Sebastian Hoffmann

Evidence-based in vitro toxicology

Dissertation

		reference standard		
		non-toxic	toxic	Σ
test	non-toxic	a	b	a + b
	toxic	c	d	c + d
Σ		a + c	b + d	a + b + c + d

Universität Konstanz

Januar 2005

Evidence-based in vitro toxicology

Dissertation

zur Erlangung des akademischen Grades

des Doktors der Naturwissenschaften an

der Universität Konstanz (Fachbereich Biologie)

vorgelegt von

Sebastian Hoffmann

Tag der mündlichen Prüfung: 13. Mai 2005

1. Referent: Prof. Dr. Thomas Hartung

2. Referent: Prof. Dr. Albrecht Wendel

Universität Konstanz

Januar 2005

Für meine Eltern

List of publications

Major parts of this thesis are published or submitted for publication:

- **Hoffmann, S.,** and Hartung, T. Diagnosis: toxic! - Trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations -. *Tox Sci*, in press

- **Hoffmann, S.,** Cole, T., and Hartung, T. (2005). Skin Irritation: Prevalence, variability and regulatory classification of existing in vivo data from industrial chemicals. *Regul Toxicol Pharmacol,* in press.

- **Hoffmann, S.,** Peterbauer, A., Schindler, S., Fennrich, S., Poole, S., Mistry, Y., Montag-Lessing, T., Spreitzer, I., Löschner, B., van Aalderen, M., Bos, R., Gommer, M., Nibbeling, R., Werner-Felmayer, G., Loitzl, P., Jungi, T., Brcic, M., Brügger, P., Frey, E., Bowe, G., Casado, J., Coecke, S., de Lange, J., Mogster, B., Næss, L.M., Aaberge, I.S., Wendel, A. and Hartung, T. International validation of novel pyrogen tests based on human monocytoid cells. *J Immunol Methods*, in press.

- **Hoffmann, S.,** Lüderitz-Püchel, U., Montag-Lessing, T., and Hartung, T. (2005). Optimisation of pyrogen testing in parenterals according to different pharmacopoeias by probabilistic modeling. *J Endotoxin Res,* in press.

Further publications, not integrated into this thesis:

- **Hoffmann, S.,** Hartung, T., and Beran, J. (2002). Comments on the use of bootstrap resampling to assess the uncertainty of Cooper statistics. *Altern Lab Animals* **30**, 551-554.

- Langezaal, I., **Hoffmann, S.,** Hartung, T., and Coecke, S. (2002). Evaluation and Prevalidation of an Immunotoxicity Test Based on Human Whole-blood Cytokine release. *Altern Lab Animals* **30,** 581-595.

- von Aulock, S., Schröder, N., Gueinzius, K., Traub, S., **Hoffmann, S.,** Graf, K., Dimmeler, S., Hartung, T., Schumann, R., and Hermann, C. (2003). Heterozygous toll-like receptor 4 polymorphism Asp299Gly has no influence on inflammatory serum markers or the ex vivo inducible release of pro-inflammatory cytokines of leukocytes. *J Inf Dis* **188,** 938-943.

- Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., **Hoffmann, S.,** Janusch Roi, A., Prieto, P., Sabbioni, E., Scott, L., Worth, A., and Zuang, V. (2004). A modular approach to the ECVAM test principles on test validity. *Altern Lab Animals* **32***,* 467-472.

- **Hoffmann, S.,** Ponti, J., Munaro, B., Fischbach, M., and Sabbioni, E. An optimised data analysis for the Balb/c 3T3 cell transformation assay applied to some metal compounds. (submitted to *Regul Toxicol Pharmacol*)

Acknowledgement

The work presented in this thesis was carried out between January 2001 and January 2005 at the Chair of Biochemical Pharmacology of the University of Konstanz under the supervision of Prof. Dr. Dr. Thomas Hartung.

I especially want to thank Prof. Dr. Dr. Thomas Hartung for introducing me into this field, for his continuous support and motivation and for providing excellent facilities. Furthermore, he gave me the opportunity to attend many exciting meetings.

My thank goes to Prof. Dr. Albrecht Wendel for welcoming me into the group and his steady support.

Special thanks to the all members of the 'Z'-group, who welcomed me warmly and formed an extraordinary team I was lucky to work with. Especially I would like to thank Stefanie Schindler who patiently introduced me into the field of pyrogens.
In the same way, I want to thank all members of ECVAM for welcoming me as warmly and introducing me to the multi-facetted worlds of toxicology, validation, regulation, and Italy.
I thank Anne for all the love that kept me, although always over a distance, easily going.
Finally, my very special thanks go to my family who always backed me up.

Abbreviations

BET	bacterial endotoxin test
CI	confidence interval
DL	developing laboratory
ECB	European chemicals bureau
ECS	European classification system
ECVAM	European centre for the validation of alternative methods
EINECS	European inventory of existing commercial substances
ELC	endotoxin limit concentration
ELISA	enzyme-linked immunosorbent assay
EP	European pharmacopoeia
EU	endotoxin units
GHS	globally harmonized system
GLP	good laboratory practice
IC	inhibiting concentration
ICCVAM	Interagency coordinating committee on the validation of alternative methods
IFNγ	interferon γ
IL	interleukin
JP	Japanese pharmacopoeia
LAL	Limulus amoebocyte lysate
LD	lethal dose
LoD	limit of detection
LPS	lipopolysaccharide
LTA	lipoteichoic acid
MM6	MONO MAC 6
MVD	maximum valid dilution
NCD	new chemicals database
NL	naive laboratory
NOV	negative predictive value

OECD	Organisation for economic co-operation and development
PBMC	peripheral blood mononuclear cells
PBS	phosphate buffered saline
PPV	positive predictive value
(Q)SAR	(quantitative) structure activity relationship
ROC	receiver operation curve
TLR	toll-like receptor
TNFα	tumor necrosis factor α
USP	United States pharmacopoeia
WBT	whole blood test

Table of contents

1 General introduction

Toxicology studies adverse effects of substances affecting man, animals or the environment, which in the presented thesis this are narrowed down to adverse health effects to humans. In order to asses and understand the absence or presence of such effects toxicologists construct models. Traditionally these are *in vivo* methods, i.e. methods making use of higher living organisms. With the better understanding of toxicological mechanisms and the strongly evolving biotechnologies more and more *in vitro* methods are developed. The basis of these methods ranges from lower organisms, e.g. the Ames test using bacteria to cultured cells.

Similarly to the impact of *in vitro* methods on toxicology, the emerging field of evidence-based medicine changes clinical sciences. Being introduced only a decade ago, it is usually defined as the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients.

In this thesis principles and tools of evidence-based medicine are systematically applied to *in vitro* toxicology, where emphasis is put on the assessment of the relevance of toxicological methods.

2 Diagnosis: toxic!
- Trying to apply approaches of clinical diagnostics and prevalence in toxicology considerations -

Sebastian Hoffmann and Thomas Hartung

European Commission, JRC - Joint Research Centre

Institute for Health & Consumer Protection

ECVAM – European Centre for the Validation of Alternative Methods

21020 Ispra, Italy

Toxicological Sciences, in press

2.1 Summary

The assessment of relevance of toxicological testing was compared with approaches of diagnostic medicine, a discipline that faces a comparable situation. Considering the work of a toxicologist as setting a diagnosis for compounds, assessment tools for diagnostic tests were transferred to toxicological tests. In clinical diagnostics, test uncertainty is well accepted and incorporated in this assessment. Furthermore, prevalence information is considered to evaluate the gain in information resulting from the application of a test. Several common toxicological scenarios, in which test uncertainty and prevalence are combined, are discussed including the interdependence of test accuracy, prevalence and predictive values or the sequential application of a screening and a confirmatory test. In addition, real prevalences from prevalences determined by an imperfect test are presented. We conclude that

information on prevalences of toxic health effects is required to allow a complete assessment of the relevance of toxicological test. In this process, lessons can be learned from evidence-based approaches in clinical diagnostics.

2.2 Thought starter

Let's assume you are a doctor asked to carry out an HIV test on a healthy European without any specific risk factors. You choose the best test available, which is 99.9% accurate. Unfortunately, the result is positive. Bad news for your patient? Not yet: The prevalence of HIV-infection, i.e. proportion of infected people in the general population, in Europe, is about 1:10,000 inhabitants. This means, if testing 10,000 people you will pick up one real-positive but your best available test will show 10 false-positives. A positive result will thus only be correct in one out of 11 cases, i.e. the probability that your patient is really HIV-infected is about 9%. A similar reasoning can be found in Gigerenzer et al. (1998) pointing out the need to communicate carefully any diagnosis to these patients.

What does this teach us as toxicologists? Our problem is in most cases, that we do not even know how accurate our test methods are (certainly less then 99.9%), and have no indication of the prevalence, i.e. the proportion of toxic chemicals, for a given health effect in specific populations of chemicals.

This thought prompted us to elaborate, what diagnostic medicine can teach toxicology in handling the uncertainty of setting the diagnosis of a substance exerting a given toxic effect in our test methods.

2.3 The accuracy of the diagnosis – translation to toxicology

Setting a diagnosis in clinics is an art, which involves three aspects: the patient, the physician and the diagnostic measures (Haynes et al., 1996; Haynes et al., 2002). Difficulties in setting a diagnosis arise from incompatibilities and limitations of each component. When identifying a toxic hazard of a chemical, similar to the role of the physician, expertise of the assessor has to overcome the many limitations of our

insight into the nature of the phenomena. In both cases a number of problems have to be considered (Table 1).

It is impossible to judge the relative contribution of these factors. The conclusion is simple: Our tools to assign a toxic health effect are imperfect. This is well accepted in the field of clinical diagnostics (Boyko et al., 1988; Sackett et al., 1991; Hunink et al., 2001; Knottnerus et al., 2002;). However, in toxicology, we are not used to estimate and incorporate the uncertainty but we base the conclusion (i.e. labelling/classification as well as use or not-use of the substance for certain purposes) on this imperfect assessment.

Noteworthy, this review omits any discussion of the link of the relationship of a toxicity test to any adverse human health effect. The analysis is based only on the interplay of test quality and prevalence for a given test. This is principally applicable to any test, be it diagnostic in humans, in animals or in vitro.

Clinical Medicine	Toxicology
Patient-related problems	**Substance-related problems**
Mix-up of patients	Mix-up of substance identities
General health status	Purity of the substance including synergistic and antagonistic effects of bystanders/impurities
Possibility to carry out invasive diagnostics	Stability of the substance
Individual reasons not permitting a diagnostic test (no consent of the patient, phobias, allergies)	Chemicophysical properties of the substance interfering with the test; size of crystals and solubility in carrier as well as biological fluids
Comorbidity	Further toxic effects of the substance beside the tested one
Temporal changes: compliance; stage of disease; fluctuation of symptoms	Bioavailability, i.e. in vivo toxicokinetics or in vitro biokinetics (describing the fate of a substance in cell culture including unspecific binding)
Test-related problems	**Test-related problems**
The variability and reproducibility of the test	The variability and reproducibility of the test
The reliability of reference standards employed	The reliability of reference standards and materials employed
The data interpretation procedure, e.g. the thresholds for classifications	The data interpretation procedure, e.g. the thresholds for classifications
The pathophysiological relevance of the test used	The mechanistic relevance of the test used
The limit of detection of the test	The limit of detection of the test
The predictive capacity of the test for a disease (accuracy, sensitivity, specificity)	The predictive capacity of the test for the human health effect (accuracy, sensitivity, specificity)
Personnel-related problems	**Personnel-related problems**
The correct conductance of the test linked to the quality and training of personnel	The correct conductance of the test linked to the quality and training of personnel
The bias of the technical personnel and the physician when interpreting the test	The bias of the experimentator/assessor with regard to the outcome of the test for a given substance

Table 1. Problems of setting a diagnosis in clinical medicine and toxicology

2.4 The quality of our 'diagnostic' tools

The field of carcinogenicity, namely the rodent bioassay's 50% positive rate triggered a detailed discussion of its restrictions and limitations regarding the predictive capacity for humans (Ames and Gold, 1990; Gold et al., 1998). However, few toxicity tests have been studied with this scrutiny. The area of validation of alternative methods has

pioneered the assessment of the quality of methods employed in toxicology (Balls et al., 1990; Balls et al., 1995). The crucial achievement here was the concept of relevance, i.e. assessing not only the reliability/reproducibility, but also the predictive capacity of a method. This implies, however, to have a point of reference (usually termed 'reference standard'). In clinical diagnostics this reference is often included in systematic studies (Walter et al., 1999; Knottnerus and Muris, 2003) and new assessment tools have even been developed for study evaluation (Whiting et al., 2003). In toxicology this optimal way of direct comparison most often is not applied due to cost or animal welfare considerations. If at all, a retrospective analysis in comparison to the reference (e.g. from databases) is carried out (Fentem et al., 1998). In both disciplines, this reference standard is usually but not necessarily another test. In toxicology a consensus of experts on a list of substances' toxicological properties or classification could substitute here. In clinical diagnostics, sometimes similarly the reference standard to assess the performance of a diagnostic measure is established by consensus of an independent expert panel using patients' diagnosis established by clinical criteria (Weller and Mann, 1997; Knottnerus and Muris, 2003).

It is of utmost importance to understand that validation assesses the reliability and relevance of methods. Figure 1 illustrates the validation process and which type of information constitutes the validity of a test method. Since often the primary test result is not expressed identical to the reference test, a prediction model is required to translate one into the other (Worth and Balls, 2001). For example, results of the test method might be continuous but must be classified into positive/negative or negative/mild/moderate/severe employing thresholds. In case of alternatives to animal experiments, the prediction model would translate from the in vitro result, e.g. an IC_{50} value, to an in vivo endpoint, e.g. LD_{50}.

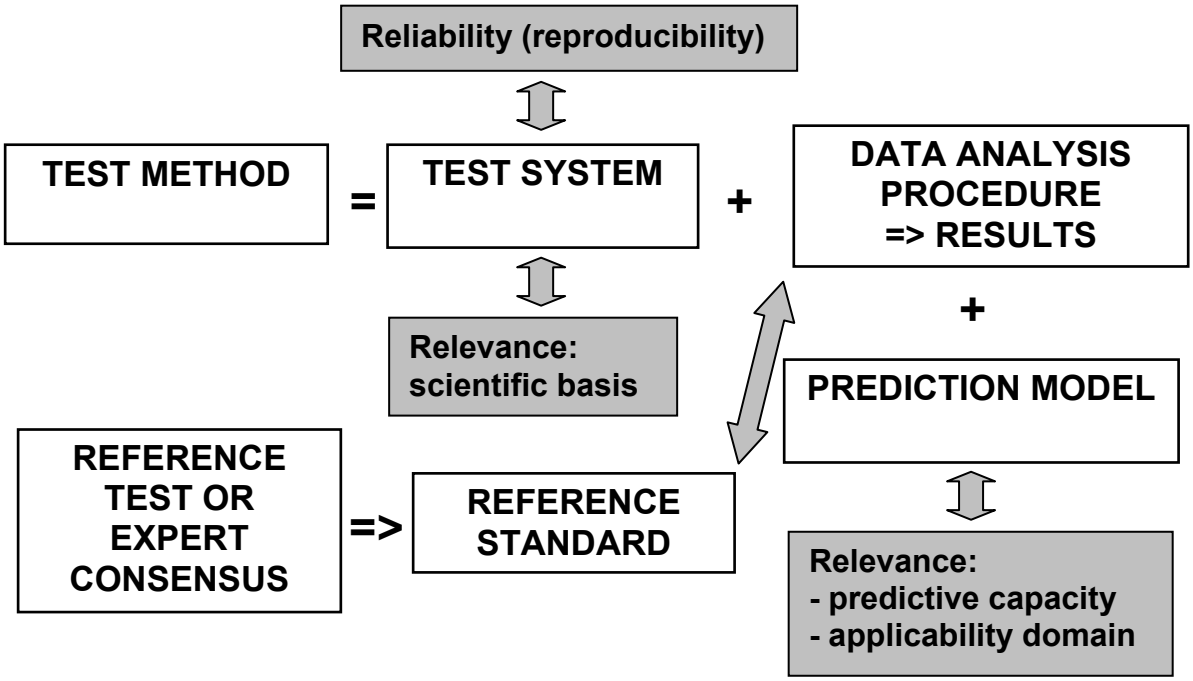

Figure 1. The process of validation

The graphic presentation highlights the three aspects of test validity. The test to be validated consists of the test system and its analysis procedure. Via a prediction model, the results have to be converted into the results of the reference standard to be compared with.

The quality and the adjustment of the prediction model is a key determinant of the predictive capacity of a test. In clinical diagnostics, designs and the sample sizes often allow threshold determination from the study data themselves (Sackett and Haynes, 2002). In contrast, the prediction model in toxicology is normally developed before the validation study using a training-set of substances (Bruner, 1996). The quality and properties of this set pre-determines the quality of results and applicability of the test method. Similarly, the selection of patients to establish a diagnostic method determines its quality for its intended use, the so-called patient spectrum (Irwig et al., 2002). If the selection is not representative or somehow flawed, e.g. only including severe cases, the relevance of the test will be impaired or restricted. Alike, it is instrumental that validation studies in toxicology include a sufficient number of weak toxicants. Optimally, although often a dichotomous, i.e. positive vs. negative, test outcome is chosen the selection should representatively cover the whole range of

toxic potency. This allows a better test assessment by expressing probabilities, e.g. of being positive or negative, for each chemical: A highly toxic compound will be classified as such more likely as a moderate toxic compound. Another difference in prediction model development/threshold definition in clinical diagnostics is that the sample sizes are usually substantially larger. This eases biometrical assessment, but several advantages of setting a diagnosis of toxicity compared to clinical medicine can compensate here:

- Testing of substances can be synchronised
- Testing can be repeated
- Positive and negative controls are readily available
- Replicates and related substance testing is feasible
- The number of toxic health effects is limited

2.5 The impact of prevalence

2.5.1 Incorporating prevalence into test assessment

As demonstrated in our thought starter, the prevalence of a disease is a key determinant of the practical value of a diagnostic measure (Buck and Gart, 1966; Linnet, 1988; Grimes and Schulz, 2002). If you are looking for something rare, even the best test will produce too many false-positives to rely on the result. It is therefore crucial to use descriptors of test relevance, which take the prevalence into account. In most simple cases (two outcomes of the test as well as for your reference standard), which will be mainly considered here for reasons of simplicity, this means to describe the relevance by the PPV (positive predictive value) and NPV (negative predictive value) instead of sensitivity, i.e. the probability of a correct negative result, and specificity, i.e. is the probability of a correct positive result. It would be challenging but also demanding to expand this concept of including prevalence information into multiple-class outcomes, as we have recently demonstrated for the case of skin irritation (Hoffmann et al., 2005). The predictive values estimate the proportion of correct positives/negatives test outcomes in all positives/negatives and are thus an

indication for the reliability of a positive/negative test result. However, in toxicology we often forget that our panel of test compounds is not reflecting the real world, but was designed to produce efficiently reliable estimates for sensitivity and specificity. For example, we choose 20 negatives and 20 positives not caring what the toxic effect is. Thus, the predictive values based on the artificial study prevalence are only telling us the predictive capacity of the test if the same distribution of positives and negatives is found in the real world. This is usually not the case. Unfortunately, for most toxicities we have no idea about their actual prevalence for example in chemicals of general use. Not only are we lacking complete information on basic toxic properties for a large number of high production volume chemicals on the market (Allanou et al., 1999; EPA, 1998), but even for the existing datasets no such analysis is available. Therefore, efforts should be spent in order to retrieve reliable estimates of those prevalences for the most relevant areas of toxicology.

Taking the example of skin irritation, this problem was recently approached (Hoffmann et al., 2005). Although in this work also a detailed distribution of skin irritating potential was presented and analysed, we restrict ourselves here to the prevalence analysis of the dichotomous outcome, i.e. irritant vs. non-irritant. In the New Chemicals Database of the European Chemical Bureau, including 3121 chemicals mainly notified in the last 15 years, the prevalence of skin irritating substances (according to EU-regulation) assessed by an animal experiment was 7.9%. The applicability domain with this prevalence would be the population of newly developed chemicals, while its use for other domains would have to be discussed. Since the database contains only results from one test in one laboratory for each chemical, the predictive capacity of the in vivo experiment could only be modelled for the outcome of a repetition of the same experiment. This resulted in a specificity of 99.7%, i.e. three out of 1000 non irritating chemicals would be classified false positive, and sensitivity of 94.1%, i.e. 59 out of 1000 irritating chemicals would be classified false negative, and thus in a NPV of 99.5%, i.e. only one out of 200 chemicals classified negative would in fact be an irritant, and a PPV of 96.8%, i.e. out of 1000 chemicals classified as irritating 32 would be not irritating. It is evident, that modelling further aspects of variability, e.g. the within- and between-laboratory reproducibility, would decrease the predictive capacity

estimates resulting in the respective decrease of the predictive values. For some combinations, the effect of prevalences and test accuracy, i.e. assuming that sensitivity equals specificity, on the predictive values is illustrated (Table 2). The most important consequence of these considerations is that for rare toxic events we can rely on the negative, not on the positive test results.

Positive predictive value (PPV) of a test [%]						
Prevalence of toxicity	**Test accuracy**					
	99.9%	**99%**	**95%**	**90%**	**80%**	**70%**
0.01%	9.1	1.0	0.2	0.1	0.04	0.02
0.1%	50.0	9.0	1.9	0.9	0.4	0.2
1%	91.0	50.0	16.1	8.3	3.9	2.3
10%	99.1	91.7	67.4	50.0	30. 8	20.6
Negative predictive (NPV) value of a test [%]						
Prevalence of toxicity	**Test accuracy**					
	99.9%	**99%**	**95%**	**90%**	**80%**	**70%**
0.01%	100	100	100	100	100	100
0.1%	100	100	100	100	100	100
1%	100	100	100	100	99.7	99.6
10%	100	100	99.4	98.8	97.3	95.5

Table 2. Predictive values for combinations of prevalence and test accuracy

Assuming a given test accuracy resulting from equal sensitivity and specificity, the consequences of different prevalences for the predictive values are calculated.

As the negative predictive value is always close to 100%, this shows that the idea to control negative test results in a second test, e.g. confirm negative in vitro results in vivo as suggested in the field of skin corrosion and irritation (OECD, 2002a), makes no sense at all for rare toxicities. Since the events are rare anyway, in most cases both test will be done, although the negative predictive value is high already. On the contrary, the positive ones have to be challenged, i.e. in regulatory toxicology avoiding

over-classification of and unnecessary restrictions for substances. In this context, we are well aware of the most crucial safety aspect of false-negative classifications, but the calculation shows that a second test can hardly improve the NPV, which is impaired by the false-negatives, of the first test. For example, for skin irritation even a test with only 70% accuracy will identify negatives correctly in more than 96% of the cases.

2.5.2 Consequences of inaccurate tests for prevalence determinations

An important question often overlooked is: How reliable are prevalences of rare diseases/toxicities, if assessed with our imperfect tools? If we agree that an in vivo experiment is not 100% accurate, many false-positives will populate our databases in case of rare toxicities. This means, that rare toxicities are rarer than we believe. For illustration, we present some combinations of prevalence determined by a test with a given accuracy (Table 3).

Real prevalences, if assessed with imperfect tests [%]						
Determined prevalence	**Test accuracy**					
	99.9%	**99%**	**95%**	**90%**	**80%**	**70%**
0.1%	0	n.d.	n.d.	n.d.	n.d.	n.d.
1%	0.9	0	n.d.	n.d.	n.d.	n.d.
5%	4.9	4.1	0	n.d.	n.d.	n.d.
10%	9.9	9.2	5.6	0	n.d.	n.d.
20%	19.9	19.4	16.7	12.5	0	n.d.
30%	30.0	29.6	27.8	25	16.7	0

Table 3: Real prevalences for some combinations of test accuracy and determined prevalences

The table presents the real underlying prevalences, when rare prevalences are determined with imperfect tests assuming accuracy = sensitivity = specificity; n.d. = not defined, because for a given test accuracy, e.g. 90%, the determined prevalence cannot be smaller than 100% - accuracy, e.g. 100% - 90% = 10%.

For example, applying a 90% accurate and finding a prevalence of 20%, means that the true prevalence is only 12.5%, i.e. only 5 out of 8 positive test substances are truly positive. Similarly, if we assume that the rabbit skin irritation test is 95% accurate, more than half of the selected skin irritants (prevalence 7.9%) would be false-positives. An obvious consequence is that the usefulness of databases for selecting the proper reference standard data is limited. If we assume that the rabbit skin irritation test is 95% accurate, more than 50% of the selected skin irritants (prevalence 7.88%) would be false-positives. As long as confirmatory testing in vivo is not carried out, it might be favourable to rely on the fewer, but more extensively studied substances from the scientific literature.

2.5.3 The use of confirmatory tests

It is common practice to apply a second test to confirm or challenge the results of a first test. When retesting positives, the specificity of the test procedure can be improved, when retesting negatives, the sensitivity of the test procedure can be improved. As we have seen above, this makes sense for relatively low prevalences only for positives, the NPV being almost optimal anyway. However, sensitivity and specificity of a test are interdependent: By defining for example the threshold value for a classification as positive or negative, one can be increased on cost of the other, usually demonstrated with receiver operation curves (ROC) as illustrated in Figure 2 (McNeil et al., 1975; van der Schouw et al., 1995). This offers the opportunity to render tests extremely sensitive accepting impaired specificity, a situation typical for screening tests.

