Research Article

Associations Between Clinical Signs and Pathological Findings in Toxicity Testing

Antero V. Silva[1], Ulf Norinder[2], Elin Liiv[3], Björn Platzack[4], Mattias Öberg[1#] and Elin Törnqvist[1#]

[1]Unit of Integrative Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; [2]Department of Computer and Systems Sciences, Stockholm University, Kista, Sweden; [3]Department of Pharmaceutical Biosciences, Drug Safety and Toxicology, Faculty of Pharmacy, Uppsala University, Uppsala, Sweden; [4]Chemical and Pharmaceutical Safety Unit, Research Institutes of Sweden (RISE), Södertälje, Sweden

Abstract

Animal testing for toxicity assessment of chemicals and pharmaceuticals must take the 3R principles into consideration. During toxicity testing *in vivo*, clinical signs are used to monitor animal welfare and to inform about potential toxicity. This study investigated possible associations between clinical signs, body weight change and histopathological findings observed after necropsy. We hypothesized that clinical signs and body weight loss observed during experiments could be used as early markers of organ toxicity. This represents a potential for refinement in terms of improved study management and decreasing of pain and distress experienced during animal experiments. Data from three sequential toxicity studies of an anti-cancer drug candidate in rats were analyzed using the multivariate partial least squares (PLS) regression method. Associations with a predictive value over 80% were found between the occurrence of mild to severe clinical signs and histopathological findings in the thymus, testes, epididymides and bone marrow. Piloerection, eyes half shut and slightly decreased motor activity were most strongly associated with the pathological findings. A 5% body weight loss was found to be a strong empirical predictor of pathological findings but could also be predicted accurately by clinical signs. Thus, we suggest using mild clinical signs and a 5% body weight loss as toxicity markers and as a non-invasive surveillance tool to monitor research animal welfare in toxicity testing. These clinical signs may also enable reduction of animal use due to their informative potential to support scientific decisions regarding drug candidate selection, dose setting, study design, and toxicity assessment.

1 Introduction

Every year, approximately 10 million laboratory animals are used for scientific purposes in the European Union (EU) only (EC, 2019). In the United States (US), the Animal Welfare Act (USDA, 1966) excludes rats and mice, but the Humane Society of the US has estimated that 25 million vertebrate animals are used annually for research purposes. The majority of research animals are mice and rats. In the EU, most laboratory animals are used for the study of human and animal diseases, and about 2 million animals are used annually for regulatory testing required for the marketing of chemicals and pharmaceutical substances (EC, 2019).

In the EU, risk assessment of chemicals follows the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) legislation (EC Regulation No 1907/2006, see EC, 2006), which requires animal testing to be performed if no suitable alternatives are available. In the pharmaceutical industry, *in vivo* safety assessment studies are performed during the non-clinical phase of the pharmaceutical development process (EMA, 2010; ICH, 2009). Most candidate drugs are thus tested *in vivo* at doses high enough to identify adverse effects and dose-response relationships (Hornberg et al., 2014; Sewell et al., 2014; Sparrow et al., 2011). Assessment of toxicity during animal studies is used to predict safe exposure levels in humans, balancing the risk-benefit of the chemical exposure to support the decision-making process. This principle is also applicable to non-pharmaceutical chemical testing (Olson et al., 2000).

Numerous regulations and guidelines are to be considered for toxicity testing using animal models, for example the Organisation for Economic Co-operation and Development's (OECD)

[#] contributed equally

Received March 31, 2020; Accepted October 26, 2020;
Epub October 26, 2020; © The Authors, 2021.

ALTEX 38(2), 198-214. doi:10.14573/altex.2003311

Correspondence: Antero Silva
Institute for Environmental Medicine
Karolinska Institutet, Box 210
171 77 Stockholm, Sweden
(antero.silva@ki.se)

Guidelines for the Testing of Chemicals, Section 4 and the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) *Safety Guidelines*. Most *in vivo* studies follow strict scientific, ethical and regulatory requirements, providing valuable information regarding the safety of the candidate drugs, their potential side effects, and their mechanism(s) of action (EMA, 2010; EU, 2010; ICH, 2009; NC3Rs, 2009). The scope of these legislations is to ensure high-quality data that are reproducible and comparable between studies, as well as to protect laboratory animals from unnecessary suffering while not compromising the informative value of the study.

Since 2013, all use of animals for scientific purposes in the EU has to be performed in accordance with Directive 2010/63/EU, which emphasizes the 3Rs – replacement, reduction and refinement (EU, 2010). Indeed, certain animal models used for toxicity testing have been replaced with cell or computer-based methods within the area of risk assessment, for example the *in vitro* ARE-Nrf2 luciferase test method (OECD, 2018) and the *in chemico* direct peptide reactivity assay (DPRA) (OECD, 2019) for skin sensitization evaluation purposes. Work towards reduction and refinement includes improvement of, e.g., project and study design, animal housing, and experimental procedures (Ringblom et al., 2017a; Kalantari et al., 2017; Zidar et al., 2019). Progress has been made in the area of toxicity testing, for example, in terms of the refined use of body weight loss assessment for decisions regarding the maximum tolerated dose (MTD) (Chapman et al., 2013), reduction of animal use by microsampling of low blood volumes (Jonsson et al., 2012) and by including fewer recovery animals (Sewell et al., 2014; Sparrow et al., 2011). Systematic 3R approaches reveal a major reduction potential for the use of animals in pharmaceutical toxicity testing (Törnqvist et al., 2014). However, such systematic 3R developments and implementations are rarely seen in academic research where animal models are not regulated by guidelines. Still, efforts to guide academic researchers and laboratory animal facilities are being made in the EU. One example is the European Commission's publication of a Severity Assessment Framework, which includes animal model descriptions and clinical and behavioral monitoring sheets aiming to improve animal welfare and reduce suffering (EC, 2012).

According to international guidelines and Good Laboratory Practice (GLP) for safe drug development, clinical signs should be thoroughly monitored, registered and reported in animal experiments (OECD, 2008; WHO, 2009). Clinical signs are used for the assessment of the animals' general condition and for implementing humane endpoints, defined as the point at which a research animal is pre-terminally sacrificed to avoid further suffering that is not justified by scientific benefit (Morton, 1997; OECD, 2000). In the pharmaceutical industry, clinical signs are also used for dose-setting and study design purposes (NC3Rs, 2009; Sewell et al., 2015). Although clinical signs are registered, they are rarely used as informative endpoints of toxicity for risk assessment purposes. For example, neither the operating procedures for setting acute exposure guideline levels nor subsequent reference doses mention clinical signs (NAC, 2001). The WHO states in the criteria document for risk assessment of chemicals in food that "*other findings*", such as clinical signs and changes in body weight, may suggest a need to establish an acute reference dose (WHO, 2009). An example of this is that the Joint FAO/WHO Expert Committee on Food Additives (JECFA) reported "*clinical signs of toxicity [and], reduction in body weight*" as a basis to establish a NOAEL for a study with flavoring agents (JECFA, 2016). To the best of our knowledge, clinical signs are very rarely discussed or used as a critical endpoint to establish reference doses.

In the present study, we hypothesized that mild clinical signs and body weight loss observed in animal studies can be used as early markers of toxicity, i.e., pathological findings detected after necropsy. These associations do not reflect the underlying biological mechanisms, as clinical signs are not necessarily associated with a specific organ. However, it is possible to use clinical signs as a marker or potential key event in an adverse outcome pathway network, where the adverse outcome, in this case, would be organ pathology. To facilitate the future use of clinical signs as early markers of toxicity in areas other than pharmaceutical safety assessment, we created a shortlist of clinical signs and tested whether these would be similarly accurate in predicting pathological findings compared to using all registered clinical signs. Partial least squares regression (PLS), a multivariate data analysis method, was employed to analyze pre-existing data from three non-clinical and sequential *in vivo* toxicity studies in rats testing an anti-cancer candidate drug.

2 Animals, materials and methods

Animal studies

Data from animal studies are presented according to the ARRIVE (Animal Research: Reporting of *In Vivo* Experiments) guidelines (Kilkenny et al., 2010; NC3Rs, 2010). The analyzed data were reused from previously performed studies, thus avoiding any new animal experiments for the present investigation. Data were collected from three *in vivo* safety assessment studies performed at the Swedish Toxicology Sciences Research Center (Swetox), which were conducted following the *Guideline on repeated dose toxicity* (EMA, 2010) and the *Guidance for Industry – M3(R2) Nonclinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorization for Pharmaceuticals* (US FDA, 2010). All studies had been approved by the Southern Stockholm Ethical Committee for Research Animals (ethical permit number S7-15) and were performed according to Swedish animal welfare legislation L1 (SFS, 1988) and L150 (SJVFS, 2012 for study I; SJVFS, 2015 for study II and III) in the Swetox facilities. The tested compound was a proprietary anti-cancer candidate drug intended for human therapy through oral administration.

Two dose-range finding studies (study I and II) were performed in order to decide acceptable oral doses for a 28-day repeated dose toxicity study (study III). Study I consisted of a 7-day oral toxicity study, study II was a 13-day oral toxicity study, and study III was a 28-day oral toxicity study (Tab. 1). These studies were performed in rats in accordance with the regulatory guidelines (EU, 2001) to set doses for the first human trials.

Tab. 1: Dose groups, animal distribution, daily dose levels and dosing scheme

	Group	Animal number	Daily dose levels[a]
Study I	1 (Vehicle[b])	5 males + 5 females	0 mg/kg
	2	5 males + 5 females	60 mg/kg
	3	5 males + 5 females	180 mg/kg
	4	5 males + 5 females	360 mg/kg
Study II	1 (Vehicle[b])	2 males + 2 females	0 mg/kg
	2 M	3 males	30 mg/kg
	2 F	3 females	20 mg/kg
	3 M	3 males	90 mg/kg
	3 F	3 females	60 mg/kg
Study III	1 (Vehicle[b])	10 males + 10 females	0 mg/kg
	2 M	10 males	8.3 mg/kg
	2 F	10 females	3.3 mg/kg
	3 M	10 males	25 mg/kg
	3 F	10 females	10 mg/kg
	4 M	10 males	75 mg/kg
	4 F	10 females	30 mg/kg

[a] All doses and concentrations are expressed in terms of the tested substance. All doses were administered at dose volumes of 10 mL/kg according to the dosing scheme.
[b] The vehicle formulation used in Group 1 was the same %w/v as the vehicle used for group 4.

