Organisation for Economic Co-operation and Development

ENV/CBC/MONO(2022)10

Unclassified

English - Or. English

16 January 2025

ENVIRONMENT DIRECTORATE
CHEMICALS AND BIOTECHNOLOGY COMMITTEE

Cancels & replaces the same document of 19 September 2022

Supporting Document for Evaluation and Review of Test Guideline 467 on Defined Approaches (DAs) for Serious Eye Damage / Eye Irritation

Series on Testing and Assessment
No. 354

JT03558377

OECD Environment, Health and Safety Publications
Series on Testing & Assessment
No. 354

Supporting Document for Evaluation and Review of Test Guideline 467 on Defined Approaches (DAs) for Serious Eye Damage / Eye Irritation

IOMC
INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS
A cooperative agreement among FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD

Environment Directorate
ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT
Paris 2022

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 38 industrialised countries in North and South America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in eleven different series: **Testing and Assessment; Good Laboratory Practice and Compliance Monitoring; Pesticides; Biocides; Risk Management; Harmonisation of Regulatory Oversight in Biotechnology; Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents;** and **Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (www.oecd.org/chemicalsafety/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

This publication is available electronically, at no charge.

Also published in the Series on Testing and Assessment: link

For this and many other Environment,

Health and Safety publications, consult the OECD's

World Wide Web site (www.oecd.org/chemicalsafety/)

or contact:

OECD Environment Directorate,

Environment, Health and Safety Division

2 rue André-Pascal

75775 Paris Cedex 16

France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

Supporting Document for Evaluation and Review of Test Guideline 467 on Defined Approaches (DAs) for Serious Eye Damage / Eye Irritation

Foreword

This document contains the Supporting Document for the Defined Approaches (DAs) for Serious Eye Damage and Eye Irritation, published as Guideline No. 467. It provides detailed information on the process for selecting high quality reference chemicals, predictive performance, uncertainty in the DAs and their individual data information sources, as well as on the analysis of individual DIPs explored for the DAs in GL 467. This Supporting Document and its annexes were used as a basis to develop the GL 467, and they were circulated to the Working Party of the National Coordinators of the Test Guidelines Programme for reviews and comments in July and December 2021. The WNT approved the GL 467 and endorsed the supporting document in its 34th meeting in April 2022, on the basis of a project led by France. This document is published under the responsibility of the Chemicals and Biotechnology Committee.

Table of contents

Tables

Figures

List of acronyms

BCOP: Bovine Corneal Opacity and Permeability

CASRN: Chemical Abstracts Service Registry Number

Cat. 1: UN GHS classification for chemicals causing irreversible effects on the eye/serious damage to the eye

Cat. 2: UN GHS classification for chemicals causing reversible effects on the eye/eye irritation

CC: conjunctival chemosis

CO: corneal opacity

CON4EI: CONsortium for *in vitro* Eye Irritation testing strategy

Conj: conjunctival effects

CR: conjunctival redness

DA: Defined approach

DAL: defined approach liquids

DIP: data interpretation procedure

DRD: Draize eye test Reference Database

ECHA: European Chemicals Agency

EIT: Eye Irritation Test

EURL ECVAM: European Union Reference Laboratory on Alternatives to Animal Testing

FNR: false negative rate

FPR: false positive rate

GL: guideline

HCE: Human Corneal Epithelium

IATA: Integrated Approaches to Testing and Assessment

IR: iritis

LLBO: laser light-based opacitometer

LogP: octanol-water partition coefficient

MoA: Modes of Action

MSDS : material safety data sheet

MW: molecular weight

No Cat.: Not requiring UN GHS Classification

OECD: Organisation for Economic Co-operation and Development

OFG: Organic Functional Groups

RhCE: Reconstructed human Cornea-like Epithelium

SPSF: Standard Project Submission Form

ST: surface tension

STE: Short Time Exposure

UN GHS: United Nations Globally Harmonized System

VP: vapour pressure

VRM: Validated Reference Method

WNT: Working Group of National Co-ordinators of the Test Guidelines programme

WS: water solubility

1. Introduction

1. On November 2017 and 2018 two defined approaches (DAs) for serious eye damage/eye irritation were introduced to the Meeting of Expert Group on Eye/Skin Irritation/Corrosion and Phototoxicity. These two DAs received interest from the Expert Group in 2017 and support in 2018 when a draft Standard Project Submission Form (SPSF) was submitted by France in collaboration with Cosmetics Europe, in November 2018, ahead of the OECD Expert Group Meeting on Eye/Skin Irritation/Corrosion & Phototoxicity in view of WNT 31 (Working Group of National Co-ordinators of the Test Guidelines programme, April 2019) for a possible inclusion in the OECD WNT Workplan. The SPSF submitters updated the draft SPSF based on recommendations from the Expert Group and this version was shared with the WNT. On April 2019, the WNT accepted the SPSF on two DAs.

2. According to the UN GHS classification system, Category 1 (serious eye damage) refers to the production of tissue damage in the eye, or serious physical decay of vision, which is not fully reversible, occurring after exposure of the eye to a substance or mixture. Category 2 (eye irritation) refers to the production of changes in the eye, which are fully reversible, occurring after the exposure of the eye to a substance or mixture. Based on this definition, the hazard potential of a test chemical is determined in the Draize eye test (OECD TG 405, 2020) based on its effect on corneal opacity (CO), iritis (IR), conjunctival redness (CR), and conjunctival chemosis (CC). Based on the severity of effects and/or the timing of their reversibility, classifications are derived according to the serious eye damage/eye irritation classification criteria defined by the United Nations (UN) Globally Harmonized System of Classification and Labelling of Chemicals (GHS) (UN 2021). Effects not fully reversed at the end of the 21 day observation period of the Draize test are considered irreversible (Category 1) or not (Category 2). This category may be divided into the optional Categories 2A (effects fully reversible within 21 days) and 2B (effects fully reversible within 7 days). When none of the Cat. 1 or Cat. 2 classification criteria are met, the chemical does not require classification which corresponds with No Category (No Cat.). Note that every time reference is made to *in vivo* Cat. 1, Cat. 2, and No Cat. in this background review document, those classifications have been derived from testing in albino rabbits according to the Draize eye test method (OECD TG 405). The main data source of the historical data was the Draize eye test Reference Database (DRD) published by Cosmetics Europe (Barroso et al., 2017; see §15).

3. A comprehensive analysis to address the main *in vivo* ocular tissue effects that drive UN GHS classification was conducted and the outcomes were used to evaluate the performances of the two DAs described in the present document. The analyses identified nine different criteria from the four *in vivo* tissue effects (CO, IR, CR, and CC) that can each independently drive the classification of a chemical (Barroso et al., 2017). Of note, CR and CC were not reported separately but were reported together as conjunctival effects (Conj) because previous analyses revealed that CC rarely drives the classification of chemicals in the absence of CR effects (Adriaens et al., 2014; Barroso and Norman, 2014). Chemicals classified as Cat. 1 were grouped based on (i) severity (mean scores of days 1–3); (ii) persistence of any ocular effect on day 21 in the absence of severity; or (iii) CO = 4 (at any observation time during the study)

in the absence of both severity and persistence (or if unknown). Chemicals classified as Cat. 2 were allocated to one of the three following groups based on the main endpoint leading to Cat. 2 classification, i.e. “CO”, “Conj”, and “IR”. Studies with chemicals not requiring classification for serious eye damage/eye irritation (No Cat.) were distributed in four different groups depending on whether they showed CO scores equal to 0 in all animals and all observed time points (CO = 0 and CO = 0**) or not (CO > 0 and CO > 0**). No Cat. studies for which at least one animal had a mean of the scores of days 1–3 above the classification cut-off for at least one endpoint but not in enough animals to generate a classification (borderline cases) were marked with ** (CO = 0**, CO > 0**). A detailed description of the drivers of classification and use of the terms CO, IR and Conj to describe key effects is provided in the paper of Barroso and co-workers (2017).

1.1. References

Adriaens E, Barroso J, Eskes C, Hoffmann S, McNamee P, Alépée N, Bessou-Touya S, De Smedt A, De Wever B, Pfannenbecker U, Tailhardat M, Zuang V (2014) Retrospective analysis of the Draize test for serious eye damage/eye irritation: importance of understanding the *in vivo* endpoints under UN GHS/EU CLP for the development and evaluation of *in vitro* test methods. Arch Toxicol 88:701–723.

Barroso J, Norman K (2014). REACHing for alternatives to animal testing. A webinar series on modern testing strategies for REACH. Webinar 3 of 6 on “Serious Eye Damage and Eye Irritation”, December 4, 2014. http://www.piscltd.org.uk/reaching-alternatives-animal-testing/ (accessed 17.06.2016).

Barroso J, Pfannenbecker U, Adriaens E, Alépée N, Cluzel M, De Smedt A, Hibatallah J, Klaric M, Mewes KR, Millet M, Templier M, McNamee P (2017). Cosmetics Europe compilation of historical serious eye damage/eye irritation *in vivo* data analysed by drivers of classification to support the selection of chemicals for development and evaluation of alternative methods/strategies: the Draize eye test Reference Database (DRD). Arch Toxicol (2017) 91:521–547.

OECD (2020). Test No. 405: Acute Eye Irritation/Corrosion. In: OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris, https://doi.org/10.1787/9789264185333-en

UN (2021). United Nations Globally Harmonized System of Classification and Labelling of Chemicals (GHS). ST/SG/AC.10/30/Rev.9, Ninth Revised Edition, New York and Geneva: United Nations. Available at [https://unece.org/transport/standards/transport/dangerous-goods/ghs-rev9-2021]

2. Presentation of the DAs analysed

2.1. Introduction

4. This document supports a draft Guideline (GL) covering two DAs developed by Cosmetics Europe for eye hazard identification, i.e. addressing both serious eye damage and eye irritation (or the absence thereof), for non-surfactant liquid test substances. These DAs are DAL-1 for non-surfactant liquids combining physicochemical properties, Reconstructed human Cornea-like Epithelium (RhCE) test method (OECD TG 492), and Bovine Corneal Opacity and Permeability (BCOP) test method (OECD TG 437) and DAL-2 for non-surfactant liquids combining Short Time Exposure (STE) test method (OECD TG 491) and BCOP test method (OECD TG 437) (OECD 2019a; 2020a; 2020b). In both DAs, the BCOP laser light-based opacitometer (LLBO) is used, as described within the OECD TG 437 (2020). Within DAL-1 and DAL-2, only the opacity is used to identify liquids that cause serious eye damage. Assessment of permeability does not increase the method's predictivity and can therefore be omitted in the DAL-1 and DAL-2. Whenever in this background document the term "BCOP LLBO" is mentioned, this means that only the opacity as measured with the LLBO is used as endpoint to identify Cat. 1.

5. The applicability domain of the DAL-1 and DAL-2 is restricted to non-surfactant liquids. The DAs that are proposed in the current document are refinements of initially proposed defined approaches that resulted from the CONsortium for *in vitro* Eye Irritation testing strategy (CON4EI) project (Adriaens et al., 2018). During the CON4EI project, 80 chemicals (liquids and solids) were tested with 8 different alternative test methods (including OECD TG 437, TG 491 and TG 492 test methods). The chemicals were chosen in collaboration with Cosmetics Europe from the Draize eye test Reference Database (DRD) developed by Cosmetics Europe (Barroso et al., 2017). Additional analyses performed by Cosmetics Europe on an enlarged set of chemicals showed that the predictivity was better for liquids as compared to solids (Alépée et al., 2019a; Alépée et al., 2019b). This resulted in the development of two DAs for liquids. The rationale for the restriction of the chemical set to non-surfactant liquids was based on the fact that the false negative rate (FNR) for *in vivo* Cat. 1 surfactants (N=12) was high (42%) for the BCOP LLBO (based on opacity > 145).

6. DAL-1 and DAL-2 have been shown to be useful for making predictions across the whole range of UN GHS categories i.e., Category 1 (Cat. 1) on "serious eye damage"; Category 2 (Cat. 2) on "eye irritation" and No Category (No Cat.) for chemicals "not requiring classification and labelling" for eye irritation or serious eye damage (UN GHS, 2019). Whilst non-animal accepted OECD TGs can be used to identify Cat. 1 (e.g. OECD TG 437, TG 491) and chemicals that do not require classification for eye irritation or serious eye damage (No Cat.; e.g. OECD TG 491, TG 492), the two DAs also allow the classification into Cat. 2. However, the two DALs are not designed to distinguish between Categories 2A and 2B.

7. The two DAs described in this document follow recommendations and combinations of modules as stipulated in the Guidance Document No. 263 on Integrated Approaches to Testing and Assessment

(IATA) for serious eye damage and eye irritation, originally adopted by the OECD in 2017 (OECD, 2019b). This supporting document provides information on the evaluation of the two proposed DAs for hazard identification of serious eye damage and eye irritation potential of test chemicals (or the absence thereof), which are under consideration for inclusion in the OECD TG on DAs for serious eye damage/eye irritation. Information resulting from the application of the DAs contained in the final GL will be used, either on its own or in conjunction with other information, to meet regulatory data requirements for serious eye damage/eye irritation and will be covered under the agreement on the mutual acceptance of data (MAD). Much of the information provided in this supporting document has been published in peer reviewed journals (Alépée et al., 2019a, 2019b). Further, chapter 3 provides details on the criteria applied for the selection of the reference chemicals used to assess the DALs performance, and as agreed during the OECD Expert Group on Eye/Skin Irritation/Corrosion and Phototoxicity meeting of 2019. The decision on which DA will be used depends on the applicability domain of the individual test methods of the DAs (§54 and §63). In addition, the end user might prefer DAL-1 or DAL-2 because of familiarity with certain test methods (e.g., RhCE or STE).

2.2. DAL-1

8. The DAL-1 presented in this document describes the combination of one and/or three physicochemical properties with the results of two *in vitro* test methods (RhCE and BCOP LLBO) for the identification of the eye hazard potential of non-surfactant liquid substances primarily for the purposes of classification and labelling without the use of animal testing. The RhCE models that are part of DAL-1 are the EpiOcular™ Eye Irritation Test (EIT) and the SkinEthic™ Human Corneal Epithelium (HCE) EIT and are referred to in this GL as Validated Reference Methods - EpiOcular™ EIT (VRM1) and SkinEthic™ HCE EIT (VRM2), respectively (corresponding with OECD TG 492).

9. The data interpretation procedure (DIP) applied uses the readout of the prediction models of the individual test method as defined by the Test Guidelines and/or information on the physicochemical properties retrieved from publicly available databases (§10). A scheme of DAL-1 is presented in Figure 2.1. Physicochemical property exclusion rules based on water solubility (WS) or a combination of octanol-water partition coefficient (LogP), vapour pressure (VP) and surface tension (ST) of the neat liquid are used in a first step to identify liquid chemicals with no serious eye damage or eye irritation potential (details are provided in section 5.1.2.). Liquids for which the exclusion rules are not met, are evaluated based on a RhCE test method (VRM1 or VRM2) in Step 2. Liquids that result in a tissue viability > 60% are classified No Cat. Liquids that result in a tissue viability ≤ 60% are evaluated based on the BCOP LLBO test method in a third step. Liquids that result in an opacity > 145 are predicted Cat. 1 and the remaining liquids are assigned Cat. 2. Note that it is also possible to start with a RhCE method, followed by the physicochemical property exclusion rules in case the tissue viability measured with VRM1 or VRM2 > 60% (Figure 2.2). Furthermore, when a RhCE method is used as a first step and if the tissue viability > 60%, the prediction is based on the stand-alone method.

10. Physicochemical properties for the reference chemicals were extracted from the following sources (listed in order of priority): the European Chemicals Agency (ECHA) website that contains information on chemicals from registration dossiers submitted to ECHA

(https://echa.europa.eu/information-on-chemicals/registered-substances), the EPA Chemistry Dashboard website (https://comptox.epa.gov/dashboard), the PubChem website (https://www.ncbi.nlm.nih.gov/pccompound), the ChemSpider website (http://www.chemspider.com/) and other sources like e.g. Scientific Committee on Consumer Safety (SCCS) opinions publications. Furthermore, highest priority was given to experimentally derived measurements followed by computational methods (e.g. Quantitative Structure-Activity Relationships (Q)SAR) used to determine physicochemical properties. For the predicted values, highest priority was given to (Q)SAR models that were developed based on the five OECD principles (OECD, 2014). An overview of the OECD GLs and (Q)SAR models is provided in Annex E.

11. The performance of DAL-1 for the same set of chemicals was also calculated in case the BCOP LLBO was replaced with the BCOP OP-KIT (IVIS ≥ 55) to identify Cat. 1. Further, the performance was also provided not taking into account the physicochemical properties (start immediately from Step 2, Figure 2.1). The performance criteria were not met for any of these combinations. This information is provided in Annex A.

Figure 2.1. Scheme of the DAL-1; step 1 physicochemical exclusion rules (WS: water solubility in mg/mL; or LogP: octanol-water partition coefficient / VP: vapour pressure in mm Hg / ST: surface tension in dyne/cm of the neat liquid) to identify No Cat., step 2 RhCE test method used to identify No Cat., and step 3 BCOP LLBO used to identify Cat. 1

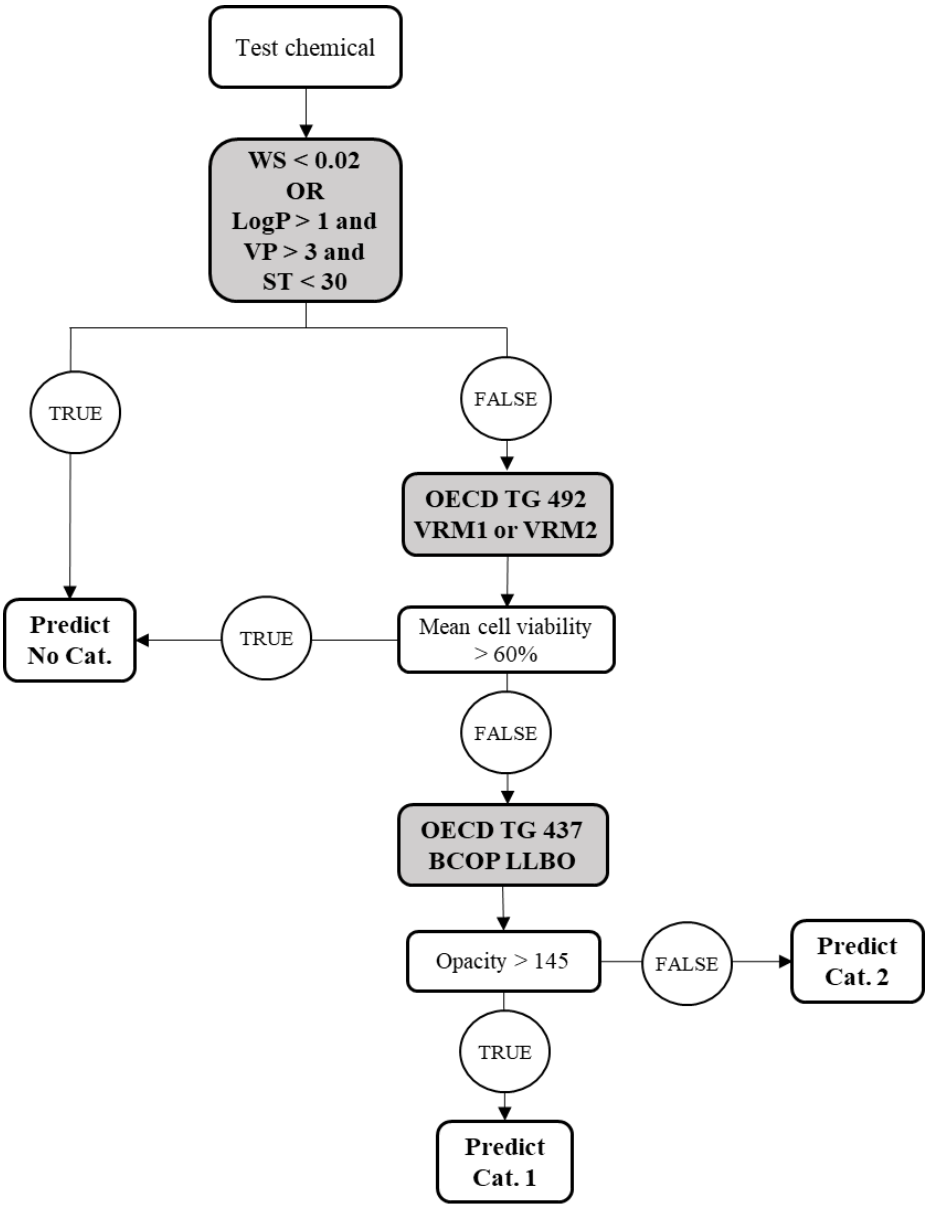
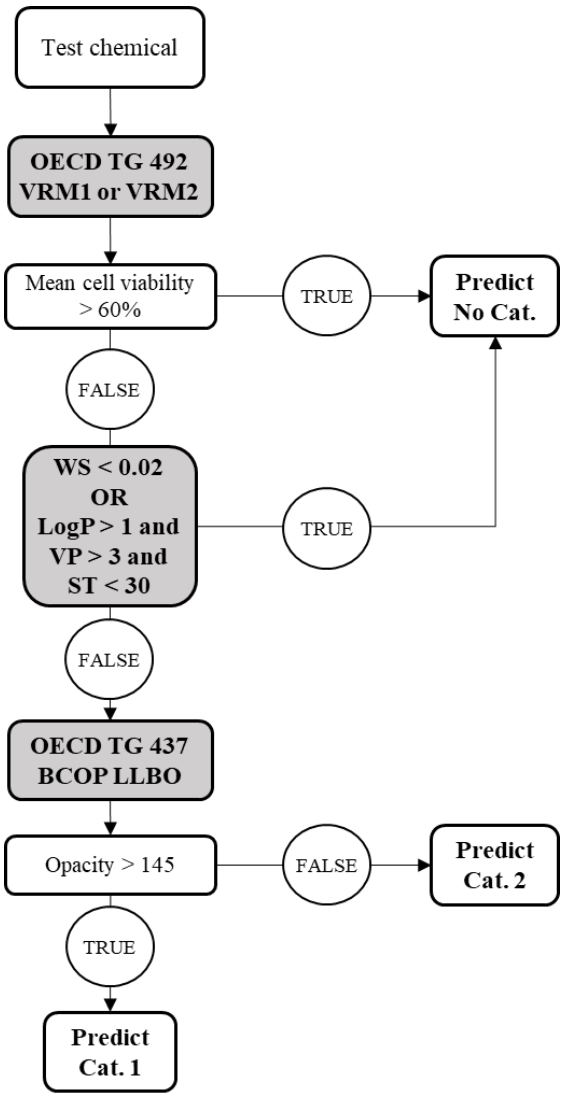


Figure 2.2. Scheme of the DAL-1; step 1 RhCE test method used to identify No Cat., step 2: physicochemical exclusion rules (WS: water solubility in mg/mL; or LogP: octanol-water partition coefficient / VP: vapour pressure in mm Hg / ST: surface tension in dyne/cm of the neat liquid) to identify No Cat., and step 3 BCOP LLBO used to identify Cat. 1



2.3. DAL-2

12. The DAL-2 presented in this document describes the combination of two *in vitro* test methods (STE and BCOP LLBO) for the identification of the eye hazard potential of non-surfactant liquid substances (neat liquids and liquids and solids dissolved in water), primarily for the purposes of classification and labelling without the use of animal testing. Whenever in this supporting document the term “non-surfactant liquids” is mentioned with respect to DAL-2, this refers to neat liquids and liquids and solids dissolved in water.

