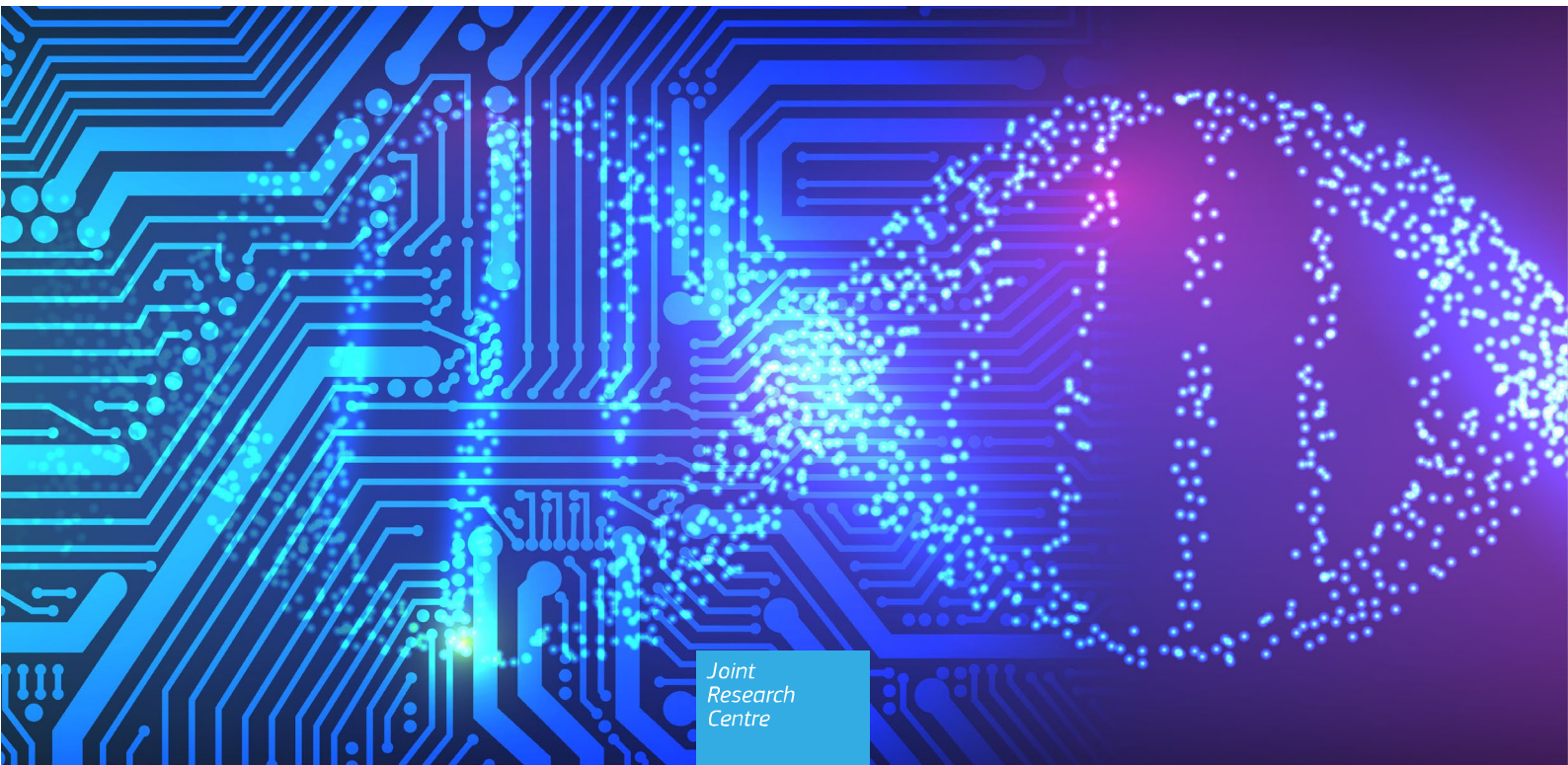
European Commission

ESAC Opinion

on the

Scientific Validity of the GARDskin and GARDpotency Test Methods

ESAC Opinion No. 2021-01
of 8 July 2021



The independent scientific peer review of the GARDskin and GARDpotency *in vitro* test methods described in this report was organised by the Joint Research Centre's EU Reference Laboratory for alternatives to animal testing (EURL ECVAM) and conducted by the EURL ECVAM Scientific Advisory Committee (ESAC).

The ESAC peer review was coordinated by João Barroso and Silvia Casati on behalf of JRC / EURL ECVAM.

Contact information
European Commission, Joint Research Centre (JRC), Chemical Safety and Alternative Methods Unit (F3)
Address: via E. Fermi 2749, I-21027 Ispra (VA), Italy
Email: JRC-F3-ENQUIRIES@ec.europa.eu

EU Science Hub
https://ec.europa.eu/jrc

JRC125963

PDF ISBN 978-92-76-40345-6 doi:10.2760/626728

Luxembourg: Publications Office of the European Union, 2021

How to cite this report: EURL ECVAM Scientific Advisory Committee. *ESAC Opinion on the Scientific Validity of the GARDskin and GARDpotency Test Methods*. Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-40345-6, doi:10.2760/626728, JRC125963. Available at: http://publications.jrc.ec.europa.eu/repository/handle/JRC125963.

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
JOINT RESEARCH CENTRE
Directorate F - Health, Consumers and Reference Materials
European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM)

EURL ECVAM
SCIENTIFIC
ADVISORY
COMMITTEE
(ESAC)

ESAC OPINION

on the

Scientific Validity of the GARDskin and GARDpotency Test Methods

ESAC Opinion No.	2021-01
Relevant ESAC Request No.	2020-01
Date of Opinion	08/07/2021

Table of contents

Abstract

ESAC, the EURL ECVAM Scientific Advisory Committee, advises EURL ECVAM on scientific issues. Its main role is to conduct independent peer review of validation studies of alternative test methods and to assess their scientific validity for a given purpose. The committee reviews the appropriateness of study design and management, the quality of results obtained and the plausibility of the conclusions drawn. ESAC peer reviews are formally initiated with a EURL ECVAM Request for ESAC Advice, which provides the necessary background for the peer-review and establishes its objectives, timelines and the questions to be addressed. The peer review is normally prepared by specialised ESAC Working Groups. ESAC's advice to EURL ECVAM is formally provided as 'ESAC Opinions' and 'Working Group Reports' at the end of the peer review. ESAC may also issue Opinions on other scientific issues of relevance to the work and mission of EURL ECVAM but not directly related to a specific alternative test method.

The ESAC Opinion expressed in this report relates to the peer-review of the GARDskin and GARDpotency *in vitro* test methods for skin sensitisation.

Ispra, 8 July 2021

ESAC Opinion

In April 2020, the EURL ECVAM Scientific Advisory Committee (ESAC) (Annex 1) was formally asked by EURL ECVAM to review the available evidence supporting the scientific validity of GARDskin and GARDpotency *in vitro* test methods. GARDskin and GARDpotency are separate assays with different intended purposes. GARDskin provides hazard identification of skin sensitisers (Cat. 1 vs not classified), while GARDpotency provides hazard characterisation by subcategorisation of skin sensitisers (Cat. 1A vs Cat. 1B), according to the United Nations Globally Harmonized System of Classification and Labelling of Chemicals (UN GHS) (UN, 2021). These two methods are also proposed to be used in combination, where a positive result in GARDskin is analysed with GARDpotency to determine the UN GHS subcategory.

An ESAC Working Group (WG) was established to assess the scientific validity of GARDskin and GARDpotency (Annex 1). The ESAC WG decided to proceed to the peer-review of the scientific evidence regarding:

1. GARDskin transferability, reproducibility and ability to support the discrimination between sensitisers (Cat. 1) and non-sensitisers (not classified), in line with the methods adopted within OECD Test Guidelines 442C, 442D and 442E (OECD, 2018a, 2018b, 2021).
2. GARDpotency transferability, reproducibility and ability to support the identification of Cat. 1A and Cat. 1B skin sensitisers.
3. The combined use of GARDskin and GARDpotency as a stand-alone strategy to identify Cat. 1A, Cat. 1B and not classified.

Based on its independent assessment, the ESAC WG delivered a detailed ESAC WG report (Annex 2) to support the development of this opinion. The analysis and conclusions of the ESAC WG were based primarily on the GARDskin and GARDpotency files submitted to EURL ECVAM, including all the relevant study Annexes and supporting documents. The assessment also included direct requests from the ESAC WG to the test developer for supporting information.

At its 47th meeting, held virtually on 8-9 March 2021, the members of the ESAC drafted this opinion. The final version of the opinion was unanimously endorsed by the ESAC by written procedure on 8th July 2021. Based on the available information, the existing scientific literature and the experts' own extensive experience as detailed in the ESAC WG report, the ESAC unanimously concluded the following:

Biological relevance for the evaluation of skin sensitisation

The biological system used in both GARDskin and GARDpotency is a cultured human myeloid dendritic-like cell line (SenzaCells – subcloned from MUTZ-3 cells). SenzaCells are used as a surrogate for dendritic cells, in a similar way to other cell lines used in OECD adopted non-animal methods for skin sensitisation (e.g., THP1 and U937). The gene signature that forms the basis for both test methods was identified by an automated selection process, based on transcriptome data generated with a set of skin sensitisers and non-sensitisers. The gene signature of GARDskin is composed of 196 genes when using nanoString data (200 genes were originally selected from Affymetrix data) and that of GARDpotency is composed of 51 genes when using nanoString data (52 genes were originally selected from Affymetrix data). Even though the selection process was empirically driven, a subsequent analysis of the final gene signature by the test developer has provided a mechanistic rationale for many of the selected genes and their respective relevance to skin sensitisation. Many of the identified transcription pathways, such as oxidative stress, immune responses, dendritic cell maturation and cytokine responses, are in line with mechanisms described under key events of the skin sensitisation Adverse Outcome Pathway (AOP) (OECD, 2012). The ESAC considers that the information provided by the test developer was sufficient to evaluate the biological relevance of GARDskin and GARDpotency. The ESAC concludes that the GARDskin and GARDpotency methods are biologically relevant for the evaluation of skin sensitisation.

Verification of the Support Vector Machine (SVM) algorithms

Using the code, explanations and data provided by the test developer, the ESAC was able to replicate the SVM algorithms, and verify and reproduce the steps from raw data to prediction for both GARDskin and GARDpotency, obtaining the same Decision Values for every main stimulation. The ESAC also evaluated the online tool for data analysis, the GARD Data Analysis and Application (GDAA) suit, and found it to be functional and user-friendly.

In addition, the ESAC used the training set to generate GARDskin SVM models using 1 up to 196 genes. The results showed that models trained with considerably fewer genes have similar performance to the GARDskin SVM, which uses 196 genes. Even though the current model is considered appropriate for its purpose, the ESAC considers it overly complex and that it could benefit from simplification. A simpler model would be cheaper to conduct and easier to assess and understand.

As far as the ESAC is aware, this is the first time a machine-learning algorithm has been independently reviewed for application in the field of regulatory toxicology. Two specific Appendixes detailing this work are available in the ESAC WG Report.

The ESAC concludes, therefore, that the evidence supporting the GARDskin and GARDpotency SVMs is sufficient and adequate.

Appropriateness of the study designs

The ESAC noted that a large proportion of the chemicals used as test set in the blind multilaboratory trial of the validation study were also part of the training set used to build the GARDskin and GARDpotency SVM prediction models. The ESAC considers that selecting a large proportion of the chemicals used to train a model (training set) for the validation study (test set) is inappropriate. When a machine-learning algorithm (e.g., SVM) is used, it is even more important that the principle of keeping training and test sets separate is strictly adhered to, due to the risk of overfitting.

In the case of GARDskin, some extra data were available to augment the confidence on the assessment of reproducibility and predictive capacity of the method, so that the study design was considered to be sufficient. Nonetheless, it would have been useful to have included a larger number of test chemicals in the multilaboratory study (especially non-sensitisers) which had not been used to train the model.

For GARDpotency, the overall amount of data available was smaller than for GARDskin. This was due primarily to the fact that only samples that were positive in GARDskin were subsequently analysed in GARDpotency. This study design led to an incomplete data matrix, in which chemicals had a varying number of repetitions in the validation study. Moreover, a majority of the sensitisers from the multilaboratory study analysed with GARDpotency were also part of the training set used to build the GARDpotency prediction model. Therefore, the ESAC considers that the limitations in the study design hindered the assessment of the validation status of the GARDpotency method.

Regarding the combined GARDskin + GARDpotency strategy, the ESAC considers that its assessment was also adversely affected by the above-mentioned limitations. Furthermore, no rationale was provided for the number of chemicals used to evaluate this combined strategy to predict three skin sensitisation categories (versus two predicted categories by GARDskin and by GARDpotency).

Reproducibility

GARDskin

The ESAC concludes that the reproducibility of GARDskin is appropriate. The within-laboratory reproducibility (WLR) was in the range of 78.6-89.2% and the between-laboratory reproducibility (BLR) was 82.1% (95%-Confidence Interval (CI): 63.1%-93.9%), when considering all chemicals, including those claimed to have technical issues by the test method developer. Both WLR and BLR are close or above the set target of at least 80% concordance.

GARDpotency

The ESAC considers that the reproducibility of GARDpotency is not sufficient at this time. The target was set at 75% by the Validation Management Group (VMG) for both WLR and BLR, as opposed to 80% in GARDskin. However, insufficient justification for this reduction in

the target was provided. The WLR was 62.5%, 83.3% and 88.9% in a best case scenario, and 50%, 77.8% and 77.8% in a worst case scenario (for more details, please refer to the ESAC WG Report in Annex 2), being below the set target of at least 75% concordance for one of the laboratories. The BLR for the 18 chemicals with valid results in at least two laboratories was 61.1% (95%-CI: 35.7%-82.7%), which was also below the set target. The ESAC also calculated the BLR for the 14 chemicals for which a prediction could be derived from all three laboratories, even if based only on two concordant experiments in a laboratory (instead of three), and this was still below the set target: 71.4% (95%-CI: 41.9-91.6%).

Combined GARDskin + GARDpotency strategy

The ESAC considers that the issues with GARDpotency stated above must be addressed before a combined approach can be assessed for its reproducibility.

Predictive capacity

GARDskin

The ESAC considers that the GARDskin assay has a performance that appears to be comparable to *in vitro*/*in chemico* methods currently adopted to support skin sensitisation hazard identification, although the Confidence Intervals for specificity are very large.

The ESAC evaluated the predictive capacity of GARDskin on the basis of individual experiments and taking into account only chemicals that were not used for training the SVM algorithm. Based on this, the sensitivity of GARDskin ranged from 76% to 100% across experiments and laboratories, with the width of the 95%-CIs between 22 and 45 percentage points and all lower Confidence limits above 50%. The specificity ranged from 40% to 100%, with the width of the 95%-CIs being at least 63 and up to 80 percentage points. The accuracy ranged from 73% to 100%, with the width of the 95%-CIs being at least 21 and up to 36 percentage points and all lower Confidence limits above 50%.

The ESAC concludes that the predictive capacity of GARDskin is appropriate to support the discrimination between skin sensitisers (Cat. 1) and chemicals not classified for skin sensitisation (No Cat.) according to UN GHS. Nonetheless, it would have been useful to have included a larger number of chemicals (especially non-sensitisers), which had not been used to train the model, to increase the precision of the estimates.

GARDpotency

The ESAC evaluated the predictive capacity of GARDpotency on the basis of individual experiments and taking into account only chemicals that were not used for training the SVM algorithm. The correct Cat. 1A classification rate ranged from 0% to 100% across experiments and laboratories, with the width of the 95%-CIs between 53 and 97 percentage points and all lower Confidence limits below 50%. The correct Cat. 1B

classification rate ranged from 50% to 100% with the width of the 95%-CIs between 52 and 98 percentage points and all lower Confidence limits below 50%.

The ESAC concludes that the available data are not sufficient to support the use of GARDpotency to discriminate Cat. 1A and Cat. 1B skin sensitisers at this time.

Combined GARDskin + GARDpotency strategy

The ESAC considers that the issues with GARDpotency stated above must be addressed before a combined approach can be assessed for its performance.

Applicability domain

The ESAC considers that the GARD platform validation studies were not properly designed to investigate and define limitations in the applicability domain of the test methods. Nevertheless, the GARD is likely to have similar limitations to those of the currently adopted *in vitro* methods based on submerged cultures, such as solubility, stability in culture media and volatility. As more diverse chemicals are tested in the GARD platform, this will enable a better characterisation of the applicability domain. While current data on the use of GARDskin for the identification of pro-haptens look promising, the ESAC recommends further characterisation of the metabolic capacity of the test system, both from relevant protein expression and enzymatic activity perspectives.

It is important to identify or foresee possible limitations to properly inform test method users. While chemicals may not be excluded without experimentation, caution must be used in the case of negative results. For the time being, as the applicability domains and limitations have not been fully explored, negative results should be carefully investigated in combination with other evidence.

Conclusions and Recommendations

The ESAC concludes that the evidence provided on GARDskin is sufficient and adequate to support its scientific validity. Thus, the ESAC considers that GARDskin is ready to progress to further consideration by the OECD for Test Guideline development. GARDskin can contribute to skin sensitisation hazard identification in a weight-of-evidence approach. Depending on the regulatory context, positive results obtained with GARDskin may be used stand-alone to identify skin sensitisers. However, a negative result obtained with this assay may not be sufficient stand-alone evidence to identify non-sensitisers and should be considered together with additional evidence.

In contrast, the ESAC does not consider the information currently available on GARDpotency to be sufficient, at present, to recommend its use for regulatory purposes (in combination with any other assay). The use of the GARDpotency assay for discriminating Cat. 1A and Cat. 1B sensitisers is currently prevented by

the identified issues in reproducibility and predictive capacity due to the design of the validation study.

References

OECD (2012) The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. OECD Environment, Health and Safety Publications; Series on Testing and Assessment, No. 168. Organisation for Economic Co-operation and Development, Paris. ENV/JM/MONO(2012)10.

OECD (2018a) Test Guideline No. 442D - Key Event-Based Test Guideline for in vitro skin sensitisation assays addressing the AOP Key Event on Keratinocyte Activation. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.

OECD (2018b) Test Guideline No. 442E - Key Event-Based Test Guideline for in vitro skin sensitisation assays addressing the Key Event on Activation of Dendritic Cells on the Adverse Outcome Pathway for skin sensitisation. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.

OECD (2021) Test Guideline No. 442C - Key Event-Based Test Guideline for in chemico skin sensitisation assays addressing the Adverse Outcome Pathway Key Event on Covalent Binding to Proteins. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.

UN (2021) Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Ninth revised edition, United Nations, New York and Geneva. ST/SG/AC.10/30/Rev.9.

