Regulatory Toxicology and Pharmacology 101 (2019) 79–90

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

A comparative assessment of the CLP calculation method and *in vivo* testing for the classification of plant protection products

Denise Kurth*, Korinna Wend, Sarah Adler-Flindt, Sabine Martin

Federal Institute for Risk Assessment, Max-Dohrn-Str. 8-10, 10589, Berlin, Germany

ARTICLE INFO

Keywords:
CLP calculation method
Additivity approach
Weight of evidence approach
Hazard assessment
Pesticides
Plant protection products

ABSTRACT

In Europe, animal testing for the purpose of regulatory plant protection product (PPP) assessment should be undertaken only as a last resort. Nevertheless, there is a need to improve the acceptance of alternative methods, which has been slow due to a lack of data regarding the predictivity of *in vivo* effects. The CLP calculation method is an alternative method based on the concentration addition of all adverse substances in a mixture. It is often applied as a conservative approach for the estimation of toxicodynamic interactions. However, PPPs consist of pesticides and co-formulants, which in combination can also exhibit altered toxicokinetic properties. Our analysis revealed that oral and inhalation toxicity was underestimated for approximately 45% of the *in vivo* classified products by the CLP calculation method as compared to *in vivo* testing. With regard to skin and eye irritation, the CLP calculation method underestimated the irritating potential in 22% and 6% of PPPs, respectively. Based on specific concentration limits, skin sensitisation was underestimated in 34% of PPPs. Similar false negative rates have been reported for PPP *in vitro* testing. Hence, we suggest the development of an integrated assessment strategy, weighing all available information and considering relevant parameters influencing predictivity and uncertainty.

1. Introduction

In Europe, pesticides and their products are strictly regulated since they can be toxic to target as well as non-target organisms. Human health and safety is ensured by the comprehensive evaluation of inherent toxicological properties, the derivation of toxicological reference values from dose-response relations, classification and labelling, and the exposure assessment of humans, companion animals, and livestock (Solecki and Ritz, 2018). According to Regulation (EC) No 1107/2009, pesticides (i.e. active substances) are evaluated and approved for use in the European Union (EU). Based on EU legal requirements, a set of studies on potential acute, chronic, and sub-chronic effects of the active substance are required. Plant protection products (PPPs) containing the approved active substance as well as co-formulants are assessed on a zonal level and approved by each member state. Co-formulants can have various functions in PPPs. For instance, they act as wetting, antifoaming or dispersing agents, solvents, preservatives, emulsifiers, or antioxidants (Federal Office of Consumer Protection and Food Safety, 2018). Notably, co-formulants can influence the toxicokinetic properties of the active substance(s) (Damalas and Eleftherohorinos, 2011). Furthermore, multiple active substances in a PPP can partake in toxicodynamic interactions, causing additive or even aggravated toxicity (Hernández et al., 2017). Therefore, applicants who seek PPP authorisation must provide sufficient evidence that their product can be safely applied. In the past, this evidence commonly included six *in vivo* studies on acute oral, dermal, and inhalation toxicity as well as studies on skin and eye irritation and skin sensitisation.

However, *in vivo* testing may cause severe distress for animals, it is subject to inter- and intra-species variability, and it is time and cost-intensive (Creton et al., 2010; Hamm et al., 2017; Sewell et al., 2017). In 2016, more than 33,000 animals were used to meet PPP legislation in Germany (Bundesministerium für Ernährung und Landwirtschaft, 2017), even though current legislation promotes the reduction of animal testing to the necessary minimum. The submission of alternative testing methods to address the above endpoints is of increasing relevance. According to Regulation (EC) No 1107/2009, tests on vertebrate animals shall only be undertaken as a last resort (EU, 2009). However, alternative methods need to be fit for purpose, i.e. reproducible and reliable in predicting PPP adverse effects in humans (Sewell et al., 2017). For the majority of alternative methods, especially data on the prediction of adverse effects in humans is lacking or difficult to provide, since human reference data is not available (Hamm et al., 2017). Currently, available alternative methods include *in vivo* studies which have already been conducted on similar PPPs,

* Corresponding author.
E-mail address: denise.kurth@bfr.bund.de (D. Kurth).

https://doi.org/10.1016/j.yrtph.2018.11.012
Received 15 August 2018; Received in revised form 29 October 2018; Accepted 20 November 2018
Available online 22 November 2018

List of abbreviations

BCOP	Bovine Corneal Opacity and Permeability
BfR	Bundesinstitut für Risikobewertung (German Federal Institute for Risk Assessment)
CLP	Regulation (EC) No 1272/2008 for the classification, labelling and packaging of substances and mixtures
DPRA	Direct Peptide Reactivity Assay
ECHA	European Chemicals Agency
EFSA	European Food Safety Authority
EIT	Reconstructed Human Cornea-Like Epithelium (RhCE) test
EU	European Union
GHS	Globally Harmonised System
GLP	Good Laboratory Practice
GPMT	Guinea Pig Maximisation Test
h-CLAT	Human Cell Line Activation Test
ICE	Isolated Chicken Eye
LC_{50}	Lethal Concentration at which 50% of the test animals died
LD_{50}	Lethal Dose at which 50% of the test animals died
LLNA	Local Lymph Node Assay
M&K	Magnusson and Kligman
OECD	Organisation for Economic Co-operation and Development
PPP	Plant protection product
RAC	Risk Assessment Committee
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
RhCE	Reconstructed Human Cornea-Like Epithelium
RHE	Reconstructed Human Epidermis
TER	Transcutaneous Electrical Resistance
WHO	World Health Organisation
WoE	Weight of evidence

harmonised *in vitro* studies, and the CLP calculation method according to Regulation (EC) No 1272/2008 on the classification, labelling and packaging (CLP) of substances and mixtures (EU, 2008).

The CLP calculation method is based on the additivity principle and assumes a similar mode of action for all ingredients in a mixture (Altenburger et al., 2003; Finney and Tattersfield, 1952; Hoel, 1987). Regardless of dissimilarly acting chemicals present in PPPs, the additivity approach is often accepted as a worst-case estimation of chemical interaction (Backhaus et al., 2000; Backhaus et al., 2004; Junghans et al., 2006). Moreover, it is fast and inexpensive, since it may rely on data collected for the purpose of pesticide authorisation and chemical registration under the REACH Regulation (EU, 2006). Although providing a reasonable worst-case for toxicodynamic interaction, the CLP calculation method does not generally account for toxicokinetic interaction between active substance(s) and co-formulants (Van Cott et al., 2018). It should therefore be carefully assessed for its suitability to predict PPP classification.

Such assessments have been conducted by Van Cott et al. (2018) and Corvaro et al. (2016) for systemic toxicity and by Corvaro et al. (2017) for skin and eye irritation and skin sensitisation. The authors evaluated the CLP calculation method with different datasets and a focus on the underestimation of adverse effect by the CLP calculation method (false negative prediction) as well as the overestimation of adverse effect (false positive prediction). They concluded that while the CLP calculation method may provide appropriate overall accuracy, there is a concern for the underestimation of adverse effects for more toxic products.

In this study, we therefore focused the underestimation of effects by evaluating the prediction of classification for all six endpoints by the CLP calculation method. To specifically assess false negative predictions, we limited our dataset to *in vivo* classified PPPs authorised in Germany. Within the context of PPP hazard assessment according to Regulation (EC) No 1272/2008 (EU, 2008), we discuss available and necessary data for the validation of alternative methods and propose the development of an alternative strategy to address the required endpoints. Although the calculation method is also applied in other regulatory systems, for example the Globally Harmonised System (GHS) (UN, 2017) or the World Health Organisation (WHO) classification system (WHO, 2010), this study focuses on European legislation to conclude on current challenges for the regulatory acceptance of alternative methods and discuss a strategy intended to reduce animal testing in Europe.

2. Materials and methods

2.1. Classification

Throughout this study, the derived classification from CLP calculation method and *in vivo* testing refers to CLP classification categories (EU, 2008) as summarised in Table 1. Classification into sub-categories was not conducted, since they cannot be predicted by the CLP calculation method.

2.2. In vivo methods

No new vertebrate tests were conducted for this study. Instead, available evaluations of tests submitted for the purpose of PPP authorisation to the German Federal Institute for Risk Assessment (BfR) were used or re-evaluated, if the former BfR evaluation was not according to current guidelines. These tests included the acute oral toxicity test in rats according to OECD TG 401 and 423, the acute dermal toxicity test in rats according to OECD TG 402, the acute inhalation toxicity test in rats according to OECD TG 403 and 436, the dermal irritation/corrosion test in rabbits according to OECD TG 404, the eye irritation/corrosion test in rabbits (Draize test) according to OECD TG 405, the skin sensitisation tests in guinea pig by Magnusson and Kligman (M&K or Guinea Pig Maximisation Test, GPMT) and Buehler according to OECD TG 406, and the Local Lymph Node Assay (LLNA) for skin sensitisation in mice according to OECD TG 429 (OECD, 1987, 1992, 2002, 2009a; 2009b, 2010, 2015, 2017a; 2017b).