13. The DIP applied uses the readout of the prediction models of the individual test methods as defined by the Test Guidelines. A scheme of DAL-2 is presented in Figure 2.3. The STE test method is used to identify non-surfactant liquid chemicals with no serious eye damage or eye irritation potential (No Cat.: liquids that result in a mean cell viability > 70% at a 5% and 0.05% concentration) or to identify non-surfactant liquids that cause serious eye damage/eye irritation (Cat. 1: liquids that result in a mean cell viability ≤ 70% at a 5% and 0.05% concentration). For non-surfactant liquids that result in a mean cell viability ≤ 70% at 5% concentration but > 70% at 0.05%, the BCOP LLBO is needed. Liquids that result in an opacity > 145 are predicted as Cat. 1 and the remaining liquids are assigned to Cat. 2. Note that it is also possible to start with the BCOP LLBO followed by the STE test method, this scheme of DAL-2 is presented in Figure 2.4.

14. The performance of DAL-2 for the same set of chemicals was also calculated in case the BCOP LLBO was replaced with the BCOP OP-KIT (IVIS ≥ 55) to identify Cat. 1. This information is provided in Annex A. In this case, the performance criteria was not met.

Figure 2.3. Scheme of the DAL-2 option 1: start with the STE test method followed by the BCOP LLBO test method

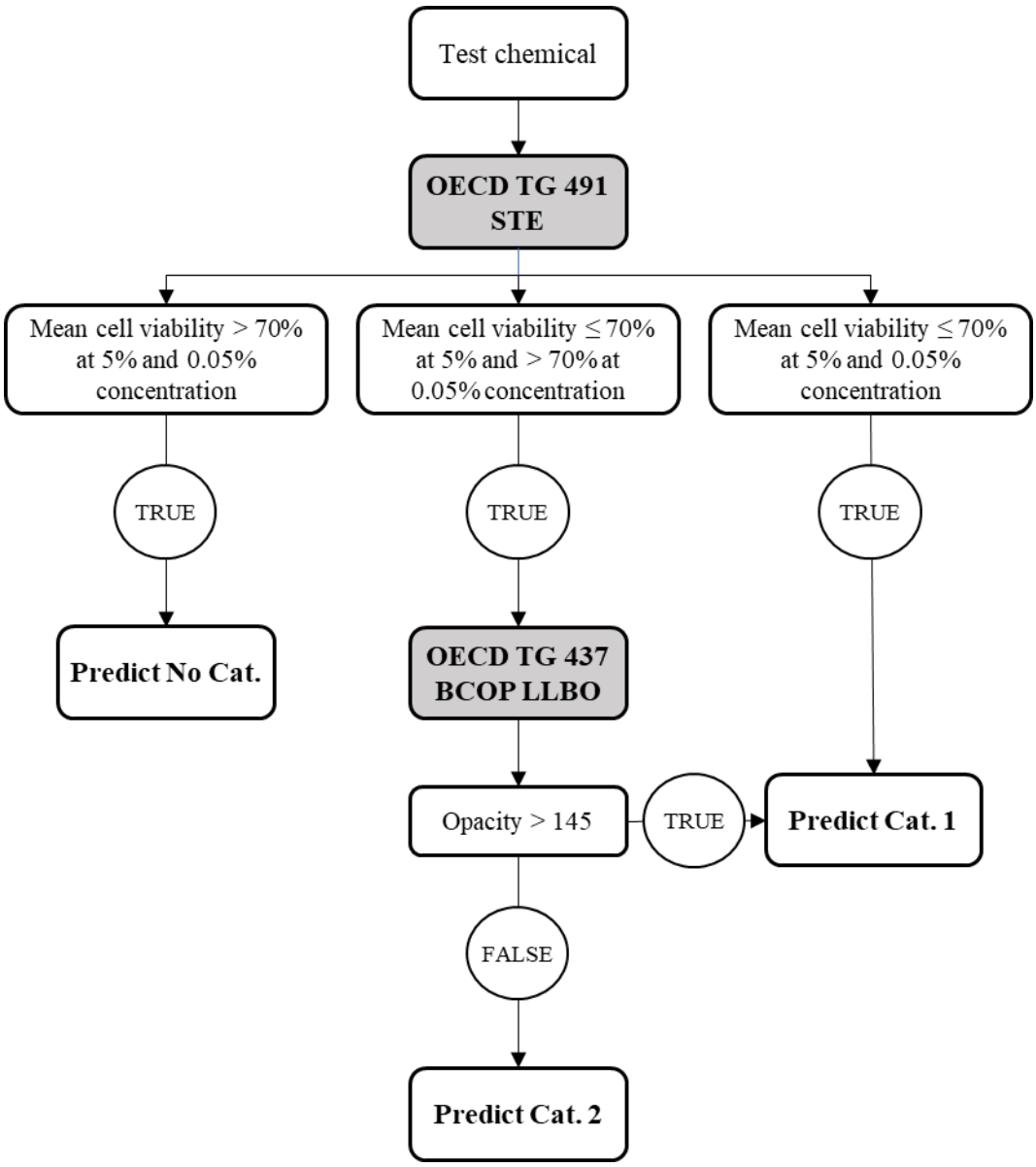
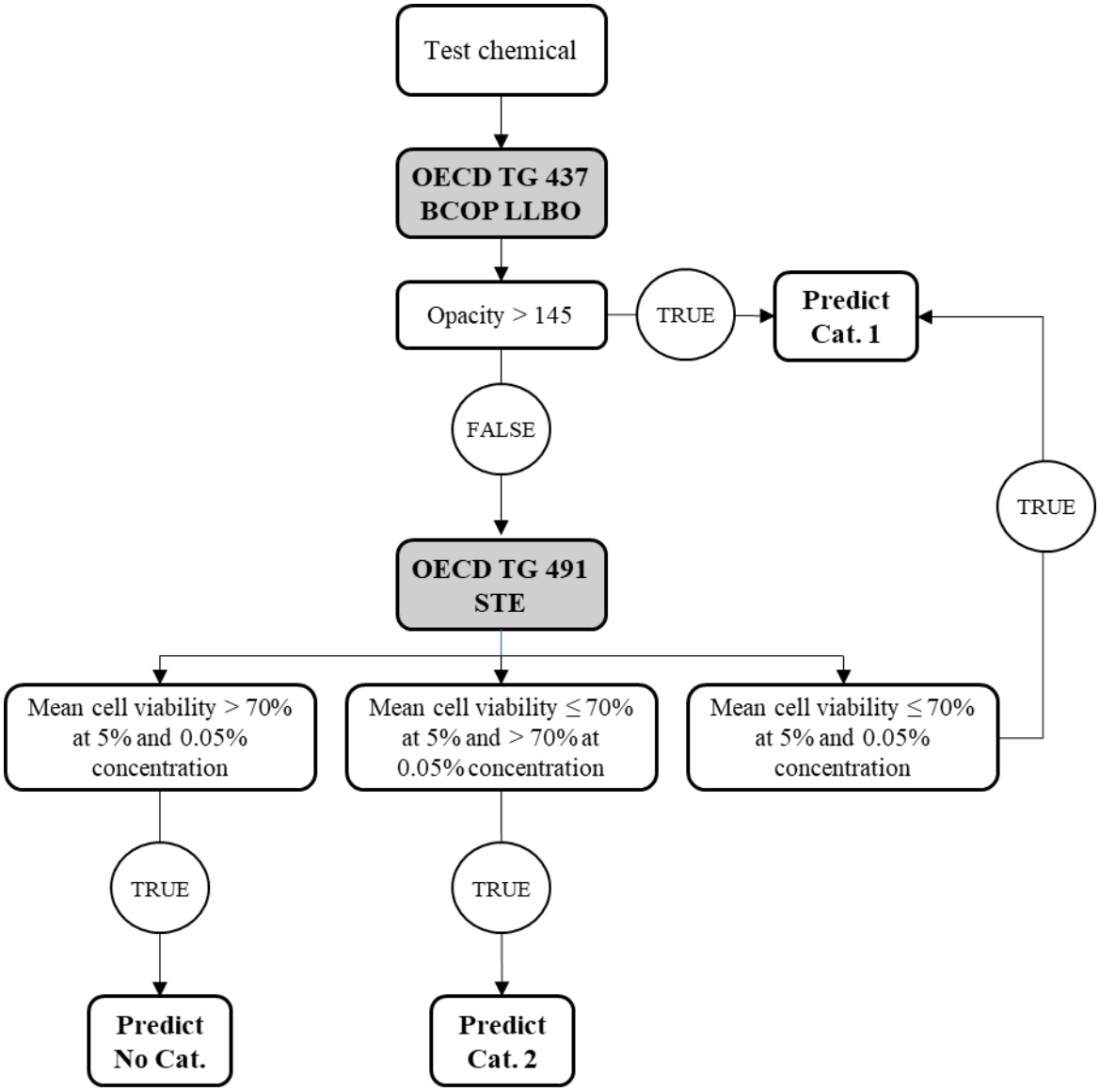


Figure 2.4. Scheme of the DAL-2 option 2: start with the BCOP LLBO test method followed by the STE test method



2.4. References

Adriaens E, Verstraelen S, Alépée N, Kandarova H, Drzewiecka A, Gruszka K, Guest R, Willoughby JA, Van Rompay AR (2018). CON4EI: Development of testing strategies for hazard identification and labelling for serious eye damage and eye irritation of chemicals. Toxicol. in Vitro 49, 99–115. https://doi.org/10.1016/j.tiv.2017.09.008.

Alépée N, Adriaens E, Abo T, Bagley D, Desprez B, Hibatallah J, Mewes KR, Pfannenbecker U, Sala A, Van Rompay AR, Verstraelen S, McNamee P. (2019a). Development of a defined approach for eye irritation or serious eye damage for neat liquids based on Cosmetics Europe Analysis of *in vitro* RhCE and BCOP test methods. Toxicology In Vitro (2019) 59: 100-114. doi: 10.1016/j.tiv.2019.04.011.

Alépée N, Adriaens E, Abo T, Bagley D, Desprez B, Hibatallah J, Mewes KR, Pfannenbecker U, Sala A,

Van Rompay AR, Verstraelen S, McNamee P. (2019b). Development of a defined approach for eye irritation or serious eye damage for liquids, neat and in dilution, based on cosmetics Europe analysis of in vitro STE and BCOP test methods. Toxicology In Vitro (2019) 57: 154-163. doi: 10.1016/j.tiv.2019.02.019.

Barroso J, Pfannenbecker U, Adriaens E, Alépée N, Cluzel M, De Smedt A, Hibatallah J, Klaric M, Mewes KR, Millet M, Templier M, McNamee P (2017). Cosmetics Europe compilation of historical serious eye damage/eye irritation *in vivo* data analysed by drivers of classification to support the selection of chemicals for development and evaluation of alternative methods/strategies: the Draize eye test Reference Database (DRD). Arch Toxicol (2017) 91:521–547.

OECD (2019a) Guideline for testing of chemicals no. 492: Reconstructed human cornea-like epithelium (RhCE) test method for identifying chemicals not requiring classification and labelling for eye irritation or serious eye damage. In: OECD Guidelines for the Testing of Chemicals, Section 4. Organisation for Economic Co-operation and Development, Paris. https://doi.org/10.1787/9789264242548-en.

OECD (2019b). No 263: Guidance Document on an Integrated Approach on Testing and Assessment (IATA) for Serious Eye Damage and Eye Irritation. Organisation for Economic Cooperation and Development, Paris, France. Available at: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV-JM-MONO(2017)15/REV1%20&doclanguage=en

OECD (2020a). Guideline for testing of chemicals no. 437: Bovine corneal opacity and permeability test method for identifying (i) chemicals inducing serious eye damage and (ii) chemicals not requiring classification for eye irritation or serious eye damage. In: OECD Guidelines for the Testing of Chemicals, Section 4. Organisation for Economic Co-operation and Development. https://doi.org/10.1787/9789264203846-en

OECD (2020b). Guideline for testing of chemicals no. 491: Short Time Exposure In Vitro Test Method for Identifying i) Chemicals Inducing Serious Eye Damage and ii) Chemicals Not Requiring Classification for Eye Irritation or Serious Eye Damage. In: OECD Guidelines for the Testing of Chemicals, Section 4. Organisation for Economic Co-operation and Development. https://doi.org/10.1787/9789264242432-en

UN (2019). United Nations Globally Harmonized System of Classification and Labelling of Chemicals (GHS). ST/SG/AC.10/30/Rev.7, Seventh Revised Edition, New York and Geneva: United Nations. Available at [https://read.un-ilibrary.org/environment-and-climate-change/globally-

OECD (2014), Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment, No. 69, OECD Publishing, Paris, https://doi.org/10.1787/9789264085442-en.

3. *In vivo* reference data (Draize eye test)

15. The main data source was the Draize eye test Reference Database (DRD) published by Cosmetics Europe (Barroso et al., 2017). The DRD contains 681 independent Draize rabbit eye test studies and was compiled using various sources of historical Draize eye test data, produced according to OECD TG 405, which were created to support previous validation activities (Barroso et al., 2017). Detailed information on the UN GHS category, the driver of classification, the organic functional groups present, and the identification of chemicals that should not be used for the evaluation of alternative methods and/or testing strategies can be retrieved from the DRD Supplementary Material 1. The second source of Draize eye test studies was historical eye irritation data on 122 test substances, used for the evaluation of the Hen's Egg Test Chorioallantoic Membrane (HET-CAM) test method (Scheel et al., 2011). However, the Draize eye test data of only one chemical (CAS RN 770-35-4) were used from this source.

3.1. Criteria applied for the selection of the reference chemicals for the DALs

16. The following criteria, as identified by Barroso and co-workers (2017), were considered when selecting the reference chemicals: (1) the expected applicability of the DALs in terms of UN GHS prediction (No. Cat., Cat. 2, and Cat. 1; §6), (2) important drivers of classification, (3) purity of the chemicals, and (4) relevance of the chemicals in terms of their representative functional and chemical classes and industrial use.

3.1.1. Drivers of classification

17. The chemical selection was performed by taking into account several key criteria that were identified by Barroso and co-workers (2017). One of the key criteria is that the pool of reference chemicals needs to address the main ocular tissue effects that drive classification. In the Draize rabbit eye test, the hazard potential of a test chemical is determined based on its effect on corneal opacity (CO), iritis (IR), conjunctival redness (CR), and conjunctival chemosis (CC). Based on the severity of effects and/or the timing of their reversibility, classifications are derived according to the serious eye damage/eye irritation classification criteria defined by UN GHS (UN 2019).

As described by Barroso and co-workers (2017), there are nine different criteria derived from the four tissue effects (CO, IR, CR, and CC) that can each independently drive the classification of a chemical (Table 3.1). Of note, CR and CC were not reported separately but were reported together as conjunctival effects (Conj) because previous analyses revealed that CC rarely drives the classification of chemicals in the absence of CR effects (Adriaens et al., 2014; Barroso and Norman, 2014).

Table 3.1. Drivers of UN GHS classification

Category 1 Irreversible effects on the eye/serious eye damage						Category 2 Reversible effects on the eye/eye irritation		
Severity (Mean scores of Days 1-3) [a]		Persistence on Day 21			Severe CO	Severity (Mean scores of Days 1-3) [a]		
CO mean ≥ 3	**IR mean > 1.5**	**CO**	**Conj**	**IR**	**CO=4**	**CO mean ≥ 1**	**Conj mean ≥ 2**	**IR mean ≥ 1**
in ≥ 60% of the animals	in ≥ 60% of the animals	in at least one animal	in at least one animal	in at least one animal	in at least one animal	in ≥ 60% of the animals	in ≥ 60% of the animals	in ≥ 60% of the animals

CO: corneal opacity; IR: iritis; Conj: conjunctival redness (CR) and/or conjunctival chemosis (CC)

Drivers with a greyed background correspond with the most important drivers

[a] Mean scores are calculated from gradings at 24, 48, and 72 hours after instillation of the test chemical

3.1.2. Key criteria considered when selecting reference chemicals

18. According to Barroso and co-authors (2017), corneal opacity is the most important endpoint driving Cat. 1 classification, corneal opacity and conjunctival effects are the most important endpoints driving Cat. 2 classification. The most important drivers of Cat. 1 (3 different criteria) and Cat. 2 (2 different criteria) classification are shown with a greyed background in Table 3.1 and are listed below. Note that for chemicals that were classified based on the driver CO persistence on day 21 or CO = 4, this effect should be present in at least 60% of the animals as advised by Barroso and co-authors (2017).

Drivers of classification for Cat. 1, by order of importance:

1. CO mean ≥ 3 (days 1 – 3) in ≥ 60% of the animals;
2. CO persistence on day 21 (D21) in ≥ 60%[a] of the animals (with CO mean < 3);
3. CO = 4 in ≥ 60% [b] of the animals in the absence of both severity and persistence or if unknown.

[a] Note that he 60% criterion applied for selection of the reference chemicals differs from the GHS criteria for the Draize test, in which a substance that produces in at least one animal effects on the cornea, iris or conjunctiva that are not expected to reverse or have not fully reversed within an observation period of normally 21 days.

[b] Cat. 1 is also adopted for substances that result in grade 4 cornea lesions and other severe reactions (e.g. destruction of cornea) observed at any time during the test.

Drivers of classification for Cat. 2, by order of importance:

4. CO mean ≥ 1;
5. CR mean ≥ 2 (with CO mean < 1)

   Subgroups for chemicals that do not require classification (No Cat.):
6. CO > 0 (minor effects on CO observed)
7. CO = 0 (clear negative results)

8. CO = 0 ** and CO > 0 ** (only a few chemicals should be included)

CO = 0: CO scores equal to 0 in all animals and all observed time points

CO > 0: in at least one observation time in at least one animal and all animals showing mean scores of days 1–3 below the classification cut-offs for all endpoints

** Indicates at least one animal with a mean score of days 1–3 above the classification cut-off for at least one endpoint

In total, the list of reference chemicals contained three substances that were identified in the DRD list (Barroso et al., 2017) as “should not be used”. Those chemicals were part of the training set (CON4EI project). At the time of the chemical selection, the guidelines on the selection of reference chemicals were not finalized yet and the identification of chemicals not recommended for testing was still ongoing. The following three chemicals are part of the DAL-1 training set: CASRN 3121-61-7 (No. 125), CASRN 2365-48-2 (No.126), and CASRN 109-99-9 (No. 132). The following two chemicals are part of the DAL-2 training set: CASRN 3121-61-7 (No. 125) and CASRN 2365-48-2 (No.126). For two out of those three chemicals the Cat. 1 triggering effect (CO=4) was not observed in the majority of the animals (No. 126 and No. 132) and for one chemical (No. 125) CO=4 was observed on day 1 and reversed to 0 by day 3 in 2/3 animals and was equal to 1 on day 14 in 1/3 animals (study terminated on day 14). For all the remaining chemicals of the DAL-1 and DAL-2 reference set, the selection criteria were fulfilled.

3.1.3. Purity of the chemicals

19. According to OECD GD 34, the reference chemicals should have a well-defined chemical structure and purity. The chemicals tested should, where possible, be of the highest available purity, or be of known composition.