Annex 1

COMPOSITION OF THE ESAC AND ESAC WORKING GROUP

Composition of the ESAC and ESAC Working Group

EURL ECVAM Scientific Advisory Committee (ESAC)

Core Members

- Dr. Chantra ESKES (ESAC Chair)
- Prof. Paula M. ALVES
- Dr. Rebecca CLEWELL
- Prof. Emanuela CORSINI
- Prof. Ian COTGREAVE
- Prof. Annette KOPP-SCHNEIDER
- Dr. José Maria NAVAS ANTÓN
- Prof. Aldert PIERSMA
- Dr. Carl WESTMORELAND

ESAC Working Group (WG)

- Prof. Emanuela CORSINI (WG Chair)
- Dr. Rebecca CLEWELL
- Prof. Ian COTGREAVE
- Dr. Chantra ESKES
- Prof. Annette KOPP-SCHNEIDER
- Dr. Carl WESTMORELAND

EURL ECVAM (Secretariat)

- Dr. João BARROSO (ESAC Coordinator)
- Dr. Silvia CASATI (ESAC WG Coordinator)
- Dr. David ASTURIOL
- Prof. Maurice WHELAN (Head of Unit)

Annex 2

ESAC WORKING GROUP REPORT

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
JOINT RESEARCH CENTRE
Directorate F - Health, Consumers and Reference Materials
European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM)

EURL ECVAM
SCIENTIFIC
ADVISORY
COMMITTEE
(ESAC)

ESAC WORKING GROUP REPORT

on the

Scientific Validity of the GARDskin and GARDpotency Test Methods

Title page information	
File name	ESAC_WG_Report_GARD.docx
Abbreviated title of ESAC request	GARD
Relating to ESAC REQUEST Nr.	2020-01
Request discussed through	Written procedure following ESAC 46 (December 2019)
Report to be handed over to ESAC Chair and EURL ECVAM Coordinator by	Emanuela Corsini (Working Group Chair)

Version tracking

Date	**Version**	**Author(s)**	**Description**
08/03/2021	V1.0	ESAC WG	First agreed draft of ESAC WG Report
06/07/2021	V2.0	ESAC WG	Final approved draft of ESAC WG Report sent to ESAC for endorsement

Table of Contents

ESAC Working Group

Full title: ESAC Working Group on the GARD Test Methods

Abbreviated title: ESAC WG GARD

The ESAC Working Group (WG) was established in May 2020 by written procedure to assist in the production of an ESAC Opinion on the scientific validity of the GARDskin and GARDpotency *in vitro* test methods. GARDskin and GARDpotency are separate assays with different intended purposes. GARDskin provides hazard identification of skin sensitisers (UN GHS Cat. 1) and non-sensitisers (UN GHS No Cat.), while GARDpotency provides hazard characterisation by subcategorisation of skin sensitisers (UN GHS Cat. 1A vs Cat. 1B). These two methods are also proposed to be used in combination, where a positive result in GARDskin is analysed with GARDpotency to determine the UN GHS subcategory.

This report was prepared at the request of EURL ECVAM by the "ESAC Working Group on the GARD" (ESAC WG), which was charged with conducting a detailed scientific peer review of the external validation study of the GARDskin and GARDpotency *in vitro* test methods. The basis for the scientific peer review was the EURL ECVAM Request for ESAC Advice approved by the ESAC by written procedure following the ESAC46 plenary meeting of December 2019 (ESAC request 2020-01).

The ESAC WG met virtually on 29/06/2020; 18 and 24/09/2020; 9, 21 and 22/10/2020; 23 and 24/11/2020; 10 and 15/12/2020; 8 and 18/01/2021; 19 and 24/02/2021; 17 and 31/03/2021; 27/04/2021; 21/05/2021 to conduct its peer review. This ESAC WG Report was endorsed by the ESAC WG on 07/07/2021 and represents its consensus view. The Report was endorsed by the ESAC on 07/07/2021.

The ESAC WG had the following members:

- Prof. Emanuela CORSINI (ESAC Core Member, WG Chair)
- Dr. Rebecca CLEWELL (ESAC Core Member)
- Prof. Ian COTGREAVE (ESAC Core Member)
- Dr. Chantra ESKES (ESAC Core Member)
- Prof. Annette KOPP-SCHNEIDER (ESAC Core Member)
- Dr. Carl WESTMORELAND (ESAC Core Member)

EURL ECVAM (Secretariat):

- Dr. João BARROSO (ESAC Coordinator)
- Dr. Silvia CASATI (ESAC WG Coordinator)
- Dr. David ASTURIOL

ABBREVIATIONS USED IN THE DOCUMENT

- **AOP** Adverse Outcome Pathway
- **BLR** Between-laboratory reproducibility
- **BRT** Burleson Research Technologies
- **CI** Confidence Interval
- **DA** Defined Approach
- **DC** Dendritic Cell
- **DV** Decision Value
- **EG DASS** Expert Group on Defined Approaches for Skin Sensitisation
- **ESAC** EURL ECVAM Scientific Advisory Committee
- **ESAC WG** ESAC Working Group
- **EURL ECVAM** European Union Reference Laboratory for Alternatives to Animal Testing
- **GARD** Genomic Allergen Rapid Detection
- **GD** Guidance Document
- **GDAA** GARD Data Analysis and Application
- **GLP** Good Laboratory Practice
- **GPPS** GARD Potency Prediction Signature
- **GPS** GARD Prediction Signature
- **JRC** Joint Research Centre
- **KE** Key Event
- **NS** Non-Sensitiser
- **OECD** Organisation for Economic Co-operation and Development
- **QA** Quality Assurance
- **QC** Quality Control
- **SOP** Standard Operating Procedure
- **S** Sensitiser
- **SVM** Support Vector Machine
- **TG** Test Guideline
- **TGP** Test Guidelines Programme
- **UN GHS** United Nations Globally Harmonized System of Classification and Labelling of Chemicals
- **VMG** Validation Management Group
- **VSR** Validation Study Report
- **WLR** Within-laboratory reproducibility

1. Study objective and design

1.1 Analysis of the clarity of the study objective's definition

(a) ESAC WG summary of the study objective as outlined in the Validation Study Report

GARDskin and GARDpotency are separate assays with different intended purposes. GARDskin provides hazard identification of skin sensitisers (Cat. 1) and non-sensitisers (No Cat.), while GARDpotency provides hazard characterisation by subcategorisation of skin sensitisers (Cat 1A vs Cat. 1B), according to the United Nations Globally Harmonized System of Classification and Labelling of Chemicals (UN GHS) (UN, 2021). These two methods are also proposed to be used in combination, where a positive result in GARDskin is analysed with GARDpotency to determine the UN GHS subcategory.

In the validation study report (VSR) for GARDskin the overall objective for the validation study was: 1) To demonstrate the transferability and reproducibility of the Genomic Allergen Rapid Detection (GARD) platform and 2) to provide evidence supporting the GARDskin method as being a reliable tool for assessing skin sensitisation hazard, with added value to any integrated testing and assessment strategy in which it is included.

The GARDpotency method was proposed as being a second tier for assessing the potency (Cat. 1A/Cat. 1B) of the skin sensitisation hazard, with added value to any integrated testing and assessment strategy in which it is included. There were 2 specific objectives: (1) To demonstrate reproducibility of the method in view of integration into testing strategies for risk assessment and (2) to provide evidence that the GARDpotency discriminates Cat 1A and Cat 1B skin sensitisers.

In subsequent correspondence, the test developer indicated that the data presented supports the use of the GARDskin and GARDpotency assays as a stand-alone testing strategy, should there be sufficient scientific support for this.

(b) Appraisal of clarity of study objective as outlined in the Validation Study Report

Since the ESAC WG received two separate reports (GARDskin and GARDpotency), it was not initially clear that a stand-alone approach was proposed. Following a request for clarification, the test method developers initially stated that the two methods were to be used as part of testing strategies/Defined Approaches (DAs), but later stated that: “It is the opinion of the method developers that available evidence scientifically supports the potential adaptation of the GARDskin and/or GARDpotency assay(s) as a stand-alone strategy”. Due to the conflicting purposes stated by the test developer, the ESAC WG decided to proceed to the peer-review of the scientific evidence regarding:

1. GARDskin results to support the discrimination between sensitisers (S) and non-sensitisers (NS), in line with the purposes of the validation studies conducted to assess reproducibility and transferability of the methods adopted within OECD Test Guidelines (TGs) 442C, 442D and 442E (OECD, 2018a, 2018b, 2021a).
2. GARDpotency results to support the identification of Cat 1A and Cat 1B sensitisers.
3. The use of the combination of GARDskin and GARDpotency as a stand-alone strategy. No information was provided about the potential use of GARDskin/potency data in the context of DAs as defined by the OECD in Guidance Documents (GDs) N°255 and N°256 (OECD, 2016a, 2016b).

1.2 Quality of the background provided concerning the purpose of the test method

GARDskin and GARDpotency are proposed for regulatory testing for skin sensitisation hazard identification and potency assessment, but could also have application as a screening tool during early product development.

(a) Analysis of the scientific rationale provided in the Validation Study Report

The ESAC WG considers that the information provided in the GARDskin and GARDpotency submissions is of sufficient quality to evaluate the scientific rationale of the methods. Both the GARDskin and the GARDpotency are based on the transcriptional changes of selected genomic biomarkers referred to as the GARD Prediction Signature (GPS) in the GARDskin and GARD Potency Prediction Signature (GPPS) in the GARDpotency.

(a) The selection of genes used in the GARDskin (200 genes were originally selected from Affymetrix data but only 196 of those genes are currently measured with nanoString in the latest version of the method) and GARDpotency (52 genes were originally selected from Affymetrix data but only 51 of those genes are currently measured with nanoString in the latest version of the method) was not based on scientific understanding of the mechanisms underpinning *in vivo* skin sensitisation, but was driven by the statistical analysis of gene expression changes (from whole genome array analysis) in the test system following treatment with sensitisers and non-sensitisers. Subsequent analysis by the test developer of the genes used has provided a mechanistic rational for many of the genes selected and their respective relevance to different key events in the skin sensitisation Adverse Outcome Pathway (AOP) (OECD, 2012) (i.e., not solely related to Key Event (KE) 3). However, as multiple cell types are involved with the induction of skin sensitisation, the ESAC WG believes that both the biological system(s) and the genes induced must be considered independently.

(b) The selection of skin sensitisers/non-sensitisers for gene identification were used for prediction of skin sensitisation as an upstream event for allergic contact dermatitis.

(b) Analysis of the regulatory rationale provided in the Validation Study Report

The regulatory rationale provided in the VSRs is sufficient. The GARDskin and GARDpotency test methods were proposed to extend the existing toolbox for testing of potential skin sensitising substances currently including OECD TGs 406, 429, 442A, 442B, 442C, 442D and 442E (OECD, 2010a, 2010b, 2018a, 2018b, 2018c, 2021a, 2021b), by providing a test method for KE3.

The application of the GARDskin and GARDpotency test methods is anticipated to meet regulatory and data requirements in the context of relevant EU legislations such as REACH (EC, 2006), Cosmetic Products (EC, 2009a), Classification, Labelling and Packaging (CLP) (EC, 2008), Plant Protection Products (EC, 2009b) and Biocidal Products (EU, 2012).

Furthermore, the GARDskin and GARDpotency test methods can contribute to hazard identification and potency determination in a weight-of-evidence or read-across approach, and to reduce the number of animals when applied for screening and early decision-making during product development.

1.3 Appraisal of the appropriateness of the study design

The overall conduct of the validation study was generally in accordance with accepted practices as described in OECD GD N°34 (OECD, 2005). The study coordination and organisation, chemical selection, purchasing, coding, distribution and biostatistical evaluation were conducted independently of the test developer. Five laboratories participated in the validation study. Three laboratories performed cell culture exposures and preparation of RNA samples (the lead laboratory

SenzaGen AB, Burleson Research Technologies (BRT) and Eurofins BioPharma Products), which is standard practice for assessing reproducibility and transferability. However, SenzaGen performed nanoString nCounter analysis in-house, whereas the other two laboratories shipped their samples to Covance Genomics Laboratory and KIGene for nanoString nCounter analysis. This is unconventional, but acceptable if quality assurance (QA) controls are maintained and documented across the additional laboratories. Regarding the number of chemicals tested, no documentation was provided on power calculations conducted before the GARDskin and GARDpotency validation study started.

To this end, a post-hoc analysis, conducted by the test developer to estimate the sample size required for both GARDskin and GARDpotency, was shared following a request from the ESAC WG. This analysis showed that the number of required samples for GARDskin reproducibility was 26 and for GARDpotency predictive performance was estimated as 57 samples, albeit that the parameters chosen for these calculations were not clearly justified and the results of the statistical hypothesis tests for which these power calculations were performed were not shown.

GARDskin

The number of laboratories (3 laboratories), number of chemicals (19 UN GHS sensitisers and 9 non-sensitisers) and distribution of chemicals was comparable to those from previous validation studies that assessed the reproducibility of the test methods falling within TGs 442C, 442D and 442E, to support skin sensitisation hazard identification. In addition, the number of chemicals used was discussed in a teleconference of the OECD Expert Group on Skin Sensitisation, held on 12 September 2016, who recommended alignment with the previous validation studies. However, the ESAC WG notes that a large number of the sensitisers (6 out of 19) and the majority of the non-sensitisers (5 out of 9) tested in the three laboratories during the validation study of GARDskin, were also part of the training set used to build the GARDskin prediction model (for details see Table 1).

In addition to the data generated in multiple laboratories, another set of 31 coded chemicals were tested within the SenzaGen laboratory only, for GARDskin and subsequent analysis also with GARDpotency, to further assess the predictive capacity of the two test methods. These chemicals comprised 11 Cat. 1A chemicals, 12 Cat. 1B chemicals and 8 non-sensitisers, which were used by the ESAC WG to assess the GARDskin's predictive capacity. Only one Cat. 1A (2-nitro-1,4-phenylenediamine), one Cat. 1B (penicillin G potassium salt) and three No Cat. (4-aminobenzoic acid, 4-hydroxybenzoic acid and sodium lauryl sulphate) out of these 31 additional chemicals tested at SenzaGen were part of the training set used to build the GARDskin prediction model (see Table 1).

In addition to the data generated in the validation study, the submission mentions further data generated on 25 sensitisers and 16 non-sensitisers (Forreryd et al., 2016) and data obtained with "difficult to test" chemicals (Johansson et al., 2014). The data generated by the test developer and published by Johansson et al. (2014) are however of less relevance as they were generated using Affymetrix analysis, and not by the nanoString platform. Further data generated using the nanoString platform were produced in a blind Cosmetics Europe study (Johansson et al., 2017).

In conclusion, the study design based on number of laboratories and chemicals used to assess the within-laboratory reproducibility (WLR), between-laboratory reproducibility (BLR) and predictive capacity of GARDskin was considered to be sufficient by the ESAC WG, as it is comparable to currently adopted test methods. Nonetheless, it would have been useful to have included a larger number of test chemicals in the multilaboratory study (especially non-sensitisers) which had not been used to train the model (see Section 10).

Table 1: Chemicals used in the validation phase that had been used in the training phase of the prediction model.

Chemical name	CAS number	Chemical set	GHS Category	Used to train GARDskin	Used to train GARDpotency and define GPPS	Used to train the tiered approach (GARDskin + GARDpotency)
4-Nitrobenzyl bromide	100-11-8	Ring trial	Cat. 1A	N	Y (both)	Y (only GARDpotency)
Cinnamal	104-55-2	Ring trial	Cat. 1A	N	Y (both)	Y (only GARDpotency)
Formaldehyde	50-00-0	Ring trial	Cat. 1A	Y	Y (both)	Y (both)
Lauryl gallate	1166-52-5	Ring trial	Cat. 1A	N	Y (both)	Y (only GARDpotency)
Propyl gallate	121-79-9	Ring trial	Cat. 1A	N	Y (only to train SVM)[a]	Y (only GARDpotency)
Isoeugenol	97-54-1	Ring trial	Cat. 1A	Y	Y (both)	Y (both)
2-Mercaptobenzothiazole	149-30-4	Ring trial	Cat. 1A	Y	Y (only to define GPPS)	Y (both)
4-(Methylamino)phenol sulphate	55-55-0	Ring trial	Cat. 1A	N	Y (both)	Y (only GARDpotency)
Diethyl maleate	141-05-9	Ring trial	Cat. 1B	N	Y (only to train SVM)[a]	Y (only GARDpotency)
Ethylene diamine	107-15-3	Ring trial	Cat. 1B	Y	Y (both)	Y (both)
Benzyl benzoate	120-51-4	Ring trial	Cat. 1B	N	Y (only to train SVM)[a]	Y (only GARDpotency)
Cinnamyl alcohol	104-54-1	Ring trial	Cat. 1B	Y	Y (both)	Y (both)
Citral	5392-40-5	Ring trial	Cat. 1B	N	Y (both)	Y (both)
Eugenol	97-53-0	Ring trial	Cat. 1B	Y	Y (both)	Y (both)
Dextran	9004-54-0	Ring trial	No Cat.	N	Y (only to define GPPS)[b]	Y (only GARDpotency)
Glycerol	56-81-5	Ring trial	No Cat.	Y	Y (only to define GPPS)[b]	Y (both)
Hexane	110-54-3	Ring trial	No Cat.	N	Y (only to define GPPS)[b]	Y (only GARDpotency)
Isopropanol	67-63-0	Ring trial	No Cat.	Y	Y (only to define GPPS)[b]	Y (both)
Kanamycin	70560-51-9	Ring trial	No Cat.	N	Y (only to define GPPS)[b]	Y (only GARDpotency)
Lactic acid	50-21-5	Ring trial	No Cat.	Y	Y (only to define GPPS)[b]	Y (both)
Propylene glycol	57-55-6	Ring trial	No Cat.	Y	N	Y (only GARDskin)
Salicylic acid	69-72-7	Ring trial	No Cat.	Y	Y (only to define GPPS)[b]	Y (both)
2-Nitro-1,4-phenylendiamine	5307-14-2	Extra chemicals	Cat. 1A	Y	Y (both)	Y (both)
p-Benzoquinone	106-51-4	Extra chemicals	Cat. 1A	N	Y (both)	Y (only GARDpotency)
Penicillin G Potassium salt	113-98-4	Extra chemicals	Cat. 1B	Y	N	Y (only GARDskin)
4-aminobenzoic acid	150-13-0	Extra chemicals	No Cat.	Y	N	Y (only GARDskin)
4-Hydroxybenzoic acid	99-96-7	Extra chemicals	No Cat.	Y	N	Y (only GARDskin)
Sodium lauryl sulphate	151-21-3	Extra chemicals	No Cat.	Y	Y (only to define GPPS)[b]	Y (both)

[a] Gradin et al., 2020

[b] Zeller et al., 2017

GPPS: GARD Potency Prediction Signature

GARDpotency

Frozen mRNA samples from the GARDskin multilaboratory study were used for analysis in the GARDpotency. The ESAC WG notes that a majority of the sensitisers from the multilaboratory study analysed with GARDpotency were also part of the chemical set used to train the GARDpotency

Support Vector Machine (SVM) algorithm (13 out of 19, i.e., 7 out of 11 UN GHS Cat. 1A and 6 out of 8 UN GHS Cat. 1B). In addition, 11 of the 19 tested sensitisers were used to define the GPPS (for details see Table 1). In total, 14 out of 19 sensitisers were part of the training set used to build the GARDpotency SVM algorithm and/or define the GPPS. Furthermore, only samples identified as sensitisers by the GARDskin, within each laboratory and within each run, were subsequently analysed in GARDpotency. This led to differences in the number of chemicals analysed with GARDpotency by the different laboratories, as well as in differences in the number of runs available for each chemical (see below). For example, of the three runs conducted in the GARDskin for benzyl benzoate at the SenzaGen laboratory, only one run gave a sensitiser result and that was the only sample processed in the GARDpotency assay (see GARDskin and GARDpotency Data.xls in the GARDpotency submission). Finally, for some chemicals, the quality acceptance criteria for GARDskin or GARDpotency were not met.

As a consequence, the number of runs available for each chemical varied. For 12 of the 19 tested sensitisers, data were available for the foreseen 9 runs (i.e., 3 runs/laboratory tested in 3 laboratories). In contrast, for 6 out of the 19 tested sensitisers, a total of 5 to 8 runs were available (with at least one run per laboratory). Furthermore, for one chemical (i.e., ethylene diamine), only data from one run (and one laboratory) was available, due to the fact that this chemical was mostly predicted to be non-sensitiser with GARDskin in most laboratories and runs. In addition, in some cases, GARDpotency data was generated on non-sensitiser chemicals predicted to be sensitisers by GARDskin. This was the case for 6 of the 9 non-sensitisers tested, where only a minority of runs lead to data being generated on GARDpotency (1 run out of 9) corresponding to four chemicals (i.e., isopropanol, kanamycin, lactic acid and salicylic acid). Furthermore, two chemicals had two runs leading to data being generated on GARDpotency, one of which (vanillin) had two runs in the same laboratory leading to GARDpotency prediction, whereas the second chemical (propylene glycol) had one overpredicted run in each of two laboratories.