Table 1
Classification and labelling for acute toxicity, skin irritation/corrosion, eye irritation/damage, and skin sensitisation according to Regulation (EC) No 1272/2008.

	Acute toxicity			Skin irritation	Eye irritation	Skin sensitisation
	Oral	Dermal	Inhalation			
Category 1	H300	H310	H330	H314	H318	H317
Category 2	H300	H310	H330	H315	H319	
Category 3	H301	H311	H331			
Category 4	H302	H312	H332			

2.3. CLP calculation method according to regulation (EC) No 1272/2008

Throughout this document, the term CLP calculation method refers to the additivity approach, a non-testing strategy detailed in CLP guidance for the endpoints acute oral, dermal, and inhalation toxicity and skin and eye irritation. In addition, for skin sensitisation the term is applied to refer to specific or generally applying concentration limits in agreement with CLP guidance (ECHA, 2017).

2.3.1. Acute toxicity

For acute oral, dermal, and inhalation toxicity, the CLP calculation method is based on two different formulas, depending on the cumulative percent of ingredients without reliable toxicity information.

If toxicity information is available for more than 90% of the composition, the following formula derived by Finney and Tattersfield (1952) is to be applied:

$$\frac{100}{ATE_{mix}} = \sum_{n} \frac{C_i}{ATE_i} \quad (1)$$

ATE_{mix} describes the acute toxicity estimate of the mixture, which is composed of n ingredients. C_i is the concentration of ingredient i.

At more or equal to 10% concentration of ingredients of unknown toxicity, the above formula is corrected to:

$$\frac{100 - \sum C_{unknown}\ if > 10\%}{ATE_{mix}} = \sum_{n} \frac{C_i}{ATE_i} \quad (2)$$

2.3.2. Irritation

The calculation of cumulative skin and eye irritation requires the addition of the concentration of ingredients with skin irritating and corrosive or eye irritating and damaging properties.

A mixture is classified as corrosive to the skin, if ≥ 5% of the concentration of its ingredients are Category 1 corrosive substances. It is classified as irritant to the skin, if the sum of the concentrations of Category 2 skin irritants and Category 1 corrosives, corrected by a factor of 10, is ≥ 10%.

Similarly, a mixture is classified as damaging to the eye, if the sum of the concentration of Category 1 skin corroding and Category 1 eye damaging ingredients is ≥ 3%. To assess whether a mixture is irritant to the eye, the concentration of substances, which are Category 1 skin corrosive or eye damaging, is again corrected by a factor of 10. If the sum of this concentration factor and the concentration of Category 2 eye irritants is ≥ 10%, the mixture is classified as Category 2 eye irritant.

2.3.3. Skin sensitisation

In contrast to the other endpoints under evaluation, the concentration of skin sensitising components of a mixture is not added up. Instead, concentration limits apply to single substances. If the mixture contains ≥1% of a Category 1 or 1B or ≥0.1% of a Category 1A skin sensitiser, it is considered a Category 1 skin sensitiser. In deviation to the above limits, some potent chemical sensitisers possess specific concentration limits. For example, the common PPP preservative 1,2-benzisothiazol-3(2H)-one (CAS No 2634-33-5) has a specific concentration limit of 0.05%.

2.4. Data analysis

The dataset is comprised of all PPPs authorised in Germany in July 2017, which were classified for acute oral, dermal or inhalation toxicity, skin or eye irritation or skin sensitisation. Authorised PPPs not classified for one of the above endpoints were not considered for evaluation. Hence, this dataset allows for conclusions on correct positive and false negative while not addressing correct negative predictions. False positive predictions are limited to classification into stricter categories.

To establish the dataset for evaluation, a list of authorised and classified PPPs was generated from ProSafe-Pesticides, a hazard evaluation and documentation tool for PPP assessment, generated by the German Federal Institute for Risk Assessment (BfR). As ProSafe-Pesticides contains information on classification by different European and German authorities as well as manufacturers and applicants, the obtained dataset required refinement. In a first step, the dataset was limited to PPPs classified for the specific endpoints by the German Federal Office of Consumer Protection and Food Safety (BVL), which is the competent risk management authority in Germany. Hence, manufacturer or applicant classification was not considered for this evaluation. Afterwards, PPPs with classifications not derived from laboratories certified for good laboratory practice (GLP) and OECD test guideline compliant vertebrate studies were removed from the dataset. This step was necessary to ensure the comparability of the CLP calculation method to internationally accepted *in vivo* studies. Next, formulations which were not adequate for assessment by the CLP calculation method, such as gas-generating products or combination products which require the mixing of different formulations prior to use were excluded. Finally, PPPs classified based on studies conducted with the formulation under investigation were separated from those classified based on studies conducted with similar formulations. Formulations were considered similar, if no additional toxic, irritant or sensitising properties were expected from the comparable formulation and if composition was not considerably altered according to the EU Guidance document on significant and non-significant changes of the chemical composition of authorised PPPs (EU, 2012). The predictability of the CLP calculation method was assessed using two datasets: one containing only those PPPs for which study and CLP calculation method were assessed with the identical formulation and another which contained all PPPs classified based on *in vivo* testing, including studies conducted with similar formulations.

For PPPs classified for acute oral, dermal, and inhalation toxicity, LD/LC_{50} values were estimated based on the CLP calculation method and *in vivo* studies. For the other endpoints *in vivo* scores and calculated reference contents were evaluated. Classification according to Regulation (EC) No 1272/2008 (EU, 2008) was derived from *in vivo* and calculated toxicity scores and compared. Sub categorisation – if available for the specific endpoint – was not conducted. In case of identical classification, predictions were considered correct positive. If classification based on *in vivo* testing was less severe compared to the CLP calculation method, predictions were termed false positive. In turn, false negative predictions lead to less severe classification by the CLP calculation method as compared to *in vivo* testing.

For active substances, the following data was taken into account for calculation in that specific order:

a) *In vivo* LD/LC_{50} data from the European Food Safety Authority (EFSA) conclusion (acute toxicity)
b) Route-to-route extrapolation of *in vivo* LD_{50} data from EFSA conclusion (acute toxicity)
c) Legal classification by the European Chemicals Agency (ECHA) (all endpoints)
d) Classification proposal by EFSA (all endpoints)
e) Route-to-route extrapolation of ECHA or EFSA classification (acute toxicity)

For co-formulants, the following information was considered:

a) *In vivo* LD/LC_{50} data from material safety data sheets (MSDS) for the endpoint under evaluation (acute toxicity)
b) *In vivo* LD/LC_{50} data from REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) registration dossiers for the endpoint under evaluation (acute toxicity)
c) Route-to-route extrapolation of *in vivo* LD_{50} data from MSDS (acute

toxicity)
d) Legal classification by the European Chemicals Agency (ECHA) (all endpoints)
e) Manufacturer self-classification (all endpoints)
f) Route-to-route extrapolation of ECHA or manufacturer classification (acute toxicity)
g) Classification derived from studies described in REACH registration dossiers (all endpoints)
h) Majority (≥50%) ECHA notifier classification (all endpoints)

It should be noted that MSDS effect concentrations were only taken into account, if they were derived from *in vivo* testing and conducted with the specific substance. Read-across as well as information on "slight irritation" not resulting in classification was not considered. In case of multiple lines of evidence for a substance from one group of sources, e.g. LD_{50} data from several MSDSs by different manufacturers, the most plausible or conservative value was taken into account based on expert judgement. This step could have resulted in the underestimation of false negative classification in rare cases. If available, information on co-formulant mixtures was preferred over single substance information. However, in case co-formulant mixture information was not in accordance with single substance information, expert judgment was applied. Furthermore, expert judgement was also applied, when data from different sources (a) to i)) was considered inconsistent or implausible.

3. Results

Throughout this paper, all values refer to the comparison of *in vivo* test results with the calculation method for identical formulations. Tables are provided for both datasets, including and excluding results obtained from *in vivo* tests conducted with similar PPPs.

3.1. Dataset characterisation

PPP studies were evaluated for more than 900 product-endpoint combinations (see Table 2). After the removal of entries from the dataset, which did not support the comparison of the calculation with *in vivo* methods (as described in Section 2.4), the remaining PPPs were evaluated with the CLP calculation method.

Based on *in vivo* testing for acute oral and inhalation toxicity, approximately 90% of all PPPs were classified into Category 4 (H302 or H332). When the CLP calculation method was applied to the dataset, about 50% of all PPPs were classified into Category 4 (H302 or H332) while roughly 40% were not classified (see Fig. 1). Acute dermal toxicity could not be evaluated based on the limited number of classified products for this endpoint. In total, 13 PPPs authorised in Germany are classified for acute dermal toxicity based on *in vivo* testing. Of those, eleven PPPs were classified based on active substance toxicity studies. In addition, they included many gas-generating products, which are not considered within the applicability domain of the CLP calculation method.