20. The set of reference chemicals to support the review of DAL-1 was composed of 108 mono-constituents non-surfactant liquids. The set of reference chemicals to support the review of DAL-2 was composed of 164 mono-constituents non-surfactant liquids (147 neat liquids and 8 liquids and 9 solids dissolved in water). The purity of the chemicals should be as high as possible and ideally ≥ 95% (Barroso et al., 2017). The purity reported in the DRD supplementary Material 1 applies to available purity for the commercial source as indicated in the DRD, in fact the commercial source in the DRD is provided as an example (Barroso et al., 2017). Annex B.1 includes the detailed Draize eye test data that were used for DAL-1 and DAL-2 and the purity of the chemical as tested in the *in vivo* study was reported in case this was known. For the *in vitro* methods, the highest purity that was commercially available, was tested.

3.1.4. Chemical class, functional groups and uses

21. The set of liquids covers a broad range of uses and chemical classes, containing small and large molecules, and hydrophobic and hydrophilic chemicals, with a wide range of organic functional groups represented (79 different OFGs) defined according to OECD QSAR Toolbox analysis version 3.2; https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm). The most common OFGs are listed in chapter 7.3.

3.1.5. Criteria used for chemicals that should not be selected according to DRD paper (Barroso et al., 2017)

22. Health and safety issues relating to transport and handling of chemicals were also taken into account. The chemical selection avoided substances known to have critical toxicological and/or unstable physical properties (*e.g.* carcinogens, mutagens, lethal by inhalation) evident from official classifications and material safety data sheet (MSDS) information.

23. In general, chemicals that were classified based on one of the following criteria only (based on Draize eye test) were not included in the reference set of chemicals since they were identified as *"should not be used"* in either prospective studies or retrospective evaluations:

- Chemicals classified as Cat. 1 based only on persistence of CR and/or CC equal to 1 on day 21.

  The reasoning behind this is that in terms of biological relevance, persistence of low-level conjunctival effects (CR/CC = 1 on day 21) in the absence of any other Cat. 1 triggering effects should not have resulted in a Cat. 1 classification. A Cat. 1 classification in case of a repeat study is highly unlikely, especially when the effect is observed in only 1 out of 6 animals and where the other 5 animals are completely recovered by day 21. The DRD contains in total 3 liquids, all surfactants (outside the applicability domain of the DAs) where the Cat. 1 classification was driven by conjunctival persistence on day 21 in 1/6 animals. Those substances are identified as "*should not be used*" in the DRD.

  The DRD contains in total 7 substances (3 liquids and 4 solids) that were classified Cat. 1 based on the single driver conjunctival (CR and CC) persistence on day 21; all substances were surfactants, as such they do not belong to the applicability domain of the DALs.
- Chemicals that are classified as Cat. 1 based only on CO = 4 and/or persistent effects appearing in a minority (<60 %) of the animals.
- Chemicals that are classified Cat. 1 in the absence of any other Cat. 1 triggering effect (none of the Cat. 1 drivers listed in Table 3.1 could be assigned, those chemicals are listed in the DRD supplementary Material 1 https://static-content.springer.com/esm/art%3A10.1007%2Fs00204-016-1679-x/MediaObjects/204_2016_1679_MOESM1_ESM.pdf with the label "other observations" in the column "Specific observations") should in general not be selected as this accounts for a very limited number of studies in the DRD (9 substances in total – 5.5% of the Cat. 1 substances in the DRD – with only 1 liquid, CASRN 122321-04-4).
- Chemicals in the DRD identified as having repeat *in vivo* study data available resulting in discordant classifications were excluded from the reference list. For example for ethanol (CASRN 64-17-5), Draize eye irritation data of 4 studies are provided in the DRD (No. 57, 171, 201, and 602) and this resulted in inconsistencies of classification (1x Cat. 1, 2x Cat. 2, and 1x at least Cat. 2 but the study criteria were not met).

3.2. Key elements for evaluation of the DALs versus the Draize eye test

24. In 2005, an expert meeting was held by the European Union Reference Laboratory on Alternatives to Animal Testing (EURL ECVAM) and the outcome was reported in a peer review paper (Scott et al., 2010). As of today, no single *in vitro* test method has been validated to fully replace the Draize eye test for regulatory purposes. The difficulty to predict the middle category (UN GHS Cat. 2) was recognized and a testing strategy using the Bottom-Up and Top-Down approach was developed from this meeting. The experts (test developers/users) were requested to nominate *in vitro* eye irritation methods that could be considered as a basis for such testing strategy. Test methods were evaluated/categorized

based on their proposed applicability domain e.g., (i) categories of irritation severity, (ii) modes of action, (iii) chemical class, (iv) physicochemical compatibility. The main outcome of the EURL ECVAM expert meeting was the development of the Bottom-Up and Top-Down testing strategy (Scott et al., 2010; OECD, 2019b). However, having the knowledge on the applicability domain of the *in vitro* test methods and applying the Bottom-Up or Top-Down approach, successful full replacement of the *in vivo* Draize eye test was still not achieved.

25. Later on, a different approach was used in order to better understand the reason why still only partial replacement of the regulatory Draize eye test was achieved (Adriaens et al., 2014; Barroso et al., 2017). It was recognized that determination of the most relevant *in vivo* endpoint(s), in particular the effects on cornea, iris or conjunctiva, is extremely important for the development of adequate *in vitro* methods and will allow better understanding of the relationship between the *in vitro* and the *in vivo* data. A comprehensive in-depth analysis of historical *in vivo* rabbit eye data provided insight into which of the observed *in vivo* effects are important in driving the classification of chemicals for serious eye damage/eye irritation according to the UN GHS, concluding that full replacement of *in vivo* testing for eye hazard will require accounting for the impact of the *in vivo* tissue effects which drive classification (Adriaens et al. 2014); by taking into account the key drivers of classification, the DAs move closer to fully replacing the Draize test. Further, the uncertainty (variability) of the *in vivo* reference data is also recognized as a challenging factor that may hinder the successful development of non-animal approaches and should be allowed for when evaluating/validating *in vitro* test methods and strategies. For example, it has been shown that, for the rabbit eye test, the likelihood of achieving the same classification upon repeat testing is <50% for substances which fall into the mild to moderate irritation range (Luechtefeld et al., 2016). It is therefore challenging to align the results from *in vitro* methods to the *in vivo* rabbit test for the middle category (UN GHS Cat 2). Next, a database of Draize data was compiled (Cosmetics Europe Draize eye test Reference Database, DRD) and an evaluation of the various *in vivo* drivers of classification compiled in the database was performed to establish which of these are most important from a regulatory point of view (Barroso et al., 2017). These analyses established the most important drivers for Cat. 1 and Cat. 2 classification and the distribution in different groups for the chemicals that do not require classification. Further, a number of key criteria were identified that should be taken into consideration when selecting reference chemicals for the development, evaluation and/or validation of alternative methods and/or strategies for serious eye damage/eye irritation testing.

26. In November (Nov 3, 2020) a teleconference was held with a subgroup of the Expert Group on Skin and Eye irritation to discuss the issue regarding the Modes of action (MoA). It was concluded that the MoA are unknowable for the majority of the chemicals and most test substances would fall into multiple chemical classes. As such an analysis based on the MoA will not provide additional insight in explaining the performance of test methods and defined approaches and it is not possible to assess whether the DAs cover all relevant mode of actions. In addition to ensuring that the key *in vivo* drivers of classification have been covered by the selected reference chemicals, analysis of the OFGs present across the reference test chemicals show that a wide range of functionality has been covered over the UN GHS Cat. 1, Cat. 2 and No Cat. classified chemicals.

27. In conclusion, the assessment of the performance of the individual test methods and DAs against the Draize eye test has been conducted based on reference chemicals selected according to key criteria, as defined by Barroso and co-workers (2017), such that the important drivers of classification for each UN GHS category and a wide range of organic functional groups are represented.

3.3. References

Adriaens E, Barroso J, Eskes C, Hoffmann S, McNamee P, Alépée N, Bessou-Touya S, De Smedt A, De Wever B, Pfannenbecker U, Tailhardat M, Zuang V (2014) Retrospective analysis of the Draize test for serious eye damage/eye irritation: importance of understanding the *in vivo* endpoints under UN GHS/EU CLP for the development and evaluation of *in vitro* test methods. Arch Toxicol 88:701–723.

Barroso J, Pfannenbecker U, Adriaens E, Alépée N, Cluzel M, De Smedt A, Hibatallah J, Klaric M, Mewes KR, Millet M, Templier M, McNamee P (2017). Cosmetics Europe compilation of historical serious eye damage/eye irritation *in vivo* data analysed by drivers of classification to support the selection of chemicals for development and evaluation of alternative methods/strategies: the Draize eye test Reference Database (DRD). Arch Toxicol 91:521–547.

Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu H, Hartung T, 2016. Analysis of Draize Eye Irritation Testing and its Prediction by Mining Publicly Available 2008–2014 REACH Data. ALTEX 33, 123-134. doi:10.14573/altex.1510053.

OECD (2019a) Guideline for testing of chemicals no. 492: Reconstructed human cornea-like epithelium (RhCE) test method for identifying chemicals not requiring classification and labelling for eye irritation or serious eye damage. In: OECD Guidelines for the Testing of Chemicals, Section 4. Organisation for Economic Co-operation and Development, Paris. https://doi.org/10.1787/9789264242548-en.

OECD (2019b). No 263: Guidance Document on an Integrated Approach on Testing and Assessment (IATA) for Serious Eye Damage and Eye Irritation. Organisation for Economic Cooperation and Development, Paris, France. Available at: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV-JM-MONO(2017)15/REV1%20&doclanguage=en

OECD (2020). Test No. 405: Acute Eye Irritation/Corrosion. In: OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris, https://doi.org/10.1787/9789264185333-en

Scheel J, Kleber M, Kreutz J, Lehringer E, Mehling A, Reisinger K, Steiling W (2011). Eye irritation potential: usefulness of the HET-CAM under the globally harmonized system of classification and Labeling of chemicals (GHS). Regul. Toxicol. Pharmacol. 59, 471–492. https://doi.org/10.1016/J.YRTPH.2011.02.003.

Scott L, Eskes C, Hoffmann S, Adriaens E, Alépée N, Bufo M, Clothier R, Facchini D, Faller C, Guest R, Harbell J, Hartung T, Kamp H, Varlet BL, Meloni M, McNamee P, Osborne R, Pape W, Pfannenbecker U, Prinsen M, Seaman C, Spielmann H, Stokes W, Trouba K, Berghe CV, Goethem FV, Vassallo M, Vinardell P, Zuang V (2010) A proposed eye irritation testing strategy to reduce and replace *in vivo* studies using bottom-up and top-down approaches. Toxicol In Vitro 24:1–9. doi:10.1016/j.tiv.2009.05.019

UN (2019). United Nations Globally Harmonized System of Classification and Labelling of Chemicals (GHS). ST/SG/AC.10/30/Rev.7, Seventh Revised Edition, New York and Geneva: United Nations. Available at [https://read.un-ilibrary.org/environment-and-climate-change/globally-harmonized-system-of-classification-and-labelling-of-chemicals-ghs_f8fbb7cb-en#page1]

4. Evaluation of the Draize eye test uncertainty and reproducibility

28. The Draize eye test is the *in vivo* animal reference test used for benchmarking the predictive performance of serious eye damage/eye irritation DAs.

29. This document reports an assessment of the Draize eye test reproducibility that was based on two published comprehensive analyses (Adriaens et al. 2014; Barroso et al. 2017) on the inherent variability of the Draize eye test. The variability of the animal data has to be considered in the evaluation of the uncertainties when comparing DAs' predictions to the benchmark animal predictions.

4.1. Within-test variability

30. The impact of the uncertainty of *in vivo* reference data on the evaluation/validation of alternative methods was illustrated by the resampling analysis (within-test variability using individual rabbit data) presented by Adriaens et al. (2014). In total, 2089 studies were used for this analysis.

31. The resampling probabilities were estimated based on the individual rabbit data. Only studies with individual data on at least three rabbits were taken into account. In the resampling approach used in this study, simulated chemicals were created by randomly grouping together three animals that may have been tested with different chemicals.

- First, the different studies were pooled according to UN GHS classification of the tested chemicals. In this way, it was assured that the rabbits used in the various resampling always came from studies with chemicals classified with the same UN GHS category (i.e. No Cat., Cat. 2, or Cat. 1).
- Next, separate resampling analyses were then performed on each of the three individual data pools (the pool of studies within each UN GHS category). Data on 10,000 simulated chemicals were generated, i.e. a random sample of three rabbits was drawn 10,000 times from the data pool without replacement. This means that each animal entered a simulated chemical only once.
- Finally, the UN GHS classification criteria were applied for these simulated chemicals and the predictive capacity (correct classification) was calculated by comparing the theoretical classification (resulting from the resampling approach) with the observed classification.

32. This analysis strongly suggests a high over-predictive power of the Draize eye test. The resampling analyses based on the simulated chemicals demonstrated an overall probability of

- at least **15%** that liquids classified as Cat. 1 by the Draize eye test could be equally identified as Cat. 2 and none of them were identified as No Cat.
- about **10%** for Cat. 2 liquids to be equally identified as No Cat.
- the over-classification error for No Cat. and Cat. 2 liquids was negligible **(<1 %)**

4.2. Between-test variability

33. Cosmetics Europe has compiled a database of Draize data (Draize eye test Reference Database, DRD) from external lists that were created to support past validation activities (Barroso et al. 2017). This database contains 681 independent *in vivo* studies on 634 individual chemicals representing a wide range of chemical classes.

34. For the purpose of this document, an evaluation of the Draize eye test between-test variability was considered. Such analyses were based on liquids for which more than one independent study was performed by different laboratories. However, one must take into account the low number of repeat studies and therefore, generalization of the reproducibility is not possible.

- The reproducibility of the repeat studies (set of 24 chemicals) evaluated in terms of agreement of classifications:
  - **16.7%** (1/6) of the non-surfactant liquids with at least one Cat. 1 study could be equally identified as Cat. 2
  - **75.0%** (3/4) of the non-surfactant liquids with at least one Cat. 2 study (highest classification Cat. 2) could be equally identified as No Cat.
  - 14 non-surfactant liquids showed a concordant No Cat. outcome in two repeated studies.

4.3. Performance metrics

35. To date, no target values are defined to assess the performance of defined approaches for eye hazard identification to distinguish between the three UN GHS categories. Cosmetics Europe proposed target values that considered the uncertainty of the Draize eye test by taking into account the within- and between-test variability (Adriaens et al. 2014; Barroso et al. 2017). After discussion with the OECD expert group a consensus was reached on the performance criteria to assess the predictivity of DAs. The values are reported in Table 4.1.

Table 4.1. Performance metrics for assessment of the predictivity of a DA of non-surfactant liquid test substances for eye hazard identification

UN GHS	Defined Approach		
	Cat. 1	Cat. 2	No Cat.
Cat. 1	≥ 75%	≤ 25%	≤ 5%
Cat. 2	≤ 30%	≥ 50%	≤ 30%
No Cat.	≤ 5%	≤ 30%	≥ 70%

4.4. References

Adriaens E, Barroso J, Eskes C, Hoffmann S, McNamee P, Alépée N, Bessou-Touya S, De Smedt A, De

Wever B, Pfannenbecker U, Tailhardat M, Zuang V (2014) Retrospective analysis of the Draize test for serious eye damage/eye irritation: importance of understanding the *in vivo* endpoints under UN GHS/EU CLP for the development and evaluation of *in vitro* test methods. Arch Toxicol 88: 701–723.

Barroso J, Pfannenbecker U, Adriaens E, Alépée N, Cluzel M, De Smedt A, Hibatallah J, Klaric M, Mewes KR, Millet M, Templier M, McNamee P (2017). Cosmetics Europe compilation of historical serious eye damage/eye irritation *in vivo* data analysed by drivers of classification to support the selection of chemicals for development and evaluation of alternative methods/strategies: the Draize eye test Reference Database (DRD). Arch Toxicol 91: 521–547.

5. Analyses of the DAs performance

36. Chapter 5 of this document includes information on the DAL-1 and DAL-2 developed by Cosmetics Europe that was originally presented at the OECD Expert Group on Eye/Skin Irritation/Corrosion & Phototoxicity (2017) and was supported by the OECD Expert Group to be considered in their programme. The information in this chapter was organised according to the evaluation framework proposed for the Defined Approaches for Skin Sensitisation (OECD, 2019a). Further, chapter 3 provides details on the criteria applied for the selection of the reference chemicals used to assess the DALs performance, and as agreed during the OECD Expert Group on Eye/Skin Irritation/Corrosion and Phototoxicity meeting of 2019.

5.1. DAL-1

5.1.1. Background information

37. Based on the outcome of the CON4EI project, a two-step approach was proposed for assessing the eye hazard potential of chemicals using an RhCE test method for No Cat. identification and the BCOP LLBO (opacity only) for Cat. 1 identification (Adriaens et al., 2018). An important change with regard to the BCOP LLBO was that instead of using an LIS > 125 to identify Cat. 1 (Verstraelen et al., 2013), the prediction model was optimized and, compared to TG 437, opacity only was used to identify Cat. 1 (opacity > 145; Adriaens et al., 2017). This approach formed the basis of the DAL-1 developed by Cosmetics Europe. Additional analyses performed on an enlarged set of chemicals showed that the predictivity was better for liquids as compared to solids (Alépée et al., 2019a). Further analysis (Principal Component Analysis, PCA) revealed that liquids with certain physicochemical properties did correspond with liquids that were categorized as No Cat. based on the *in vivo* Draize eye test and that were predicted positive (False Positive, FP) with the RhCE test method. The DAL-1 combines four physicochemical properties with the results of two *in vitro* test methods (RhCE and BCOP LLBO, Figure 2.1). A more detailed explanation on the selection of relevant physicochemical properties is provided in the following section. Note that the analysis performed on the physicochemical properties of the reference chemicals was a separate analysis and that the training (N=135) and test set (N=32) for the physicochemical properties analysis are not per se the same training set (N=46) and test set (DAL-1 with VRM1 and VRM2, N=48 and N=40) used for the analysis of the combination of the *in vitro* test methods in DAL-1.

5.1.2. Selection of physicochemical properties

38. The relationship between several physicochemical properties and UN GHS classification was investigated using principal component analysis (PCA). A complete set of records was available for 148 liquids (20 Cat. 1, 23 Cat. 2, and 105 No Cat. chemicals) for the following physicochemical parameters of interest: molecular weight (MW), octanol-water partition coefficient (LogP), water solubility (WS), melting point (MP), vapour pressure (VP) and surface tension (ST) of the neat liquid. Since the distributions of the WS and VP were strongly skewed to the right, the data were log-transformed for further analysis. The

outcome of the PCA showed that physicochemical properties such as WS, LogP, VP and ST may improve the correct identification of neat liquids (Alépée et al., 2019a). In a next step, limit values for WS, LogP, VP and ST that can be used for prediction of chemicals not likely to cause serious eye damage/eye irritation were identified based on classification trees. The optimal threshold value for a physicochemical property was found which best splits the data into two groups (UN GHS No Cat. versus Cat. 1 + Cat. 2). For this purpose, the original dataset was enlarged (19 chemicals were added) and split into a training set (N=135) used to compute the threshold value, and a test set (N=32) to assess the accuracy of the algorithm. No changes were made to the threshold values after assessing the performance of the test set. Physicochemical properties of additional non-surfactant liquids (extra sets) were retrieved to assess the robustness of the exclusion rules. The distribution of training, test set, and extra sets is shown in Table 5.1.

Table 5.1. Distribution of the chemical sets over the UN GHS categories

Data seta	Cat. 1	Cat. 2	No Cat.	Total
Training set DRD	17	23	95	135
Test set DRD	4	3	25	32
Extra set DRD	4	2	14	20
Extra set ECHA	1	6	0	7
All	26	34	134	194

Source: [a] 135 substances for training set, 32 substances for test set, 20 extra substances from the DRD (Barroso et al., 2017) and 7 extra substances from ECHA

39. The distribution of the chemicals according to the physicochemical property limits is presented in Table 5.2 (training and test set) and Table 5.3 (extra set). This table shows for each physicochemical property and for the combination of properties, the proportion of neat liquids that have this characteristic within each UN GHS category. Of note, none of the *in vivo* Cat. 1 and Cat. 2 liquids have the combined properties of LogP > 1 and VP > 3 and ST < 30. As such, the combination of these physicochemical property limits can be used as an exclusion rule to identify chemicals with no serious eye damage or eye irritation potential. For example, if LogP > 1 was used alone as an exclusion rule for physicochemical properties, this would result in a high FN rate, for *in vivo* Cat. 1 (57.1%) and Cat. 2 (46.2%) liquids (Table 5.2). It is therefore important that all three parameters are used in combination.

Table 5.2. Proportion of liquids among the UN GHS categories that have values below (WS and ST) or above (LogP and VP) the physicochemical property limit (training and test set).