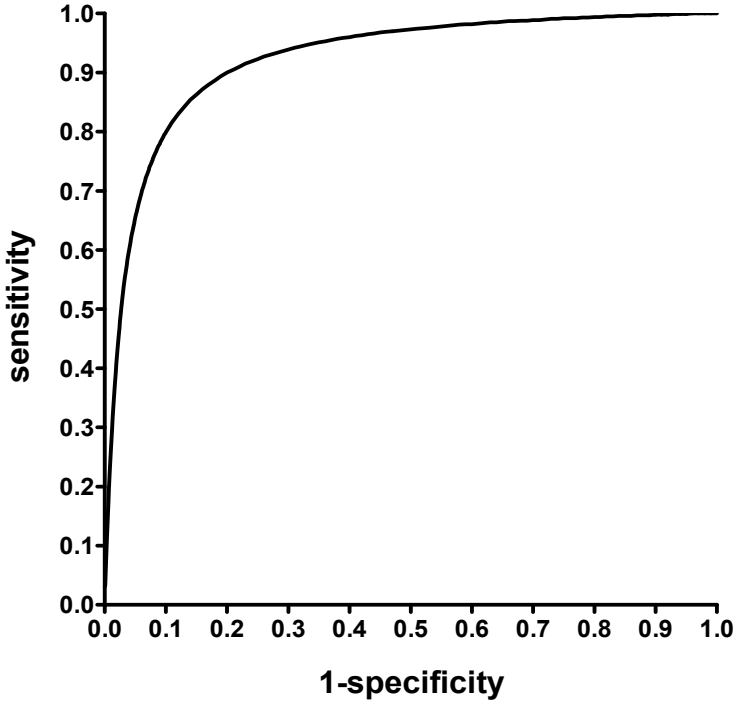

Figure 2: Illustration of a receiver operation curve (ROC)

The interdependence of sensitivity and specificity of a test by moving the classification threshold can best be visualized by ROC graphs. The steeper the curve ascends, the better the test, i.e. combining high sensitivity with high specificity.

A commonly applied and simple strategy in clinics as well as in toxicology is the combination of a screening test with a subsequently performed confirmatory test. With an oversensitive test a population is screened in order to detect as many positives as possible. Inevitably, this approach produces a lot of false-positive results in the first step. In a second step, all positively screened patients/substances are tested with a confirmatory test, which should be able to discriminate positives from negatives. The advantage of this strategy is often a reduction of costs, as screening tests with their lower overall predictive capacity are often substantially cheaper than their associated confirmatory tests. Nevertheless, the usefulness of this approach again strongly depends on the prevalence of the health effect (Buck and Gart, 1966) and the tests' dependence (Marshall, 1989). If the prevalence is low, a screening test will result only in minor increase of the number of positives due to false-positives. By this the overall testing costs are strongly decreased, but the positive predictive value does not change

tremendously, when compared to the PPV of the confirmatory test: For example let us assume a prevalence of 1%, an extremely sensitive screening assay with a sensitivity of 100%, but a specificity of only 50% and a good confirmatory assay with an accuracy of 95%. Testing 10000 substances, of which according to the assumed prevalence 100 are positive, the screen reduces the number of substances subjected to the confirmation test by 4950, i.e. all classified negatives. Applying now the confirmatory test reduces the overall NPV from 100% to 99.95%, i.e. 9652 of 9657 negatives are true negative, but results only in a PPV of 27.7% (Table 4), i.e. 95 of the 343 positives are true positives, compared to a PPV of the confirmatory test alone of 16.1% (Table 2). This means, for rare health effects a sufficient specificity of the screening has to be maintained. Even close-to perfect confirmatory tests cannot compensate. In the given example, improved screening test specificity 80% would result in a PPV of 49.0% and a value of 90% in a PPV of 65.7% (Table 4).

Prevalence = 1%: PPV of a combination of a screening and a confirmatory test [%]				
Screening test specificity (sensitivity = 100%)	**Accuracy of confirmatory test**			
	99.9%	**99%**	**95%**	**90%**
90%	99.0	90.9	65.7	47.6
80%	98.1	83.3	49.0	31.3
70%	97.1	76.9	39.0	23.3
60%	96.2	71.4	32.4	18.5
50%	95.3	66.7	27.7	15.4
40%	94.4	62.5	24.2	13.2
Prevalence = 10%: PPV of a combination of a screening and a confirmatory test [%]				
Screening test specificity (sensitivity = 100%)	**Accuracy of confirmatory test**			
	99.9%	**99%**	**95%**	**90%**
90%	99.9	99.1	95.5	90.9
80%	99.8	98.2	91.3	83.3
70%	99.7	97.3	87.6	76.9
60%	99.6	96.5	84.1	71.4
50%	99.6	95.7	80.9	66.7
40%	99.5	94.8	77.9	62.5

Table 4: PPV of combination of a screening test and a confirmatory test for prevalences of 1 and 10%

These calculations show the impact of the accuracy (assumed sensitivity = specificity) of a confirmatory test on the overall PPV when combined with a screening test with 100% sensitivity, but varying specificity.

A solution to further increase PPV is the application of a series of complementary tests in a sequence. This solution bares the problem of loss in cost reduction and of the evaluation of the dependencies between tests as complementary screens might be difficult to find. Furthermore the efficacy of combining tests for rare health effects is limited even under optimal conditions, i.e. assuming test independency (Table 5).

	PPV [%]				
Prevalence	**n = 1**	**n = 2**	**n = 3**	**n = 4**	**n = 5**
0.1%	0.9	7.5	42.2	86.8	98.3
1%	8.3	45.0	88.0	98.5	99.8
5%	32.1	81.0	97.5	99.7	100
10%	50.0	90.0	98.8	99.9	100

Table 5: PPV resulting from combining n 90% accurate and independent tests in sequence

This table shows how the PPV can be increased by combining several tests with 90% accuracy (assuming sensitivity = specificity).

When combining a screening and a confirmatory test, the screening test balancing sensitivity and specificity has to be carefully designed and adjusted as false-negative results in this step impair patients' health or consumers' safety. A more detailed insight into the prevalence is needed here to estimate consequences. One has also to consider the strength of a response and not only the dichotomized classification: It makes an enormous difference whether a large proportion of actual results is borderline to a given threshold or whether the negatives and positives are clearly distinct (Brenner and Gefeller, 1997; Bruner et al., 2002b). Especially in case of low prevalences, an extremely skewed distribution towards the negative end of the scale can be expected.

2.6 Prevalence in distinct chemical classes

So far we have handled the chemicals from the chemical universe only as a single entity. But they are related to other chemicals, e.g. by chemical structure, physiochemical properties or mechanism of action. If information of a related chemical is available, there is an increased probability that the considered chemical behaves toxicologically similar. This can be compared to the clinical situation where the integration of family anamnesis might change the probability of a diagnosis dramatically. Considering for example mutagenicity, in contrast of being a relatively

rare toxic effect among all chemicals, it is more common for the chemical group of nitrosamines. Making use of this kind of a-priori information, structural alerts or read-across approaches should help to assign chemicals to families with high or low prevalence of health effects. This is by no means new but reflects practices of priority setting by (Q)SAR and similar computational approaches or simply experience of the risk assessor. What we, however, desperately need are measures of how close two substances are related. As long as these are missing, groups of chemicals should be considered mainly as classes with different prevalences and thus different certainty of test results allowing applying tests with different sensitivities and specificities. This calls for test strategies, which take into account general prevalence, chemical classes with their individual prevalences and the use of proper tests with suitable predictive capacities.

2.7 Conclusions

Toxicological test were reviewed as diagnostic tools to assess the toxicological properties of substances. Taking this point of view parallels between diagnostic medicine and toxicology can be found, where evidence-based medicine methodology should be explored and possibly adopted. To properly assess toxicological tests, their reliability and relevance need to be explored, a process, in which the reference standard is of crucial importance. If no appropriate reference standard exists, expert consensus can be a valid alternative. Nevertheless, reference standards will always be imperfect. Accounting for this imperfectness is crucial for a complete test evaluation. Here, the extent to which imperfect reference data might populate databases was demonstrated, especially with false-positives for low prevalence cases. Furthermore, a systematic assessment of prevalences of toxic health effects in the chemical universe as well as in defined classes of chemicals is required. Only when combined with prevalence considerations, the ‘diagnostic’ value of a test can be estimated. For example, in low prevalence situations negative predictive values are almost optimal which challenges the approach of confirming negative results. In addition, we highlighted that the use of confirmatory tests strongly depends on

prevalence and test accuracy. Additional information on substances, including chemicophysical properties, chemical structure or classes, might affect the prevalence. Therefore, for toxicological hazard identification and testing strategies integration of prevalence information is crucial. In this process, lessons can be learned from medical diagnosis setting and especially from the evidence-based evaluation of diagnostic measures as a step towards evidence-based toxicology.

3 Skin irritation: Prevalence, variability and regulatory classification of existing in vivo data from industrial chemicals

Sebastian Hoffmann[a], Thomas Cole[b], Thomas Hartung[a]

[a] European Commission, JRC - Joint Research Centre; Institute for Health & Consumer Protection; ECVAM – European Centre for the Validation of Alternative Methods, 21020 Ispra, Italy

[b] European Commission, JRC - Joint Research Centre; Institute for Health & Consumer Protection; ECB – European Chemicals Bureau, 21020 Ispra, Italy

Regulatory Toxicology and Pharmacology (2005), in press

3.1 Summary

In vivo rabbit data for skin irritation registered in the European New Chemicals Database (NCD) and an ECETOC database were evaluated to characterise the distribution of irritation potential among chemicals and to assess the variability of the animal test. These databases could be used to determine experimental and rudimentarily within-laboratory variability, but not between-laboratory variability. Our evaluation suggests that experimental variability is small. Using two classification systems - the system currently used in Europe and the Globally Harmonised System (GHS) - the prevalence of skin irritation data obtained from NCD was analysed. This analysis revealed that out of 3121 chemicals tested, less than 10% showed an irritation potential in rabbits, which would require an appropriate hazard label, and 64% did not cause any irritation. Furthermore, it appears that in practical use the

European classification system introduces bias towards over-classification. Based on these findings, we conclude, that the classification systems should be refined taking prevalence into account. Additionally, prevalence should be incorporated into the design and analysis of validation studies for in vitro test methods and in the formulation of testing strategies.

3.2 Introduction

In order to manufacture, transport or market a chemical or finished product, the assessment of skin corrosive and skin irritation potential forms part of toxicological routine evaluation. To conduct this assessment, most regulatory authorities require a standardized in vivo test in which, having first excluded skin corrosion potential, the test substance is applied to the skin of a maximum of three rabbits (EPA, 1998; OECD, 2002 a; EC, 2004). The ability of the test substance to induce erythema and/or oedema is scored per animal. An integer score between 0 and 4 according to the Draize scale, which increases with severity, is subjectively assigned for erythemal and oedemal effects, usually at 24, 48 and 72 hours after application of the substance (Draize et al., 1944).

Scientific concerns about the variability (Weil and Scala, 1971; Gilman 1978; Worth and Cronin, 2001) and predictive capacities of this animal test in terms of human health effects (Campbell and Bruce, 1981; Calvin, 1992; Robinson, 2000) are raised regularly. In addition, animal welfare and more recently political pressure in Europe, e.g. chemicals and cosmetics legislation, decree the development of appropriate and validated alternative in vitro test methods (Hartung et al., 2003). Therefore, in the last ten years, considerable scientific effort has been directed at developing valid in vitro skin models as a replacement for the animal test. For skin corrosion this effort resulted in a successful validation of in vitro tests (Fentem et al., 1998), which were recently accepted as Organization for Economic Co-operation and Development (OECD) guidelines (OECD, 2004 b, c). Currently, a similar validation of in vitro methods for skin irritation is in progress (Zuang et al., 2002). In this study, a critical review of the in vivo rabbit skin irritation test according to current regulations was performed, focusing

on the assessment of historical in vivo data. For this purpose, two databases were identified and analysed in close cooperation between the European Centre for the Validation of Alternative Methods (ECVAM), the European Chemicals Bureau (ECB) and the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) in the United States.

Additionally, and in parallel, regulators have been optimising the assessment of skin irritation in a broader context. An international effort has resulted in a harmonised classification scheme for chemicals known as the Globally Harmonised System (GHS) including skin irritation, which is generally based on the Draize scores obtained in the traditional animal test (OECD, 1998; UN, 2003a). Furthermore, testing strategies incorporating historical data, (quantitative) structure activity relationship ((Q)SAR) methods, in vitro and also in vivo tests, have been developed which should significantly reduce or replace animal testing (Calvin, 1992; Robinson et al., 2002; Gerner and Schlede, 2002; OECD, 2002 a; UN, 2003; EC, 2004). Hitherto however, only a limited technical literature is available, e.g., an approach modelling a strategy for skin corrosion (Worth et al., 1998) and an OECD review considering the GHS approach for skin irritation (OECD, 1999). This paper analyses available in vivo data from a review perspective on proposed testing strategies and the impact of different classification systems.

3.3 Material and Methods

The databases containing in vivo data were the ECETOC database (Bagley et al., 1996; ECETOC, 1995) and the New Chemical Database (NCD) of the ECB (European Commission Joint Research Centre). NCD comprises chemicals introduced commercially, relatively recently, i.e. notified after 1981, where registered data have been derived according to regulatory standards, including GLP and predominantly according to official test methods (Annex V of Directive 67/548/EEC) (EC, 1992). The ECETOC database was established in 1995 (ECETOC, 1995). Most of the ECETOC database chemicals and substances are registered in the European INventory of Existing Commercial Substances (EINECS) (http://ecb.jrc.it), where no information on

skin irritation is available. Thirteen chemicals are officially classified with regard to skin irritation (Annex I of Directive 67/548/EEC) (EC, 1992).

Two current dermal hazard classification systems of relevance are the GHS (OECD, 2002 a; UN, 2003) and the European Classification System (EC, 2001) illustrated schematically in Figure 1.

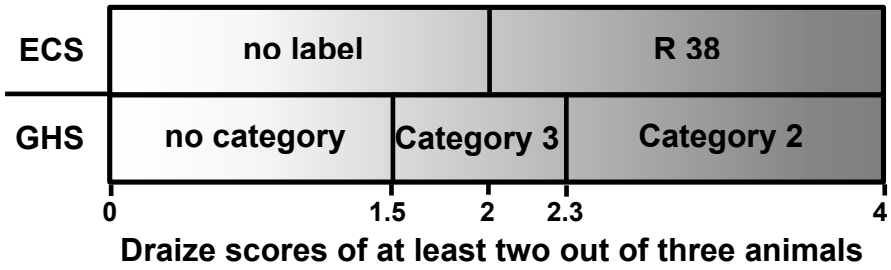

Figure 1. Schematic comparison of the European classification scheme and the Globally Harmonised System for skin irritation based on Draize scores of three rabbits.

The European Classification System (ECS) has a dichotomous outcome distinguishing skin irritant chemicals with assignment of the risk phrase R38 'irritating to skin' (EC, 1992) from non-irritant chemicals with no label. GHS introduces a classification that distinguishes irritants (category 2), mild irritants (category 3) and non-irritants (no category) by assigning increasing category numbers for decreasing severity of the adverse health effect (category 1 is assigned to corrosives). Today, both schemes are mainly based on the Draize scores of three animals. Erythemal and oedemal effects of each animal are scored according to the Draize scale, 24, 48 and 72 hours after chemical application. For each dermal effect these scores are summarised by calculating an individual mean score over the three time points per animal. The resulting individual means constitute the information which is used for classification in both schemes. In ECS, a chemical is labelled R38, if at least two of the three rabbits show an individual mean score ≥ 2.0 for at least one dermal effect. In GHS, besides two secondary criteria covering persistent effects and singular events of strong irritation, a chemical is classified into Category 2 if at least two rabbits show an individual mean score ≥ 2.3 for erythemal and/or oedemal effects. Chemicals inducing mean scores in the range from 1.5 to 2.3 for at least one dermal effect in two or three

animals are classified into Category 3. For example, a chemical tested on three rabbits, causing individual erythema mean scores of 1.7, 2 and 2.3, and individual oedema mean scores of 1, 1.3 and 1.7, would be classified due to its erythema scores, as irritant (R38) according to ECS or as mild irritant (Category 3) according to GHS. Until recently, in Europe the number of test animals was not restricted to three. In the case of studies with more than three animals, ECS further summarised the individual mean scores of the rabbits by calculating their overall mean for each dermal effect. A chemical was labelled R38 when the overall mean for at least one dermal effect was ≥ 2.0 and not labelled otherwise. However, this option has become obsolete, since testing in more than three animals has discontinued (OECD Test Guideline 404) (OECD, 2002 a; EC 2004).

The statistical tools, descriptive and resampling techniques, i.e. random generation of data samples from existing data, were applied primarily in an exploratory manner by means of the software package S-Plus 6.1 (© Insightful, Seattle, WA). For an assessment of prevalence, i.e. the proportion of irritant chemicals among all considered chemicals, chemicals were classified by applying both classification schemes. The results on test performances were summarised with the probabilities for correct classifications and the predictive values, i.e. in the case of ECS sensitivity and specificity, respectively, the positive and the negative predictive value (NPV and PPV). To facilitate this process and to allow a more detailed insight into prevalence, animal test data were grouped by the dominant median of the chemicals. The dominant median is a concept developed for this analysis. It is determined by calculating the median of the individual rabbit mean scores for each dermal effect and then choosing the larger, i.e. dominant one. This median allows classification of a chemical according to both schemes by comparison with the classification cut-off points, i.e. 2 in the case of ECS or 1.5 and 2.3 in the case of GHS (Figure 1). In the example given above, the erythema median is 2 and the oedema median is 1.3, resulting in a dominant median of 2 and thus in a classification as irritant R38 (ECS) or Category 3 (GHS).

3.4 Results

3.4.1 New Chemicals Database (NCD)

At the time of this analysis there were almost 3900 individual substances registered in NCD. Excluding gases and substances produced at < 100 kg/annum, for which, according to Annex VII C of the Directive 67/548/EEC (EC, 1992), skin irritation information is not available, 3278 substances of interest remained. This set represented the starting point for our analysis (Figure 2). Among the 3278 chemicals, 148 (4.5%) were classified as corrosives, indicated by the hazard symbol ‘C’, including the risk phrases R34 (‘causes severe burns’) and R35 (‘causes burns’). Nine chemicals were classified as toxicants with the hazard symbol ‘T’, which hierarchically includes burns as an endpoint. Of the 3278 chemicals, 246 (7.5%) were labelled with the risk phrase R38 as skin irritants. Accordingly, 2875 chemicals, i.e. 87.7%, were neither corrosives nor skin irritants. From 1990 onwards, more than 150 chemicals per year with information on skin effects were registered in NCD. The proportion of R38 labelled substances varied from 4.9% to 12.8% with no evident trend over time (data not shown).

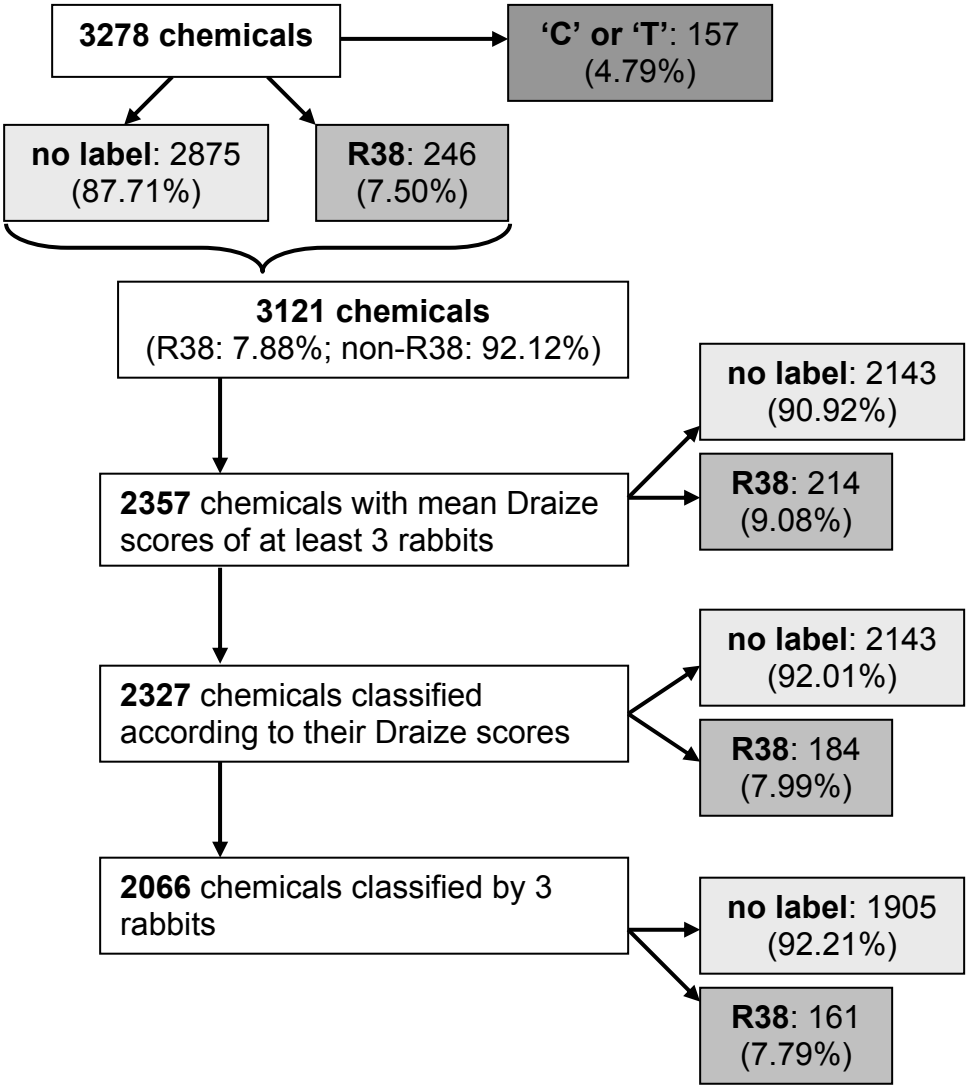

Figure 2. Flowchart of NCD evaluation with regard to skin corrosive and skin irritation labelling.

Excluding the 157 chemicals with the hazard symbols 'T' or 'C', which, in routine toxicity testing are assigned prior to testing for skin irritation, reduced the total number of relevant chemicals from 3278 to 3121. Of these, 2357 (75.2%) were registered with Draize scores from at least three rabbits, which were summarised according to ECS: If three rabbits were employed, the individual mean score over time is given per animal for each dermal effect; if more than three animals were used, only the mean of all

individual mean scores per effect was given, so that in these cases a measure of variation could not be derived from the data. Hence, the database included 764 (24.8%) chemicals which had either no, incomplete or non-conforming in vivo data. Focusing on the 246 substances labelled R38, 62 (25.2%) of these were classified as such, not based on the Draize score, although for 30 of these, Draize scores were given. Investigating the 62 chemicals further by reviewing regulator comments available within the notifications, NCD revealed justifications for the R38 classifications. Persistent effects on the skin were the most frequent cause (17 cases). For other chemicals, the R38 classification was, e.g., based on data from in vivo studies not conducted according to standard methodology (6 cases), on safety considerations (8 cases), i.e. when the scores were indicating irritancy close to, but below the classification threshold, or on read-across methods (4 cases), i.e. the interpolation of toxicological properties of a given substance from chemically closely related substances with known toxicological properties.

Having filtered NCD that way, the prevalence of Draize score based skin irritation in terms of ECS was calculated with 2327 chemicals. Of these, 184 (8.0%) were labelled R38 and 2143 (92.0%) were not labelled with regard to skin irritation. This subset is considered as highly representative for the 3121 chemicals of interest, because the prevalence of skin irritation according to ECS is almost identical in both data sets (Figure 2). Evaluating the two dermal effects in detail revealed that 101 (54.9%) of these 184 chemicals were labelled due to erythemal damage, twelve (6.5%) due to oedemal damage and 71 (38.6%) due to both endpoints. In order to allow a more detailed description of the distribution of irritation potential, we developed the concept of the dominant median. As both ECS and GHS are primarily based on the two highest individual means of three rabbits, the median individual mean of the three rabbits can be used equivalently. In doing so, a single value for each dermal effect, i.e. the median, resulting in identical classifications was easily derived. According to the Draize scale, ranging from 0 to 4, each dermal effect median of three animals can take thirteen different values. These are, rounded to the first digit, 0.0, 0.3, 0.7, 1.0, 1.3,…, 3.7 and 4.0 which cover the entire response range of the animal experiment. For the 261 chemicals in NCD that were tested using more than three animals,

resulting in the reduction from 2327 to 2066 chemicals in Figure 2, the reported means were rounded to the next dominant median. By choosing the larger of the extracted medians of erythemal and oedemal effects, we could reduce all available in vivo information on a chemical to a single characteristic value, i.e. the dominant median.

Calculating in a first step the median of the 2327 chemicals registered in NCD for both erythema and oedema resulted in a two-dimensional distribution of irritation potential (Figure 3 A).

