Pharmacology based toxicity assessment : towards quantitative risk prediction in humans

Sahota, T.

CHAPTER 1

Challenges in the assessment and prediction of safety pharmacology and drug toxicity in humans

Tarjinder Sahota, Meindert Danhof and Oscar Della Pasqua

Abstract

Despite ongoing efforts to better understand the mechanisms underlying safety and toxicity, approximately 30% of the attrition in drug discovery and development is still due to safety concerns. Changes in current practice regarding the assessment of safety and toxicity are required to reduce late stage attrition and enable effective development of novel medicines. This review focuses on the implications of empirical evidence generation for the evaluation of safety and toxicity during drug development. A shift in paradigm is proposed to 1) ensure that pharmacological concepts are incorporated into the evaluation of safety and toxicity; 2) facilitate the integration of historical evidence and thereby the translation of findings across species; and 3) promote the use of experimental protocols tailored to address specific safety and toxicity questions.

Based on historical examples, we highlight the challenges for the early characterisation of the safety profile of a new molecule and discuss how model-based methodology can be applied for the design and analysis of experimental protocols. Issues relative to the scientific rationale are categorised and presented as a hierarchical tree describing the decision making process. Focus is given to four different areas, namely, optimisation, translation, analytical construct, and decision criteria. From a methodological perspective, nonlinear-mixed effects modelling is recommended as a tool to account for such requirements. Its use in the evaluation of pharmacokinetics (PK) and pharmacokinetic-pharmacodynamic relationships (PKPD) has enabled the advance of quantitative approaches in pharmacological research in recent decades. Comparable benefits can be anticipated for the assessment of safety and toxicity.

1. Introduction

The assessment of the safety and toxicity profile of new chemical or biological entities is an integral part of drug development. Despite ongoing efforts to better understand the mechanisms underlying safety and toxicity, approximately 30% of the attrition in drug discovery and development is still due to safety concerns (1,2). Such a high attrition rate is further compounded by the empiricism and entrenched belief which prevails among industry scientists and regulators about the level of evidence and requirements for determining acceptable risk in humans.

In addition to its contribution to the attrition rate, safety and toxicity findings have business, legal and societal consequences, which often lead to speculations and even more empiricism in the evaluation and interpretation of experimental data. Whilst a positive benefit-risk ratio should be anticipated and subsequently demonstrated when administering new drugs to humans, the basis upon which inferences are made still lacks the scientific clarity and rigour one would endeavour. The efficiency and value of current paradigm for the evaluation of safety and toxicity, which relies primarily on standard battery tests at supra-therapeutic exposure levels of the investigational drug, is not questioned by the scientific community. Rather, it is mandated by regulators as a mechanism to minimise liabilities.

A shift in paradigm is required that 1) enables the introduction of pharmacological concepts to the evaluation of safety and toxicity; 2) facilitates the integration of historical evidence and thereby the translation of findings across species; and 3) promotes the value of experimental protocols tailored to address specific safety and toxicity questions.

In this review we will focus on the implications of current practice for drug development and consider the scientific and ethical requirements for the evaluation of safety and toxicity. Of particular interest for us is to demonstrate that despite the assumption that preclinical safety testing, toxicity findings are generally seen as predictive of human toxicity (3), inefficiencies in the experimental design violate the principle of the 3 Rs (reduction,

refinement and replacement) (4). Empirical evidence must be replaced by a model-based approach.

Two recent examples can be used to illustrate the issues with the current paradigm for the evaluation of safety and toxicity, namely the serious adverse events observed with TGN1412 and the increased incidence of myocardial infarction in patients who were prescribed rofecoxib. These two cases encompass most of the critical issues one attempts to address prior to making a commitment to clinical development and subsequently to regulatory submission and marketing of a medicinal product. Albeit neglected in the assessment of the clinical findings and in the subsequent reports in the published literature, the use of a mechanism-based approach in conjunction with some basic pharmacology concepts would be sufficient to predict the consequences of the treatment, whether given as single dose to healthy subjects or chronically to patients; i.e., both examples reflect the immediate consequences of target engagement and the corresponding changes due to the mechanism of action and (patho)physiological pathways. Yet, the experimental evidence generated pre-clinically for these two compounds does not take into account target engagement or exposure-response relationships as the basis for the interpretation of the findings. Instead, it is the characterisation of the maximum tolerated dose (MTD) and /or no-adverse effect level (NOAEL) that ultimately drives the design of safety pharmacology and toxicity experiments. The empirical evidence of MTD and NOAEL does not provide insight into the underlying mechanisms and often obscures the translation of findings across species.

According to published reports, the serious adverse events observed after intravenous administration of TGN1412, a novel monoclonal T-cell agonist, could not have been "predicted" or inferred from non-clinical data. The empiricism in the design of the experimental protocol and in the interpretation of the findings clearly shows the disconnection between pharmacology and toxicology, despite extremely high degree of selectivity and specificity of the biologicals. The failure to predict a systemic inflammatory response by rapid induction of cytokines (a "cytokine storm") with catastrophic multi-organ failure (5) is not surprising when structure homology, target occupancy and pharmacokinetic

principles are disregarded. Despite the availability of in vitro binding assays, there was no attempt to correlate or integrate the results from different experiments with each other. Most importantly, the effects observed with the proposed dosing regimen could have been anticipated even without any experimental data. Knowledge of receptor agonism theory and drug disposition properties would have been sufficient to make inferences about target activation and pharmacological effects.

Tragedies like this provoke reactive measures from industry and regulator (6-11). New guidelines for the assessment of preclinical data were released by regulatory authorities. However, none of them tackle the problem from a scientific, mechanistic perspective. Similarly, changes have been introduced to the design of first-in-man studies (6), which reflect mitigation measures for process-related consequences of safety and toxicity findings. A framework that ensures critical appraisal of the scientific rationale, based on pharmacological concepts and expected biological activity (i.e., target engagement) is still missing.

Rofecoxib, a selective COX-2 inhibitor prescribed to more than 107 million patients in the US (12), is another example of withdrawal from the market because of so-called “unexpected” long-term safety findings. Despite the debate that followed the evidence from clinical trial data on the increased risk of myocardial infarction (13), little effort was made to incorporate very basic pharmacological concepts into the evaluation of the findings and provide a mechanism-based interpretation, which could easily disentangle the core issue: whether this is a class-effect or whether that was a compound specific toxicity. Paracelsus highlighted the importance of the dose more than 500 years ago, and yet none of the published reports considered this critical question: were patients receiving the optimal dose and dosing regimen for the proposed indications? Clinical and scientific experts dwelled on the realm of toxicity as the result of an off-target event, without exploring in a systematic manner the (obvious) connection to dosing regimen, target exposure, the time course of pharmacological effect, the duration of treatment and physiological role of the substrates for COX2 in the heart and other tissues. Evidence of concentration-effect relationship was

not gathered, neither used as basis for interpreting those findings. Instead, allegations of misconduct followed that overruled any comprehensive scientific debate (12).

From a clinical pharmacology perspective, the aforementioned examples reflect the failure in exploring causality and anticipating the biological consequences of target engagement, i.e., in establishing the correlation between target-related events and drug exposure, as defined by the evidence of pharmacokinetic-pharmacodynamic relationships. Post-market withdrawals are not an uncommon occurrence: between 1975 and early 2000 there have been 26 withdrawals from the US market due to safety issues (14). In fact, the withdrawal of a medicinal product seems to have become the expected *course of action* for regulators and industry who are faced with 'unexpected' safety findings. Interestingly, dosage changes, due to safety occurred in approximately one out five drugs in the period from 1980 to 1999 (15,16). On the other hand, from a clinical perspective, the aforementioned landscape appears to result from the lack of a formalised assessment of the benefit-risk ratio in which efficacy and safety are evaluated in an integrated manner. Different stakeholders appraise the problem from a distinct point-of-view without acknowledging the intrinsic, albeit indirect, link between dosing regimen, exposure, target engagement and clinical events.

The incorporation of model-based concepts and pharmacokinetic-pharmacodynamic relationships into the rationale for the design, analysis and interpretation of safety pharmacology and toxicology protocols is vital for the future of screening of novel compounds and for an effective shift in the assessment of safety and acceptable risk in drug discovery and development. More than just enabling a framework for modernisation of outdated methods and techniques, a model-based approach challenges the mainstream scientific views about the role of experimental evidence as the sole basis for the assessment of non-clinical safety; it unravels the strength of inferential methods and evidence synthesis.

In this review, we aim therefore at identifying the pitfalls in current approaches to estimating and predicting safety pharmacology and toxicity in humans. Focus is given to the estimation of safety thresholds and decision making, with special emphasis on the

underlying methodological issues. Our objectives may intersect with the message from other reviews of safety in humans (17,18). However, our concerns go beyond technical aspects of experimental and statistical methods; the objective of the larger research to be presented in this thesis, is to detail improved techniques for data analysis and study design, as well as to illustrate how a mechanism-based approach for risk assessment can formally be applied to support more accurate decision making.

In the subsequent sections, we will cover a wide range of methodological and conceptual issues, starting with low level problems, which usually comprise experimental aspects or relate to the statistical methods. Given their technical nature, implementation of the proposed recommendations requires little effort and can be relatively straightforward, as compared to higher level problems, which involve conceptual features and require a different attitude towards the generation, analysis and interpretation of experimental data regarding safety and toxicity. From a theoretical perspective, different facets of the same problem will be discussed, which relate to four seminal areas of scientific research: 1. **optimisation** (e.g., accuracy, precision), 2. **translation** (e.g., sensitivity, biological substrate, relevance), 3. **analytical construct** (e.g. choice of parameterisation) and 4. **decision criteria** (e.g., acceptable risk level). Each of these points will be addressed separately.

As shown in Figure 1, on the most basic level of the hierarchical tree is the choice of the measure of drug exposure and endpoint selected for the assessment of safety. These issues are compounded by the use of point estimates and by statistical inferences regarding the reporting of safety thresholds. Experimental design considerations in relation to type I and II errors constitute the next level of attention. The drawbacks of the use of empirical approaches as opposed to mechanism-based approaches will be covered. Empiricism here relates to data analysis methods which are primarily descriptive rather than explanatory of the observed phenomena. Of particular interest is the current dichotomisation of the problem using safety thresholds. This will be followed by a critique of allometric scaling to predict exposure in humans and then more generally the manner in which risk is translated into decisions.