For skin irritation, all PPPs classified based on *in vivo* studies were Category 2 (H315) irritants, as depicted in Fig. 2. According to the CLP calculation method, the same products lead to Category 2 classification in 70% of the cases. Approximately 20% of PPPs were classified as non-irritating based on the CLP calculation method, while almost 10% were classified as Category 1 (H314) corrosives. With regard to eye irritation, 40% of PPPs were eye damaging (Category 1, H318) based on *in vivo* testing. In comparison, the CLP calculation method overestimated the number of corrosive PPPs, as two thirds of all PPPs in the dataset were attributed to Category 1. For this endpoint, the CLP calculation method resulted in few non-classified PPPs.

With regard to skin sensitisation, all *in vivo* tests had to result in Category 1 classification (H317, see Fig. 3). In total, approximately 35% of PPPs were not classified by the CLP calculation method.

3.2. CLP calculation method evaluation

3.2.1. Acute toxicity

When classification by *in vivo* testing and the CLP calculation method was compared, acute oral and inhalation toxicity displayed an almost identical pattern. Approximately 45% of PPPs classified for those endpoints were predicted into the same classification category by the CLP calculation method as by *in vivo* testing (see Fig. 4). Nonetheless, an almost equal amount of PPPs was attributed to a classification category of lower toxicity, thus underestimating *in vivo* toxicity. The majority of false negative predictions for oral and inhalation toxicity were attributed to Category 4 (H302 or H332) by *in vivo* testing, whereas the CLP calculation method resulted in no classification (see Table 3 and Table 4). Hence, the CLP calculation method could not reliably differentiate between Category 4 and non-toxic PPPs in this dataset. Since few products in the dataset were classified into Categories 1–3 by *in vivo* testing for acute oral and inhalation toxicity, no reliable conclusions can be drawn on the ability of the CLP calculation method to predict true positive values for highly toxic products. However, with regard to acute systemic toxicity it can be summarised that classification into Categories 1–3 is required for < 10% of classified PPPs authorised in Germany.

Due to the acceptance of multiple data sources of varying resolution and reliability (please see Section 2.4), the percentage of substances of unknown toxicity was rather low. While correction for unknown toxicity according to Equation (2) was not necessary for oral toxicity, correction was conducted for 8 of 38 PPPs for inhalation toxicity. Mean (and median) percentage of unknown toxicity was 0.6% (0%) and 5% (1.5%) for oral and inhalation toxicity, respectively.

Table 2
Overview of dataset for evaluation. The number of registered and classified PPPs was extracted from the ProSafe Pesticides database. The dataset was refined by removing entries which did not support CLP calculation method evaluation. It was differentiated, whether the *in vivo* study triggering classification was conducted with the PPP formulation currently registered in Germany or with a similar or former formulation.

	Oral toxicity	Dermal toxicity	Inhalation toxicity	Skin irritation	Eye irritation	Skin sensitisation
# registered, classified PPPs	167	19	151	120	242	242
Removed entries[a]	43	17	75	53	74	90
Similar PPP[b]	29	1	38	35	78	64
Identical PPP[c]	95	1	38	32	90	88

[a] Database classification not based on authority proposal for classification under Regulation (EC) 1272/2008 or classification not based on *in vivo* study, but e.g. on the CLP calculation method, *in vitro* studies or critical ingredients of concern.
[b] Animal study has been conducted with a similar formulation, which has been evaluated to be representative of the classification of the product under evaluation, i.e. composition as currently registered in Germany.
[c] Animal study and CLP calculation method were conducted with identical formulations.

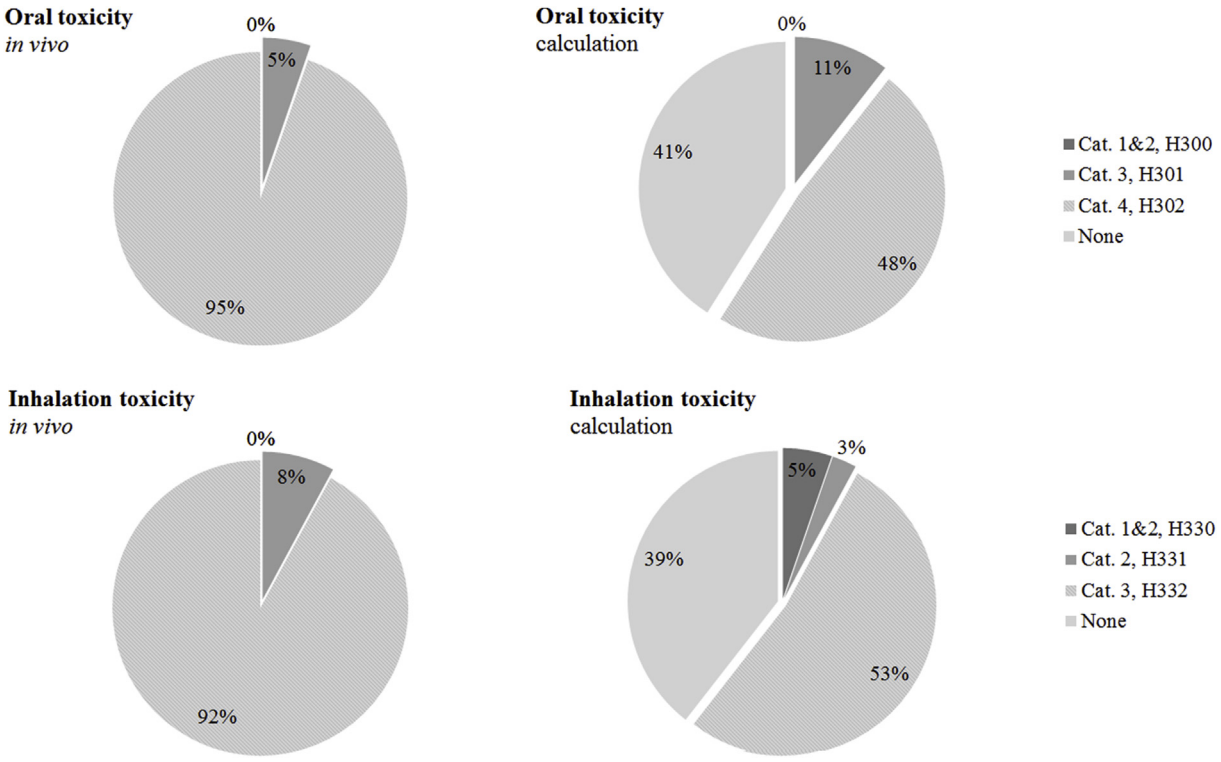

Fig. 1. Distribution of the derived PPP classification for acute oral (n = 95) and inhalation toxicity (n = 38) based on *in vivo* testing and the CLP calculation method (identical formulations).

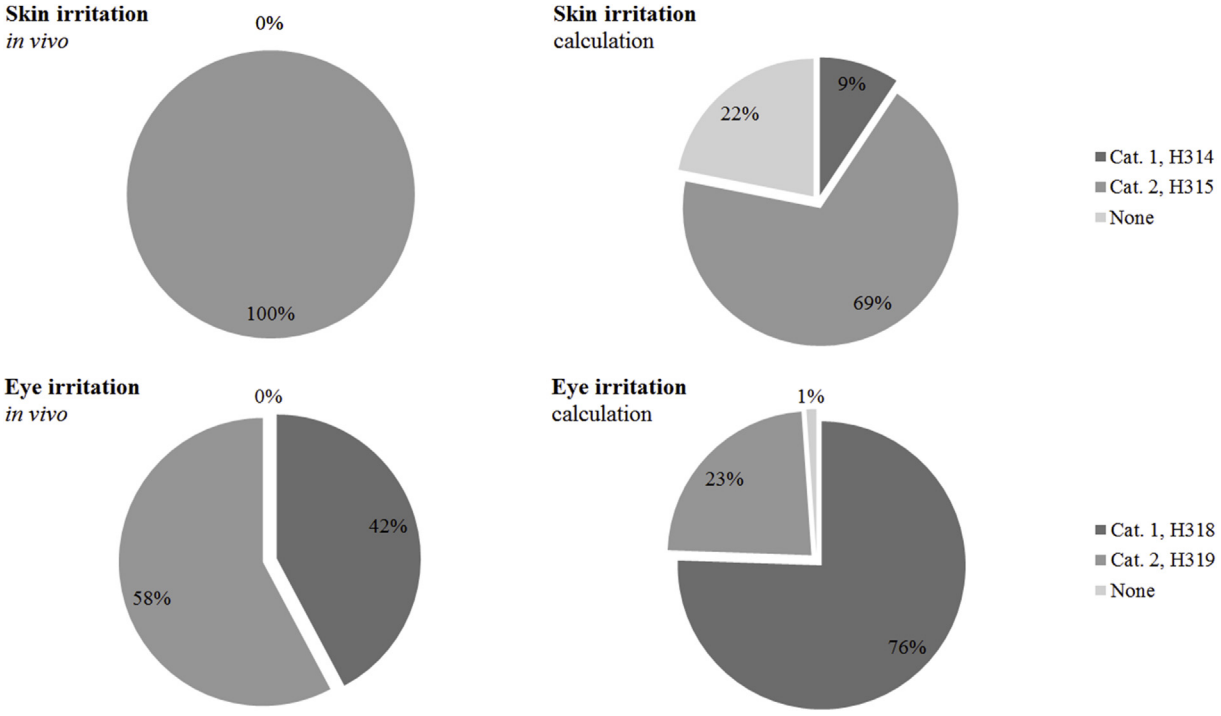

Fig. 2. Distribution of the derived PPP classification for skin (n = 32) and eye (n = 90) irritation based on *in vivo* testing and the CLP calculation method (identical formulations).