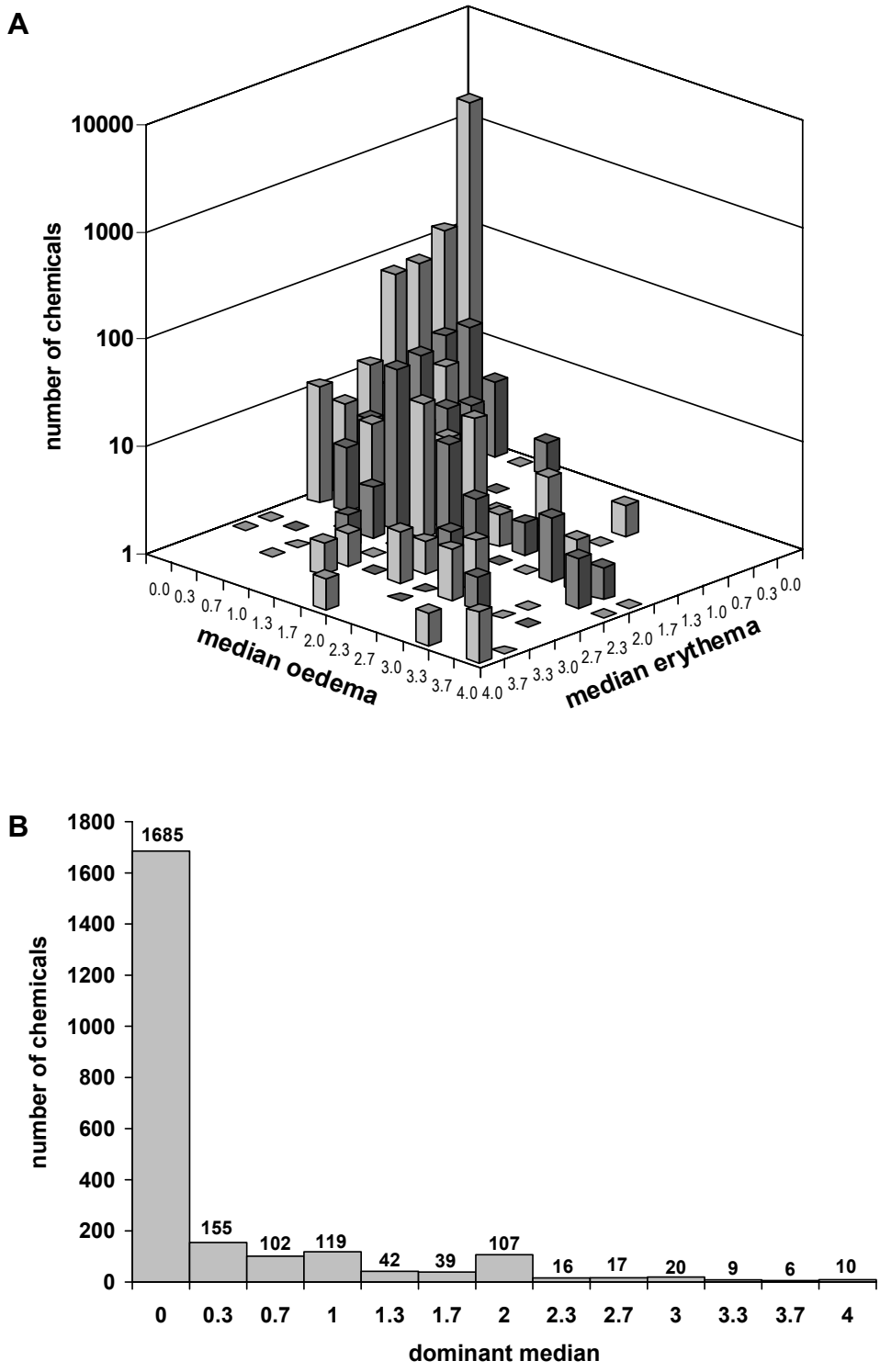

Figure 3. A: Two-dimensional histogram of chemical numbers (log-scale) of erythema and oedema in vivo medians in NCD (n = 2327).
B: Histogram of the prevalence of dominant medians in NCD (n = 2327). For each median class the total number of chemicals is given.

For 1756 (75.5%) of these chemicals the medians of the two endpoints were equal, where 1685 (72.4%) chemicals had medians of 0 for both. For 529 (22.7%) chemicals the erythemal effect was dominant and for 42 (1.8%) chemicals the oedemal effect was dominant. Applying in a second step the dominant median approach, this distribution was reduced to a one-dimensional histogram (Figure 3 B). Again, a dominant median of 0 was the most prominent class (1685, i.e. 72.4%). Of these 1685, 1482 (63.7%) chemicals had scores of 0 for all animals and both effects. Of the 2327, 2061 (88.6%) showed a dominant median ≤ 1.0. The group that includes the cut-off of ECS, i.e. the dominant median class 2.0, represented a peak in the classes larger than 1.0. This approach allowed calculation of the prevalence for every classification scheme that is based on three rabbits and on their individual mean scores over time. Considering GHS, being such a scheme, a chemical would be classified into no category for dominant medians up to 1.3. It would be classified into Category 3 for dominant medians of 1.7 and 2.0 and into Category 2 for a dominant median ≥ 2.3. The prevalence according to GHS in NCD is thus composed of 2103 (90.4%) non-irritants (no category), of 146 (6.3%) mild irritants (Category 3) and of 78 (3.3%) irritants (Category 2).

3.4.2 ECETOC database

We screened the ECETOC database, including 176 entries, for substances using several criteria. Corrosive chemicals, preparations and chemicals that did not have complete scores at 24, 48 and 72 hours after application for at least three animals or that were applied less than four hours were excluded. After filtering, 129 different chemicals with 162 entries remained. Twenty-three of the 129 chemicals had more than one entry, where the repeated experiments of each of these chemicals were conducted in the same laboratory. Per entry, the number of animals differed from three to six. Of the 106 chemicals with one entry, 50 were tested with three animals, 50 were tested using four rabbits and six were tested using six animals. Considering the 23 chemicals tested in replicate, data from seven to 15 rabbits were available. From each of these chemicals, the entry with the highest Draize scores was used to derive the prevalence of skin irritation in the ECETOC database according to the

dominant median approach (Figure 4). For cases with more than three animals, the mean over all rabbits was calculated and rounded to the next dominant median. In equivocal cases, the rounding was carried out conservatively, i.e. to the next larger dominant median. The same rounding was applied to scores of 0.5, 1.5, 2.5 and 3.5, which were present, although not in line with the Draize scale.

In contrast to NCD, prevalence in the ECETOC database showed less frequent occurrence of the dominant median 0.0. Chemicals with a dominant median of 2.0 were the most prominent class whilst chemicals with a dominant median value of 1.0 to 1.7 were common. We considered the distribution of dominant median in NCD the more relevant to analysis of prevalence in irritation scores.

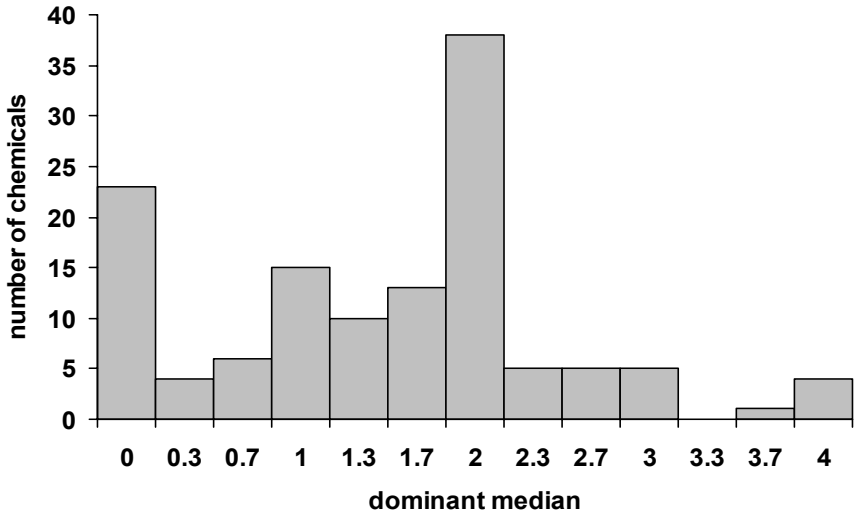

Figure 4. Histogram of the prevalence of dominant medians in the ECETOC (n = 129).

3.4.3 Within-test variability of Draize-scoring

Within-test variability of Draize-scoring refers to variation between assigned scores, when all animals are examined in one experiment, i.e. not independently. This analysis was carried out with the 2066 chemicals, for which individual animal data were available. This subset showed almost the same basic prevalence (Figure 2). For these chemicals, the individual animal means of the dominant dermal effect were taken, reducing the data to three values per chemical. The median of these values,

i.e. the dominant median, was used to place each chemical into one of the 13 groups of the above-described scale from 0 to 4. The distribution of the three values for chemicals with a dominant median of 0 or 4 is by definition asymmetric because they can vary in one direction only. Due to this property of the data, we employed the range of the three values, expressed in terms of scale intervals, to evaluate the variability. For example, a chemical with three individual means 0, 0 and 0.7 for the dominant dermal effect and thus a dominant median of 0, would range over two scale intervals, i.e. from 0 to 0.3 and from 0.3 to 0.7. To summarise this information, we calculated the median of this variability measure for each dominant median group (Figure 5).

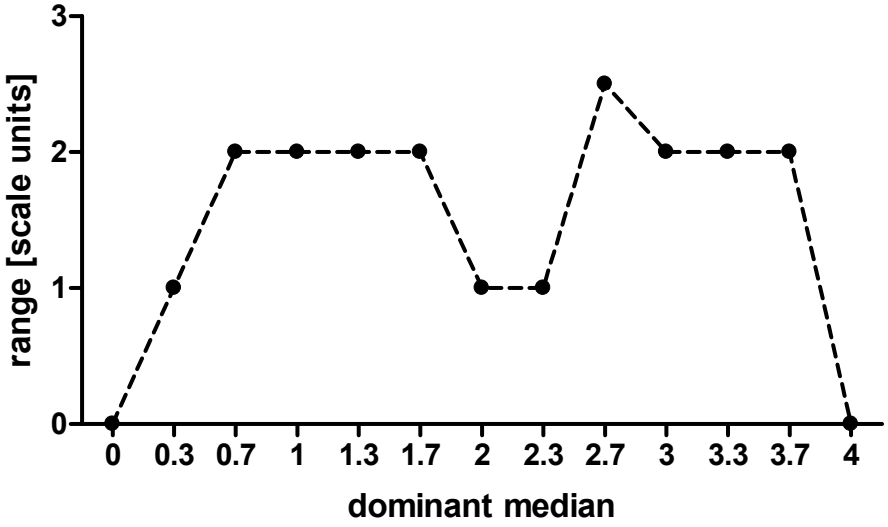

Figure 5. Within-test variability of Draize-scoring for each dominant median expressed as median number of scale units.

In contrast to the expectation that the variability would be largest in the centre of the scale, it was decreased for the dominant median groups 2.0 and 2.3. Notably 2.0 represents the threshold score of ECS, by which the NCD chemicals were assessed, indicating a certain bias in scoring borderline irritation effects.

To be able to assess the impact of this variability on ECS- and GHS-classifications, the approach of the dominant median was combined with statistical re-sampling techniques to analyse the NCD data. As the information per chemical, consisting of

only the minimum data needed for classification, i.e. from three animals, did not allow a straightforward analysis, we pooled the data in each of the dominant median groups. This resulted, e.g., for the dominant median group 0, in 4593 values from 1531 chemicals. For each of the thirteen classes, the animal data were re-sampled to model predictive capacity. For this purpose, 500,000 theoretical chemicals were generated by randomly drawing 500,000 combinations of three single values out of the respective values, i.e. in the example, out of the 4593 values. Subsequently, each theoretical chemical was classified according to ECS or GHS. By comparing each result with the original classification, which is defined by the dominant median group, we could calculate the probability of a correct classification per group. Together with the size n, for each group, these probabilities are given for both classification schemes (Table 1).

dominant median	sample size	ECS		GHS		
		no label	R38	no category	Category 3	Category 2
0.0	4593	**1**	0	**0.9980**	0.0020	0
0.3	369	**0.9998**	0.0002	**0.9995**	0.0005	0
0.7	246	**0.9999**	0.0001	**0.9962**	0.0038	0
1.0	315	**0.9915**	0.0085	**0.9745**	0.0255	0
1.3	105	**0.9793**	0.0207	**0.8961**	0.1039	0
1.7	87	**0.9002**	0.0998	0.1310	**0.8674**	0.0016
2.0	288	0.0953	**0.9047**	0.0374	**0.9460**	0.0165
2.3	30	0.0032	**0.9968**	0.0047	0.0741	**0.9254**
2.7	42	0.0389	**0.9611**	0.0052	0.1642	**0.8306**
3.0	57	0.0009	**0.9991**	0	0.0536	**0.9464**
3.3	21	0	**1**	0	0	**1**
3.7	15	0	**1**	0	0	**1**
4.0	30	0	**1**	0	0	**1**

Table 1. Re-sampling probabilities of classifications according to ECS and GHS for each dominant median for NCD data based on the within-test variability of Draize-scoring

bold: probability for correct classification ; light grey: probability of over-prediction; dark grey: probability of under-prediction

Taking for example the dominant median 1.3, in which 105 values were pooled originating from 35 originally non-labelled chemicals, 97.9% of the 500,000 theoretical chemicals would consistently not have been labelled when applying ECS. Only 2.1% of these would have been classified wrongly as R38. In terms of GHS, the probability for a correct classification (no category) was 89.6%. Of the 500,000 theoretical chemicals, 10.4% would have been classified into Category 3 and none into Category 2. Regardless of the classification system, the resampling approach showed that the within-test variability of the Draize-scoring rarely resulted in misclassifications. However, as the probabilities of incorrect classification are largest around the classification borders and decrease towards the limits of the dominant median scale,

introducing a second threshold by means of a third class can generally be expected to increase the overall likelihood of misclassification.

The ECETOC data were also employed to assess aspects of variation. Twenty-three substances were tested in replicate: 15, twice; six, three times; and two chemicals, four times. The in vivo data, usually employing three or four animals per test, were analyzed with regard to within-laboratory reproducibility, which includes the above modeled within-test variability. For each substance, the dominant median was extracted for each single test. We calculated the maximum differences between these medians, expressed as numbers of scale intervals. Twelve substances were classified identically in all experiments. There was a difference of one interval for seven chemicals, a difference of two intervals for two chemicals and differences of three and four classes once each. Due to the limited information available, the results should be considered only as an indicator of the within-laboratory reproducibility of the animal experiment.

3.4.4 Performance and comparison of classification schemes

Due to an absence of data, the between-experiment, and the within-/between-laboratory variation of the animal experiment remained un-assessed, implying substantial underestimation of misclassification probabilities. Nevertheless, the prevalence in NCD (Figure 3) was integrated with the results of the re-sampling, modelling within-test variability of Draize-scoring, in order to evaluate the utility of the approach. Table 2 summarises prevalence, predictive capacity and probability for false classification. The predictive capacities and probabilities for false classifications were calculated by means of the prevalence and the probabilities of correct, over- and under-prediction for the respective dominant medians. Considering, e.g., Category 3 of GHS, which comprises the two dominant medians 1.7 and 2.0, first their proportions were extracted, i.e. 39/2327 and 107/2327, from Figure 3. Multiplication of these proportions with the respective probabilities for correct classification, i.e. 0.8674 and 0.9460 (Table 1), and dividing their sum by the prevalence of Category 3, resulted in a predictive capacity of 0.925 and a probability for false classifications of 1 - 0.925 = 0.075. For GHS, the probabilities for false classifications were subdivided into the two

other categories (Table 2). Finally, the predictive values were calculated from the prevalences and the predictive capacities. Comparing both classification schemes, GHS Category 3 has the lowest probability for correct classifications and the smallest predictive value, where in particular, under-predictions (i.e. false negatives) may cause safety problems. Thus, the introduction of a third class would affect the overall performance of the classification of chemicals regarding skin irritation.

System	Classes	Prevalence	Predictive capacity	False classifications		Predictive value
ECS	no label	0.921	Specificity: 0.997	false positive: 0.003		NPV: 0.995
	R38	0.079	Sensitivity: 0.941	false negative: 0.059		PPV: 0.968
GHS	no category (NI)	0.904	0.996	0.004	MI: 0.004 I: 0	0.996
	Category 3 (MI)	0.063	0.925	0.075	NI: 0.062 I: 0.013	0.913
	Category 2 (I)	0.033	0.934	0.067	NI: 0.002 MI: 0.065	0.975

Table 2. Prevalence, predictive capacity and predictive values of the rabbit test according to ECS and GHS, based on NCD taking into account only the within-test variability of Draize-scoring

3.5 Discussion

While classification schemes simplify skin irritation to categories, the effect is continuous, where this study highlights response distribution of rabbits. As the majority of chemicals tested has no or low irritation potential, accurate classification of these substances should have priority. Currently, regulatory policy often allows safe negative classification only by in vivo methods. However, as already demanded

(Stitzel, 2002), this position should be reconsidered, enabling better implementation of animal welfare. For example, an alternative in vitro method developed to identify chemicals with no or slight irritant potential would not impair hazard assessment or consumer's safety, whilst reducing in vivo experimentation.

Our analysis of skin irritation presented here suggests that classification schemes have an influence on experimental scoring. High prevalence and variability of the threshold ECS dominant median, 2.0, demonstrate a bias towards assigning this score. In particular, the results suggest that many chemicals with Draize scores close to but below the threshold are classified as irritants. Subjective judgement fosters safety, but poses problems when comparing with new approaches or when comparing classification schemes. In the context of ECS and GHS, the consequence of introducing a third class increases misclassifications. Furthermore, there is an increased potential for scoring bias when additional classifications are introduced.

Although some principal aspects of within- and between-laboratory variability could not be modelled, the presented approach demonstrates relevance to assessment and comparison of classification schemes. In particular, some inconsistency is evident in the GHS classification categories for skin irritation potential, indicating scope for optimisation of the system. GHS, as established by the OECD (OECD, 1999) is based on three different, but equally arbitrary classification schemes, which were compared using archive databases. The occurrence of Draize scale non-conforming data in these databases seems to have influenced the lower threshold definition of Category 3 in GHS being set at ≥ 1.5. As according to the latest regulation 1.5 cannot be scored in future by definition, any eventual refinement of skin irritation potential definitions in GHS should reconsider a revision of this threshold score.

3.6 Conclusion

Our analysis indicates how data modelling may contribute to refinement of classification schemes. The evaluation of prevalence, here of skin irritation potential of chemicals, illustrates data distribution patterns relevant to defining testing strategies. Acknowledging applicability, costs and relevance, toxicity testing of a chemical often

follows a tiered approach. Refinement of these strategies by integrating prevalence could reduce costs and improve safety and animal welfare. Identifying in this process gaps in the strategies could stimulate new approaches to the validation of in vitro alternatives including the design of alternative methods, validation studies and (Q)SAR prediction models or the implementation of human data. Further, incorporation of cost-benefit analysis into hazard and risk assessment of industrial chemicals would be facilitated.

Our finding that classification schemes bias the underlying experiment adds to the problem of the imperfect reference standard which in vitro toxicology faces regularly (Gettings et al., 1996; Worth and Cronin, 2001). The animal test is imperfect, since assessment of its predictive capacity for human health effects is limited by absence of relevant comparative human data. Ideally, the relevance assessment of new alternatives should take relevance and reliability of the reference standard into account. With regard to our analysis, the available data allowed only evaluating the within-test variability, which constitutes only a small part of the reliability assessment of the animal experiment. It is evident that within- and between-laboratory variability of the animal experiment, as it is evaluated in validation studies, will substantially reduce reliability estimates. Ultimately, extrapolation to humans would introduce an additional variability due to heterogeneity of response among individuals and populations. In support of previous observations (Purchase, 1990; Stitzel, 2002) a key element to progress lies in sharing available data, facilitating analysis of reliability and relevance of reference standards and providing a basis for strategic development of assessment methods.

3.7 Acknowledgements

We thank Joseph Haseman whose earlier analysis of the ECETOC database prompted our study. We are also grateful to Joaquin Baraibar Fentanes (ECB) for the data extraction from the ECB New Chemicals Database and Ray Tice (NICEATM) for critical review of the manuscript.

4 International validation of novel pyrogen tests based on human monocytoid cells

Sebastian Hoffmann[a,h§], Anja Peterbauer[e*§], Stefanie Schindler[a§], Stefan Fennrich[a], Stephen Poole[b], Yogesh Mistry[b], Thomas Montag-Lessing[c], Ingo Spreitzer[c], Bettina Löschner[c], Mirjam van Aalderen[d], Rogier Bos[d], Martin Gommer[d], Ria Nibbeling[d], Gabriele Werner-Felmayer[e], Petra Loitzl[e], Thomas Jungi[f], Marija Brcic[f], Peter Brügger[g], Esther Frey[g], Gerard Bowe[h], Juan Casado[h], Sandra Coecke[h], Jan de Lange[h], Bente Mogster[i], Lisbeth M. Næss[i], Ingeborg S. Aaberge[i], Albrecht Wendel[a], Thomas Hartung[a,h#]

[a] Institute of Biochemical Pharmacology and Steinbeis Center InPuT, University of Konstanz, Universitätsstrasse 10, D-78457 Konstanz, Germany

[b] NIBSC, National Institute for Biological Standards and Control, Blanche Lane, South Mimms, Potters Bar, Herts EN6 3QG, England, UK

[c] Paul Ehrlich Institute, Paul-Ehrlich Strasse 51-59, D-63225 Langen, Germany

[d] RIVM, National Institute of Public Health and the Environment, A. van Leeuwenhoeklaan 9, P.O.Box 1, 3720 BA Bilthoven, The Netherlands

[e] Institute of Medical Chemistry and Biochemistry, Fritz-Pregl-Strasse 3, A-6020 Innsbruck, Austria

[f] Institute of Veterinary Virology, Länggass-Strasse 122, University of Bern, CH-3012 Bern, Switzerland

[g] Biological Analytics, Novartis Pharma AG, CH-4002 Basel, Switzerland

[h] European Centre for the Validation of Alternative Methods (ECVAM), Institute for Health & Consumer Protection, European Commission Joint Research Centre, Via Fermi 1, I-21020 Ispra, Italy

[i] Division of Infectious Disease Control, Norwegian Institute of Public Health, P.O. Box 4404 Nydalen, NO-0403 Oslo, Norway

* Present address: Ludwig Boltzmann Institute for Experimental and Clinical Traumatology/Blood Transfusion Service for Upper Austria, Blumauerstr. 3-5, A-4020 Linz, Austria

[§] S. Hoffmann, A. Peterbauer, S. Schindler contributed equally to the work presented here

Journal of Immunological Methods, in press

4.1 Abstract

Parenteral medicines are required to be tested for pyrogens (fever-causing agents) in one of two animal-based tests: the rabbit pyrogen test and the bacterial endotoxin test. Understanding of the human fever reaction has led to novel non-animal alternative tests based on in vitro activation of human monocytoid cells in response to pyrogens. Using 13 prototypic drugs, clean or contaminated with pyrogens, we have validated blindly six novel pyrogen tests in ten laboratories. Compared with the rabbit test, the new tests have a lower limit of detection and are more accurate as well as cost and time efficient. In contrast to the bacterial endotoxin test, all tests are able to detect Gram-positive pyrogens. The validation process showed that at least four of the tests meet quality criteria for pyrogen detection. The here validated in vitro pyrogen tests overcome several shortcomings of animal-based pyrogen tests. Our data suggest that animal testing could be completely replaced by these evidence-based pyrogen tests and highlight their potential to further improve drug safety.

4.2 Introduction

Pyrogens, a chemically heterogeneous group of fever-inducing compounds, are derived from bacteria, viruses, fungi or the host himself. Monocytes/macrophages react to microbial products during an immune response by producing endogenous pyrogens such as prostaglandins and the pro-inflammatory cytokines interleukin-1 (IL-1), interleukin-6 (IL-6) and tumor necrosis factor-α (TNFα) (Dinarello, 1999). Depending on the type and amount of pyrogen challenge and the sensitivity of an individual, life-threatening shock-like conditions can be provoked. Consequently, to assure the quality and safety of any pharmaceutical product for parenteral application in humans, pyrogen testing is mandatory.

Depending on the drug, one of two animal-based pyrogen tests is currently prescribed by the health authorities and Pharmacopoeias, i.e., for more than sixty years, the rabbit pyrogen test or the bacterial endotoxins test (BET), often referred to as Limulus amebocyte lysate test (LAL). For the rabbit pyrogen test, sterile test substances are injected intravenously into rabbits and any rise in body temperature is measured. This in vivo test detects various pyrogens but alone the fact that large numbers of animals are required to identify the rare pyrogen-containing samples in routine practice argues against its use if valid alternatives are available. In the past two decades, the declared intention to refine, reduce and replace animal testing, the 3Rs concept (Russel and Burch, 1959) that was implemented e.g. into European legislation in 1986 (European Union, 1986), has led to a reduction in rabbit pyrogen testing by 80% by allowing the BET as an in vitro alternative pyrogen test for many parenteral products.