An additional set of 31 chemicals were tested within the SenzaGen laboratory for GARDskin and subsequently GARDpotency. These chemicals comprised 11 Cat. 1A chemicals, 12 Cat. 1B chemicals and 8 non-sensitisers. Only two Cat. 1A (2-nitro-1,4-phenylenediamine and p-benzoquinone, see Table 1) out of the 23 additional sensitisers tested at SenzaGen were part of the training set used to build the GARDpotency SVM algorithm. These two same Cat. 1A chemicals were also used to define the GPPS (see Table 1). Furthermore, the test developer shared during the peer-review process, the WLR results obtained for these additional chemicals. Based on the classification obtained with GARDskin, data on GARDpotency were generated for 21 of the 23 sensitisers (1,2-cyclohexane dicarboxylic anhydride and benzyl salicylate were identified as non-sensitisers by the GARDskin in all the three runs), out of which three chemicals had GARDpotency results from only one run (phthalic anhydride, maleic anhydride and penicillin G potassium salt) and two chemicals from two runs (benzylcinnamate and diethyl acetaldehyde). The remaining 16 chemicals had GARDpotency results from three runs. In addition, GARDpotency data was generated for 5 of the 8 non-sensitiser chemicals predicted to be sensitisers by GARDskin in at least one run. One non-sensitiser (hydrocortisone) was consistently overpredicted as a sensitiser by GARDskin in all three runs, one other non-sensitiser (phenoxyethanol) was overpredicted as a sensitiser in two of the three runs, and finally three non-sensitisers (sulphanilic acid, sodium lauryl sulphate and triethanolamine) were overpredicted in only one run by GARDskin.

Based on the above observations, a total of 12 sensitisers (out of 19) had a complete dataset (i.e., 3 runs/laboratory in 3 laboratories) for assessing the WLR and BLR of GARDpotency assay, and an additional 3 chemicals had data on 8 out of the 9 foreseen runs. In addition, within-laboratory data on 3 runs was available within the SenzaGen laboratory for an additional set of 16 sensitisers and one non-sensitiser (hydrocortisone) consistently classified as sensitiser in the three runs, resulting in a total of 29 chemicals having data on three runs for WLR assessment.

The number of chemicals having a full dataset for assessing BLR with the GARDpotency was, therefore, lower than those used for GARDskin (12 to 15 samples vs 28), but a similar number of

chemicals had a full dataset available for assessing WLR, albeit in only one laboratory (i.e., 28 in both cases).

Regarding predictive capacity, information was available on a total of 40 sensitisers, out of which 16 were part of the training set used to build the GARDpotency SVM algorithm and/or define the GPPS, resulting in a total of 24 new sensitisers assessed during the validation study.

In conclusion, the ESAC WG expresses concerns on the design of the GARDpotency validation study, due both to the inconsistent number of runs available for each chemical for assessing the method's reproducibility, and to the limited number of new chemicals (testing set) assessed for predictive capacity.

GARDskin+GARDpotency

Regarding the combination of GARDskin and GARDpotency as a stand-alone strategy, no power calculation was shared with the ESAC WG on the number of chemicals necessary to assess the overall performances of the combined strategy (GARDskin + GARDpotency), to predict the three UN GHS categories, i.e., Cat. 1A, Cat. 1B and No Cat. ***The ESAC WG has concerns about the lack of a clear justification for the number of chemicals tested, taking into account that the strategy aims to predict three categories, i.e., UN GHS Cat. 1A, UN GHS Cat. 1B and UN GHS No Cat. Furthermore, there are concerns regarding the number of training chemicals included in the validation dataset (17 out of 42 S, and 11 out of 17 NS).***

1.4 Appropriateness of the statistical evaluation

Several of the chemicals used for training set were also used to evaluate predictive capacity. While the validation study Chemicals Selection Group was aware of the fact that training chemicals were included as test chemicals, they argued that this does not affect the value of the determined WLR and BLR values, but rather that it affects the credibility of the performance values and they provided arguments for retaining this selection. The ESAC WG considers that even WLR and BLR may be affected by the use of training chemicals, since these may be easier to classify. However, the ESAC WG recognises that previous validation studies have included substances used during the development of the test method to assess WLR and BLR.

Most importantly, the ESAC WG considers that unbiased estimation of predictive capacity of an algorithm based on machine learning requires the use of chemicals not contained in the training set for gene selection and model building. The ESAC WG hence derived the predictive capacity excluding the chemicals used during training of the SVM algorithms, i.e., the chemicals used to define the GPS and train GARDskin (Johansson et al., 2011) and the chemicals used to define the GPPS and train GARDpotency (Zeller et al., 2017 and Gradin et al., 2020) (results shown in Section 10).

Confidence intervals (CIs) should be given to assess the variability of the estimates or WLR, BLR and predictive capacity. This is especially important for low numbers of test chemicals. Since no confidence intervals were provided at all in the reports, the ESAC WG calculated all relevant 95%-CIs for the proportions by the Clopper-Pearson method (using the function binomCI of the R Package DescTools with R Version 4.0.2).

For the sample size justification, given *post hoc*, there was no clear rationale provided for the selection of parameters (alpha, power, null hypothesis value and expected proportion), and the values chosen for power are non-standard, making the sample size calculation somewhat arbitrary. Further, a sample size was calculated on the basis of achieving the specified power for statistical hypothesis testing given alpha, null hypothesis value and expected proportion. However, results of statistical hypothesis testing, e.g., p-values, were not given in the statistical analysis reports. Therefore, it is not clear to the ESAC WG how the power calculations support the number of samples tested in the validation study.

2. Collection of existing data

2.1 Existing data used as reference data

According to the GARDskin VSR, the primary source of reference data came from peer-reviewed publications such as Natsch et al. (2013) and Basketter et al. (2014), that together provided LLNA, human as well as *in vitro* data on 145+ chemicals.

For the GARDpotency assay, the validation study re-used the mRNA samples from the GARDskin validation study. In addition, 31 chemicals were selected for coded testing by the lead laboratory using the same criteria used for the GARDskin validation study (see Appendix 1 from original submission).

2.2 Existing data used as testing data

GARDskin

In addition to the data generated during the validation study, in-house data were used to assess the predictive capacity of GARDskin, as published by Forreryd et al. (2016), which were obtained using the same protocol version as the one agreed after the training phase of the validation study (Standard Operating Procedure (SOP) v. 04.04). This version of the protocol represents an earlier version as the version used during the validation, in which further updates were included after the transferability phase (SOP v. 05.01) (differences of protocols explained in p. 29-30 of GARDskin VSR).

Furthermore, data obtained with an older protocol of the GARDskin method, where 37 coded chemicals were tested, were also used for comparative purposes (Johansson et al., 2014).

GARDpotency

In the case of GARDpotency, only the chemicals tested and coded during the validation study were used. There were no additional retrospective data used to establish the predictive capacity of the assay.

2.3 Search strategy for retrieving existing data

Not applicable.

2.4 Selection criteria applied to existing data

No selection criteria were applied to the existing data because all data were either generated in the context of the validation study or by the test developer.

3. Quality aspects relating to data generated during the study

3.1 Quality assurance systems used when generating the data

In terms of quality aspects relevant to the data generated during this validation study, perhaps the most important factor is understanding the course taken from raw data, through processed data towards a classification, using the GARD Data Analysis and Application (GDAA) suit. Initial concerns of the work-flow in data generation and data collation on the proprietary data interface were clarified by the test developer. The test developer provided all the information necessary for the ESAC WG to replicate the proprietary algorithm, including normalisation procedures and classification process, as well as the results of the prediction model. Concerns regarding the cloud solution for test data analysis, which could allow for data corruption in transfer, are being addressed by the OECD Working Group of the National Coordinators for the Test Guidelines Programme (WNT) and Working Group on Good Laboratory Practice (GLP). The ESAC WG recommends the test developer to follow directions of this joint expert group.

With respect to the QA in the study as performed, the validation was performed at GLP accredited laboratories, but was agreed to be performed non-GLP, according to practices in individual laboratories involved. There was a set of GLP-like quality assurance measures implemented that were described in the VSRs (GARDskin: “GARDskin Validation Final” p. 12; GARDpotency: “Updated Validation Clean”, p.13).

3.2 Quality check of the generated data prior to analysis

It is stated that the Coordinator is responsible for quality control (QC) audits and that the Validation management Group (VMG) is responsible for the evaluation of the outcome of these audits. Despite this, much responsibility hinges on the role of a “biostatistician” who was responsible for data templates used in transfer. In addition to data transfer between the laboratories involved, RNA samples were transferred from the two remote laboratories to the test developer for nanoString analysis. In all cases, QC audits on raw data generation and data transfer should be evident and a statement issued by the coordinator. This is not formally evident from the reports but is implicit in the work process description provided in the VSRs and the SOPs which support this.

4. Quality of data used for the purpose of the study (existing and newly generated)

4.1 Overall quality of the evaluated testing data (newly generated or existing)

According to the VSRs, all data generated during the validation study were recorded directly, promptly and legibly by the responsible individual(s) in compliance with the following set of (non-GLP) QA measures considered essential for the acceptance of information and data produced in the validation process:

- Qualified personnel and appropriate facilities, equipment and materials to be provided.
- Records of the qualifications, training and experience, and a job description for each professional and technical individual, to be kept.
- For each study, an individual with appropriate qualifications, training and experience to be appointed responsible for its overall conduct and for any report issued.
- Instruments used for the generation of experimental data to be inspected regularly, cleaned, maintained and calibrated according to established protocols, if available, or according to manufacturer's instructions.
- Reagents to be labelled, as appropriate, to indicate their source, identity, concentration, expiring date and specific storage conditions.
- All data generated during the study to be recorded directly, promptly and legibly by the responsible individual(s). These entries shall be attributable and dated.
- All changes to data to be identified with the date and the identification of the individual responsible, and a reason for the change to be provided and explained in writing.

Prior to the start of the study, a standard reporting template was distributed to the participating laboratories. This template was developed for the study by the biostatistician, in collaboration with the lead laboratory. The template contained pre-evaluated formulae to assure the quality of the template before it was distributed.

Each document was subjected to a formal quality check procedure. The quality check focused on the acceptance criteria for the run for each chemical to assure the validity of the results. Once completed, the checklists were scanned as a PDF file and added to the Excel sheet as an item. The checked document was added to the official results folder of the study.

For statistical analysis, a summary template was designed by the biostatistician who transferred the results to this template. Preparations of the summary contained internal checks that ensured that no transcription errors were made in the transfer of the results. As an additional check, the final conclusions for each chemical were compared with the conclusions in the reports sent by the laboratories.

The following documents have been shared with the ESAC WG:

- Standard reporting template (in PDF format).
- The raw data obtained in each laboratory, as well as the historical data published by Forreryd et al. (2016).
- Excel documents summarising all data obtained with both GARDskin and GARDpotency assays, including the coding of the tested substances.
- A statistical report on the analysis conducted on for GARDskin validation study.
- A statistical report on the analysis conducted on for GARDpotency validation study.
- A report on encountered substance issues for the GARDpotency validation study.

The quality of the testing data seems therefore to be sufficient and adequate for conducting the present peer-review.

4.2 Quality of the reference data for evaluating relevance[1]

GARDskin

Regarding GARDskin, the validation study included 28 chemicals tested blind, for which different types of reference data have been provided: LLNA potency, Guinea Pig outcome, GHS potency and human potency. Although the original sources of reference data for each chemical were not provided, the VSR states that the primary source for chemicals selection were peer-reviewed publications such as Natsch et al. (2013) and Basketter et al. (2014). Furthermore, the VSR states that the selection of chemicals was based on the availability of robust *in vivo* data (LLNA and human data). In order to ensure the robustness of the reference data provided, the ESAC WG compared the LLNA potencies provided, with those recently agreed by the OECD Expert Group on Defined Approaches for Skin Sensitisation (EG DASS) (OECD, 2021c):

- Out of the 19 chemicals reported to be skin sensitisers, all were also considered to be skin sensitisers based on the LLNA *in vivo* data curated by the OECD EG DASS.
- Out of 9 chemicals reported to be UN GHS No Cat., Kanamycin sulphate (CAS 70560-51-9) was not tested in the DAs and Dextran (CAS 9004-54-0) was excluded by the OECD EG DASS because of its undefined structure. Of the remaining 7 chemicals, 6 were reported to be LLNA GHS No Cat. according to the OECD EG DASS, whereas one was reported to be LLNA Cat. 1B (Salicylic acid, CAS 69-72-7). However, regarding salicylic acid (CAS 69-72-7), it should be noted that a recent SCCS weight-of-evidence assessment concluded salicylic acid a non-sensitiser (SCCS, 2019). Furthermore, other literature sources have also considered salicylic acid to be a non-sensitiser (Casati et al., 2009; Natsch et al., 2013).
- It is to be noted that the OECD EG DASS Database was not available at the time of the conduct of the GARDskin validation study. Therefore, the OECD DASS EG observations should not overrule the historical data available and also cited above. Furthermore, it is noted that no quality check was conducted on the other reference data sources than the LLNA *in vivo* data (i.e., human data, Guinea pig). This is due to the fact that the validation and statistical reports on GARDskin did not use these other reference data when evaluating the performances of GARDskin. Based on the above observations, the *in vivo* LLNA UN GHS classifications reported in the GARDskin validation report were considered to be appropriate.

GARDpotency

Of the 28 chemicals tested in the GARDskin, only those leading to a GARDskin positive outcome were tested in the GARDpotency. It is understood that 'GHS potency' has been used as the main reference to compare the study results with and that it refers to the LLNA obtained UN GHS classification. However, no clear mention about this could be found in the materials provided to the ESAC WG. In order to ensure the robustness of the reference data provided, the ESAC WG compared the LLNA potencies provided, with those recently compiled by the OECD EG DASS (OECD, 2021c) and made the following observations:

- One of the chemicals, 4(methylamino)phenol sulphate (Metol, CAS 55-55-0), was reported to be classified as Cat. 1B in the GARD statistical report and summary excel sheet of all GARD results, but it is reported to be classified as Cat. 1A in the GARDpotency validation study report. Thus, in total 11 GHS potency Cat. 1A and 8 GHS potency Cat. 1B were considered for the validation of the GARDpotency.
- Out of 11 chemicals reported to have a GHS potency Cat. 1A, 10 were reported to be LLNA GHS Cat. 1A according to the OECD EG DASS, whereas one was reported to be LLNA Cat. 1

[1] OECD Guidance Document No. 34 on validation defines relevance as follows: "Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of accuracy (concordance) of a test method."

with no possibility of distinction between 1A and 1B (formaldehyde, CAS 50-00-0). However, this chemical has also been reported to be a human UN GHS Cat. 1A by the OECD DASS EG, to be a LLNA strong sensitiser (EURL ECVAM, 2012; Natsch et al., 2013) and to be a human potency Cat. 2 (Basketter et al., 2014). Therefore, it is considered to be appropriate to assign it a UN GHS Cat. 1A classification.

- Out of 8 chemicals reported to have a GHS potency Cat. 1B, all of them were reported to be LLNA GHS Cat. 1B according to the OECD EG DASS.
- It is to be noted that the OECD EG DASS Database was not available at the time of the conduct of the GARDpotency validation study. Therefore, the OECD DASS EG observations should not overrule the historical data available and also cited above. Furthermore, it is noted that no quality check was conducted on the other reference data sources than the LLNA *in vivo* data (i.e., human data, Guinea pig). This is due to the fact that the validation and statistical reports on GARDpotency did not use these other reference data when evaluating the performances of GARDpotency. Based on the observations above-listed nevertheless, the *in vivo* LLNA UN GHS classifications reported for the set of 28 chemicals tested in the multilaboratory trial were considered to be appropriate.

In addition to these 28 chemicals, 31 chemicals were selected and tested blind, in three runs, by one laboratory. Here again, only those chemicals leading to a GARDskin positive outcome were evaluated with GARDpotency, but not those leading to non-sensitising outcome. In order to ensure the robustness of the reference data provided, the ESAC WG compared the LLNA potencies provided, with those recently compiled by the OECD EG DASS (OECD, 2021c) and made the following observations:

- Out of 11 chemicals reported to have a GHS potency Cat. 1A, 9 were reported to be LLNA GHS Cat. 1A according to the OECD EG DASS. One chemical, Beryllium sulphate (CAS 7787-56-6), was not found in the OECD DASS EG Database, but was considered to be a UN GHS Cat. 1A in the EURL ECVAM validation study of the DPRA assay (EURL ECVAM, 2012). Furthermore, 1-phenyl-1,2-propanedione (CAS 579-07-7) was reported to be LLNA Cat. 1 with no possibility of distinction between 1A and 1B by the DASS EG, but was considered to be a LLNA moderate sensitiser according to Natsch et al. (2013). It may be therefore considered as a borderline Cat. 1A/1B.
- Out of 12 chemicals reported to have a GHS potency Cat. 1B, 8 were reported to be LLNA GHS Cat. 1B according to the OECD EG DASS, whereas one was reported to be LLNA Cat. 1 with no possibility of distinction between 1A and 1B (Benzylidene acetone, CAS 122-57-6) but was considered to be a moderate sensitiser in previous studies (EURL ECVAM, 2012; Natsch et al., 2013). Furthermore, two chemicals were not included in the DASS EG database due to unclear characterisation (1-Thioglycerol, CAS 96-27-5 and Glyceryl monothioglycolate, CAS 30618-84-9). However, 1-Thioglycerol (96-27-5) was considered to be a UN GHS Cat. 1B during the EURL ECVAM validation study of the DPRA assay (EURL ECVAM, 2012). Finally, two chemicals were not found in the OECD DASS EG Database (Penicillin G Potassium salt, CAS 113-98-4 and Diethyl acetaldehyde, 105-57-1).
- Out of 8 chemicals reported to be UN GHS No Cat., only 3 were reported to be LLNA GHS No Cat. according to the OECD EG DASS. One chemical, sodium lauryl sulphate (CAS 151-21-2), was considered to be UN GHS Cat. 1B although it is also reported to be a human No Cat. by the DASS EG and a Cat. 6 by Basketter et al. (2014). It could be a false positive in the DASS EG database due to the fact that it is a known irritant that could result in a LLNA response. Furthermore, four chemicals had LLNA data considered to be not robust enough to assign No Cat (CAS 150-13-0, 121-57-3, 50-23-7 and 122-99-6). However, 4-aminobenzoic acid (CAS 150-13-0) was considered to be a LLNA non-sensitiser in the EURL ECVAM validation study of the DPRA assay (EURL ECVAM, 2012); and sulphanilic acid (CAS 121-57-3) was considered to be a LLNA non-sensitiser in Natsch et al. (2013).

It is to be noted that the OECD DASS EG applied very stringent criteria, especially for identifying UN GHS No Cat. chemicals, so that the DASS database has very few No Cat. chemicals (OECD, 2021c). Furthermore, such a database was not available at the time of the conduct of the GARDpotency validation study. Finally, although the GARDpotency VSR did not provide supporting reference for the LLNA *in vivo* classifications for the set of 31 additional chemicals, the test developer provided supporting references for all of the 31 additional chemicals tested (cf. table 8 in Appendix 4 of letter to ESAC as follow-up to the TC from 18 September 2020). Three Cat. 1A (out of 12) and two No Cat. (out of 8) did not have a LLNA supporting reference, but the test developer provided information and supporting reference on human sensitisation effects. Based on these observations, it is considered that the quality of the reference data supporting the chemicals selection mentioned above is mostly appropriate.

4.3 Sufficiency of the evaluated data in view of the study objective

Regarding GARDSkin, it is believed that the quality of both the *in vitro* and *in vivo* reference data were sufficient, adequate and appropriate to draw conclusions regarding the scientific validity of GARDskin when compared to the *in vivo* LLNA reference data.

Regarding GARDpotency however, it is important to take into consideration the discrepancy observed above before making conclusions regarding the scientific validity of GARDpotency when compared to the *in vivo* LLNA reference data.

