0	0.01	10%	1
Sodium lauryl sulphate	not pt	16	1.2	1%	0	0.05	17%	2	7	1.3	1%	0	0.01	0%	0
Terephthalidene dicamphor s.a.	not pt	—	—	—	—	—	—		8	1.6	**63%**	5	0.01	9%	1
Acridine, free base	pt	11	605.7	100%	0	0.56	100%	0	—	—	—	—	—	—	
Acridine hydrochloride	pt	13	1033.6	100%	0	0.56	100%	0	8	545.7	100%	0	0.60	100%	0
Amiodarone	pt	15	4.6	78%	3	0.22	75%	4	7	9.3	100%	0	0.41	100%	0
Anthracene	pt	14	409.2	79%	3	0.49	71%	4	8	190.3	100%	0	0.59	100%	0
Bergamot oil	pt	16	15.2	79%	3	0.37	91%	2	—	—	—	—	—	—	
Bithionol	pt	16	16.2	100%	0	0.42	100%	0	8	15.4	100%	0	0.36	100%	0
Chlorpromazine	pt	16	65.3	100%	0	0.51	100%	0	8	36.0	100%	0	0.44	100%	0
Demeclocycline	pt	15	356.7	100%	0	0.51	100%	0	8	295.8	73%	2	0.39	75%	2
Fenofibrate	pt	14	48.3	93%	1	0.39	86%	2	—	—	—	—	—	—	
Furosemide	pt	14	91.8	**43%**	8	0.15	54%	6	—	—	—	—	—	—	

Table 6: continued

		Phase II							Phase III					
			PIF			MPE				PIF			MPE		
Chemical	***In vivo***	**No. runs**	**Mean PIF**	**p-value**	**No. false classified runs**	**Mean MPE**	**p-value × 100%**	**No. false classified runs**	**No. runs**	**Mean PIF**	**p-value**	**No. false classified runs**	**Mean**	**p-value × 100%**	**No. false classified runs**
Ketoprofen	pt	14	1188.1	99%	0	0.53	87%	2	8	405.9	100%	0	0.57	100%	0
5-Methoxypsoralene (5-MOP)	pt	13	39.3	80%	3	0.37	85%	2	—	—	—	—	—	—	
6-Methylcoumarine	pt	14	96.6	100%	0	0.54	100%	0	—	—	—	—	—	—	
Musk ambrette	pt	14	19.2	86%	2	0.34	86%	2	9	33.4	100%	0	0.46	100%	0
Nalidixic acid, free acid	pt	14	25.7	64%	5	0.31	88%	2	—	—	—	—	—	—	
Nalidixic acid, sodium salt	pt	16	14.5	100%	0	0.41	100%	0	—	—	—	—	—	—	
Neutral red	pt	14	10,773.9	99%	0	0.96	100%	0	—	—	—	—	—	—	
Norfloxacin	pt	14	28.5	100%	0	0.43	100%	0	—	—	—	—	—	—	
Ofloxacin	pt	13	7.6	77%	3	0.41	86%	2	—	—	—	—	—	—	
Promethazine	pt	16	59.5	100%	0	0.61	100%	0	—	—	—	—	—	—	
Promethazine hydrochloride	pt	—	—	—	—	—	—		8	81.3	100%	0	0.50	100%	0
Protoporphyrine IX, free acid	pt	13	20,356.0	100%	0	0.69	100%	0	—	—	—	—	—	—	
Protoporphyrine IX, sodium salt	pt	15	5788.7	100%	0	0.65	100%	0	3	118.9	100%	0	0.64	100%	0
Rose bengal	pt	16	57.6	100%	0	0.51	100%	0	—	—	—	—	—	—	
Tiaprofenic acid	pt	14	7172.1	93%	1	0.65	100%	0	—	—	—	—	—	—	

MPE = mean photo effect; PIF = photo-irritancy factor; pt = phototoxic.