3.2.2. *Irritation*

The highest number of accurate predictions by the CLP calculation method was observed for skin irritation (see Fig. 5). Although approximately 70% of the products were accurately classified by the CLP calculation method when compared to *in vivo* testing, about one fourth of the products were still underestimated with regard to their skin irritating potential. Few products in the dataset were classified as Category 1 corrosives (see Table 5). The major source of false classification was the underestimation of Category 2 irritants. Hence, the CLP calculation method cannot be used to reliably differentiate between products requiring or not requiring classification for skin irritation.

In contrast, the CLP calculation method is able to accurately classify

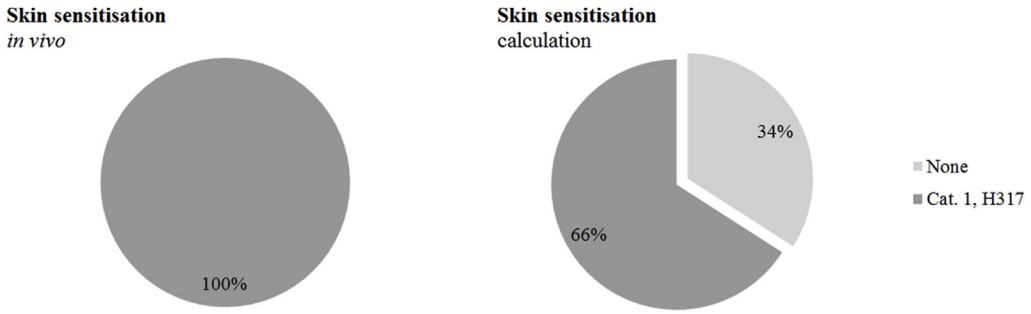

Fig. 3. Distribution of the derived PPP classification for skin sensitisation (n = 88) based on *in vivo* testing and the CLP calculation method (identical formulations).

in vivo Category 1 eye damaging PPPs. However, many *in vivo* irritant PPPs are also inadequately classified into Category 1 by the CLP calculation method (see Fig. 5 and Table 6). Due to the high false positive rate of 40%, the CLP calculation method is prone to overestimating the potential for eye damage (Category 1).

3.2.3. Skin sensitisation

For skin sensitisation, several OECD TGs are available: the LLNA, the M&K (or GPMT) test, and the 3-induction and 9-induction Buehler tests, which were summarised for this assessment. For seven of 88 PPPs, multiple tests were available. When LLNA and M&K results were compared, both methods resulted in classification for two of three products. For PPPs with information on M&K and Buehler tests, three of four were classified based on M&K while two of four were classified based on Buehler tests. In total, classification derived from different test systems for the same PPP only matched in three of seven cases (40%).

Due to the high variability of *in vivo* results from different test systems, CLP calculation method results were displayed for each test system and in total (see Table 7). Since the CLP calculation method does not distinguish between Category 1A and 1B sensitisers, it can only predict classification or no classification. Overall, calculation and *in vivo* methods resulted in different classification for approximately 35% of PPPs in this dataset with slightly higher agreement between the CLP calculation method and Buehler tests compared to the other test methods.

Classification according to the CLP calculation method for skin sensitisation is based on specific or generally applying concentration limits of single substances. For the majority of PPPs in this dataset, active substances were the drivers for classification (see Fig. 6, left). The 30 PPPs (34%) not classified by the CLP calculation method can be separated into 13% not containing any known sensitisers and 21% containing sensitisers below their (specific) concentration limits. As depicted in Fig. 6 (right), 39% of all PPPs were composed of equal to or less than 10% known sensitisers, while 29% and 23% contained equal to or less than 1% and 0.1% known sensitisers, respectively. In sum, half the products in the dataset were composed of approximately 30% known sensitisers. Their incidence in one PPP in this dataset ranged from zero to seven with a mean and median of 2.

4. Discussion

4.1. Acute toxicity

Comparing *in vivo* study and CLP calculation method-derived classification, we observed a high rate of false negative predictions, i.e. the underestimation of acute *in vivo* toxicity classification, by the CLP calculation method. While our dataset was limited to PPPs classified for the specific endpoints based on *in vivo* testing, other studies were published which focused on datasets including *in vivo* non-classified PPPs. For example, Corvaro et al. (2016) evaluated a dataset of 213 PPPs for acute oral toxicity, which contained only 20% classified PPPs based on *in vivo* testing. In contrast, Van Cott et al. (2018) published an evaluation of the acute oral toxicity CLP calculation method with a dataset of 210 PPPs, 70% of which were classified based on *in vivo* testing. Reducing the mentioned datasets to positive *in vivo* results only, we may compare the results of this study to published literature (Table 8). At false negative predictions ranging from 38 to 54%, all studies indicate the substantial underestimation of PPP acute oral toxicity by the CLP calculation method. When the entire datasets in the studies by Corvaro et al. (2016) and Van Cott et al. (2018) are considered, acute oral toxicity false negative prediction amounts to 8% and 37%, respectively. This diverging outcome has already been discussed by Van Cott et al. (2018), who attributed the deviation to the higher number of *in vivo* non-classified products investigated by Corvaro et al. (2016). Although false positive prediction, i.e. the overestimation of toxicity resulting in more severe classification and labelling, can be problematic for the risk perception of users, it occurs more rarely. In contrast, false negative prediction appears frequently and may reduce the awareness of users for possible hazards while leading to inadequate protection measures. Hence, further efforts should be focused on the

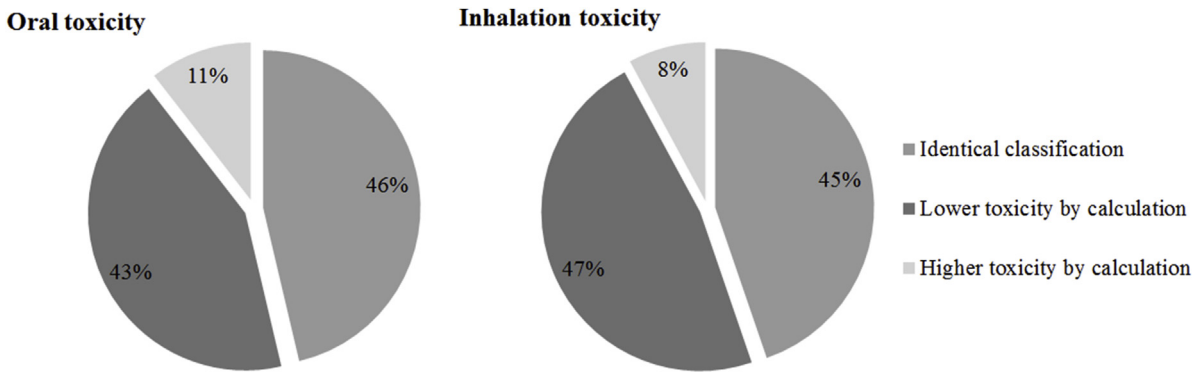

Fig. 4. Comparison of *in vivo* and calculation-derived classification for acute oral and inhalation toxicity (identical PPPs, oral toxicity: n = 95, inhalation toxicity: n = 38).

Table 3
Contingency table for the *in vivo* and CLP calculation method derived classification of acute oral toxicity.

Identical PPPs		Calculation					Sum
		Category 1	Category 2	Category 3	Category 4	No class.	
In vivo	Category 1	**0**	0	0	0	0	0
	Category 2	0	**0**	0	0	0	0
	Category 3	0	0	**0**	2	3	5
	Category 4	0	0	10	**44**	36	90
Sum		0	0	10	46	39	95

Incl. similar PPPs		Calculation					Sum
		Category 1	Category 2	Category 3	Category 4	No class.	
In vivo	Category 1	**0**	0	0	0	0	0
	Category 2	0	**1**	0	0	0	1
	Category 3	0	0	**1**	4	3	8
	Category 4	0	0	15	**57**	43	115
Sum		0	0	16	61	46	124

minimisation of false negative prediction for acute oral and inhalation toxicity.