Finally, it is noted that although reference to the human data is given for the 28 chemicals used in the multilaboratory trial, such information was not initially provided for the 31 additional chemicals tested in the GARDpotency validation report. In a follow-up letter to ESAC, the test developer provided human categories for 19 of the 31 additional chemicals tested (cf. table 8 in Appendix 4 of letter to ESAC as follow-up to the TC from 18 September 2020). Furthermore, the test developer provided an evaluation of the performances of GARDpotency compared to the 'GHS potency according to scientific literature'. However, it is unclear how the 'GHS potency according to scientific literature' was assigned, as it seems to take into account sometimes the LLNA UN GHS category, and other times the human category. It would have been useful, in order to assess the performances of GARDpotency and GARDskin, that separate analyses of the performances of both GARDskin and GARDpotency compared to the human data were conducted.

5. Test definition (Module 1)

5.1 Quality and completeness of the overall test definition

Test Definition

The test definition describes a two-tiered approach for defining: 1) If a compound is likely to be a skin sensitiser (GARDskin), and 2) those who are shown to be so, in which potency class they reside (GARDpotency). The GARDskin test method is defined as an additional method to the suite of existing non-animal methods addressing KE3 (dendritic cell (DC) activation, maturation and migration), to predict the skin sensitising potential of test chemicals and to be used as part of integrated approaches (i.e., IATAs and DAs). It is also suggested that GARDskin can be considered as a stand-alone method for use in Classification and Labelling. It is unclear as to why, as the test is defined as addressing KE3 of skin sensitisation, it could be considered a stand-alone method for Classification and Labelling. Once a chemical has been identified as a sensitiser, the purpose of the GARDpotency method is to categorise the chemical according to CLP/GHS as Cat. 1A or Cat. 1B. When used in tandem with GARDskin, this can be done through additional analysis of stored samples from the GARDskin test method.

The two test method are based on the use of gene expression change "signatures" in SenzaCells, a surrogate for DCs, caused when exposed to a test compound. Despite being described as addressing one of the KE of the current Skin Sensitisation AOP (i.e., KE3), which outlines multiple key events in different cell types (OECD, 2012), the claim of the test developer is that the methods successfully identify skin sensitisers and their potency based solely on this cell-type model. The gene signatures that form the basis for both test methods were identified by a process of elimination from the entire "expressome", based on the well-characterised annotation of a battery of skin sensitisers and non-sensitisers.

The choice of biomarkers for the GARDskin test method (expression levels of 196 genes when using nanoString) was determined statistically following testing of 18 sensitisers and 22 non-sensitisers, rather than by informed selection based on mechanistic relevance to KE3. The biological/mechanistic relevance of these gene changes was studied subsequently and many of the genes have been shown to be of mechanistic relevance to the known biology of skin sensitisation, e.g., oxidative stress, immune responses, DC maturation and cytokine responses. The choice of biomarkers for the GARDpotency test method (expression levels of 51 genes when using nanoString and input concentration) was also determined statistically following testing of 23 1A sensitisers, 25 1B sensitisers and 22 non-sensitisers (Zeller et al., 2017).

It was noted that some of the genes used in the SVM algorithms have a low frequency of selection in the bootstrap exercise (Johansson et al., 2011; Zeller et al., 2017). The ESAC WG felt that there might be a potential to decrease the number of descriptors used in the SVM algorithms (see "Prediction Model" below in this Section). This was discussed with the test developer who felt that the current model was appropriate.

Test System

The biological system used in both test methods is relevant to the proposed use of the methods. It is a cultured human myeloid dendritic-like cell line (SenzaCells – subcloned from MUTZ-3 cells used in original gene expression studies; stocks of these cells are deposited with ATCC, but are available for purchase only through SenzaGen). SenzaCells are used as a surrogate for DCs, in a similar way to other cell lines used in other non-animal methods for addressing KE3 (e.g., THP1 and U937). The metabolic competency of SenzaCells has not been fully characterised, but information on the expression of mRNA transcripts from 68 genes associated with metabolic processes was provided to the ESAC WG by the test developer.

Test Acceptance Criteria / SOPs

Four versions of the SOP were used throughout the validation studies. Following each phase, the SOP was updated for clarity, etc. The procedure for normalisation changed between versions 4.03 (also referred to as version 4.3) and 5.01 of the SOP, with the use of 3 benchmarks instead of the original 12 benchmarks, though the prediction model used throughout remained the same. The versions of the SOP used during the validation studies were:

- SOP 4.03/4.3 Training of naïve laboratories.
- SOP 4.04 Transfer phase.
- SOP 5.01 Blind validation phase for GARDskin / generation of samples for GARDpotency.
- SOP 6.0/6.01 Blind validation phase of GARDpotency.

The Acceptance Criteria were updated in SOP 5.01 and did not change thereafter. SOP v6.00 (also referred to as v. 6.01) is the final protocol for future use.

GARD assay overview

Prior to conducting the main test which is referred to as "GARD main stimulation" in the GARD final protocol (SOPv6.00/6.01), every test chemical is screened for cytotoxic effects. During this step (GARD input finder), cells are exposed to serial dilution concentrations to determine the GARD input concentration to be used in the main test (GARD main stimulation).

The GARD main stimulation is repeated three times, either in parallel or sequentially, using three different batches of cells, to achieve three "replicate samples" as they are named in the SOP. The three replicate samples are used to generate three decision values (DV) for each test chemical in the GDAA. To classify a substance, in the GARDskin the mean DV is used whereas in the GARDpotency the median DV is used.

In the GARD final protocol (SOPv6.00/6.01), it is recommended that the GARD assay is carried out in campaigns, assaying 1-30 test chemicals at a time. Each GARD campaign, in addition to the test chemicals, also includes replicates of the positive and negative benchmark controls and the unstimulated control.

Acceptance Criteria are set for each GARD 'Campaign'. If these criteria are not met, all test chemicals and benchmark controls tested during this 'Campaign' must be retested.

1. All stimulations must pass the phenotypic quality control both to ensure cells are maintained in an inactivated state and to detect phenotypic drift.
2. ≥3 Replicates of unstimulated control samples must pass absolute viability, RNA and nanoString Quality Control criteria.
3. ≥2 Replicates of each positive (PPD) and negative (DMSO) control samples must pass the absolute viability, RNA and nanoString Quality Control criteria.
4. The positive and negative benchmark controls must be accurately classified as 'sensitiser' and 'non-sensitiser' respectively.

Acceptance Criteria are also set for each test chemical. If these criteria are not met, the Test Substance (including benchmark controls) must be retested.

1. The test chemical must be soluble in either DMSO or water.
2. 3 replicates for each test chemical must pass the Relative Viability QC of 90% Relative Viability (84.5-95.4% or 500 µM/maximum solubility if non-cytotoxic). N.B. If, after 5 GARD Main Stimulations, 2 replicate samples have passed, a test chemical is considered to have passed this criterion.
3. ≥2 replicate samples must pass the RNA QC criteria.
4. ≥2 replicate samples must pass the nanoString QC criteria.

The SOP also describes the actions to take if some of the samples fail to pass the QC. Up to 5 Main Stimulations can be conducted to generate the 3 biological samples needed to generate a prediction for a given chemical. If only two biological samples have passed the QC after 5 GARD Main Stimulations, then it is considered acceptable to use these two main stimulations to derive a prediction. The ESAC WG noted that there were two cases (2-Bromo-2-glutaronitrile, CAS 35691-65-7 and Methylisothiazolinone, CAS 2682-20-4) in the validation study where classification of a substance was based on only one biological sample in the Burleson laboratory.

Rate of experiment failure

The rate of experiment failure was not provided in the submission, but the submitted data (excel file Els_SenzaGen ValExp XX GARDpotency CASI DMSO norm.xlsx) contained the experiments performed by the SenzaGen laboratory with indication of the successful runs (referred to as OK in the file) and those that did not pass the quality check, including the reason for the experiment failure, e.g., cell viability too high or too low (>95% or <80%), low RNA concentration and others (See Appendix I to this report). The most common reason for failure was the Relative Viability Quality Check, which accounted for more than half of the failures (average=56%, median=52%). The reasons for failure are variable and dependent on the experiment, therefore it is difficult to describe a trend. The only data that can be easily extracted is that the rate of failure oscillates between 17% and 47% of the samples. The average rate of failure being 27%, and the median 23%.

On the basis of the analysis performed on data from SenzaGen, an experimental failure rate of approximately 25% is expected.

The ESAC WG considers that this protocol and its accompanying Acceptance Criteria are appropriate for the stated purposes.

Prediction Model

GARDskin and GARDpotency prediction models are implemented in the GDAA cloud platform, which consists of an online shinyapps.io application, coded in R. This app uses the gene expression data generated with the nanoString platform for each test chemical to predict its skin sensitisation hazard and/or potency. The prediction step is carried out by applying the pre-trained SVM algorithm on the normalised and standardised quality-checked input data. Therefore, before the application of the SVM algorithm, the app carries out a quality check on the input data to identify invalid runs that may lead to low quality predictions. The samples that pass the quality check steps undergo subsequent normalisation for use in the SVM algorithm. These normalisation steps reduce variability within genes and possible batch-to-batch effects (Forreryd et al., 2016; Gradin et al., 2019). The classifications are obtained using the mean (GARDskin) or median (GARDpotency) DV of biological replicate samples (main stimulations). If the mean or median DV is ≥ 0, the substance is classified as a Sensitiser (GARDskin) or as Cat. 1A (GARDpotency). If the mean or median DV of biological replicate samples is < 0, the substance is classified as a Non-sensitiser (GARDskin) or Cat. 1B (GARDpotency).

The SVM algorithms of GARDskin and GARDpotency were defined using linear kernels. This means that the SVM algorithms consist of a linear combination of the normalised and standardised features (i.e., mRNA transcript counts).

The predictive algorithm has evolved over time, with changes made to the algorithm up to the time of the transferability study. The ESAC WG has successfully reproduced the SVM algorithm and results for the validation study data (see Appendix I to this report). It is important that any future changes to the analysis pipeline are transparent and fully documented, including the SVM algorithm and pre-processing of data.

GARDskin

The initial number of features available to build the SVM algorithm was 29141 gene transcripts obtained from 137 main stimulations performed on 38 chemicals, 20 non-sensitisers and 18 sensitisers using Affimetrix data (Johansson et al., 2011). The set of 29141 gene transcripts was reduced to a subset of 1010 based on the corresponding p-values (p<2E-6) obtained in an ANOVA analysis. Subsequently, these 1010 genes were reduced to 200 by using a "backward feature elimination" strategy based on the Kullback-Leibler divergence. These 200 gene transcript set defined the GPS, which would be used to train the GARDskin model. The nanoString platform only provides 196 genes of the 200 defined with Affimetrix.

The ESAC WG used the training set to generate GARDskin SVM algorithms using 1 up to 196 genes. The results show that SVM algorithms trained with considerably fewer genes show similar performance to the GARDskin SVM algorithm which uses 196 genes (see Appendix II to this report). Even though the current model is considered appropriate for its purpose, the ESAC WG considers it to be overly complex and that it could benefit from simplification. A simpler model would be cheaper to conduct and easier to assess and understand.

GARDpotency

The initial set of chemicals used to train the GARDpotency subcategorisation model (GHS subcategories 1A and 1B) was comprised of the 40 chemicals that were used to train the GARDskin (Johansson et al., 2011), with an additional 48 chemicals, which sum up to a total of 88 chemicals. Of these, 70 were used as training set and 18 as test set. The training set contained chemicals of three different categories; 22 non-sensitisers, 23 Cat. 1A sensitisers (extreme), and 25 Cat. 1B sensitisers (weak). The test set contained 6 chemicals of each category.

The initial number of features available to train the model were 33297 gene transcripts obtained from several main stimulations performed on the 70 chemicals of the training set. The normalised and batch corrected transcript intensities from individual samples of the training set were fed into a random forest model (RF) (Zeller et al., 2017), combined with a backward elimination procedure in the varSelRF R package (Diaz-Uriarte, 2007), which reduced the features down to a total of 52 gene transcripts, which defined the GPPS. The GARDpotency SVM algorithm also consisted of a linear kernel and was trained to differentiate Cat. 1A from Cat. 1B chemicals in Gradin et al. (2020), using nanoString data for 51 chemicals, 51 out of the 52 gene transcripts identified in Zeller et al. (2017) (one of them was not available in nanoString) and the concentration at which the GARD experiments are conducted.

The ESAC WG was able to independently reproduce the results of the GARDskin and GARDpotency validation study (for full details see Appendix I to this report). The ESAC WG evaluated the online tool for data analysis (i.e., the GDAA) and found it to be functional and user-friendly.

5.2 Quality and completeness of the documentation concerning SOPs and prediction models

The final SOP was derived iteratively, including a platform change for the gene expression analysis. In general, the descriptions allow the reader to follow the experimental work and there are a number of good tips and advice to the experimentalist included. However, there are number of areas where "in-house experience" may, to some extent, undermine transfer to other laboratories, without fuller explanations. This is, for example, evident in the manner by which cell batches are characterised for batch-to-batch variation and inclusion or exclusion from use. These criteria are based on FACS analysis and use a variety of markers established to probe purity/homogeneity and viability. Inclusion criteria are based primarily on in-house experience, without any documentation of how the boundaries of the variables were achieved.

Whilst it is stated that the 3 biological samples for each test chemical are from separate main stimulations, the ESAC WG noted that Appendix 3 of the GARD assay v.06.0 'Procedures for failed Relative viability Quality Control' shows that it is possible to include duplicate samples from a single main stimulation if a previous main stimulation has failed the relative viability criteria. If this is correct, the ESAC WG considers that it should be made clear for a test chemical how many main stimulations were used to generate data for a chemical as well as the number of biological samples analysed.

Clarification is needed on the procedure to be used in cases where the molecular weight of a test chemical is unknown, an example should be provided. On page 12 of the SOP (Appendix 6_SOPv.06.01), in the Note box, the following two phrases are reported:

- If the molecular weight is not available, use best available knowledge to approximate molecular weight of the test substance.
- If the purity of the substance is not available, use best available knowledge to approximate the purity of the test substance.

The ESAC WG considers these two phrases to be unacceptable from a scientific point-of-view as it is unclear what is meant by 'best available knowledge'. The test developer has agreed to modify these statements.

The prediction model for GARD potency is described in Appendix 7 of the GARDpotency submission in the file "Amendment to the GARD assay SOP v.06.01: Predicting skin sensitiser potency using the GARD Data Analysis Application". The ESAC WG recommends that this Appendix is incorporated in a new version of the SOP to avoid having multiple documents.

Test Definition: Overall Conclusion

The ESAC WG considers that the Test Definition is, in general, appropriate for the stated purpose and has made some recommendations for future improvement of some of the documentation concerning SOPs and prediction models.

6. Test materials

6.1 Sufficiency of the number of evaluated test items in view of the study objective

GARDskin

The multilaboratory study included 19 sensitisers and 9 non-sensitisers. However, 6 of the sensitisers (formaldehyde, isoeugenol, 2-mercaptobenzothiazole, ethylenediamine, cinnamyl alcohol and eugenol) and 5 of the non-sensitisers (glycerol, isopropanol, lactic acid, propylene glycol and salicylic acid) had been used for training the prediction model. This leaves 13 sensitisers and 4 non-sensitisers as completely new chemicals to evaluate predictive capacity. As stated in the VSR "While the presence of 'training chemicals' among the test chemicals of this study does not affect the value of the determined WLR and BLR values, it affects the credibility of the performance values". The number of additional novel chemicals (13 sensitisers and 4 non-sensitisers) is not sufficient to obtain reliable estimates either for sensitivity, and/or more even so for specificity, as the following calculations show: based on 13 new sensitisers, the length of the 95%-CI (Clopper-Pearson method) for sensitivity can be estimated with a precision of about 47 percentage points if sensitivity is 80%, i.e., it would be expected to range from 49% to 96%. Based on 4 new non-sensitisers, the length of the 95%-CI (Clopper-Pearson method) for specificity can be estimated with a precision of about 77 percentage points if specificity is 80%, i.e., it would be expected to range from 23% to 100% (see Table 2).

Nevertheless, an additional set of 31 coded chemicals were tested within the SenzaGen laboratory only to further assess the predictive capacity of the GARDskin and GARDpotency test methods. These chemicals comprised 11 Cat. 1A chemicals, 12 Cat. 1B chemicals and 8 non-sensitisers, which were used by the ESAC WG to assess the GARDskin's predictive capacity. Only one Cat. 1A (2-nitro-1,4-phenylenediamine), one Cat. 1B (Penicillin G Potassium salt) and three No Cat. (4-aminobenzoic acid, 4-hydroxybenzoic acid and sodium lauryl sulphate) out of these 31 additional chemicals tested at SenzaGen were part of the training set used to build the GARDskin prediction model.

For assessment of WLR and BLR, the overlap of chemicals between training and test phase is a less serious concern. However, even the informative values of WLR and BLR may be affected by reusing chemicals used for training the classifier, as WLR and BLR are based on the model prediction. Model prediction on the chemicals that were used for training might be less ambiguous than model prediction on completely new chemicals.

With 28 chemicals to assess BLR and WLR, the 95%-CI (Clopper-Pearson method) for BLR and WLR can be estimated to have a length of about 32 percentage points if BLR and WLR both are 80%, i.e., they would be expected to range from 61% to 93% (see Table 2). In any case, the additional set of 31 chemicals can also be used to further assess the WLR of GARDskin.

GARDpotency

In the multilaboratory study, 28 chemicals were used, of which 11 Cat. 1A, 8 Cat. 1B and 9 non-sensitisers. However, 8 of the Cat. 1A chemicals (4-nitrobenzylbromide, cinnamal, formaldehyde, lauryl gallate, propyl gallate, isoeugenol, 2-mercaptobenzothiazole and 4-(methylamino)phenol sulphate) and 6 of the Cat. 1B chemicals (diethyl maleate, ethylene diamine, benzyl benzoate, cinnamyl alcohol, citral and eugenol) had been used for training the GARDpotency SVM algorithm and/or defining the GPPS. This only leaves three Cat. 1A and two Cat. 1B chemicals, not used in former experiments, to assess the predictive performance in the ring trial. Reliable estimates for sensitivity and specificity of distinction between Cat. 1A and Cat. 1B cannot be obtained from such small numbers. Using the same setup as above (i.e., assuming true sensitivity and specificity of 80%), the expected range of the 95%-CI (Clopper-Pearson method) for sensitivity (Cat. 1A) would be 16% to 100% while for specificity (Cat. 1B), the range would be 8% to 100% (see Table 2).

For assessment of WLR and BLR, the overlap of chemicals between training and test phase is a less serious concern. However, even WLR and BLR may be affected by using training chemicals, as WLR and BLR are based on the model prediction. Model prediction on the chemicals that were used for training might be less ambiguous than model prediction on completely new chemicals.

In addition to the 28 chemical used in the ring trial, 31 chemicals were selected and evaluated by SenzaGen alone, 11 Cat. 1A, 12 Cat.1B and 8 non-sensitisers. This additional set of chemicals can be used to further assess WLR as well as the predictive capacity in SenzaGen. However, of the Cat. 1A, two chemicals had been used to train the classifier and define the GPSS (2-nitro-1,4-phenylendiamine and p-benzoquinone). Moreover, no data were generated with GARDpotency for one of the Cat. 1A (1,2-cyclohexane dicarboxylic anhydride) and one of the Cat. 1B (benzyl salicylate) chemicals, because they were identified as non-sensitisers by the GARDskin in all the three runs. Therefore, they could not be used to assess predictive capacity of GARDpotency.

Taking the two chemical lists together (multilaboratory study list and extra chemicals list) for SenzaGen, but subtracting the chemicals used for training the GARDpotency SVM algorithm and defining the GPPS, 3+8=11 Cat. 1A and 2+11=13 Cat. 1B chemicals are available for assessing the accuracy of GARDpotency for predicting Cat. 1A (=sensitivity) and Cat. 1B (=specificity). For evaluation of the combined GARDskin and GARDpotency approach, 25 sensitisers and 6 non-sensitisers not used to train the prediction models of either method are available (from the two chemical lists together). For different numbers of chemicals, the length of the 95%-CI (Clopper-Pearson method) and its expected range, given the true value of 80% for correct classification, is summarised in Table 2.