Bacterial endotoxin comprised largely of lipopolysaccharide (LPS) from the cell wall of Gram-negative bacteria that stimulates monocytes/macrophages via interaction with CD14 and toll-like receptor 4 (TLR4) (Beutler and Rietschel, 2003) is the pyrogen of major concern to the pharmaceutical industry due to its ubiquitous sources, its stability and its high pyrogenicity (Mascoli and Weary, 1979a; Mascoli and Weary, 1979b; Twohy et al., 1984). With the BET, endotoxin is detected by its capacity to coagulate the amoebocyte lysate from the haemolymph of the American horseshoe crab, *Limulus polyphemus*, or the Japanese horseshoe crab, *Tachypleus tridentatus*, a

principle recognized some 40 years ago (Levin and Bang, 1964). In the United States, *Limulus* crabs are generally released into nature after drawing about 20% of their blood and therefore most of these animals survive. However, the procedure still causes mortality of about 30.000 horseshoe crabs per year, which adds to the even more severe threats of the horseshoe crab population such as its use as bait for fisheries, habitat loss and pollution (http://www.horseshoecrab.org). As with the rabbit test the general problem of translation of the test results to the human fever reaction persists. Moreover, although it is highly sensitive, the failure of the BET to detect non-endotoxin pyrogens as well as its susceptibility to interference by, for example, high protein levels of test substances or by glucans impedes full replacement of the rabbit pyrogen test (Roslansky and Novitsky, 1991; Fennrich et al., 1999). Hence, an estimated 200.000 rabbits per year are still used for pyrogen testing in the European Union.

A test system that combines the high sensitivity and in vitro performance of the BET with the wide range of pyrogens detectable by the rabbit pyrogen test is therefore required in order to close the current testing gap for pyrogens and to avoid animal-based tests. With this intention and due to improved understanding of the human fever reaction (Dinarello, 1999), test systems based on in vitro activation of human monocytoid cells have been developed. First efforts date back about 20 years, when peripheral blood mononuclear cells (PBMC) were used to detect endotoxin by monitoring the release of pyrogenic cytokines (Dinarello et al., 1984; Duff and Atkins, 1982). Subsequently, a number of different test systems, using either whole blood, PBMC or the monocytoid cell lines MONO MAC 6 (MM6) (Ziegler-Heitbrock et al., 1988) or THP-1 (Tsuchiya et al., 1980) as a source for human monocytes and various read-outs have been established and were recently reviewed (Poole et al., 2003). Here, the six most prominent of these test systems were formally validated with the aim of developing an evidence-based tool for safer, animal-free and more efficient pyrogen detection and allowing their regulatory acceptance. Formal validation of in vitro methods, i.e. the evaluation of reliability and relevance of a method, was developed by the European Centre for the Validation of Advanced and Alternative

Methods (ECVAM) and is now internationally accepted (Balls et al., 1990; Balls et al., 1995; Worth and Balls, 2002).

4.3 Methods

4.3.1 Rabbit pyrogen test

For this study data from 171 rabbits (kindly provided by Dr. U. Lüderitz-Püchel) accumulated over several years at the Paul Ehrlich Institute, the German Federal Agency for Sera and Vaccines in Langen, were used for analysis. For these experiments, Chinchilla Bastards (Charles River) were injected with 0, 5, 10, 15, 20 EU in 1 ml/kg of *E. coli* LPS (EC5 (Poole and Mussett, 1989) or EC6 (Poole et al., 1997)) in saline (corresponding to 0, 0.5, 1.0, 1.5 and 2.0 EU/kg in 10 ml, the largest volume allowed for injection in rabbits). The fever threshold in rabbits was defined as a body temperature increase of 0.55°C during 180 min after injection. This value represents the mean individual rabbit value at the threshold of 6.6°C of the EP when the maximum of twelve animals is tested (Council of Europe, 2001b).

4.3.2 In vitro monocyte-based tests

Good laboratory practice concordant Standard Operating Procedures of the various methods were made available by ECVAM (www.ecvam.jrc.it). The test systems are summarized by Hartung et al. (Hartung et al., 2001) and detailed in previous work (Eperon and Jungi, 1996; Hartung and Wendel, 1996; Peterbauer et al., 1999; Peterbauer et al., 2000; Poole et al., 1988; Taktak et al., 1991).

4.3.3 Reagents and consumables for all methods

The 2nd International WHO Standard for endotoxin (from *E. coli* O113:H10:K(-) (94/580), which is identical to FDA/USP standard EC6/Lot G was used as the standard endotoxin (Poole et al., 1997). Test materials for validation are specified in the Results section. All consumables were purchased as sterile and pyrogen-free and not specified reagents were pro analysis grade.

4.3.4 PBMC-IL6

4.3.4.1 Blood Collection and preparation of PBMC

Blood donors had to describe themselves as being in good health, not suffering from any bacterial or viral infections for at least one week prior to the donation of blood and not to be taking drugs known to influence the production of cytokines. Using a heparinized (50 µl Fragmin at 10000 IU, Dalteparin, Pharmacia) syringe, 30 ml blood were collected. Within two hours, PBMCs were isolated from 20 ml Lymphoprep (Nycomed, Oslo, Norway), 15 ml PBS and 15 ml of heparinized whole blood by centrifuging at 340 x g for 45 min at room temperature. The PBMC-layer was washed twice with PBS centrifuging at 340 x g for 15 min. The sediment was suspended with RPMI-C (RPMI 1640, Life Technologies™, Paisley, Scotland) with 10 ml/l human serum AB from clotted human male whole blood (Sigma), 10 ml/l L-Glutamine (Life Technologies™), 200 mM, and 20 ml/l Penicillin/Streptomycin solution (Seromed, Vienna, Austria)) after counting in a Neubauer haemocytometer to 1 mio cells/ml. The cells shall be incubated with samples within four hours after blood withdrawal.

4.3.4.2 Protocol for PBMC-IL6

In quadruplicate per each of four blood donors, 100 µl of RPMI-C, 50 µl of samples/controls and 100 µl of gently swirled PBMC were incubated in a 96-well tissue culture plate (Falcon Microtest, Becton Dickinson Labware) at 37°C for 16- 24 hours in an atmosphere of 5% CO_2 in humidified air. After incubation, 50 µl of supernatant from each of the wells was transferred on the ELISA plate ensuring that cells are not aspirated by angling the assay plate.

4.3.4.3 ELISA for PBMC-IL6

2.5 µg/ml coating mouse monoclonal anti-IL-6 antibody (Novartis in-house Clone 16) was added at 200 µl to each well of a 96-well microtitre plate (Nunc-Immuno 96-well plate MaxiSorp, F96; Life Technologies™) at 15 - 25°C for 16 - 24 hours. The washed plate was coated with 200 µl blocking buffer (24.2 g/l Tris(hydroxymethyl)aminomethane, 0.2 ml/l Kathon MW/WT (Christ Chemie AG,

Reinach, Switzerland) and 10.0 g/l bovine serum albumine). Plates were incubated with 200 µg/ml horseradish peroxidase conjugated to sheep anti-IL-6 antibodies (Novartis, in-house) for 2-3 hours at 20-25°C. Shortly before use, 90 ml substrate buffer and 4.5 ml TMB solution (240 mg 3,3',5,5'Tetramethylbenzidine in 5 ml acetone, 45 ml ethanol and 0.3 ml Perhydrol (30% H_2O_2)) were mixed and 200 µl pipetted into each well. After 10-15 minutes, the enzyme reaction was stopped by 50 µl of 5.4% H_2SO_4 per well. The absorbance was measured at 450 nm using 540-590 nm as reference wavelength.

4.3.5 WBT-IL1

4.3.5.1 Blood Collection for WBT-IL1

Blood donors should show no evidence of disease or need of medication during the last two weeks. Blood was collected into heparinized tubes (Sarstedt S-MONOVETTE 7.5 ml, 15 IU/ml Li-Heparin) and used within four hours (Schins et al., 1996).

4.3.5.2 Protocol for WBT-IL1

In this order and in quadruplicates per single blood donor, 1000 µl saline, 100 µl sample/control and 100 µl blood were added to pyrogen-free reaction tubes (Greiner Bio-one tubes, 1.2 ml (polystyrene) or 1.5 ml (polypropylene), Frickenhausen, Germany). Closed tubes were mixed gently, inverted once or twice and then incubated in an incubator or a heating block at 37°C ± 1°C for 10-24 hours. The incubation tubes were mixed thoroughly by inverting them. Incubations were centrifuged for 2 minutes at 10.000 g and the clear supernatant, taking aliquots of ≥ 150 µl, was used for the ELISA (ENDOSAFE-IPT, Charles-River Endosafe, Charleston, USA) following the manufacturer's procedure.

4.3.6 WBT-IL6

4.3.6.1 Blood Collection for WBT-IL6

Blood donors were selected as described for PBMC-IL6. 30 ml blood were drawn and immediately transferred into a 50 ml sterile centrifuge tube containing 300 IU heparin (Fragmin, Pharmacia, diluted 1/10 with saline). The closed tubes were inverted slowly five times to ensure thorough mixing without vortexing and used within four hours (Schins et al., 1996).

4.3.6.2 Protocol for WBT-IL6

In quadruplicate per each of four blood donors, 50 µl of saline, 50 µl of gently mixed blood, 50 µl of samples/controls and 100 µl of saline were incubated in a 96-well tissue culture plate (Falcon Microtest, Becton Dickinson Labware) at 37°C for 16-24 hours in a humid atmosphere of 5% CO_2. After incubation, 50 µl of supernatant from each of the wells was transferred on the ELISA plate ensuring that cells are not aspirated by angling the assay plate. The same IL-6 ELISA as for PBMC-IL6 was used.

4.3.7 MM6-IL6

4.3.7.1 Cell culture for MM6-IL6

The human monocytoid cell line MonoMac-6 was obtained from Prof. H.W.L. Ziegler-Heitbrock (Institute for Immunology, University of Munich, Munich, Germany). Frozen cells from liquid nitrogen were thawed on ice. Cells were transferred to a 50 ml centrifuge tube, 10 ml RPMI (+4°C) (e.g. Life Technologies™) added and then centrifuged at 100 x g for 5 min at +4°C. Afterwards the cells were resuspended in 10 ml RPMI-M (containing 10% ml heat-inactivated low-pyrogen foetal calf serum, 2 mM L-Glutamine, 0.1 mM MEM non-essential amino acid, 0.23 IU/ml Bovine insulin, 1 mM Oxaloacetic acid, 1 mM Sodium pyruvate, 20 mM HEPES). After a wash step, cells were transferred to a 25 cm^2 tissue culture flask and incubated at 37°C, with 5% CO_2 and high humidity. The number of viable cells was determined by Trypan blue

exclusion using a haematocytometer. The cells were passaged with 2 x 10^5 cells/ml twice a week.

4.3.7.2 Protocol for MM6-IL6

To pre-incubate the cells for a test, 30-50 ml of cell suspension were centrifuged at 100 x g for 8 min at room temperature and resuspended in RPMI-C (as RPMI-M, but only 2% heat-inactivated foetal calf serum) at a final concentration of 4x10^5 cells/ml. The cells were incubated approximately 24 hours at 37°C, 5% CO_2 and high humidity. Cells were washed and counted as above, diluting to 2.5 x 10^6 viable cells/ml, just prior to addition to the culture plate. In quadruplicates, 50 µl of samples/controls, 100 µl of RPMI-C and 100 µl of gently swirled MM6 were incubated in 96-wells tissue culture plates at 37°C for 16-24 hours with 5% CO_2 and humidified air. After incubation, 50 µl of supernatant from each of the wells was transferred on the ELISA plate ensuring that cells are not aspirated by angling the assay plate. The same IL-6 ELISA as for PBMC-IL6 was used.

4.3.8 THP-Neo

4.3.8.1 Cell culture for THP-Neo

THP-1 cells were obtained from the American Type Culture Collection (ATCC, TIB-202). 6 x 10^6 cells were seeded in 60 ml medium (RPMI 1640 supplemented with 10% (v/v) FCS (high-quality lots with the lowest endotoxin content available (< 30 pg/ml) were chosen, e.g. Biochrom, Berlin, Germany) in 75 cm^2 culture flasks. Flasks were incubated in upright position at 37° C with 5% CO_2 and humidified air. On the fourth day of culture, further 30 to 60 ml (depending on the culture doubling time) of culture medium were added and cells were incubated for another three days. If cells from freshly thawed stocks are used, they have to be grown for two to three weeks in order to ensure that they divide properly before using them for tests. Furthermore, cells should not be kept in culture for more than four months but new cultures should be started from frozen stocks at regular intervals. Cells were counted with a hemocytometer and cell viability by trypan blue exclusion was ≥ 90%. Tubes with 2.5 x

10^7 cells (for one plate) were centrifuged at 400 x g and 20° C for 7 min and resuspended in 20 ml medium, 2 mM L-glutamine and 50 μM 2-mercaptoethanol.

4.3.8.2 Protocol for THP-Neo

100 μl IFNγ (human, recombinant, endotoxin content < 0.1 EU/mg; Gammaferon 50, Rentschler Biotechnologie, Laupheim, Germany) stock solution (6250 U in 100 μl medium, 110 μl aliquots) were added to 20 ml of cell suspension and mixed well. 200 μl/well of mixed cell suspension were added to a 96-well cell culture microtiter plate. After incubation for 30 min, 50 μl of vortexed samples/controls (in quadruplicate) were added and put on an orbital plate shaker for 2 min at room temperature and 500 rpm. After 18-22 hours of incubation, 150 μl of supernatant were collected and frozen and/or directly processed with the neopterin ELISA (Elitest Screening, Brahms Diagnostica, Berlin, Germany) according to the manufacturer's protocol.

4.3.9 THP-TNF

4.3.9.1 Protocol for THP-TNF

THP-1 cells (obtained from ATCC or ECACC) were used. Subclones from this cell line prepared in-house showed a higher sensitivity towards LPS. Cells were cultured in RPMI (1% L-glutamine, 1% HEPES, 1% Penicillin/streptomycin solution, 1% Sodium pyruvate, all from Biochrom (Berlin, Germany), 1% nonessential aminoacids for MEM, 0.4% MEM vitamin solution, 0.5% β-mercaproethanol (10 mM), all from Invitrogen (Basle, Switzerland), and 12% heat-inactivated low-pyrogen FCS in 6-well plates or T25 flasks at 37°C in a humidified 5% CO_2 incubator. They were passaged once weekly. When new cells are required for an assay, cells from a cryovial were thawed two to three weeks before use. For the last passage prior to the test, terminal differentiation was induced by cultivating the cells in the presence of sterile-filtered calcitriol (1,25-dihydroxy vitamin D3, Sigma or Hoffmann-La Roche, Basle, Switzerland) (10 μg/ml) for 44-48 hours. Cells were collected, centrifuged and resuspended in culture medium containing calcitriol (final concentration 100 ng/ml). They were counted and adjusted to 1 to $1.25x10^6$ cells/ml. Cells were cultured for 44-

48 hours in T25 flasks. Then, terminally differentiated cells were harvested and counted using a haematocytometer and trypan blue. Cells were diluted to $1.25x10^6$ cells/ml and 200 µl of suspension were dispensed into each well of the above 96-well cell culture plate containing already 50 µl of sample/control in quadruplicates. Plates were incubated for 16-24 hours at 37°C and 5%CO_2.

4.3.9.2 TNFα ELISA for THP-TNF

Non-sterile plates Dynex PF microtiter 'flat bottom' styrene 96-well plates (Dynex Tech., Worthing, UK) were rinsed extensively with pyrogen-free PBS. The plates were coated with 1 µg/ml monoclonal antibody 101-4 against human TNFα (a generous gift from Dr. T Meager, Division of Immunobiology, NIBSC, UK) at 100 µl/well and 4 ºC overnight. 50 µl of sample/control (in quadruplicates) or duplicates of TNFα standards (250, 62.5, 15.6, 3.9, 0.98, 0.24, 0 U/ml, NIBSC) were added for 16-24 hours at 37°C and 5% CO_2. An aliquot of the detecting antibody (biotinylated goat-anti-human TNF-α from the Duoset kit, R&D) was diluted 180-fold, using dilution buffer (0.1% bovine serum albumin, 0.1% Tween 20, in 20 mM Tris, 100 mM NaCl, pH 7.2-7.4). 100 µl were dispensed to each well for two hours at room temperatue. After washing, 100 µl Streptavidin-peroxidase conjugate (R&D) was added for 20 min. After washing, 100 µl of TMB (Sigma) were dispensed and incubated in the dark before reading at 650 nm. Incubation time was chosen so that 250 U/ml TNFα value had an OD $\geq$ 1.5.

4.3.10 Data analysis

The rabbit fever reaction was modeled by regression techniques applied to the logarithmically transformed data. The within- and between-laboratory reproducibility were assessed comparing the resulting classifications by means of simple matching, i.e. the proportions of identically classified samples, as a measure of similarity. In case of the within-laboratory reproducibility, where three independent but identical runs were performed, the mean similarity was calculated.

A one-sided t-test, assuming hazard and thus designed to proof safety of a tested compound, was employed as a so-called prediction model (PM) to dichotomize the

test results into a classification of either ‘pyrogenic’ or ‘non-pyrogenic’. The t-test compares the data of a given sample against the data of the standard positive control of 0.5 EU/ml, which is performed in parallel. It is calculated with the log-transformed data and a local significance level of 1% was chosen in order to increase safety. If this test resulted in a significant p-value, i.e. smaller than 1%, then the considered sample was classified as non-pyrogenic, and as pyrogenic otherwise. This means that a negative sample had to be significantly lower than 0.5 EU/ml. The levels of contaminations chosen were 0, 0.25, 0.5 (twice) and 1 EU/ml. According to the rabbit model, 0 and 0.25 EU/ml were considered as non-pyrogenic samples and 0.5 and 1 EU/ml as pyrogenic samples. Having thus defined the reference standard, i.e. the ‘true’ contamination level, we calculated via 2x2-contingency tables the performance parameters sensitivity, i.e. the probability of a correct positive classification, and specificity, i.e. the probability of a correct negative classification. Confidence intervals for these parameters were calculated with the Clopper and Pearson method based on the F distribution (Clopper and Pearson, 1934).

4.4 Results

4.4.1 The limit of endotoxin detection in rabbits

Employing regression techniques, the temperature data from 171 rabbits could be modeled by the equation $y = 0.217 * (EU + 1)^{0.508}$, where y is the expected temperature increase for a given concentration EU/ml (Fig. 1). This approach was recently described in more detail and further exploited (Hoffmann et al., in press).

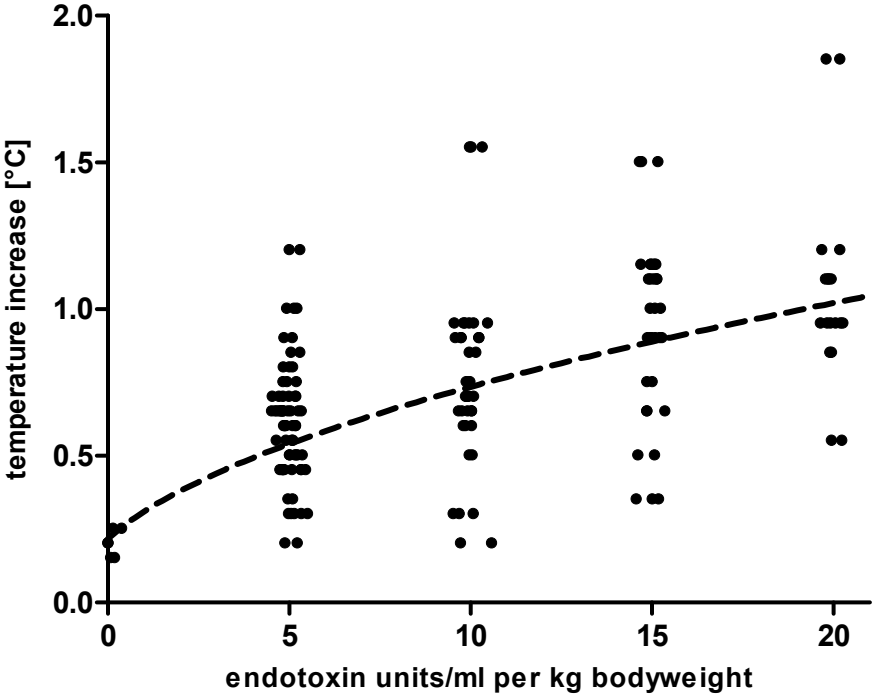

Figure 1. Temperature increase of 171 rabbits upon endotoxin injection with a fitted regression line

The maximum temperature increase in ºC within 180 minutes after endotoxin injection of 171 rabbits is presented. The mean temperature increase, modeled with regression techniques, is indicated by the dotted line.

The model indicated that 50% of the animals develop fever, i.e. showing a 0.55°C rise of body temperature within 180 min after injection, in response to 5.22 EU per kg body weight of endotoxin with a 95%-confidence interval of 4.24 to 6.21 EU/ml. Only at 20 EU per kg of body weight, all animals showed an increase in temperature of 0.55°C or more. We deduced from these data that a sample concentration of 0.5 EU/ml represents the required limit of detection (LoD) that alternative pyrogen tests must meet. This assumption takes into account the fact that the largest volume allowed for injection into rabbits is 10 ml per kg, corresponding to 0.5 EU/ml for injections at 10 ml/kg. Thus, the concentration of 0.5 EU/ml was defined as the threshold between pyrogenic and non-pyrogenic samples.

4.4.2 Prevalidation of the novel in vitro pyrogen tests

Before prevalidation, the test-developing laboratories that took part in the study compiled standard operating procedures for the alternative tests. This required an intensive phase of test optimization and standardization in order to allow the transfer of the tests. A standard curve of endotoxin in saline including the 0.5 EU/ml concentration as the threshold for pyrogenicity was included in all tests. Only if the 0.5 EU/ml endotoxin standard was detectable, did the test run qualify for analysis. Before prevalidation was started, the naive laboratories proved evidence of successful transfer of the respective test systems (data not shown). Prevalidation was then carried out with twelve blinded samples. These consisted of three drugs spiked with either pyrogen-free saline (clinical grade 0.9% NaCl) or with reference endotoxin. Two negative, i.e. pyrogen-free samples, and two LPS-containing, i.e. pyrogenic samples (0.5 EU/ml and 1.0 EU/ml sample concentration, respectively) were tested. The concentration of 0.5 EU/ml was the limit of detection defined for the rabbit pyrogen test (see above). The drugs used were Gelafundin®, a volume-replacement therapy for transfusion with high protein (gelatine) content (B. Braun Melsungen AG, Melsungen, Germany), Jonosteril®, an electrolyte infusion (Fresenius AG, Bad Homburg, Germany) and Haemate®, a factor VIII preparation (Aventis Behring GmbH, Marburg, Germany). In addition, a positive control (0.5 EU/ml LPS in saline) and a negative control (endotoxin-free saline) were included. Each test was performed three times in the respective developing laboratory (DL) as well as in two naive laboratories (NL).

Table 1 summarizes the six novel test systems used, their major characteristics, their performance regarding reproducibility, which was assessed before the blinding code was broken, as well as sensitivity and specificity.

Test	System	Read-out	Ref.	Within-laboratory reproducibility (%)	Between-laboratory reproducibility (%)	Sensitivity (%)	Specificity (%)
WBT-IL6	whole blood	IL-6	(Poole et al., 2003)	DL: 83.3 NL1: 94.4 NL2: 100	DL-NL1: 72.2 DL-NL2: 72.2 NL1-NL2: 96.3	72.2	92.6
WBT-IL1	whole blood	IL-1β	(Hartung et al., 2001)	DL: 88. 9 NL1: 95.8 NL2: 94.4	DL-NL1: 91.7 DL-NL2: 76.8 NL1-NL2: 67.8	72.0	100.0
PBMC-IL6	PBMC	IL-6	(Hartung and Wendel, 1996)	DL: 94.4 NL1: 100 NL2: 94.4	DL-NL1: 80.6 DL-NL2: 86.1 NL1-NL2: 88.9	87.0	98.1
MM6-IL6	MM6 (Ziegler-Heitbrock et al., 1988)	IL-6	(Poole et al., 2003)	DL: 100 NL1: 94.4 NL2: 94.4	DL-NL1: 97.2 DL-NL2: 88.9 NL1-NL2: 86.1	72.2	100.0
THP-TNF	THP-1 clone	TNFα	(Taktak et al., 1991)	DL: 94.4 NL1: 83.3 NL2: 55.5	DL-NL1: 90.7 DL-NL2: 67.6 NL1-NL2: 65.7	66.7	88.9
THP-Neo	THP-1 parental (Tsuchiya et al., 1980)	neopterin	(Eperon and Jungi, 1996)	DL: 100 NL1: 94.4 NL2: 77.7	DL-NL1: 97.2 DL-NL2: 50.0 NL1-NL2: 51.8	88.9	72.2

Table 1. Novel pyrogen tests and their performance in prevalidation

Protocols for all methods are listed in Poole et al. (2003) and in the Methods section. All tests include dilution of the sample by 1:5 with the exception of the WBT-IL-1 test that requires a 1:12 dilution of the sample. The WBT-IL6 and the PBMC-IL6 tests combine data from three respectively four blood-donors per run, the WBT-IL1 from one donor per run. Samples and controls were tested in quadruplicate in each of the tests. DL denotes developing laboratory, NL1 and NL2 the two naive laboratories. The sample size analyzed for sensitivity and specificity was 108 for all tests besides WBT-IL1 (100 samples). Sensitivity describes the probability to correctly classify positive samples and specificity describes the probability to correctly classify negative samples.

As can be seen, the predictive capabilities of the various tests were encouraging, particularly in the light of the restricted stability of endotoxin spikes at the borderline concentration of 0.5 EU/ml. Although all tests were successfully transferred to the naive laboratories during the preparatory phase of prevalidation, this optimal performance could not be maintained for the two test systems using THP-1 cells, as is

reflected by the comparatively low between-laboratory reproducibility between the developing laboratory and one of the naive laboratories for each. The lower specificity of the THP-Neo test was entirely caused by misclassification in NL2. Furthermore, prevalidation also revealed that, despite preceding interference testing and diluting of the drugs accordingly, interference/recovery problems persisted in some cases, as is reflected by the values for sensitivity.