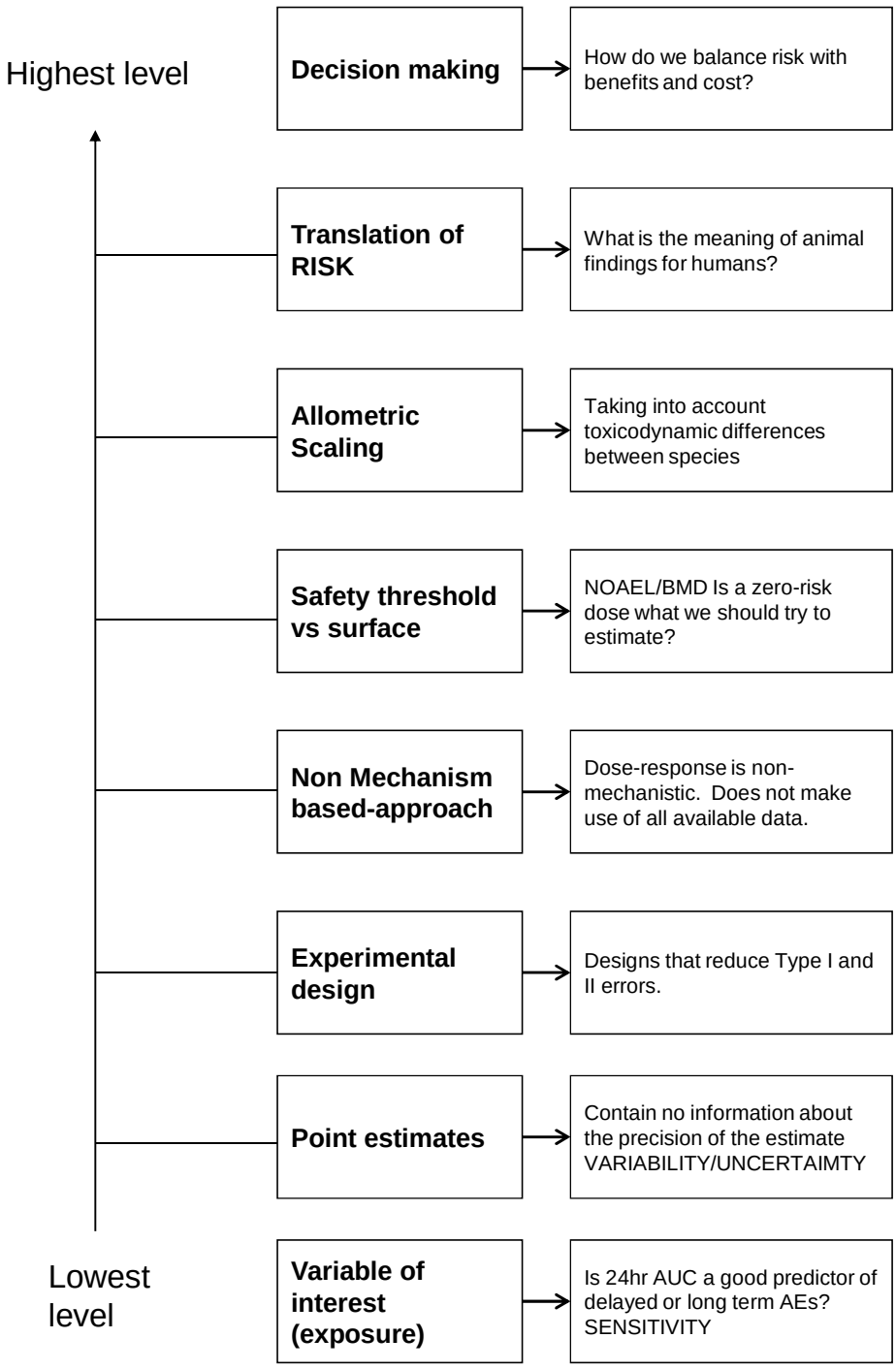

Figure 1: A hierarchical tree describing the different levels and issues underpinning decision making during the assessment of safety and toxicity profile of a new chemical entity.

2. Nonclinical evaluation of safety and toxicity

2.1. Defining variables of interest.

The development of a pharmaceutical is a stepwise process involving an evaluation of both animal and human efficacy and safety information. The goals of the nonclinical safety evaluation generally include a characterisation of toxic effects with respect to target organs, dose dependence, relationship to exposure, and, when appropriate, potential reversibility. This information is used to estimate an initial safe starting dose and dose range for the human trials and to identify parameters for clinical monitoring for potential adverse effects. Toxicity occurs when the drug-induced alteration of biological function overcomes normal repair and homeostatic mechanisms. Toxicity can be measured by its effects on the target (organism, organ, tissue or cell) or indirectly by measuring altered biological function downstream after acute, sub-chronic or chronic exposure to a chemical or biological entity. Drug exposure is then used as a proxy or surrogate for the undesirable effects. It should be noted that an adverse event is any undesirable experience associated with the use of a medical product, irrespective of the evidence of a causal relationship between drug and adverse event. However, from a drug development perspective, different aspects of safety and toxicity need to be evaluated experimentally, which encompass the expected therapeutic and supra-therapeutic dose levels. Although different experimental protocols must be implemented during the development of a new compound, the evaluation of immunotoxicity, genotoxicity, carcinogenicity, phototoxicity, abuse liability and reproductive performance and developmental toxicity are beyond the scope of this review. The nonclinical safety and toxicity studies should be adequate to characterise potential adverse effects that might occur under the conditions of the clinical trial to be supported. Serious nonclinical findings can influence the continuation of the development programme and of clinical trials.

Despite the different protocols for the assessment of safety and toxicity and the myriad of adverse events one may come across, a common practice in this field of research is the assessment of empirical safety thresholds such as the no observed adverse effect level

(NOAEL), which are no more than *qualitative* indicators of acceptable risk. Support for the existence of thresholds has been argued on biological grounds (19-21). The argument is that although any exposure to a chemical will cause some change in the biological system, the change must override homeostatic mechanisms in order for it to be biologically significant. In contrast to the maximum tolerated dose (MTD), which remains the primary endpoint of choice in the evaluation of chronic toxicity, the NOAEL is one of the main indicators of risk in nonclinical safety assessment. Definitions of the NOAEL vary from source to source, however the basis behind all of them is the estimation of "the highest experimental point, without biologically significant adverse effects that are above baseline" (22). In fact, the experimental findings are used to reflect another threshold, i.e., the underlying no adverse event level (NAEL). The calculation involves determination of the lowest observed adverse effect level (LOAEL) which is the lowest observed dosing level for which AEs are recorded. The NOAEL is the dosing level below this. If no LOAEL is found, then the NOAEL cannot be determined. In these cases the LOAEL/10 is sometimes used in place of the NOAEL.

Drug exposure and risk can be represented by a variety of different experimental measures. Usually, in the NOAEL approach, the measures used are dosing level, area-under-concentration-time-curve (AUC) and/or maximum concentration (CMAX). On the other hand, the benchmark dose (BMD) is an alternative to the NOAEL. The method involves the construction of a model of the exposure-AE relationship to predict the dosing level that corresponds to the threshold between non-significant and significant risk of AEs. The quantity is usually expressed as a dose level rather than an AUC or CMAX, but the BMD remains of limited use in Industry (23).

Another common measure is the human equivalent dose (HED), which represents the estimated dose level in humans yielding equivalent drug exposure as observed in animals at the safety threshold (23). In addition, recommendations have been made for the use of the maximum recommended starting dose (MRSD) for the selection of the starting doses in first-in-human studies. The MRSD is believed to minimise the chance of serious adverse events in early clinical studies (7,23). Recently, the minimum anticipated biological effect level

(MABEL) has also been introduced to assist in selection of doses for first in man studies and to supplement existing approaches. MABEL describes the exposure that is anticipated, prior to clinical testing, to produce a minimum biological effect level (24,25).

Given the empirical nature of such safety thresholds, errors in the prediction of safety may arise. Despite the various options, there is still a real safety concern when using these thresholds to extrapolate drug exposure levels from animals to humans and to make inferences from short to long term effects. Unfortunately, instead of pursuing a more mechanistic approach, empirical methods continue to be used. To cope with inaccuracy and poor precision, safety factors, also known as uncertainty factors, have been incorporated on the top of empirical thresholds. Their application in drug development has become widespread (26) and is detailed within the regulatory guidelines. The purpose of such safety factors is to account for variability potentially greater toxicity in humans than predicted by the HED using existing approaches. This is to ensure that the safety threshold is beneath the true threshold. The default safety factor is 10, but it can by modified by considering it as a product of more refined uncertainty factors. These comprise; interspecies uncertainty, UF_A, interindividual uncertainty UF_H, subchronic to chronic uncertainty, UF_S, LOAEL to NOAEL uncertainty, UF_L, and data adequacy UF_D, for when chronic toxicity studies in at least two different species are unavailable (27,28). There is also a modification factor where there is a perceived greater risk of toxicity in humans.

It should be noted that even when safety factors are factored into the estimation of thresholds, the actual risk a treatment represents to humans can be overlooked. Over-conservative attitude may give the wrong perception of caution. Accurate assessment of risk can simply not be performed without some degree of understanding of target engagement and nature of the ligand (i.e., agonistic or antagonistic interaction with the target).

2.2 Measures of drug exposure used as descriptors of acceptable risk

A consequence of the use of safety thresholds is the estimation of drug exposure or dose levels that can be correlated with the adverse events observed beyond that specific threshold, for which the risk for humans is deemed unacceptable. Numerous assumptions are however required to ensure accurate translation of such findings from animals to humans. To be predictive, the exposure levels and the adverse events must reflect pathophysiological processes and pharmacokinetics in humans.