In this study and in others, it was demonstrated that the majority of positive findings for acute oral and inhalation toxicity were attributed to Category 4 (H302 or H332) by *in vivo* testing. In result, the majority of false negative predictions were associated with the failure to discriminate between Category 4 and no classification (see Tables 3 and 4). This was considered to be of minor importance by Corvaro et al. (2016), as oral exposure is not an intended route of exposure. The authors furthermore argued that neither does classification for acute oral toxicity trigger the use of additional protective measures, nor does knowledge on the LD_{50} in rats provide medically relevant information for humans in case of intended or unintended oral exposure to PPP due to high inter-species variability. Moreover, for both the oral and inhalation route, the authors pointed out that exposure corresponding to the classification range of Category 4 would be unrealistically high. Hence, they proposed to evaluate the likelihood and relevance of exposure for hazard assessment considerations. Although those arguments can be validly discussed on a scientific basis, the current European regulatory framework does require the adequate characterisation of hazard; even at unrealistically high exposure conditions (EU, 2008). On the one hand, this requirement was instituted to identify interaction between active substances and co-formulants. On the other hand, it informs users of chemical products about potential risks and ensures their cautious use and handling. Additionally, in the case of inhalation toxicity, classification with Category 4 can lead to the implementation of additional safety measures. Thus, the adequate characterisation of hazard requires accurate toxicity prediction. In view of the presented results in this study and based on the findings by Corvaro et al. (2016) and Van Cott et al. (2018), we conclude that the CLP calculation method does not limit false negative prediction sufficiently for PPP regulatory purposes.

In addition to the CLP calculation method, bridging to similar formulations with adequate data available or *in vitro* testing can be conducted as alternatives to vertebrate testing. Notably, there currently exist no harmonised *in vitro* methods for acute toxicity.

Table 4
Contingency table for the *in vivo* and CLP calculation method derived classification of acute inhalation toxicity.

Identical PPPs		Calculation					Sum
		Category 1	Category 2	Category 3	Category 4	No class.	
In vivo	Category 1	**0**	0	0	0	0	0
	Category 2	0	**0**	0	0	0	0
	Category 3	0	0	**0**	3	0	3
	Category 4	0	2	1	**17**	15	35
Sum		0	2	1	20	15	38

Incl. similar PPPs		Calculation					Sum
		Category 1	Category 2	Category 3	Category 4	No class.	
In vivo	Category 1	**0**	0	0	0	0	0
	Category 2	0	**0**	0	0	0	0
	Category 3	0	0	**2**	5	0	7
	Category 4	0	3	6	**33**	27	69
Sum		0	3	8	38	27	76

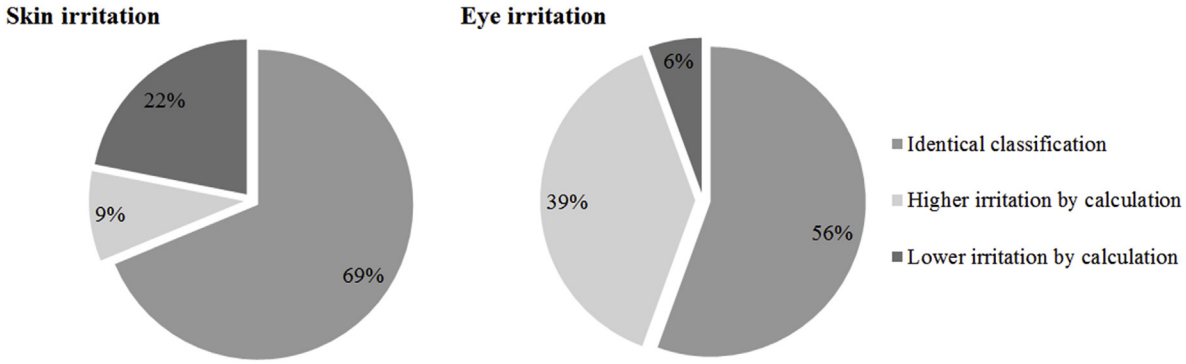

Fig. 5. Comparison of *in vivo* and calculation-derived classification for skin and eye irritation (identical PPPs, skin irritation: n = 32, eye irritation: n = 90).

Table 5
Contingency table for the *in vivo* and CLP calculation method derived classification of skin irritation.

Identical PPPs		Calculation			Sum
		Category 1	Category 2	No class.	
In vivo	Category 1	**0**	0	0	0
	Category 2	3	**22**	7	32
Sum		3	22	7	32

Incl. similar PPPs		Calculation			Sum
		Category 1	Category 2	No class.	
In vivo	Category 1	**0**	3	0	3
	Category 2	5	**45**	12	62
Sum		5	48	12	65

Table 6
Contingency table for the *in vivo* and CLP calculation method derived classification of eye irritation.

Identical PPPs		Calculation			Sum
		Category 1	Category 2	No class.	
In vivo	Category 1	**34**	4	0	38
	Category 2	34	**17**	1	52
Sum		68	21	1	90

Incl. similar PPPs		Calculation			Sum
		Category 1	Category 2	No class.	
In vivo	Category 1	**61**	8	0	69
	Category 2	66	**29**	4	99
Sum		127	37	4	168

Table 7
Contingency table for the *in vivo* and CLP calculation method derived classification of skin sensitisation with different test systems.

Identical PPPs		Calculation		Sum
In vivo		Category 1	No class.	
Cat. 1	LLNA	**22 (65%)**	12 (35%)	34
	M&K	**22 (58%)**	16 (42%)	38
	Buehler	**14 (74%)**	5 (26%)	19
	All data	**58 (66%)**	30 (34%)	88/91[a]

Incl. similar PPPs		Calculation		Sum
In vivo		Category 1	No class.	
Cat. 1	LLNA	**32 (70%)**	14 (30%)	46
	M&K	**49 (68%)**	23 (32%)	72
	Buehler	**31 (78%)**	9 (22%)	40
	All data	**110 (72%)**	42 (28%)	152/158[b]

[a] For three PPPs positive *in vivo* tests with two different test systems were available.

[b] For six PPPs positive *in vivo* tests with two or more different test systems were available.

4.2. Irritation

4.2.1. CLP calculation method

In this study the CLP calculation method resulted in the highest number of correct *in vivo* predictions (almost 70%) for skin irritation while false negative prediction was rather high at more than 20% of all *in vivo* classified products (see Table 9). In a similar study by Corvaro et al. (2017) analysed the CLP calculation method for skin and eye irritation using a dataset consisting of both classified and non-classified products. When this dataset is reduced to classified products only, it may be compared to the data in this study (Table 9). With regard to skin irritation, the authors reported a contrasting pattern of accurate and false prediction by the CLP calculation method for skin irritation. While they arrived at approximately one third accurate and two thirds false negative predictions, the pattern is almost exactly reversed in this study. False positive predictions however were low in this study as well as the study by Corvaro et al. (2017). This deviating pattern could indicate the uncertainty of predictions by the CLP calculation method or the high variability of *in vivo* testing.

Better agreement between the *in vivo* classified sub-dataset of the study by Corvaro et al. (2017) and this study was achieved for eye irritation. The high number of false positive and the comparably low number of false negative predictions indicates that the CLP calculation method is more conservative for this endpoint. As summarised in Table 6, products which were classified by *in vivo* testing as eye damaging in Category 1 were well detected with the CLP calculation method. In addition, within this dataset the CLP calculation method could be successfully applied to identify the need for classification, but did not reliably distinguish between irritant and corrosive (Category 1 and 2) products. However, as worst-case, Category 1 classification could be an acceptable application of the precautionary principle.

While for the acute toxicity endpoints *in vivo* and calculated LD/LC_{50} values can be compared, skin and eye irritation classification categories are derived from *in vivo* test scores and calculated cumulative substance content. An assessment of the CLP calculation method can

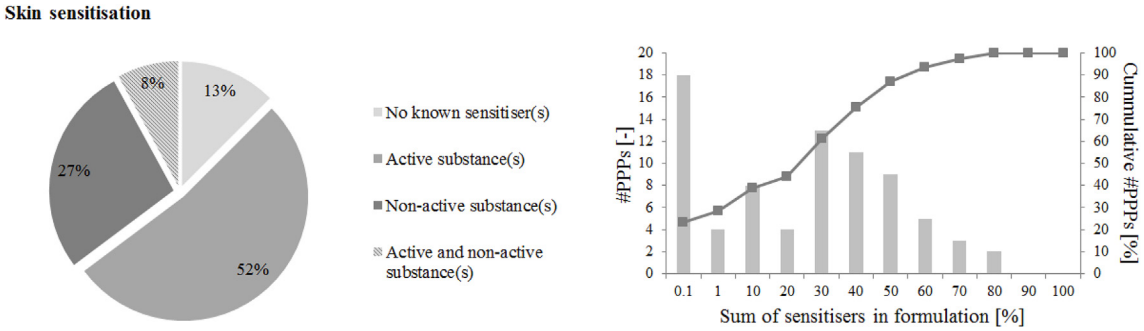

Fig. 6. Left: Known sensitisers in PPPs triggering classification based on their (specific) concentration limits (identical PPPs, n = 88); Right: Histogram of additive contents of sensitisers in PPPs classified for skin sensitisation and containing sensitising components in percent (light grey bars, identical PPPs, n = 77). Bins on the x-axis range from greater than the lower limit to equal to or less than the upper limit. The dark grey line depicts the cumulative percent of PPPs in each bin.