Bold-face figures indicate conflicting cases (discordance between phototoxicity reported in vivo *and phototoxicity assessed by* in vitro *method).*

quality of predictions. In this case, the analysis was confined to an internal study on UV-filter chemicals, because this was the only study organised by the management team to ensure a consistent use of solvents by the participating laboratories. The dose interval for the calculation of PIF and MPE values was restricted to maximal concentrations of 100µg/ml and 1000µg/ml, respectively. For both limitations, the rate of false-positive classifications (MPE 1.4%, PIF 0.5%) was significantly lower than the rate of false-positive classifications obtained without dose limitations (MPE 5.2%, PIF 9.3%). The rate of false-negative classifications (MPE 2.7%, PIF: 2.9%) was not affected by variations in the highest test dose. Hence, testing at doses higher than 1000µg/ml seems to increase the risk of false-positive classifications.

Received 26.11.01; received in final form 23.4.02; accepted for publication 26.4.02.

References

1. Spielmann, H., Balls, M., Döring, B., Holzhütter, H.G., Kalweit, S., Klecak, G., L'Eplattenier, H., Liebsch, M., Lovell, W.W., Maurer, T., Moldenhauer, F., Moore, L., Pape, W., Pfannenbecker, U., Potthast, J., De Silva, O., Steiling, W. & Willshaw, A. (1994). EEC/COLIPA project on *in vitro* phototoxicity testing: first results obtained with a Balb/c 3T3 cell phototoxicity assay. *Toxicology in Vitro* **8**, 793–796.
2. Spielmann, H., Balls, M., Dupuis, J., Pape, W.J.W., Pechovitch, G., De Silva, O., Holzhütter, H-G., Clothier, R., Desolle, P., Gerberick, F., Liebsch, M., Lovell, W.W., Maurer, T., Pfannenbecker, U., Potthast, J-M., Csato, M., Sladowski, D., Steiling, W. & Brantom, P. (1998). EU/COLIPA "*in vitro* phototoxicity" validation study, results of phase II blind trial, part 1: the 3T3 NRU phototoxicity test. *Toxicology in Vitro* **12**, 305–327.
3. Spielmann, H., Balls, M., Dupuis, J., Pape, W., de Silva, O., Holzhütter, H-G., Gerberick, F., Liebsch, M., Lovell, W. & Pfannenbecker, U. (1998). A study on UV filter chemicals from Annex VII of European Union *Directive 76/768/EEC* in the *in vitro* 3T3 NRU phototoxicity test. *ATLA* **26**, 679–708.
4. OECD (1996). *OECD Test Guidelines Programme, ENV/MC/CHEM/TG 96)9: Final Report of the OECD Workshop on Harmonisation of Validation and Acceptance Criteria of Alternative Toxicological Test Methods*. Paris, France: OECD Publications Office.
5. Borenfreund, E. & Puerner, J.A. (1985). Toxicity determination *in vitro* by morphological alterations and neutral red absorption. *Toxicology Letters* **24**, 119–124.
6. Holzhütter, H-G. (1997). A general measure of *in vitro* phototoxicity derived from pairs of dose–response curves and its use for predicting the *in vivo* phototoxicity of chemicals. *ATLA* **25**, 445–462.
7. Holzhütter, H-G. & Colosimo, A. (1990). SIMFIT: a microcomputer software toolkit for modellistic studies in biochemistry. *Computer Applications in the Biosciences* **6**, 23–28
8. Clothier, R., Willshaw, A., Cox, H., Garle, M., Bowler, H. & Combes, R. (1999). The use of human keratinocytes in the EU/COLIPA international *in vitro* phototoxicity test validation study and the ECVAM/COLIPA study on UV filter chemicals. *ATLA* **27**, 247–259.
9. Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, 436 pp. London, UK: Chapman & Hall.
10. Holzhütter, H-G., Archer, G., Dami, N., Lovell, D., Saltelli, A. & Sjöström, M. (1996). Recommendations for the application of biostatistical methods during the development and validation of alternative toxicological methods. *ATLA* **24**, 511–530
11. Holzhütter, H-G. & Quedenau, J. (1995). Mathematical modeling of cellular responses to external signals. *Journal of Biological Systems* **3**, 127–138.