Different measures of exposure are used in reports. The most basic of these is dosing level, which is usually expressed in terms of daily dose (e.g., mg/day). Dose, however may be a poor indicator of response since it does not account for confounders such as bioavailability, differences in metabolic capacity, or other pharmacokinetic processes that alter target exposure despite comparable dose. For this reason, parameters derived from the assessment of systemic drug concentrations are preferred (e.g., AUC and C_{MAX}). The choice for those parameters relies on the assumption that rapid equilibration occurs between systemic circulating drug and the target tissue. Given the fragmented process used for the evaluation of pharmacology and toxicology data, the validity of this assumption is questioned even when evidence from pharmacological and pharmacokinetic data indicates otherwise. Nonlinearity in drug disposition is another important pharmacokinetic aspect which is not accurately captured by the use of dose as a measure of drug exposure. Differences in systemic and target exposure can be large in the case of metabolic saturation, when small increments in dose can produce disproportionately large increases in AUC. This can lead to deceptively safe estimates even if the dose is divided by a safety factor. Conversely, the occurrence of metabolic induction may lead to overly conservative dose selection.

In addition to the aforementioned points, it is also critical to understand the implications of the use of systemic levels as compared to target tissue or target organ exposure. Time-dependent processes take place which cannot be neglected or inferred from conventional measures of exposure. First, one should realise that given that pharmacokinetic equilibration

between plasma and tissue may not always be assumed. Unbound drug concentrations are primarily distributed into tissues. The extent and rate of distribution depend on physicochemical as well as receptor binding properties. The implications of such processes are that irreversible binding, slowly reversible binding and tissue accumulation may not be easily correlated with circulating total concentrations. From a pharmacodynamic perspective, the same considerations must be made when signal transduction and downstream mechanisms are rate limiting for the onset and maintenance of effects (i.e., adverse events). Consequently, the use of AUC and C_{MAX}, expressed over a single day may not accurately reflect the underlying relationship between exposure and adverse event. The implicit assumption that there is a correlation between "daily" drug exposure and risk is suitable mainly for direct and reversible processes; however it is insufficient to account for the complex nature of indirect effects, slowly reversible and irreversible binding.

These complexities can be illustrated by permetrexed-induced neutropenia. Absolute neutrophil count (ANC) is reduced by inhibition of thymidylate synthase, dihydrofolate reductase and glycinamide ribonucleotide formyltransferase (29). The trough of the ANC curve occurs between 8 and 9.6 days after dosing (30), and is followed by an overshoot effect once levels return to baseline (Figure 2). Empirical approaches are in principle able to quantify the PK exposure associated with a particular ANC minimum, however this ignores the complexity of the ANC-curve. The time below a threshold ANC may be a more relevant descriptor of risk and will require a different measure (i.e., parameterisation) of drug exposure. Most importantly, in these circumstances the time course of drug effects (onset, duration and washout) often does not correlate with daily systemic exposures.

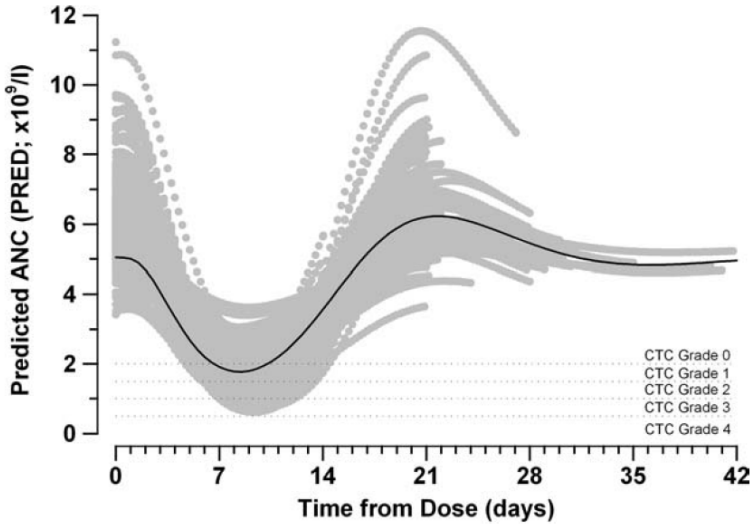

Figure 2. Time course of predicted absolute neutrophil counts (PRED) following 500 mg/m^2 pemetrexed. Lines: Solid black curve the overall ‘‘typical’’ patient in the analysis dataset (i.e., median values for each of the covariates contained in the final PKPD model); gray shading predictions based on the population PKPD model for each of the patients in the analysis dataset, assuming a 500 mg/m^2 dose; dashed horizontal lines hematologic toxicity grades (grade 1 <2, grade 2 <1.5, grade 3 <1, grade 4 <0.5) (30).

Likewise, irreversible binding mechanisms cause drug accumulation at the effect site yielding adverse events that depend primarily on the treatment duration, rather than on daily exposure. Measures that do not capture the cumulative nature of these processes may lead to poor correlation between species. Measures such as cumulative AUC may provide better prediction than 24-hour AUC since the entire dosing history is used. Figure 3 shows an example of such an effect is tardive dyskinesia produced by neuroleptic drugs (31).

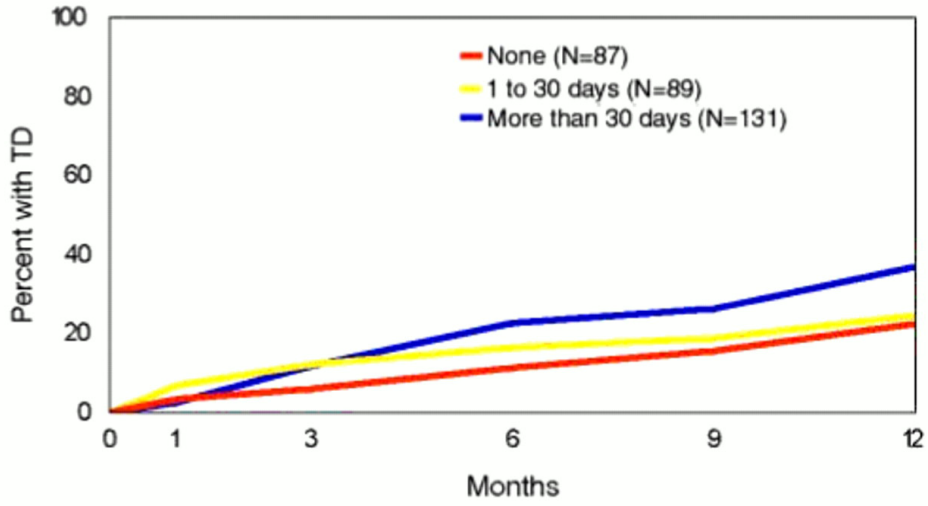

Figure 3. Curve showing incidence of tardive dyskinesia given cumulative neuroleptic exposure. Patients with more than 30 days of neuroleptic use at baseline had a trend for a greater cumulative incidence of tardive dyskinesia than those with 0-30 days of neuroleptic use (31).

Although safety factors have been used to account for possible inaccuracies in the estimates of safety thresholds, there are translational aspects that cannot be factored in by such an empirical approach. A systematic, rational translation of findings across species requires the use of mechanism-based approaches to assess the implications of differences in pharmacokinetics, pharmacodynamics as well as in pathophysiology. Of particular importance is the fact that between species variability in metabolic rate and capacity can lead to completely different safety profiles across specie if metabolites are the moiety underlying adverse events. Likewise, molecules that are substrate to active transporters, carrier-mediated processes and other distribution mechanisms with known species-specific differences will show discrepancies in safety profile.

2.3 Statistical and biological limitations of point estimates

From a statistical perspective, safety thresholds are often presented as point estimates to describe the population. This ignores variability which can be decomposed into two parts; variability associated with estimation methods and real variability in response between and within subjects. There is also lack of best practice in statistical inference. Risk is inferred from toxicology results using statistics that may be imprecise or inaccurate. A statistic is a random variable which is typically a function of the experimental data (such as, e.g., a mean or an observed rate). Statistics are intended to provide an estimate of underlying parameters reflecting physiological processes and/or pharmacokinetics. The implications of such practice can be illustrated by the comparison between *sample* standard deviation and *population* standard deviation. The former is a statistic and the latter is the inferred parameter. The equivalent for NOAEL is the no adverse effect level (NAEL). The term "NOAEL estimate" is a misnomer in that it is the NAEL, which is being estimated by the NOAEL (see figure 4). Based on statistical concepts, it can be shown that meaningful and useful reporting of toxicology findings should be of the estimate of NAEL with its precision (standard errors). However, an empirical approach prevents the estimation of uncertainty in the NAEL.

The use of statistics, in place of model parameters for decision making can treat the estimate as if it were of sufficient precision to give sufficiently narrow confidence intervals. This limitation is believed to be mitigated by the incorporation of safety factors, an assumption which we dispute. As can be seen in Table 1, the parameter precision for the probability of an AE varies from 1587 to 67%, depending on group size and risk. The number of animals in a group exhibiting an adverse event is often reported however, the performance of this estimator is highly dependent on the underlying risk of the AE in question and the sample size.