The rats were ordered based on strain (Wistar Hannover Galas, Charles River Laboratories, Denmark for study I; RccHan: Wistar, Harlan Laboratories, The Netherlands for studies II and III), weight and age. Upon arrival, the rats were approximately 10 weeks old, and 240-280 g and 170-200 g for males and females, respectively. All animals were thoroughly examined to ensure good condition and subsequently randomized into cages and dose groups. The used BK Rat Cage dimensions were 30 cm width x 42 cm length x 21 cm height. Water from the site drinking water supply and RM1 (P) SQC pelleted diet supplied by Special Diets Services Ltd., England were provided *ad libitum*. All cages were enriched with wood chip bedding, nesting material, a plastic tube (tunnel), a carton toilet paper roll and wooden sticks for gnawing.

The rats were habituated to handling and experimental procedures during the acclimatization period, which has been shown to greatly reduce stress during the experiment. Each rat was individually handled for 1-2 min daily on 5 days per week for 2-3 weeks before the study start and trained for dosing by administering sham doses of tap water by oral gavage during the last two weeks. During all 3 studies, the animals were group-housed, separated by gender and dose group, keeping the environment they were introduced to during the acclimatization period. The dosing vehicle solution for the test compound was 2-hydroxypropyl-β-cyclodextrin in acetate buffer (pH 4.5-4.8), a non-toxic and common vehicle solution for lipophilic drugs (Gould and Scott, 2005).

In study I, the rats were dosed for a total of 7 d (Tab. 1). Due to the declining general condition of some animals, reaching the predetermined humane endpoint (impaired food and feed intake, as well as poor general condition, all animals in dose groups 3 and 4 were pre-terminally sacrificed. The remaining animals were sacrificed, as planned, 24 h after the last dose. Animals scheduled for sacrifice were terminated by exsanguination through the carotid artery/vein under isoflurane and oxygen anesthesia.

In study II, the doses were adjusted based on toxicokinetic data from study I (Tab. 1). The animals were dosed for a total of 13 d. All animals were sacrificed 24 h after receiving the last dose.

In study III, the doses were adjusted based on toxicokinetic data from studies I and II (Tab. 1). The rats were dosed for a total of 28 d. All animals were sacrificed 24 h after receiving the last dose.

Independent variables

As independent variables, all available registered clinical observations, the toxicokinetic parameter C_{max}, body weight change and gender were used. Gender was included to detect any significant differences between the two sexes. The clinical signs were used as indicators of potential toxicity during the animal studies and to monitor animal welfare in general. Clinical signs were registered by trained animal technicians during all 3 animal studies. The technicians performing the dosing procedure were not blinded. The signs to be registered were decided before the study start, based on an in-house reference list with general and organ/physiology-related clinical signs of adverse outcomes (Tab. 2). Observations were registered daily, either immediately after the dosing procedure or in the morning of dose-free days. In addition, scheduled repeated observations, in which all animals were observed at five different timepoints during a 24-h period, were performed twice during all 3 studies.

For clinical signs and body weight variables, a binary scoring system was employed. The presence of clinical signs was scored with 1, absence with 0. Body weight measurements were performed regularly during all studies (Tab. 3). For detection of body weight changes, the arithmetic mean of the control group was determined for each weighing. Each individual rat's weight was then compared to the average of the latest weighing of the control group, for the respective gender. If the difference was greater than 5%, i.e., if the animal did not display a body weight ≥ 95% of the control group's average, a value of 1 was set. This value indicates an abnormal body weight loss compared to the control group, regardless of when it occurred. Animals with a body weight ≥ 95% of the average of the control group were scored with a 0. A body weight loss of 5% was chosen as an endpoint after testing 3, 5 and 10%, and was tested both as an independent as well as a dependent variable.

C_{max} is a toxicokinetic parameter that denotes the individual's highest measured substance concentration in a specific compartment, in this case rat serum. Blood samples (75 µL) were drawn

Tab. 2: Description of codes used for reporting of clinical observations
Clinical observations scored in binary format, i.e., either 0 (no finding) or 1 (finding).

Variable	Definition
Blood on the gavage probe	The probe has blood on the tip directly after the gavage dosing procedure.
Difficulty dosing	Rat resists during gavage dosing procedure.
Eye(s) half shut/shut	Rat has one or both eyes partially shut/shut. This is usually linked to poor condition or opacities.
Gender	Categorization into male or female animal.
Hairless patches	Rat displays hairless patches on the limbs.
Hair loss general	Rat displays a general hair loss body condition.
Hunched posture	Rat's back is abnormally arched in a concave manner.
Loose feces	Rat displays altered feces consistency, i.e., softer than normal.
Slightly decreased motor activity	Rat shows slightly decreased motor activity, compared to normal activity.
Pale	Rat has pale extremities, skin or mucosa. This observation excludes pale eyes and gums.
Pale eyes	Rat has pale eyes or paler than normal.
Piloerection	Rat has erected fur. Can be observed in connection with dosage (related to substance's flavor).
Ploughing	Rat ploughs its nose in the cage bedding. Can be observed in connection with dosage (related to substance's flavor), but also in undosed animals.
Reflux	Rat has gastroesophageal reflux.
Salivation increased	Rat has an increased rate of salivation observed after administration of the dose. Usually observed in connection with dosage.
Salivation reflex	Rat has an increased rate of salivation, usually before or during dosing procedure.
Stained eyes	Rat has porphyria around the eyes.
Stained nostrils	Rat has porphyria in the nostrils. Often stress- or illness-related, can spread to other fur areas.
Stiff body	Rat body is stiff during handling.
Struggling during handling	Rat struggles excessively during handling and/or dosing procedure.
Tiptoe gait	Rat walks on tiptoes. Usually associated with poor condition.
Trembling	Rat is trembling.
Vocalization during handling	Rat vocalizes when handled. This is usually observed in connection with oral dosing, but can be observed in unhandled animals.

Tab. 3: Body weight measurement days, per study

Study number	Weighing days
Study I	All animals were weighed on study day -1 (before the first dose) and day 3. Additionally, group 1 and group 2 males were weighed on study days 6 and 8 (before sacrifice), and group 1 and group 2 females were weighed on study days 4, 7 and 8 (before sacrifice). Group 3 males were weighed on days 6 and 7, and group 3 females were weighed on days 4 and 5. Group 4 males and females were weighed on day 4.
Study II	All animals were weighed on study day -1 (before the first dose), day 5, 9 and 13. Group 3 animals were additionally weighed on study day 10.
Study III	All animals were weighed on study day -1 (before the first dose), day 4, 22 and 29. Additionally, all males were weighed on study days 10 and 16, and all females on study days 8, 14 and 18.

from the tail vein using capillary tubes, followed by centrifugation and separation into aliquots. A total of 5, 6 and 9 blood samples were drawn at different timepoints on the first and last day of dosing for studies I, II and III, respectively. C_{max} was quantified at the end of each study, using an in-house developed liquid chromatography and tandem mass spectrometry (LC-MS/MS) method. The toxicokinetic profiling and calculations were done by PKxpert AB (Stockholm, Sweden). The C_{max} estimate, i.e., no conversion, was used for the PLS modelling.

Dependent variables

All animals were sacrificed the day after being given the last dose of the candidate drug, except for study I where necropsy was performed after the pre-terminal sacrifice of animals in dose groups 3 and 4. Organs were collected, fixed and processed to wax blocks, sectioned and stained according to standard operating procedures. The slides were given an ID and then analyzed for microscopic pathology by a qualified veterinary pathologist in a blinded manner. A pathological finding of any kind and severity was scored as 1 for that organ and individual, and a 0 was scored if no pathological finding was observed. The following organs and injuries were recorded and included in the analysis, regardless of the severity of the pathological findings:

- Epididymides: cellular debris (males only)
- Liver: diverse findings merged (glycogen depletion, increased hepatocellular mitosis and single hepatocellular necrosis)
- Lymph node: lymphoid depletion
- Large intestines (caecum, colon and rectum): mucosal gland necrosis, atrophy or dilatation
- Bone marrow in the sternum: decreased cellularity
- Testes: tubular atrophy (males only)
- Thymus: lymphoid depletion

Body weight loss of 5% was also tested as a dependent variable.

Multivariate analysis

The partial least squares (PLS) regression is a multivariate data analysis method (Wold, 1975). PLS regression tests possible relationships between a set of independent and dependent variables and is used to predict the outcome of the dependent variables on a new sample (Eriksson et al., 2005; Hubert and Branden, 2003). This method performs well with missing data points in both independent and dependent variables (Eriksson et al., 2013). The data analysis and randomization were performed using Simca software (version 15, Sartorius Stedim Data Analytics AB, Umeå, Sweden). All data were mean-centered and auto-scaled prior to analysis. A default 7-fold cross-validation was used for model development in order to determine the significant number of components (latent variables).

In this study, two types of training and independent test sets were tested in order to investigate the predictive ability of models based on the entire set of data (all three studies) as well as future forecasting capacity on new data (study III based on models trained on studies I and II data):

- Setup 1: Data from all three studies were merged and then randomly divided into a training set (two-thirds of the data) and an independent test set (one-third of the data);
- Setup 2: Data from studies I and II were employed as the training set; study III was used as the independent test set.

Two sets of clinical signs were used in the present study; all registered clinical signs ("full list") and a subset of signs ("shortlist") (Tab. 6). The shortlist was compiled *a priori* based on the authors' previous experience of pharmaceutical toxicity testing in rodents and inspired by the LASA guidance on dose level selection (NC3Rs, 2009). The shortlist of clinical observations, i.e., piloerection, decreased motor activity, stained eyes and nostrils, hunched posture, trembling, vocalization during handling, and 5% body weight loss, comprises general signs of toxicity commonly observed in toxicity testing. These particular clinical signs are also used for severity assessment and for setting humane endpoints in rodents used in other research areas. The aim was to create and test a shortlist of signs that would be easy to observe and use in toxicity studies as well as in other research areas to facilitate early detection of toxicity. In addition, gender and C_{max} were included in the shortlist used in this study.