UN GHS	WS < 0.02 (mg/mL)	LogP > 1	VP > 3 (mm Hg)	ST < 30 (dyne/cm)	LogP > 1 & VP > 3	LogP > 1 & VP > 3 & ST < 30
Cat. 1 (N=21)	0.0	57.1	9.5	52.4	4.8	0.0
Cat. 2 (N=26)	3.8	46.2	30.8	50.0	3.8	0.0
No Cat. (N=120)	27.5	70.0	41.7	53.3	30.0	22.5

Table 5.3. Proportion of liquids among the UN GHS categories that have values below (WS and ST) or above (LogP and VP) the physicochemical property limit. (Extra set)

UN GHS	WS < 0.02 (mg/mL)	LogP > 1	VP > 3 (mm Hg)	ST < 30 (dyne/cm)	LogP > 1 & VP > 3	LogP > 1 & VP > 3 & ST < 30
Cat. 1 (N=5)	0.0	20.0	20	20	0.0	0.0
Cat. 2 (N=8)	0.0	62.5	0	37.5	0.0	0.0
No Cat. (N=14)	35.7	92.9	14.3	35.7	7.1	7.1

40. The robustness of the exclusion criteria is evaluated for each physicochemical property separately. The results are shown in violin plots (Figure 5.1 to Figure 5.4). Regarding water solubility (Figure 5.1), one *in vivo* Cat. 1 liquid had a WS close to the threshold value tributyltin oxide, CASRN 56-35-9, WS = 0.071 mg/mL) and one *in vivo* Cat. 2 liquid had a WS < 0.02 mg/mL (Bioallethrin, CASRN 28434-00-6). On the other hand, for several *in vivo* No Cat. liquids, WS was below 0.02 mg/mL (Table 5-2 (33/120) and Table 5-3 (5/14) combined 28.4% (38/134)).

41. The distribution of the LogP values by UN GHS category are shown in Figure 5.2. The symbols in red represent liquids with a LogP ≤ 1 and VP > 3 and ST < 30, in fact the liquids for which 2 out of 3 exclusion criteria are met. One *in vivo* Cat. 1 liquid and 6 *in vivo* Cat. 2 liquids have a VP > 3 and ST < 30 with a LogP value ≤ 1. Note that the LogP value for this Cat. 1 liquid is 0.45 (< 1) and the highest LogP value among the Cat. 2 liquids is 0.77 (< 1).

42. The distribution of the VP (mm Hg) values by UN GHS category are shown in **Figure 5.3**. The symbols in red represent liquids with a VP ≤ 3 and a LogP > 1 and ST < 30. Three *in vivo* Cat. 1 liquids and 6 *in vivo* Cat. 2 liquids have a LogP > 1 and ST < 30 with a VP ≤ 3. The highest VP value among the Cat. 1 and Cat. 2 liquids is 0.90 (< 3 mm Hg) and 2.73 (< 3 mm Hg).

43. The distribution of ST (dyne/cm) values by UN GHS category are shown in **Figure 5.4**. The symbols in red represent liquids with a ST ≥ 30 and LogP > 1 and VP > 3, in fact the liquids for which 2 out of 3 exclusion criteria are met. One *in vivo* Cat. 1 and one *in vivo* Cat. 2 liquid have a LogP > 1 and VP > 3 with a ST ≥ 30 (ST = 32 and 70, respectively).

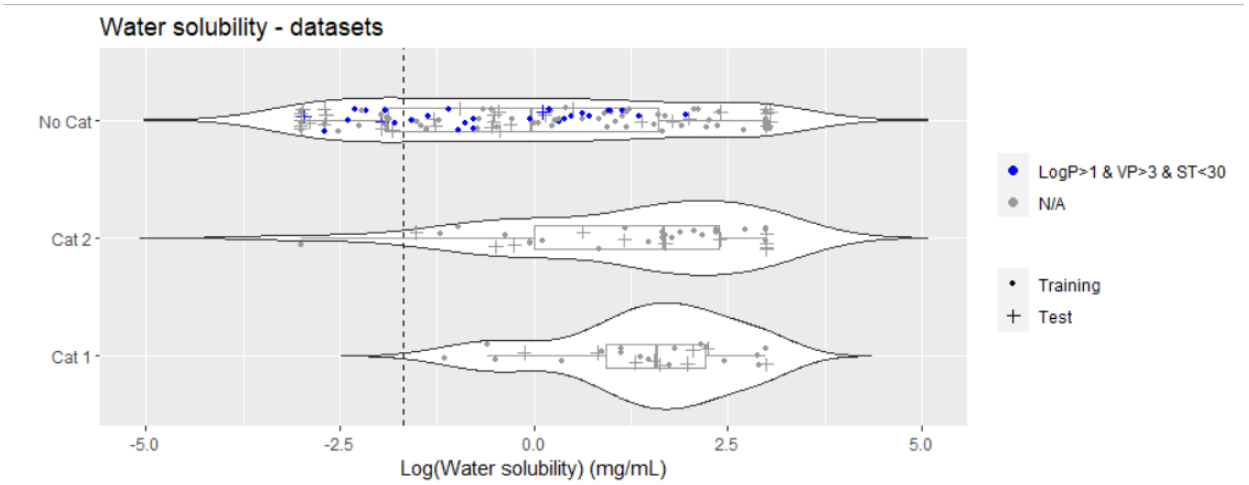


Figure 5.1. Distribution of Log(WS) values showing the liquids used in the training set (●) and test set (+). In blue the liquids for which the exclusion criteria (LogP > 1 & VP > 3 & ST < 30) were met, in grey all other combinations for LogP, VP, and ST (N/A). The dotted line corresponds with the cut-off of Log(0.02 mg/mL) = -1.7 mg/mL, liquids with a water solubility < 0.02 mg/mL are predicted No Cat.

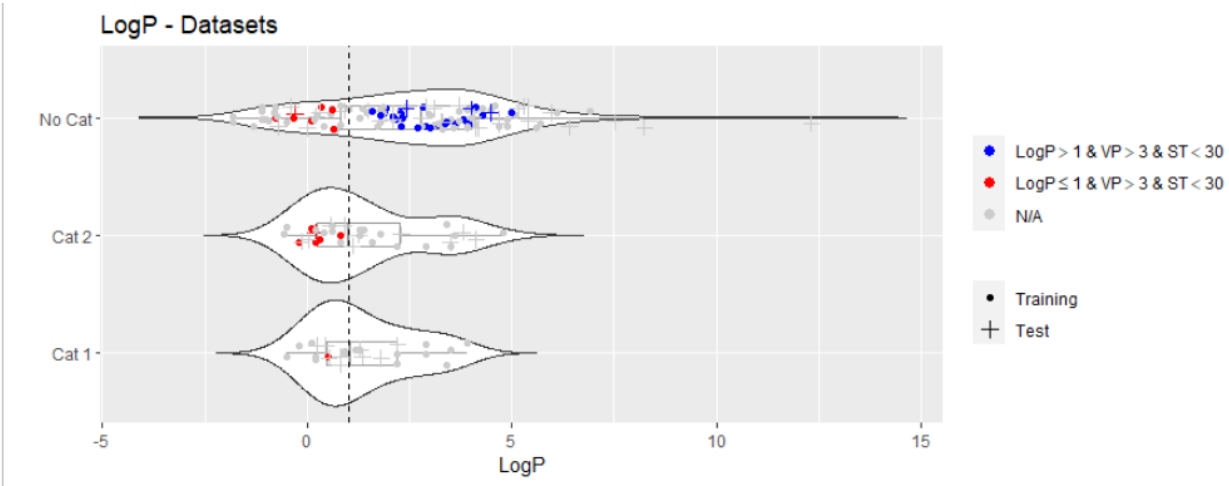


Figure 5.2. Distribution of the octanol water partition coefficient (LogP) showing the liquids used in the training set (●) and test set (+). In blue the liquids for which the exclusion criteria (LogP > 1 & VP, > 3 and ST < 30) were met, in red the liquids with LogP ≤ 1 and VP > 3 and ST < 30, in grey the remaining liquids (N/A).

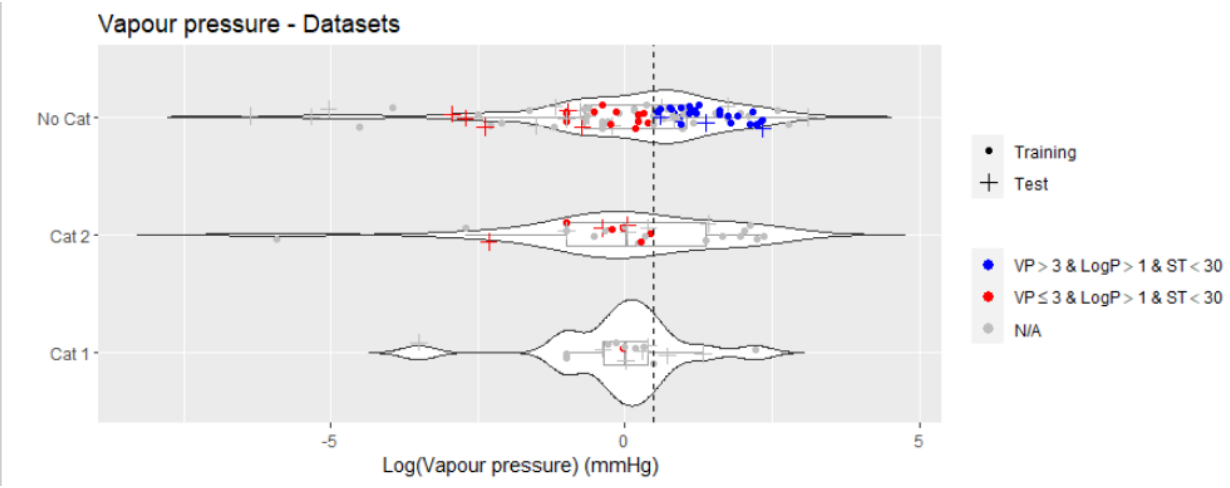


Figure 5.3. Distribution of Log(Vapour Pressure) showing the liquids used in the training set (●) and test set (+), in blue the liquids for which the exclusion criteria (LogP > 1 & VP, > 3 and ST < 30) were met, in red the liquids with VP ≤ 3 and LogP > 1 and ST < 30, in grey the remaining liquids.

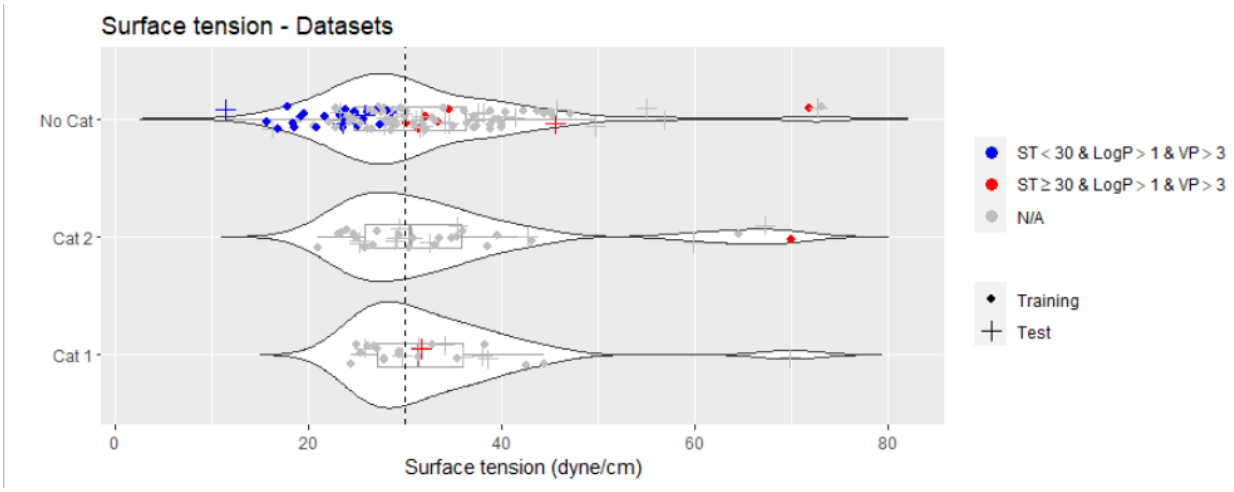


Figure 5.4. Distribution of ST showing the liquids used in the training set (●) and test set (+), in blue the liquids for which the exclusion criteria (LogP > 1 & VP, > 3 and ST < 30) were met, in red the liquids with ST ≥ 30 and LogP > 1 and VP > 3, in grey the remaining liquids.

5.1.3. Development of the DAL-1

44. The DAL-1 was developed based on the results of 46 non-surfactant neat liquids (training set) that were available for the different components of the DAL-1. The training set contained the liquids that were tested during the CON4EI project. There were only 7 No Cat. liquids from the subgroup CO = 0 that

were tested during the CON4EI project (Adriaens et al., 2018), so an additional 13 No Cat. chemicals were selected from the subgroups: CO = 0, CO > 0, and CO > 0** (an explanation of the subgroups is provided in § 3).

45. In a next step, the performance of DAL-1 was assessed for the test set. Regarding the selection of chemicals for the test set, nearly the maximum number of Cat. 1 and Cat. 2 non-surfactant liquids identified from the DRD publication (Barroso et al., 2017) were included in the test set. No changes were made to the DIP after assessing the performance of the test set since no further improvement to the DIP was possible based on the performance of the training and test set results. The identification of the substances that were used in the training set and the test set is available in Annex B (spreadsheet Annex_B.2) of the current background review document. The distribution of the liquids by UN GHS category and chemical set is provided in Table 5.4.

Table 5.4. Distribution of the reference chemicals: number of chemicals tested

UN GHS	Training set	Test seta	Total seta
Cat. 1	14	3/3 (3)	17/17 (17)
Cat. 2	12	10/11 (10)	22/23 (22)
No Cat.	20	35/26 (13)	55/46 (33)
Total	46	48/40 (26)	94/86 (72)

Note: [a] n/n (n): number of liquids tested with DAL-1 with VRM1 / DAL-1 with VRM2 (number of liquids tested in common)
The RhCE EIT test methods using VRM1 and VRM2 are able to correctly identify chemicals not requiring classification and labelling for eye irritation or serious eye damage according to UN GHS (OECD TG 492, 2019). Based on the validation studies and their independent peer review, similar performance in terms of reproducibility and reliability was reported for both VRMs.

46. The full set of substances evaluated with DAL-1 (total 108 different substances, 72 were tested in common, 94 with VRM1 and 86 with VRM2) is reported in Annex B (spreadsheet Annex_B.3). Summary statistics describing the space of the chemicals tested using DAL-1 are provided in Table 5.5.

Table 5.5. Summary of the physicochemical property ranges that describe the chemical space of the chemicals tested using DAL-1

Property	DAL-1 VRM1 Min-Max	DAL-1 VRM2 Min-Max
MW(g/mol)	58 – 596	58 – 596
logP	-1.75 – 7.59	-1.87 – 7.43
Water Solubility (mg/mL)	0 – 1000	0 – 1000
Vapour pressure (mmHg)	0 – 221	0 – 180
Surface tension of the neat liquids (dyne/cm)	0 – 71.6	0 – 71.6

47. For the set of 108 substances, high quality Draize eye test data were described in Supplementary Material 1 (https://static-content.springer.com/esm/art%3A10.1007%2Fs00204-016-1679-x/MediaObjects/204_2016_1679_MOESM1_ESM.pdf) of the Cosmetics Europe Draize eye test Reference Database (Barroso et al., 2017).

5.1.4. Predictive capacity for the overall set[1]

48. The predictive performance considering the three UN GHS categories (Cat. 1, Cat. 2, No Cat.) of DAL-1 is reported for 94 liquids with VRM1 (Table 5.6) and for 86 liquids with VRM2 (Table 5.7), respectively. Details on the calculations of the class-specific performance metrics for the 3x3 contingency matrix are provided in Annex C. Results of the BCOP LLBO test method do not exist for 40/55 (VRM1) and 34/46 (VRM2) No Cat. liquids respectively. As such, it was not possible to identify the final prediction for *in vivo* No Cat. liquids that are mispredicted (positive result for RhCE and exclusion criteria not met). It is however very unlikely that a large number of mis-predictions will be predicted Cat. 1 by the BCOP LLBO. This assumption is based on BCOP LLBO data that were published by Adriaens and co-authors (2020) and results on the proficiency chemicals (not published). Out of 22 *in vivo* No Cat. liquids, no liquid was mispredicted as Cat. 1 based on the BCOP LLBO (opacity (Lux/7) < 145). Furthermore, only 1 out of 24 No Cat. solids (4.2%) was mispredicted as Cat. 1. These data suggest a low likelihood of No Cat. mispredictions as Cat. 1 by the BCOP LLBO.

Table 5.6. Performance of the DAL-1 based on physicochemical properties, VRM1 and BCOP LLBO (N = 94 liquids)

UN GHS	DAL-1 with VRM		
	Cat 1	Cat 2	No Cat
Cat. 1 (N=17), % [a] (n/N)	**76.5% (13.0/17.0)[a]**	23.5% (4.0/17.0)	0.0% (0.0/17.0)
Cat. 2 (N=22) , % [a] (n/N)	27.3% (6.0/22.0)	**59.1% (13.0/22.0)**	13.6% (3.0/22.0)
No Cat. (N=55) , % [a] (n/N)	29.5% (16.2/55.0) [b]		70.5% (38.8/55.0)
68.7% balanced accuracy			

[a] The proportions in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1). See Annex D for explanation. To improve the readability of the numbers in the table, the numbers n/N have been rounded, so they may deviate slightly from the percentage corresponding to the weighted calculation. [b] BCOP LLBO data are not available for the majority (40/55) of the in vivo No Cat. liquids, as such, for liquids that were not identified as No Cat. with VRM1, it was not possible to distinguish between Cat. 1 (based on BCOP LLBO) and Cat. 2. Therefore, the false positives are presented between Cat. 1 and Cat. 2

Note: The performance is the same for the two versions of the DIP (Fig. 2.1 and Fig 2.2).

Class-specific performance metrics

	Cat. 1	**Cat. 2**	**No Cat.**
Correct within-class predictions (%)	76.5	59.1	70.5
Correct outside-class predictions (%)	92.2	73.1	92.3
Balanced Accuracy (%)	84.4	66.1	81.4

[1] The predictive capacity for the training set and the test set was similar, and thus the results for the overall set are provided.

Note: The balanced accuracy is the average of the correct within-class predictions and the correct outside-class predictions. The formulas for the class-specific performance metrics are provided in Annex C.

Table 5.7. Performance of the DAL-1 based on physicochemical properties, VRM2 and BCOP LLBO (N = 86 liquids)

UN GHS	DAL-1 with VRM2		
	Cat. 1	Cat. 2	No Cat.
Cat. 1 (N=17), % [a] (n/N)	**76.5% (13.0/17.0)[a]**	23.5% (4.0/17.0)	0.0% (0.0/17.0)
Cat. 2 (N=23), % [a] (n/N)	30.4% (7.0/23.0)	68.7% (15.8/23.0)	0.9% (0.2/23.0)
No Cat. (N=46), % [a] (n/N)	20.3% (9.3/46.0) [b]		79.7% (36.7/46.0)
75.0% balanced accuracy			

[a] The proportions in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1). To improve the readability of the numbers in the table, the numbers n/N have been rounded, so they may deviate slightly from the percentage corresponding to the weighted calculation.
[b] BCOP LLBO data are not available for the majority (34/46) of the in vivo No Cat. liquids, as such, for liquids that were not identified as No Cat. with VRM2, it was not possible to distinguish between Cat. 1 (based on BCOP LLBO) and Cat. 2. Therefore, the false positives are presented between Cat. 1 and Cat. 2
Note: The performance is the same for the two versions of the DIP (Fig. 2.1 and Fig 2.2).

Class-specific performance metrics

	Cat. 1	Cat. 2	No Cat.
Correct within-class predictions (%)	76.5	68.7	79.7
Correct outside-class predictions (%)	89.9	79.3	99.5
Balanced Accuracy (%)	83.2	74.0	89.6

Note: The balanced accuracy is the average of the correct within-class predictions and the correct outside-class predictions. The formulas for the class-specific performance metrics are provided in Annex C.

49. Detailed information on the proportion of correct and mispredicted liquids by DAL-1 is reported in Annex B (spreadsheet Annex_B.3;)

5.1.5. Limitations of individual sources of information

50. The strengths and limitations on individual test methods are described in the corresponding OECD Test Guidelines (OECD TG 437 and TG 492). The limitations of the individual methods for measuring the physicochemical properties are specified in their respective GLs. Assessment on how accuracy of predictions from the different QSAR models (OPERA and T.E.S.T.) can be evaluated, is provided in Annex E.

51. It is important to separate these limitations into:

- technical limitations
- limitations in the predictivity for UN GHS categories

52. The technical limitations may make a chemical not testable in one or more component methods of DAL-1 and may thus limit its applicability domain.

53. The predictivity limitations of some individual test methods for UN GHS categories do not necessarily limit the predictivity of an overarching DA; one of the advantages of DAs is that they are designed to overcome predictivity limitations of single test methods, i.e. the DAs can predict Cat. 2.

5.2. DAL-2

5.2.1. Development of the DIP

54. The DAL-2 was developed based on the results of 45 non-surfactant liquids (training set) that were available for the different components of the DAL-2. The training set contained the liquids that were tested during the CON4EI project. There were only 7 No Cat. liquids from the subgroup CO = 0 that were tested during the CON4EI project (Adriaens et al., 2018), so an additional 13 No Cat. chemicals were selected from the subgroups: CO = 0, CO > 0, and CO > 0** (an explanation of the subgroups is provided in § 3).