4.4.3 Validation phase

For the validation phase, 10 drugs with five blinded spikes each (0 (i.e. pyrogen-free), 0.25, 0.5 (twice) and 1 EU/ml) were tested, again in three laboratories, i.e. the DL of a test and the two NLs, respectively. To avoid the possibility that different dilutions of the drugs were tested depending on their different interference with different test systems, all drugs were tested at their maximum valid dilution (MVD), thus adopting the rationale of the pharmacopoeial BET reference (limit) test. The MVD is calculated from the endotoxin limit concentration (ELC in EU/ml) defined for a drug by the European Pharmacopoeia (Council of Europe, 2001a), divided by the threshold of pyrogenicity as the limit of detection (LoD), i.e. 0.5 EU/ml. Drugs, sources, ELCs and MVDs (= ELCs/LoD, where LoD=0.5) are summarized in Table 2.

Drug	Source	Agent	Indication	ELC (EU/ml)	MVD (-fold)
Glucose 5% (w/v)	Eifelfango GmbH	glucose	nutrition	35	70
Ethanol 13% (w/v)	B.Braun AG	ethanol	diluent	17.5	35
MCP®	Hexal AG	metoclopramid	antiemetic	175	350
Orasthin®	Aventis Pharma GmbH	oxytocin	initiation of delivery	350	700
Binotal®	Aventis Pharma GmbH	ampicillin	antibiotic	70	140
Fenistil®	Novartis Consumer Health GmbH	dimetindenmaleat	antiallergic	87.5	175
Sostril®	GlaxoSmithKline GmbH	ranitidine	antiacidic	70	140
Beloc®	Astra Zeneca GmbH	metoprolol tartrate	heart dysfunction	70	140
Drug A		0.9% NaCl		17.5	35
Drug B		0.9% NaCl		35	70

Table 2. Test substances for the validation phase

Drugs were selected by a selection committee which excluded the developing laboratories and included experts. Drugs A and B which were saline only were included as further controls using notional ELCs.
Drugs were obtained from Eifelfango GmbH (Bad Neuenahr-Ahrweiler, Germany), B. Braun AG (Melsungen, Germany), Hexal AG (Holzkirchen, Germany), Aventis GmbH (Bad Soden, Germany), Novartis GmbH (München, Germany), GlaxoSmithKline GmbH (München, Germany) and Astra Zeneca GmbH (Wedel, Germany). ELCs of drugs were calculated according to European Pharmacopoeia (Council of Europe, 2001a).

While the tests using whole blood, PBMC and MM6 cells performed well in all three test laboratories in terms of reproducibility (Table 3), technical problems with the two tests using THP-1 cells were obvious. For the THP-TNF test this was caused by a batch of TNFα-ELISA plates sent out to the two NLs that did not satisfy the quality criteria with regard to detection limit when used with cells. For the THP-Neo test, the

technical problems in NL2 persisted such that the quality criteria defined in the SOP were not met. The tests could not be repeated due to the limited time frame of validation and for logistical reasons. Therefore, for the THP-TNF assay only the data from the DL and for the THP-Neo assay only the data from the DL and from NL1 could be analyzed. Sensitivity and specificity were 76.7% and 78.9% for the THP-TNF assay (sample size = 40) and 93.3% and 47.5% for the THP-Neo assay (sample size = 100). The data for the other four tests are summarized in Table 3.

Test	Between-laboratory reproducibility (%)	Sample size: sensitivity[#]	Sensitivity (%)	Sample size: specificity	Specificity (%)
WBT-IL6	DL-NL1: 85.4 DL-NL2: 85.4 NL1-NL2: 92.0	89	88.9	59	96.6
WBT-IL1	DL-NL1: 72.9 DL-NL2: 81.6 NL1-NL2: 70.2	88	72.7	59	93.2
PBMC-IL6	DL-NL1: 84.0 DL-NL2: 86.0 NL1-NL2: 90.0	90	92.2	60	95.0
MM6-IL6	DL-NL1: 90.0 DL-NL2: 89.6 NL1-NL2: 83.3	89	95.5	59	89.8
Rabbit[†]	-	-	57.9	-	88.3

Table 3. Validation of the predictive capability of novel pyrogen tests

[#] sample sizes are reduced by outlier exclusion defined in the study protocol
[†] parameters calculated by the fitted regression model

Almost all misclassifications, either false negatives or false positives, occurred around or at the defined classification threshold, i.e. for the contaminations of 0.25 and 0.5 EU/ml. Confidence intervals (CI) with a significance level of 5% were calculated for sensitivity and specificity. By focusing on the lower bounds of CI (Fig. 2), a worst-case scenario can be conducted by which the likelihood of underestimation of pyrogen

content is maximized and thus possible negative consequences for health can be estimated.

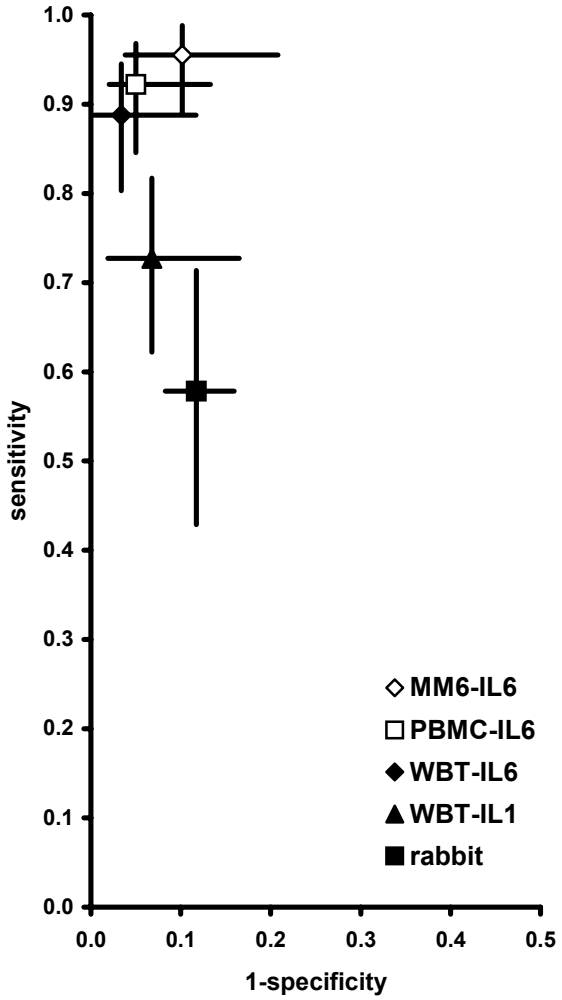

Figure 2. Sensitivity and specificity of four in vitro assays in the validation study and modeled rabbit test performance with 95%-confidence intervals

The sensitivity and specificity resulting from the pre-defined prediction model and considering samples with 0 and 0.25 EU/ml as non-pyrogenic and with 0.5 and 1 EU/ml as pyrogenic are presented with their corresponding 95% confidence intervals for four validated tests. Similarly, the respective parameters were calculated with the rabbit model. As performance improves towards the upper left of the graph, all validated tests outperform the rabbit test.

The lower predictive capability of the WBT-IL1 test as compared to the WBT-IL6 and the PBMC-IL6 test can be explained by the one-donor approach used for the WBT-IL1 test and the multiple-donor approach used for the other tests that is more conservative

and laborious, but decreases the probability for false-negative classification. For the THP-TNF assay, the lower bounds of CI for sensitivity and specificity were 60.6% and 55.2%, respectively. For the THP-Neo assay, the respective lower bounds were 78.0% and 38.7%. Applying this kind of analysis to the in vivo assay employing the regression model based on the data from the rabbit pyrogen test yields a sensitivity of 57.8% and a specificity of 88.3% (Table 3) with confidence intervals also presented in Fig. 2. Thus, the novel pyrogen tests listed in Table 3 show parameters of performance outperforming the rabbit pyrogen test.

An additional analysis, which could be conducted with the available data, supports this conclusion. According to their SOPs, the four systems included an uncontaminated negative control, i.e. saline, and another positive control of 1 EU/ml. For each of these two controls we adapted the prediction model described above: First, we compared the blinded samples against the response of 1 EU/ml control. Therefore, we constructed a modified prediction model using the 1 EU/ml control response instead of the positive control of 0.5 EU/ml, which is denoted in the following by PM1. In doing so, the true classification of the samples changed, as now only the samples spiked with 1 EU/ml were considered as pyrogenic and the other samples as non-pyrogenic. Second, with a modified prediction model, denoted as PM0, classifying a sample as pyrogenic when the response was significantly larger than the negative control response (significance level 1%), we compared all spikes against this control. Again, the true classification of the samples needed to be adjusted considering the contaminated samples (0.25, 0.5, 1.0 EU/ml) as pyrogenic and the unspiked samples as non-pyrogenic. The resulting sensitivities and specificities are summarized together with the results from the original PM for the four test systems in Fig. 3. All tests performed best for PM0, where the sum of these two parameters was at least 1.90, while WBT-NI even resulted in the maximum sum of 2.

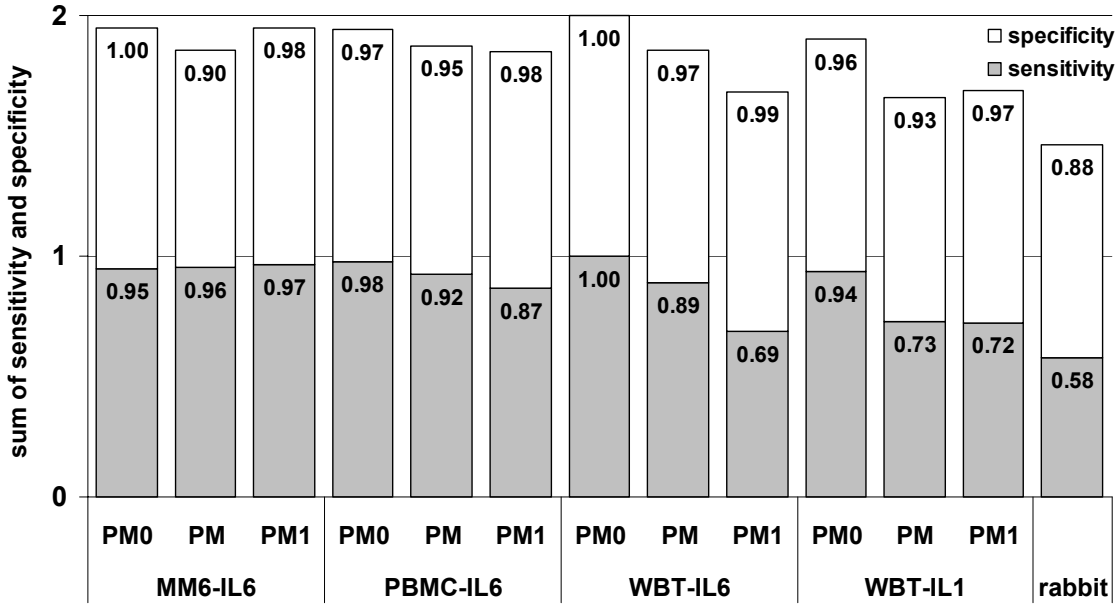

Figure 3. Sum of sensitivity and specificity resulting from three prediction models for four in vitro assays in the validation study

The validation data of four tests were analysed with three prediction model employing different controls for comparison and thus defining the true classification of the samples (non-pyrogenic vs pyrogenic) accordingly. The test accuracies is described for each test and prediction model by the sum of specificity and sensitivity allowing also for individual parameter assessment. For comparison, the rabbit test performance according to the pre-defined prediction model is added.

The DLs also tested lipoteichoic acid (LTA) from *Bacillus subtilis*, a BET-negative Gram-positive compound that activates cytokine release from human monocytes (Morath et al., 2001; Morath et al., 2002) prepared according to Morath et al. (2002), which was clearly detectable by the novel tests.

4.5 Discussion

Previous work (Eperon and Jungi, 1996; Hartung et al., 2001; Hartung and Wendel, 1996; Nakagawa et al., 2002; Peterbauer et al., 1999; Poole et al., 1988; Poole et al., 2003; Taktak et al., 1991) had established that different sources of human monocytoid cells are valuable tools for mimicing the human fever reaction in vitro. Not only can these cells detect the important pyrogen LPS from *E. coli* and other Gram-negative

bacteria but also a number of compounds involved in the immune response to Gram-positive bacteria such as LTA (Morath et al., 2001; Morath et al., 2002; Schindler et al., 2003), exotoxins (Eperon et al., 1997; Peterbauer et al., 1999), cell wall components like muramyl dipeptide (Eperon et al., 1997) or peptidoglycan (Harada et al., 1968), *S. aureus* Cowan (SAC) (Nakagawa et al., 2002) or DNA (Peterbauer et al., 1999) as well as poly(I:C) (Nakagawa et al., 2002), a synthetic double-stranded RNA used as a virus model compound in fever research. It was also established that these novel test systems overcome limitations of the BET and yield results comparable to the rabbit pyrogen test (Eperon et al., 1997; Nakagawa et al., 2002; Peterbauer et al., 1999; Spreitzer et al., 2002). For the first time, six of these monocytoid-cell based in vitro pyrogen tests were formally validated in the present study. For this purpose, a harmonized analysis procedure was established that allowed the direct comparison of the different tests and incorporated various safety aspects. A conservative statistical approach showed that four test systems met the criteria for safe detection of pyrogens. The two test systems based on the use of THP-1 cells posed problems in performance. These were related to insufficient transfer to one naive laboratory (THP-Neo) and to use of an ELISA batch for the one-plate assay format (THP-TNF) that, although qualifying for the detection of TNFα did not qualify for the use with cells and caused their prestimulation. Both problems became obvious only during validation and could not be overcome within the tight schedule of validation. Thus, for these two systems additional validation processes would be required. However, the data obtained for the other four test systems clearly suggest that these have reached a stage of development that makes them suitable for use in pyrogen testing as replacements for the rabbit pyrogen test.

For the purpose of this study, a threshold value of 0.5 EU/ml was chosen on the basis of historical data from rabbit tests carried out in a national control authority. This approach was conservative as only 50% of animals of the very sensitive strain used showed a febrile reaction at this concentration. Additionally, in order to be classified negative, the samples had to be, according to the PM, significantly lower than 0.5 EU/ml. On the one hand, the enormous challenge to the models by placing two samples at the threshold of 0.5 EU/ml, which had to be classified positive, resulted in

reduced sensitivities. On the other hand including a sample with 0.25 EU/ml, which had to be identified as negative, was the reason for almost all false-positive classifications resulting in the reduced specificities without representing any safety concern. When tested against the negative control (PM0), i.e. when samples which are not significantly different from the negative control were considered as pyrogen-free, the tests performed even better, i.e. with increased sensitivity. However, this approach increases consumers' safety on the cost of rejecting drugs, whose minor pyrogenic contamination would not induce adverse health effects in humans. At the same time, this reflects the fact that the study design put main emphasis on the threshold of 0.5 EU/ml. Similarly, decreased sensitivity when given the task of identifying 1 EU/ml as threshold value shows that the tests were especially designed for the threshold of 0.5 EU/ml. Since the test performance when changing the threshold is still acceptable or even better, the robustness of the alternative tests is underlined.

In summary, this study provides thus evidence of the validity of these tests and should facilitate the regulatory acceptance of these novel tests and lead to their introduction into Pharmacopoeias.

4.6 Acknowledgements

We thank U. Lüderitz-Püchel from the Paul Ehrlich Institute, Langen, Germany, for providing rabbit pyrogen test data.

This work was supported by the European Union [QLRT-1999-00811].

5 Optimisation of pyrogen testing in parenterals according to different pharmacopoeias by probabilistic modelling

Sebastian Hoffmann[a], Ursula Lüderitz-Püchel[b], Thomas Montag-Lessing[b], Thomas Hartung[a]

[a] European Commission, JRC - Joint Research Centre; Institute for Health & Consumer Protection; ECVAM – European Centre for the Validation of Alternative Methods; 21020 Ispra, Italy

[b] Paul-Ehrlich-Institut; Federal Agency for Sera and Vaccines; 63225 Langen (Germany)

Journal of Endotoxin Research (2005), in press

5.1 Summary

The rabbit test to detect pyrogenic contamination in parenterals is crucial to ensure patient safety. The pharmacopoeial tests in Europe, the United States and Japan are based on the fever reaction of rabbits, but differ in their experimental design and in their algorithms to assess contamination. Employing an international reference endotoxin, fever can be induced in rabbits. Data from 171 rabbits built the base for probabilistic modelling of the fever reaction and for the comparison of the pharmacopoeial tests. The rabbit fever reaction could be modelled as a function of the amount of injected endotoxin (per kg bodyweight) by linear regression. Combining the pharmacopoeial algorithms of the rabbit pyrogen test with the developed model allowed analysis of differences regarding test results and animal consumption. This showed that the assessment of pyrogenic contamination strongly depends on the respective pyrogen test stipulated by regulations. Additionally, the approach was used

to develop a new experimental design. Two specific versions of this design resulted in a reduction of the number of animals used by about 30% while the safety of the test was maintained. A need for harmonisation is evident, allowing optimisation of the experimental design, which promotes animal welfare.

5.2 Introduction

Pyrogens, a chemically heterogeneous group of fever-inducing compounds, originate from bacteria, viruses, fungi or the host himself. Depending on the amount and type of pyrogen challenge as well as the sensitivity of an individual, life-threatening shock-like conditions can be provoked. Consequently, pyrogen testing is mandatory for any pharmaceutical product for parenteral application in humans in order to assure its quality and safety.

An international endotoxin standard, WHO-LPS 94/580 (*E.coli* O113:H10:K) (Poole et al., 1997; Poole and Mussett, 1989), is available, which is derived from Lipopolysaccharide of the Gram-negative bacteria *E.coli*, representing the most prominent and a highly potent pyrogen (Poole and Mussett, 1989; Weary et al., 1982: Mascoli and Weary, 1979a; Mascoli and Weary, 1979b) . In the pharmacopoeia of the United States (USP) (The United States Pharmacopeial Convention, 2000), of Europe (EP) (Council of Europe, 2001b) and of Japan (JP) (Society of Japanese Pharmacopoeia, 1996) the reference method for pyrogen testing is a rabbit experiment. This animal experiment measures the fever reaction of rabbits after injection of a parenteral drug in order to assess the drugs safety. The classification of a drug as either pyrogenic or non-pyrogenic is based on the maximum rise in body temperature of rabbits, which is recorded during three hours after injection. Nevertheless, the tests in the Pharmacopoeias differ in the experimental design and the analysis of the data. Nowadays, due to high quality standards in manufacturing, pyrogenic samples are found rarely. Therefore, the need of an animal experiment still consuming several hundred thousands animals per year world-wide is questioned.

Much effort, including the field of pyrogen testing, was spent in recent years to implement the 3R of Russel and Burch (Russel and Burch, 1959) - i.e. refinement,

reduction and replacement of animal experiments - in toxicity and safety testing. Focusing on the reduction aspects in this process, especially biometric modelling and simulations optimising and redesigning animal experiments were applied to lower animal consumption. Examples can be found in the field of ecotoxicology (Hutchinson et al., 2003), acute toxicity (Bruce, 1987; Diener et al., 1998; Schlede et al., 1995; Stallard et al., 2003; Stallard et al., 2004; van den Heuvel et al., 1990) and biologicals (Weisser and Hechler, 1997). A study by ECVAM, the European Centre of Validation of Alternative Methods, which was established in 1991 in order to guide, promote and support the 3R, to validate alternatives to the rabbit pyrogen test (Hartung et al., 2001), which are based on the human fever reaction, triggered an in-depth analysis of the respective rabbit in vivo experiment. As it represents the reference method for pyrogen testing, it was also employed in the validation study as the reference standard. Different numbers of rabbits, experimental designs, and criteria for drug safety assessment between pharmacopoeias prompted us to investigate retrospectively the rabbit fever reaction and these differences.

5.3 Methods

5.3.1 Data and statistical methods

The Paul Ehrlich Institut (PEI), the German Federal Agency for Sera and Vaccines, provided a high-quality set of body temperature data from 171 Chinchilla Bastards from Charles River, Kisslegg, Germany, injected with reference endotoxins EC-6 (Poole et al., 1997) or *E.coli* O113 from Associates of Cape Cod, Falmouth, Massachusetts. They were produced during internal quality controls in several studies over the past five years. The rabbits were injected with 0, 5, 10, 15, 20 endotoxin units (EU) in 1 ml/kg bodyweight of standard endotoxin dissolved in saline. Animals receiving an injection of endotoxin-free saline, which is denoted here as 0 EU/kg, can be regarded as negative controls. The maximum temperature increase in ºC, denoted as y, was calculated for each of the 171 animals.

We modelled the temperature rise induced by standard endotoxin expressed in endotoxin units per kg rabbit bodyweight by fitting a linear regression to the

logarithmically transformed data, i.e. ln(y) = intercept + ln(EU+1)*slope + error, with the statistical software package S-PLUS 6.1 (© Insightful, Seattle, WA). Retransforming this relationship allowed e.g. the calculation of the mean endotoxin concentration, denoted as $\overline{E}\,\overline{U}$, causing a fixed temperature increase y by the equation $\overline{E}\,\overline{U} = e^{((\ln(y)-\text{intercept})/\text{slope})-1}$. Estimation of the regression parameters for the provided data resulted in an intercept of –1.526 and in a slope of 0.508. The transformed data together with the regression line are presented in Figure 1, in which both axes are logarithmically scaled. The coefficient of determination of the regression, denoted as R^2, was 0.486. As an analysis of residuals did not reveal any major systematic deviation from the model, which is indicated by the symmetrical spread of the data around the regression line, the coefficient of determination was strongly influenced by the biological variability of the rabbit fever reaction. The variability, expressed by the residual standard error of the regression, was calculated as 0.389. Identical modelling of four larger subsets of the data, each produced in a single experiment, with sample sizes of 18 up to 54, resulted in similar parameter estimation. The estimated intercepts lay in between –1.659 and –1.383 and the estimated slopes in between 0.462 and 0.560. As the standard errors ranged from 0.291 to 0.409, it could particularly be concluded that the combination of the data from all subsets did not increase the standard error.

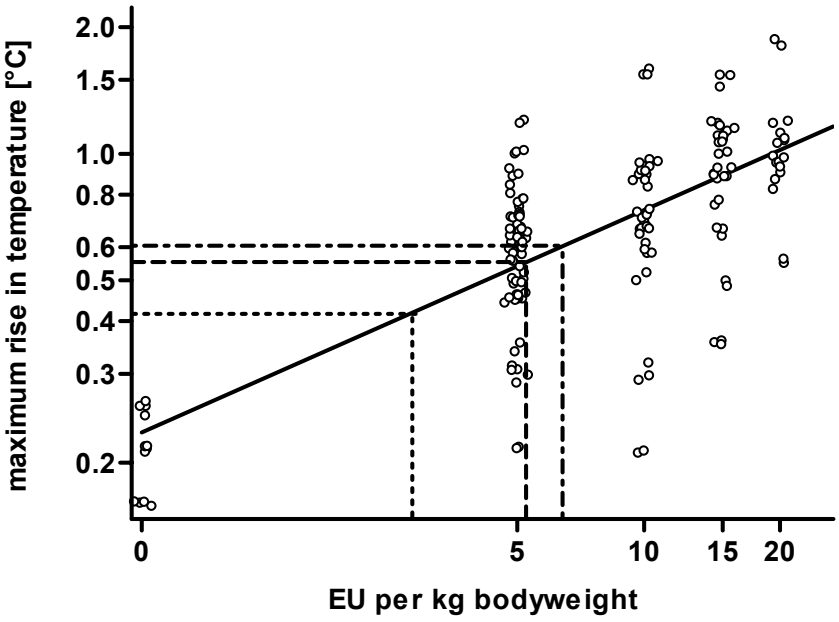

Figure 1. Biometrical model of rabbit fever describing the temperature rise induced by standard endotoxin per kg bodyweight based on the data of 171 rabbits (circles) with the regression line (straight) and the pharmacopoeial rabbit fever thresholds (USP: ····· ; EP: – – ; JP: – · –). A jitter-effect was used to present the data to make identical data distinguishable.

To be able to confirm the appropriateness of the model, a second, smaller data set from the Department of Industrial Pharmacy at the Federal University of Santa Maria (UFSM), Brazil, generously provided by Dr. Sergio Dalmora, was analysed with the same technique. It comprised 54 New Zealand White rabbits injected with six different endotoxin concentrations of the international WHO-standard. The endotoxin, dissolved in saline, was administered per kg bodyweight up to a maximum of 60 EU. The estimated intercept was –2.909, the slope 0.808 and the residual standard error 0.176. A residual analysis revealed a good fit of the model, while R^2 was 0.931. The difference of the coefficients of determination of the two data sets was mainly due to the different variabilities. In contrast to the first data set, the temperature rise in the second rabbit group was less variable. Furthermore, these rabbits were less sensitive. Due to the properties of the rabbit fever model, the sensitivity to endotoxin could not be derived directly from the parameters. However, it became evident in a direct

comparison, as the first regression curve (PEI data) was superior to the second (UFSM data) up to 100 EU/kg (not shown). Similar findings relating differences between results of a Pyrogen Test and the rabbit strain were already described in two large-scale collaborative studies (Dabbah et al., 1980; Poole and Mussett, 1989).
With the developed model, rabbit fever responses could be simulated by random sampling from a normal distribution for each transformed EU-concentration. For this simulation the EU-concentration determined the mean and the residual standard error of the regression the standard deviation of the respective normal distribution. Fever reaction data of 360.000 rabbits per concentration were generated in a grid of 0.2 EU/kg from 0 to 12 EU/kg and in a grid of 0.5 EU/kg up to 25 EU/kg. In parallel, the algorithms of the Pyrogen Tests of the three pharmacopoeias were employed in S-PLUS. Each algorithm was simulated 100.000 times for each endotoxin concentration by randomly recruiting the required numbers of animals from the generated data. This allowed us to calculate the probability of a classification as pyrogenic as well as the number of animals used to terminate the algorithms as functions of the concentration.