Table 2: Expected width, lower and upper 95%-CI limits for proportions, given that the true proportion is 80%, for number of chemicals used for GARDskin and GARDpotency. Calculations performed with PASS software, version 15 (NCSS, 2017).

Number of chemicals	**Width of 95%-CI (percentage points)**	**Lower CI limit**	**Upper CI limit**
2	92	8	100
3	84	16	100
4	77	23	100
5	71	28	100
6	66	33	99
11	51	46	97
12	49	48	97
13	47	49	96
14	45	51	96
25	34	59	93
28	32	61	93

6.2 Representativeness of the test items with respect to applicability

For the 28 chemicals used in the multilaboratory trial, all three UN GHS categories were represented (10 Cat. 1A, 8 Cat. 1B and 9 No Cat. when taking into account the considerations stated in Section 4.2). Furthermore, these 28 chemicals represent a variety of LLNA and human potencies, as well as Toxtree classes (cf. Appendix 4 of the test submission).

Regarding the 31 additional chemicals used to assess GARDskin and GARDpotency, here again a good representation of the three UN GHS categories was found (i.e., 11 Cat. 1A, 12 Cat. 1B and 8 No Cat.). However, less detailed information was given regarding the LLNA and human potencies of these additional chemicals. Furthermore, no information was provided on their respective Toxtree classes.

7. Within-laboratory reproducibility (WLR) (Module 2)

7.1 Assessment of repeatability and reproducibility in the same laboratory

Based on what is reported in the SOP, once the GARD Input concentration for each test chemical is established, a GARD Main Stimulation (run) should be repeated three times, using three cell batches, on all test chemicals and benchmark controls to achieve three biological replicate samples. In this document, the ESAC WG uses the term "run" as defined in the SOP, so that a run is a single determination or biological replicate and an "experiment" consists of at least 3, and no more than 5, independent runs. Additional runs (up to 5) were allowed if invalid results were observed (see Acceptance Criteria listed in SOPv.06.01). In general, concordance should be established on the basis of a minimum of three independent experiments. For those chemicals for which only two valid experiments were available, if these were concordant, the test developer considered the conclusion for the chemicals to be concordant for the calculation of the WLR (best case scenario). However, should a third experiment have been conducted, it is possible that this may or may not have been concordant with the previous two. Thus, in addition to the values presented by the test developer, the ESAC WG also calculated a worst case scenario assuming that a third experiment would have been discordant.

GARDskin

WLR was assessed on the entire set of 28 chemicals (19 sensitisers and 9 non-sensitisers), using only valid data. Three independent experiments were conducted in each laboratory. The target was set at 80% concordance of the prediction, a target which is commonly used in these types of validation studies.

GARDpotency

WLR was performed using the mRNA samples produced by the GARDskin validation study. As only chemicals predicted by the GARDskin to be sensitisers were analysed with GARDpotency, the number of chemicals used for WLR was lower compared to the one used for the GARDskin. This lower number of chemicals used to assess the GARDpotency WLR was not adequately justified. Only data from valid experiments were considered for statistical analysis, while failed runs and experiments were documented to report their occurrence. WLR assessment was based on the concordance of the individual predictions, as Cat. 1A or cat.1B sensitiser, which were determined using a computer-based prediction algorithm. Three independent experiments were conducted in each laboratory. The target was set at 75% for concordance of the prediction. It is not clear why, in this case, concordance was set at 75% and not 80% as for GARDskin, as both methods use binary classification (S/NS or Cat. 1A/Cat. 1B).

7.2 Conclusion on within-laboratory reproducibility as assessed by the study

WLR of GARDskin

The observed WLR based on experiments considered to have passed the acceptance criteria, as summarised on page 4 of the VSR on the GARDskin, ranged from 82.1% to 89.2%, which is considered acceptable (target at least 80%). The ESAC WG noted that slightly different percentages were achieved when considering, in a consistent manner, either the entire set of 28 chemicals (78.6-89.2%) or the dataset, where runs not complying with cell viability acceptance criteria are excluded (82.1-88.9%). However, the ESAC WG considers that, in either case, this level of WLR is acceptable.

Additional details:

- At BRT, a consistent prediction was obtained for twenty of the twenty-four chemicals that were considered to have passed the acceptance criteria in the three independent experiments, resulting in a WLR of 83.3%. Chemicals excluded were 2-bromo-2-glutaronitrile, 4-(methylamino)phenol sulphate, citral and dextran. Considering all 28 chemicals, the overall WLR was 78.6% (22/28, 95%-CI: 59.0-91.7%).
- At Eurofins BioPharma, a consistent prediction was obtained for twenty-four out of the twenty-seven chemicals that were considered to have passed the acceptance criteria in the three independent experiments, resulting in a WLR of 88.9%. The only chemical excluded was dextran. Considering all 28 chemicals, the overall WLR was 89.2% (25/28, 95%-CI: 71.8-97.7%).
- At SenzaGen AB, a consistent prediction was obtained for twenty-three out of the twenty-eight chemicals that were considered to have passed the acceptance criteria in the three independent experiments, resulting in a WLR of 82.1% (23/28, 95%-CI: 63.1-93.9%). No chemicals were excluded.

WLR of GARDpotency

At BRT, of the twenty-eight chemicals, sixteen were predicted by the GARDskin to be a sensitiser, based on at least two out of three experiments being concordant. Two sensitisers produced invalid results: 2-Bromo-2-glutaronitrile (#2) and citral (#17). Ethylene diamine (#12) was wrongly classified as non-sensitiser. For these chemicals, the acquired mRNA samples were therefore not considered for further processing. No non-sensitiser was wrongly identified as being a sensitiser. A consistent prediction was obtained for ten of the sixteen sensitising chemicals that were considered to have passed the acceptance criteria in the three independent experiments, resulting in a WLR of 62.5% (10/16, 95%-CI: 35.3-84.8%), which is below the target of 75%. However, the ESAC WG noted that in the GARDpotency WLR – BRT table, for two chemicals (#6 4-methylamino phenol sulfate and #15 benzyl benzoate), the final prediction was derived on the basis of only two valid experiments and calculated a worst case scenario assuming that a third experiment would have been discordant, resulting in a WLR of 50.0% (8/16, 95%-CI: 24.7-75.3%).

At Eurofins BioPharma, of the twenty-eight chemicals, eighteen were predicted by the GARDskin to be a sensitiser based on at least two out of three experiments being concordant. Ethylene diamine (#12) was wrongly identified as non-sensitiser. No non-sensitisers were wrongly identified as sensitisers. A consistent prediction was obtained for fifteen of the eighteen identified sensitisers that were considered to have passed the acceptance criteria in the three independent experiments, resulting in a WLR of 83.3% (15/18, 95%-CI: 58.6-96.4%). However, the ESAC WG noted that in the GARDpotency WLR – Eurofins table, for one chemical (#15 benzyl benzoate), the final prediction was derived on the basis of only two valid experiments and calculated a worst case scenario assuming that a third experiment would have been discordant, resulting in a WLR of 77.8% (14/18, 95%-CI: 52.4-93.6%).

At SenzaGen, of the 28 chemicals, 18 were predicted by the GARDskin to be a sensitiser. Ethylene diamine (#12) and benzyl benzoate (#15) were wrongly identified as non-sensitisers and vanillin was wrongly identified as a sensitiser. A consistent prediction was obtained for sixteen of the eighteen sensitising chemicals with a WLR of 88.9% (16/18, 95%-CI: 65.3-98.6%). However, the ESAC WG noted that in the GARDpotency WLR – SenzaGen table, for two chemicals (#14 2-mercaptobezothiazole and #28 vanillin), the final prediction was derived on the basis of only two valid experiments and calculated a worst case scenario assuming that a third experiment would have been discordant, resulting in a WLR of 77.8% (14/18, 95%-CI: 52.4-93.6%).

The WLRs obtained at Eurofins and SenzaGen met the acceptance criteria (83.3% and 88.9%, respectively) (Tables 9, 11 of the GARDpotency VSR). However, for BRT, the reported WLR was 62.5%, which is considered not acceptable (*a priori* target at least 75%). This reported WLR excluded data from invalid results. If all data were included, the WLR would be even lower.

Conclusions reported by the VMG that the overall WLR met the target (at least 75%), under the assumption that that the first experiment by BRT was negatively affected by the technical issues at Covance, is questionable. To calculate WLR, three independent experiments are needed, which makes the data obtained by BRT incomplete (only two data sets were available due to technical issues described above) and not usable to evaluate the WLR and, therefore, the whole study is compromised. The test developer speculated that, since pairwise WLR was around 75% for the two experiments, a WLR of 75% for three experiments "must be feasible". The ESAC WG does not agree with the statistical reasoning of this argument. WLR among three experiments, as it is commonly performed, is a more stringent requirement than WLR in two experiments as was performed here. Thus, the ESAC WG does not consider this hypothetical argument sufficient to demonstrate the WLR at BRT, because of the problem with the first experiment. An additional experiment should have been run to acquire the data needed to reach an overall conclusion on WLR.

Furthermore, the ESAC WG notes that, in some cases, the data sets were incomplete in each of the laboratories, due to negative results of chemicals in the GARDskin assay. Because GARDpotency was only run on samples that tested positive in the GARDskin assay, some samples were not tested in the GARDpotency WLR studies. These represent 3/16 chemicals at BRT, 1/18 at Eurofins, and 2/18 at SenzaGen.

The ESAC WG believes that the WLR of GARDpotency has not been appropriately characterised for the following reasons:

- ***Three runs were not completed to assess WLR due to a problem with experiment 1 in one of the three participating laboratories. Therefore, the entire study was compromised. An additional experiment should have been conducted to acquire the full dataset needed to reach a more robust conclusion on WLR.***
- ***The target was set at 75% by the VMG for both WLR and BLR, as opposed to 80% in GARDskin. However, insufficient justification for this reduction in the target was provided. WLR was 62.5%, 83.3%, 88.9% for best case scenario, and 50%, 77.8% and 77.8% for the worst case scenario, being below the target for at least one of the laboratories. Therefore, the WLR is not considered sufficient at this time.***

8. Transferability (Module 3)

8.1 Quality of design and analysis of the transfer phase

GARDskin

The training of the laboratory's personnel and the transferability of the GARDskin assay were performed in accordance with OECD GD N°34 (OECD, 2005), under the supervision of the VMG. The assay was transferred to two naïve laboratories (BRT and Eurofins). Prior to initiation of the Transferability Study, a written document describing the study plan and criteria for successful transfer was developed (Appendix 2 – GARDskin Validation Study Plan). Quality criteria for personnel, equipment and facilities were defined *a priori.* SenzaGen provided an SOP (Appendix 14 - SOPv04.03) to BRT and Eurofins and performed in-person training with both laboratories. Training focused on the biological system (i.e., cells handling and stimulation), RNA isolation, nanoString data analysis (using data produced by the lead laboratories on the five chemicals) and generation of predictions. Cells, reagents and chemicals were either provided by SenzaGen or obtained from providers recommended by SenzaGen. The training session led to minor adjustments of the protocol, included in SOPv04.04. During the transfer phase, additional comments were raised by the participating laboratories resulting in a revised SOP (SOPv5.01). Revisions included:

- A minor clarification of the text was necessary after the first Transfer experiment at BRT, resulting in v04.04, which was distributed to both external laboratories and was used throughout the remainder of the Transfer study. Further clarification of the text and addition of descriptive figures were added based on the Transfer study.
- Phenotypic Quality Control - Historical data were originally used to set the acceptance range for each phenotypic biomarker with a BD FACSCanto flow cytometer. When data from the external laboratories showed that the expressions of some markers were outside the specified range, this was believed to be due to the difference in flow cytometer instruments and their individual settings. It was decided that, for all but PI, CD86 and CD80 (markers for SenzaCell viability and activation), a more general acceptance criterion of a positive population should be applied and the specified ranges were removed.
- "Cells" population - Again, due to differences in equipment, the specified range was removed and exchanged with a general notification of paying attention to the movement of the cell population.
- Change of positive control - 2-hydroxyacrylate was replaced with p-phenylenediamine due to issues with variability and chemical instability.
- Reduced number of benchmark controls - A 3-benchmark normalisation method (the Updated solution) rather than the originally prescribed 12-benchmark normalisation method due to the increased efficiency of protocols and improved results (Appendix 11_GARD Results Report Transfer Eurofins).

Successful transfer was defined *a priori*, the laboratories were asked to correctly classify five chemicals as sensitisers or non-sensitisers, in 3 consecutive valid experiments, with 100% agreement and correct predictions (see Appendix 11_GARD Results Report Transfer Eurofins and Appendix 10_GARD Results Report Transfer BRT).

The ESAC WG noted that, in the summary tables included in the GARDskin VSR (i.e., Table 7 and Table 9 -*12-benchmark calibration method (SOPv04.4)),* in both naïve laboratories, resorcinol was classified as NS in at least one 'Transfer' experiment, thus, not reaching the criterion for success ('three consecutive valid experiments with 100% concordance and correct prediction'). The reduction of the number of benchmark chemicals from 12 to 3 improved the test performance, allowed the naïve laboratories to meet the requirement of 'three consecutive valid experiments with 100% concordance and correct prediction'.

GARDpotency

The transferability of the GARDpotency was considered being demonstrated by the transferability of the GARDskin method. Training was provided in the context of the GARDskin validation study. The same protocols were used during the GARDpotency validation study.

8.2 Conclusion on transferability to a naïve laboratory / naïve laboratories as assessed by the study

The ESAC WG considers that the study design for transfer and implementation was acceptable. While changes were made to acceptance criteria and the SOP during the transfer study, these changes were well documented and scientifically justifiable, were necessary to improve the transferability of the assay to naïve laboratories and were not detrimental to assay reliability.

9. Between-laboratory reproducibility (BLR) (Module 4)

9.1 Assessment of reproducibility in different laboratories

GARDskin

The validation study was performed in compliance with OECD GD N°34 (OECD, 2005). BLR was assessed on the entire set of 28 chemicals, using only valid data (a maximum of five runs were allowed when cell viability issues occurred). Three independent experiments were conducted in each laboratory.

The ESAC WG considers that the method for assessing GARDskin BLR was acceptable.

GARDpotency

The GARDpotency was performed as a second tier to subcategorise chemicals that were identified as sensitisers by the GARDskin method, into the GHS/CLP Cat. 1A or Cat. 1B (i.e., strong and other sensitisers, respectively). The study was performed using the mRNA samples produced by the GARDskin validation study. Only data from the valid experiments were considered for statistical analysis, while failed runs and experiments were documented to report their occurrence. The BLR was assessed based on the data for GARDskin positive chemicals (among the 28 chemicals). The focus of the evaluation of the BLR was 'majority of predictions' (at least two out of three) as Cat. 1A or Cat. 1B sensitiser.

The maximum number of chemicals with a full data set (i.e., three valid results per laboratory) for which potency could be compared across all three laboratories is 12, after classification by GARDskin. Moreover, a prediction could be derived from all three laboratories for 14 chemicals, even if based only on two concordant experiments in a laboratory (instead of three). The ESAC WG calculated the estimates of BLR and their 95%-CI based on these various results.

9.2 Conclusion on between-laboratory reproducibility as assessed by the study

BLR of GARDskin

For the purpose of calculating BLR, an overall (majority) conclusion on class (sensitiser/non-sensitiser) was drawn from concordance of at least 2 out of 3 independent experiments, in each laboratory. For example, if two experiments resulted in 'sensitiser' and one resulted in 'non-sensitiser', the chemical would be considered a sensitiser.

The observed BLR for all three laboratories was either 82.1%, considering all 28 chemicals (23/28, 95%-CI: 63.1-93.9%), or 92.0% (23/25, 95%-CI: 74.0-99.0%), excluding chemicals with technical issues (# 2 2-bromo-2-glutaronitrile, #6 4-(methylamino)phenol sulphate, #17 citral and #20 dextran). In both cases, these values for BLR are considered acceptable (i.e., above the *a priori* target of 80%).

BLR of GARDpotency

For the purpose of calculating BLR, an overall (majority) conclusion on category (1A/1B) was drawn from concordance of at least 2 out of 3 independent experiments in each laboratory. For example, if two experiments resulted in 'Cat. 1A' and one resulted in 'Cat. 1B', the chemical would be considered Cat. 1A in that laboratory (BRT generated two inconsistent results for one chemical and a majority prediction could not be derived; therefore, BLR was considered not concordant for this chemical). Taking this approach into account, the overall BLR among the three laboratories was 61.1% (11/18) for the 18 chemicals with valid results in at least two laboratories (reported on page 37 of the VSR on the GARDpotency), and with a 95%-CI of 35.7-82.7% as calculated by the ESAC WG. This BLR was below the target (at least 75%) that was set before the start of the study, therefore not

meeting the required criterion. The ESAC WG also notes that the maximum number of chemicals having a full data set (i.e., three valid results per laboratory) for which potency could be compared across all three laboratories is 12. The BLR of these 12 chemicals was calculated by the ESAC WG to be of 75% (9/12, 95%-CI: 42.8-94.5%). However, as stated in Section 7, insufficient justification was given for having the target at 75% by the VMG for both WLR and BLR, as opposed to 80% in GARDskin. Furthermore, when considering the 14 chemicals for which a prediction could be derived from all three laboratories, even if based only on two concordant experiments in a laboratory (instead of three), a BLR of 71.4% (10/14, 95%-CI: 41.9-91.6%) was obtained, which is again below the 75% target value.

An additional analysis on BLR (and WLR) was provided as complimentary material by the test developer for identifying UN GHS Cat. 1A vs the rest, whereby the negative results from GARDskin were considered as Cat. 1B (Table 6 of Appendix 4 in the follow-up comments from the method developer following the ESAC WG meeting of Sept 18, 2020). It is unclear whether the Cat. 1B was an assumption or a result of experimental testing. Because of this, the ESAC WG could not take into account this additional information when evaluating the BLR of the GARDpotency.

BLR of Combined GARDskin and GARDpotency

The test developer provided additional information on the overall BLR of the combined GARDskin and GARDpotency, which showed 66.7% concordance (18/27 chemicals) with the three categories (non-sensitiser, Cat. 1A and Cat. 1B) (Table 5 of Appendix 4). The ESAC WG calculated that the 95%-CI for this BLR ranges from 46.0% to 83.5%. However, there were neither predefined acceptance criteria for the BLR nor justification for the numbers of chemicals assessed for the combined approach. Furthermore, the BLR of GARDpotency fell below its own acceptance criterion and there were only a small number of chemicals with complete data sets. The impact of these issues on the overall BLR of the combined strategy are unclear. The ESAC WG believes that the issues with the GARDpotency BLR must be addressed before a combined approach can be characterised in terms of BLR.

10. Predictive capacity and overall relevance (Module 5)

10.1 Adequacy of the assessment of the predictive capacity in view of the purpose

The predictive capacities of the test methods were assessed by the accurate prediction of the blinded test compounds sent to each test site in the multilaboratory study and/or an extra set of coded compounds tested by the lead laboratory only. In the VSRs, the data are presented both in terms of the results from individual test sites and as a pooled value obtained from all three sites in the multilaboratory study and/or an extra set of coded compounds tested by the lead laboratory only. Predictive Capacity is assessed in terms of accuracy, sensitivity and specificity of the predictions.

The modular approach to validation (Hartung et al., 2004) and the OECD GD on the Validation of (Q)SAR Models (OECD, 2007), state the importance that the performances of the substances used to train prediction models are clearly described and separate from the performances of the substances used to test the models. The ESAC WG considers that, since a machine-learning model (i.e., SVM) is used for both GARDskin and GARDpotency, it is even more important that this principle is strictly adhered to, due to the risk of overfitting when using SVM/machine learning models.