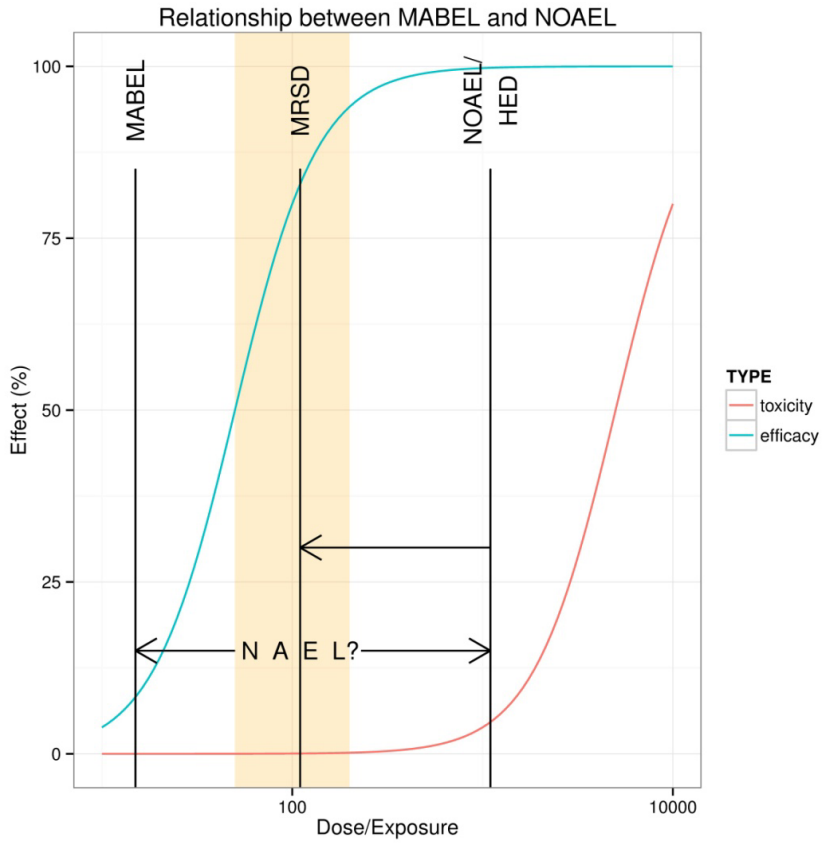

Figure 4. Relationship between MABEL and NOAEL/HED. Shaded region indicates the expected therapeutic range.

Table 1: Parameter precision for probability of adverse events

Risk of AE	n=4	n=8	n=10	n=16	n=20
0.10%	1578.77%	1116.75%	998.08%	790.59%	707.92%
1.00%	497.41%	351.67%	314.71%	248.64%	222.50%
5.00%	217.98%	154.14%	137.83%	108.98%	97.46%
10.00%	150.03%	106.07%	94.87%	74.99%	67.08%

AEs were assumed to be independent binary events. The estimator is the number of animals as a percentage of n. Values depicted coefficients of variation.

The other aspect to missing variability is the real variability in the data. Since sampling in toxicology is often very sparse, exposure levels are calculated from satellite groups which mirror dosing of the animals assigned to the primary treatment group. This ignores real differences that may be present between the two groups. It is equivalent to assuming that all animals have the same exposure and variability in exposure or the underlying physiological processes is not responsible for variability in response.

Given that human variability is typically larger, it is important to understand the role of the different sources of variability. Without quantification of variability and identification of covariates it becomes difficult to predict which groups are more prone to overexposure or more sensitive to adverse events. Furthermore, depending on the actual distribution of drug exposure, the distribution of AEs in this group may not be representative of the risk posed to the overall target population. This is not limited to pharmacokinetics, pharmacodynamic differences have the greater potential for harm and can be more variable than pharmacokinetic differences. In this case, hypersensitive subpopulations can be completely missed. This is the case of abacavir-induced rash and other dose-independent reactions associated with receptor or target polymorphism.

Finally, it should be noted that empirical approaches remain prone to bias. For example, the mean NOAEL is only unbiased if its underlying distribution is symmetrical. This practice ignores that such a summary violates current understanding of pharmacokinetic processes, which are best described by lognormal distributions. Without clear assumptions of the underlying distribution, the choice of measure for central tendency remains unjustifiable and may lead to bias.

2.4 Mechanism-based assessment of safety, toxicity and risk

Whilst the introduction of regulatory policies for the non-clinical evaluation of medicinal products in humans, at a time when understanding about receptor pharmacology and pathophysiology was very limited, partly explains the historical evolution of current

standards and practice in safety and toxicity research, its perpetuation is no longer justifiable. It is evident that the concept of safety thresholds as well as the measures of exposure used as *proxy* for acceptable risk cannot be deemed absolute: they rely upon numerous assumptions, which may not hold true in a considerable number of cases. In principle, information regarding the causal chain between target engagement and adverse events should be used as basis for relevant measures of exposure and risk. This concept can be implemented even in the absence of evidence for the actual target or mechanism underlying a given adverse event or undesirable effect. Sufficient evidence exists to support the use of concentration-effect relationships to identify the rate limiting step in the chain of events from dose to response. In conjunction with tailored experimental protocols and pharmacokinetic-pharmacodynamic modelling, a mechanism-based evaluation of safety findings provides the basis for characterising safety and toxicity. Moreover, it should be noted that safety and toxicity findings may not solely depend on pharmacokinetic drug exposure, but also on the extent of target activation or inhibition, post-receptor amplification and signal transduction processes as well as homeostatic mechanisms. For instance, drug concentrations may be a poor predictor of risk relative to the relevant biomarker concentrations when signal transduction is the rate limiting step for a given response. It is unfortunate that despite the wide discussion regarding the use of biomarkers in the literature (32-35), the focus has primarily been on the assessment of efficacy, not safety.

In addition, experimental protocols and data analysis have not advanced in the same way risk management concepts have evolved over the last decade. Causality has become pivotal for the characterisation of adverse drug reactions, which in contrast to adverse events, are defined as any noxious unintended and undesired effects of a drug that occur at doses used for prevention, diagnosis or treatment. This subtle difference in definition has major consequences for the evaluation of safety, toxicity and risk, including experimental protocol requirements. Rawlins and Thompson devised a classification scheme in 1991, which continues to be the most frequently used in clinical research, which could be used as the basis for the assessment of nonclinical safety. Their scheme, shown in Table 2, defines

adverse drug reactions according to seven different categories, which account for the underlying chain of events. The different categories nicely match the mechanistic classification of biomarkers proposed by Danhof et al., and could form the basis for a new paradigm for the evaluation of nonclinical safety and toxicity (34).

Table 2. Classification of adverse drug reactions, as proposed by Rawlins and Thompson.

Type "A": Predictable, common and related to Pharmacological action of the drug		
Toxicity of overdose:	e.g. hepatic failure	paracetamol
Side effects:	e.g sedation	Antihistaminergic drugs
Secondary effects:	e.g. development of	antibiotic therapy
Drug interaction:	e.g. Theophylline toxicity	erythromycin therapy
Type "B": Unpredictable, uncommon, usually not directly related to the mechanism or pharmacological actions of the drug.		
Intolerance:	e.g. tinnitus	Aspirin
Hypersensitivity:	e.g. anaphylaxis	penicillin
Pseudoallergic:	(Non-Immunological)	radio contrast dye reaction
Idiosyncratic reaction:	e.g. anaemia due to glucose-	anti-oxidant drugs
Type "C": These reactions are associated with long-term drug therapy e.g. Benzodiazepine dependence and Analgesic nephropathy. They are well known and can be anticipated.		
Type "D": These reactions refer to carcinogenic and teratogenic effects. These reactions are delayed in onset and are very rare since extensive mutagenicity and carcinogenicity studies are done before drug is licensed.		
Type "E" : The end of treatment or rebound effects		
Type "F" : Failure of treatment		
Type "G" : due to genetic polymorphism, not immunologically mediated		

As it can be seen from Figure 5, biomarkers can be associated or linked to one of the reaction types in Rawlins and Thompson's classification. Undoubtedly, these concepts allow the causal chain of events to be correlated to the time course of overt symptoms and signs in a quantitative manner. Such an integrated approach is essential for accurate (mechanistic) interpretation of risk in humans. In the next paragraphs, we will highlight how distant these concepts are from the approaches currently used in the assessment of nonclinical safety and toxicity.

Figure 5. Mechanistic classification of biomarkers.

2.5 Minimum Anticipated Biological Effective Level

A first attempt to implement mechanism-based measures of exposure has evolved over the last decade, which relies on the assessment of the minimum anticipated biological effect level (MABEL). In the calculation of the MABEL any biomarker can be used, for example receptor occupancy or even downstream markers such as physiological mediators (25). This has the advantage of allowing measures that correlate to any target-related toxicity (i.e., including off-target or secondary target) when pharmacokinetic processes are not the rate limiting step. The concept relies on the assumption of some knowledge of the putative targets underlying the adverse event to accurately interpret (patho)physiological response and assess causality. However, since the MABEL is defined in terms of biological effect, not toxicity, it is not a measure of risk and not a replacement of the NOAEL. Current practice is therefore to use the NOAEL as a measure of risk to guide maximum doses in dose escalation

studies, but the maximum recommended starting dose in FTIH should now be no higher than both the MABEL and the NOAEL-derived MRSD. If the NOAEL with the addition a safety factors were indeed protective, such a measure would be an unnecessary. Yet, one needs to acknowledge that the MABEL is simply a retrospective risk-mitigation measure that can account for some of the deficiencies of the NOAEL approach.

2.6. Limitations in experimental design

There are methodological aspects that need to be addressed to allow wider use of MABEL or any other mechanism-based measures of 'acceptable risk'. The predictive or prognostic value of statistical correlations depends on satisfying five important criteria, namely: selectivity, specificity, sensitivity, reproducibility and clinical relevance. Currently, despite the characterisation of a correlation between biomarker and response, very little effort has been made to quantify estimators such as false positive and false negative rates. For instance, liver enzyme levels provide an example of a biomarker which has high sensitivity but poor specificity. Interestingly, despite the aforementioned limitations clinical scientists and pathologists will defend the value of ALT, AST and bilirubin as better predictors of risk, as compared to drug exposure. Another aspect of interest is the fact that according to current practice, if e.g., elevated liver enzymes are observed in one individual and acute liver failure in another, an empirical framework ignores the correlation between these adverse events. It may be treated as the same adverse event (i.e., 100% correlation), or a two different adverse events (i.e., uncorrelated). The statistical methods and summary measures of toxicity are unable to account for partial correlation or interaction between events within or between individuals.

So the question is why does one not go further along the causal chain of toxicity for all adverse events, instead of relying on measures of systemic exposure? The answer probably lies in that pharmacokinetics is seen as the primary step along the way for most adverse drug reactions. It is a simple, general purpose measure which fits the criterion of providing

predictive value for many adverse events, despite the exceptions, for which it will perform poorly, with low predictivity.