5.3.2 The pharmacopoeial algorithms

The EP algorithm consists of a maximum of four consecutive steps with three rabbits each, in which all steps allow for pyrogenic or non-pyrogenic classification (Table 1). Additionally, the first three steps include temperature ranges, which demand proceeding to the next step. In the EP, the criterion for classification is the sum of the temperature rises of all rabbits tested. Focusing for example on the first step, the crucial sums are 1.15°C and 2.65°C: a sum of temperature rises below or equal to the lower limit results in a negative and above the upper limit in a positive classification. Sums in-between demand the testing of additional three rabbits in a second step. In the last step the tested product fails, if the sum of temperature rises of all twelve rabbits exceeds 6.60°C, and passes otherwise. The JP algorithm consists of two steps with three and five animals (Table 1). If in the first step at least two of the three rabbits show an individual temperature rise of 0.60°C or more, the test is considered positive. The second step is carried out if a single animal's temperature rise is at least 0.60°C or if the sum of the three increases exceeds 1.40°C. The second step results in a

positive classification if at least two of the five additionally tested rabbits show an individual temperature rise of 0.60°C. Also the USP algorithm is composed of two steps, which foresee three and five rabbits, respectively (Table 1). If all three individual temperature rises in the first step are below 0.50°C, the tested product passes; otherwise the second step with five additional animals has to be conducted. If then at least three of the eight rabbits in total show a temperature increase of at least 0.50°C or if all eight increases sum up to at least 3.30°C, the test is positive. In contrast to the JP and EP, the Pyrogen Test of the USP lacks an explicit criterion for failure in the first step. This prompted us to modify the application of the USP, denoted in the following USP-2, which can be assumed to be more realistic: it can at least be anticipated that, if the three rabbits in the first step show already a total increase in temperature of 3.30°C or more, which is the threshold in the second step of the algorithm, the test will be terminated and the substance tested will fail.

Pyrogen Test	Cumulative number of rabbits	Product passes / negative test result	Product fails / positive test result
EP	3	if summed response ≤ 1.15 ºC	if summed response > 2.65 ºC
	6	if summed response ≤ 2.80 ºC	if summed response > 4.30 ºC
	9	if summed response ≤ 4.45 ºC	if summed response > 5.95 ºC
	12	if summed response ≤ 6.60 ºC	if summed response > 6.60 ºC
JP	3	if all individual responses < 0.60 ºC and if summed response ≤ 1.40 ºC	if 2 or 3 individual responses ≥ 0.60 ºC
	5#	if 4 or 5 individual responses < 0.60 ºC	if 2 or more individual responses ≥ 0.60 ºC
USP*	3	if all individual responses < 0.50 ºC	- (USP-2: If summed response > 3.30 ºC)
	8	if summed response ≤ 3.30 ºC and if not more than 3 individual responses ≥ 0.50 ºC	if summed response > 3.30 ºC or if 4 or more individual responses ≥ 0.50 ºC

Table 1. Experimental designs and criteria for product assessment of the Pyrogen Tests of the EP, JP and USP

* Although the USP lacks a criterion for failure in the first step, it can be anticipated (USP-2) that if already the three rabbits in this step show a summed response larger than 3.30 ºC, which is a criterion for failure in the second step, the test will be terminated.
# Five additional animals are tested and the test result is determined only considering these.

From the three basic algorithms the rabbit fever threshold for failure, i.e. discriminating pyrogenic from non-pyrogenic samples for a rabbit, and their corresponding EU were deduced. The thresholds range from 0.41°C for the USP, to 0.60°C for the JP, while the EP considers 0.55°C as the crucial increase. With regard to the USP and EP, they were calculated as the mean rabbit temperature, which demands a pyrogenic classification in the respective final step. This approach resulted e.g. for the USP in dividing 3.30°C by 8, the number of rabbits tested. The JP threshold, which focuses on the individual temperature increase, was defined as 0.60°C. The corresponding

endotoxin concentration inducing these temperature rises in our model were 2.53 EU/kg for the USP, 5.22 EU/kg for the EP and 7.38 EU/kg for the JP (indicated by the dotted lines in Figure 1).

5.4 Results

Curves giving the probability for a pyrogenic classification for every endotoxin concentration were calculated for the three pharmacopoeial tests and the USP-2 with the developed tools (Figure 2 A). The two main criteria to assess and compare the different algorithms of the Pyrogen Tests were derived from these curves: On the one hand, the size of the range of equivocal results was considered. We chose the range of concentrations inducing a pyrogenic classification with a probability of at least 0.01 and at maximum of 0.99. The EP-range spanned 6.46 EU/kg from 2.30 to 8.76 and the JP-range spanned 9.48 EU/kg (from 0.83 to 10.31). The respective length for the USP and USP-2 was 7.82/kg (from 0.67 to 8.49), showing that the modification in the USP-2 did not affect this criterion. However, these ranges of equivocal classification depend on the number of steps and the number of rabbits per algorithm step, so that the EP, with a maximum of twelve animals tested, is least equivocal.

On the other hand, the range of equivocal classifications should provide a reasonable balance between false-negative classification, which impairs the safety of the patient, and false-positive classifications. To be able to assess these misclassifications, a clear-cut definition of an endotoxin concentration, representing the threshold between pyrogenic and non-pyrogenic, was needed. As the derived rabbit fever thresholds differ between the pharmacopoeias, the corresponding concentrations differed as well (see above), i.e. between 2.53 and 7.38 EU/kg. These concentrations and their corresponding probabilities for a pyrogenic classification were included in Figure 2 A. For the investigated data set, the different definitions of rabbit fever had a tremendous impact. When an endotoxin concentration of 7.38 EU/kg, which results in a mean temperature rise in rabbits of 0.60°C, was considered pyrogenic, all pharmacopoeial algorithms resulted with a probability of at least 95.0% in a pyrogenic classification.

But if the threshold was 2.53 EU/kg, i.e. 0.41°C, the probabilities of pyrogenic classification differed between 2.5% for the EP and 53.8% for the USP.

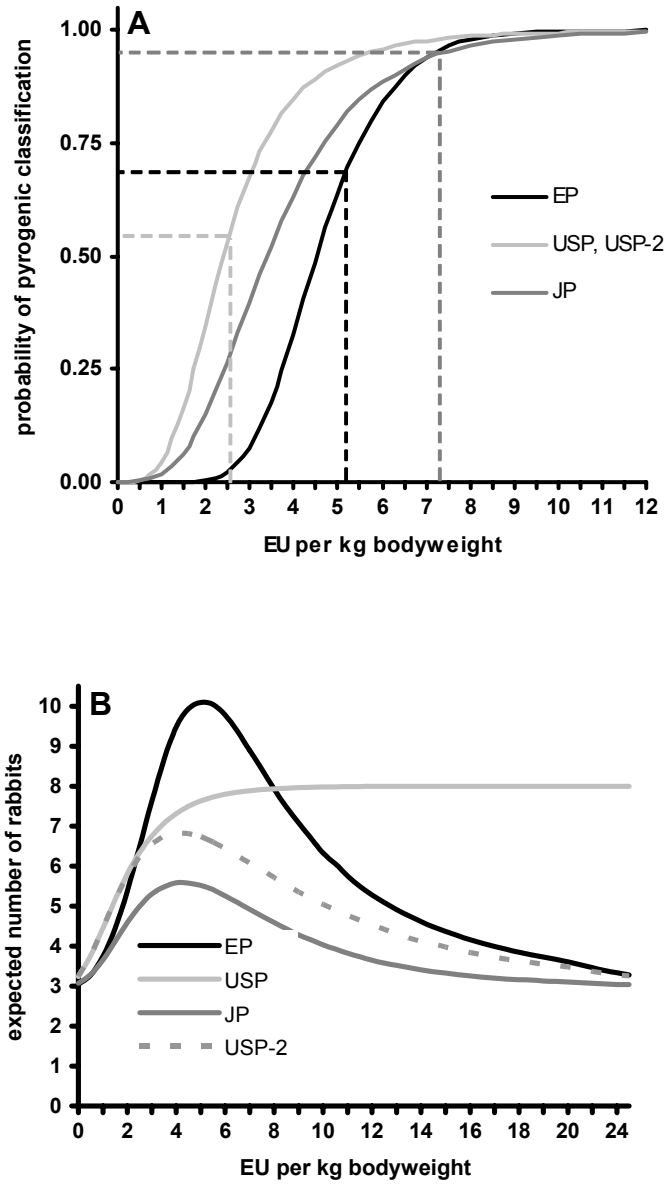

Figure 2. A: Probability for pyrogenic classification depending on endotoxin concentration according to EP, USP, USP-2 and JP based on the modelled data. For each pharmacopoeial test, the dotted line indicates the EU-concentration causing the rabbit temperature increase that is considered as the threshold discriminating pyrogenic from non-pyrogenic samples and the corresponding probabilities to classify this threshold concentration as pyrogenic.
B: Expected number of rabbits used depending on endotoxin concentration according to EP, USP, USP-2 and JP based on the modelled data.

Both criteria exposed severe differences between the pharmacopoeial algorithms in the way the crucial safety test for pyrogenicity is approached and in the way it is assessed. It is recommended to harmonise the Pyrogen Tests with regard to these properties and the definition of the rabbit fever threshold. In the process of harmonisation, which is for example encouraged by the EP (Council of Europe, 2001c), the presented model could serve as an objective tool to find a generally accepted approach.

In parallel to the classification curves, the curves of the numbers of rabbits expected to complete the algorithms for every endotoxin concentration were calculated (Figure 2 B). According to the algorithms, all curves started with about three rabbits for 0 EU/kg. The respective USP curve increased fast for low and approximated the maximum of eight rabbits for increasing endotoxin concentrations. Here, the modification resulting in the USP-2 approach had a strong impact as it reduced the number of animal for increasing EU-concentration. The EP-algorithm resulted in a curve with a high peak consumption of rabbits around the threshold concentration of 5.23 EU/kg, which slowly decreased for increasing concentrations. The JP-algorithms showed the lowest animal numbers over the whole range of EU-concentrations with a peak around 4 EU/kg. Especially the different levels and shapes of the curves in the range between 2 and 10 EU/kg indicated the need for harmonisation of the pharmacopoeial algorithms regarding animal numbers.

As the simulation process could be adjusted for the three pharmacopoeial algorithms, it also offered a tool for developing and optimising new experimental designs. The new approaches could either aim at an optimisation of the curves of classification (Figure 2 A) by increasing their slopes, i.e. minimising the area of equivocal test results, or at a decrease of the number of animals used. Assuming acceptable performances of the current algorithms with regard to misclassifications, the focus was set on the reduction of the number of animals used while maintaining a similar classification curve. Based on the step-wise approach of the experimental design, a varying number of steps with a reduction of animal numbers per step were considered. Among several new designs simulated with the model, an approach with three steps and two rabbits per step performed particularly well. Two versions, which

were based on the sum of rabbit temperature increases, are presented here (Table 2). Regarding the classification curves, the approach 2-2-2-A was designed as a compromise between the EP and the JP algorithm, while the approach 2-2-2-B was constructed to mimic the USP.

Approach	Cumulative number of rabbits	Product passes/negative test result if summed response	Product fails/positive test result if summed response
2-2-2-A	2	≤ 0.65 ºC	> 1.50 ºC
	4	≤ 1.75 ºC	> 2.35 ºC
	6	≤ 3.10 ºC	> 3.10 ºC
2-2-2-B	2	≤ 0.60 ºC	> 1.40 ºC
	4	≤ 1.60 ºC	> 2.20 ºC
	6	≤ 2.50 ºC	> 2.50 ºC

Table 2. Two versions of a new approach to pyrogen testing with criteria for product safety assessment

Their curves of classification are given together with the curves from Figure 2 A for illustration (Figure 3 A). The respective expected numbers of animals to be used are shown in Figure 3 B, where the corresponding curves for the EP, JP and USP were added. Additionally, a comparison of the current EP-algorithm with the 2-2-2-A and a comparison of the USP with the 2-2-2-B are presented as reduction of animal numbers in percent depending on the endotoxin concentration (Figure 3 C).

The endotoxin range of equivocal classification of the approach 2-2-2-A spanned 8.27 EU from 1.28 to 9.55 EU/kg, which was in-between the approaches of the JP and the EP. The respective range of the 2-2-2-B approach was 8.98 EU/kg (from 0.57 to 9.55 EU/kg), which was slightly larger than the USP-range of 7.82 EU/kg. Furthermore, the probabilities for positive classifications at the threshold concentration were comparable, which ranged for 2-2-2-A between 0.156 for 0.41°C and 0.948 for 0.60°C and for 2-2-2-B between 0.535 and 0.994. These competitive properties of the classification curve were achieved although the numbers of rabbits employed were

severely reduced. The curves of the expected rabbit numbers used were below the ones of the current pharmacopoeial algorithms over the entire concentration range (Figure 3 B). Although the reduction of animals was most extreme in the area of the peaks, attention should be paid to the offset of the curves. Reducing the number of animals in the first step from three to two animals, shifted the origin of the curves down to 2.20 (2-2-2-A) and 2.31 (2-2-2-B) animals, respectively. Both comparisons, i.e. the approach 2-2-2-A with the EP-algorithm and the 2-2-2-B approach with the USP-2, resulted in a reduction of 28% for a pyrogen-free sample, i.e. 0 EU/kg (Figure 3 C). As it can be assumed, that in practice the majority of tested batches are pyrogen-free (Twohy et al., 1984), this presented the most important aspect of the comparison. However, also for pyrogenic batches a decrease in the number of animals was achieved over the entire endotoxin range. For the comparison of the 2-2-2-A with the EP, the reduction was between 18% and 55%, approximating 33% for increasing concentrations. In contrast, the reduction of the comparison of the approach 2-2-2-B with USP approximated 75%, while a minimum of 22% was reached for 2 EU/kg. These properties of the new approach, i.e. its comparable safety and the reduction in animal numbers, hold true for all reasonable combinations of the three input parameters of the developed model, i.e. the regression intercept, slope and standard error.

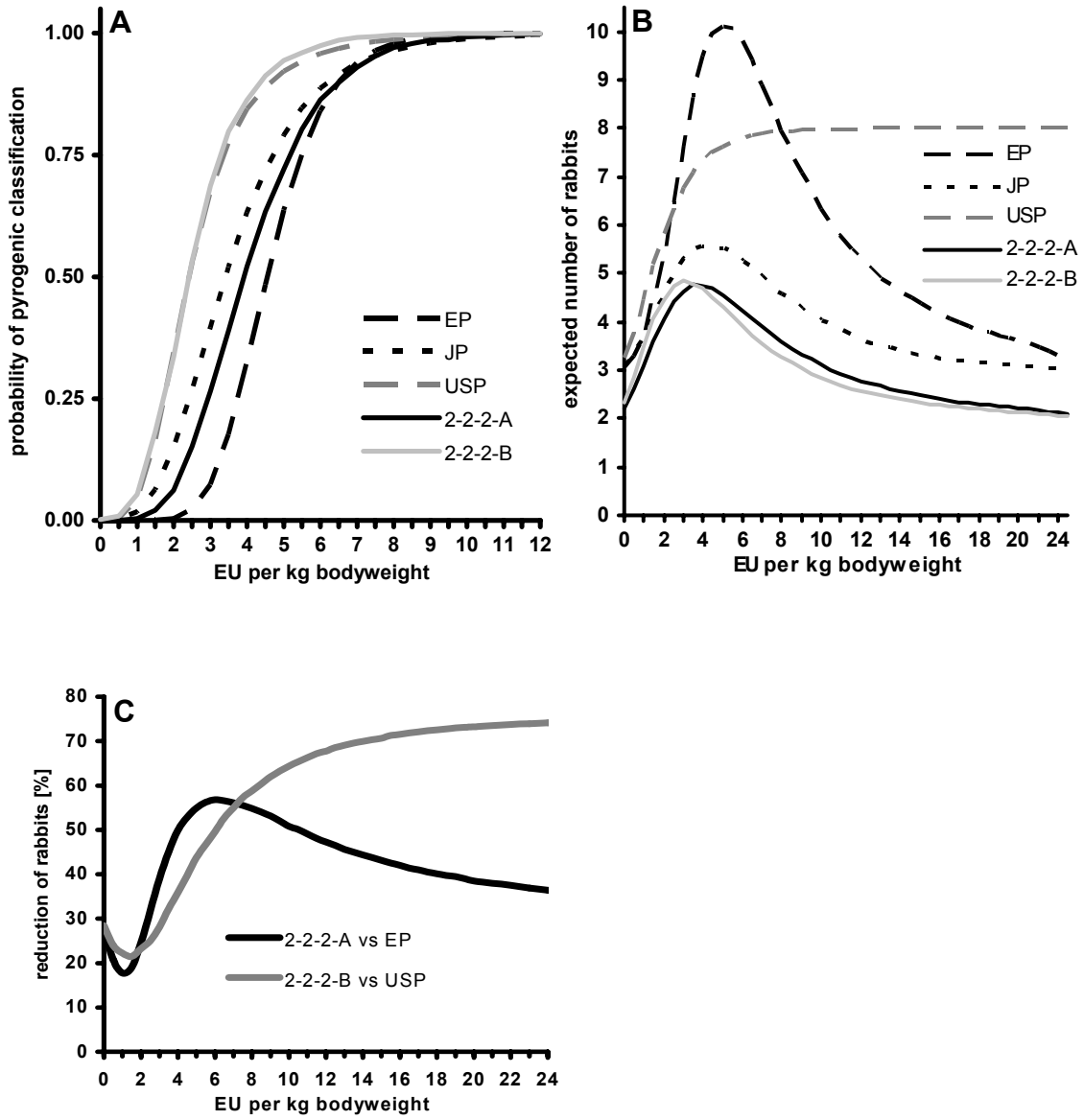

Figure 3. A: Probability of a pyrogenic classification depending on the endotoxin concentration according to the versions 2-2-2-A and 2-2-2-B of the suggested approach based on the modelled data. To simplify comparison, the EP-, JP- and USP-curves from Figure 2 A were added.
B: Expected number of rabbits required depending on endotoxin concentration for the two versions of the new 2-2-2-approach based on the modelled data. To simplify comparison, the EP-, JP- and USP-curves from Figure 2 B were added.
C: Relative comparison of the approach 2-2-2-A with the EP and of the approach 2-2-2-B with the USP regarding the reduction of rabbits used in %.

5.5 Discussion

Recently statistical modelling approaches to reduce animal consumptions in animal experiments according to Russell and Burch's 3R gave promising results in different fields of toxicity and safety testing. Especially in acute oral toxicity, this resulted in regulatory adoption by the OECD (OECD, 2001a-d) of three different approaches, comprising a fixed-dose (van den Heuvel et al., 1990) a stepwise (Schlede et al., 1995), and an up-and-down procedure (Bruce, 1987). Considering the safety rabbit pyrogen test, in contrast to a recent modelling approach (Tschumi, 2003), which was based on theoretical temperature rises and assumed variability, the crucial links to reference endotoxin data as well as a realistic estimation of the variation are provided here. With the developed rabbit fever model and its computational implementation to simulate current pharmacopoeial algorithms, differences between pharmacopoeias were highlighted and a need for harmonisation became evident. Additionally, the approach presented here supplies a broadly applicable tool for optimising in vivo pyrogen testing. Its flexibility allows simulating new experimental designs taking the level of safety and animal welfare into account. Our specifically developed designs clearly show that pyrogen testing can be optimised. As the number of rabbit tests worldwide has been estimated to be around 400.000 animals per year, it appears timely to reconsider the algorithms employed in order to refine and harmonise the test and reduce animal numbers by 30%.

5.6 Acknowledgement

The authors wish to thank Dr. S. Dalmora from the Department of Industrial Pharmacy at the Federal University of Santa Maria, Brazil, for the provision of the second data set, and Dr. S. Poole from the NIBSC, UK, for helpful suggestions.

6 Summarising discussion

Being introduced to major extent only in the 1990-ies, the concept of validation had a tremendous impact on *in vitro* toxicology. It was developed in response to politics, which demanded in order to promote animal welfare the implementation of the 3R principles of Russell and Burch (1959), i.e. the refinement, reduction and replacement of animal experiments. Validation was thus defined as the process of establishing the reliability and relevance of a method for a given purpose, for which subsequently principles and guidance were developed (Balls et al., 1990; Balls et al., 1995; Bruner et al, 1998; Worth and Balls, 2001; Worth and Balls, 2002; Hartung et al., 2004). To be able to overcome substantial doubts in the scientific and regulatory community regarding the general usefulness of *in vitro* approaches in toxicology, highest standards and scientific rigour were aimed for. Nowadays, the value of validation for toxicology in general and in particular for hazard assessment is recognised, e.g. the OECD currently develops a guidance document on validation for all types of test methods, in which especially the emphasis on relevance is considered the major advance.

A basic element for test relevance assessment is the prevalence of a toxic health effect: It makes a tremendous difference for the assessment of any kind of tests (*in vitro* as well as *in vivo*) whether 1, 10 or 50 out of 100 tested substances are toxic as optimal test performance varies with this prevalence (Grimes and Schulz, 2002). However, in establishing a toxicological test, e.g. by means of a validation study, the prevalence of the considered health effect was often disregarded. Only by describing or modelling a realistic setting, in which a test is applied or will be applied, allows drawing conclusions regarding its relevance for practical use, i.e. the given purpose.

Relevance of an *in vitro* test method incorporates two additional aspects. On the one hand, it is based on current scientific knowledge about potential mechanism of toxic effects and on the way and extent these are reflected by a given test. On the other hand, relevance is established by a comparison with a reference standard (Fentem et

al., 1998; Walter et al., 1999; Knottnerus and Muris, 2003), i.e. usually the current *in vivo* test in use. However, resulting views, which consider these two aspects, might nevertheless differ regarding the relevance conclusions for a test. Therefore, a harmonised approach for relevance assessment is needed in order to achieve a scientifically sound, unambiguous and transparent relevance statement. Although difficult for the mechanistic basis of a test, which steadily develops, this is feasible for the comparison with the reference standard. With regard to the *in vitro* side of this comparison, the validation process constitutes a highly objective tool providing the information needed, which, however, can still be optimised. In contrast, information of similar quality on the reference test is usually lacking. But also the reference test is neither optimally reproducible nor perfectly true having thus an enormous impact on the relevance assessment of the *in vitro* test (Bruner et al., 2002a; Bruner et al., 2002b). So far, if considered at all, this aspect was in *in vitro* validation studies at most only rudimentarily addressed (Gettings et al., 1996). Therefore, new and objective approaches evaluating first of all the reference test would substantially support a complete comparison of both tests.

Collecting, analysing and summarising objective information for these three aspects, i.e. the prevalence, the *in vitro* test and the reference standard, would finally allow a complete and evidence-based test assessment. In this process, statistical approaches and methodology from evidence-based medicine offer useful tools.

6.1 Evidence-based approaches applied to in vitro toxicology

6.1.1 The reference standard

In validation exercises, be it clinical or toxicological, the relevance of a test is assessed by comparison against a pre-defined reference standard. Ideally, this is a perfect and thus true reference standard, which, however, rarely does exist (Garrett et al., 2002). In the scientific literature, in this context often the term ‘gold standard’ is used, while its definition is not clear: It either describes the best currently available test or test combination (Barthel and Everett, 1990; Hawkins et al., 2001), or it refers to a

perfectly true, i.e. a theoretical test (Garrett et al., 2002; Pouillot et al., 2002). Noteworthy, it has to be realised, that even tests in clinical diagnostics for crucial diseases like HIV are not always true, although the tests are close to perfect and usually repeated for positive outcomes (Gigerenzer et al., 1998).

In validation of *in vitro* alternatives as well as in clinical diagnostic test assessment, the reference standard usually is the current test of choice to test for a given health effect (Fentem et al., 1998; Walter et al., 1999; Fentem et al. 2001; Knottnerus and Muris, 2003). As toxicity tests on humans, which would be closest to a perfect test, are not or only very limited available, the current test of choice for validating alternative methods is most often an *in vivo* test. Nevertheless, if no reference test is available, the reference could also be another *in vitro* test or be established by an expert panel, as it is an option in clinical diagnostics (Weller and Mann, 1997; Knottnerus and Muris, 2003). In any case, when finally processing study results, again methodology from clinical diagnostics allowing to address and implement the imperfectness of the reference standard might provide guidance and could, if feasible, be adapted to toxicological validation (Walter et al., 1999; Enoe et al., 2000).