GARDskin

For the GARDskin assay, in the submission, the predictive capacity was evaluated for each individual laboratory using the majority voting for each chemical within the predictions of the 3 experiments. For instance, if a chemical was classified in SenzaGen Exp1, SenzaGen Exp2 and Senzagen Exp3 as: S, S, and NS; the prediction considered for the calculation was S. The accuracy, sensitivity and specificity calculated in the submission, included results from chemicals that had been used to train the prediction model. Accuracy for the 3 laboratories ranged from 89.3% to 96.3%, sensitivity ranged from 89.5% to 94.7% and specificity from 88.9% to 100%, for the chemical sets used (from 17-19 sensitisers and from 8-9 non-sensitisers). In addition, the test developer accumulated the data over the three laboratories to report a cumulative performance of 93.8% accuracy, 92.7% sensitivity and 96.0% specificity, where a total of 84 predictions, representing the 28 chemicals (19 S and 9 NS) tested in 3 laboratories, were considered. In addition to the data generated in the multilaboratory study, another set of 31 coded chemicals (23 S and 8 NS) were tested within the SenzaGen laboratory only to further assess the predictive capacity of the method. The GARDskin predicted this set of chemicals with an accuracy of 77.4%, a sensitivity of 78.3% (18 of 23) and a specificity of 75.0% (6 of 8).

In addition to the data generated in the validation study, the GARDskin submission also mentioned data generated on a further 25 sensitisers and 16 non-sensitisers. This was used as a testing set to confirm the GARDskin prediction model and was based on a protocol corresponding to the protocol agreed after the training phase and further modified after the transferability phase. The performance values associated with this data set were reported to be 97.6% accuracy, 100% sensitivity and 93.3% specificity (Forreryd et al., 2016). There were also references to other data generated by the test developer, but these are of less relevance as they were generated using Affymetrix analysis and not by the nanoString platform (Johansson et al., 2014). According to the test developer, due to the use of 'difficult to test' chemicals, the predictions in Johansson et al. (2014) were somewhat lower when compared to the multilaboratory trial dataset (i.e., 89% vs 94% accuracy, 89% vs 93% sensitivity and 88% vs 96% specificity for Johansson et al. (2014) and the multilaboratory trial, respectively). Further data generated using the nanoString platform were produced in a blind Cosmetics Europe study. The performance values associated with this data set were reported to be 76% accuracy, 90% sensitivity and 45% specificity when compared to LLNA results, or 83% accuracy, 93% sensitivity and 56% specificity when compared to a composite of LLNA and human reference data (Johansson et al., 2017).

ESAC WG Analysis of Predictive Capacity

The ESAC WG noted that, amongst the 28 chemicals tested in the multilaboratory study to derive accuracy, sensitivity and specificity of GARDskin, a number of chemicals (6 sensitisers, 5 non-sensitisers) were also present in the list of chemicals used to build the GARDskin prediction model (see Sections 1.3 and 6). Similarly, the performance values of the additional studies discussed in the VSR were also obtained for datasets which included substances that had been used to train the SVM algorithm. In particular, 2 sensitisers and 3 non-sensitisers out of the 31 extra coded chemicals tested by the SenzGen laboratory only were also part of the training set used to build the GARDskin prediction model (see Sections 1.3 and 6). The ESAC WG, therefore, considered it important to calculate also the performance values of GARDskin using only those chemicals that had not been used to train the prediction model.

Furthermore, when using the GARDSkin for testing new chemicals, a single experiment would be used for prediction. The ESAC WG considers therefore that the evaluation of predictive capacity should be performed in this setting (i.e., considering all individual predictions obtained in the validation study to calculate predictive capacity instead of applying the majority voting to the multiple predictions obtained by each laboratory).

The ESAC WG re-calculated the performance values (accuracy, sensitivity, specificity) of GARDskin based on the above considerations, and taking into account their respective 95%-CIs, as shown in Figure 1 (calculated by Clopper-Pearson method with the R Package DescTools and function binomCI). The predictive capacity was calculated using results from all chemicals tested in the validation study ('all chemicals') and also calculated using only those chemicals that had not been used to train the prediction model ('test chemicals'). Performance values using only the test chemicals were: Sensitivity ranged from 76% to 100% (with the width of the 95%-CIs between 22 and 45 percentage points and all lower Confidence limits above 50%); Specificity ranged from 40% to 100% (with the width of the 95%-CIs at least 63 and up to 80 percentage points); Accuracy ranged from 73% to 100% (with the width of the 95%-CIs at least 21 and up to 36 percentage points and all lower Confidence limits above 50%).

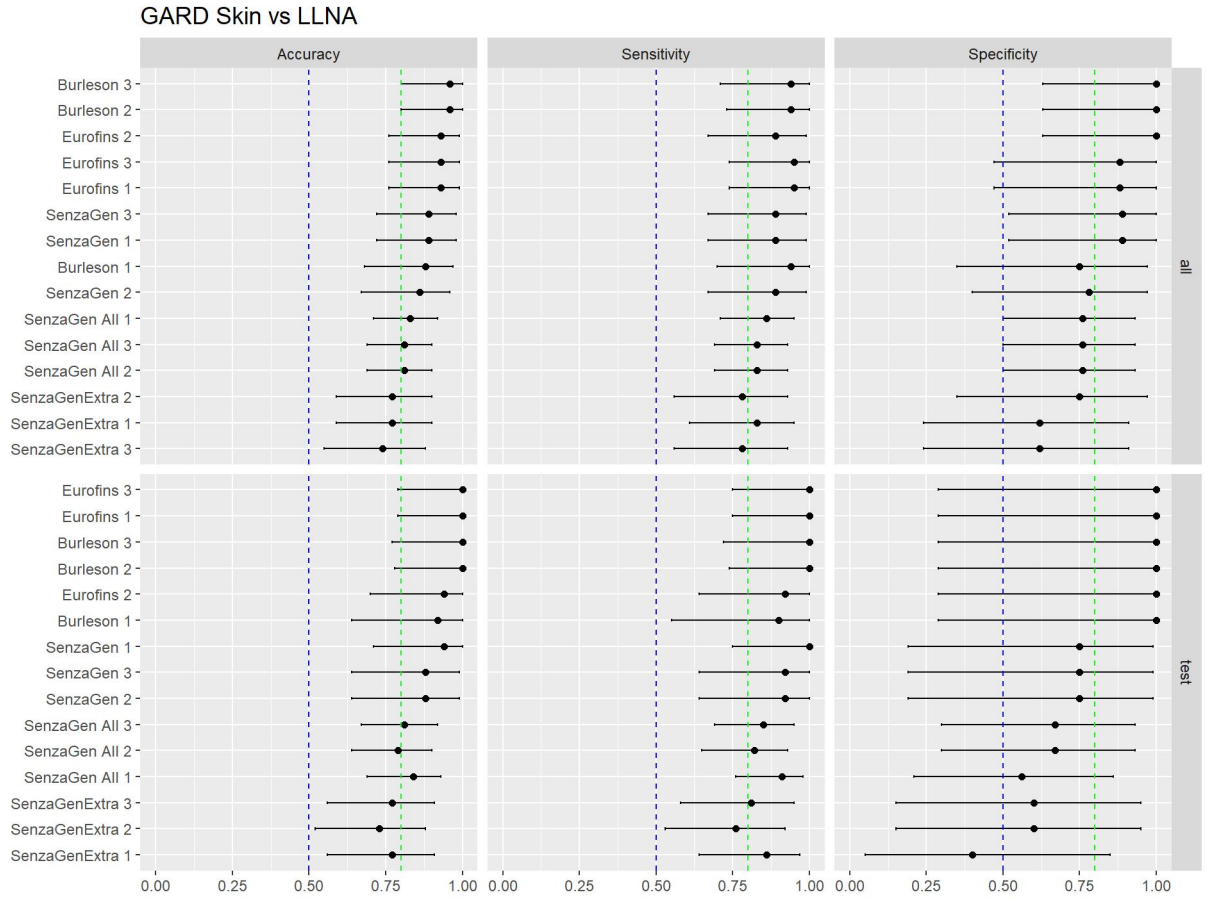


Figure 1: Predictive performance of the GARDskin algorithm for all chemicals (top) and for test chemicals, i.e., excluding chemicals used for training of the GARDskin method (bottom). Each calculation was performed for a single experiment. The dotted blue vertical line highlights 50%, i.e., blind random classification into two categories. The dotted green vertical line shows the value of 80%, i.e., a desirable value for correct classification. The order of the experiments is by Youden index = sensitivity + specificity -1, a summary value for predictive capacity that is independent of class prevalence. The confidence intervals show the uncertainty of the estimates and give the full range of values in accordance with observed data. For specificity, the confidence intervals in many settings cover 0.5, specifically when excluding the training chemicals, i.e., specificity is not significantly better than random class allocation. "SenzaGen Extra" denotes the additional chemicals set (up to 31 chemicals) that were tested only in the SenzaGen laboratory, "SenzaGen All" denotes all chemicals, ring trial + additional chemicals.

In addition to the calculation of performance values described above, and in order to summarise the performance across laboratories and experiments of the GARDskin, the ESAC WG also calculated accuracy, sensitivity and specificity of GARDskin using a weighted calculation. In this weighted calculation each individual prediction (from the three laboratories participating in the validation study) was captured as an independent prediction in the calculations but the same weight was given to each chemical independently of the number of predictions available for each chemical. In summary, the "NS", "S", "Cat. 1B" or "Cat. 1A" predictions for each chemical (obtained by the three laboratories participating in the study) was divided by the total number of available predictions to determine the number of correct and under- or over-predictions for that chemical (as fractions of 1) and these were then used to calculate accuracy, sensitivity and specificity so that all chemicals contribute with an equal weight of 1 in the calculations. This approach using weighted calculation is applied by the OECD for the development of Test Guidelines (OECD, 2015, 2017). No confidence intervals have been derived for this weighted calculation. As above, these calculations were performed using 'all chemicals' and 'test chemicals' and in addition also show the results of the training chemicals for comparative purposes. These results are shown in Table 3.

Table 3: Weighted calculation of accuracy, sensitivity and specificity of GARDskin[a].

	Multilaboratory ring trial			SenzaGen Extra			Ring trial + SenzaGen Extra		
	All Chemicals (n=28)	Training Set Chemicals (n=11)	Test Set Chemicals (n=17)	All Chemicals (n=31)	Training Set Chemicals (n=5)	Test Set Chemicals (n=26)	All Chemicals (n=59)	Training Set Chemicals (n=16)	Test Set Chemicals (n=43)
TP	*17.6*	*5.0*	*12.6*	*18.3*	*1.3*	*17.0*	*35.9*	*6.3*	*29.6*
TN	*8.1*	*4.4*	*2.9*	*5.3*	*2.7*	*2.7*	*13.4*	*7.1*	*6.3*
FP	*0.9*	*0.6*	*0.3*	*2.7*	*0.3*	*2.3*	*3.6*	*0.9*	*2.7*
FN	*1.4*	*1.0*	*0.4*	*4.7*	*0.7*	*4.0*	*6.1*	*1.7*	*4.4*
Accuracy	92%	86%	95%	76%	80%	76%	84%	84%	83%
Sensitivity	92%	83%	97%	80%	67%	81%	85%	79%	87%
Specificity	90%	89%	90%	67%	89%	53%	79%	89%	70%

[a] The first part of the table (numbers in *italics*) corresponds to the weighted number of chemicals. Since a weighted calculation is used to derive the values of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), these are not integers. The values in the second part of the table (accuracy, sensitivity and specificity) are reported as percentages. For the data produced by SenzaGen, a number of analyses were conducted, reflecting that several datasets were available: "SenzaGenExtra" denotes the chemicals of the additional chemicals set (up to 31 chemicals) that were tested only in the SenzaGen laboratory.

Based on all analyses described above, the ESAC WG concluded that the GARDskin assay has a performance that is comparable to in vitro/in chemico methods adopted to support skin sensitisation hazard identification. Nonetheless, it would have been useful to have included a larger number of chemicals (especially non-sensitisers), which had not been used to train the model, to increase the precision of the estimates.

The ESAC WG concluded that the predictive capacity of GARDskin is therefore appropriate and in line with other in vitro tests for prediction of skin sensitisation hazard to be used in combination with other assays.

GARDpotency

For the GARDpotency assay, in the submission, the predictive capacity was determined by the validity of assignment of a test chemical to a particular UN GHS Category of sensitiser, i.e., Cat. 1A or Cat. 1B. This was performed on the chemicals predicted as sensitisers by GARDskin. The predictive capacity was evaluated for each individual laboratory using the majority voting for each chemical within the predictions of the 3 experiments. For instance, if a chemical was classified in SenzaGen Exp1, SenzaGen Exp2, and Senzagen Exp3 as: Cat. 1A, Cat. 1A, and Cat. 1B; the prediction considered for the calculation was Cat. 1A. The accuracy obtained in the 3 laboratories ranged from 76.5% to 94.4%, with 100% of correct Cat. 1A classification in all 3 laboratories and Cat. 1B correct classifications ranging from 42.9% to 87.5%, for the chemicals having GARDpotency data (8-10 Cat. 1A and 7-8 Cat. 1B). The test developer accumulated the data over the three laboratories, to report a cumulative performance of 88.0% accuracy, 100% correctly predicted Cat. 1A and 72.7% correctly predicted Cat. 1B, where a total of 50 predictions were considered (28 Cat. 1A and 22 Cat. 1B predictions, representing 8-10 Cat. 1A and 7-8 Cat. 1B chemicals tested in 3 laboratories).

In addition, the submission reported the results obtained for an extra set of 31 coded chemicals (11 Cat. 1A, 12 Cat. 1B and 8 No Cat.) tested in a blind manner by the lead laboratory only. Out of the 23 sensitisers in this set of chemicals, five were not considered by the test developer for the calculation of the GARDpotency predictive capacity as they were identified as non-sensitisers in GARDskin based on the majority vote of three runs. The resulting 18 sensitisers were predicted with an accuracy of 66.7%, with 62.5% correct prediction for Cat. 1A (5 of 8) and 70% correct prediction for Cat. 1B (7 of 10).

ESAC WG Analysis of Predictive Capacity

The design of the validation study for the GARDpotency assay was dependent on the results obtained from GARDskin, which led to data gaps in the matrix of results generated within the study (see Section 1.3). In an ideal study design, all the chemicals (sensitisers and non-sensitisers) would have been tested in GARDpotency, to be able to both determine the performance of GARDpotency for sensitisers, as well as to determine the effects of testing non-sensitisers in GARDpotency. However, due to the study design, not all tested sensitisers had a complete dataset available, i.e., 3 replicates per experiment and per laboratory, and some sensitisers had no GARDpotency data. When evaluating the performance of GARDpotency, these chemicals may be considered as "no data", but it is possible that if these chemicals had been tested in GARDpotency they could have resulted in either correct predictions, or incorrect predictions. The ESAC WG calculated the performance of GARDpotency not based on the majority of predictions as suggested in the validation report but using only the sensitisers that had been tested in GARDpotency and taking into account the confidence interval and weighted predictive values as described below. This calculation was supplemented with two additional analyses, one considering the missing data as correct predictions (best case), and another one considering the missing data as incorrect predictions (worst case). Finally, based on the study design used, only a few non-sensitisers (those that were false positive in GARDskin) were processed with GARDpotency. Therefore, it is impossible to assess the level of overprediction with non-sensitisers when using GARDpotency with other first tier methods instead of GARDskin.

The ESAC WG noted that, amongst the chemicals tested to derive the accuracy, sensitivity and specificity of GARDpotency, a large number (8 out of 11 Cat. 1A, and 6 out of 8 Cat. 1B) were also present in the chemical list used to train the GARDpotency SVM algorithm and/or defining the GPPS. Therefore, in the validation study, there were only three Cat. 1A and two Cat. 1B that had never been seen by the model. In the extra set of 31 chemicals tested in a blind manner by the lead laboratory only, two Cat. 1A had been used to fit the GARDpotency SVM algorithm or define the GPPS. The ESAC WG therefore additionally calculated the performance values of GARDpotency using only those chemicals that had not been used to train the prediction model both in the multilaboratory ring trial and in the lead laboratory.

The ESAC WG noted that, when using the GARDpotency for testing new chemicals, a single experiment would be used for prediction. The ESAC WG considers therefore that the evaluation of predictive capacity should be performed in this setting (i.e., considering all individual predictions obtained in the validation study to calculate predictive capacity instead of applying the majority voting to the multiple predictions obtained by each laboratory). The ESAC WG therefore calculated accuracy, Correctly Predicted Cat. 1A rate (defined below as CP1Arate) and Correctly Predicted Cat. 1B rate (defined below as CP1Brate) for single experiments in all the laboratories. The predictive capacity was calculated using results from all chemicals tested in the validation study ('all chemicals') and also calculated using only those from chemicals that had not been used to train the prediction model ('test chemicals'). These calculations are shown in Figure 2. Performance values for GARDpotency using only the test chemicals were: CP1Arate ranged from 0% to 100% (with the width of the 95%-CIs between 53 and 97 percentage points and all lower Confidence limits below 50%); CP1Brate ranged from 50% to 100% (with the width of the 95%-CIs at least 52 and up to 98 percentage points and all lower Confidence limits below 50%).

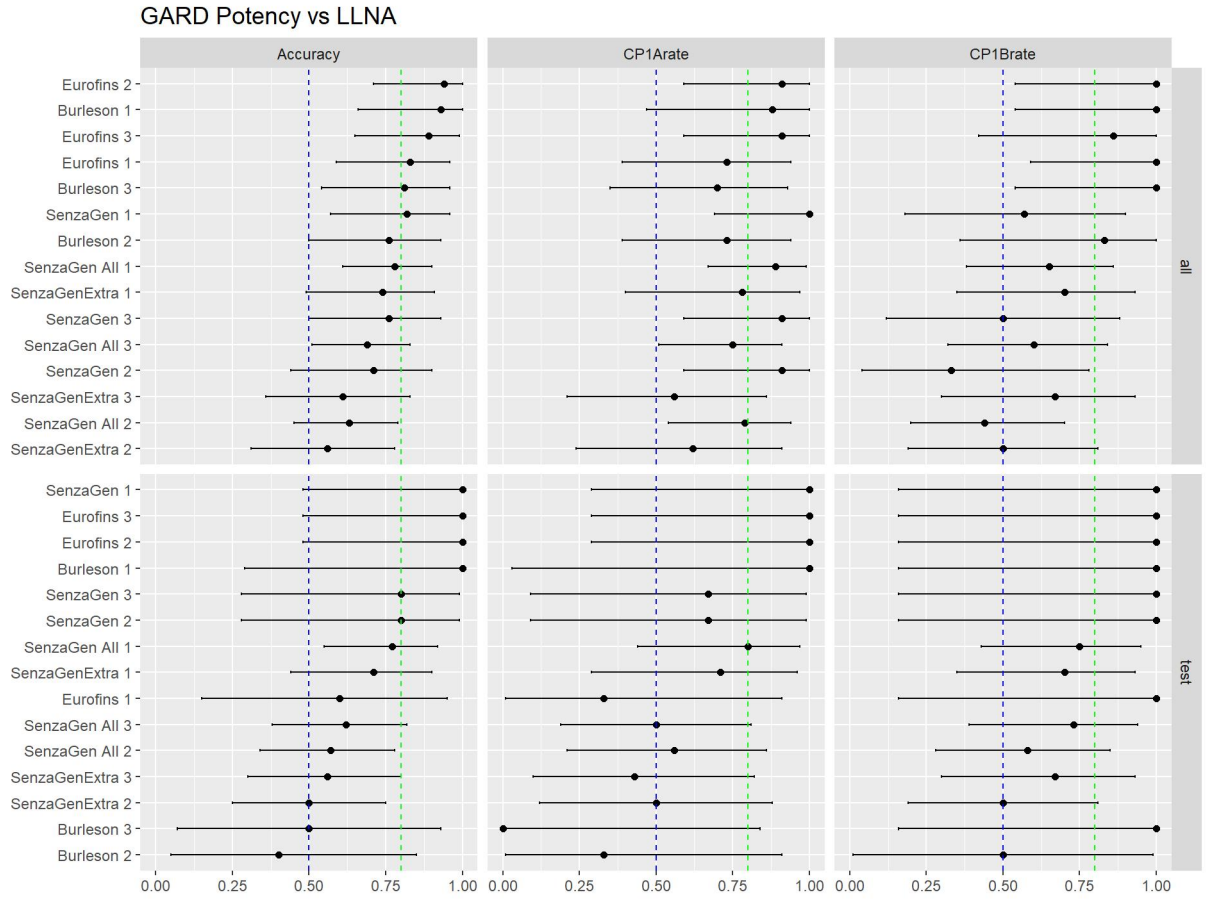


Figure 2: Predictive performance of the GARDpotency for all chemicals (top) and for test chemicals, i.e., excluding chemicals used for training of the GARDpotency method (bottom). Each calculation was performed for a single experiment. The dotted blue vertical line highlights 50%, i.e., blind random classification into two categories. The dotted green vertical line shows the value of 80%, i.e., a desirable value for correct classification. The order of the experiments is by generalised Youden index = CP1Arate + CP1Brate -1, a summary value for predictive capacity that is independent of class prevalence. The confidence intervals show the uncertainty of the estimates and give the full range of values in accordance with observed data. The confidence intervals in many settings cover 0.5, and more so when excluding the training chemicals, i.e., the performance values are not significantly better than random class allocation. "SenzaGen Extra" denotes the additional chemicals set (up to 31 chemicals) that were tested only in the SenzaGen laboratory, "SenzaGen All" denotes all chemicals, ring trial + additional chemicals.