Table 8
Comparison of CLP calculation method predicitivity in different studies restricted to (sub-)datasets of PPPs classified for acute oral or inhalation toxicity based on *in vivo* testing.

	Corvaro et al. (2016)	Van Cott et al. (2018)	This study
Oral toxicity	n = 45	n = 145	n = 95
Accurate	58%	45%	46%
False negative	38%	54%	43%
False positive	4%	1%	11%
Inhalation toxicity	n = 9	n = 69	n = 38
Accurate	33%	61%	45%
False negative	44%	19%	47%
False positive	22%	20%	8%

Table 9
Comparison of CLP calculation method predicitivity in different studies only including (sub-)datasets of PPPs classified for skin or eye irritation based on *in vivo* testing.

	Corvaro et al. (2017)	This study
Skin irritation	n = 42	n = 32
Accurate	26%	69%
False negative	74%	22%
False positive	0%	9%
Eye irritation	n = 98	n = 90
Accurate	38%	56%
False negative	15%	6%
False positive	47%	39%

therefore be considered as an assessment of how well the defined scores match. Although they are based on expert knowledge, the number of false negatives may be reduced by either shifting the scores for *in vivo* or for calculation-derived classification. However, more conservative classification thresholds for the CLP calculation method could potentially not result in more correct, but rather more false positive predictions. If necessary, the limits should therefore be optimised to promote clear differentiation between classified (Category 1 and 2) and non-classified products.

4.2.2. Other alternatives to in vivo testing

In addition to *in vivo* testing and the calculation method, bridging to similar formulations for which *in vivo* data is available can be applied based on expert judgement. Furthermore, for skin and eye irritation harmonised *in vitro* methods exist (see Table 10). Several studies were conducted to verify skin and eye irritation *in vitro* methods for PPPs (Kolle et al., 2012, 2015, 2017a, 2017b; Schrage et al., 2010).

For skin irritation, a pioneer assessment was conducted by Kolle et al. (2012), who compared *in vivo* and *in vitro* derived classification for 134 substances and mixtures according to OECD TG 431 and for 38 substances and mixtures according to OECD TG 439 with reconstructed human epidermis. The datasets contained 46 and 16 PPPs, respectively. OECD TG 431 is designed to differentiate between corrosive and non-corrosive substances and mixtures. For this test, the authors observed an accuracy of 91% in the subset of PPPs only, and 88% for the dataset excluding PPPs (n = 88). However, none of the tested formulations was classified for skin corrosion based on *in vivo* testing. Thus, good accuracy was only demonstrated for the correct identification of non-corrosive PPPs, while there is still a lack of data for corrosive PPPs. Complementary to OECD TG 431, the requirement for classification versus no classification can be tested following OECD TG 439. For this test, the authors found a low accuracy of 56% with an almost equal number of false positive and negative predictions. Notably, the authors observed better accuracy (68%) in the dataset excluding PPPs. Those findings were later confirmed by Kolle et al. (2017b) with a larger dataset of 81 PPPs (OECD TG 431) and 55 PPPs (OECD TG 439) for the two tests, respectively. Hence, while OECD TG 431 was demonstrated to perform well for non-corrosive PPPs, OECD TG 439 cannot be used to reliably predict PPP *in vivo* skin irritation. So-far, no evaluation with PPPs was published for other harmonised *in vitro* methods.

With regard to the *in vitro* eye irritation methods according to OECD TG 437 (Bovine Corneal Opacity and Permeability Test, BCOP) and 492 (Reconstructed Human Cornea-Like Epithelium (RhCE) test, EIT), verification has been conducted by Kolle et al. (2015) with 97 PPPs. The BCOP test can identify substances and mixtures to be classified as eye

Table 10
Overview of existing *in vitro* methods for skin and eye irritation and skin sensitisation.

Endpoint	Test system	Test guideline
Skin irritation	Transcutaneous Electrical Resistance (TER) Test Method	OECD TG 430
	Reconstructed Human Epidermis (RHE) test	OECD TG 431, OECD TG 439
	In Vitro Membrane Barrier Test Method for Skin Corrosion	OECD TG 435
Eye irritation	Bovine Corneal Opacity and Permeability (BCOP) Test	OECD TG 437
	Isolated Chicken Eye (ICE) Test	OECD TG 438
	Fluorescein Leakage Test Method	OECD TG 460
	Short Time Exposure *In Vitro* Test	OECD TG 491
	Reconstructed Human Cornea-Like Epithelium (RhCE) test (EIT)	OECD TG 492
Skin sensitisation	Direct Peptide Reactivity Assay (DPRA)	OECD TG 442C
	ARE-Nrf2 Luciferase Test Method	OECD TG 442D
	Human Cell Line Activation Test (h-CLAT)	OECD TG 442E

damaging (Category 1) as well as those not requiring classification. In addition, it may result in further testing, if no prediction can be made for a substance or mixture. According to Kolle et al. (2015), 76% of the PPPs for which prediction was possible (n = 45) were predicted accurately while 16% were falsely classified as non-irritant. The EIT is able to identify substances and mixtures not to be classified for eye irritation. Like the BCOP, it may result in further testing, if prediction is not possible. Kolle et al. (2015) observed that 86% of the PPPs for which a prediction could be made (n = 36) were accurately predicted as non-irritants, while 14% were falsely attributed to the no classification category. In general, the number of PPPs attributed to the category for which prediction by the *in vitro* tests was not possible, ranged from 54 to 63% in this study and 73–90% in a subsequent study by Kolle et al. (2017a). It should be stressed, that a result in this category is intended to trigger further testing and does not exclude either Category 1 or no classification. Therefore, we propose to disregard this category for the validation of eye irritation *in vitro* methods, as the implied rate of false negative or positive predictions would otherwise be distorted. In addition to the BCOP test and the EIT, Kolle et al. (2017a) also evaluated the comparability of *in vitro* to *in vivo* classification by the isolated chicken eye test (ICE) according to OECD TG 438. However, due to the low number of PPPs for which prediction was possible (1 of 10 PPPs), no more detailed conclusion can be drawn from the study with regard to the accuracy of the methods for PPPs.

4.3. Skin sensitisation

For this endpoint, several *in vivo* test systems are available, which have been widely applied in the past. Overall, approximately 65% of PPPs were classified by the CLP calculation method in agreement with *in vivo* testing. In the study by Corvaro et al. (2017) classification between the CLP calculation method and *in vivo* tests corresponded for 54% of PPPs, when *in vivo* classified products were considered only.

In addition to the CLP calculation method, bridging and *in vitro* testing may provide alternative or complementary information. It should be noted, however, that *in vitro* testing for skin sensitisation may only show molecular or cellular effects, while not addressing the impact on the organ level (OECD, 2014). Up to now, the *in vitro* skin sensitisation of PPPs was only verified by Settivari et al. (2015), who compared the KeratinoSens™ results of 10 PPPs to two Buehler, three M&K, and five LLNA tests. Based on this limited dataset, good agreement between *in vitro* and *in vivo* skin sensitisation was observed. However, a larger dataset will need to be compared to each *in vivo* method to draw conclusions on the applicability of the KeratinoSens™ assay to PPPs. Moreover, similar information will be required on other skin sensitisation *in vitro* assays and the key events they address.

4.4. Reasons for false negative prediction by the CLP calculation method

A thorough discussion of the causes for false prediction by the CLP calculation method was already provided by Van Cott et al. (2018). Briefly, the authors name (1) the inadequate application of the additivity principle for PPPs due to expected dissimilar modes of action, (2) the neglection of the influence of co-formulants on the bioavailability of toxic substances, (3) the integration of data of different resolution or quality (e.g. LD_{50} values determined by an acute toxic class method as compared to dose-response modelling or the use of classification instead of effect data), and (4) the combination of data generated with test vehicles which do not match the PPP under evaluation.

Van Cott et al. (2018) argued, that especially formulation parameters, which influence PPP toxicokinetics are commonly disregarded when applying the CLP calculation method and are thus a plausible cause for false prediction. In consequence, the authors proposed the use of formulation-type specific information for toxicity prediction by the CLP calculation method. In practice, this could be challenging to implement, since available toxicity data, for example from Material Safety Data Sheets (MSDS), does not commonly entail the required test formulation information. Moreover, adequate data would have to be generated, leading to additional testing.