55. In a next step, the performance of DAL-2 was assessed for the test set. Regarding the selection of chemicals for the test set, nearly the maximum number of Cat. 1 and Cat. 2 non-surfactant liquids identified from the DRD publication (Barroso et al., 2017) were included in the test set. No changes were made to the DIP after assessing the performance of the test set. The identification of the substances that were used in the training set and the test set is available in Annex B (spreadsheet Annex_B.2) of the current background review document. The distribution of the liquids by UN GHS category and chemical set is provided in Table 5.8. Distribution if the reference chemicals: number of chemicals tested

UN GHS	Training set	Test set	Total set
Cat. 1	13	4	17
Cat. 2	12	12	24
No Cat.	20	103	123
Total	45	119	164

56.

Table 5.8. Distribution if the reference chemicals: number of chemicals tested

UN GHS	Training set	Test set	Total set
Cat. 1	13	4	17
Cat. 2	12	12	24
No Cat.	20	103	123
Total	45	119	164

57. The full set of liquids (neat and in dilution) evaluated with DAL-2 (total 164 different substances: 148 neat liquids and 16 liquids tested in dilution) is reported in Annex B (spreadsheet Annex_B.4).

58. For the set of 164 chemicals, high quality Draize eye test data were described in Supplementary Material 1 (https://static-content.springer.com/esm/art%3A10.1007%2Fs00204-016-1679-x/MediaObjects/204_2016_1679_MOESM1_ESM.pdf) of the Cosmetics Europe Draize eye test Reference Database (Barroso et al., 2017).

5.2.2. Predictive capacity for the overall set [1]

59. The predictive performance considering the three UN GHS categories (Cat. 1, Cat. 2, No Cat.) of DAL-2 is reported for 164 liquids (Table 5.9). Details on the calculations of the class-specific performance metrics for the 3x3 contingency matrix are provided in Annex C. Further, results of the BCOP LLBO test method do not exist for 108/123 No Cat. liquids. As such, it was not possible to identify the final prediction for *in vivo* No Cat. liquids that are mispredicted. It is however very unlikely that a large number of mispredictions will be predicted Cat. 1 by the BCOP LLBO. This assumption is based on BCOP LLBO data that were published by Adriaens and co-authors (2020) and results on the proficiency chemicals (not published). Out of 22 *in vivo* No Cat. liquids, no liquid was mispredicted as Cat. 1 based on the BCOP LLBO (opacity (Lux/7) < 145). Furthermore, only 1 out of 24 No Cat. solids (4.2%) was mispredicted as Cat. 1. These data suggest a low likelihood of No Cat. mispredictions as Cat. 1 by the BCOP LLBO.

Table 5.9. Performance of the DAL-2 based on STE and BCOP LLBO (N = 164 liquids)

UN GHS	DAL-2		
	Cat. 1	Cat. 2	No Cat.
Cat. 1 (N=17), % a (n/N)	**81.2% (13.8/17.0)**	17.6% (3.0/17.0)	1.2% (0.2/17.0)
Cat. 2 (N=24), % a (n/N)	30.2% (7.2/24.0)	**56.3% (13.5/24.0)**	13.5% (3.2/24.0)
No Cat. (N=123), % [a] (n/N)	14.7% (18.1/123.0) [b]		**85.3% (104.9/123.0)**
74.3% balanced accuracy			

[a] The proportions in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1). To improve the readability of the numbers in the table, the numbers n/N have been rounded, so they may deviate slightly from the percentage corresponding to the weighted calculation.
[b] BCOP LLBO data are not available for the majority (108/123) of the in vivo No Cat. liquids, as such, for liquids that were not identified as No Cat. with the modified STE test method, it was not possible to distinguish between Cat. 1 (based on BCOP LLBO) and Cat. 2. Therefore, the false positives are presented between Cat. 1 and Cat. 2
Note: The performance reported in Table 5-9 was obtained using the version of the DIP provided in Fig. 2.4.

Class-specific performance metrics

	Cat. 1	Cat. 2	No Cat.
Correct within-class predictions (%)	81.2	56.3	85.3
Correct outside-class predictions (%)	95.1	84.9	91.6
Balanced Accuracy (%)	88.1	70.6	88.5

Note: The balanced accuracy is the average of the correct within-class predictions and the correct outside-class predictions. The formulas for the class-specific performance metrics are provided in Annex C.

[1] The predictive capacity for the training set and the test set was similar, so the results for the overall set are provided.

5.2.3. Limitations of individual sources of information

60. The strengths and limitations on individual test methods are described in the corresponding OECD Test Guidelines (OECD TG 437 and TG 491).

61. It is important to separate these limitations into:

- technical limitations
- limitations in the predictivity for UN GHS categories

62. The technical limitations may make a chemical not testable in one or more component methods of DAL-2 and may thus limit its applicability domain.

63. The predictivity limitations of some individual test methods for UN GHS categories do not necessarily limit the predictivity of an overarching DA; one of the advantages of DAs is that they are designed to overcome predictivity limitations of single test methods, i.e. the DAs can predict Cat. 2.

5.3. References

Adriaens E, Verstraelen S, Alépée N, Kandarova H, Drzewiecka A, Gruszka K, Guest R, Willoughby JA, Van Rompay AR (2018). CON4EI: Development of testing strategies for hazard identification and labelling for serious eye damage and eye irritation of chemicals. Toxicol. in Vitro 49, 99–115. https://doi.org/10.1016/j.tiv.2017.09.008.

Adriaens E, Verstraelen S, Desprez B, Alépée N, Abo T, Bagley D, Hibatallah J, Mewes KR, Pfannenbecker U, Van Rompay AR (2020). Overall performance of Bovine Corneal Opacity and Permeability (BCOP) Laser Light-Based Opacitometer (LLBO) test method with regard to solid and liquid chemicals testing. Toxicol. in Vitro 70, 105-044

Alépée N, Adriaens E, Abo T, Bagley D, Desprez B, Hibatallah J, Mewes KR, Pfannenbecker U, Sala A, Van Rompay AR, Verstraelen S, McNamee P. (2019a). Development of a defined approach for eye irritation or serious eye damage for neat liquids based on Cosmetics Europe Analysis of *in vitro* RhCE and BCOP test methods. Toxicology in Vitro (2019) 59, 100-114. doi: 10.1016/j.tiv.2019.04.011.

Barroso J, Pfannenbecker U, Adriaens E, Alépée N, Cluzel M, De Smedt A, Hibatallah J, Klaric M, Mewes KR, Millet M, Templier M, McNamee P (2017). Cosmetics Europe compilation of historical serious eye damage/eye irritation *in vivo* data analysed by drivers of classification to support the selection of chemicals for development and evaluation of alternative methods/strategies: the Draize eye test Reference Database (DRD). Arch Toxicol 91, 521–547.

OECD (2019). Guideline for testing of chemicals no. 492: Reconstructed human cornea-like epithelium (RhCE) test method for identifying chemicals not requiring classification and labelling for eye irritation or serious eye damage. In: OECD Guidelines for the Testing of Chemicals, Section 4. Organisation for Economic Co-operation and Development, Paris. https://doi.org/10.1787/9789264242548-en.

OECD (2020a). Guideline for testing of chemicals no. 437: Bovine corneal opacity and permeability test method for identifying (i) chemicals inducing serious eye damage and (ii) chemicals not requiring classification for eye irritation or serious eye damage. In: OECD Guidelines for the Testing of Chemicals, Section 4. Organisation for Economic Co-operation and Development. https://doi.org/10.1787/9789264203846-en

OECD (2020b). Guideline for testing of chemicals no. 491: Short Time Exposure In Vitro Test Method for Identifying i) Chemicals Inducing Serious Eye Damage and ii) Chemicals Not Requiring Classification for Eye Irritation or Serious Eye Damage. In: OECD Guidelines for the Testing of Chemicals, Section

4. Organisation for Economic Co-operation and Development. https://doi.org/10.1787/9789264242432-en

Verstraelen, S., Jacobs, A., De Wever, B., Vanparys, P. (2013). Improvement of the Bovine Corneal Opacity and Permeability (BCOP) assay as an *in vitro* alternative to the Draize rabbit eye irritation test. Toxicol. in Vitro 27, 1298–1311.

6. Analyses of the DALs uncertainty and reproducibility

64. The objective of this analysis is to evaluate the uncertainty associated with the performance of the individual test methods (VRM1, VRM2, STE, and BCOP LLBO) and defined approaches DAL-1 (with VRM1 or with VRM2) and DAL-2. The aim was to assess the reproducibility of each information source, and how that propagates to the DAL-1 and DAL-2 overall.

65. The evaluation is based on 35 chemicals (DAL-1, Table 6.1) and 26 chemicals (DAL-2, Table 6.4) for which multiple results are available for each test method and are therefore suitable for reproducibility analysis. The reference benchmark is the UN GHS classification based on the Draize eye test (UN GHS column in Table 6.1 and Table 6.4, respectively).

66. A recognised method for incorporating uncertainty assessment into performance evaluation is also to apply a bootstrap approach. Bootstrapping is a resampling procedure by which a single dataset is randomly resampled over a high number of times. Each random resample is obtained from the original dataset (i.e., resampling with replacement) creating many simulated samples. Here, bootstrapping was used to produce a distribution of DA predictions based on 100,000 replicates and compared to benchmark reference classification (Draize eye reference data).

67. Example: Multiple predictions obtained by individual methods for 2-Methyl-1-pentanol (No. 20) are reported in Table 6.1. For VRM2 there are 11 classifications available based on existing data, i.e. 10 positive (Cat. 1/ Cat. 2) and 1 negative (No Cat.). Bootstrapping allows generation of "new" data either Cat. 1/Cat. 2 or No Cat. The probability of generating Cat. 1/Cat. 2 is proportional to the occurrence of Cat. 1/Cat. 2 in existing data, i.e. 10/11. Therefore, bootstrapping can be used to generate an arbitrarily large number of "new" classifications where the frequency of Cat. 1/Cat. 2 will be 10/11 (90.9%) whereas No Cat. 1/11 (9.1%) (similar to weighted calculation approach).

68. Bootstrapping is used to generate a full matrix of classifications for three single methods by resampling the data for all chemicals (N=35 for DAL-1 and N=26 for DAL-2). Resampling is repeated 100,000 times and resulting performance measures are averaged across 100,000 bootstrap replicates. The resulting performance values are shown in Table 6.2 (DAL-1) and Table 6.5 (DAL-2) and are based on predictions reported in Table 6.1 and Table 6.4 (35 and 26 chemicals with multiple results for the individual *in vitro* test methods). In addition, since the number of liquids with multiple results was rather small, the performance values for the set of liquids that were tested during the RhCE and STE validation studies and the BCOP LLBO evaluation study are reported in Table 6.3 and Table 6.6 respectively.

Table 6.1. Prediction for the individual test methods (proportion of correct predictions, TRUE pred. %).

DA predictions are derived by applying the associated data interpretation procedure (DIP) to predictions from a single method. [TRUE pred., proportion of correctly predicted results: VRM1 and VRM2 = No Cat. versus Cat. 1 + Cat. 2 and BCOP LLBO = Cat. 1 versus Cat. 2 + No Cat.]. The last two columns correspond with the proportions of correct predictions within each UN GHS Category (Cat. 1, Cat. 2 and No Cat.) for DAL-1 with VRM1 and DAL-1 with VRM2.

				SINGLE METHODS									DAL-1 VRM1	DAL-1 VRM2
	Chemicals	**CAS#**	**UN GHS**	**RhCE: VRM1**			**RhCE VRM2**			**BCOP LLBO Opacity >145**				
				Cat. 1 + Cat. 2 N	**No Cat. N**	**TRUE pred %**	**CAT. 1 + CAT. 2**	**NO CAT.**	**TRUE PRED**	**CAT. 1**	**CAT. 2 + NO CAT.**	**TRUE PRED %**	**TRUE pred %**	**TRUE pred %**
1	2-Hydroxy iso-butyric acid ethyl ester	80-55-7	Cat 1	2	0	100	11	0	100	2	0	100	100	100
2	Benzensulphonylchloride	98-09-9	Cat 1	2	0	100	2	0	100	2	0	100	100	100
3	bis-(3-Aminopropyl)-tetramethyldisiloxane	2469-55-8	Cat 1	2	0	100	2	0	100	2	0	100	100	100
4	Diethylethanolamine	100-37-8	Cat 1	3	0	100	2	0	100	2	0	100	100	100
5	Ethylhexyl acid phosphate ester	12645-31-7	Cat 1	2	0	100	2	0	100	2	0	100	100	100
6	Lactic acid	50-21-5	Cat 1	3	0	100	11	0	100	2	0	100	100	100
7	Hydroxyethyl acrylate	818-61-1	Cat 1	2	0	100	11	0	100	2	0	100	100	100
8	1-Chloroctan-8-ol	23144-52-7	Cat 1	2	0	100	11	0	100	0	2	0	0	0
9	(3-Aminopropyl)triethoxy silane	919-30-2	Cat 1	2	0	100	5	0	100	2	0	100	100	100
10	Methoxyethyl acrylate	3121-61-7	Cat 1	4	0	100	2	0	100	2	0	100	100	100
11	Methyl thioglycolate	2365-48-2	Cat 1	11	0	100	11	0	100	2	0	100	100	100
12	n-Octylamine	111-86-4	Cat 1	2	0	100	2	0	100	2	0	100	100	100
13	Tetrahydrofuran	109-99-9	Cat 1	3	0	100	2	0	100	2	0	100	100	100
14	Tributyltin oxide	56-35-9	Cat 1	2	0	100	2	0	100	0	2	0	0	0
15	2,6-Dichlorobenzoyl chloride	4659-45-4	Cat 2	3	0	100	9	0	100	0	3	100	100	100
16	Cyclopentanol	96-41-3	Cat 2	12	0	100	11	0	100	2	0	0	0	0
17	Propasol Solvent P	1569-01-3	Cat 2	11	0	100	11	0	100	2	0	0	0	0
18	iso-Propanol	67-63-0	Cat 2	4	0	100	5	0	100	1	1	50	50	50

19	Methyl cyanoacetate	105-34-0	Cat 2	1	2	33.3	5	0	100	0	2	100	33.3	100
20	2-Methyl-1-pentanol	105-30-6	Cat 2	13	0	100	10	1	90.9	0	2	100	100	90.9
21	3-Chloropropionitrile	542-76-7	Cat 2	13	0	100	11	0	100	1	1	50	50	50
22	Butyl Dipropasol Solvent	29911-27-1	Cat 2	12	0	100	11	0	100	0	2	100	100	100
23	Ethyl-2-methyl acetoacetate	609-14-3	Cat 2	10	3	76.9	8	1	88.9	0	3	100	76.9	88.9
24	iso-Butanal	78-84-2	Cat 2	11	0	100	11	0	100	0	2	100	100	100
25	n-Butanal	123-72-8	Cat 2	2	0	100	5	0	100	0	2	100	100	100
26	2-Propanol, 1-phenoxy-	770-35-4	Cat 2	2	0	100	2	0	100	0	2	100	100	100
27	2-Pseudoionone	141-10-6	Cat 2	0	2	0	2	0	100	0	2	100	100	100
28	iso-Propyl acetoacetate	542-08-5	Cat 2	10	1	90.9	11	0	100	0	2	100	90.9	100
29	1,2,6-Hexanetriol	106-69-4	No Cat	2	0	0	0	5	100	0	2	100	0	100
31	2-Ethylhexyl p-dimethylamino benzoate [a]	21245-02-3	No Cat	0	3	100	0	2	100	0	2	100	100	100
32	Dioctyl ether [a]	629-82-3	No Cat	0	11	100	0	11	100	0	2	100	100	100
33	Dodecane [a]	112-40-3	No Cat	0	4	100	0	2	100	0	2	100	100	100
34	Glycerol	56-81-5	No Cat	0	8	100	0	11	100	0	2	100	100	100
35	Propylene glycol	57-55-6	No Cat	0	3	100	0	11	100	0	2	100	100	100
31	2-Ethylhexyl p-dimethylamino benzoate [a]	21245-02-3	No Cat	0	3	100	0	2	100	0	2	100	100	100

Note: [a] Correspond with liquids for which the physicochemical exclusion rules (water solubility < 0.02 mg/mL) are met, those liquids were also predicted No Cat. with VRM1 and VRM2

Table 6.2. Performance measures based on 100,000 Bootstrap replicates, of individual methods and DAL-1 against UN GHS classifications (individual data Table 6.1)

	Reproducibility	Accuracy	Specificity	Sensitivity	
7 No Cat. / 28 Cat. 1 + Cat. 2					
RhCE: VRM1	97.2%	94.3%	85.7%	96.5%	
RhCE: VRM2	99.4%	99.4%	100%	99.3%	
14 Cat. 1 / 21 Cat. 2 + No Cat					
BCOP LLBO	97.1%	85.7%	85.7%	85.7%	

UN GHS (N=35)		Accuracy	TRUE No Cat. (N=7)	TRUE Cat. 2 (N=14)	TRUE Cat. 1 (N=14)
DAL-1 with VRM1	94.3%	80.0%	85.7%	71.5%	85.7%
DAL-1 with VRM2	96.6%	85.1%	100%	77.1%	85.7%

Table 6.3. Performance values from the validation studies (liquids only)

	Reproducibility WLR/BLR	**Accuracy**	**Specificity**	**Sensitivity**
No Cat. / Cat. 1 + Cat. 2	Mean values 3 independent runs in 3 laboratories			
VRM1 [a]	97.5% / 94.4% (N=53)	81.9% (N=53)	66.7% (N=27)	98.3% (N=26)
VRM2 [b]	92.2% / 93.3% (N=60)	84.4% (N=105)	68.5% (N=50)	99.0% (N=55)
Cat. 1 / Cat. 2 + No Cat.	Mean value 2 independent runs			
BCOP LLBO (based on opacity) [c]	93.2% / NA (44/NA)	79.2% (N=81)	81.3% (N=50)	76.0% (N=31)

[a] Barroso et al., 2014; [b] Alépée et al., 2016 ; [c] Adriaens et al., 2020

Table 6.4. Prediction for the individual test methods (proportion of correct predictions, TRUE pred.).

DA predictions are derived by applying the associated data interpretation procedure (DIP) to predictions from a single method. [TRUE pred., proportion of correctly predicted results: STE = No Cat. versus Cat. 1 + Cat. 2 and STE and BCOP LLBO = Cat. 1 versus Cat. 2 + No Cat.]. The last column corresponds with the proportions of correct predictions within each UN GHS Category (Cat. 1, Cat. 2 and No Cat.) for DAL-2.

				SINGLE METHODS									DAL-2
	Chemicals	**CAS#**	**UN GHS**	**STE**			**STE**			**BCOP LLBO Opacity >145**			
				Cat. 1 + Cat. 2	**No Cat.**	**TRUE pred %**	**Cat. 1**	**Cat. 2 + No Cat.**	**TRUE pred %**	**Cat 1**	**Cat. 2 + No Cat.**	**TRUE pred %**	**TRUE pred**
1	2-Hydroxy iso-butyric acid ethyl ester	80-55-7	Cat 1	2	0	100	0	2	0	2	0	100	100
2	Diethylethanolamine	100-37-8	Cat 1	2	0	100	0	2	0	2	0	100	100
3	Ethylhexyl acid phosphate ester	12645-31-7	Cat 1	2	0	100	2	0	100	2	0	100	100
4	Lactic acid	50-21-5	Cat 1	4	0	100	0	4	0	2	0	100	100
5	Hydroxyethyl acrylate	818-61-1	Cat 1	4	1	80	0	5	0	2	0	100	80
6	1-Chloroctan-8-ol	23144-52-7	Cat 1	2	0	100	0	2	0	0	2	0	0
7	(3-Aminopropyl)triethoxy silane	919-30-2	Cat 1	2	0	100	0	2	0	2	0	100	100
8	Methoxyethyl acrylate	3121-61-7	Cat 1	4	0	100	0	4	0	2	0	100	100
9	n-Octylamine	111-86-4	Cat 1	2	0	100	2	0	100	2	0	100	100
10	2,6-Dichlorobenzoyl chloride	4659-45-4	Cat 2	0	2	0	0	2	100	0	3	100	0
11	Cyclopentanol	96-41-3	Cat 2	4	0	100	0	4	100	2	0	0	0
12	Propasol Solvent P	1569-01-3	Cat 2	3	1	75	0	4	100	2	0	0	0
13	Methyl cyanoacetate	105-34-0	Cat 2	2	2	50	0	4	100	0	2	100	50
14	Butyl Dipropasol Solvent	29911-27-1	Cat 2	3	1	75	0	4	100	0	2	100	75
15	Ethyl-2-methyl acetoacetate	609-14-3	Cat 2	4	0	100	0	4	100	0	3	100	100
16	iso-Butanal	78-84-2	Cat 2	4	0	100	0	4	100	0	2	100	100
17	n-Butanal	123-72-8	Cat 2	3	0	100	0	3	100	0	2	100	100
18	2-Pseudoionone	141-10-6	Cat 2	2	0	100	0	2	100	0	2	100	100
19	iso-Propyl acetoacetate	542-08-5	Cat 2	2	0	100	0	2	100	0	2	100	100
20	Glycolic acid (10%)	79-14-1	Cat 2	2	0	100	0	2	100	2	0	0	0

21	1,2,6-Hexanetriol	106-69-4	No Cat	0	2	100	0	2	100	0	2	100	100
22	1-Ethyl-3-methylimidazolium ethyl sulphate	342573-75-5	No Cat	0	2	100	0	2	100	0	2	100	100
23	2-Ethylhexyl p-dimethylamino benzoate	21245-02-3	No Cat	0	14	100	0	14	100	0	2	100	100
24	Dodecane	112-40-3	No Cat	0	2	100	0	2	100	0	2	100	100
25	Glycerol	56-81-5	No Cat	0	10	100	0	10	100	0	2	100	100
26	Propylene glycol	57-55-6	No Cat	0	9	100	0	9	100	0	2	100	100

Table 6.5. Performance measures based on 100,000 Bootstrap replicates, of individual methods and DAL-2 against UN GHS classifications (individual data Table 6.4)

	Reproducibility	Accuracy	Specificity	Sensitivity	
6 No Cat. / 20 Cat. 1 + Cat. 2					
STE	95.4%	91.5%	100%	89.0%	
9 Cat. 1 / 17 Cat. 2 + No Cat.					
STE	100%	73.1%	100%	22.0%	
BCOP LLBO	100%	84.6%	82.4%	88.9%	
UN GHS (N=26)		**Accuracy**	**TRUE No Cat. (N=6)**	**TRUE Cat. 2 (N=11)**	**TRUE Cat. 1 (N=9)**
DAL-2	100%	77.1%	100%	72.7%	86.7%

Table 6.6. Performance values from the validation studies (liquids only)

	Reproducibility WLR	Accuracy	Specificity	Sensitivity
No Cat. / Cat. 1 + Cat. 2				
STE [a]	NA	85.9% (N=92)	82.0% (N=50)	90.5% (N=42)
Cat. 1 / Cat. 2 + No Cat.				