The employed binary scoring system (0 or 1) scored a 1 for each finding in a given variable, regardless whether it was a dependent or independent variable. Conversely, for each individual with no finding in a given variable, a 0 was scored. The arithmetic mean value is thus 0.5, the standard cut-off value of the dependent variable for model performance classification. This cut-off value can change depending on the model's imbalance with respect to the classes (0 or 1) of the input data. The training set cross-validation procedure was used to determine the cut-off value for assigning the prediction as pathological (> cut-off) or no pathological finding (< cut-off). This was done by setting a cut-off value that maximized balanced accuracy given by the function:

$$Balanced\ accuracy = 0.5 * (specificity + sensitivity)$$

$$Balanced\ accuracy = 0.5 * \left[\left(\frac{number\ of\ true\ negatives}{total\ number\ of\ negative\ events}\right) + \left(\frac{number\ of\ true\ positives}{total\ number\ of\ positive\ events}\right)\right]$$

A cut-off level of 0.8 was employed for model performance classification, i.e., the ability to accurately predict ≥ 80% of the dependent variable findings in the test set. This threshold level is a commonly used value and was found by Ekins and colleagues (2018) to be the ideal value. Poor model performance thus is defined as a lack of the ability to predict at least 80% of the events.

The balanced accuracy is composed of two terms, specificity and sensitivity. Specificity is defined as the rate of true negatives predicted, i.e., the proportion of events correctly predicted as negative when the true result was negative (class 0). Sensitivity is the rate of true positive results predicted, i.e., the proportion of correctly predicted positive events when the true result was positive (class 1). The same cut-off level of 0.8 was used for classification of the model performance regarding specificity and sensitivity.

The importance ranking of the different independent variables is given in terms of variable importance in projection (VIP) score. The VIP score describes the contribution of an independent variable to the outcome (dependent variable) of the derived PLS model. It is obtained by estimating the weighted sum of the squared

correlations between the independent and the dependent variables. The greater a VIP score is, the more information-bearing and predictive power it possesses. The VIP method is implemented in the SIMCA-P computer package; VIP scores with values ≥ 1 are classified as important variables in the model (Lazraq et al., 2003).

3 Results

3.1 Pathology predictions based on Setup 1

Using the full list of registered clinical signs to predict pathology in any organ at necropsy (Tab. 2) when using Setup 1, the predictive models showed a balanced accuracy ≥ 0.8 for four (thymus, bone marrow, testes and epididymides) out of seven organs with pathological findings and for 5% body weight loss (Tab. 4). The predictions were made with an accuracy between 81% and 98% when using all clinical signs, including registered body weight loss. Similar levels of accuracy were observed when using the shortlist of clinical signs (80% to 96%) (Tab. 4) for the same four organs as well as for the large intestine.

The models resulted in a borderline acceptable performance for prediction of the pathological findings in the large intestines using the full list of clinical signs (balanced accuracy ≈ 0.8) and poor performance for prediction of the pathological findings in the liver and lymph node using either the full list or the shortlist (balanced accuracy < 0.8) (Tab. 4). For the liver and lymph node organ predictions, poor model performance in the test sets was anticipated due to the low balanced accuracies observed in the training sets (Tab. 4). Expectably, a similar model performance was observed when using the shortlist in comparison to the inclusion of all clinical signs for the liver, lymph node and large intestines, i.e., it did not improve (Tab. 4).

The predictive power of the individual clinical signs (used as independent variables) to describe the pathological findings (dependent variables) varied. For example, when describing pathological findings of the bone marrow, piloerection and body weight loss were rated as the most information-bearing predictors, both when using the full list (Fig. 1a) as well as the shortlist of clinical signs (Fig. 1b). Eyes half shut was also among the top-ranked signs when using the full list of clinical signs (Fig. 1a). However, this endpoint was not included in the shortlist.

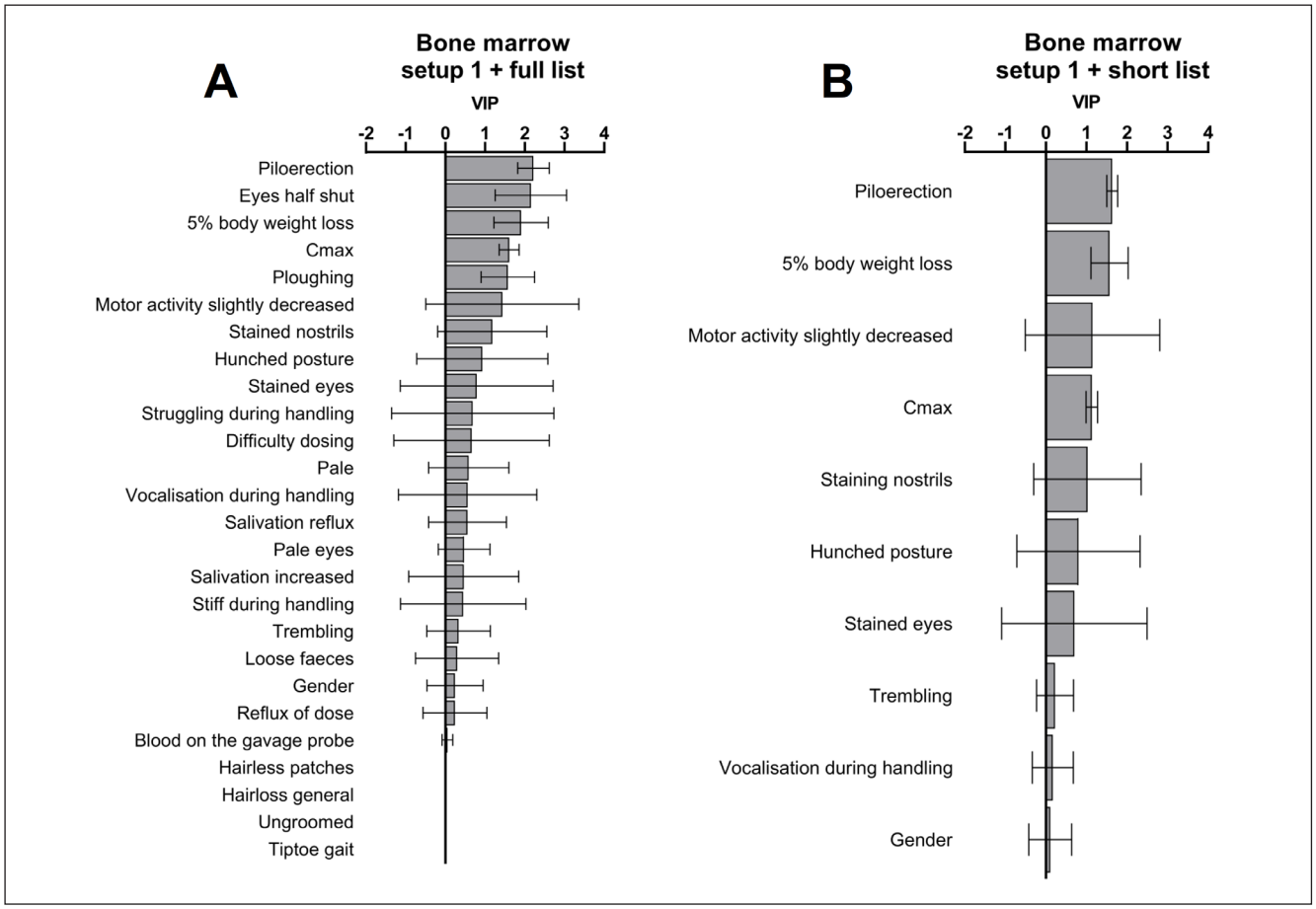

Fig. 1: Model prediction for bone marrow pathology described using VIP scores using Setup 1 and the full list (left) or the shortlist (right) of clinical signs

Error bars represent standard deviations (SDs). Balanced accuracies were 0.985 and 0.955, respectively.

Tab. 4: Prediction results using Setup 1 for the full list or shortlist of clinical signs

Dependent variable	Sensitivity	Specificity	Balanced accuracy	ROC AUC[a]	Dataset
Full list					
5% body weight loss	0.800	0.827	**0.813**	0.964	Training
	0.933	0.694	**0.814**	0.952	Test
Bone marrow	0.846	0.875	**0.861**	0.942	Training
	1.000	0.970	**0.985**	0.996	Test
Epididymides	0.929	0.905	**0.917**	0.976	Training
	0.846	0.933	**0.890**	0.979	Test
Large intestines	0.882	0.882	**0.882**	0.825	Training
	0.769	0.786	**0.777**	0.870	Test
Liver	0.659	0.659	**0.659**	0.728	Training
	0.556	0.643	**0.599**	0.681	Test
Lymph node	0.722	0.717	**0.719**	0.788	Training
	0.500	0.719	**0.609**	0.666	Test
Testes	0.889	0.923	**0.906**	0.979	Training
	0.857	0.905	**0.881**	0.973	Test
Thymus	0.929	0.930	**0.929**	0.957	Training
	0.955	0.778	**0.866**	0.932	Test
Shortlist					
5% body weight loss	0.800	0.800	**0.800**	0.925	Training
	0.933	0.750	**0.842**	0.933	Test
Bone marrow	0.846	0.847	**0.847**	0.937	Training
	1.000	0.909	**0.955**	1.000	Test
Epididymides	0.857	0.857	**0.857**	0.942	Training
	0.769	1.000	**0.885**	0.897	Test
Large intestines	0.697	0.846	**0.772**	0.868	Training
	0.857	0.815	**0.836**	0.744	Test
Liver	0.591	0.585	**0.588**	0.695	Training
	0.407	0.429	**0.418**	0.504	Test
Lymph node	0.722	0.617	**0.669**	0.762	Training
	0.313	0.625	**0.469**	0.566	Test
Testes	0.889	0.846	**0.868**	0.974	Training
	0.714	0.905	**0.810**	0.932	Test
Thymus	0.762	0.884	**0.823**	0.808	Training
	0.773	1.000	**0.886**	0.851	Test

[a]ROC AUC (receiver operating characteristic area under the curve, also known as the concordance statistic) was used as a secondary assessment measure for model comparison purposes.