In addition to the calculation of performance values described above, and in order to summarise the performance across laboratories and experiments of the GARDpotency, the ESAC WG also calculated CP1Arate, CP1Brate and accuracy using a weighted calculation, following the same approach described above for GARDskin. No confidence intervals have been derived for this weighted calculation. As above, these calculations were performed using 'all chemicals' and 'test chemicals' and in addition also show the results of the training chemicals for comparative purposes. These results are shown in Table 4.

Table 4: Weighted calculation of accuracy, Correctly Predicted Cat. 1A (CP1Arate) and Correctly Predicted Cat. 1B (CP1Brate) of GARDpotency, considering missing data as: N/A, incorrect classification or correct classification[a].

	Multilaboratory ring trial			**SenzaGen Extra**			**Ring trial + SenzaGen Extra**		
	All Chemicals (n=19)	**Training Set Chemicals (n=14)**	**Test Set Chemicals (n=5)**	**All Chemicals (n=21-23)[b]**	**Training Set Chemicals (n=2)**	**Test Set Chemicals (n=19-21)[b]**	**All Chemicals (n=40-42)[b]**	**Training Set Chemicals (n=16)**	**Test Set Chemicals (n=24-26)[b]**
Correctly Predicted Cat. 1A	*9.25 (9.18 - 9.29)*	*7.29 (7.22 - 7.33)*	*1.96*	*5.67 (5.67 - 8.00)*	*2.00*	*3.67 (3.67 - 6.00)*	*14.91 (14.85 - 17.29)*	*9.29 (9.22 - 9.33)*	*5.62 (5.62 - 7.96)*
Correctly Predicted Cat. 1B	*6.44 (5.00 - 6.67)*	*4.56 (3.11 - 4.78)*	*1.89*	*7.33 (6.00 - 8.33)*	*0.00*	*7.33 (6.00 - 8.33)*	*13.78 (11.00 - 15.00)*	*4.56 (3.11 - 4.78)*	*9.22 (7.89 - 10.22)*
Over-predicted Cat. 1B	*1.56 (3.00 - 1.33)*	*1.44 (2.89 - 1.22)*	*0.11*	*3.67 (6.00 - 3.67)*	*0.00*	*3.67 (6.00 - 3.67)*	*5.22 (9.00 - 5.00)*	*1.44 (2.89 - 1.22)*	*3.78 (6.11 - 3.78)*
Under-predicted Cat. 1A	*1.75 (1.82 - 1.71)*	*0.71 (0.78 - 0.67)*	*1.04*	*4.33 (5.33 - 3.00)*	*0.00*	*4.33 (5.33 - 3.00)*	*6.09 (7.15 - 4.71)*	*0.71 (0.78 - 0.67)*	*5.38 (6.38 - 4.04)*
Accuracy	82.6% (74.6% - 84.0%)	84.6% (73.8% - 86.5%)	76.9%	61.9% (50.7% - 71.0%)	100%	57.9% (46.0% - 68.3%)	71.7% (61.5% - 76.9%)	86.5% (77.1% - 88.2%)	61.9% (52.0% - 69.9%)
CP1Arate	84.1% (83.4% - 84.5%)	91.1% (90.3% - 91.7%)	65.2%	56.7% (51.5% - 72.7%)	100%	45.8% (40.7% - 66.7%)	71.0% (67.5% - 78.6%)	92.9% (92.2% - 93.3%)	51.1% (46.9% - 66.3%)
CP1Brate	80.6% (62.5% - 83.3%)	75.9% (51.9% - 79.6%)	94.4%	66.7% (50.0% - 69.4%)	N/A	66.7% (50.0% - 69.4%)	72.5% (55.0% - 75.0%)	75.9% (51.9% - 79.6%)	70.9% (56.3% - 73.0%)

[a] The first part of the table (numbers in *italics*) corresponds to the weighted number of chemicals. Since a weighted calculation is used to derive the values of correctly predicted Cat. 1A, correctly predicted Cat. 1B, overpredicted Cat. 1B and underpredicted Cat. 1A, these are not integers. The values in the second part of the table (accuracy, CP1Arate and CP1Brate) are reported as percentages. Where three values are reported, the main value (not in brackets) was calculated based on the available GARDpotency data. The first value within brackets assumes sensitisers with no GARDpotency data as 'wrong predictions' (worst case scenario) and the second value within brackets assumes sensitisers with no GARDpotency data as 'correct predictions' (best case scenario).

[b] In the validation study, there are overall 42 sensitisers but, due to the study design, only 40 sensitisers have data for GARDpotency. This is due to the fact that two of the 42 sensitisers (i.e., 1,2-cyclohexane dicarboxylic anhydride and benzyl salicylate) were predicted as non-sensitisers in GARDskin (from the SenzaGen Extra dataset) and, therefore, were not assessed in GARDpotency. The main reported values in the table do not take into account these two chemicals that have no data on GARDpotency. However, these two chemicals were taken into account for the calculation of the best case and worst case scenarios as described above. It should be noted that when doing this, the number of chemicals used for the main value will be different to that of chemicals reported between brackets for the worst and best case scenarios.

For the ring trial chemicals, the accuracy, CP1Arate and CP1Brate are all >80%. However, these are based on 14/19 chemicals which had been used to train the model. The 5 test set chemicals of the ring trial show different performances for CP1Arate and CP1Brate (but similar accuracy) as compared to all chemicals, but the small number of chemicals does not allow a sound conclusion to be drawn. The performance figures for the extra set of chemicals tested by SenzaGen, which had never been seen by the model, i.e., 21/23, are Accuracy=58%, CP1Arate=46%, CP1Brate=67%. Slightly better values are obtained if one considers the test set chemicals of the ring trial + test set chemicals of extra set, with Accuracy=62%, CP1Arate=51%, CP1Brate=71%. These data show a low rate of correct Cat. 1A (51%) prediction and a large difference in the rate of correct Cat. 1A predictions between the chemicals used to train the model and the test set chemicals (Accuracy

diff=25%, CP1Arate diff=42%, CP1Brate diff=5%). This cannot be explained only by the difference in chemical types between the two sets, but rather indicates a possible overfitting of the model. Thus, no sound conclusions on the performance of GARDpotency can be made.

The ESAC WG concluded that the available incomplete dataset is not sufficient at this time to draw a sound conclusion on the predictive capacity of GARDpotency to distinguish UN GHS Cat. 1A from UN GHS Cat. 1B sensitisers.

Combined GARDskin and GARDpotency

For the tiered approach (GARDskin + GARDpotency), in the submission, the predictive capacity was determined by the validity of assignment of a test chemical to a particular UN GHS Category, i.e. Cat. 1A, Cat. 1B or No Cat. The predictive capacity was evaluated for each individual laboratory using the majority voting for each chemical within the predictions of the 3 experiments. The accuracy, calculated including those chemicals that were used for training the SVM prediction model or defining the GPPS, ranged from 75.0% to 92.6%, with 100% of correct Cat. 1A classification in all 3 laboratories, Cat. 1B correct classifications ranging from 33.3% to 77.8% and No Cat. correct identification ranging from 88.9% to 100%. The test developer also accumulated the data over the three laboratories to report a cumulative performance of 86.1% accuracy, 100% correctly predicted Cat. 1A, 61.5% correctly predicted Cat. 1B and 96% correctly predicted No Cat.

In addition, the submission reported the results obtained for an extra set of 31 coded test chemicals (11 Cat. 1A, 12 Cat. 1B and 8 No Cat.) tested in a blind manner by the lead laboratory only (i.e., SenzaGen). These substances were predicted with an accuracy of 58.1%, 45.5% correct prediction for Cat. 1A, 58.3% correct prediction for Cat. 1B and 75% correct prediction for No Cat.

When combined with the multilaboratory ring trial chemicals, the overall accuracy was 66.1% (n=59), with 71.4% correct prediction for Cat. 1A (n=21), 47.4% correct prediction for Cat. 1B (n=21) and 82.4% correct prediction for No Cat. (n=7). These values were slightly lower when considering the data generated only by the lead laboratory, with 58.1% accuracy (n=31), 45.5% correct prediction for Cat. 1A (n=11), 58.3% correct prediction for Cat. 1B (n=12) and 75% correct prediction for No Cat. (n=8).

ESAC WG Analysis of Predictive Capacity

The ESAC WG noted that 22 chemicals of the ring trial and 6 of the extra set were used to train the GARDskin and/or GARDpotency models.

Furthermore, when using the combined GARDskin and GARDpotency approach, a single experiment would be used for prediction of new chemicals. The ESAC WG considers therefore that the evaluation of predictive capacity should be performed in this setting (i.e., considering all individual predictions obtained in the validation study to calculate predictive capacity instead of applying the majority voting to the multiple predictions obtained by each laboratory). The ESAC WG therefore calculated accuracy, Correctly Predicted Cat. 1A rate (CP1Arate), Correctly Predicted Cat. 1B (CP1Brate) rate and Correctly Predicted No Cat. rate (defined below as CPNoCatrate) for single experiments in all the laboratories. A calculation was made with all chemicals included (all) and another excluding chemicals that were used for training the GARDskin and/or the GARDpotency classifiers (test). Resulting estimates and confidence intervals are provided in Figure 3. Performance values using only the test chemicals were: CP1Arate ranged from 0% to 100% (with the width of the 95%-CIs between 55 and 97 percentage points and all but one lower Confidence limits below 33%); CP1Brate ranged from 36% to 100% (with the width of the 95%-CIs at least 52 and up to 98 percentage points and all but one lower Confidence limits below 33%); CPNoCatrate ranged from 0% to 100% (with the width of the 95%-CIs between 74 and 98 percentage points and all lower Confidence limits below 33%).

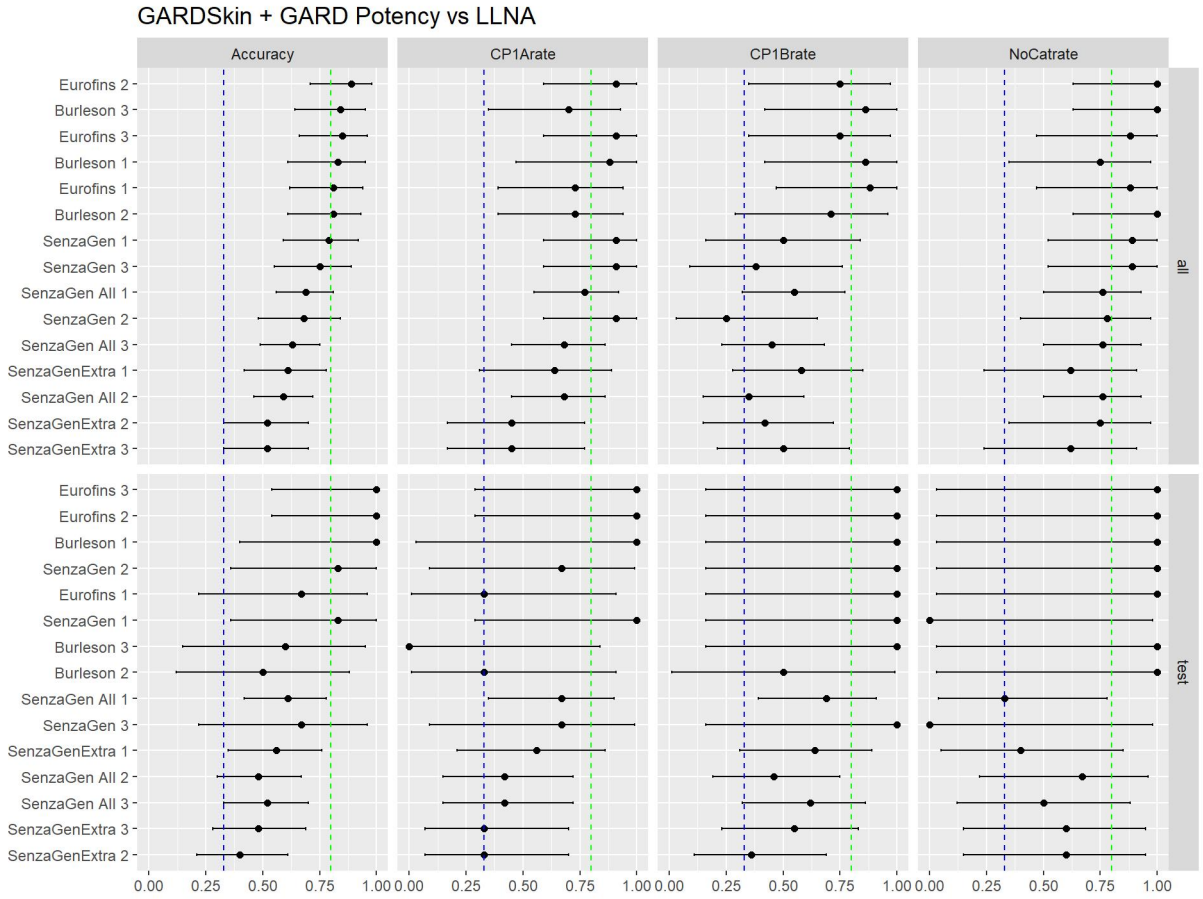


Figure 3: Predictive performance of the combination of GARDskin and GARDpotency for all chemicals (top) and for test chemicals, i.e., excluding chemicals used for training of the GARDskin algorithm as well as chemicals used for training the GARDpotency method (bottom). Each calculation was performed for a single experiment. The dotted blue vertical line highlights 33%, i.e., blind random classification into one of three categories. The dotted green vertical line shows the value of 80%, i.e., a desirable value for correct classification. The order of the experiments is by generalised Youden index = CP1Arate + CP1Brate + CPNoCatrate -1, a summary value for predictive capacity that is independent of class prevalence. The confidence intervals show the uncertainty of the estimates and give the full range of values in accordance with observed data. The confidence intervals in many settings cover 0.5, specifically when excluding the training chemicals, i.e., the performance values are not significantly better than random class allocation. "SenzaGen Extra" denotes the additional chemicals set (up to 31 chemicals) that were tested only in the SenzaGen laboratory, "SenzaGen All" denotes all chemicals, ring trial + additional chemicals.

In addition to the calculation of performance values described above, and in order to summarise the performance across laboratories and experiments of the tiered approach, the ESAC WG re-evaluated the performance using the same weighted methodology as explained before for GARDskin and GARDpotency. No confidence intervals have been derived for this weighted calculation. As above, these calculations were performed using 'all chemicals' and 'test chemicals' and in addition also show the results of the training chemicals for comparative purposes. Table 5 shows the obtained results.

Table 5: Weighted calculation of accuracy, Correctly Predicted Cat. 1A (CP1Arate), Correctly Predicted Cat. 1B (CP1Brate) and Correctly Predicted No Cat. (CP1NoCatrate) of the GARDskin + GARDpotency combined approach[a].

	Multilaboratory ring trial			**SenzaGen Extra**			**Ring trial + SenzaGen Extra**		
	All Chem. (n=28)	**Training Set Chem. (n=22)**	**Test Set Chem. (n=6)**	**All Chem. (n=31)**	**Training Set Chem. (n=6)**	**Test Set Chem. (n=25)**	**All Chem. (n=59)**	**Training Set Chem. (n=28)**	**Test Set Chem. (n=31)**
Correctly Predicted 1A	*9.2*	*7.2*	*2.0*	*5.7*	*2.0*	*3.7*	*14.8*	*9.2*	*5.6*
Correctly Predicted 1B	*5.1*	*3.2*	*1.9*	*6.0*	*0.3*	*5.7*	*11.1*	*3.6*	*7.6*
Correctly Predicted NoCat	*8.1*	*7.3*	*0.8*	*5.3*	*2.7*	*2.7*	*13.4*	*10.0*	*3.4*
1A-->Predicted as 1B	*1.7*	*0.6*	*1.0*	*3.0*	*0.0*	*3.0*	*4.7*	*0.7*	*4.0*
1A-->Predicted as NoCat	*0.1*	*0.1*	*0.0*	*2.3*	*0.0*	*2.3*	*2.4*	*0.1*	*2.3*
1B-->Predicted as 1A	*1.6*	*1.4*	*0.1*	*3.7*	*0.0*	*3.7*	*5.2*	*1.4*	*3.8*
1B-->Predicted as NoCat	*1.3*	*1.3*	*0.0*	*2.3*	*0.7*	*1.7*	*3.7*	*2.0*	*1.7*
NoCat-->Predicted as 1A	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*
NoCat-->Predicted as 1B	*0.9*	*0.6*	*0.2*	*2.7*	*0.3*	*2.3*	*3.6*	*1.0*	*2.6*
CP1Arate (CP1A/1A)	83%	90%	65%	52%	100%	41%	67%	92%	47%
CP1Brate (CP1B/1B)	64%	54%	94%	50%	33%	52%	56%	51%	58%
CPNoCatrate (CPNocat/NoCat)	90%	92%	78%	67%	89%	53%	79%	91%	57%
Accuracy ((CP1A+CP1B+CPNoCat)/ all)	80%	81%	77%	55%	83%	48%	67%	81%	54%
Rate of 1A Predicted as 1B or NoCat (Sens 1A)	17%	10%	35%	48%	0%	59%	33%	8%	53%
Rate of 1A Predicted as 1B	16%	8%	35%	27%	0%	33%	21%	7%	34%
Rate of 1A Predicted as NoCat	1%	1%	0%	21%	0%	26%	11%	1%	19%
Rate of 1B Predicted as 1A or NoCat (Sens 1B)	36%	46%	6%	50%	67%	48%	44%	49%	42%
Rate of 1B Predicted as 1A	19%	24%	6%	31%	0%	33%	26%	21%	29%
Rate of 1B Predicted as NoCat	17%	22%	0%	19%	67%	15%	18%	29%	13%
Rate of NoCat Predicted as 1A or 1B (Sens NoCat)	10%	8%	22%	33%	11%	47%	21%	9%	43%
Rate of NoCat Predicted as 1A	0%	0%	0%	0%	0%	0%	0%	0%	0%
Rate of NoCat Predicted as 1B	10%	8%	22%	33%	11%	47%	21%	9%	43%

[a] The first part of the table (numbers in *italics*) corresponds to the weighted number of chemicals. Since a weighted calculation is used, the values are not integers. The values in the second part of the table are reported as percentages and correspond to the rates of correctly predicted, over- and under predicted chemicals.