STE [a]	NA	85.4% (N=89)	98.6% (N=72)	29.0% (N=17)
BCOP LLBO (based on opacity) [b]	93.2% (N=44, 2 independent runs)	79.2% (N=81)	81.3% (N=50)	76.0% (N=31)

[a] STE review document (ICCVAM, 2013); [b] Adriaens et al., 2020

6.1. References

1. Adriaens E, Verstraelen S, Desprez B, Alépée N, Abo T, Bagley D, Hibatallah J, Mewes KR, Pfannenbecker U, Van Rompay AR (2020). Overall performance of Bovine Corneal Opacity and Permeability (BCOP) Laser Light-Based Opacitometer (LLBO) test method with regard to solid and liquid chemicals testing. Toxicol. in Vitro 70, 105-044
2. Alépée N, Leblanc V, Adriaens E, Grandidier MH, Lelièvre D, Meloni M, Nardelli L, Roper CS, Santirocco E, Toner F, Van Rompay AR, Vinall J, Cotovio J (2016). Multi-laboratory validation of SkinEthic HCE test method for testing serious eye damage/eye irritation using liquid chemicals. Toxicol. in Vitro 31, 43-53. http://dx.doi.org/10.1016/j.tiv.2015.11.012
3. Barroso J. (2014). The EURL ECVAM - Cosmetics Europe prospective validation study of Reconstructed human Cornea-like Epithelium (RhCE)-based test methods for identifying chemicals not requiring classification and labelling for serious eye damage/eye irritation. EUR - Scientific and Technical Research Reports, Publications Office of the European Union
4. ICCVAM (2013). Short Time Exposure (STE) Test Method Summary Review Document, NIH. Available at: [http://www.ntp.niehs.nih.gov/iccvam/docs/ocutox_docs/STE-SRD-NICEATM-508.pdf]

7. Detailed performance analysis of individual methods and DAs against Draize eye test

69. This chapter analyses the performance of the individual methods EpiOcular™ EIT (Validated Reference Method, VRM1), SkinEthic™ HCE EIT (VRM2), BCOP LLBO, STE and that of DAL-1 and DAL-2, against the curated Draize Eye test reference data. The following methods and DAs were analysed:

- EpiOcular™ EIT (RhCE) = VRM1
- SkinEthic™ HCE EIT (RhCE) = VRM2
- BCOP LLBO
- STE
- DAL-1 with VRM1
- DAL-1 with VRM2
- DAL-2

70. The performance of these methods with respect to the whole dataset, by driver of classification, and by chemical class, is presented in the next chapters, with a specific focus on mispredictions. The analyses are meant to provide considerations and support recommendations regarding the use of the different DALs based on the performance observed in this dataset. More details regarding the drivers of classification and the OFG are provided in section 3.2. (Key criteria for evaluation of the DALs versus the *in vivo* Draize eye test) and section 7.3 (Analysis of the performance for specific Organic Functional Groups (OFG) with DALs).

7.1. All chemicals

71. The full set of substances evaluated with DAL-1 (total 108 different substances, 72 were tested in common, 94 with VRM1 and 86 with VRM2) is reported in Annex B (spreadsheet Annex_B.3). The full set of liquids evaluated with DAL-2 (total 164 different substances: 147 neat liquids and 8 liquids and 9 solids dissolved in water) is reported in Annex B (spreadsheet Annex_B.4).

72. The prevalence of *in vivo* classified liquids (i.e., UN GHS Cat. 1 and Cat. 2) is 41.5% (39/94) and 46.5% (40/86) for DAL-1 with VRM1 and VRM2, respectively. The prevalence of *in vivo* classified liquids (i.e., UN GHS Cat. 1 and Cat. 2) is 25% (41/164) for DAL-2. It should be

noted that 37 *in vivo* classified liquids (16 Cat. 1 and 21 Cat. 2) and 31 *in vivo* No Cat. liquids were tested in common with DAL-1 and DAL-2.

73. The performance of the individual test methods for identifying chemicals not requiring classification for eye irritation or serious eye damage (UN GHS No Cat.) is shown in Table 7.1. The individual test methods have accuracies which range from 72.0-86.9%, with specificities and sensitivities which range from 57.6-85.3% and 91.6-99.5%, respectively.

Table 7.1. Predictive Capacity of individual in vitro test methods for identifying chemicals not requiring classification for eye irritation or serious eye damage [UN GHS No Cat. versus Not No Cat. (Cat. 1 + Cat. 2)]

	Accuracy		UN GHS Cat. 1 + Cat. 2			UN GHS No Cat.		
	N	% [a]	N	Sensitivity (%)	FN (%)	N	Specificity (%)	FP (%)
VRM1	94	72.0	39	92.3	7.7	55	57.6	42.4
VRM2	86	79.7	40	99.5	0.5	46	63.3	37.4
STE	164	86.9	41	91.6	8.4	123	85.3	14.7

[a] The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1).

74. The performance of the individual test methods for identifying chemicals inducing serious eye damage (UN GHS Cat. 1) is shown in Table 7.2.

Table 7.2. Predictive Capacity of individual in vitro test methods for identifying chemicals inducing serious eye damage [UN GHS Cat. 1 versus Not Cat. 1 (Cat. 2 + No Cat.)]

	Accuracy		UN GHS Cat. 1			UN GHS Cat. 2 + No Cat.		
	N	% [a]	N	Sensitivity (%)	FN (%)	N	Specificity (%)	FP (%)
BCOP LLBO	56	79.5	17	76.5	23.5	39	80.8	19.2
STE	164	91.5	17	23.5	76.5	147	99.4	0.6

[a] The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1)

75. An overview of the liquids which are mispredicted by DAL-1 with VRM1, DAL-1 with VRM2, and DAL-2 are listed in Table 7.3. In *vivo* Cat. 1 liquids which are under-predicted are the result of an under-prediction by the BCOP LLBO (DAL-1) or an under-prediction by the BCOP LLBO and/or STE (DAL-2). One liquid (tributyltin oxide; CASRN 56-35-9) was under-predicted with the BCOP LLBO (DAL-1) but was correctly predicted with the STE (DAL-2). *In vivo* Cat. 2 liquids which are over-predicted are the result of an over-prediction by the BCOP LLBO (DAL-1 and DAL-2), none of the *in vivo* Cat. 2 liquids are over-predicted by the STE test method. False negative *in vivo* Cat. 2 liquids and false positive *in vivo* No Cat. liquids are the result of a misprediction by VRM1 (DAL-1), VRM2 (DAL-1) or STE (DAL-2).

Table 7.3. Mis-predicted liquids in comparison with UN GHS categories

DRD No.	Chemical	CASRN	UN GHS	Driver	DAL-1 prediction with VRM1	DAL-1 prediction with VRM2	DAL-2 prediction
5	Cyclohexanol	108-93-0	Cat 1	CO mean ≥ 3	UP (1)	UP (1)	UP (1)
34	Hydroxyethyl acrylate	818-61-1	Cat 1	CO pers D21	TP (1)	TP (1)	TP/FN[a] (0.80/0.20)
47	1-Chloroctan-8-ol	23144-52-7	Cat 1	CO pers D21	UP (1)	UP (1)	UP (1)
49	Benzyl alcohol	100-51-6	Cat 1	CO pers D21	UP (1)	UP (1)	UP (1)
133	Tributyltin oxide	56-35-9	Cat 1	CO = 4	UP (1)	UP (1)	TP (1)
166	2,6-Dichlorobenzoyl chloride	4659-45-4	Cat 2	CO mean ≥ 1	TP (1)	TP (1)	FN (1)
167	2-Ethyl-1-hexanol	104-76-7	Cat 2	CO mean ≥ 1	TP (1)	TP (1)	TP/FN (0.75/0.25)
168	Acetone	67-64-1	Cat 2	CO mean ≥ 1	OP (1)	OP (1)	OP (1)
170	Cyclopentanol	96-41-3	Cat 2	CO mean ≥ 1	OP (1)	OP (1)	OP (1)
172	Ethyl trans-3-ethoxyacrylate	5941-55-9	Cat 2	CO mean ≥ 1	FN (1)	TP (1)	FN (1)
176, 225	Methyl acetate	79-20-9	Cat 2	CO mean ≥ 1	OP (1)	OP (1)	OP (1)
177	Methyl ethyl ketone	78-93-3	Cat 2	CO mean ≥ 1	OP (1)	OP (1)	OP (1)
183	Propasol Solvent P	1569-01-3	Cat 2	CO mean ≥ 1	OP (1)	OP (1)	OP/FN (0.75/0.25)
200	Allyl alcohol	107-18-6	Cat 2	CO mean ≥ 1	-	OP (1)	-
203	iso-Propanol	67-63-0	Cat 2	Conj mean ≥ 2	OP/TP (0.50/0.50)	OP/TP (0.50/0.50)	OP/TP (0.50/0.50)
204	Methyl cyanoacetate	105-34-0	Cat 2	Conj mean ≥ 2	TP/FN (0.33/0.67)	TP (1)	TP/FN (0.50/0.50)
210	Sodium hydroxide (1%)	1310-73-2	Cat 2	Conj mean ≥ 2	-	-	OP (1)
218	2-Methyl-1-pentanol	105-30-6	Cat 2	CO mean ≥ 1	TP (1)	TP/FN (0.91/0.09)	TP (1)
219	3-Chloropropionitrile	542-76-7	Cat 2	CO mean ≥ 1	OP/TP (0.50/0.50)	OP/TP (0.50/0.50)	-
221	Butyl Dipropasol Solvent	29911-27-1	Cat 2	CO mean ≥ 1	TP (1)	TP (1)	TP/FN (0.75/0.25)
223	Ethyl-2-methyl acetoacetate	609-14-3	Cat 2	CO mean ≥ 1	TP/FN (0.77/0.23)	TP/FN (0.89/0.11)	TP (1)
233	2-Pseudoionone	141-10-6	Cat 2	Conj mean ≥ 2	FN (1)	TP (1)	TP (1)
234	iso-Propyl acetoacetate	542-08-5	Cat 2	Conj mean ≥ 2	TP/FN (0.91/0.09)	TP (1)	TP (1)
237	Glycolic acid (10%)	79-14-1	Cat 2	Conj mean ≥ 2	-	-	OP (1)
245	2,2-Dimethyl-3-pentanol	3970-62-5	No Cat	CO > 0 **	FP (10)	FP (9)	TN (2)
247	3-Phenoxybenzyl alcohol	13826-35-2	No Cat	CO > 0 **	FP (1)	FP (1)	-
251	Cyclohexanone	108-94-1	No Cat	CO > 0 **	FP (1)	-	FP (1)
252	Ethyl thioglycolate	623-51-8	No Cat	CO > 0 **	FP (1)	FP (1)	TN (1)
254	Glycidyl methacrylate	106-91-2	No Cat	CO > 0 **	FP (1)	FP (1)	FP (1)
255	Lactic acid (10%)	50-21-5	No Cat	CO > 0 **	-	-	FP (1)
257, 284	Methyl amyl ketone	110-43-0	No Cat	CO > 0 **	TN (1)	-	FP/TN (0.13/0.87)
276	2,4-Pentanedione	123-54-6	No Cat	CO > 0	FP (1)	-	FP (1)
277	2-Ethylhexanoyl chloride	760-67-8	No Cat	CO > 0	-	-	FP (1)
279	Dimethyl carbonate	616-38-6	No Cat	CO > 0	-	-	FP (1)
280	Dimethyl sulfoxide	67-68-5	No Cat	CO > 0	FP (1)	TN/FP (0.33/0.67)	TN (1)
281	Ethyl acetate	141-78-6	No Cat	CO > 0	FP (1)	FP (1)	TN (1)
282	Ethyl acetoacetate	141-97-9	No Cat	CO > 0	-	-	FP (1)

288	Propiconazole	60207-90-1	No Cat	CO > 0	FP (1)	-	FP (1)
292, 468	Triethanolamine	102-71-6	No Cat	CO > 0 // CO = 0	FP (1)	-	TN (1)
329	1,2,6-Hexanetriol	106-69-4	No Cat	CO = 0	FP (1)	TN (1)	TN (1)
330	1,3-Dibromopropane	109-64-8	No Cat	CO = 0	-	FP (1)	TN (1)
338	1-Bromo-4-chlorobutane	6940-78-9	No Cat	CO = 0	FP (1)	TN/FP (0.78/0.22)	TN (1)
341	1-Nitropropane	108-03-2	No Cat	CO = 0	FP (1)	-	-
343	2,4-Dicyano-1-butene	1572-52-7	No Cat	CO = 0	-	-	FP (1)
345	2-(2-Ethoxy ethoxy) ethanol	111-90-0	No Cat	CO = 0	FP (1)	TN/FP (0.56/0.44)	TN (1)
347	2-Ethoxyethyl methacrylate	2370-63-0	No Cat	CO = 0	TN/FP (0.22/0.78)	FP (1)	TN (1)
351	2-Methoxy-3,4-dihydropyran	4454-05-1	No Cat	CO = 0	-	-	FP (1)
390	gamma-Chloropropyltrimethoxy silane	2530-87-2	No Cat	CO = 0	-	-	FP (1)
391, 392	gamma-Glycidyloxypropyltrimethoxy silane	2530-83-8	No Cat	CO = 0	FP (1)	-	TN (1)
397	Tris(2-chloroethyl) phosphate)	115-96-8	No Cat	CO = 0	-	-	FP (1)
398	Glycediol	556-52-5	No Cat	CO = 0	-	-	FP (1)
409	iso-Nonylaldehyde	5435-64-3	No Cat	CO = 0	-	-	FP (1)
429	Methyl triglycol	112-35-6	No Cat	CO = 0	-	TN/FP (0.33/0.67)	TN (1)
432	m-Methoxthybenzaldehyde		No Cat	CO = 0	-	-	FP (1)
458	p-Methyl thiobenzaldehyde	3446-89-7	No Cat	CO = 0	TN/FP (0.56/0.44)	TN/FP (0.67/0.33)	-
470	Triethylene glycol	112-27-6	No Cat	CO = 0	-	TN/FP (0.67/0.33)	TN (1)
482	Vinyltrimethoxy silane	2768-02-7	No Cat	CO = 0	-	-	FP (1)
485	Acid Red 92 (1%)	18472-87-2	No Cat	CO = 0	-	-	FP (1)
496	Trichloroacetic acid (3%)	76-03-9	No Cat	CO = 0	-	-	FP (1)

[a] five results are available for the STE test method with 1/5 a No Cat. prediction (FN) and 4/5 a No

Prediction Can be Made. This liquid was correctly identified as Cat. 1 with the BCOP LLBO

TN: True Negatives (cells with grey background); TP: True Positives (cells with grey background); FN: False Negatives (in vivo Cat. 2 predicted as No Cat.); UP: Under-Predicted chemicals (in vivo Cat. 1 predicted as Cat. 2); FP: False Positives (in vivo No Cat. predicted as Cat. 2); OP: Over-Predicted chemicals (in vivo Cat. 2 predicted as Cat. 1); the number in parentheses corresponds with the proportion of the prediction, for some chemicals two fractions are provided, this is because the predictions differ between the in vitro study result available.

CO > 0: in at least one observation time in at least one animal and all animals showing mean scores of days 1–3 below the classification cut-offs for all endpoints, ** Indicates at least one animal with a mean score of days 1–3 above the classification cut-off for at least one endpoint (see §18).

7.2. Analyses of the performance by driver of classification

76. This section focuses on the performance of the individual test methods and DALs by driver of classification. Details on the driver of classification are presented in Chapter 3. The results of the individual test methods are presented in tables summarising the TP (sensitivity),

TN (specificity) and accuracy as compared to the Draize eye test benchmark data. The results of the DALs are presented in tables summarising TP (true Cat. 1 and true Cat. 2), TN (true No Cat.), over-predictions (OP, *in vivo* Cat. 2 predicted as Cat. 1), under-predictions (UP, *in vivo* Cat. 1 predicted as Cat. 2), FN (*in vivo* Cat. 1 or Cat. 2 predicted as No Cat.) and accuracy as compared to the Draize eye test benchmark data.

77. VRM1, VRM2, and the STE test method can be used to identify chemicals that do not require classification for eye irritation or serious eye damage with DAL-1 or DAL-2. The TN rate per No Cat. subgroup is shown in Table 7.4. No Cat. liquids from the subgroup CO > 0 and CO > 0 ** resulted in 100% FPs for VRM1 and 58.3-100% FPs for VRM2 whereas the FP rate for the STE test method range from 20.8-33.3%. A lower FP rate was observed for the subgroup CO = 0 with 21.9% and 26.5% for VRM1 and VRM2 respectively and 11.0% for the STE test method. FN results were not observed for the most important drivers of Cat. 1 classification (one exception for the STE test method, a FN prediction was observed for 1/5 test results for hydroxyethyl acrylate, CASRN 818-61-1, Table 7.3). Differences in TP rate for the Cat. 2 drivers of classification were observed between the different test methods.

Table 7.4. Predictive Capacity of individual in vitro test methods for identifying chemicals not requiring classification for eye irritation or serious eye damage [UN GHS No Cat. (True Negative, TN) versus Not No Cat. (Cat. 1 + Cat. 2 = True Positive, TP)]

Parameter	UN GHS	VRM1		VRM2		STE	
	Driver of classification	**N**	**Correct prediction [a]**	**N**	**Correct prediction [a]**	**N**	**Correct prediction [a]**
TP	**Cat. 1**	**17**	100	**17**	100	**17**	98.8
	CO mean ≥ 3	7	100	7	100	7	100
	CO pers D21	4	100	4	100	5	96
	CO = 4	6	100	6	100	5	100
	Cat. 2	**22**	86.4	**23**	99.1	**24**	86.5
	CO mean ≥ 1	17	92.8	18	98.9	17	83.8
	Conj mean ≥ 2	5	64.8	5	100	7	92.9
TN	**No Cat.**	**55**	57.6	**46**	63.3	**123**	85.3
	CO > 0 **	7	0	5	0	15	79.2
	CO > 0	8	0	4	41.7	15	66.7
	CO = 0 **	2	100	1	100	2	100
	CO = 0	38	78.1	36	73.5	91	89.0
Accuracy		94	72.0	86	80.1	164	86.9

[a] The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1).

** Indicates at least one animal with a mean score of days 1–3 above the classification cut-off for at least one endpoint.

78. The BCOP LLBO and the STE test methods can be used to identify chemicals requiring classification for serious eye damage with DAL-1 (BCOP LLBO) or DAL-2 (BCOP LLBO and STE). The TP rate and TN rate by driver of classification is shown in Table 7.5. A high FN rate was found in STE, with 100% for the driver CO pers D21 and 71.4% and 60% for

the drivers CO mean ≥ 3 and CO = 4, respectively. On the other hand, the FP rate of the STE was almost 0%. The FN rate for the BCOP LLBO was low for the drivers CO mean ≥ 3 (FN = 14.3%) and CO = 4 (16.7%) and was 40% for CO pers D21. The FP rate for both drivers of Cat. 2 classification was around 35% for the BCOP LLBO.

Table 7.5. Predictive Capacity of individual in vitro test methods for identifying chemicals inducing serious eye damage [UN GHS Cat. 1 (True Positive, TP) versus Not Cat. 1 (Cat. 2 + No Cat. = True Negative, TN)]

Parameter	UN GHS Driver of classification	LLBO		STE	
		N	Correct prediction [a]	N	Correct prediction [a]
TP	**Cat 1**	**18**	77.8	**17**	23.5
	CO mean ≥ 3	7	85.7	7	28.6
	CO pers D21	5	60.0	5	0.0
	CO = 4	6	83.3	5	40.0
TN (Not Cat. 1)	**Cat 2**	**26**	65.4	**24**	100
	CO mean ≥ 1	19	65.8	17	100
	Conj mean ≥ 2	7	64.3	7	100
	No Cat	**15**	100	**123**	99.3
	CO > 0 **	-	-	15	94.2
	CO > 0	-	-	15	100
	CO = 0 **	-	-	2	100
	CO = 0	16	100	91	100
Accuracy		**59**	78.0	**164**	91.5

a The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1).