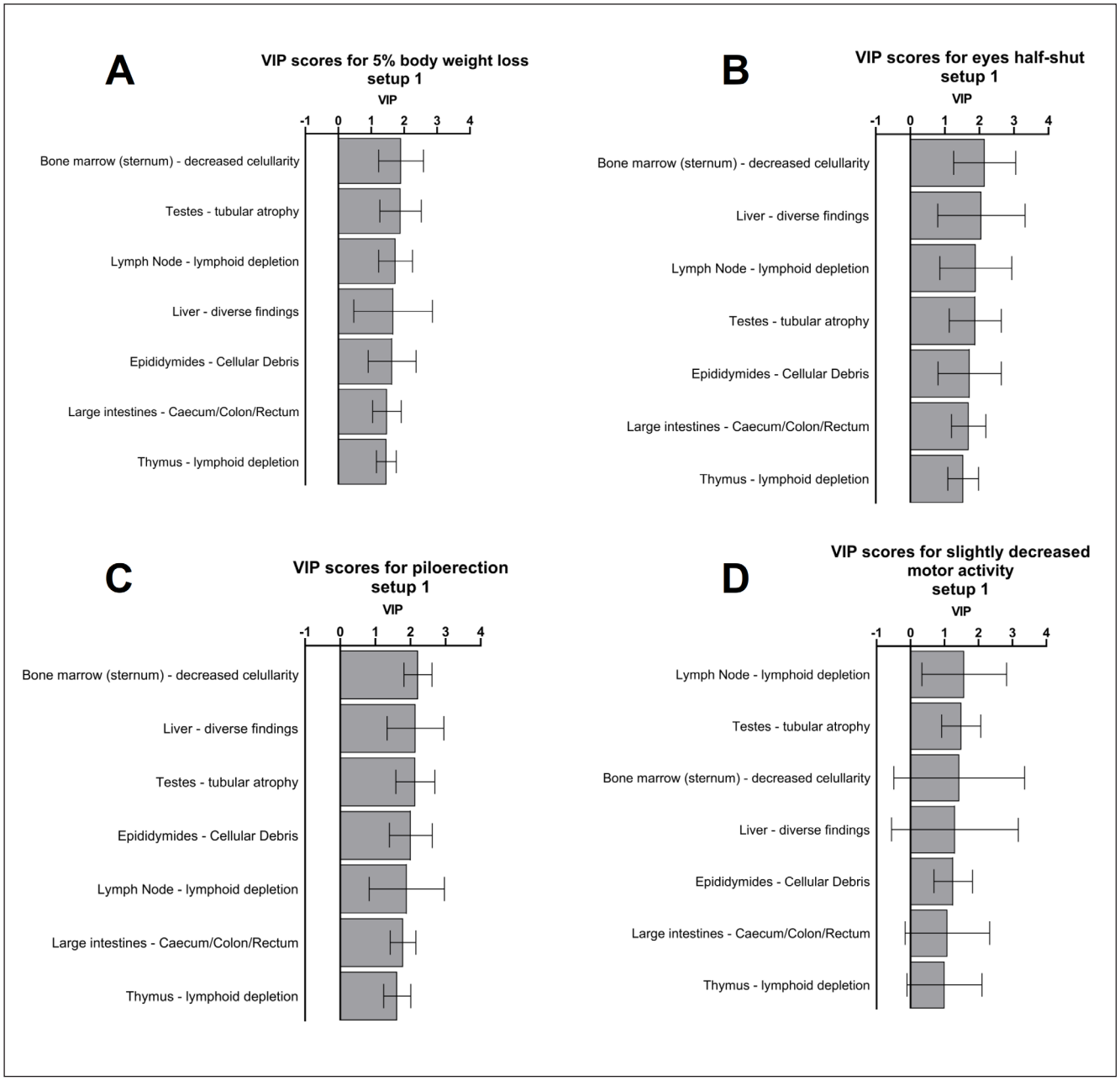

Fig. 2: Extracted mean VIP scores using a 5% body weight loss (2A), piloerection (2B), eyes half shut (2C) and decreased motor activity (2D) as a predictor for pathology findings using Setup 1
Error bars represent SDs.

A 5% body weight loss, piloerection, and eyes half shut were among the most information-bearing variables predicting pathological findings in all seven organs (Fig. 2a,b,c). Slightly decreased motor activity also showed high VIP scores for all organs, however with higher standard deviations (Fig. 2d).

Body weight loss was shown to be not only useful as an independent variable to predict organ pathology (Fig. 2a) but also to be accurately predicted by other clinical signs, both when using the full list and the shortlist (Fig. 3a,b).

3.2 Pathology predictions based on Setup 2

When studies I and II were used as the training set to predict study III (Setup 2), the results indicated an overall good model performance (balanced accuracy ≥ 0.8) for most of the investigated organs (Tab. 5). Using the full list of clinical signs, acceptable predictions between 83 to 100% were seen in the thymus, testes, bone marrow, and epididymides (Tab. 5). The prediction of a 5% body weight loss also showed a high balanced accuracy of 85 to 92%. The large intestines had acceptable model performances for the training sets

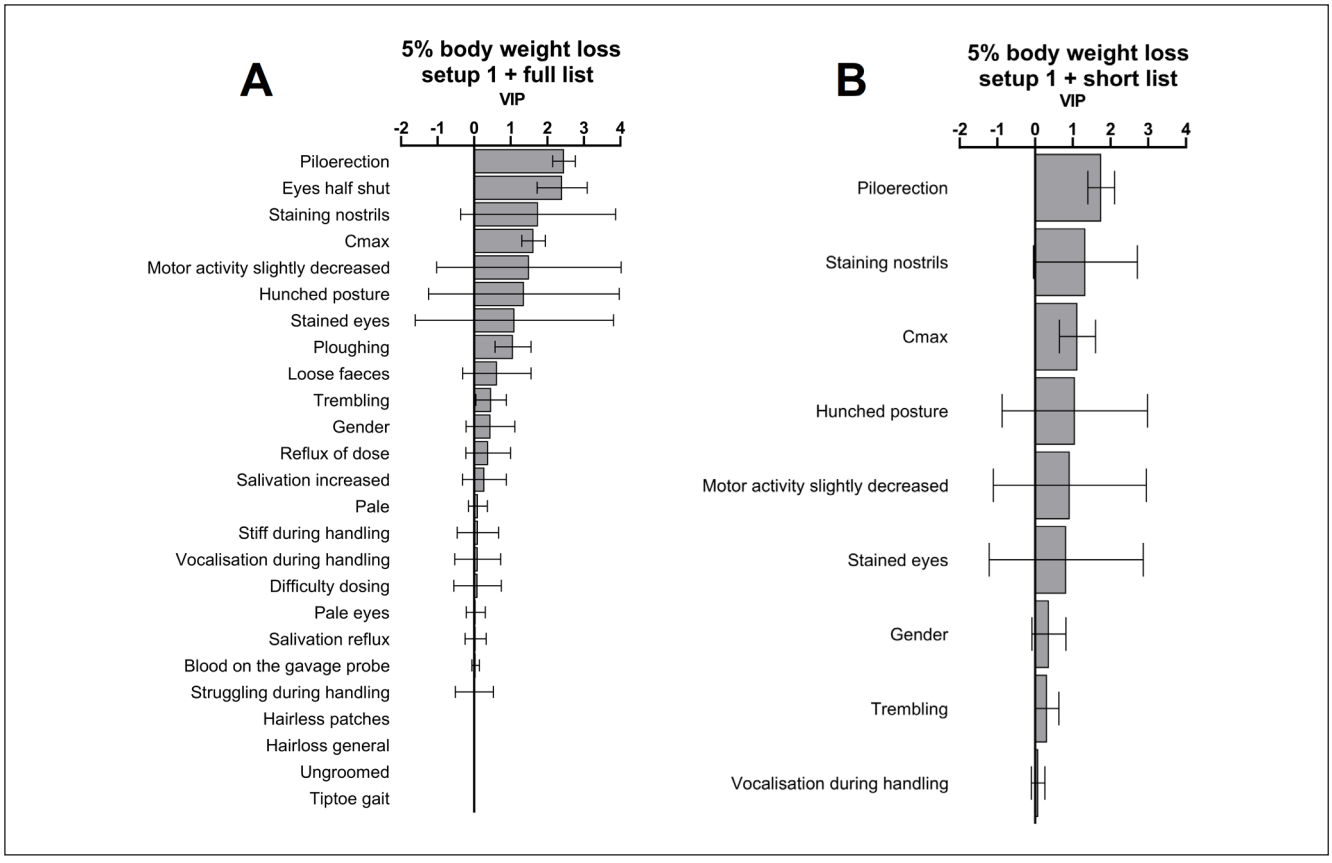

Fig. 3: Model prediction for a 5% body weight loss described using VIP scores using Setup 1 and the full list (left) or shortlist (right) of clinical signs

Error bars represent SDs. Balanced accuracies were 0.814 and 0.842, respectively.

but not for the test sets (93 and 66% prediction, respectively). Low balanced accuracies were obtained from modelling attempts for the pathology registered in the liver and lymph node (Tab. 5).

Overall, the model performances were better when using the full list as compared to models based on the shortlist of clinical signs. The shortlist resulted in low balanced accuracies for all endpoints except body weight loss (Tab. 5).

Piloerection, body weight loss, slightly decreased motor activity and, in the full list, eyes half shut showed the highest VIP scores also when using Setup 2, e.g., when predicting pathology in the bone marrow (Fig. 4a,b).