In contrast to the aspect of relevance of the reference standard, which will, due to limitations regarding human information, usually not be properly accessible, there is an optimal, though provocative solution to assess the reference standard's reliability. Employing this standard, i.e. usually an *in vivo* test, in parallel in a validation study of *in vitro* tests and applying the same quality criteria, e.g. blinding, would result in data of equal quality for both tests. When appropriate designs minimising animal use and suffering are applied, the use of animals might be justifiable by the substantially increased chance to (partly) replace the animal experiment. Although methodologies from evidence-based medicine and/or clinical trials might be helpful (European Medicines Agency, 1998; Knottnerus and Muris, 2003), study designs balancing the comparability of the two tests versus animal consumption and suffering would need to be developed and discussed. Furthermore, it has to be considered that such a proposal will be difficult to set into practice as it will most probably meet strong political resistance, where even legal constraints and hurdles might have to be overcome. The most prominent example, where the imperfectness of the *in vivo* reference standard is

strongly criticised for decades, is the field of eye irritation (Weil and Scala, 1971; Freeberg et al., 1986; Gettings et al., 1996). In the 1990-ies, several international validation studies with immense costs were conducted without major successes, e.g. by Balls et al. (1995), by Gettings et al. (1996) or by Chamberlain et al. (1997). It is evident, that in these studies the imperfectness of the reference standard, an *in vivo* experiment on rabbits (Draize et al., 1944), was not the only reason for failure. Nevertheless, proper parallel incorporation of the animal experiment into one of these studies might at least have substantiated the claim of imperfectness or may even have enabled more differentiated and realistic relevance assessment.

6.1.2 Prevalence

If an *in vitro* tests aims to replace an *in vivo* test, this could be done via performance parameter of predictive capacity, such as sensitivity and specificity, i.e. the probability for correct positive, respectively negative results. Unfortunately, most *in vivo* test were never rigorously validated so that in many cases their outcomes are simply considered as true. As the assumption is wrong, this results in an overestimation of their predictive capacity. However, combining for the individual tests parameter for predictive capacity with corresponding information on prevalence to calculate predictive values (NPV and PPV) allows a more realistic comparison and assessment (Grimes and Schulz, 2002). Although the predictive capacity in terms of sensitivity and specificity of a test might be sub-optimal, the predictive values might lead to different conclusions regarding a test's reliability and relevance. For example, applied to a population with 1% prevalence, a test with 90% sensitivity and specificity and a test with 95% specificity and only 60% sensitivity would have comparable predictive values, i.e. is a negative predictive value of almost 100% and a positive predictive value around 10%. Test assessment by predictive capacity alone, would have favoured the first test and unjustifiably rejected the second test. In cases when a test does not seem to be able to replace an *in vivo* test, predictive values might trigger other uses of the test, e.g. testing strategies. In the same manner as for a single test, the development of such a strategy should be based on prevalence information.

This aspect opens up a so far disregarded field for *in vitro* test and their use, e.g. for regulatory purposes. Most often the aim of *in vitro* validation studies was to replace an animal experiment, where successes (Spielmann and Liebsch, 2002), but also failures, especially in the field of eye irritation (Balls et al., 1995; Brantom et al., 1997; Chamberlain et al., 1997), can be listed. Describing by means of prevalence the toxic potential of substances to be tested in sufficient details allows to identify and characterise important aspects and to design *in vitro* tests for special purposes addressing these. For example, if a test lacks discriminative power in the middle range of toxic potential, it might be optimised to identify the non-toxic, the highly toxic substances or both by protocol modifications. Following these initial ideas, a so far lacking framework for the validation of test strategies incorporating prevalence information could be developed triggering specific *in vitro* development by identifying realistic needs. Ultimately, this could lead to an optimal way to define test strategies by risk-benefit analysis integrating safety aspects, animal welfare and economical consideration using all available sources of information, i.e. *in silico*, *in vitro* and *in vivo* tests.

6.2 The reference standard for skin irritation

The two concepts of prevalence and the imperfectness of the reference standard were applied in this thesis to the toxic effect of skin irritation in order to support an ongoing validation of *in vitro* methods for skin irritation (Zuang et al., 2002; Botham, 2004). In the presented work, a critical review of the *in vivo* rabbit skin irritation test according to current regulations, which was defined as the reference standard, was conducted. For this purpose, historical *in vivo* data from two databases were employed. The prevalence of skin irritation potential of industrial chemicals was described according to regulatory classification schemes, where less than 10% of the substances required a hazard label. As skin irritation is a continuous effect with regard to toxic potency and extent of effect, also a more detailed distribution of skin irritation was presented. The resulting frequencies did not only allow comparison of regulatory classification schemes, but especially the detailed distribution revealed a bias introduced by such a

scheme. However, the extent to which the results can be transferred from the applicability domain 'industrial chemicals notified in Europe after 1981' to other applicability domains needs further discussion.

This leads inevitably to another problem associated with applicability domains. Once a domain is defined, a systematic search for useful information to assess its prevalence should be conducted. Basically, two kinds of information sources will be found: databases and scientific literature, whose representativity for the applicability domain has to be discussed and assessed in both cases. In comparison to inquiring databases, searching the scientific literature is a much more laborious task, which has to be well organised in order to assure its completeness and quality. It should be conducted independently to avoid any form of bias, again a topic where the field of evidence based-medicine might give guidance (Knottnerus et al., 2002; Horvath and Pewsner, 2004). In cases when no information on prevalence or no consensus regarding its representativity can be found, several prevalence values covering the respective interval from worst case scenarios, i.e. high prevalence, to best case scenarios, i.e. low prevalence, can be assumed (Bruner et al., 2002a; Bruner et al., 2002b). In addition, this study might trigger similar analysis for other toxicological endpoints. Regarding the New Chemicals Database of the European Chemicals Bureau, one of the databases used here, this is foreseen. However, the more evidence is made available, e.g. from other databases or from industry (Purchase, 1990; Stitzel, 2002), the more complete and differentiated the prevalence information will be.

Furthermore, by employing re-sampling techniques it was shown how the reproducibility of the reference standard affects its predictive capacity. The variation among rabbit responses within an experiment, which usually includes three animals, reduced the test performance in predicting regulatory classes. Here, the detailed prevalence distribution allowed taking the distance to the classification threshold(s) into account: for example, the less irritating a chemical is, the bigger its distance from the classification threshold and thus the higher is the probability that it is classified as a non-irritant. Additional sources of variation such as different operators, time or different laboratories, which could not be assessed with the available data, will further

reduce the predictive capacity of reference standard. As the relevance of the rabbit experiment for human skin irritation is difficult to assess, at least the way how its experimental variabilities impair the predictive capacity should be accounted for when validating and assessing respective *in vitro* tests.

6.3 Validation of alternative pyrogen tests

Starting some 20 years ago, several *in vitro* tests, either based on cell lines or human blood, were developed for the detection of pyrogens making use of cytokine release, an early indicator of the human fever reaction (Dinarello, 1999). Having reached a mature state, six of these assays were put forward to formal validation (Hartung et al., 2001; Poole et al., 2003). This validation study was designed to maximise the evidence to be gained by focusing on the chosen reference test, i.e. the rabbit pyrogen test, and by harmonising the *in vitro* tests' set-ups, e.g. the number of replicates or the definition and use of negative and positive controls. The harmonisation allowed applying almost identical analyses procedures to all tests generating comparable evidence. Especially the harmonised prediction model, which converted test results into statements of pyrogenic contamination, represented a major benefit. Furthermore, thorough construction of the prediction model as a statistical test, which was designed to proof and thus foster safety, reduced unfavourable impact of test variability (Hauschke and Hothorn, 1998; Hauschke et al., 2003). Making use of an evaluation of the reference standard, a highly demanding study design was developed: The pyrogen threshold concentration, which the novel *in vitro* test would have to detect, was defined as the concentration, which 50% of *in vivo* tests with a very sensitive rabbit strain would classify as pyrogenic. Additionally, further spikes containing concentrations closely above and below this threshold were chosen challenging the detection limits of the tests and the discriminative power of the prediction model.

This approach also allowed modelling the rabbit test performance with regard to the chosen study design. Thus the *in vitro* tests and the reference standard could be directly compared showing the superior predictive capacities of four new tests.

Calculating the tests' performance for other threshold concentrations generated further evidence favouring the *in vitro* tests by underlining their robustness.

Tasks ahead concerning biological and validation aspects for these alternative tests are further test improvement, i.e. mainly the validation of methods based on cryopreserved blood (Schindler et al., 2004), and evaluation of the tests' performances for a broad range of pyrogens and for specific applicability domains, e.g. vaccines (Ochia et al., 2002). Regarding the methodology, the applied tools and their applicability needs to be further evaluated, implemented and communicated. Especially the receiver operation curve-techniques (McNeil et al., 1975; van der Schouw et al., 1995; Edler and Ittrich, 2003), a standard tool in clinical diagnostics, which was here only rudimentarily introduced, needs to be further explored. It could be most supportive for the assessment of *in vitro* toxicological test, not at least as it only requires a general prediction model which could be optimized anytime during a study.

6.4 Optimisation and harmonisation of the rabbit pyrogen test

In addition to its crucial importance for the relevance assessment of new *in vitro* test for pyrogen testing, the thorough analysis of the rabbit pyrogen test opened up an opportunity for exploring reduction of animals used in this test. This is a common approach to animal reduction in animal experiments usually achieved by modelling of the performance of experimental settings using fewer animals with (bio-)statistical tools (Hutchinson et al., 2003; Stallard et al., 2004). Regarding pyrogen testing, first a discrepancy between the different international pharmacopoeias in the way pyrogenic contamination is assessed and in animal consumption was demonstrated highlighting the need for international harmonisation. Making in a second step use of the developed statistical model, experimental designs consuming lesser animals were constructed while the safety properties of the current tests were maintained. Although only scarce (Twohy et al., 1984), prevalence information was incorporated in the construction of these designs: As pyrogenic samples are very rare (<< 5%), most animals, i.e. about 30%, could be saved by focusing on the baseline animal response, i.e. to pyrogen-free samples.

As such biostatistical approaches to animal reduction have proven to work (OECD, 2001b-e), their application to areas where promising *in vitro* approaches are lacking could result in a substantially lower animal consumptions. Especially the areas of biologicals (Weisser and Hechler, 1997), chronic toxicity (Prieto, 2002) and carcinogenicity (Spielmann, 2002) should be explored.

Biometrical tools form an essential part of evidence-based medicine, which aims for the objective assessment of opportunities and limitations of diagnostic as well as therapeutic measures. Similarly, the application of such approaches to toxicology promises to develop an evidence-based toxicology. The foundation has been laid with the principles of validation and good practice quality assurance systems. The thesis presented aimed to expand this concept to a critical analysis of the reference standard based on existing data and modelling. Immediate conclusions for a more rational use of the *in vivo* test itself could be drawn.

The important concept of prevalence was adopted and consequences for the assessment of methods demonstrated. Finally, combining this with the biometrical guidance of a validation study, it was for the first time possible to demonstrate that *in vitro* approaches can actually outperform the animal experiment considered as the reference standard.

7 Summary

Although basic principles promoting laboratory animal welfare were developed almost 50 years ago, only nowadays with better understanding of toxicological mechanisms and new biological technologies tools are at hand to efficiently implement these principles into toxicology.

For most of the important toxicological endpoints *in vitro* approaches were or are developed, and for some health effects, e.g. skin corrosion or phototoxicity, animal tests could already be replaced. Once a test is developed and is promising, the final step to assess its reliability and relevance is usually a multi-laboratory validation study. However, validation, having proven to be an appropriate tool, still requires optimization and further development.

In the first part of this thesis, the field of evidence-based medicine in clinical diagnostics was identified as a topic closely related to toxicology and validation, whose methodologies are worthwhile to explore and, if indicated to be adapted. Basically two concepts, the prevalence and the imperfectness of the reference standard, were transferred to the field of toxicology. Considering typical toxicological situations, several ways to implement these concepts, especially assessment of toxicological tests, were explored.

- Prevalence, a key determinant for the practical value of a diagnostic measure for a disease in clinical diagnostics, was considered in the context of a toxicological health effect. Using it in the simple dichotomous case, the added value of the Bayesian parameters negative and positive predictive value, compared to sensitivity and specificity for test assessment was demonstrated.
- The impact of the imperfectness of the reference standard, i.e. usually an *in vivo* experiment, on the prevalence of a toxicological effect was modelled. For rare toxicities, this meant that these are even rarer than we believe. The thus wrong assumptions regarding toxic potential of substances have to be considered in *in vitro* test assessment.

- Modelling several parameter combinations of accuracies and prevalence for a dichotomous test outcome challenged the common use of confirmatory tests. The results highlighted how the benefit of a confirmatory test depends on these parameters.

Prevalence according to the reference standard was evaluated in detail for the skin irritation potential of chemical substances. Two databases containing *in vivo* data of the rabbit experiment demanded by regulators were analysed focusing on different prevalence related aspects.

- The prevalence according to regulatory classification schemes was described in detail, where less than 10% of more than 3000 substances showed irritation potential requiring a hazard label.
- A simple integrated measure of skin irritation with minimal loss of information, which results from the rabbit experiment, was developed. This measure allowed describing skin irritation potential according to the animal experiment on an almost continuous scale.
- Analysis of the within-experiment reproducibility of the rabbit experiment revealed that this is only a minor source of variability.
- The prevalence was combined with the results of the modelling of the within-test variability in order to evaluate the utility of the approach and compare two classification schemes.

A large-scale international validation study for novel pyrogenicity tests focusing on the chosen reference standard was guided with regard to design and biometry and was thoroughly evaluated. In this study six *in vitro* test based on the human fever reaction were validated against the pharmacopoeial rabbit pyrogen test.

- Analysing historical, high quality data from a very sensitive rabbit strain showed that the rabbit pyrogen test would detect 0.5 Endotoxin units/ml in 50% of all cases. This concentration was defined as the threshold level novel pyrogen test would have to meet.

- Making use of this threshold a very challenging study was designed to evaluate the performance of the novel test for low, but crucial contamination levels.
- In the prevalidation phase all six novel tests showed good reliability and good predictive capacities for twelve coded samples from three drugs.
- In the validation phase (50 coded samples from 10 drugs) four tests performed reliably with predictive capacity outperforming the rabbit test, for which corresponding results were modelled for the chosen study design.
- The results from modifications of the prediction models confirmed the robustness of these four tests.

The evaluation of the rabbit pyrogen test initiated by the validation study triggered an in-depth analysis. Based on a probabilistic model of the rabbit fever reaction, this analysis focused on international harmonization of pharmacopoeial approaches and on modelling of new experimental designs of the rabbit test.

- Comparing three pharmacopoeial tests differing in their design and data interpretation strongly demonstrated need for harmonization. As variable rabbit temperature increases are considered as indicators for pyrogenicity, the probability curves for pyrogenic classification differed substantially between Pharmacopoeias.
- Modelling the corresponding animal consumption revealed substantial differences between the pharmacopoeial tests caused by different test algorithms.
- Using the rabbit model, two new test designs were developed. While maintaining safety levels of current pharmacopoeial tests, animal consumption could be reduced by 30% taking prevalence information on pyrogenicity into account.

In summary, with several studies implementing different principles of evidence-based medicine the usefulness of these principles for *in vitro* toxicology was demonstrated. Prevalence is a most important information for a comprehensive and complete test

assessment. Similar, detailed evaluation of reference tests in validation studies is mandatory to allow an optimal relevance assessment for both reference and *in vitro* test. Introducing and employing these approaches systematically into the field of *in vitro* toxicology would constitute a first step towards an evidence-based toxicology.

8 Zusammenfassung

Obwohl die grundlegenden Prinzipien des Tierschutzes in der Wissenschaft bereits vor fast 50 Jahren definiert wurden, erlauben erst das heutige Verständnis toxikologischer Mechanismen und moderne Bio-Technologien die effiziente Implementierung dieser Prinzipien in die Toxikologie.

In-vitro-Testmethoden wurden und werden immer noch für fast alle wichtigen toxikologischen Endpunkte entwickelt. Für einige Endpunkte, zum Beispiel Hautkorosivität oder Phototoxizität, konnten bereits die dort üblicherweise verwendeten Tierversuche ersetzt werden. Dabei steht am Ende der Entwicklung eines vielversprechenden *in-vitro*-Tests in der Regel eine multi-zentrische Validierungsstudie, um die Verlässlichkeit und Relevanz des Tests beurteilen zu können. Da sich diese Vorgehensweise bewährt hat, ist es nun an der Zeit, diesen Prozess zu optimisieren und weiterzuentwickeln.

Im ersten Abschnitt dieser Arbeit wurde die evidenz-basierte Medizin in der klinischen Diagnose als ein zur Validierung in der *in vitro* Toxikologie eng verwandtes Gebiet identifiziert. Grundprinzipien der Methodik wurden in die Toxikologie übertragen, wobei der Aspekt der Prävalenz und der Einfluss der Fehlerhaftigkeit des Referenzstandards im Vordergrund standen. Anhand typischer toxikologischer Situationen wurden mehrere Optionen untersucht, wie diese Konzepte in die Toxikologie und besonders deren Testbeurteilung eingeführt werden könnten.

- Prävalenz, ein Schlüsselparameter für den praktischen Nutzen eines diagnostischen Tests, wurde auf die Häufigkeit des Auftretens eines toxikologische Endpunkts bezogen. Für den einfachen dichotomen Fall wurden die Vorteile der Bayes'schen Parameter negativer und positiver prädiktiver Wert im Vegleich zu Sensitivität und Spezifität dargestellt.
- Es wurde der Einfluss der Fehlerhaftigkeit des Referenzstandards, in der Regel ein *in-vivo*-Test, auf die Prävalenz eines toxikologischen Endpunkts modelliert. Für seltene Toxizitäten bedeutete dies, dass diese sogar seltener sind als angenommen. Die damit verbundenen falschen Annahmen bezüglich der

Toxizität bestimmter Substanzen muss bei der Beurteilung von *in-vitro*-Tests berücksichtigt werden.

- Der gebräuchliche Einsatz von Bestätigungstests, bzw. Konfirmationstests, wurde für verschiedene Kombinationen von Prävalenzen und Testgenauigkeiten im dichotomen Fall modelliert. Die Ergebnisse zeigten, dass der Nutzen eines Bestätigungstests stark von diesen Parametern abhängt.

Die Prävalenz von hautreizenden Eigenschaften chemischer Substanzen, beurteilt nach dem *in-vivo*-Referenzstandard, wurde untersucht. Zwei Datenbanken, die Daten des von Regulatoren geforderten Kaninchenexperiments enthalten, wurden in Hinsicht auf verschiedene Prävalenzaspekte analysiert.

- Die Prävalenz in Bezug auf regulatorische Klassifizierungssysteme wurde im Detail beschrieben, wobei weniger als 10% von 3000 Substanzen ein zu kennzeichnendes Irritationspotential zeigten.
- Es wurde ein einfaches Maß entwickelt, das die im Tierversuch generierten Daten mit nur sehr geringem Verlust von Information kondensiert.
- Die Analyse der Reproduzierbarkeit innerhalb eines Tierexperiments zeigte, dass diese nur ein geringe Variabilität mit sich bringt.
- Die Prävalenz wurde zusammen mit den Resultaten der experimentellen Variabilität kombiniert, um den generellen Nutzen dieser Vorgehensweise zu verdeutlichen und um schließlich zwei Klassifizierungssysteme zu vergleichen.

Eine große, internationale Validierungsstudie von neuartigen Pyrogenitätstests, die stark auf den gewählten Referenzstandard fokusiert war, wurde biometrisch betreut und detailliert analysiert. In dieser Studie wurden sechs *in-vitro*-Tests, die auf der menschlichen Fieberreaktion basieren, mit dem regulatorisch vorgeschriebenen Kaninchenpyrogentest verglichen.

- Eine Analyse historischer, hochqualitativer Daten von einem sehr sensitiven Kaninchenstamm zeigte, dass diese Tiere 0.5 Endotoxin Einheiten/ml in 50%

aller Fälle entdecken würden. Diese Konzentration wurde als der Wert definiert, den die neuen *in vitro* Test verlässlich detektieren müssten.

- Um diesen Wert wurde ein anspruchsvolles Studiendesign konstruiert, um die Eigenschaften der neuen Tests für diese niedrigen, aber kritischen pyrogenen Konzentrationen beurteilen zu können.
- In der Prävalidierungsphase zeigten alle sechs validierten Tests eine gute Verlässlichkeit sowie gute Genauigkeit für zwölf Proben von drei pharmazeutischen Produkten.
- In der Validierungsphase (50 Proben von zehn Produkten) zeigten vier Tests verlässliche Ergebnisse, die besser als die entsprechenden, modellierten Kaninchentestergebnisse waren.
- Die guten Resultate, die mit modifizierten Vorhersagemodellen erzielt wurden, unterstrichen die Robustheit dieser vier Tests.

Die Evaluierung des Kaninchenpyrogentest, die durch die Validierungsstudie initiiert wurde, zog eine ausführliche Analyse nach sich. Auf dem entwickelten Kaninchenfiebermodell basierend, konzentrierte sich diese Untersuchung auf die internationale Harmonisierung von regulatorischen Pyrogentests und auf die Modellierung von neuen Versuchdesigns für diesen Kaninchentests.

- Der Vergleich von drei regulatorischen Tests, die sich in ihrem Design und ihrer Dateninterpretation unterscheiden, zeigte dringenden Harmonisierungsbedarf auf. Da unterschiedliche Kaninchentemperaturanstiege als Indikator für Pyrogenität dienen, unterschieden sich ebenfalls die resultierenden Wahrscheinlichkeitskurven pyrogener Klassifizierung.
- Auch die Modellierung der entsprechenden, zu verwendenden Tierzahlen machte Unterschiede zwischen den regulatorischen Tests deutlich, welche auf die verschiedenen Testalgorithmen zurückzuführen waren.
- Mit Hilfe des Kaninchenfiebermodels wurden zwei neue Versuchsdesigns entwickelt. Bei gleichzeitiger Beibehaltung des momentanen Sicherheitsniveaus, reduzierten diese die benötigten Tierzahlen um 30%,

wobei Prävalenzinformationen über die Häufigkeit pyrogener Verunreinigungen berücksichtigt wurden.

Anhand meherer Studien, die stark auf verschiedenen Prinzipien der evidenz-basierten Medizin fokusiert waren, konnte der Nutzen dieser Prinzipien für die *in- vitro*-Toxikologie gezeigt werden. Prävalenz ist eine notwendige und wichtige Information, die erst ein umfassende Beurteilung eines Tests ermöglicht. In ähnlicher Weise erlaubt erst eine detaillierte Evaluierung des Referenztests in Validierungsstudien eine optimierte Beurteilung der Relevanz des Referenztests als auch des *in-vitro*-Tests. Die systematische Einführung und Verwendung dieser Betrachtungsweisen würde ein erster Schritt in Richtung einer evidenz-basierten *in-vitro*-Toxikologie darstellen, aus der sich für die Toxikologie allgemeingültige evidenz-basierte Prinzipien entwickeln könnten.

9 References

Allanou, R., Hansen, B.G., and van der Bilt, Y. (1999). Public availability of data on EU High Production Volume Chemicals. EUR 18996 EN (http://ecb.jrc.it).

Ames, B.N., and Gold, L.S. (1990). Chemical carcinogenesis: Too many rodent carcinogens. *Proc Natl Acad Sci USA* **87,** 7772-7776.

Bagley, D.M., Gardner, J.R., Holland, G., Lewis, R.W., Regnier, J.F., Stringer, D.A., and Walker, A.P. (1996). Skin Irritation: Reference Chemicals Data Bank. *Toxicol In Vitro* **10**, 1-6.

Balls, M., Blaauboer, B.J., Brusick, D., Frazier, J., Lamb, D., Pemberton, M., Reinhardt, C., Roberfroid, M., Rosenkranz, H., Schmid, B., Spielmann, H., Stammati, A.L., and Walum, E. (1990). Report and recommendations of the CAAT/ERGATT workshop on validation of toxicity test procedures. *Alt Lab Animals* **18,** 303-337.

Balls, M., Botham, P.A., Bruner, L.H., and Spielmann, H. (1995). The EC/HO international validation study on alternatives to the Draize eye irritation test. *Toxicol in Vitro* **9,** 871-929.

Balls, M., Blaauboer, B.J., Fentem, J.H., Bruner, L., Combes, R.D., Ekwall, B., Fielder, R.J., Guillouzo, A., Lewis, R.W., Lovell, D.P., Reinhardt, C.A., Repetto, G., Sladowski, D., Spielmann, H., and Zucco, F. (1995). Practical aspects of the validation of toxicity test procedures. The report and recommendations of ECVAM workshop 5. *Al. Lab Animals* **23,** 129-147.

Barthel, J.S., and Everett, E.D. (1990). Diagnosis of Campylobacter pylori infections: the “gold standard” and the alternatives. *Rev Infect Dis* **12 Suppl 1,** S107-S114.

Beutler, B., and Rietschel, E.T. (2003). Innate immune sensing and its roots: the story of endotoxin. *Nat Rev Immunol* **3,** 169-176.

Botham, P.A. (2004). The validation of in vitro methods for skin irritation. *Toxicol Lett* **149,** 387-90.

Boyko, E.J., Alderman, B.W., and Baron, A.E. (1988). Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. *J Gen Intern Med* **3,** 476-481.

Brantom, P.G., Bruner, L.H., Chamberlain, M., De Silva, O., Dupuis, J., Earl, L.K., Lovell, D.P., Pape, W.J.W., Uttley, M. et al. (1997). A summary report of the COLIPA international validation study on alternatives to the Draize rabbit ete irritation test. *Toxicol in Vitro* **11,** 141-179.

Brenner, H., and Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med* **16,** 981-991.