** Indicates at least one animal with a mean score of days 1–3 above the classification cut-off for at least one endpoint.

79. The performance by driver of classification (Cat. 1 and Cat. 2) or by subgroup (No Cat.) with DAL-1 and DAL-2 is shown in Table 7.6. The UP rate for the Cat. 1 drivers of classification CO mean ≥ 3 and CO = 4 was low and ranged from 0-16.7%. The UP (or FN, see §76 for explanation) for CO pers D21 was 44% with DAL-2 and 50.0% with DAL-1. The TP rate for Cat. 2 liquids that were classified based on CO mean ≥ 1 was similar for the different DALs and ranged from 55.9-62.8%. A more important difference between the DALs was observed for Cat. 2 liquids that were classified based on Conj mean ≥ 2, where the TP rate ranged from 54.8-90.0%. The FP rate for liquids from the subgroup CO = 0 was low for the different DALs and ranged from 11.0-16.4%. Differences in performance between the DALs were observed for the subgroups CO > 0 ** and CO > 0.

Table 7.6. Predictive performance considering the three UN GHS categories (Cat. 1, Cat. 2, No Cat.) of DALs with BCOP LLBO

Cat. 1	CO mean ≥ 3		CO pers D21		CO=4	
	DAL-1	DAL-2	DAL-1	DAL-2	DAL-1	DAL-2
N	**7**	**7**	**4**	**5**	**6**	**5**

TP (%) [a]	85.7	85.7	50.0	56.0	83.3	100
UP (%) [a]	14.3	14.3	50.0	40.0	16.7	0
FN (%) [a]	0	0	0	4.0	0	0

Cat. 2	CO mean ≥ 1		Conj mean ≥ 2		Cat. 2	
	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2
N	**17**	**18**	**17**	**5**	**5**	**7**
OP (%) [a]	32.4	36.1	27.9	10.0	10.0	35.7
TP (%) [a]	60.4	62.8	55.9	54.8	90.0	57.1
FN (%) [a]	7.2	1.1	16.2	35.2	0.0	7.1

No Cat.	CO > 0 **			CO > 0			CO = 0 **			CO = 0		
	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2
N	**7**	**5**	**15**	**8**	**4**	**15**	**1**	**1**	**2**	**38**	**36**	**91**
FP (%) [a]	71.4	80.0	20.8	62.5	33.3	33.3	0	0	0	16.4	11.1	11.0
TN (%) [a]	28.6	20.0	79.2	37.5	66.7	66.7	100	100	100	83.6	88.9	89.0

[a] The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1).
** Indicates at least one animal with a mean score of days 1–3 above the classification cut-off for at least one endpoint.

7.3. Analyses of the performance for specific Organic Functional Groups (OFG) with DALs

80. This section focuses on the performance of the DAL-1 and DAL-2 by organic functional group. The results of the DALs are presented in tables summarising TP (true Cat. 1 and true Cat. 2), TN (true No Cat.), over-predictions (OP, *in vivo* Cat. 2 predicted as Cat. 1), under-predictions (UP, *in vivo* Cat. 1 predicted as Cat. 2), FN (*in vivo* Cat. 1 or Cat. 2 predicted as No Cat.) and accuracy as compared to the Draize eye test benchmark data.

81. Only performance metrics for the most frequent OFG's, being at least 5 chemicals per allocated OFG are discussed. Therefore, the whole set of 175 liquids that were evaluated with DAL-1 and/or DAL-2 was considered. The distribution according to the UN GHS category is shown in Table 7.7.

Table 7.7. Number of liquids with a specific OFG according to the UN GHS category

OFG	% of total N (=175)	n	UN GHS (n)		
			Cat. 1	Cat. 2	No Cat.
Alcohol	21.1	37	7	11	19
Ether	14.9	26	1	3	22
Carboxylic acid ester	12.6	22	2	4	16
Alkyl halide	11.4	20	1	1	18
Aryl	10.3	18	1	1	16
Alkoxy	9.1	16	2	3	11
Alkane, branched with tertiary or quaternary carbon	9.1	16		2	14
Isopropyl	7.4	13		3	10
Dihydroxyl group	6.3	11			11
Aliphatic Amine	5.7	10	5	1	4
Allyl	5.7	10		2	8
AlkoxySilane	5.7	10	2		8
Ketone	5.1	9		5	4
Carboxylic acid	4.6	8	2	1	5
Acetoxy	4.0	7	1	1	5
Aldehyde	3.4	6		2	4
Aryl halide	3.4	6		1	5
Cycloalkane	3.4	6	1	1	4
Thioalcohol	3.4	6	1		5
Methacrylate	2.9	5			5
Saturated heterocyclic fragment	2.9	5	1		4

Note that a single chemical may have more than one organic functional group.

82. The performance of the different DALs by OFG is shown in Table 7.8. Only OFGs for which at least 5 liquids were evaluated for a specific UN GHS category are discussed. Liquids with an alcohol or aliphatic amine are the only *in vivo* Cat. 1 chemicals with at least 5 liquids. The TP rate for Cat. 1 liquids with an alcohol function (N=7) was similar for the different DALs

and ranged from 54.3-57.1%. The TP rate for Cat. 1 liquids with an aliphatic amine function (N=5) was 100% for the different DALs. Liquids with an alcohol or ketone function are the only *in vivo* Cat. 2 chemicals with at least 5 liquids being tested with the DAL-1 and DAL-2. The TP rate for Cat. 2 liquids with an alcohol function (N=8-10) was similar for the different DALs and ranged from 60.0-68.5%, misprediction are mostly over-predictions. The TP rate for Cat. 2 liquids with ketone function (N=5) ranged from 33.6-60.0%, 40.0% (2/5) was over-predicted. Regarding UN GHS No Cat. chemicals, for VRM1 (DAL-1), a FP rate ≥ 40% was observed for liquids with an alcohol (N=6), ether (N=7), or ketone (N=5) function (66.7%, 54.0%, and 40.0%, respectively). The FP rate for VRM2 (DAL-1) was 30.6% for alcohols (N=8) and 34.4% for ethers. No liquids with a ketone function were tested with VRM2. The FP rate for STE (DAL-2) was generally < 20%, except for the following OFGs with a FP ≥ 25%: ketone (52.1%, N=6), carboxylic acid (40%, N=5), aryl halide (40%, N=5) and allyl (25%, N=8).

83. In the absence of any indication that a particular substance is out of domain, test guidelines, and the DAs which rely on them, should be assumed to be broadly applicable. While the number of substances per OFG with results for DAL-1 or DAL-2 is somewhat limited, this analysis of predictive performance does not indicate any particular issue relating to specific OFGs. As detailed in Section 5.1.6, substances known to clearly fall outside the applicability domain of the individual information source methods, as defined within the respective TGs, should be excluded..

Table 7.8. Predictive performance considering the three UN GHS categories (Cat. 1, Cat. 2, No Cat.) of DALs with BCOP LLBO

UN GHS	Predicted class	Alcohol			Ether			Carboxylic acid ester		
		DAL-1	DAL-1	DAL-2	DAL-1	DAL-1	DAL-2	DAL-1	DAL-1	DAL-2
		RhCE: VRM1	RhCE: VRM2		RhCE: VRM1	RhCE: VRM2		RhCE: VRM1	RhCE: VRM2	
		N = 21	N = 24	N = 33	N = 11	N = 14	N = 24	N = 16	N = 13	N = 18
Cat. 1	Cat. 1	4.0/7	4.0/7	3.8/7	1.0/1	1.0/1	1.0/1	2.0/2	2.0/2	1.0/1
	Cat. 2	3.0/7	3.0/7	3.0/7	0.0/1	0.0/1	0.0/1	0.0/2	0.0/2	0.0/1
	No Cat.	0.0/7	0.0/7	0.2/7	0.0/1	0.0/1	0.0/1	0.0/2	0.0/2	0.0/1
Cat. 2	Cat. 1	2.5/8	3.5/9	3.3/10	1.0/3	1.0/3	0.75/3	1.0/4	1.0/4	1.0/4
	Cat. 2	5.5/8	5.4/9	6.0/10	1.0/3	2.0/3	0.75/3	2.0/4	2.9/4	2.5/4
	No Cat.	0.0/8	0.1/9	0.8/10	1.0/3	0.0/3	1.5/3	1.0/4	0.1/4	0.5/4
No Cat.	Cat. 1/Cat. 2	5.0/6	3.4/8	2.0/16	3.8/7	3.4/10	2.0/20	2.0/10	2.0/7	1.0/13
	No Cat.	10/6	4.6/8	14.0/16	3.2/7	6.6/10	18.0/20	8/10	5.0/7	12.0/13

UN GHS	Predicted class	Alkyl halides			Aryl			Alkoxy		
		DAL-1	DAL-1	DAL-2	DAL-1	DAL-1	DAL-2	DAL-1	DAL-1	DAL-2
		RhCE: VRM1	RhCE: VRM2		RhCE: VRM1	RhCE: VRM2		RhCE: VRM1	RhCE: VRM2	
		N = 7	N = 10	N = 19	N = 9	N = 7	N = 16	N = 10	N = 10	N = 16
Cat. 1	Cat. 1	0.0/1	0.0/1	0.0/1	1.0/1	1.0/1	1.0/1	2.0/2	2.0/2	2.0/2

	Cat. 2	1.0/1	1.0/1	1.0/1	0.0/1	0.0/1	0.0/1	0.0/2	0.0/2	0.0/2
	No Cat.	0.0/1	0.0/1	0.0/1	0.0/1	0.0/1	0.0/1	0.0/2	0.0/2	0.0/2
Cat. 2	Cat. 1	0.5/1	0.5/1	NA	0.0/1	0.0/1	0.0/1	1.0/3	1.0/3	0.8/3
	Cat. 2	0.5/1	0.5/1	NA	1.0/1	1.0/1	1.0/1	1.0/3	2.0/3	0.7/3
	No Cat.	0/1	0/1	NA	0.0/1	0.0/1	0.0/1	1.0/3	0.0/3	1.5/3
No Cat.	Cat. 1/Cat. 2	1.0/5	1.2/8	3.0/18	2.4/7	1.3/5	2.0/14	1.8/5	1.4/5	0.0/11
	No Cat.	4.0/5	6.8/8	15.0/18	4.6/7	3.7/5	12.0/14	3.2/5	3.6/5	11.0/11

UN GHS	Predicted class	Alkane, branched with tertiary or quaternary carbon			Isopropyl			Dihydroxyl group		
		DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2
		N = 8	N = 6	N = 13	N =6	N = 11	N = 11	N = 5	N = 5	N = 11
Cat. 1	Cat. 1	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Cat. 2	NA	NA	NA	NA	NA	NA	NA	NA	NA
	No Cat.	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cat. 2	Cat. 1	0.0/2	0.0/2	0.0/2	0.5/3	0.5/3	0.5/3	NA	NA	NA
	Cat. 2	2.0/2	1.9/2	1.7/2	2.5/3	2.4/3	2.5/3	NA	NA	NA
	No Cat.	0.0/2	0.1/2	0.3/2	0.0/3	0.7/3	0.0/3	NA	NA	NA
No Cat.	Cat. 1/Cat. 2	1.0/6	1.0/4	1.0/11	0.0/3	0.1/8	0.0/8	0.0/5	0.0/5	0.0/11
	No Cat.	5.0/6	3.0/4	10.0/11	3.0/3	7.9/8	8.0/8	5.0/5	5.0/5	11.0/11

UN GHS	Predicted class	Aliphatic amine			Allyl			AlkoxySilane		
		DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2
		N = 6	N = 5	N = 10	N = 4	N = 4	N = 9	N = 3	N = 4	N = 10
Cat. 1	Cat. 1	5.0/5	5.0/5	5.0/5	NA	NA	NA	2.0/2	2.0/2	2.0/2
	Cat. 2	0.0/5	0.0/5	0.0/5	NA	NA	NA	0.0/2	0.0/2	0.0/2
	No Cat.	0.0/5	0.0/5	0.0/5	NA	NA	NA	0.0/2	0.0/2	0.0/2
Cat. 2	Cat. 1	NA	NA	0.0/1	0.0/1	1.0/2	0.0/1	NA	NA	NA
	Cat. 2	NA	NA	1.0/1	0.0/1	1.0/2	1.0/1	NA	NA	NA
	No Cat.	NA	NA	0.0/1	1.0/1	0.0/2	0.0/1	NA	NA	NA
No Cat.	Cat. 1/Cat. 2	1.0/1	NA	0.0/4	0.0/3	0.0/2	2.0/8	1.0/1	0.0/2	1.0/8
	No Cat.	0.0/1	NA	4.0/4	3.0/3	2.0/2	6.0/8	0.0/1	2.0/2	7.0/8

UN GHS	Predicted class	Ketone			Carboxylic acid			Acetoxy		
		DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2	DAL-1 RhCE: VRM1	DAL-1 RhCE: VRM2	DAL-2
		N = 10	N = 5	N = 11	N = 1	N = 1	N = 8	N = 4	N = 4	N = 7
Cat. 1	Cat. 1	NA	NA	NA	1.0/1	1.0/1	2.0/2	NA	NA	NA
	Cat. 2	NA	NA	NA	0.0/1	0.0/1	0.0/2	NA	NA	NA
	No Cat.	NA	NA	NA	0.0/1	0.0/1	0.0/2	NA	NA	NA
Cat. 2	Cat. 1	2.0/5	2.0/5	2.0/5	NA	NA	1.0/1	1.0/1	1.0/1	1.0/1
	Cat. 2	1.7/5	2.9/5	3.0/5	NA	NA	0.0/1	0.0/1	0.0/1	0.0/1
	No Cat.	1.3/5	0.1/5	0.0/5	NA	NA	0.0/1	0.0/1	0.0/1	0.0/1
No Cat.	Cat. 1/Cat. 2	2.0/5	NA	3.1/6	NA	NA	2.0/5	1.0/3	1.0/3	0.0/5

	No Cat.	3.0/5	NA	2.9/6	NA	NA	3.0/5	2.0/3	2.0/3	5.0/5

UN GHS	Predicted class	Aldehyde			Aryl halide			Cycloalkane		
		DAL-1	DAL-1	DAL-2	DAL-1	DAL-1	DAL-2	DAL-1	DAL-1	DAL-2
		RhCE: VRM1	RhCE: VRM2		RhCE: VRM1	RhCE: VRM2		RhCE: VRM1	RhCE: VRM2	
		N = 4	N = 4	N = 6	N = 2	N = 1	N = 6	N = 4	N = 3	N = 6
Cat. 1	Cat. 1	NA	NA	NA	NA	NA	NA	0.0/1	0.0/1	0.0/1
	Cat. 2	NA	NA	NA	NA	NA	NA	1.0/1	1.0/1	1.0/1
	No Cat.	NA	NA	NA	NA	NA	NA	0.0/1	0.0/1	0.0/1
Cat. 2	Cat. 1	0.0/3	0.0/3	0.0/3	0.0/1	0.0/1	0.0/1	1.0/1	1.0/1	1.0/1
	Cat. 2	3.0/3	3.0/3	3.0/3	1.0/1	1.0/1	0.0/1	0.0/1	0.0/1	0.0/1
	No Cat.	0.0/3	0.0/3	0.0/3	0.0/1	0.0/1	1.0/1	0.0/1	0.0/1	0.0/1
No Cat.	Cat. 1/Cat. 2	0.4/1	0.3/1	2.0/3	1.0/1	NA	2.0/5	1.0/2	0.0/1	1.0/4
	No Cat.	0.6/1	0.7/1	1.0/3	0.0/1	NA	3.0/5	1.0/2	1.0/1	3.0/4

UN GHS	Predicted class	Methacrylate			Saturated heterocyclic fragment			Thioalcohol		
		DAL-1	DAL-1	DAL-2	DAL-1	DAL-1	DAL-2	DAL-1	DAL-1	DAL-2
		RhCE: VRM1	RhCE: VRM2		RhCE: VRM1	RhCE: VRM2		RhCE: VRM1	RhCE: VRM2	
		N = 2	N = 3	N = 5	N = 4	N = 2	N = 5	N = 3	N = 2	N = 4
Cat. 1	Cat. 1	NA	NA	NA	1.0/1	1.0/1	1.0/1	1.0/1	1.0/1	NA
	Cat. 2	NA	NA	NA	0.0/1	0.0/1	0.0/1	0.0/1	0.0/1	NA
	No Cat.	NA	NA	NA	0.0/1	0.0/1	0.0/1	0.0/1	0.0/1	NA
Cat. 2	Cat. 1	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Cat. 2	NA	NA	NA	NA	NA	NA	NA	NA	NA
	No Cat.	NA	NA	NA	NA	NA	NA	NA	NA	NA
No Cat.	Cat. 1/Cat. 2	1.8/2	2.0/3	1.0/5	1.0/3	1.0/1	3.0/4	2.0/3	1.0/1	0.0/4
	No Cat.	0.2/2	1.0/3	4.0/5	2.0/3	0.0/1	1.0/4	1.0/3	0.0/1	4.0/4

8. Annex A

8.1. Variations of the DIP and the effect on the performance of DAL-1

84. In case physicochemical properties are not available for the test chemical, the combination of the *in vitro* test methods is still applicable (immediately start from Step 2, Figure 2.1) as described in the IATA Guidance Document No. 263. The effect on predictive performance of eliminating the physicochemical properties when considering the three UN GHS categories (Cat. 1, Cat. 2, No Cat.) of the *in vitro* testing approach is provided in Table 8.1 (VRM1 and BCOP LLBO) and Table 8.2 (VRM2 and BCOP LLBO). Omitting the physicochemical exclusion rules from DAL-1 resulted in a decrease of the specificity from 70.5% (Table 5.6) to 57.6 (with VRM1, Table 8.1) and from 79.7 (Table 5.7) to 63.3% (with VRM2, Table 8.2), as such the performance criteria provided in Table 4.1 were not met. Results for the BCOP LLBO test method do not exist for 40/55 (VRM1) and 34/46 (VRM2) No Cat. liquids respectively. As mentioned before, it is unlikely that a large number of VRM FPs will be predicted Cat. 1 by the BCOP LLBO (§49).

Table 8.1. Performance of the testing strategy with the in vitro test methods: VRM1 and BCOP LLBO (N = 94 liquids)

UN GHS	VRM1 with BCOP LLBO		
	Cat 1	Cat 2	No Cat
Cat. 1 (N=17), % [a] (n/N)	**76.5% (13.0/17.0)**	23.5% (4.0/17.0)	0.0% (0.0/17.0)
Cat. 2 (N=22), % [a] (n/N)	27.3% (6.0/22.0)	**59.1% (13.0/22.0)**	13.6% (3.0/22.0)
No Cat. (N=55), % [a] (n/N)	42.4% [b] (23.3/55.0)		**57.6% (31.7/55.0)**
64.4% balance accuracy			

[a] The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1). To improve the readability of the numbers in the table, the numbers n/N have been rounded, so they may deviate slightly from the percentage corresponding to the weighted calculation.
[b] BCOP LLBO data are not available for the majority (40/55) of the in vivo No Cat. liquids, as such, for liquids that were not identified as No Cat. with VMR1, it was not possible to distinguish between Cat. 1 (based on BCOP LLBO) and Cat. 2. Therefore, the false positives are presented between Cat. 1 and Cat. 2

Class-specific performance metrics

	Cat. 1	**Cat. 2**	**No Cat.**
Correct within-class predictions (%)	76.5	59.1	57.6
Correct outside-class predictions (%)	92.2	73.1	92.3
Balanced Accuracy (%)	84.4	66.1	75.0

Note: The balanced accuracy is the average of the correct within-class predictions and the correct outside-class predictions. The formulas for the class-specific performance metrics are provided in Annex C.

Table 8.2. Performance of the testing strategy with the in vitro test methods: VRM2 and BCOP LLBO (N = 86 liquids)

UN GHS	VRM2 with BCOP LLBO		
	Cat 1	Cat 2	No Cat
Cat. 1 (N=17), % [a] (n/N)	**76.5%** **(13.0/17.0)**	23.5% (4.0/17.0)	0.0% (0.0/17.0)
Cat. 2 (N=23), % [a] (n/N)	30.4% (7.0/23.0)	**68.7%** **(15.8/23.0)**	0.9% (0.2/23.0)
No Cat. (N=46), % [a] (n/N)	36.7% [b] (16.9/46.0)		**63.3%** **(29.1/46.0)**
69.5% balance accuracy			

[a] The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1). To improve the readability of the numbers in the table, the numbers n/N have been rounded, so they may deviate slightly from the percentage corresponding to the weighted calculation.
[b] BCOP LLBO data are not available for the majority (34/46) of the in vivo No Cat. liquids, as such, for liquids that were not identified as No Cat. with VRM2, it was not possible to distinguish between Cat. 1 (based on BCOP LLBO) and Cat. 2. Therefore, the false positives are presented between Cat. 1 and Cat. 2.

Class-specific performance metrics

	Cat. 1	Cat. 2	No Cat.
Correct within-class predictions (%)	76.5	68.7	63.3
Correct outside-class predictions (%)	89.9	66.8	99.5
Balanced Accuracy (%)	83.2	67.8	81.4

Note: The balanced accuracy is the average of the correct within-class predictions and the correct outside-class predictions. The formulas for the class-specific performance metrics are provided in Annex C.