Comparing Setup 1 and 2, overall a higher balanced accuracy was obtained in Setup 1 (Fig. 5), both for the full list and for the shortlist of clinical observations (Fig. 5). The same pattern was observed for model sensitivity (Fig. 6) but not for specificity (Fig. 7), when comparing the use of the shortlist of clinical signs to the full list, and Setup 1 vs 2.

3.3 Importance of clinical observations in the derived models

The VIP score method was used to identify the most informative predictors of toxicity. The rank order of importance of the independent variables was identified by the PLS models.

The most important across Setups 1 and 2 were piloerection, 5% body weight loss, decreased motor activity, and, in case of the full list, eyes half shut and ploughing (Tab. 6). Additionally, the C_{max} predictive power was highly ranked by the PLS models, as expected, as higher exposure usually correlates well with organ toxicity.

Although ploughing, observed in connection with the dosing procedure, was highly ranked, it was noted that this is often related to a substance's flavor and was mostly observed in the high dose groups, where the concentrations of the test substance, and therefore the flavor intensity, were highest. Therefore, this sign was regarded as related to the concentration of the test substance and its flavor rather than connected to the pathogenesis. For this reason, ploughing and eyes half shut were purposely excluded from the short list although they were well associated with pathological findings when using the full list of clinical signs (Tab. 6).

In sum, the clinical signs piloerection, 5% body weight loss, decreased motor activity and eyes half shut were the most important predictors of pathology. Ploughing and eyes half shut were excluded from the shortlist but were considered informative by the PLS models when using the full list of clinical signs (Tab. 4). Finally, the rank order of the variables was similar regardless of the setup tested (Tab. 4).

Tab. 5: Prediction results using Setup 2 for the full list or shortlist of clinical signs

Dependent variable	Sensitivity	Specificity	Balanced accuracy	ROC AUC[a]	Dataset
Full list					
5% body weight loss	0.933	0.903	0.918	0.963	Training
	1.000	0.700	0.850	0.970	Test
Bone marrow	0.875	0.900	0.888	0.969	Training
	1.000	0.813	0.907	0.896	Test
Epididymides	0.857	0.889	0.873	0.976	Training
	1.000	0.815	0.907	0.989	Test
Large intestines	0.900	0.962	0.931	0.972	Training
	0.370	0.943	0.657	0.686	Test
Liver	0.760	0.762	0.761	0.820	Training
	0.413	0.765	0.589	0.616	Test
Lymph node	0.923	0.879	0.901	0.937	Training
	0.429	0.695	0.562	0.582	Test
Testes	1.000	1.000	1.000	1.000	Training
	1.000	0.719	0.859	0.926	Test
Thymus	1.000	1.000	1.000	1.000	Training
	0.788	0.870	0.829	0.914	Test
Shortlist					
5% body weight loss	0.867	0.871	0.869	0.856	Training
	0.900	0.925	0.913	0.935	Test
Bone marrow	0.875	0.900	0.888	0.935	Training
	0.000	0.987	0.493	0.893	Test
Epididymides	0.857	0.889	0.873	0.857	Training
	0.538	0.963	0.751	0.862	Test
Large intestines	0.850	0.846	0.848	0.940	Training
	0.222	1.000	0.611	0.675	Test
Liver	0.720	0.619	0.670	0.783	Training
	0.239	0.794	0.517	0.496	Test
Lymph node	0.769	0.788	0.779	0.816	Training
	0.048	0.983	0.515	0.450	Test
Testes	1.000	1.000	1.000	1.000	Training
	0.375	0.969	0.672	0.863	Test
Thymus	0.742	1.000	0.871	0.772	Training
	0.303	0.978	0.641	0.635	Test

[a]ROC AUC (receiver operating characteristic area under the curve) was used as a secondary assessment measure for model comparison purposes.

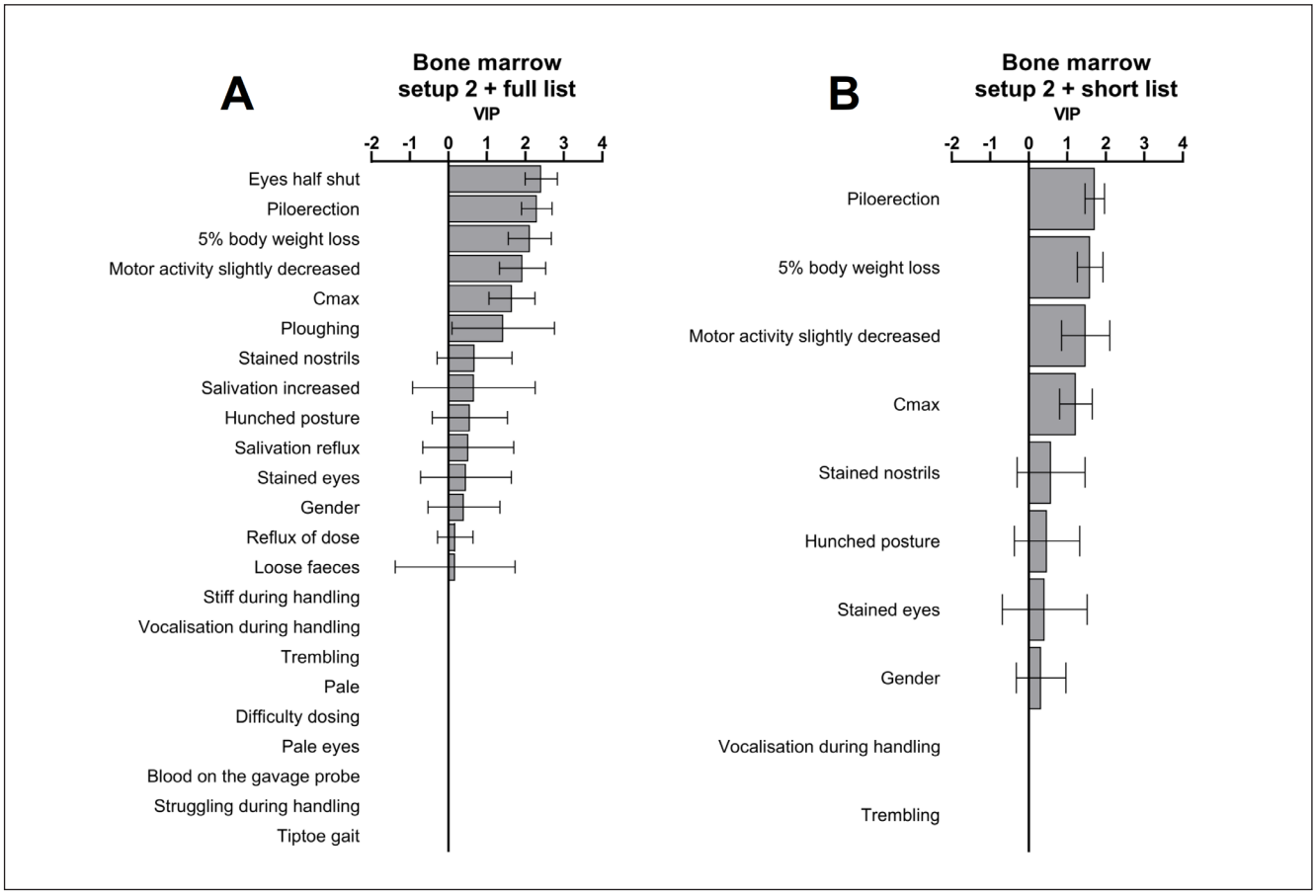

Fig. 4: Model prediction for bone marrow pathology described as VIP scores using Setup 2 and the full list (left) or shortlist (right) of clinical signs

Error bars represent SDs. Balanced accuracies were 0.907 and 0.493, respectively.

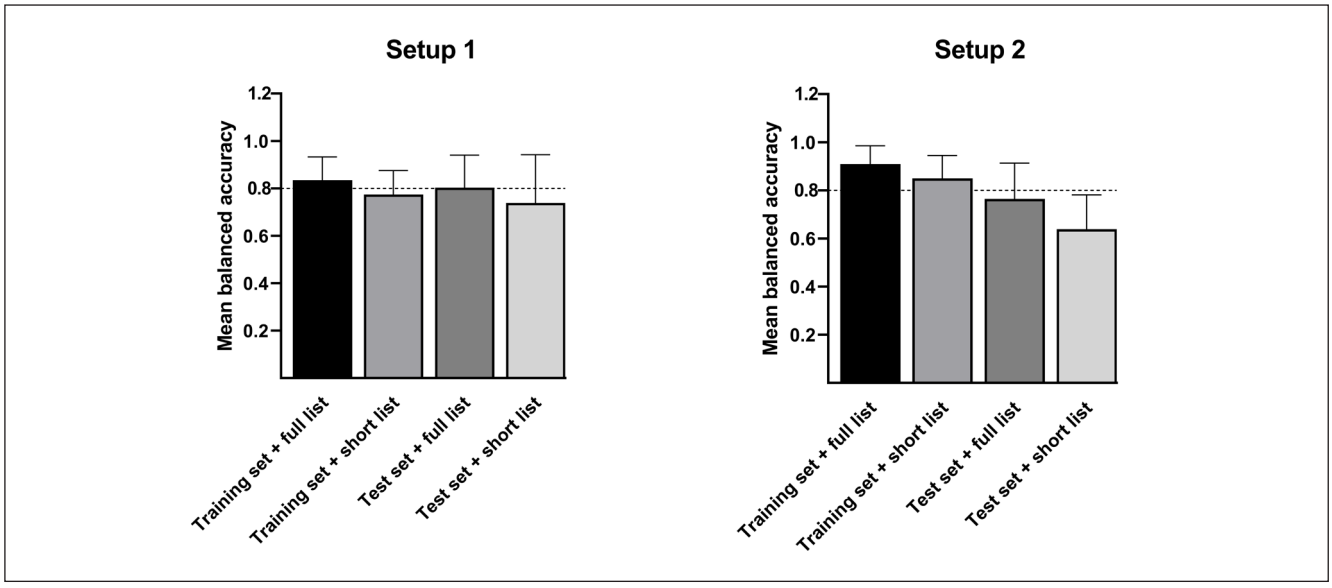

Fig. 5: Comparison of the mean balanced accuracy across all pathological endpoints using Setup 1 versus Setup 2

Error bars represent SDs; the dotted line represents the cut-off value (0.8) for performance evaluation.