From a theoretical point of view, it should be highlighted that empirical approaches perform poorly when incidence of a type of adverse event is low (Table 3). This means that pooling data across different types of adverse events is necessary, and this is the root cause behind the choice for a single measure of exposure, rather than more predictive ones.

Table 3: Probability of detection of adverse events with low incidence. Summary data is reflect the occurrence of adverse events according to a Bernoulli random variable. For different incidence rates, value depicts its probability of occurrence given an experimental group size of n.

Risk of adverse events	n=4	n=8	n=10	n=16	n=20
0.10%	0.40%	0.80%	1.00%	1.59%	1.98%
1.00%	3.94%	7.73%	9.56%	14.85%	18.21%
5.00%	18.55%	33.66%	40.13%	55.99%	64.15%
10.00%	34.39%	56.95%	65.13%	81.47%	87.84%

As indicated previously, empirical data analysis does not provide uncertainty estimates to properly account for Type I (false positive) and Type II (false negative). In addition, experimental findings are evaluated in an experiment by experiment basis. This leads to misrepresentation of the estimated population characteristics, which imposes the need for conservative safety factors to account for bias and uncertainty. An immediate consequence of this is illustrated by safety levels identified for tolcapone (36), cerevastin (37), and ximelagatran (38), which were deemed “well-tolerated” at the predefined dose levels, but were later shown to be unsafe (39). It should therefore become clear that the use of the term tolerability ignores the high incidence of false negative results in standard designs.

Based on empirical methods, the absence of adverse events within the experimental group implies that risk is not present all.

Another problem is the fixed design used for in the estimation of the safety thresholds, which relies on a set of arbitrary selection of the dose levels. Consequently, the NOAEL is limited to one of the experimental dose levels. This results in the dose selection having a heavy influence on the precision and accuracy of the NOAEL estimate. Unfortunately, attempts to overcome the uncertainty and bias in the results may prove ineffective even if the number of animal is increased per group. In addition to the dose selection, the duration of the experiments also requires careful consideration and must be factored accordingly into the estimation of safety thresholds. Current approaches consider treatment duration as a constant factor, irrespective of the nature of the underlying adverse event. In general, high doses administered over shorter periods of time are deemed comparable to therapeutic doses administered chronically. This has little pharmacological foundation where time-to-onset may bear little relationship to dose (e.g. neutropenia). At high doses, effects may merely be due to secondary pharmacology. On the other hand, certain effects that can occur at therapeutic levels may be overlooked at higher exposures. Furthermore, if toxicity is delayed, then the likelihood of false negatives will increase if recording of adverse events stops at the end of dosing. A historical example is the case of methylmercury-induced dendritic degeneration in cats (40). Daily dosing for two months results in no differences from control groups, up to month five, when a significant difference becomes evident. If observations had ceased at month two, this effect would have been missed. In brief, the experimental limitations of current approaches can be summarised not only in terms of imprecision and inaccuracy, but also in terms of the lack of integration of the information contained within and between experiments.

From a statistical perspective, the occurrence of an adverse event can be viewed as a multidimensional random process over time, with one dimension for each type of adverse event. In practice, these dimensionalities are reduced to binary processes, leading to loss of information for data arising from continuous processes. Data loss also occurs when all these

binary processes are combined and reduced to a single binary number for each individual: *the animal either had an adverse event or it didn't*. Information about which adverse event occurred, the time-to-onset, duration, frequency and severity is all lost. On top this a further reduction happens at group level whereby the binary numbers for each individual are combined and reduced to a single binary number: *an adverse event occurs at a given dosing level, or it does not*. This approach prevents the use of quantitative methods, as it removes the evidence arising from the number of animals which exhibited adverse events. Data are further reduced by the very definition of NOAEL, which requires only the lowest dose to be considered in the estimation of the NOAEL: *the NOAEL is highest treatment level exhibiting no adverse events*. As a consequence of all the aforementioned steps, important information about the relationship between dose and exposure and adverse events may be lost.

By contrast, an approach which involves longitudinal statistical modelling of continuous and categorical data has the potential use all information in the production of estimates without any loss in information. However, an alternative to the NOAEL, the benchmark dose (BMD) approach has been proposed (41), which permits better use of experimental data. The BMD yields evidence about the entire dose-response curve, rather than a single point. Typically there are also large reductions at an individual and group level, but on a smaller scale. Relevant data across experimental groups are not collated and analysed together (42).

2.7. Additional flaws in the empirical evaluation of safety and toxicity

From a scientific and clinical point of view, one of the main disadvantages of empirical approaches is that extrapolation beyond experimental setting is often unreliable. Paradoxically, the ability to extrapolate or make inferences is central for the evaluation of safety and toxicity. Nonclinical data are generated with the primary objective of data extrapolation in mind.

Another limitation which cannot be easily circumvented is the inability to parameterise risk in a systematic manner, accounting for what is observed and what can be inferred from an intervention, irrespectively of the experimental evidence. Consequently, for instance, one fails to assess the implications of an adverse event arising from two different mechanisms of actions. To make accurate extrapolations, any relevant differences in the mechanism of action must be incorporated into the analysis and interpretation of the data. A similar problem arises in the case of nonlinear kinetics, when extrapolation to dose ranges outside of experimental ranges can lead to very different exposure levels, as compared to those expected from linear kinetics. Hence, it is evident that extrapolations derived from safety factors are doomed to remain inaccurate without further understanding of the mechanisms underlying the overt symptoms and signs.

Lastly, it is important to bear in mind that empirical methods often do not lend themselves well to integrating data and combining results from multiple experiments. This situation forces one to rely on clinical judgment to decide which findings can be deemed relevant. This inflexibility represents another inherent weakness of current approaches for the evaluation of safety, which clashes with one of the primary objectives of the drug development process, i.e., to reduce uncertainty about the safety and efficacy of a compound (43). In theory, more information should to lead to improve precision rather bias.

2.8. Safety threshold vs. risk or hazard surface

Currently, the use of fixed thresholds as a metric of safety ignores the variable nature of continuous processes and potentially prevents accurate interpretation of the underlying phenomena. For example, gastric ulceration is dependent on membrane permeability. Interindividual differences in tissue permeability are perceived as interindividual differences in sensitivity to drug effects, i.e., in the exposure which is required to reach a threshold. Based on current practice, the factor driving such differences often remains obscure. More

sophisticated approaches have been proposed to incorporate toxicodynamic differences through use of a sensitivity parameter (44). However, this suffers from the same weakness as the use of a threshold. Furthermore, thresholds offer no mechanistic basis for extrapolation across species. For example, there is no way to account for interspecies differences in membrane permeability. As such, interspecies differences can only be handled by safety factors.

Another immediate difficulty is the lack of consensus on what is defined as adverse events and how definitions vary across species. These definitions lead to different safety levels, meaning that safety thresholds are sensitive to definitions of events as adverse or non-adverse rather than the risk associated with them. Therefore, it should be noted that even with agreed definitions, the relevance of a threshold for the assessment of risk is questionable since it mostly relates only to the presence of an adverse event, rather than its severity. In this context, the shape and slope of the exposure-risk relationship is an important consideration. Yet, the use of thresholds incurs the danger that risk is treated and thought of as a binary endpoint. Since the only way to truly eliminate risk is to cease the hazard-causing activity, this is at odds with the binary treatment of it. Safety thresholds can also obfuscate more complicate U-shaped or bell-shaped relationships which may be relevant characteristics for consideration in a risk-benefit analysis.

In summary, it should be clear that despite the dichotomous nature of thresholds, all (patho)physiological processes underlying an adverse event are continuous processes. In fact, increasing understanding of the mechanisms underlying drug-target interactions (e.g. receptor pharmacology theory) as well as the identification of downstream pathways (i.e., factors determining post-receptor events) imposes revisiting the utility and relevance of thresholds as basis for the evaluation of drug response, irrespective of whether it involves efficacy or safety. The continuous nature of ligand-target relationships, based upon which target exposure must approach a certain order or magnitude in order to block or transducer a signal, offers the possibility of exploring signal using multidimensional response surfaces, rather than thresholds.

2.9. Translational toxicology: allometric scaling

All the undertaking required to implementing experimental protocols in safety pharmacology and toxicity implies the validity of a set of assumptions regarding the correlation between findings in animals and humans. Unfortunately, these assumptions do not take into account the prerequisite of construct validity to ensure direct comparability of the findings across species.

As indicated previously, uncertainty about differences between species and lack of understanding about the relevance of certain effects in humans, have lead to the introduction of safety factors the estimation of safety thresholds. Whilst many supporters of the approach envisage this as a plausible, cautionary measure, it cannot be ignored that in many cases over-conservatism will prevent the development of compounds that otherwise could be innocuous in humans. The challenge is therefore to identify a mechanistic basis for translating nonclinical safety findings or at least making inferences about drug action based on the results in a different species or experimental system (e.g., in vitro or cell culture). Five different dimensions need to be considered for that purpose: 1) differences in pharmacokinetics (i.e., accounting for physiological processes determining drug absorption, distribution, metabolism and elimination); 2) differences in pharmacodynamics (i.e., accounting for variation or differences in receptor engagement, activation and downstream amplification of the biosignal); 3) differences in homeostasis (i.e., accounting for functional capacity and feedback mechanisms which may compensate for drug-induced changes in physiological processes); 4) differences in response during health vs. disease conditions and 5) differences due to drug delivery properties.

It can be anticipated that accurate assessment of causality is essential for making inferences from one species to another. Furthermore, it is rather evident that in most cases all five dimensions need to be factored in the interpretation of nonclinical findings. However, currently, more focus is given to differences in pharmacokinetics more than any other aspect. As a matter of fact, extrapolation of findings between species often relies on the use of allometric scaling principles (45,46). Allometry requires assumptions about the

relationships between physiological function (e.g., metabolic capacity) and body size. In principle, this concept can also be applied to differences in pharmacodynamics (46,47), but the use of this technique in drug development is usually restricted to pharmacokinetic parameters, and more specifically to volume of distribution and clearance.