Bruce, R.D. (1987). A confirmatory study of the up-and-down method for acute oral toxicity testing. *Fundam Appl Toxicol* **8,** 97-100.

Bruner, L.H. (1996). No prediction model, no validation study. *Alt Lab Animals* **24,** 139-142.

Bruner, L.H., Carr, G.J., Curren, R.D., and Chamberlain, M. (1998). Validation of alternative methods for toxicity testing. *Environ Health Perspect* **106 Suppl 2,** 477-484.

Bruner, L.H., Carr, G.J., Harbell, J.W., and Curren, R.D. (2002a). An investigation of new toxicity test method performance in validation studies: 2. comparison of three measures of toxicity test performance. *Hum Exp Toxicol* **21,** 313-323.

Bruner, L.H., Carr, G.J., Harbell, J.W., and Curren, R.D. (2002b). An investigation of new toxicity test method performance in validation studies: 3. sensitivity and specificity are not independent of prevalence or distribution of toxicity. *Hum Exp Toxicol* **21,** 325-334.

Buck, A.A., and Gart, J.J. (1966). Comparison of a screening test and a reference test in epidemiologic studies. I. Indices of agreement and their relation to prevalence. *Am J Epidemiol* **83,** 586-592.

Calvin, G. (1992). New approaches to the assessment of eye and skin irritation. *Toxicol Lett* **64/65**, 157-164.

Campbell, R.L., and Bruce, R.D. (1981). Comparative Toxicology. I. Direct comparison of rabbit and human primary skin irritation responses to isopropylmyristate. *Toxicol Appl Pharmacol* **59,** 555-563.

Chamberlain, M., Gad, S.C., Gautheron, P., and Prinsen, M.K. (1997). IRAG working group 1. Organotypic models for the assessment/prediction of ocular irritation. *Food Chem Toxicol* **35,** 23-37.

Clopper, C.J., and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26,** 404-413.

Council of Europe (2001a). Biological Tests, 2.6.14. Bacterial endotoxins. In: *European Pharmacopoeia*. Council of Europe, Strasbourg, p. 140-147

Council of Europe (2001b). Biological Tests, 2.6.8. Pyrogens. In: *European Pharmacopoeia*. Council of Europe, Strasbourg, p. 131-132.

Council of Europe (2001c). General Texts, 5.8. Pharmacopoeial Harmonisation. In: *European Pharmacopoeia*. Council of Europe, Strasbourg, p. 503-503

Dabbah, R., Ferry. E. Jr, Gunther, D.A., Hahn, R., Mazur, P., Neely, M., Nicholas, P., Pierce, J.S., Slade, J., Watson, S., Weary, M., and Sanford, R.L. (1980). Pyrogenicity of E. coli 055:b5 endotoxin by the USP rabbit test--a HIMA collaborative study. *J Parenter Drug Assoc* **34,** 212-216.

Diener, W., Kayser D., and Schlede E. (1998). The dermal acute toxic class method: test procedures and biometric evaluations. *Arch Toxicol* **72,** 751-762.

Dinarello, C.A. (1999). Cytokines as endogenous pyrogens. *J Infect Dis* **179 Suppl 2,** S294-304.

Dinarello, C.A., O'Connor J.V., LoPreste G., and Swift R.L. (1984). Human leukocytic pyrogen test for detection of pyrogenic material in growth hormone produced by recombinant Escherichia coli. *J Clin Microbiol* **20,** 323-329.

Draize, J.H., Woodard, G., and Clavery, H.O. (1944). Methods for the study of irritation and toxicity of substances applied topically to the skin and mucous membranes. *J Pharmacol Exp Ther* **82,** 377-390.

Duff, G.W., and Atkins E. (1982). The detection of endotoxin by in vitro production of endogenous pyrogen: comparison with limulus amebocyte lysate gelation. *J Immunol Methods* **52,** 323-331.

EC (1992). Council Directive 92/32/EEC of April 1992 amending for the seventh time Directive 67/548/EEC on the approximation of the laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances. *Official Journal of the European Communities* **L154,** 1-29.

EC (2001). Annex VI of the Directive 67/548/EEC. General Classification and Labelling Requirements for dangerous substances and preparations. *Official Journal of the European Communities* **L225,** 263-314.

EC (2004). Commission Directive 2004/73/EC of 29 April 2004 adapting to technical progress for the twenty-ninth time Council Directive 67/548/EEC on the approximation of the laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances. *Official Journal of the European Union* **L152,** 1-311.

ECETOC (1995). Skin Irritation and Corrosion: Reference Chemicals Data Bank. Technical Report No. 66. European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels.

Edler, L., and Ittrich, C. (2003). Biostatistical methods for the validation of alternative methods for in vitro toxicity testing. *Altern Lab Animals* **31, Suppl 1,** 5-41.

Enoe, C., Georgiadis, M.P., and Johnson, W.O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev Vet Med* **45,** 61-81.

Environmental Protection Agency (EPA). (1998). Health Effects Test Guidelines. Acute Dermal Irritation. EPA, Office of Pollution Prevention and Toxics Washington DC.

Environmental Protection Agency (EPA) (1998). Chemical Hazard Data Availability Study. What do we really know about the safety of high production volume chemicals? EPA's 1998 baseline of hazard information that is readily available to the public. EPA, Office of Pollution Prevention and Toxics, Washington DC (available at: www.epa.gov/opptintr/chemtest/hazchem.htm).

Eperon, S., De Groote, D., Werner-Felmayer, G., and Jungi, T.W. (1997). Human monocytoid cell lines as indicators of endotoxin: comparison with rabbit pyrogen and Limulus amoebocyte lysate assay. *J Immunol Methods* **207,** 135-145.

Eperon, S., and Jungi, T.W. (1996). The use of human monocytoid lines as indicators of endotoxin. *J Immunol Methods* **194,** 121-129.

European Medicines Agency (1998). *CPMP/ICH/363/96: Note for Guidance on Statistical Principles for Clinical Trials* (adopted by the CPMP in March 1998), 37pp., European Medicines Agency, London.

European Union (1986). EU-Directive 86/609/EEC. *Official Journal of the European Union* **L 358**.

Fennrich, S., Fischer, M., Hartung, T., Lexa, P., Montag-Lessing, T., Sonntag, H.G., Weigandt, M., and Wendel, A. (1999). Detection of endotoxins and other pyrogens using human whole blood. *Dev Biol Stand* **101,** 131-139.

Fentem, J.H., Archer, G.E.B., Balls, M., Botham, P.A., Curren, R.D., Earl, L.K., Esdaile, D.J., Holzhütter, H.G., and Liebsch, M. (1998). The ECVAM

international study on in vitro tests for skin corrosivity. 2. Results and evaluation by the management team. *Toxicol In Vitro* **12,** 483-524.

Fentem, J.H., Briggs, D., Chesne, C., Elliott, G.R., Harbell, J.W., Heylings, J.R., Portes, P., Roguet, R., van de Sandt, J.J., and Botham, P.A. (2001). A prevalidation study on in vitro tests for acute skin irritation. results and evaluation by the Management Team. *Toxicol In Vitro* **15,** 57-93.

Freeberg, F.E., Nixon, G.A., Reer, P.J., Weaver, J.E., Bruce, R.D., Griffith, J.F., and Sanders, L.W. (1986). Human and rabbit eye responses to chemical insult. *Fundam Appl Toxicol* **7,** 626-34.

Garrett, E.S., Eaton, W.W., and Zeger, S. (2002). Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Stat Med* **21,** 1289-1307.

Gerner, I., and Schlede, E. (2002). Introduction of in vitro data into local irritation/corrosion testing strategies by means of SAR considerations: assessment of chemicals. *Toxicol Lett* **127,** 169-175.

Gettings, S.D., Lordo, R.A., Hintze, K.L., Bagley, D.M., Casterton, P.L., Chudkowski, M., Curren, R.D., Demetrulias, J.L., Dipasquale, L.C., Earl, L. K. et al. (1996). The CFTA Evaluation of Alternatives Program: An Evaluation of In Vitro Alternatives to the Draize Primary Eye Irritation Test. (Phase III) Surfactant-based Formulations. *Food Chem Toxicol* **134,** 79-117.

Gigerenzer, G., Hoffrage, U., and Ebert, A. (1998). AIDS counselling for low-risk clients. *AIDS Care* **10,** 197-211.

Gilman, M.R., Evans, R.A., and De Salva, S.J. (1978). The influence of concentration, exposure duration, and patch occlusivity upon rabbit primary dermal irritation indices. *Drug Chem Tox* **1,** 391-400.

Gold, L.S., Slone, T.H., and Ames, B.N. (1998). What do animal cancer tests tell us about human cancer risk?: Overview of analyses of the carcinogenic potency database. *Drug Metab Rev* **30,** 359-404.

Grimes, D. A., and Schulz, K. F. (2002). Uses and abuses of screening tests. *The Lancet* **359,** 881-884.

Harada, T., Misaki, A., and Saito, H. (1968). Curdlan: a bacterial gel-forming beta-1,3-glucan. *Arch Biochem Biophys* **124,** 292-298.

Hartung, T., and Wendel A. (1996). Detection of pyrogens using human whole blood. *In Vitro Toxicol* **9,** 353-359.

Hartung, T., Aaberge, I., Berthold, S., Carlin, G., Charton, E., Coecke, S., Fennrich, S., Fischer, M., Gommer, M., Halder, M., Haslov, K., Jahnke, M., Montag-Lessing, T., Poole, S., Schechtman, L., Wendel, A., and Werner-Felmayer, G. (2001). Novel pyrogen tests based on the human fever reaction. The report and recommendations of ECVAM Workshop 43. European Centre for the Validation of Alternative Methods. European Centre for the Validation of Alternative Methods. *Altern Lab Anim* **29,** 99-123.

Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., Janusch Roi, A., Prieto, P., Sabbioni, E., Worth, A.P., and Zuang, V. (2003). ECVAM's Response to the Changing Political Environment for Alternatives: Consequences of the European Union Chemicals and Cosmetics Policies. *Altern Lab Anim* **31,** 473–481.

Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., Hoffmann, S., Janusch Roi, A., Prieto, P., Sabbioni, E., Scott, L., Worth, A., and Zuang, V. (2004). A modular approach to the ECVAM test principles on test validity. *Altern Lab Animals* **32,** 467-472.

Hauschke, D., and Hothorn, L.A. (1998). Safety assessment in toxicological studies: proof of safety versus proof of hazard. In *Design and analysis of animal studies in pharmaceutical development*. Ed. Chow, S.C., and Liu, J.P., pp.197-225. Marcel Dekker, New York.

Hauschke, D., Hothorn, T., and Schäfer, S. (2003). The role of control groups in mutagenicity studies: matching biological and statistical relevance. *Altern Lab Animals* **31, Suppl 1,** 65-75.

Hawkins, D.M., Garrett. J.A., and Stephenson, B. (2001). Some issues in resolution of diagnostic tests using an imperfect gold standard. *Stat Med* **20,** 1987-2001.

Haynes, R.B., Sackett, D.L., Gray, J.M., Cook, D.J., and Guyatt, G.H. (1996). Transferring evidence from research into practice: 1. The role of clinical care research evidence in clinical decisions. *ACP J Club* **125,** A14-16.

Haynes, R.B., Devereaux, P.J., and Guyatt, G.H. (2002). Clinical expertise in the era of evidence-based medicine and patient choice. *Evid Based Med* **7,** 36-38.

Hoffmann, S., Cole, T., and Hartung, T. (2005). Skin irritation: Prevalence, variability and regulatory classification of existing in vivo data from industrial chemicals. *Regul Toxicol Pharmacol,* in press.

Hoffmann, S., Luederitz-Puechel, U., Montag-Lessing, T., and Hartung, T. (2005). Optimisation of pyrogen testing in parenterals according to different pharmacopoeias by probabilistic modelling. *J Endotoxin Res*, in press.

Horvath, A.R., and Pewsner, D. (2004). Systematic reviews in laboratory medicine: principles, processes and practical considerations. *Clin Chim Acta* **342,** 23-39.

Hunink, M., Glasziou, P., Siegel, J., Weeks, J., Pliskin, J., Elstein, A.S., and Weinstein, M.C. (2001). *Decision making in health and medicine: integrating evidence and values.* Cambridge University Press, New York.

Hutchinson, T.H., Barrett, S., Buzby, M., Constable, D., Hartmann, A., Hayes, E., Huggett, D., Laenge, R., Lillicrap, A.D., Straub, J.O., and Thompson, R.S. (2003). A strategy to reduce the numbers of fish used in acute ecotoxicity testing of pharmaceuticals. *Environ Toxicol Chem* **22,** 3031-3036.

Irwig, L., Bossuyt, P., Glasziou, P., Gatsonis, C., and Lijmer, J. (2002). Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* **324,** 669-671.

Knottnerus, J.A., van Weel, C., and Muris, J.W. (2002). Evaluation of diagnostic procedures. *BMJ* **324,** 477-480.

Knottnerus, J.A., and Muris, J.W. (2003). Assessment of the accuracy of diagnostic tests: the cross-sectional study. *J Clin Epidemiol* **56,** 1118-1128.

Levin, J., and Bang F.B. (1964). A description of cellular coagulation in the limulus. *Bull Johns Hopkins Hosp* **115,** 337-345.

Linnet, K. (1988). A review on the methodology for assessing diagnostic tests. *Clin Chem* **34,** 1379-1386.

Mascoli, C.C., and Weary M.E. (1979a). Applications and advantages of the Limulus amebocyte lysate (LAL) pyrogen test for parenteral injectable products. *Prog Clin Biol Res* **29,** 387-402.

Mascoli, C.C., and Weary M.E. (1979b). Limulus amebocyte lysate (LAL) test for detecting pyrogens in parenteral injectable products and medical devices: advantages to manufacturers and regulatory officials. *J Parenter.Drug Assoc* **33,** 81-95.

Marshall, R.J. (1989). The predictive value of simple rules for combining two diagnostic tests. *Biometrics* **45,** 1213-1222.

McNeil, B.J., Keller, E., and Adelstein, S.J. (1975). Primer on certain elements of medical decision making. *N Engl J Med.* **293,** 211-215.

Morath, S., Geyer, A., and Hartung, T. (2001). Structure-function relationship of cytokine induction by lipoteichoic acid from Staphylococcus aureus. *J Exp Med* **193,** 393-397.

Morath, S., Geyer, A., Spreitzer, I., Hermann, C., and Hartung, T. (2002). Structural decomposition and heterogeneity of commercial lipoteichoic Acid preparations. *Infect Immun* **70,** 938-944.

Nakagawa, Y., Maeda, H., and Murai, T. (2002). Evaluation of the in vitro pyrogen test system based on proinflammatory cytokine release from human monocytes: comparison with a human whole blood culture test system and with the rabbit pyrogen test. *Clin Diagn Lab Immunol* **9,** 588-597.

Ochia, M., Tamura, H., Yamamoto, A., Aizawa, M., Kataoka, M., Toyoizumi, H., and Horiuchi, Y. (2002). A limulus amoebocyte lysate activating activity (LAL activity) that lacks biological activities of endotoxin found in biological products. *Microbiol Immunol* **46,** 527-33.

OECD (1999). Detailed Review Document on Classification Systems for Skin Irritation/Corrosion in OECD Member Countries. ENV/JM/MONO(99)6, Organisation for Economic Cooperation and Development, Paris, 1-29.

OECD (2001a). Guidance Document on Acute Oral Toxicity Testing. 1-24.

OECD (2001b). OECD Guideline for Testing of Chemicals No. 420: Acute Oral Toxicity - Fixed Dose Procedure. 1-14.

OECD (2001c). OECD Guideline for Testing of Chemicals No. 423: Acute Oral toxicity - Acute Toxic Class Method. 1-14.

OECD (2001d). OECD Guideline for Testing of Chemicals No. 425: Acute Oral Toxicity: Up-and-Down Procedure. 1-26.

OECD (2002a). OECD Guideline for Testing of Chemicals No. 404: Acute Dermal Irritation/Corrosion. Organisation for Economic Cooperation and Development, Paris, 1-13.

OECD (2002b). Draft Test Guideline 430: In vitro skin corrosion: Transcutaneous electrical resistance test. Organisation for Economic Cooperation and Development, Paris, 1-12.

OECD (2002c). Draft Test Guideline 431: In vitro skin corrosion: Human skin model test. Organisation for Economic Cooperation and Development, Paris, 1-8.

Peterbauer, A., Eperon, S., Jungi, T.W., Werner, E.R., and Werner-Felmayer, G. (2000). Interferon-gamma-primed monocytoid cell lines: optimizing their use for in vitro detection of bacterial pyrogens. *J Immunol Methods* **233,** 67-76.

Peterbauer, A., Werner, E.R., and Werner-Felmayer, G. (1999). Further development of a cell culture model for the detection of bacterial pyrogens. *ALTEX* **16,** 3-8.

Poole, S., Dawson, P., and Gaines Das, R.E. (1997). Second international standard for endotoxin: calibration in an international collaborative study. *J Endotoxin Res* **4,** 221-231.

Poole, S., Mistry, Y., Ball, C., Gaines Das, R.E., Opie, L.P., Tucker, G., and Patel, M. (2003). A rapid 'one-plate' in vitro test for pyrogens. *J Immunol Methods* **274,** 209-220.

Poole, S., and Mussett, M.V. (1989). The International Standard for Endotoxin: evaluation in an international collaborative study. *J Biol Stand* **17,** 161-171.

Poole, S., Thorpe, R., Meager, A., Hubbard, A.R., and Gearing, A.J. (1988). Detection of pyrogen by cytokine release. *Lancet* **8577,** 130-131.

Pouillot, R., Gerbier, G., and Gardner, I. A. (2002). "TAGS", a program for the evaluation of test accuracy in the absence of a gold standard. *Prev Vet Med* **53,** 67-81.

Prieto, P (2002). Barriers, nephrotoxicology and chronic testing in vitro. *Altern Lab Anim* **30, Suppl 2**,101-6.

Purchase, I.F.H (1990). An international reference chemical data bank would accelerate the development, validation and regulatory acceptance of alternative toxicology tests. *Altern Lab Anim* **18,** 345.

Robinson, M.K., Osborne, R., and Perkins, M.A. (2000). In vitro and human testing strategies for skin irritation. *Ann N Y Acad Sci* **919,** 192-204.

Robinson, M.K., Cohen, C., de Fraissinette, A.B., Ponec, M., Whittle, E., and Fentem, J.H., 2002. Non-animal testing strategies for assessment of the skin corrosion and skin irritation potential of ingredients and finished products. *Food Chem Toxicol* **40,** 573-92.

Roslansky, P. F., Novitsky, T. J. (1991). Sensitivity of Limulus amebocyte lysate (LAL) to LAL-reactive glucans. *J Clin Microbiol* **29,** 2477-2483.

Russel, W M.S., and Burch, R.L. (1959). *The principles of humane experimental technique*. Methuen, London

Sackett, D.L., Haynes, R.B., Guyatt, G.H., and Tugwell, P. (1991). *Clinical epidemiology: a basic science for clinical medicine*. 2nd ed., Little Brown, Boston.

Sackett, D.L., and Haynes, R.B. (2002). The architecture of diagnostic research. *BMJ* **324,** 539-541.

Schindler, S., Bristow, A., Cartmell, T., Hartung, T., and Fennrich, S. (2003). Comparison of the reactivity of human and rabbit blood towards pyrogenic stimuli. *ALTEX* **20,** 59-63.

Schindler, S., Asmus, S., von Aulock, S., Wendel, A., Hartung, T., and Fennrich, S. J. (2004). Cryopreservation of human whole blood for pyrogenicity testing. *Immunol Methods* **294,** 89-100.

Schins, R.P., van Hartingsveldt, B., and Borm, P.J. (1996). Ex vivo cytokine release from whole blood. A routine method for health effect screening. *Exp Toxicol Pathol* **48,** 494-496.

Schlede, E., Mischke, U., Diener, W., and Kayser, D. (1995). The international validation study of the acute toxic class method (oral). *Arch Toxicol* **69,** 659-670.

Spielmann, H., and Liebsch, M. (2002). Validation successes: chemicals. *Altern Lab Anim* **30 Suppl 2,** 33-40.

Spielmann, H. (2003). Validation and regulatory acceptance of new carcinogenicity tests. *Toxicol Pathol* **31,** 54-59.

Spreitzer, I., Fischer, M., Hartzsch, K., Luderitz-Puchel, U., and Montag, T. (2002). Comparative study of rabbit pyrogen test and human whole blood assay on human serum albumin. *ALTEX* **19 Suppl 1,** 73-75.

Society of Japanese Pharmacopoeia (1996). General Tests, Pyrogen Test. In: *The Japanese Pharmacopoeia (English Version).* Tokyo, p. 78-79

Stallard, N., Whitehead, A., and Indans, I. (2003). Statistical evaluation of the fixed concentration procedure for acute inhalation toxicity assessment. *Hum Exp Toxicol* **22,** 575-585.

Stallard, N., Whitehead, A., and Indans, I. (2004). Statistical evaluation of an acute dermal toxicity test using the dermal fixed dose procedure. *Hum Exp Toxicol* **23,** 405-412.

Stitzel, K.A. (2002). Tiered Testing Strategies – Acute Local Toxicity. *ILAR J* **43,** Suppl, 21-26.

Taktak, Y.S., Selkirk, S., Bristow, A.F., Carpenter, A., Ball, C., Rafferty, B., and Poole, S. (1991). Assay of pyrogens by interleukin-6 release from monocytic cell lines. *J Pharm Pharmacol* **43,** 578-582.

The United States Pharmacopeial Convention, I. (2000). Biological Test, <151> Pyrogen Test. In: *The United States Pharmacopoeia 24*. Rockville, p. 1850-1851

Tschumi, J. (2003). Comparison of temperature rise interpretations between European and United States Pharmacopeias' pyrogen tests. *PDA J Pharm Sci Technol* **57,** 218-227.

Tsuchiya, S., Yamabe, M., Yamaguchi, Y., Kobayashi, Y., Konno, T., and Tada K. (1980). Establishment and characterization of a human acute monocytic leukemia cell line (THP-1). *Int J Cancer* **26,** 171-176.

Twohy, C.W., Duran, A.P., and Munson, T.E. (1984). Endotoxin contamination of parenteral drugs and radiopharmaceuticals as determined by the limulus amebocyte lysate method. *J Parenter Sci Technol* **38,** 190-201.

UN (2003). Skin Corrosion/Irritation, in: *UN, Globally harmonized system of classification and labelling of chemicals*. ST/SG/AC.10/30, United Nations, New York and Geneva, 123-135.

van den Heuvel, M.J., Clark, D.G., Fielder, R.J., Koundakjian, P.P., Oliver, G.J., Pelling, D., Tomlinson, N.J., and Walker, A.P. (1990). The international validation of a fixed-dose procedure as an alternative to the classical LD50 test. *Food Chem Toxicol* **28,** 469-482.

van der Schouw, Y.T., Verbeek, A L., and Ruijs, S.H. (1995). Guidelines for the assessment of new diagnostic tests. *Invest Radiol* **30,** 334-340.

Walter, S.D., Irwig, L., and Glasziou, P P. (1999). Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol* **52,** 943-951.

Weary, M., Pearson, F.C., Bohon, J., and Donohue, G (1982). The activity of various endotoxins in the USP rabbit test and in three different LAL tests. *Prog Clin Biol Res* **93,** 365-379.

Weller, S.C., and Mann, N.C. (1997). Assessing rater performance without a 'gold standard' using consensus theory. *Med Decis Making* **17,** 71-79.

Weil, C.S., and Scala, R.A. (1971). Study of intra- and interlaboratory variability in the results of rabbit eye and skin irritation tests. *Toxicol Appl Pharmacol* **19,** 276-360.

Weisser, K., and Hechler, U. (1997). *Animal welfare aspects in the quality control of immunobiologicals.* FRAME, Nottingham.

Whiting, P., Rutjes, A. W., Reitsma, J., Bossuyt, P. M., and Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* **3,** 25.

Worth, A.P., Fentem, J.H., Balls, M., Botham, P.A., Curren, R., Earl, L.K., Esdaile, D.J., and Liebsch, M. (1998). An Evaluation of the proposed OECD Testing Strategy for Skin Corrosion. *Altern Lab Anim* **26,** 709-720.

Worth, A.P., and Balls, M. (2001). The importance of the prediction model in the validation of alternative tests. *Alt Lab Animals* **29,** 135-143.

Worth, A.P., and Cronin, T.D. (2001). The use of Bootstrap Resampling to assess the variability of Draize Tissue Scores. *Altern Lab Anim* **29,** 557-573.

Worth, A.P., and Balls, M. (2002). The principles of validation and the ECVAM validation process. *Altern Lab Anim* **30 Suppl 2,** 15-21.

Ziegler-Heitbrock, H.W., Thiel, E., Futterer, A., Herzog, V., Wirtz, A., and Riethmuller, G. (1988). Establishment of a human cell line (Mono Mac 6) with characteristics of mature monocytes. *Int J Cancer* **41,** 456-461.

Zuang, V., Balls, M., Botham, P.A., Coquette, A., Corsini, E., Curren, R.D., Elliott, G.R., Fentem, J.H., Heylings, J.R., Liebsch, M., Medina, J, Roguet, R., van de Sandt, J.J., Wiemann, C., and Worth, A.P. (2002). Follow-up to the ECVAM prevalidation study on in vitro tests for acute skin irritation. European Centre for the Validation of Alternative Methods Skin Irritation Task Force report 2. *Altern Lab Anim* **30,** 109-29.