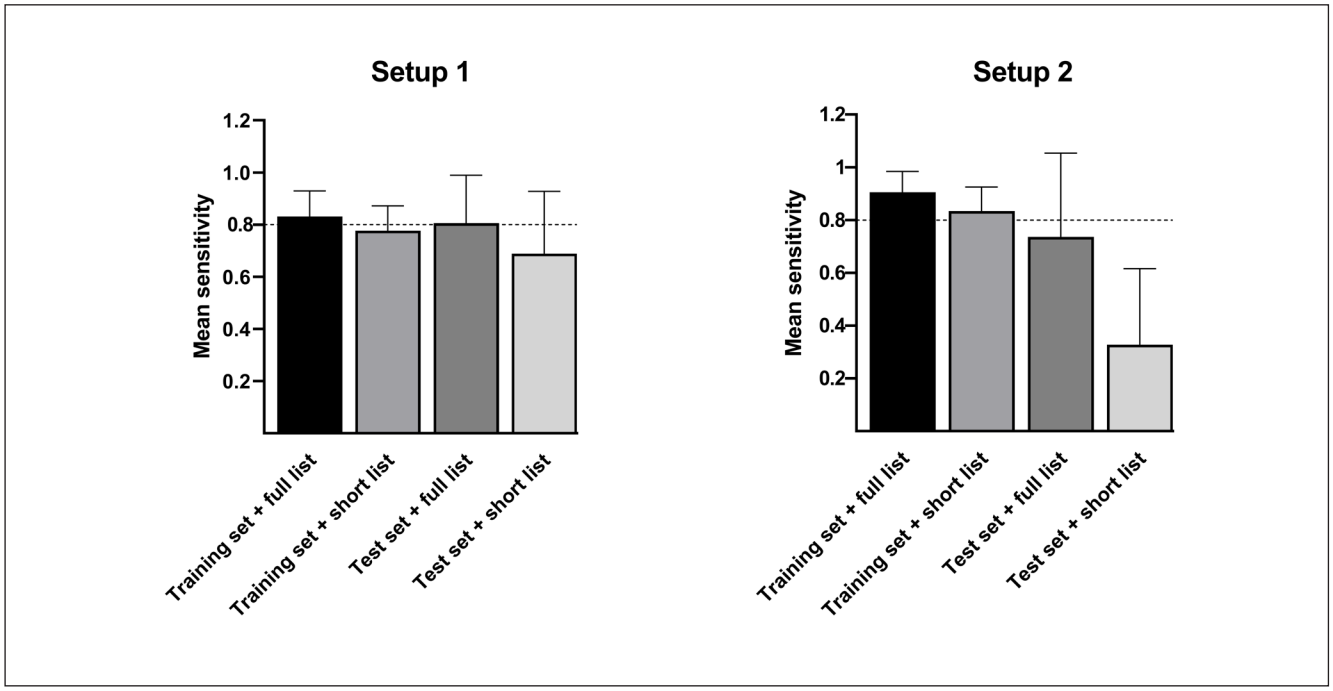

Fig. 6: Comparison of the mean sensitivity (i.e., correct positive predictions) across all pathological endpoints using Setup 1 versus Setup 2

Error bars represent SDs; the dotted line represents the cut-off value (0.8) for performance evaluation.

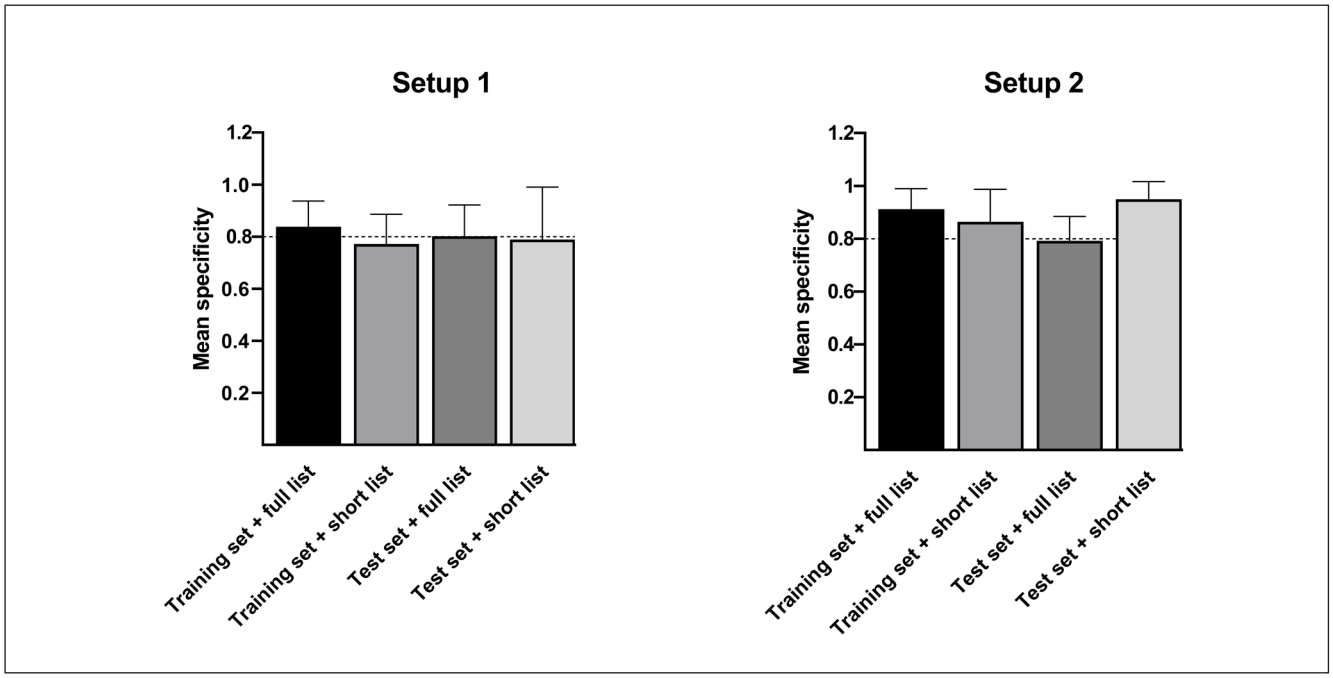

Fig. 7: Comparison of the mean specificity (i.e., correct negative predictions) across all pathological endpoints using Setup 1 versus Setup 2

Error bars represent SDs; the dotted line represents the cut-off value (0.8) for performance evaluation.

Tab. 6: Overall importance of clinical observations ranked in order of information-bearing rank for each tested setup, using merged data from all dependent variables in the derived PLS models[a]

Setup 1, full list	Setup 1, shortlist	Setup 2, full list	Setup 2, shortlist
Piloerection	**Piloerection**	**Piloerection**	**Piloerection**
Eyes half shut	**5% body weight loss**	Eyes half shut	**5% body weight loss**
Ploughing	**C_{max}**	**5% body weight loss**	**Slightly decreased motor activity**
5% body weight loss	**Slightly decreased motor activity**	**Slightly decreased motor activity**	**C_{max}**
C_{max}	**Stained nostrils**	Ploughing	**Stained nostrils**
Slightly decreased motor activity	**Hunched posture**	**C_{max}**	**Hunched posture**
Stained nostrils	**Stained eyes**	**Stained nostrils**	**Stained eyes**
Hunched posture	**Gender**	**Hunched posture**	**Gender**
Salivation reflux	**Trembling**	**Stained eyes**	**Trembling**
Pale	**Vocalization during handling**	Salivation increased	**Vocalization during handling**
Salivation increased		**Gender**	
Stained eyes		Loose feces	
Pale eyes		Salivation reflux	
Difficulty dosing		Reflux of dose	
Loose feces		Blood on the gavage probe	
Struggling during handling		Difficulty dosing	
Gender		Hairless patches	
Stiff during handling		Hair loss general	
Trembling		Pale eyes	
Vocalization during handling		Pale	
Reflux of dose		Stiff during handling	
Tiptoe gait		Struggling during handling	
Blood on the gavage probe		Tiptoe gait	
Hairless patches		**Trembling**	
Hair loss general		**Vocalization during handling**	

[a] Bold text indicates that the descriptor is part of the shortlist of descriptors. Grey background indicates that the endpoints were of equal unimportance in the tested setup.

Some of the clinical observations, i.e., vocalization during handling and trembling, which were part of the shortlist, were not particularly important in either of the tested setups, which is most likely related to the very few registered occurrences.

4 Discussion

4.1 Comparison of Setups 1 and 2

In the present study, we investigated whether clinical observations could predict pathological findings, which suggests that clinical signs can be used as early predictors for adverse outcomes observed after necropsy. Piloerection, eyes half-shut and slightly decreased motor activity showed the strongest association with the pathological findings in the thymus, testes, bone marrow and epididymides. Body weight loss also showed a high empirical association with pathological findings. Setup 1 split data from all three studies into a training (2/3 of the data) and a test set (1/3 of the data), aiming to obtain the best associations possible between clinical signs and pathology. Setup 2 used data from the two shorter dose-finding studies to predict the results of the third study, which had a longer duration and used lower doses.

Accordingly, there was a greater likelihood of more severe effects in study I and II, given the higher dosages, compared to study III.

For both tested setups, good model performances were found when associating the full list of clinical signs including 5% body weight loss with pathological findings in the thymus, testes, bone marrow and epididymides. The shortlist of clinical signs was shown to be useful in Setup 1, although a slightly better model performance was achieved when using the full list. The results show that PLS models can describe and empirically predict associations between clinical signs and pathological findings, even with limited amounts of data, such as the case of study I where two dose groups were pre-terminally sacrificed.