Another relevant reason for false prediction by the CLP calculation method is the comparison to results from *in vivo* studies, which are known to be subject to substantial inter- and intra-species variability (Hamm et al., 2017; Sewell et al., 2017). This is demonstrated by the diverging LD_{50} values of active substances tested in different test vehicles (Van Cott et al., 2018) or by the variability of different *in vivo* skin sensitisation tests in this study. Comparing different *in vivo* test systems for seven PPPs, we derived corresponding classifications in only 40%. Corvaro et al. (2017) compared Buehler and LLNA results for five PPPs and reported that for three of five products both tests lead to the same classification. In addition to inter-species variability, possible reasons for the observed variability between the different test systems include differences in test substance application and various key events addressed by the methods. Hence, especially for skin sensitisation, the comparison of alternative methods and *in vivo* testing should be interpreted carefully. Moreover, with regard to inhalation toxicity, the comparison of the *in vivo* tested with the nominal formulation could lead to false predictions, due to the potential relative change in composition during aerosol formation. Finally, since the evaluated OECD *in vivo* methods have never been validated for their prediction of human toxicity – with the exception of the LLNA, uncertainty remains concerning the relevance of *in vivo* testing for human toxicity (Hamm et al., 2017). It is therefore important to note that the accuracy of individual alternative methods is limited by *in vivo* variability and reproducibility (Hoffmann et al., 2010).

In combination, the above causes may lead to variability and uncertainty in the estimation of classification by the CLP calculation method. Hence, they should be accounted for when applying this method in regulatory hazard and risk assessment.

4.5. The weight of evidence approach

Based on the discussed information, neither the CLP calculation method nor available *in vitro* methods can currently replace PPP vertebrate *in vivo* testing as a single line of evidence while ensuring the maintenance of health and safety standards. Instead, the need for integrated testing strategies has been expressed in addition to the expert-based consideration of all available data in a weight of evidence (WoE) approach (Corvaro et al., 2017; Hamm et al., 2017; Jaworska et al., 2015; Scott et al., 2010).

The WoE approach is defined by the European Food Safety Authority (EFSA) Scientific Committee as a process in which evidence is integrated to determine its relative support for a hypothesis. The approach comprises three basic steps: assembling, weighing, and integrating evidence. The final assessment shall consider uncertainty, report the identified sources of uncertainty, and conclude on their impact on the supported outcome (EFSA, 2017).

The advantages of the WoE approach for PPP hazard assessment include the improvement of health and safety standards, the flexibility to account for future advances in the field of alternative methods, and the effortless adaptation of implemented relevance and reliability criteria, when new data or information becomes available.

For PPPs, relevant data to be assembled, weighed and integrated may include

- *in vivo* test data with PPP formulations conducted prior to the implementation of Regulation (EC) No 1107/2009 or in the context of non-EU regulatory frameworks, including studies conducted with comparable PPP formulations meeting the above criteria,
- *in vitro* studies performed with the PPP formulation for authorisation in the context of EU and non-EU regulatory frameworks,
- *in silico* calculations, including the CLP calculation method,
- information on the acute toxic, irritating, or sensitising properties of

all PPP ingredients from all appropriate sources, escpecially legal, manufacturer, and ECHA notifier information as well as published Risk Assessment Committee (RAC) opinions and REACH registration dossiers,
- information on the function of components in the product and their possible interaction with the active substance(s), and
- information on relevant physico-chemical properties.

All evidence should be provided by applicants and evaluated by regulatory experts. However, in order to weigh the individual pieces of evidence adequately, regulatory experts need to possess sufficient information on the influence of formulation parameters on the accuracy and uncertainty of different levels of information.

Since quantitative information on the uncertainty of alternative methods for PPPs is still missing, WoE assessment is currently restricted to qualitative expert judgement. In order to establish quantitative WoE assessment, adequate information on the uncertainty of each method must be generated. Such information may be transferred into quantitative reliability criteria, comparable to the Klimisch or Kaltenhäuser criteria (Kaltenhäuser et al., 2017; Klimisch et al., 1997). In effect, models may be developed, which predict hazard-based classification for PPPs by weighing multiple lines of evidence based on their relevance and reliability (Roth and Ciffroy, 2016; Suciu et al., 2018).

5. Conclusion

The findings of this study and the few available other studies on the evaluation of the CLP calculation method for PPP toxicity assessment indicate that the CLP calculation method may not be suitable to replace vertebrate testing as a single line of evidence. The focus of this initial study was on the identification of the underestimation of PPP toxicity. Therefore, the dataset was limited to PPPs authorised in Germany and classified for the specific endpoint under evaluation based on *in vivo* testing.

In comparison to the respective *in vivo* studies, the CLP calculation method resulted in the substantial underestimation of toxicity, irritation or sensitisation for almost all endpoints. At approximately 45% false negative predictions, insufficient agreement between *in vivo* and calculation-derived classification was observed for oral and inhalation toxicity. Better agreement was noted for skin irritation with almost 70% correspondence between the methods. Although the number of false negative predictions was comparably low at roughly 5% for eye irritation, a high rate of false positive Category 1 predictions was detected.

In order to implement the CLP calculation method into PPP assessment, its reliability needs to be improved. Hence, formulation parameters need to be identified which influence its uncertainty. Those may for example include physico-chemical properties or specific groups of co-formulants, which influence the toxicodynamic and toxicokinetic properties of PPPs.

To overcome the limitations of individual alternative methods for PPPs and effectively reduce the use of animals for regulatory hazard assessment, we propose that regulators should rely on multiple lines of evidence. The establishment of a WoE approach requires the initial verification of all suitable alternative methods for PPPs. Applicability domains and relevant parameters influencing the outcome and accuracy of alternative methods need to be identified. In a first step, relevant parameters influencing the false negative predictivity of the CLP calculation method need to be extracted for all six endpoints. Next, the verification of *in vitro* methods for plant protection products and the identification of parameters, influencing their reliability and relevance for PPP toxicity assessment are required. Finally, computer-based tools relying on established scientific data may be developed to quantify and harmonise WoE hazard assessment for PPPs and to ensure a high level of protection for human, animal and environmental health in the future. WoE predictions should be accompanied with uncertainty estimations. In case of high uncertainty, further testing could be initiated, specifically targeting the identified sources of uncertainty. Consequently, animal testing should be strictly conducted as a last resort only.

Acknowledgements

The authors wish to thank Dr. Roland Solecki for valuable feedback, Phuong-Chi Ly for technical assistance and five anonymous reviewers for their constructive and much valued comments.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Transparency document

Transparency document related to this article can be found online at https://doi.org/10.1016/j.yrtph.2018.11.012.

References

Altenburger, R., Nendza, M., Schuurmann, G., 2003. Mixture toxicity and its modeling by quantitative structure-activity relationships. Environ. Toxicol. Chem. 22 (8), 1900–1915. https://doi.org/10.1897/01-386.

Backhaus, T., Altenburger, R., Boedeker, W., Faust, M., Scholze, M., Grimme, L.H., 2000. Predictability of the toxicity of a multiple mixture of dissimilarly acting chemicals to Vibrio fischeri. Environ. Toxicol. Chem. *19* (9), 2348–2356. https://doi.org/10.1897/1551-5028(2000)019<2348:Pottoa>2.3.Co;2.

Backhaus, T., Arrhenius, A., Blanck, H., 2004. Toxicity of a mixture of dissimilarly acting substances to natural algal communities: predictive power and limitations of independent action and concentration addition. Environ. Sci. Technol. 38 (23), 6363–6370. https://doi.org/10.1021/es0497678.

Bundesministerium für Ernährung und Landwirtschaft, 2017. Verwendung von Versuchstieren im Jahr 2016. Retrieved from. https://www.bmel.de/DE/Tier/Tierschutz/_texte/TierschutzTierforschung.html?docId=10323474#doc10323474bodyText11.

Corvaro, M., Gehen, S., Andrews, K., Chatfield, R., Arasti, C., Mehta, J., 2016. GHS additivity formula: a true replacement method for acute systemic toxicity testing of agrochemical formulations. Regul. Toxicol. Pharmacol. 82, 99–110. https://doi.org/10.1016/j.yrtph.2016.10.007.

Corvaro, M., Gehen, S., Andrews, K., Chatfield, R., Macleod, F., Mehta, J., 2017. A retrospective analysis of *in vivo* eye irritation, skin irritation and skin sensitisation studies with agrochemical formulations: setting the scene for development of alternative strategies. Regul. Toxicol. Pharmacol. 89, 131–147. https://doi.org/10.1016/j.yrtph.2017.06.014.

Creton, S., Dewhurst, I.C., Earl, L.K., Gehen, S.C., Guest, R.L., Hotchkiss, J.A., ... Billington, R., 2010. Acute toxicity testing of chemicals-Opportunities to avoid redundant testing and use alternative approaches. Crit. Rev. Toxicol. 40 (1), 50–83. https://doi.org/10.3109/10408440903401511.