Despite its wide use in drug development, one needs to be aware of the limitations allometric methods represent to the prediction of pharmacokinetics and pharmacodynamics in humans. The first point relates to the unawareness of the underlying differences between-species. For example, total clearance can result from multiple routes; metabolism by oxidation and glucuronidation, biliary excretion, and/or renal excretion. The use of allometry assumes that when multiple physiological processes are involved, processes are scaled solely based on size differences and processes that do not scale well are considered clinically irrelevant (48). Biliary excretion is known not to scale well due to the role of ABC transporters expression levels. As such the decision to use scaling is dependent on an overall judgement of its ability to be scaled. For volume of distribution, the assumption is that distribution of drug outside system circulation occurs primarily due to passive diffusion; active transport is not accounted for either. Scaling via the more realistic physiologically based pharmacokinetic (PBPK) models (49), has been shown to account for both size-dependent and size-independent differences.

The second source of error in allometric scaling relate to the use of allometry as a monolithic extrapolation strategy: allometric relationships, even if correct, only relate to size differences between species. It is functionally equivalent to assuming that a human is a large rodent or another non-clinical species. Furthermore, the scaling of parameters assumes that size-related factors influencing systemic exposure are the only important covariate relationships governing drug effects.

Despite the clear flaw in this approach, the evaluation of alternative methods for scaling or translating pharmacokinetics and PKPD relationships remains limited. In fact, size-independent differences compose a much larger part of the differences in

pharmacodynamics and this is not accounted for with allometry. Paradoxically, there is also support for the view that size-independent differences are usually small given that adverse events in humans are predictable in the majority of cases (75%) from information obtained from preclinical experiments (50). This leads to the apparent conclusion that mechanisms of action in animals are similar to humans, however potentially serious differences may exist (51). A related problem is that clinical outcome is dependent on the underlying disease process, which may be different between species. Differences in baseline (physiological) response and in variability due to disease conditions in humans can confound the measurement of drug-induced effects, as compared to animals. Likewise, differences in target distribution can also complicate the interpretation and translation of non-clinical findings. For example, anaphylaxis is observed in the intestine and liver of rats, but in humans these symptoms are primarily observed in the lungs and blood vessels (52). The translational gap becomes even larger if one considers psychiatric or other neurological adverse events, which may not be detected in animals.

2.10. Translation of Risk

Translation of the risk associated with the experimental evidence observed in animals is the ultimate step triggering decisions related to nonclinical safety and toxicity of a novel molecule. Thus far, expert judgment is used by decision makers, which ultimately consists in the use of qualitative criteria for the assessment of risk. These criteria informally include some measure of overall uncertainty, but such an approach makes it difficult to understand the propagation of uncertainty. For instance, to infer that small physiological changes to the binding levels across species can lead to large changes the estimates of safe exposure. Clearly, accurate judgment is even more difficult when dependent on parameters for which uncertainty is unknown or not quantifiable.

Whilst the aforementioned issues have been recognised as important, regulators remain reluctant about the use of quantitative methods for risk assessment (53). There are various

reasons why qualitative risk assessment has been advocated over quantitative methods. However, many of the argued limitations do not necessarily apply when more modern statistical techniques are considered. We will address some of these points later in the next section, where model-based approaches are discussed.

The danger with a qualitative analysis is that the extent of any overall benefit will be left to human intuition. Informed decisions involve taking both benefits and risks of the drug into account. Yet, the consideration of risks and benefits based on safety thresholds is dependent on the nature of the risks in question and as such do not account for the underlying mechanisms, which in turn could be used for subsequent clinical interpretation. An encompassing inferential method is needed which accounts for underlying mechanisms and balance them against benefits. Most importantly, decision making regarding risk should include the contribution of historical data in a statistically and clinically formal manner.

3. Non-linear mixed effects modelling

The use of model-based methods has the ability to address many of the aforementioned criticisms pertinent to the design and analysis of safety pharmacology and toxicology protocols. Nonlinear-mixed effects models are a particular class of models that allow one to handle a variety of parameterisations by integrating stochastic and deterministic components of a problem. Although such models are often referred to as population models, they provide insight at the individual level, separating real variability from estimation uncertainty. They contain the necessary complexities required to assess risk in a manner that translates into scientifically rigorous decisions. In pharmacokinetic-pharmacodynamic data analysis, the use of a parametric approach based on nonlinear mixed-effects models provides a tool for handling repeated-measurement data in which the relationship between the explanatory variable and the response variable can be described by a single function, allowing model parameters to differ between subjects (54). An

immediate advantage of the approach is that the within-subject variability for a given individual can be distinguished from the differences between subjects even in the absence of balanced or frequent sampling of the data.

In hierarchical modelling, the term "mixed" refers to the use of both fixed effects (characterising the typical individual in the population) and random effects (describing the parameter distribution). The latter are divided into two levels: the difference between the individual prediction and the observation (residual error) and the variability between subjects (BSV). There may also be circumstances in which individual parameters vary longitudinally between occasions, randomly or due to some unknown physiological process. In such cases, a third level of variability can be introduced, i.e. the inter-occasion variability (IOV).

The general structure of a hierarchical model is as follows:

$$y_{ijk} = f\left(X_{ijk}, P_{ik}\right) + \varepsilon_{ijk},\ \ \varepsilon_{ijk} \sim N(0, \sigma^2) \qquad \text{Eq. 1}$$

where y_{ijk} is the j^{th} observation at occasion k in individual i. $f(\ )$ is typically a nonlinear function of individual parameter P_{ik} and independent variables X_{ijk}. In PKPD modelling, $f(\ )$ is usually then individual prediction of the observation. Independent variables are usually time, dose or drug exposure and demographic covariates. The ε_{ijk} forms the residual variability with variance σ^2. When the variance is independent of $f\left(X_{ijk}, P_{ik}\right)$, the model is said to have additive variability. On the other hand, when σ is proportional to $f(\ )$, we have a proportional error model (55).

For the i_{th} individual, the individual parameters P_{ik} can by the expression:

$$P_{ik} = \theta \cdot e^{\eta_i},\ \ \eta_i \sim N(0, \omega^2) \quad \text{Eq. 2}$$

This describes a log-normal variation of the individual parameter P, which has a typical value, Θ. The η_i and k_i are the random effects describing the differences between the typical

(population) value and the individual parameter value. η_i is assumed to be normally distributed with mean zero and variance ω^2.

Among other applications, the use of hierarchical models is justified and appropriate when the data available per individual are sparse. In addition, it is recognised as the most effective method to perform meta-analysis of data arising from different studies and to incorporate prior knowledge to the estimation of model parameters. It allows one to adjust for different variances (e.g. presence of influential factor in a given subgroup in the population) and to explore confounding correlations, when the design of the study correlates with the outcome (e.g., effect of weight vs. sex).

3.1. Estimation methods

The field statistical modelling field has developed well-established parameter estimation methods which provide the means not only to estimate the most likely value of the parameters given the data, but also to quantify uncertainty and correlation in estimated parameters and model (mis)specification. This ultimately provides us the opportunity to account for limited information and gaps in our knowledge. For example, if there is little information on the relationship between level of target occupancy and target activation, the corresponding parameters will have an appropriately high uncertainty. This feature is particularly relevant for the estimation and translation of risk as uncertainty can be propagated as high imprecision in exposure-risk relationships. Moreover, the calculation of the propagation of model uncertainty to uncertainty in the risk-benefit profile offers the prospect of efficient data collection.

The standard method for parameter estimation for nonlinear mixed effects models has been the maximum likelihood approach (56-59). This is where parameters are treated as random variables with distribution governed by the likelihood function $p(\mathbf{y}|\boldsymbol{\Theta})$, which represents the probability of the total data arising given the value of the parameters. The reported value for each parameter is the parameter at the maximum of the distribution, and associated

uncertainty given by the variance of the distribution. No data reduction is required; each raw data point directly informs parameter estimation thereby making maximal use of the available data. When multiple studies have been performed in populations which share common physiological processes or treatments, datasets may be aggregated to support integrated analyses across these studies. Furthermore, model-based analysis can handle multiple types of observations (e.g., pharmacokinetics and pharmacodynamics) as well as multiple data types (e.g., continuous and categorical).

Of particular interest for the assessment of safety and toxicity is the possibility of applying extensions of the maximum likelihood, which enable mathematically rigorous incorporation of prior parameter information (e.g., receptor occupancy or blood to plasma binding ratio in vitro to describe in vivo data). The two main methods for achieving this are the penalised likelihood method (49,60) and Bayesian estimation (61). It should also be noted that the advent of exact likelihood methods such as expectation maximisation (EM) methods (62) has provided increased reliability of PKPD analyses, especially in the presence of sparse data, often available from general toxicity protocols.

We should also emphasise that in the context of safety pharmacology and toxicity studies, trial optimisation represents proper adherence to the three R's (reduction, refinement and replacement). When prior information is available for class-specific parameters, a model-based analysis may benefit from this allowing for a reduction experimental cohort sizes or burden to animals. This is possible because model-based analyses are inferential in nature.

3.2. Model parameterisation: empirical vs. mechanistic models

Despite the increasing number of modelling examples in biomedical and pharmaceutical research, the use of pharmacokinetic and pharmacokinetic-pharmacodynamic models has remained primarily descriptive. However, the application of such models for the evaluation of safety requires further consideration of its biological plausibility and predictive or prognostic value. For example, instead of using a simple compartmental model to describe

the observed phases of drug elimination, one may need to consider a physiologically-based pharmacokinetic model (PBPK) (63). Such models can be developed by integrating prior *in vitro* data and literature information.

On the other hand, it is not unusual for components of PKPD models to be statistically correlated to some degree. Therefore, it is important that when identifying at-risk subpopulations based on collected data, covariate selection is guided by a mechanistic or physiological evaluation. There are several methods that allow such an approach (64-67). More recently, these methods have also been applied to describe disease processes (33). Statistically, these models include a response variable that characterises the disease status and its progression over time.