Piloerection, stained nostrils, and decreased motor activity previously have been identified as markers of general toxicity, often used to assess the severity of toxicity-induced distress (NC3Rs, 2009) as well as to support decisions for pre-terminal sacrifice due to general poor animal condition (Morton and Griffiths, 1985). Members of Swedish animal ethics committees also top-ranked these endpoints in relation to animal distress (75^{th} percentile weights) (Ringblom et al., 2017b). Piloerection is a general symptom often associated with toxicity but can also be related to the substance's taste or animal's discomfort independently of dose administration. It is thus dependent on the time of occurrence and may be plausibly discarded if observed immediately after dose administration. Stained nostrils are a typical sign of toxicity and decreased animal wellbeing, especially when observed recurrently. Decreased motor activity is related to poor animal condition and often linked to more severe suffering (Sewell et al., 2015). In the present study, slightly decreased motor activity was observed only in some animals in study I and II, but was important despite its low frequency. Eyes half shut was also a strong predictor for pathological findings, although it is traditionally regarded as a clinical sign that indicates pain (Langford et al., 2010). Eyes half shut is included in animal welfare assessment guidelines (e.g., Morton and Griffiths, 1985; EC, 2012) but not in the LASA guidance on dose level selection (NC3Rs, 2009). Our results strongly support that these three descriptors (piloerection, eyes half shut and decreased motor activity), even when observed as mild symptoms or with low frequency, could be predictors of general toxicity. They carry, thus, a refinement potential, in terms of study management and animal welfare monitoring. To our knowledge, there are no publications that describe the use of clinical signs in research animals to predict side effects caused by the test drug or for other purposes than as additive information in the risk assessment and for animal welfare reasons.

Body weight loss is a useful tool to determine the MTD in different species used in toxicity testing, and for rodents a 20% body weight loss has been used to identify an appropriate MTD (NC3Rs, 2009). The use of this substantial body weight loss for deciding the MTD has though been challenged (Chapman et al., 2013). For rats, a body weight loss of 10% has been suggested as a better threshold in up to 7-day MTD studies for pharmaceutical development purposes (Chapman et al., 2013) and has also been shown to be a sign of evident toxicity in acute inhalation studies (Sewell et a. 2015). In the present study, we wanted to investigate whether a less pronounced body weight loss could be predictive of toxicity. We showed that a less severe body weight loss of 5% in rats was predictive of pathology findings in the 7 up to 28-day studies analyzed. Based on the observed predictive value of body weight loss, we suggest a 5% body weight decrease to be an important predictor to consider when investigating toxicity in future animal studies. Further studies are required to support this suggested threshold for toxicity, which can be useful for decision-making on, for example, administering the next dose or as a point of departure for reference dose setting. In the present study, body weight loss was also shown to be predicted by clinical signs.

As expected, higher C_{max} concentrations, i.e., higher exposure, were predictive of pathological findings in all modelled organs. However, C_{max} does not represent a quick assessment of animal wellbeing, as it requires a resource-consuming toxicokinetic study and repeated blood sampling.

Overall accurate predictions (81 to 98% in terms of balanced accuracy) were seen in four of the seven investigated organs: thymus, bone marrow, testes and epididymides. Bone marrow, testes and epididymides could be regarded as likely target organs for side effects of an anti-cancer drug, as they are continuously proliferative organs (Remesh, 2012). The intestine is also a continuously proliferative organ, but intestinal toxicity was not as well predicted by the derived models.

Thymus and lymph node pathology and body weight loss could be related to secondary toxicity due to stress caused by drug-induced specific organ toxicity. Toxicity to the liver is also often regarded as non-specific. The pathological findings in the liver and lymph node showed an irregular pattern across the present studies, and there were too few observations for the specific liver injuries (glycogen depletion, increased hepatocellular mitoses and single hepatocellular necrosis) to be modelled individually. There were also fewer observations related to lymphoid depletion in the lymph node in study III due to the lower doses tested. In sum, the clinical signs in the present study on an anti-cancer drug seemed to be related to drug-specific side effects rather than to stress-related secondary toxicity.

4.2 Shortlist of clinical signs

We investigated whether a selection of clinical signs previously established as relevant for toxicity assessment would yield similar results compared to using all clinical signs. The shortlist was compiled to represent the most meaningful observations from a toxicological point of view. The model predictions resulted in similar balanced accuracies when the shortlist of clinical observations was used in Setup 1 but not in Setup 2, indicating a lack of generalizability for some of the models. Thus, all clinical observations are important, but some are more important than others (Tab. 6). A lower balanced accuracy was evident when the shortlist was used in combination with study III as the test set (Fig. 5). The model's performance deterioration was especially noticeable in the sensitivity results (Fig. 6), while the specificity results did not deteriorate in a similar way (Fig. 7). This could be explained by imbalances in the majority and minority classes, i.e., the negative and positive pathological findings. The lower sensitivity can be translated as overprediction of positive pathological findings, lowering the overall balanced accuracy too (Fig. 5 and 6). How-

ever, the specificity did not deteriorate in a similar way, as the models predicted more negative results (the majority class) to the cost of lower sensitivity, especially in Setup 2 (Fig. 7).

In conclusion, the model performance improved with the number of clinical endpoints included but also with the amount of data employed (the training set of Setup 1 contained more data than the training set of Setup 2), suggesting that more evident associations could be observed if data from other studies and substance classes were available. Furthermore, this analysis showed that the shortlist of clinical signs yielded similar results as the more complete assessment of clinical observations in the scenario where data from all three studies were combined.

4.3 Temporality

PLS is not time-resolved, and it disregards the time-point when a clinical sign was observed and only considers the total number of times it occurred averaged over the number of study days for comparability between different study durations. This absence of temporality may seem like a limitation but it is actually the opposite, being potentially useful in terms of creating a real-time surveillance system to predict toxicity and assess animal wellbeing. During the course of an animal study, clinical observations could be registered in a software enabled with PLS modelling, which could then perform real-time predictions and alert the investigators in advance to a likely outcome in target organs for some of the animals or for the study outcome as a whole. Although such pathological injuries can only be confirmed after necropsy, a predicted pathological outcome may support study management during its course, in terms of intended study outcome and decision-making for dose administration, simultaneously increasing animal wellbeing through reducing unnecessary suffering. As it might serve as a tool to support decision-making in case of change or interruption of the toxicity study, it represents a potential refinement action.

4.4 Future perspectives and final remarks

A wider understanding and use of clinical signs in any animal model and research area are necessary to provide important information about side effects and risks associated with the used test substance. Drugs with positive treatment effects but also even mild side effects in a disease model might not be suitable in the clinic. By making decisions that support candidate drugs with the least toxic effects early in drug development in the pharmaceutical industry and academia, many animal experiments could potentially be avoided.

Although clinical observations made during toxicity testing studies are registered, they are seldom used for decision-making in chemical risk assessment or pharmaceutical development (OECD, 2008; WHO, 2009). This suggests an underestimation of the information provided by clinical signs. In this study, we have shown that clinical signs can predict organ injuries that will potentially be observed after necropsy. Future research in this area may focus on repeating the analysis with larger amounts of data, including longer studies and other classes of substances in order to corroborate or refute the presented proof-of-principle that there are associations between clinical signs and specific pathological findings. Further elucidation of these associations could improve study management and design, promoting refinement and reduction of animal studies by enabling greater use of the information obtained during *in vivo* studies. Furthermore, prediction models based on the presented shortlist of clinical signs, extended to include eyes half shut, can be useful during efficacy studies done in the early pharmaceutical development process, as they can represent valuable information for test substance selection before regulatory toxicity testing. Eyes half shut was shown to have good predictive power and should therefore be considered in the dose setting guidelines for toxicity testing.

In conclusion, we found that clinical observations were clearly associated with pathological findings, which suggests that clinical signs may be used as early predictors of adverse outcomes observed after necropsy. For the investigated anti-cancer drug, piloerection, eyes half shut and decreased motor activity showed the strongest associations with the pathological findings in the thymus, testes, bone marrow and epididymides. Additionally, a strong empirical association was observed for a 5% body weight loss, which was an accurate predictor regarding organ injuries, but also could be predicted by the clinical signs. The empirically derived PLS regression models predicted accurately over 80% of the animals' pathological findings in the mentioned organs when building the model using all clinical observations from the three animal studies. These results show that PLS modelling represents a promising analytical method and a strong candidate for a real-time toxicity and animal welfare monitoring system.

We conclude that clinical observations can be used as early markers of toxicity, as well as to assess and improve welfare during pharmaceutical development, reducing animal use and unnecessary suffering. In addition, we suggest that signs registered during toxicity testing studies, as well as in other research areas, could be simplified using a shortlist including a 5% body weight loss, piloerection, eyes half shut and decreased motor activity. Further research is required to improve the accuracy of these predictions and to further support the proof-of-principle this analysis has presented.

References

Chapman, K., Sewell, F., Allais, L. et al. (2013). A global pharmaceutical company initiative: An evidence-based approach to define the upper limit of body weight loss in short term toxicity studies. *Regul Toxicol Pharmacol 67*, 27-38. doi: 10.1016/j.yrtph.2013.04.003

EC – European Commission (2006). Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *OJ L 396*, 1-849. http://data.europa.eu/eli/reg/2006/1907/2018-05-09 (accessed 15.11.2019)

EC (2012). Working document on a Severity Assessment Framework. Brussels, 11-12 July 2012. https://ec.europa.eu/environment/chemicals/lab_animals/pdf/Endorsed_Severity_Assessment.pdf (accessed 15.02.2020)

EC (2019). 2019 report on the statistics on the use of animals for scientific purposes in the Member States of the European Union in 2015-2017. COM (2020) 16 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1570010289634&uri=SWD:2020:10:FIN (accessed 15.02.2020)

Ekins, S., M. Clark, A. M., Perryman, A. L. et al. (2018). Accessible machine learning approaches for toxicology. In S. Ekins (ed.), *Computational Toxicology: Risk Assessment for Chemicals* (1-29). Wiley. doi:10.1002/9781119282594.ch1

EMA – European Medicines Agency (2010). Guideline on repeated dose toxicity. CPMP/SWP/1042/99 Rev 1. https://www.ema.europa.eu/documents/scientific-guideline/guideline-repeated-dose-toxicity-revision-1_en.pdf (accessed 25.10.2019)

Eriksson, L., Johansson, E. and Lundstedt, T. (2005). Regression- and projection-based approaches in predictive toxicology. In C. Helma (ed.), *Predictive Toxicology* (177-222). Taylor and Francis. doi:10.1201/9780849350351

Eriksson, L., Byrne, T., Johansson, E. et al. (2013). *Multi-and Megavariate Data Analysis Basic Principles and Applications*. 3rd edition. Umetrics Academy. ISBN 9197373052.