85. Another variation of the DIP looked at was replacing the BCOP LLBO (opacity only) with the BCOP OP-KIT (IVIS) to identify Cat. 1. The performance of DAL-1 is reported for the same set of substances, which are already the basis for the analyses shown in §49, but for which the Cat. 1 prediction was based on the standard BCOP test method that uses the OP-KIT to measure opacity instead of the BCOP LLBO. Both the BCOP OP-KIT and the BCOP LLBO had no Cat.1 substances predicted as No Cat. The main difference is in terms of Cat. 1 correct predictions for the same set of liquids, 60.8% (n=17, Table 8.3 and Table 8.4) and 76.5% (n=17, Table 5.7 and Table 5.6), that were correctly identified with the BCOP OP-KIT and BCOP LLBO test methods, respectively. Therefore the performance criteria for at least 75% correct Cat. 1 identification were not met for the variations of the DIP. For the same set of reference chemicals, the percentage of over-predicted Cat. 2 liquids was very similar when performing the measurement and prediction determination with the BCOP OP-KIT and

the BCOP LLBO methods. Based on this outcome, it was concluded that the BCOP LLBO was the preferred method to identify Cat. 1.

86. None of the variations of the DIP for DAL-1 met the performance criteria provided in section 4.3 and as such the variations do not fall under the GL.

Table 8.3. Performance of the DAL-1 based on physicochemical properties, VRM1 and BCOP OP-KIT (N = 94 liquids)

UN GHS	DAL-1		
	Cat 1	Cat 2	No Cat
Cat. 1 (N=17), % [a] (n/N)	**60.8% (10.3/17.0)**	39.2% (6.7/17.0)	0.0% (0.0/17.0)
Cat. 2 (N=22), % [a] (n/N)	25.5% (5.6/22.0)	**60.9% (13.4/22.0)**	13.6% (3.0/22.0)
No Cat. DA (N=55), % [a] (n/N)	29.5% [b] (16.2/55.0)		**70.5% (38.8/55.0)**
64.1% balance accuracy			

[a] The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1). To improve the readability of the numbers in the table, the numbers n/N have been rounded, so they may deviate slightly from the percentage corresponding to the weighted calculation.
[b] BCOP OP-KIT data: the same set of in vivo No Cat. liquids was used as for the BCOP LLBO and for the BCOP LLBO data are not available for the majority (40/55) of the in vivo no Cat. liquids, as such, for liquids that were not identified as No Cat. with VRM1, it was not possible to distinguish between Cat. 1 (based on BCOP OP-KIT) and Cat. 2. Therefore, the false positives are presented between Cat. 1 and Cat. 2.

Class-specific performance metrics

	Cat. 1	**Cat. 2**	**No Cat.**
Correct within-class predictions (%)	60.8	60.6	70.5
Correct outside-class predictions (%)	92.6	69.5	92.1
Balanced Accuracy (%)	76.7	65.0	81.4

Note: The balanced accuracy is the average of the correct within-class predictions and the correct outside-class predictions. The formulas for the class-specific performance metrics are provided in Annex C.

Table 8.4. Performance of the DAL-1 based on physicochemical properties, VRM2 and BCOP OP-KIT (N = 86 liquids)

UN GHS	DAL-1		
	Cat 1	Cat 2	No Cat
Cat. 1 (N=17), % [a] (n/N)	**60.8% (10.3/17.0)**	39.2% (6.7/17.0)	0.0% (0.0/17.0)
Cat. 2 (N=23), % [a] (n/N)	28.7% (6.6/23.0)	**70.4% (16.2/23.0)**	0.9% (0.2/23.0)
No Cat. DA (N=46), % [a] (n/N)	20.3% [b] (9.3/46.0)		**79.7% (36.7/46.0)**

70.3 % balance accuracy

[a] The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1). To improve the readability of the numbers in the table, the numbers n/N have been rounded, so they may deviate slightly from the percentage corresponding to the weighted calculation.
[b] BCOP OP-KIT data: the same set of in vivo No Cat. liquids was used as for the BCOP LLBO and for the BCOP LLBO data are not available for the majority (34/46) of the in vivo no Cat. liquids, as such, for liquids that were not identified as No Cat. with VRM2, it was not possible to distinguish between Cat. 1 (based on BCOP OP-KIT) and Cat. 2. Therefore, the false positives are presented between Cat. 1 and Cat. 2.

Class-specific performance metrics

	Cat. 1	Cat. 2	No Cat.
Correct within-class predictions (%)	60.8	70.4	79.7
Correct outside-class predictions (%)	90.4	75.1	99.5
Balanced Accuracy (%)	75.6	72.8	89.6

Note: The balanced accuracy is the average of the correct within-class predictions and the correct outside-class predictions. The formulas for the class-specific performance metrics are provided in Annex C.

8.2. Variations of the DIP and the effect on the performance of DAL-2

87. In addition, the predictive performance of DAL-2 is reported for the same set of substances, which are already the basis for the analyses shown in §71, but for which the Cat. 1 prediction was based on the standard BCOP test method that uses the OP-KIT to measure opacity instead of the BCOP LLBO. Both the BCOP OP-KIT and the BCOP LLBO had 1.2% Cat.1 substances predicted as No Cat. The main difference is in terms of Cat. 1 correct predictions for the same set of liquids, 65.5% (n=17, Table 8-5) and 81.2% (n=17, Table 5-9), that were correctly identified with the BCOP OP-KIT and BCOP LLBO test methods, respectively. Therefore the performance criteria for at least 75% correct Cat. 1 identification were not met for the variation of the DIP and as such this variation does not fall under the GL. For the same set of reference chemicals, the percentage of over-predicted Cat. 2 liquids was very similar when performing the measurement and prediction determination with the BCOP OP-KIT and the BCOP LLBO methods. Based on this outcome, it was concluded that the BCOP LLBO was the preferred method to identify Cat. 1. The effect of including physicochemical property (PCP) exclusion rules on the proportion of correct No Cat. identifications was also investigated. For 3 out of 123 No Cat. liquids that were positive with the STE test method (viability at 5% and 0.05%, < 70% and > 70%, respectively), the PCP exclusion rules could be applied resulting in a No Cat. prediction. This resulted in a slight increase of correct No Cat. predictions from 85.3% to 87.8%. Although the PCP exclusion rules are not a prerequisite for DAL-2, the amount of testing can be reduced in case of a No Cat. Prediction based on the PCP exclusion rules.

Table 8.5. Performance of the DAL-2 based on STE and BCOP OP-KIT (N = 164 liquids)

UN GHS	DAL-2		
	Cat 1	Cat 2	No Cat

Cat. 1 (N=17), % [a] (n/N)	**65.5%** **(11.1/17.0)**	33.3% (5.7/17.0)	1.2% (0.2/17.0)
Cat. 2 (N=24), % [a] (n/N)	29.6% (7.1/24.0)	**56.9%** **(13.6/24.0)**	13.5% (3.3/24.0)
No Cat. DA (N=123), % [a] (n/N)	14.7% [b] (18.1/123.0)		**85.3%** **(104.9/123.0)**
69.2 % balance accuracy			

[a] The proportion in the tables are based on weighted calculation. For each chemical, all results were taken into account and a correction factor was applied so that all chemicals had the same weight (weight of 1). To improve the readability of the numbers in the table, the numbers n/N have been rounded, so they may deviate slightly from the percentage corresponding to the weighted calculation.
[b] BCOP OP-KIT data: the same set of in vivo No Cat. liquids was used as for the BCOP LLBO and for the BCOP LLBO data are not available for the majority (108/123) of the in vivo no Cat. liquids, as such, for liquids that were not identified as No Cat. with STE, it was not possible to distinguish between Cat. 1 (based on BCOP OP-KIT) and Cat. 2. Therefore, the false positives are presented between Cat. 1 and Cat. 2.

Class-specific performance metrics

	Cat. 1	**Cat. 2**	**No Cat.**
Correct within-class predictions (%)	65.5	56.9	85.3
Correct outside-class predictions (%)	95.2	83.0	91.6
Balanced Accuracy (%)	80.3	70.0	88.5

Note: The balanced accuracy is the average of the correct within-class predictions and the correct outside-class predictions. The formulas for the class-specific performance metrics are provided in Annex C.

Annex B: Spreadsheets

[See separate file ENV/CBC/MONO(2022)10]

Annex B.1

This Annex includes the detailed Draize eye test data that were used for DAL-1 and DAL-2.

Annex B.2

Spreadsheet with the identification of the substances that were used in the training set and the test set of DAL-1 and DAL-2.

Annex B.3 (DAL-1 VRM1**,** DAL-1 VRM2**)**

This Annex includes the predictions of the individual test methods and DAs for DAL-1 with VRM1 and DAL-1 with VRM2.

Annex B.4

This Annex includes the predictions of the individual test methods and DA for DAL-2.

Annex C: Contingency matrix

The performance metrics for a 3-class classification are illustrated for UN GHS eye irritation/corrosion. Therefore, 3x3 contingency matrix or contingency table (No Cat., Cat. 2, and Cat. 1) is constructed and the number of correct and incorrect predictions in each category are evaluated. The calculations are illustrated with a fictive example (worked example).

		Actual categories		
		Cat. 1	**Cat. 2**	**No Cat.**
Predicted categories	**Cat. 1**	A	D	G
	Cat. 2	B	E	H
	No Cat.	C	F	I
	Total	**A+B+C**	**D+E+F**	**G+H+I**

Balanced accuracy: $\left(\frac{A}{A+B+C}+\frac{E}{D+E+F}+\frac{I}{G+H+I}\right)/3$

Cells with a green background (diagonal) correspond with correct predictions, cells with a grey background on the off-diagonal correspond with incorrect predictions.

Performance metrics per UN GHS category

For the calculations, one-versus-all remaining categories are applied, this reduces the 3x3 matrix to three 2x2 matrices. Note that the terminology used for No Cat. is different from what is normally used in a binary problem where there are only negatives (e.g. non-irritant) and positives (e.g. irritant). In order to obtain the performance metrics (sensitivity, specificity, and balanced accuracy) for each reference class (UN GHS category), the 3x3 matrix is converted to three 2-class matrices. The calculations are provided in the overview table below.

Overview table

Class specific Statistic	Cat. 1 versus remaining categories		Cat. 2 versus remaining categories		No Cat. versus remaining categories	
	Categories	Formula	Categories	Formula	Categories	Formula
Correct within-class predictions	True Cat. 1	$\frac{A}{A+B+C}$	True Cat. 2	$\frac{E}{D+E+F}$	True No Cat.	$\frac{I}{G+H+I}$

Correct outside-class predictions	True Cat. 2 and No Cat.	$\frac{E+F+H+I}{D+E+F+G+H+I}$	True Cat. 1 and No Cat.	$\frac{A+C+G+I}{A+B+C+G+H+I}$	True Cat. 1 and Cat. 2	$\frac{A+B+D+E}{A+B+C+D+E+F}$
Balanced accuracy		X = (Correct within + ouside-class predictions)/2		Y = (Correct within + ouside-class predictions)/2		Z = (Correct within + ouside-class predictions)/2

Worked example

The table below is a fictive example that shows the agreement in predictions between the reference test (Draize eye test) and a define approach (DA) with total number of reference chemicals = 96.

	Reference (Draize eye test)		
Predicted (based on DA)	Cat. 1	Cat. 2	No Cat.
Cat. 1	21	5	1
Cat. 2	7	18	11
No Cat.	0	3	30
Total Reference	28	26	42

Overall balanced accuracy: 71.9%

Performance UN GHS Cat. 1

For the performance of the **UN GHS Cat. 1**, a 2x2 matrix is constructed and the number of true positives (the *in vivo* Cat. 1 chemicals that are correctly predicted as Cat. 1), the number of true negatives (the number of *in vivo* Cat. 2 and No Cat. chemicals that are correctly predicted as Cat. 2 and No Cat.), the number of false negatives and false positives are calculated. Based on this 2x2 matrix, sensitivity, specificity, and balanced accuracy are calculated.

		Reference (Draize eye test)	
		Cat. 1	Remaining categories: Cat. 2 + No Cat.
Predicted (based on DA)	Cat. 1	21	5+1 = 6
	Remaining categories: Cat. 2 + No Cat.	7+0 = 7	18+3+11+30 = 62
	Total Reference	28	68

Performance UN GHS Cat. 2

For the performance of the **UN GHS Cat. 2**, a 2x2 matrix is constructed and the number of true positives (the *in vivo* Cat. 2 chemicals that are correctly predicted as Cat. 2), the number of true negatives (the number of *in vivo* Cat. 1 and No Cat. chemicals that are correctly predicted as Cat. 1 and No Cat.), the number of false negatives and false positives are calculated. Based on this 2x2 matrix, sensitivity, specificity, and balanced accuracy are calculated.

		Reference (Draize eye test)	
		Cat. 2	Remaining categories: Cat. 1 + No Cat.
Predicted (based on DA)	Cat. 2	18	7+11 =18
	Remaining categories: Cat. 1 + No Cat.	5+3 = 8	21+0+1+30 = 52
	Total Reference	26	70

Performance UN GHS No Cat.

For the performance of the **UN GHS No Cat.**, a 2x2 matrix is constructed and the number of true positives (the *in vivo* No Cat. chemicals that are correctly predicted as No Cat.), the number of true negatives (the number of *in vivo* Cat. 1 and Cat. 2 chemicals that are correctly predicted as Cat. 1 and Cat. 2), the number of false negatives and false positives are calculated. Based on this 2x2 matrix, sensitivity, specificity, and balanced accuracy are calculated.

		Reference (Draize eye test)	
		No Cat.	Remaining categories: Cat. 1 + Cat. 2
Predicted (based on DA)	No Cat.	30	0+3 = 3
	Remaining categories: Cat. 1 + Cat. 2	1+11 = 12	21+7+5+18 = 51
	Total Reference	**42**	**54**

Performance metrics per UN GHS category

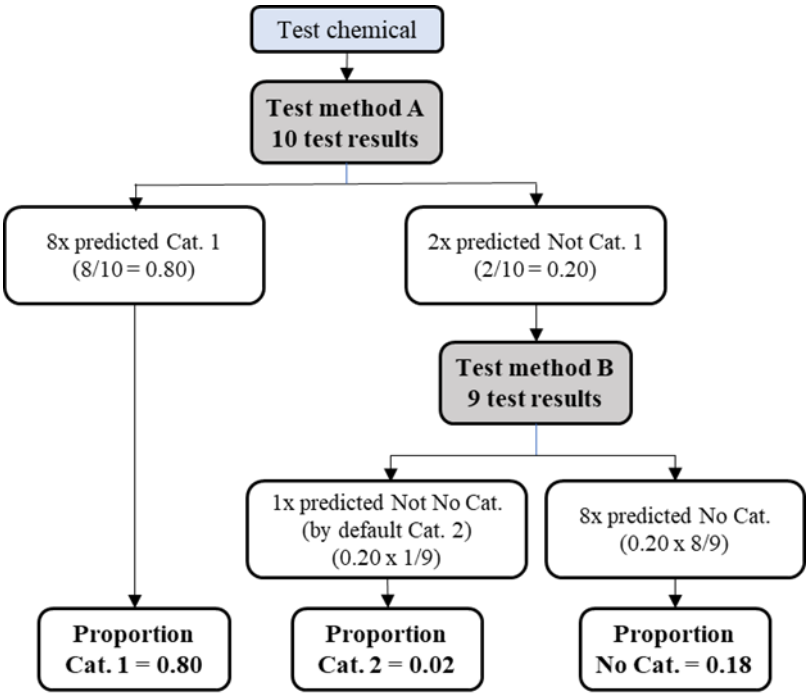
Statistic	Cat. 1	Cat. 2	No Cat.
Correct within-class predictions	21/28*100 = 75.0	18/26*100 = 69.2	30/42*100 = 71.4
Correct outside-class predictions	62/68*100 = 91.2	52/70*100 = 74.3	51/54*100 = 94.4
Balanced accuracy	(75.0+91.2)/2 = 83.1	(69.2+74.3)/2 = 71.8	(71.4+94.4)/2 = 82.9

The **average balanced accuracy**, is defined as the arithmetic mean of the class-specific balanced accuracies = (83.1+71.8+82.9)/3 = 79.3%

The confusionMatrix function of the R package “caret”, uses the same terminology (calculations) as provided in the table above for the class specific sensitivity, specificity, and balanced accuracy. https://cran.r-project.org/web/packages/caret/caret.pdf

Annex D: Weighted calculation

Note that for a single chemical multiple results were available for the different *in vitro* test methods (sometimes resulting in different predictions) and therefore a weighted calculation approach was used so that each chemical has the same weight of 1 (sum of all fractions = 1, this is illustrated in the Figure).



Annex E: Physicochemical properties

An overview of the OECD GLs for the testing of the physicochemical properties is provided in the table below. The measurement of physicochemical properties should be performed according to the OECD GLs and test reports are required corresponding with the information requested on data and reporting in each specific OECD GL. Regarding the prediction of physicochemical properties models that are based on the 5 OECD principles for QSAR models (1) and that have a QMRF (QSAR Model Reporting Format) should be used. The OPERA models are one of the QSAR tools that were developed based on those principles and a QMRF (QSAR Model Reporting Format) is available for LogP, Vapor pressure and Water solubility. The OPERA predictions are available at the NTP Integrated Chemical Environment https://ice.ntp.niehs.nih.gov/ and the EPA Comptox Dashboard https://comptox.epa.gov/dashboard. For local use, the OPERA application can also be downloaded from the NIEHS GitHub repository https://github.com/NIEHS/OPERA. Command-line and graphical user interface versions are available for Windows and Linux operating systems.

For Surface tension the toxicity-estimation-software-tool-test (T.E.S.T.) can be used and is available at https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test. The predictions for Surface tension are also available at the EPA Comptox Dashboard https://comptox.epa.gov/dashboard. A calculation report and a QMRF (Annex F) is available.

METHOD	LOGP (OCTANOL-WATER)	VAPOR PRESSURE	SURFACE TENSION	WATER SOLUBILITY
MEASURE OECD GL	GL 107 – Shake flask GL 117 – HPLC Method GL 122 – Slow-stirring	GL 104	GL 115**	GL 105
PREDICT (Q)SAR	OPERA	OPERA	T.E.S.T. (EPA)	OPERA

* The temperature of the test should be run at the temperature that is recommended in the corresponding OECD GL. For example, for water solubility and surface tension, the test should be run preferably at 20 ± 0.5° C. Note that the corresponding QSAR's predict the values for 25° C.

** Measurements should be performed on pure liquids only.

Performance statistics of the OPERA models

Performance of the selected models in fitting, cross-validation (CV) , and on the test sets (2)

Property	5-fold CV (75%)		Training (75%)			Test (25%)		
	Q^2	RMSE	Dataset	R^2	RMSE	Dataset	R^2	RMSEP

LogP	0.85	0.69	10,537	0.86	0.67	3513	0.86	0.78
VP	0.91	1.08	2034	0.91	1.08	679	0.92	1
WS	0.87	0.81	3158	0.87	0.82	1066	0.86	0.86

RMSE: root mean square error

RMSEP: root mean square error in prediction

R^2 : coefficient of determination
Q^2 : predictive squared correlation coefficient

OPERA models provide Applicability Domain (AD)-indices (global and local) that give an indication on the reliability of the prediction. Furthermore, a confidence level index ranging from 0 to 1 is provided, the higher this index, the more the prediction is likely to be reliable. This index gives the user an estimate regarding the reliability of the prediction when the query chemical is inside the AD.

QMRF reports OPERA models

Property	JRC report ID	DOI
LogP	Q17-16-0016	https://doi.org/10.13140/rg.2.2.12731.82723/1
VP	Q17-14-0013	https://doi.org/10.13140/rg.2.2.32864.48641/1
WS	Q17-13-0012	https://doi.org/10.13140/rg.2.2.16087.27041/1

Performance statistics for surface tension as predicted with the QSAR model from T.E.S.T.

The performance statistics of the QSAR model for surface tension are reported in the User's Guide for T.E.S.T (Toxicity Estimation Software Tool) Version 5.1, A Java Application to Estimate Toxicities and Physical Properties from Molecular Structure. © 2020 U.S. Environmental Protection Agency; https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

Experimental data set for surface tension: the surface tension at 25°C for 1416 chemicals was obtained from the data compilation of Jasper (3). The experimental values (at 25°C) are estimated using an empirical correlation, which is fit to experimental data from Jaspar.

Validation results for surface tension (SF in dyne/cm), the consensus method produces the best results in terms of prediction accuracy and coverage

R^2	RMSE	MAE	Coverage
0.889	2.245	1.414	0.926

R^2 : coefficient of determination

RMSE: root mean square error

MAE: mean absolute error

1. OECD (2014), Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment, No. 69, OECD Publishing, Paris, https://doi.org/10.1787/9789264085442-en.
2. Mansouri K, Grulke CM, Judson RS, Williams AJ. OPERA models for predicting physicochemical properties and environmental fate endpoints. J Chem inform. 2018 Mar 8;10(1):10. doi: 10.1186/s13321-018-0263-1. PMID: 29520515; PMCID: PMC5843579.
3. 3. Jasper, J.J., The Surface Tension of Pure Liquid Compounds. J. Phys. Chem. Ref. Data, 1972. 1: p. 841-1009.

Annex F: QMRF: T.E.S.T. Model for Surface Tension

[See separate file ENV/CBC/MONO(2022)10].