3.3. Simulations, experimental design and optimisation

Model predictions, simulated outside the experimental context are extrapolations subject to model specification bias. Since our primary goal is to show the relevance of such models to analyse data arising from pre-clinical species and eventually from healthy subjects to assess safety and toxicity in patients who will be receiving these drugs, this point is of special importance. In PKPD modelling, computer simulation involves using statistical models to predict the behaviour of the biological system described by the model (68). Clinical trial simulations (CTS) i.e. computer simulation of trials, allows for the investigation of the impact of different design characteristics on the outcome of a trial. It can also be used to investigate the implications of uncertainty and variability in pharmacokinetic and pharmacological processes for recruited individuals, thereby allowing the prior assessment of the robustness of the protocol to known uncertainty and variability (69). More generally, in a CTS it is possible to test the influence of any modelling assumption and design factor beforehand (Figure 6).

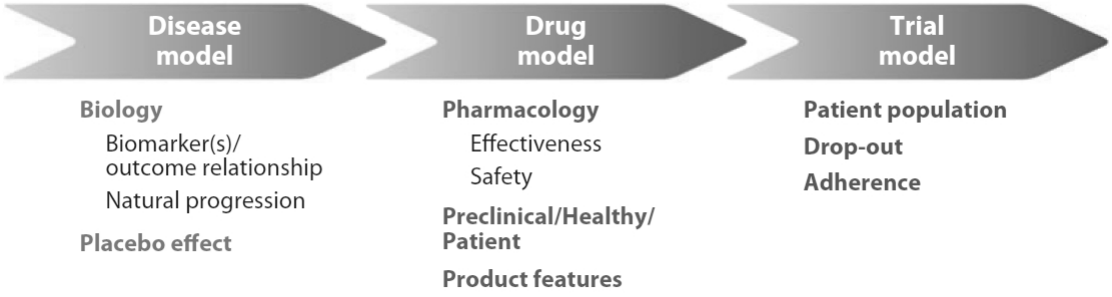

Figure 6: The diagram depicts the major components of a clinical trial simulation (CTS). In model-based drug development, CTS can be used to characterise the interactions between drug and disease, enabling among other things the assessment of disease-modifying effects, dose selection and covariate effects. In conjunction with a trial model, CTS allows the evaluation of such interactions, taking into account uncertainty and trial design factors, including the implications of different statistical methods for the analysis of the data.

Trial design can also benefit from the use of optimal design methodology. The goal of optimal design, specifically the procedure known as D-optimality, is to determine design variables (such as sampling times and dose selection) that optimise the expected information content (usually by maximising the determinant of the Fisher Information Matrix (FIM)) within the desired resource constraints. A variety of software programs exist purpose built for the estimation of PK/PD models (70). Optimal sampling schedules for toxicity experiments can help increase the precision by which drug specific parameters can be estimated and/or reduce the burden to animals by minimising the number of samples needed. This is desirable from an ethical and scientific perspective, as poor experimental design is known to result in biased estimates. Among other advantages, optimal sampling may facilitate the collection of biomarkers in conjunction with pharmacokinetic data when blood volume is limited.

Conclusions

High attrition rates due to a poor safety profile combined with inability to correctly identify risk demand revisiting of concepts and modernisation of the approaches currently used for the assessment of toxicity. Current practices fail to support decision making on multiple levels. Firstly, the parameterisation of drug exposure and available metrics of risk are often justified by historical precedent rather than by an informed scientific rationale. These measures are assumed to be predictive of drug effects in humans, despite the fact that in many cases known pharmacokinetic and pharmacological drug properties contradict such assumptions. Evidence clearly shows that empirical protocols remain primarily descriptive rather than explanatory of the observed phenomena and are therefore unsuitable for extrapolation, an important point to consider when analysing and interpreting safety pharmacology and toxicology data. Moreover, statistically, the use of point estimates and thresholds prevents understanding of the consequences of between subject variability and identification of at-risk subpopulations. Additionally, type I and II errors are also not accounted for in the design or analysis of toxicity data, both of which are critical informed decision making.

In summary, our review has highlighted the implications of empirical data generation for the evaluation of safety and toxicity during drug development. A shift in paradigm was proposed to ensure that pharmacological concepts are incorporated into the evaluation of safety and toxicity. Moreover, we indicate the urgent need to integrate historical evidence, so that findings across species can be effectively translated. Based on historical examples, we have shown some important challenges for the early characterisation of the safety profile of a new molecule and discuss how model-based methodologies can be applied for better design and analysis of experimental protocols. From a methodological perspective, nonlinear-mixed effects modelling is recommended as a tool to account for such requirements. Its use in the evaluation of pharmacokinetics (PK) and pharmacokinetic-pharmacodynamic relationships (PKPD) has enabled the advance of quantitative approaches

in pharmacological research in recent decades. Comparable benefits can be anticipated for the assessment of safety and toxicity.

Reference List

1. Kola, I., and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? Nat.Rev.Drug Discov. **3**, 711-715.

2. U.S.Food and Drug Administration (CDER). Innovation or stagnation? Challenge and opportunity on the critical path to new medical products. 2004.
URL:
http://www.fda.gov/scienceresearch/specialtopics/criticalpathinitiative/criticalpathopportunitiesreports/ucm077262.htm

3. Olson, H., Betton, G., Robinson, D., Thomas, K., Monro, A., Kolaja, G., Lilly, P., Sanders, J., Sipes, G., Bracken, W., Dorato, M., Van Deun, K., Smith, P., Berger, B., and Heller, A. (2000). Concordance of the toxicity of pharmaceuticals in humans and in animals. Regul.Toxicol.Pharmacol. **32**, 56-67.

4. Russell, W. M. S. and BURCH, R. L. The principles of humane experimental technique. 1-4-1959. London: Methuen & Co. Ltd.

5. Suntharalingam, G., Perry, M. R., Ward, S., Brett, S. J., Castello-Cortes, A., Brunner, M. D., and Panoskaltsis, N. (2006). Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. N.Engl.J.Med. **355**, 1018-1028.

6. Expert Scientific Group. Expert Scientific Group on Phase One Clinical Trials. 2006.

7. European Medicines Agency (CHMP). Guideline on strategies to identify and mitigate risks for first-in human clinical trials with investigational medicinal products. 2007.

URL:
http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002988.pdf

8. Kenter, M. J., and Cohen, A. F. (2006). Establishing risk of human experimentation with drugs: lessons from TGN1412. Lancet **368**, 1387-1391.

9. Dowsing, T., and Kendall, M. J. (2007). The Northwick Park tragedy--protecting healthy volunteers in future first-in-man trials. J.Clin.Pharm.Ther. **32**, 203-207.

10. Stebbings, R., Poole, S., and Thorpe, R. (2009). Safety of biologics, lessons learnt from TGN1412. Curr.Opin.Biotechnol. **20**, 673-677.

11. Breckenridge, A. (2010). Regulatory challenges, reimbursement, and risk-benefit assessment. Clin.Pharmacol.Ther. **88**, 153-154.

12. Krumholz, H. M., Ross, J. S., Presler, A. H., and Egilman, D. S. (2007). What have we learnt from Vioxx? BMJ **334**, 120-123.

13. Bombardier, C., Laine, L., Reicin, A., Shapiro, D., Burgos-Vargas, R., Davis, B., Day, R., Ferraz, M. B., Hawkey, C. J., Hochberg, M. C., Kvien, T. K., and Schnitzer, T. J. (2000). Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. N.Engl.J.Med. **343**, 1520-8, 2.

14. Lasser, K. E., Allen, P. D., Woolhandler, S. J., Himmelstein, D. U., Wolfe, S. M., and Bor, D. H. (2002). Timing of new black box warnings and withdrawals for prescription medications. JAMA **287**, 2215-2220.

15. Cross, J., Lee, H., Westelinck, A., Nelson, J., Grudzinskas, C., and Peck, C. (2002). Postmarketing drug dosage changes of 499 FDA-approved new molecular entities, 1980-1999. Pharmacoepidemiol.Drug Saf **11**, 439-446.

16. Heerdink, E. R., Urquhart, J., and Leufkens, H. G. (2002). Changes in prescribed drug doses after market introduction. Pharmacoepidemiol.Drug Saf **11**, 447-453.

17. Lowe, P. J., Hijazi, Y., Luttringer, O., Yin, H., Sarangapani, R., and Howard, D. (2007). On the anticipation of the human dose in first-in-man trials from preclinical and prior clinical information in early drug development. Xenobiotica **37**, 1331-1354.

18. Edler, L., Poirier, K., Dourson, M., Kleiner, J., Mileson, B., Nordmann, H., Renwick, A., Slob, W., Walton, K., and Wurtzen, G. (2002). Mathematical modelling and quantitative methods. Food Chem.Toxicol. **40**, 283-326.

19. Dybing, E., Doe, J., Groten, J., Kleiner, J., O'Brien, J., Renwick, A. G., Schlatter, J., Steinberg, P., Tritscher, A., Walker, R., and Younes, M. (2002). Hazard characterisation of chemicals in food and diet. dose response, mechanisms and extrapolation issues. Food Chem.Toxicol. **40**, 237-282.

20. Slob, W. (1999). Thresholds in toxicology and risk assessment. International Journal of Toxicology **18**, 345-367.

21. Kroes, R., Kleiner, J., and Renwick, A. (2005). The threshold of toxicological concern concept in risk assessment. Toxicol.Sci. **86**, 226-230.

22. Dorato, M. A., and Engelhardt, J. A. (2005). The no-observed-adverse-effect-level in drug safety evaluations: use, issues, and definition(s). Regul.Toxicol.Pharmacol. **42**, 265-274.

23. U.S.Food and Drug Administration (CDER). Estimating the Maximum Safe Starting Dose in Initial Clinical Trials for Therapeutics in Adult Healthy Volunteers. 2005.
URL: http://www.fda.gov/downloads/Drugs/Guidance/UCM078932.pdf

24. Agoram, B. M. (2009). Use of pharmacokinetic/ pharmacodynamic modelling for starting dose selection in first-in-human trials of high-risk biologics. Br.J.Clin.Pharmacol. **67**, 153-160.

25. Muller, P. Y., Milton, M., Lloyd, P., Sims, J., and Brennan, F. R. (2009). The minimum anticipated biological effect level (MABEL) for selection of first human dose in clinical trials with monoclonal antibodies. Curr.Opin.Biotechnol. **20**, 722-729.