EU – European Union (2001). Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to medicinal products for human use. *OJ L 311*, 67-128. https://eur-lex.europa.eu/eli/dir/2001/83/oj (accessed 25.10.2019)

EU (2010). Directive 2010/63/eu of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. Volume 53 ed. *OJ L 276*, 33-79. doi:10.3000/17252555.L_2010.276.eng (accessed 25.10.2019)

Gould, S. and Scott, R. C. (2005). 2-hydroxypropyl-β-cyclodextrin (HP-β-CD): A toxicology review. *Food Chem Toxicol 43*, 1451-1459. doi:10.1016/j.fct.2005.03.007

Hornberg, J. J., Laursen, M., Brenden, N. et al. (2014). Exploratory toxicology as an integrated part of drug discovery. Part I: Why and how. *Drug Discov Today 19*, 1131-1136. doi:10.1016/j.drudis.2013.12.008

Hubert, M. and Branden, K. V. (2003). Robust methods for partial least squares regression. *J Chemometr 17*, 537-549. doi:10.1002/cem.822

ICH – International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (2009). Guidance on nonclinical safety studies for the conduct of human clinical trials and marketing authorization for pharmaceuticals M3 (R2). International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Multidisciplinary/M3_R2/Step4/M3_R2__Guideline.pdf (accessed 05.11.2019)

JECFA – Joint FAO/WHO Expert Committee on Food Additives (2016). Evaluation of certain food additives and contaminants. 80th report of the Joint FAO/WHO Expert Committee on Food Additives. World Health Organization. IBSN 9789241209953. https://apps.who.int/iris/bitstream/handle/10665/250277/9789241210003-eng.pdf?sequence=1 (accessed 15.03.2020)

Jonsson, O., Villar, R. P., Nilsson, L. B. et al. (2012). Capillary microsampling of 25 µl blood for the determination of toxicokinetic parameters in regulatory studies in animals. *Bioanalysis 4*, 661-674. doi:10.4155/bio.12.25

Kalantari, F., Ringblom, J., Sand, S. et al. (2017). Influence of distribution of animals between dose groups on estimated benchmark dose and animal distress for quantal responses. *Risk Anal 37*, 1716-1728. doi:10.1111/risa.12741

Kilkenny, C., Browne, W. J., Cuthill, I. C. et al. (2010). Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biol 8*, e1000412. doi:10.1371/journal.pbio.1000412

Langford, D. J., Bailey, A. L., Chanda, M. L. et al. (2010). Coding of facial expressions of pain in the laboratory mouse. *Nat Methods 7*, 447-449. doi:10.1038/nmeth.1455

Lazraq, A., Cléroux, R. and Gauchi, J.-P. (2003). Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometr Intell Lab Syst 66*, 117-126. doi:10.1016/S0169-7439(03)00027-3

Morton, D. B. and Griffiths, P. H. (1985). Guidelines on the recognition of pain, distress and discomfort in experimental animals and an hypothesis for assessment. *Vet Rec 116*, 431-436. doi:10.1136/vr.116.16.431

Morton, D. B. (1997). A scheme for the recognition and assessment of adverse effects in animals. *Animal Alternatives, Welfare, and Ethics 27*, 235-240. ISBN:0-444-82424-3

NAC – National Advisory Committee for Acute Exposure Guideline Levels for Hazardous Substances (2001). *Standing Operating Procedures for Developing Acute Exposure Guideline Levels for Hazardous Chemicals*. National Research Council, Committee on Toxicology, Subcommittee on Acute Exposure Guideline Levels (AEGL). Washington, DC, USA: National Academy Press. Online ISBN 9780309570114. https://ebookcentral.proquest.com/lib/ki/detail.action?docID=3377445 (accessed 21.02.2020)

NC3Rs – National Centre for the Replacement, Refinement and Reduction of Animals in Research (2009). Guidance on dose level selection for regulatory general toxicology studies for pharmaceuticals. London: NC3Rs/LASA. http://www.lasa.co.uk/PDF/LASA-NC3RsDoseLevelSelection.pdf (accessed 05.11.2019)

NC3Rs (2010). ARRIVE Guidelines. https://www.nc3rs.org.uk/arrive-guidelines (accessed 05.11.2019)

OECD – Organisation for Economic Co-operation and Development (2000). Guidance Document on the Recognition, Assessment, and Use of Clinical Signs as Humane Endpoints for Experimental Animals Used in Safety Evaluation. *Series on Testing and Assessment, No. 19*. OECD Publishing, Paris. doi:10.1787/9789264078376-en

OECD (2008). Test No. 407: Repeated Dose 28-day Oral Toxicity Study in Rodents. *OECD Guidelines for the Testing of Chemicals, Section, 4*. OECD Publishing, Paris. doi:10.1787/9789264070684-en

OECD (2018). Test No. 442D: In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method. *OECD Guidelines for the Testing of Chemicals, Section, 4*. OECD Publishing, Paris. doi:10.1787/9789264229822-en

OECD (2019). Test No. 442C: In Chemico Skin Sensitisation. *OECD Guidelines for the Testing of Chemicals, Section, 4*. OECD Publishing, Paris. doi:10.1787/9789264229709-en

Olson, H., Betton, G., Robinson, D. et al. (2000). Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol 32*, 56-67. doi:10.1006/rtph.2000.1399

Remesh, A. (2012). Toxicities of anticancer drugs and its management. *Int J Basic Clin Pharmacol 1*, 2-12. doi:10.5455/2319-2003.ijbcp000812

Ringblom, J., Kalantari, F., Johanson, G. et al. (2017a). Influence of distribution of animals between dose groups on estimated benchmark dose and animal welfare for continuous effects. *Risk Anal 38*, 1143-1153. doi:10.1111/risa.12929

Ringblom, J., Törnqvist, E., Hansson, S. O. et al. (2017b). Assigning ethical weights to clinical signs observed during toxicity testing. *ALTEX 34*, 148-156. doi:10.14573/altex.1512211

Sewell, F., Chapman, K., Baldrick, P. et al. (2014). Recommendations from a global cross-company data sharing initiative on the incorporation of recovery phase animals in safety assessment studies to support first-in-human clinical trials. *Regul Toxicol Pharmacol 70*, 413-429. doi:10.1016/j.yrtph.2014.07.018

Sewell, F., Ragan, I., Marczylo, T.et al. (2015). A global initiative to refine acute inhalation studies through the use of 'evident toxicity' as an endpoint: Towards adoption of the fixed concentration procedure. *Regul Toxicol Pharmacol 73*, 770-779. doi:10.1016/j.yrtph.2015.10.018

SFS – Svensk författningssamling (1988). L 1 Djurskyddslag. 1988:534 [Article in Swedish]. http://rkrattsbaser.gov.se/sfst?bet=1988:534 (accessed 25.10.2019)

SJVFS – Statens Jordbruksverks Föreskrifter (2012). L 150 Statens Jordbruksverks Föreskrifter och allmänna råd om försöksdjur. 2012:26. ISSN 1102-0970 [Article in Swedish]. https://www.jordbruksverket.se/download/18.3c1967aa13afeea1eb880002406/2012-026.pdf (accessed 25.10.2019)

SJVFS (2015). L 150 Statens Jordbruksverks Föreskrifter och allmänna råd om försöksdjur SJVFS 2015:24. ISSN 1102-0970 [Article in Swedish]. https://www.jordbruksverket.se/download/18.7c05177614ea46421eb82598/1437380698357/2015-024.pdf (accessed 25.10.2019)

Sparrow, S. S., Robinson, S., Bolam, S. et al. (2011). Opportunities to minimise animal use in pharmaceutical regulatory general toxicology: A cross-company review. *Regul Toxicol Pharmacol 61*, 222-229. doi:10.1016/j.yrtph.2011.08.001

Törnqvist, E., Annas, A., Granath, B. et al. (2014). Strategic focus on 3R principles reveals major reductions in the use of animals in pharmaceutical toxicity testing. *PLoS One 9*, e101638. doi:10.1371/journal.pone.0101638

USDA – United States Department of Agriculture (1966). The Animal Welfare Act – Public Law 89-544. https://www.nal.usda.gov/awic/animal-welfare-act-publiclaw-89-544-act-august-24-1966 (accessed 15.03.2020)

US FDA – Food and Drug Administration (2010). Guidance for industry – M3 (R2) nonclinical safety studies for the conduct of human clinical trials and marketing authorization for pharmaceuticals. Rev 1. https://www.fda.gov/downloads/drugs/guidances/ucm073246.pdf (accessed 15.08.2019)

WHO – World Health Organization (2009). Principles and methods for the risk assessment of chemicals in food. EHC 240. https://www.who.int/publications-detail/principles-and-methods-for-the-risk-assessment-of-chemicals-in-food

Wold, H. (1975). Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. *J Appl Probab 12*, 117-142. doi:10.1017/S0021900200047604

Zidar, J., Weber, E. M., Ewaldsson, B. et al. (2019). Group and single housing of male mice: Collected experiences from research facilities in Sweden. *Animals 9*, 1010. doi:10.3390/ani9121010

Conflict of interest

The authors declare no conflicts of interest.

Funding

This work was supported by the Swedish Research Council for Sustainable Development (Formas, grant number 2016-01380 and 2013-01966), the Swedish Research Council (Vetenskapsrådet, grant number 2016-03085) and by the Knut and Alice Wallenberg Foundation.

Acknowledgements

The authors would like to acknowledge the *in vivo* facility staff at the former Swetox (Swedish Toxicology Sciences Research Center), currently part of the Chemical and Pharmaceutical Safety Unit of the Research Institutes of Sweden (RISE), for performing the animal studies and for the data collection.