Damalas, C.A., Eleftherohorinos, I.G., 2011. Pesticide exposure, safety issues, and risk assessment indicators. Int. J. Environ. Res. Publ. Health 8 (5), 1402–1419. https://doi.org/10.3390/ijerph8051402.

ECHA, 2017. Guidance on the Application of the CLP Criteria; Guidance to Regulation (EC) No 1272/2008 on Classification, Labelling and Packaging (CLP) of Substances and Mixtures. Version 5.0.

EFSA, 2017. Guidance on the use of the weight of evidence approach in scientific assessments. EFSA J. 15 (8).

EU, 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.

EU, 2008. Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on Classification, Labelling and Packaging of Substances and Mixtures, Amending and Repealing Directives 67/548/EEC and 1999/45/EC, and Amending Regulation (EC) No 1907/2006.

EU, 2009. Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on The Market and Repealing Council Directives 79/117/EEC and 91/414/EEC.

EU, 2012. Guidance Document on Significant and Non-significant Changes of the Chemical Composition of Authorised Plant Protection Products under Regulation (EC) No 1107/2009 of the EU Parliament and Council on Placing of Plant Protection Products on the Market and Repealing Council Directives 79/117/EEC and 91/414/EEC.

Federal Office of Consumer Protection and Food Safety, 2018. Function of co-formulants. Retrieved from. https://www.bvl.bund.de/EN/04_PlantProtectionProducts/01_ppp_tasks/08_ProductChemistry/01_ppp_coformulants_formulationChemistry/02_ppp_function_coformulants/ppp_function_coformulants_node.html.

Finney, D.J., Tattersfield, F., 1952. Probit Analysis. Cambridge University Press, Cambridge.

Hamm, J., Sullivan, K., Clippinger, A.J., Strickland, J., Shannon, B., Bhhatarai, B., ... Allen, D., 2017. Alternative approaches for identifying acute systemic toxicity: moving from research to regulatory testing. Toxicol. Vitro 41, 245–259. https://doi.org/10.1016/j.tiv.2017.01.004.

Hernández, A., Gil, F., Lacasaña, M., 2017. Toxicological interactions of pesticide mixtures: an update. Arch. Toxicol. 91 (10), 3211–3223.

Hoel, D.G., 1987. Statistical aspects of chemical mixtures. In: Methods for Assessing the Effects of Mixtures of Chemicals, pp. 369–377.

Hoffmann, S., Kinsner-Ovaskainen, A., Prieto, P., Mangelsdorf, I., Bieler, C., Cole, T., 2010. Acute oral toxicity: variability, reliability, relevance and interspecies comparison of rodent LD50 data from literature surveyed for the ACuteTox project. Regul. Toxicol. Pharmacol. 58 (3), 395–407.

Jaworska, J.S., Natsch, A., Ryan, C., Strickland, J., Ashikaga, T., Miyazawa, M., 2015. Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. Arch. Toxicol. 89 (12), 2355–2383.

Junghans, M., Backhaus, T., Faust, M., Scholze, M., Grimme, L.H., 2006. Application and validation of approaches for the predictive hazard assessment of realistic pesticide mixtures. Aquat. Toxicol. 76 (2), 93–110. https://doi.org/10.1016/j.aquatox.2005.10.001.

Kaltenhauser, J., Kneuer, C., Marx-Stoelting, P., Niemann, L., Schubert, J., Stein, B., Solecki, R., 2017. Relevance and reliability of experimental data in human health risk assessment of pesticides. Regul. Toxicol. Pharmacol. 88, 227–237. https://doi.org/10.1016/j.yrtph.2017.06.010.

Klimisch, H.J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharmacol. 25 (1), 1–5. https://doi.org/10.1006/rtph.1996.1076.

Kolle, S.N., Moreno, M.C.R., Mayer, W., van Cott, A., van Ravenzwaay, B., Landsiedel, R., 2015. The EpiOcular (TM) eye irritation test is the method of choice for the *in vitro* eye irritation testing of agrochemical formulations: correlation analysis of EpiOcular eye irritation test and BCOP test data according to the UN GHS, US EPA and Brazil ANVISA classification schemes. Alta Altern. Lab. Anim 43 (3), 181–198.

Kolle, S.N., Sullivan, K.M., Mehling, A., van Ravenzwaay, B., Landsiedel, R., 2012. Applicability of *in vitro* tests for skin irritation and corrosion to regulatory classification schemes: substantiating test strategies with data from routine studies. Regul. Toxicol. Pharmacol. 64 (3), 402–414. https://doi.org/10.1016/j.yrtph.2012.08.015.

Kolle, S.N., Van Cott, A., van Ravenzwaay, B., Landsiedel, R., 2017a. Lacking applicability of *in vitro* eye irritation methods to identify seriously eye irritating agrochemical formulations: results of bovine cornea opacity and permeability assay, isolated chicken eye test and the EpiOcular(TM) ET-50 method to classify according to UN GHS. Regul. Toxicol. Pharmacol. 85, 33–47. https://doi.org/10.1016/j.yrtph.2017.01.013.

Kolle, S.N., van Ravenzwaay, B., Landsiedel, R., 2017b. Regulatory accepted but out of domain: *In vitro* skin irritation tests for agrochemical formulations. Regul. Toxicol. Pharmacol. 89, 125–130. https://doi.org/10.1016/j.yrtph.2017.07.016.

OECD, 1987. Test No. 401: Acute Oral Toxicity.

OECD, 1992. Test No. 406. Skin Sensitisation.

OECD, 2002. Test No. 423: Acute Oral Toxicity - Acute Toxic Class Method.

OECD, 2009a. Test No. 403: Acute Inhalation Toxicity.

OECD, 2009b. Test No. 436: Acute Inhalation Toxicity – Acute Toxic Class Method.

OECD, 2010. Test No. 429. Skin Sensitisation.

OECD, 2014. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins.

OECD, 2015. Test No. 404: Acute Dermal Irritation/Corrosion.

OECD, 2017a. Test No. 402: Acute Dermal Toxicity.

OECD, 2017b. Test No. 405. Acute Eye Irritation/Corrosion.

Roth, N., Ciffroy, P., 2016. A critical review of frameworks used for evaluating reliability and relevance of (eco) toxicity data: perspectives for an integrated eco-human decision-making framework. Environ. Int. 95, 16–29.

Schrage, A., Gamer, A.O., van Ravenzwaay, B., Landsiedel, R., 2010. Experience with the HET-CAM method in the routine testing of a broad variety of chemicals and formulations. Alta Altern. Lab. Anim 38 (1), 39–52.

Scott, L., Eskes, C., Hoffmann, S., Adriaens, E., Alepée, N., Bufo, M., ... Guest, R., 2010. A proposed eye irritation testing strategy to reduce and replace in vivo studies using bottom–up and top–down approaches. Toxicol. Vitro 24 (1), 1–9.

Settivari, R.S., Gehen, S.C., Amado, R.A., Visconti, N.R., Boverhof, D.R., Carney, E.W., 2015. Application of the KeratinoSens (TM) assay for assessing the skin sensitization potential of agrochemical active ingredients and formulations. Regul. Toxicol. Pharmacol. 72 (2), 350–360. https://doi.org/10.1016/j.yrtph.2015.05.006.

Sewell, F., Doe, J., Gellatly, N., Ragan, I., Burden, N., 2017. Steps towards the international regulatory acceptance of non-animal methodology in safety assessment. Regul. Toxicol. Pharmacol. 89, 50–56. https://doi.org/10.1016/j.yrtph.2017.07.001.

Solecki, R., Ritz, V., 2018. Toxicity of selected chemicals: pesticides. In: Greim, H., Snyder, R. (Eds.), Toxicology and Risk Assessment (In Press). Wiley & Sons, Chichester.

Suciu, N.A., Panizzi, S., Ciffroy, P., Ginebreda, A., Tediosi, A., Barceló, D., Capri, E., 2018. Evolution and future of human health and environmental risk assessment. In: Modelling the Fate of Chemicals in the Environment and the Human Body. Springer, pp. 1–21.

UN, 2017. Globally Harmonized System of Classification and Labelling of Chemicals (GHS), Seventh revised edition. (New York, USA, Geneva, Switzerland).

Van Cott, A., Hastings, C.E., Landsiedel, R., Kolle, S., Stinchcombe, S., 2018. GHS additivity formula: can it predict the acute systemic toxicity of agrochemical formulations that contain acutely toxic ingredients? Regul. Toxicol. Pharmacol. 92, 407–419. https://doi.org/10.1016/j.yrtph.2017.12.024.

WHO, 2010. The WHO Recommended Classification of Pesticides by Hazard and Guidelines to Classification: 2009. Geneva, Switzerland.