26. Dourson, M. L., and Stara, J. F. (1983). Regulatory history and experimental support of uncertainty (safety) factors. Regul.Toxicol.Pharmacol. **3**, 224-238.

27. Swartout, J. C., Price, P. S., Dourson, M. L., Carlson-Lynch, H. L., and Keenan, R. E. (1998). A probabilistic framework for the reference dose (probabilistic RfD). Risk Anal. **18**, 271-282.

28. Gaylor, D. W., and Kodell, R. L. (2000). Percentiles of the product of uncertainty factors for establishing probabilistic reference doses. Risk Anal. **20**, 245-250.

29. Shih, C., Gosset, L., and Gates, S. e. a. (1996). LY231514 and its polyglutamates exhibit potent inhibition against both human dihydrofolate reductase (DHFR) and thymidylate synthase (TS): multiple folate enzyme inhibition. Annals of Oncology.

30. Latz, J. E., Rusthoven, J. J., Karlsson, M. O., Ghosh, A., and Johnson, R. D. (2006). Clinical application of a semimechanistic-physiologic population PK/PD model for neutropenia following pemetrexed therapy. Cancer Chemother.Pharmacol. **57**, 427-435.

31. Jeste, D. V., Lacro, J. P., Palmer, B., Rockwell, E., Harris, M. J., and Caligiuri, M. P. (1999). Incidence of tardive dyskinesia in early stages of low-dose treatment with typical neuroleptics in older patients. Am.J.Psychiatry **156**, 309-311.

32. Danhof, M., de Jongh, J., de Lange, E. C., Della, P. O., Ploeger, B. A., and Voskuyl, R. A. (2007). Mechanism-based pharmacokinetic-pharmacodynamic modeling: biophase distribution, receptor theory, and dynamical systems analysis. Annu.Rev.Pharmacol.Toxicol. **47**, 357-400.

33. Danhof, M., de Lange, E. C., Della Pasqua, O. E., Ploeger, B. A., and Voskuyl, R. A. (2008). Mechanism-based pharmacokinetic-pharmacodynamic (PK-PD) modeling in translational drug research. Trends Pharmacol.Sci. **29**, 186-191.

34. Danhof, M., Alvan, G., Dahl, S. G., Kuhlmann, J., and Paintaud, G. (2005). Mechanism-based pharmacokinetic-pharmacodynamic modeling-a new classification of biomarkers. Pharm.Res. **22**, 1432-1437.

35. Rolan, P., Danhof, M., Stanski, D., and Peck, C. (2007). Current issues relating to drug safety especially with regard to the use of biomarkers: a meeting report and progress update. Eur.J.Pharm.Sci. **30**, 107-112.

36. Jorga, K. M., Fotteler, B., Heizmann, P., and Zurcher, G. (1998). Pharmacokinetics and pharmacodynamics after oral and intravenous administration of tolcapone, a novel adjunct to Parkinson's disease therapy. Eur.J.Clin.Pharmacol. **54**, 443-447.

37. Bakri, R., Wang, J., Wierzbicki, A. S., and Goldsmith, D. (2003). Cerivastatin monotherapy-induced muscle weakness, rhabdomyolysis and acute renal failure. Int.J.Cardiol. **91**, 107-109.

38. Wernevik, L. C., Nystrom, P., Johnsson, G., Nakanishi, T., and Eriksson, U. G. (2006). Pharmacokinetics and pharmacodynamics of the oral direct thrombin inhibitor ximelagatran in young healthy Japanese men. Clin.Pharmacokinet. **45**, 77-84.

39. Cohen, A. (2007). Should we tolerate tolerability as an objective in early drug development? Br.J.Clin.Pharmacol. **64**, 249-252.

40. Oliveira, R. B., Gomes-Leal, W., do-Nascimento, J. L., and Picanco-Diniz, C. W. (1998). Methylmercury intoxication and histochemical demonstration of NADPH-diaphorase activity in the striate cortex of adult cats. Braz.J.Med.Biol.Res. **31**, 1157-1161.

41. Barnes, D. G., Daston, G. P., Evans, J. S., Jarabek, A. M., Kavlock, R. J., Kimmel, C. A., Park, C., and Spitzer, H. L. (1995). Benchmark Dose Workshop: criteria for use of a benchmark dose to estimate a reference dose. Regul.Toxicol.Pharmacol. **21**, 296-306.

42. Budtz-Jorgensen, E., Grandjean, P., Keiding, N., White, R. F., and Weihe, P. (2000). Benchmark dose calculations of methylmercury-associated neurobehavioural deficits. Toxicol.Lett. **112-113**, 193-199.

43. Lalonde, R. L., Kowalski, K. G., Hutmacher, M. M., Ewy, W., Nichols, D. J., Milligan, P. A., Corrigan, B. W., Lockwood, P. A., Marshall, S. A., Benincosa, L. J., Tensfeldt, T. G., Parivar, K., Amantea, M., Glue, P., Koide, H., and Miller, R. (2007). Model-based drug development. Clin.Pharmacol.Ther. **82**, 21-32.

44. Kodell, R. L., and Chen, J. J. (2007). On the use of hierarchical probabilistic models for characterizing and managing uncertainty in risk/safety assessment. Risk Anal. **27**, 433-437.

45. West, G. B., Brown, J. H., and Enquist, B. J. (1997). A general model for the origin of allometric scaling laws in biology. Science **276**, 122-126.

46. Zuideveld, K. P., Van der Graaf, P. H., Peletier, L. A., and Danhof, M. (2007). Allometric scaling of pharmacodynamic responses: application to 5-Ht1A receptor mediated responses from rat to man. Pharm.Res. **24**, 2031-2039.

47. Mager, D. E., Woo, S., and Jusko, W. J. (2009). Scaling pharmacodynamics from in vitro and preclinical animal studies to humans. Drug Metab Pharmacokinet. **24**, 16-24.

48. Bachmann, K., Pardoe, D., and White, D. (1996). Scaling basic toxicokinetic parameters from rat to man. Environ.Health Perspect. **104**, 400-407.

49. Langdon, G., Gueorguieva, I., Aarons, L., and Karlsson, M. (2007). Linking preclinical and clinical whole-body physiologically based pharmacokinetic models with prior distributions in NONMEM. Eur.J.Clin.Pharmacol. **63**, 485-498.

50. Lazarou, J., Pomeranz, B. H., and Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA **279**, 1200-1205.

51. Ju, C., and Uetrecht, J. P. (2002). Mechanism of idiosyncratic drug reactions: reactive metabolite formation, protein binding and the regulation of the immune system. Curr.Drug Metab **3**, 367-377.

52. Haley, P. J. (2003). Species differences in the structure and function of the immune system. Toxicology **188**, 49-71.

53. European Medicines Agency (CHMP). Report of the CHMP Working Group on Benefit-Risk Assessment Models and Methods. 19-1-2007.
URL:
http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2010/01/WC500069668.pdf

54. Davidian, M., and Giltinan, D. M. (1993). Analysis of repeated measurement data using nonlinear mixed effects models. Chemom.Intell.LAb.Syst. **20**, 1-24.

55. Karlsson, M. O., Beal, S. L., and Sheiner, L. B. (1995). Three new residual error models for population PK/PD analyses. J.Pharmacokinet.Biopharm. **23**, 651-672.

56. Beal, S. L. (1984). Population pharmacokinetic data and parameter estimation based on their first two statistical moments. Drug Metab Rev. **15**, 173-193.

57. Lindstrom, M. L., and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. Biometrics **46**, 673-687.

58. Pinheiro, J. C., and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. J.Comput.Graph.Stat. **4**, 12-35.

59. Baverel, P. G., Savic, R. M., Wilkins, J. J., and Karlsson, M. O. (2009). Evaluation of the nonparametric estimation method in NONMEM VI: application to real data. J Pharmacokinet.Pharmacodyn. **36**, 297-315.

60. Gisleskog, P. O., Karlsson, M. O., and Beal, S. L. (2002). Use of prior information to stabilize a population data analysis. J.Pharmacokinet.Pharmacodyn. **29**, 473-505.

61. Lunn, D. J., Best, N., Thomas, A., Wakefield, J., and Spiegelhalter, D. (2002). Bayesian analysis of population PK/PD models: general concepts and software. J.Pharmacokinet.Pharmacodyn. **29**, 271-307.

62. Kuhn E., and Lavielle M. (2004). Maximum likelihood estimation in nonlinear mixed effects models. Computational Statistics & Data Analysis **49**, 1020-1038.

63. Bois, F. Y. (2010). Physiologically based modelling and prediction of drug interactions. Basic Clin.Pharmacol.Toxicol. **106**, 154-161.

64. Lunn, D. J. (2008). Automated covariate selection and Bayesian model averaging in population PK/PD models. J.Pharmacokinet.Pharmacodyn. **35**, 85-100.

65. Wade, J. R., Beal, S. L., and Sambol, N. C. (1994). Interaction between structural, statistical, and covariate models in population pharmacokinetic analysis. J Pharmacokinet.Biopharm. **22**, 165-177.

66. Mentre, F., and Mallet, A. (1994). Handling covariates in population pharmacokinetics. Int.J Biomed.Comput **36**, 25-33.

67. Mandema, J. W., Verotta, D., and Sheiner, L. B. (1992). Building population pharmacokinetic--pharmacodynamic models. I. Models for covariate effects. J Pharmacokinet.Biopharm. **20**, 511-528.

68. Holford, N. H., Kimko, H. C., Monteleone, J. P., and Peck, C. C. (2000). Simulation of clinical trials. Annu.Rev.Pharmacol.Toxicol. **40**, 209-234.

69. Holford, N., Ma, S. C., and Ploeger, B. A. (2010). Clinical trial simulation: a review. Clin.Pharmacol.Ther. **88**, 166-182.

70. Mentre, F., Duffull, S., Gueorguieva, I., Hooker, A., Leonov, S., Ogungbenro, K., and

Retout, S. Software for optimal design in population pharmacokinetics and

pharmacodynamics: a comparison. PAGE. 2007.