The ESAC WG noted that there is a low number of chemicals within the multilaboratory ring trial that have never been seen by the models (6 out of 28; 21%). The performance for the ring trial

chemicals to predict Cat. 1A is 83%, Cat. 1B is 64% and No Cat. is 90%. However, when considering only the test set, the values for Cat. 1A and No Cat. decrease to 65% and 78%, respectively, whereas Cat. 1B increases to 94%. The performances observed with the extra chemicals tested by the lead laboratory only were even lower with CP1Arate=52%, CP1Br=50%, CPNoCatrate=67%. If only the test set chemicals of the extra set are considered, the CP1Arate=41%, CP1Brrate=52% and CPNocatrate=53%. The real performance of the model can be assessed by evaluating its performance on the test set chemicals of the ring trial and extra set together, which is CP1Arate=47% (n=12), CP1Br=58% (n=13), CPNoCatrate=57% (n=6). There are large differences in performance between training set chemicals and test set chemicals (CP1Arate diff=45%, CP1Brate diff=-7%, CPNoCatrate diff=34%). Such large differences are typical of either overfitted models or too small sample size.

When taking into consideration the above analyses, the ESAC WG notes that the number of test chemicals that were not previously used to train the prediction models represented only a minority of the chemicals used in the validation study (i.e., only 8 of 29 chemicals (28%) in the multilaboratory ring trial and 31 out of 57 chemicals (52%) in the extended set of chemicals). This resulted in low confidence in the predictions especially for non-sensitisers. Furthermore, when taking into account all datasets available, the performance for the test set chemicals was low (CP1Arate=47%, CP1Brate=58%, CPNoCatrate=57%), with large differences observed for the predictions of Cat. 1A and No Cat. when compared to the training set chemicals (CP1Arate diff=45%, CP1Brate diff=-7%, CPNocatrate diff=34%). Approximately 11% of the Cat. 1A sensitisers were incorrectly predicted as non-sensitisers when considering all tested chemicals, and 19% if only test set chemicals are considered. There were also large differences in performance between the multilaboratory ring trial results and the additional data obtained by the lead laboratory. This may be due to an insufficient amount of tested chemicals.

In conclusion, the ESAC WG considers that no sound conclusions can be made at present on the combined GARDskin and GARDpotency approach to identify UN GHS Cat. 1A, UN GHS Cat. 1B and UN GHS No Cat. chemicals as a stand-alone approach. The ESAC WG considers that the issues with GARDpotency stated above must be addressed before a combined approach can be assessed for its performance. Furthermore, it would be useful to have a study design that provides a more robust and complete dataset for evaluation, and which includes a larger number of test chemicals (especially non-sensitisers), which have not been used to train the model. This would potentially lead to a reduction in the width of the confidence intervals for specificity and allow an assessment of how false positive results from any first tier method other than GARDskin would be classified if tested in GARDpotency.

10.2 Overall relevance (biological relevance and accuracy) of the test method in view of the purpose

The purpose of the GARD platform is to provide a two-tiered approach to definition of a) if a compound is likely to be a skin sensitiser (GARDskin) and, of those shown to be positive in this prediction, b) in which CLP/GHS Category 1 potency class they reside in (GARDpotency). In terms of the first tier, the method shows predictability for separating sensitisers and non-sensitisers, despite both the caveat of the cell type being only one of several which are ultimately involved in the sensitisation process detailed in the AOP (OECD, 2012), and the claim that KE3 is the primary functionality within the test cells. There is also a question mark around the metabolic capacity of the cell type in bioactivation of pro-haptens to protein binding species, required by many skin sensitising agents, which actually occurs within the dermal epithelial cells. Attempts to define potency (GARDpotency) are based on quantitative gene expression “dose” changes within a sub-set of genes utilised in the GARDskin assay. Similar arguments regarding KE representation and cell-cell

molecular communication in skin sensitisation *in vivo* as to those in the GARDskin assay are also relevant here. It is also important to understand why this minimal gene signature is more likely to reveal a quantitative relationship to different potency classes in view of this.

11. Applicability domain (Module 6)

11.1 Appropriateness of study design to conclude on applicability domain, limitations and exclusions

Overall, the Applicability Domain of the GARD platform has not been thoroughly evaluated, as its characterisation was regarded as unnecessary by the test developer. The test developer felt difficult to exclude any chemical in view of the fact that, according to them, the GARDskin test method maintained an accuracy of 89% (sensitivity: 89%; specificity: 88%) when applied on 'difficult' substances (Johansson et al., 2014). Furthermore, the conclusion of Module 1 of the GARDpotency VSR on page 27, states that historical data demonstrate that the GARDskin can correctly predict pre- and pro-haptens, while chemicals with physico-chemical characteristics, that empirically would exclude them from the applicability domain of the test method, cannot be excluded without experimentation. However, when testing more chemicals (Conclusion on Module 5, page 44) the 66.7% accuracy obtained with the thirty-one additional chemicals seems to indicate a need to carefully define the chemical applicability domain of the method.

The ESAC considers that GARD platform validation studies were not properly designed to investigate and define limitations in the applicability domain of the test methods.

It is important to identify or foresee possible limitations to properly inform test method users. While chemicals may not be excluded without experimentation, caution must be used in the case of negative results. For the time being, as the applicability domains and limitations have not been fully explored, negative results should be carefully investigated in combination with other evidence.

The ESAC WG felt that the development of the GARD platform would certainly have benefited from a better evaluation of the applicability domain from the onset, which would have perhaps avoided a poor predictivity at the end of the validation study.

11.2 Quality of the description of applicability domain, limitations, exclusions

The Applicability Domain of the GARD platform has not been thoroughly evaluated and it is not the duty of the ESAC WG to define its applicability domain. The ESAC WG notes that this information is not sufficiently described in the VSRs. However, the ESAC WG acknowledges that this information was subsequently provided upon specific request.

A subsequent analysis by Lhasa (GARD Applicability Domain Project, Appendix 5 of the submission) evaluated the performance of GARD against an *in vivo* call using both human and mouse data. Regarding the mechanistic domains, as the number of chemicals present in some domains (SNAr, SN2) was fairly small, drawing hard conclusions is ill-advised. However, from the limited data it can be observed that the GARD assay predicts all domains assessed very well, including chemicals with no mechanistic domain assigned. Regarding the potential relationship to LogP, only 6 chemicals with a logP > 5 and 12 chemicals with a logP between 3.5 and 5 have been tested in GARDskin thus far, therefore it is difficult to fully assess the applicability of the assay for lipophilic chemicals.

It is hard to envisage that the GARD platform, based on traditional cell culture in aqueous media, will not face problems in the identification of chemicals either with poor solubility (in two of the three laboratories, one chemical (dextran) had to be excluded for solubility issue), or very lipophilic character (the misclassification of lipophilic compounds, LogP > 3.5, tested in GARDskin is 17%, in the h-CLAT is 24%), or instability in water or requiring metabolic activation (not all pro-haptens were correctly predicted, e.g., isoeugenol). Additionally, the presence of mRNAs for CYPs and phase II enzymes level is not sufficient to define the metabolic competence of the test system used.

The ESAC WG recommends a more in-depth definition of the chemical applicability domain of the GARD methods, as a better understanding of false positive and false negative results is fundamental for a proper use of the methods in the future, defining upfront the chemicals suitable

to be tested. The ESAC WG also recommends a fuller investigation of the metabolic capacity of the test cells, both from relevant protein expression and enzymatic activity perspectives.

12. Performance standards (Module 7)

12.1 Adequacy of the proposed Essential Test Method Components

Not applicable

12.2 Adequacy of the proposed Reference Chemicals

Not applicable

12.3 Adequacy of the proposed performance target values

Not applicable

13. Readiness for standardised use

13.1 Assessment of the readiness for regulatory purposes

GARDskin

The ESAC WG considers the GARDskin to be ready for regulatory use in the context of identification of skin sensitisers and chemicals not classified for skin sensitisation, based on evidence of sufficient reproducibility and transferability. Depending on the regulatory context, positive results obtained using GARDskin as a stand-alone assay may be used to identify skin sensitisers. However, a negative result in this assay may not be sufficient stand-alone evidence to identify non-sensitisers, given the limited characterisation of the applicability domain and mechanistic coverage.

GARDpotency

The ESAC does not consider the information currently available on GARDpotency to be sufficient, at present, to recommend the use of GARDpotency for regulatory purposes, whether used as a stand-alone or in combination with another assay. This is due to the issues described within this report related to i) the design of the validation study, ii) the reproducibility of GARDpotency and iii) the predictive capacity of GARDpotency for discriminating Cat. 1A and Cat. 1B sensitisers.

Combined GARDskin and GARDpotency

The ESAC considers that the issues with GARDpotency stated above must be addressed before the readiness for regulatory use of the combined GARDskin and GARDpotency approach can be assessed.

13.2 Assessment of the readiness for other uses

Both GARDskin and GARDpotency could be useful in screening for internal decision making in the industrial setting.

13.3 Critical aspects impacting on standardised use

Among the factors that may impact the implementation of the GARD methods in any laboratory, the following can be mentioned:

- When transferring the GARD assay to a new laboratory, if nanoString capability is not already established within it, the establishment of a core facility for nanoString services should be prioritised. Alternatively, cell stimulations could be performed in-house, with a subsequent shipment of RNA samples to a CRO providing nanoString services for further analysis. In this respect, the test developer offers a service for nanoString, which helps ensure access.
- In terms of transfer to a naïve laboratory, the implementation of the GARD methods requires proper training by the test developer, as the protocol used is not necessarily self-explanatory.
- Each laboratory should define its own historical/control data to ensure consistency.
- As the applicability domain (e.g., considering factors such as chemical space, metabolic activation, volatility, water solubility, etc.) has not been fully investigated, negative results should be carefully considered.
- The naïve laboratory should consider costs associated with Licensing (e.g., acquisition of the cell line).

13.4 Gap analysis

The GARD validation study was designed to assess transferability and reproducibility of the GARDskin and GARDpotency methods. The test developer decided to utilise the samples from the validation of the GARDskin for further assessing the GARDpotency. While use of samples from the GARDskin validation study gave an indication of the utility of the GARDpotency assay for classification, the ESAC WG considers that the latter would have benefited from a dedicated and appropriately powered study design. The lack of such a dedicated study design ultimately did not allow the ESAC WG to draw conclusions on the scientific validity of the GARDpotency assay.

14. Other considerations

None to mention.

15. Conclusions on the study

15.1 ESAC WG summary of the results and conclusions of the study

Conclusions common to both GARDskin and GARDpotency

The ESAC WG considers that the GARDskin and GARDpotency test definition describing the test system, biological and/or mechanistic relevance, test acceptance criteria, protocol, prediction model and SOP, was generally appropriate. Minor recommendations have been made in this document for future improvement of some of the documentation concerning SOPs and prediction models. The SVM algorithms used in GARDskin and GARDpotency were also evaluated by the ESAC WG as part of the scientific review. Based on the SVM code, data and additional explanations provided by the test developer the ESAC WG was able to independently reproduce the models and the results of the GARDskin and GARDpotency validation study, despite a few minor mismatches in data and typographical errors described above in this report. Further, the online tool for data analysis (GDAA platform) was found to be functional and user-friendly. The ESAC WG concludes, therefore, that the evidence supporting the GARDskin and GARDpotency SVM algorithms is adequate.

The ESAC WG also considers that the transferability study was properly conducted and that the GARDskin and GARDpotency methods were shown to be transferable. While changes were made to the acceptance criteria and to the SOP during the transferability study, these changes were well-documented, scientifically justified, and were not deemed to be detrimental to assay reliability.

However, the ESAC WG noted that a large proportion of the chemicals used in the blind multilaboratory trial of the validation study were also part of the training set used to build the GARDskin and GARDpotency SVM prediction models. The ESAC WG considers that selecting a large proportion of the chemicals used to train a model (training set) for the validation study (test set) is inappropriate. When a machine-learning algorithm (e.g., SVM) is used, it is even more important that the principle of keeping training and test sets separate is strictly adhered to, due to the risk of overfitting. This limitation in the study design impacted the evaluation of both GARDskin and GARDpotency.

Conclusions specifically relating to GARDskin

The overall objective of the GARDskin validation study, as described in the VSR, was to demonstrate the transferability and reproducibility of the method and to provide evidence supporting the GARDskin method as a reliable tool for assessing skin sensitisation hazard with added value to any integrated testing strategy in which it is included.

The study design, based on number of laboratories and chemicals, used to assess the WLR, BLR and predictive capacity of GARDskin, was considered to be sufficient and appropriate by the ESAC WG, as it is comparable to the study design used to evaluate currently adopted test methods.

The WLR of GARDskin was in the range of 78.6% to 89.2% and the BLR was 82.1% (95%-CI: 63.1-93.9%), when considering all chemicals, including those claimed to have technical issues by the test method developer. When considering the dataset where runs not complying with cell viability acceptance criteria are excluded, the WLR of GARDskin was in the range of 82.1% to 88.9% and the BLR was 92.0% (23/25, 95%-CI: 74.0-99.0%). In both cases, the WLR and BLR of GARDskin were considered to be appropriate by the ESAC WG.

The ESAC WG evaluated the predictive capacity of GARDskin on the basis of individual experiments and taking into account only test chemicals that were not used for training the SVM prediction model. Based on this, the sensitivity of GARDskin ranged from 76% to 100% across experiments and laboratories, with the width of the 95%-CIs between 22 and 45 percentage points, and all lower confidence limits above 50%. Specificity ranged from 40% to 100%, with the width of the 95%-CIs being at least 63 and up to 80 percentage points. Accuracy ranged from 73% to 100% (with the width of the 95%-CIs at least 21 and up to 36 percentage points and all lower Confidence

limits above 50%). Based on a weighted calculation, and taking into account only the test chemicals not used to train the SVM prediction model, from both the multilaboratory ring trial and the additional set of chemicals, a sensitivity of 87%, a specificity of 70% and an accuracy of 83% were obtained.

The ESAC WG concluded that the predictive capacity of GARDskin is appropriate to support the discrimination between skin sensitisers and chemicals not classified for skin sensitisation. The GARDskin's performance was considered comparable to other *in vitro/in chemico* methods currently adopted to support skin sensitisation hazard identification. Nonetheless, it would have been useful to have included a larger number of chemicals (especially non-sensitisers), which had not been used to train the model, to increase the precision of the estimates.

In conclusion, the ESAC WG is of the opinion that the evidence provided on GARDskin is sufficient and adequate to support its scientific validity. Thus, the ESAC considers that GARDskin is ready to progress to further consideration by the OECD for Test Guideline development. GARDskin can contribute to skin sensitisation hazard identification in a weight-of-evidence approach. Depending on the regulatory context, positive results obtained with GARDskin may be used stand-alone to identify skin sensitisers. However, a negative result obtained with this assay may not be sufficient stand-alone evidence to identify non-sensitisers and should be considered together with additional evidence.

Conclusions specifically relating to GARDpotency

The overall objective of the GARDpotency validation study, as described in the VSR, was to demonstrate the reproducibility of the method and to provide evidence supporting the GARDpotency method as a reliable second-tier to the GARDskin for assessing the potency (Cat. 1A/Cat. 1B) of a skin sensitisation hazard, with added value to any integrated testing and assessment strategy in which it is included.

When considering the study design for the GARDpotency validation study, the ESAC WG had concerns, due both to the inconsistent number of runs available for each chemical for assessing the method's reproducibility, and to the limited number of new chemicals (testing set) assessed for predictive capacity. This was due primarily to the fact that only samples that were identified as sensitisers by the GARDskin test, in each laboratory and within each run, were subsequently analysed with GARDpotency. This led to differences in the number of chemicals analysed with GARDpotency assay by the different laboratories, as well as in differences in the number of runs conducted for each chemical, resulting in an incomplete data matrix, in which chemicals had a varying number of repetitions. In addition, the ESAC WG also noted that a large proportion of the sensitisers analysed with GARDpotency were also part of the training set used to build the GARDpotency SVM algorithm or define the GPPS (16 out of 40 sensitisers). The ESAC WG considers that these limitations in the study design hindered the assessment of the reproducibility and predictive capacity of GARDpotency method.

The target for both WLR and BLR was set at 75% by the VMG for GARDpotency, as opposed to 80% for GARDskin. However, the ESAC WG found insufficient justification for this reduction in the target value. The WLR was in the range of 62.5% to 88.9% in the best case scenario (see Section 7.2 above), being below the target for one of the laboratories. The BLR for the 18 chemicals with valid results in at least two laboratories was 61.1% (95%-CI: 35.7-82.7%), which was also below the set target. The ESAC WG also calculated the BLR for the 14 chemicals for which a prediction could be derived from all three laboratories, even if based only on two concordant experiments in a laboratory (instead of three), and this was still below the target: 71.4% (95%-CI: 41.9-91.6%). Thus, despite the reduced target for reproducibility, the assay failed to meet set criteria in several instances.

The ESAC WG evaluated the predictive capacity of GARDpotency on the basis of individual experiments and taking into account only test chemicals that were not used for training the SVM algorithm or defining the GPPS. The correct Cat. 1A classification rate ranged from 0% to 100%

across experiments and laboratories, with the width of the 95%-CIs between 53 and 97 percentage points and all lower confidence limits below 50%. The correct Cat. 1B classification rate ranged from 50% to 100% with the width of the 95%-CIs between 52 and 98 percentage points and all lower Confidence limits below 50%. Based on a weighted calculation, and taking into account only the test chemicals not used to train the SVM algorithm or define the GPPS, from both the multilaboratory ring trial and the additional set of chemicals, a correct Cat. 1A classification rate (CP1Arate) of 51%, and a correct Cat. 1B classification rate (CP1Brate) of 71% were obtained. Furthermore, a large difference was observed in the Cat. 1A predictions between the chemicals used to train the model and the test set chemicals (CP1Arate diff=42%). This cannot be explained only by the difference in chemical types between the two sets, but rather indicates possible overfitting of the model. The ESAC WG concluded that the available incomplete dataset is not sufficient at this time to draw a sound conclusion on the predictive capacity of GARDpotency to distinguish UN GHS Cat. 1A from UN GHS Cat. 1B sensitisers.

In conclusion, the ESAC does not consider the information currently available on GARDpotency to be sufficient, at present, to recommend its use for regulatory purposes (in combination with any other assay). The use of the GARDpotency assay for discriminating Cat. 1A and Cat. 1B sensitisers is currently hampered by the identified issues in reproducibility and predictive capacity due to the design of the validation study.

Conclusions related to a Combined GARDskin and GARDpotency strategy

The ESAC considers that the issues with GARDpotency stated above must be addressed before a combined approach can be assessed for its performance. No sound conclusions can be made at present on the combined GARDskin and GARDpotency approach to identify UN GHS Cat. 1A, UN GHS Cat. 1B and UN GHS No Cat. chemicals, as a stand-alone approach. Furthermore, it would be useful to have a study design that provides a more robust and complete dataset for evaluation, and which includes a larger number of test chemicals (especially non-sensitisers), which have not been used to train the model. This would potentially lead to a reduction in the width of the confidence intervals for specificity and allow an assessment of how false positive results from any first tier method other than GARDskin would be classified if tested in GARDpotency.

15.2 Extent to which study conclusions are justified by the study results alone

All ESAC WG conclusions are based on the validation study data, as well as the supplemental data produced by the test developer and the additional analyses conducted by the ESAC WG.

15.3 Extent to which conclusions are plausible in the context of existing information

Not applicable

16. Recommendations

16.1 General recommendations

While current data on the use of GARDskin for the identification of pro-haptens look promising, the ESAC WG recommends further characterisation of the metabolic capacity of the test system due to the important role of metabolic activation of pro-haptens in the skin sensitisation AOP (OECD, 2012). Currently, the only evaluation of metabolic capacity has been based on mRNA expression. Further evaluation via enzyme activity and protein expression would increase confidence in the use of the test system to account for bioactivated chemicals.

To adequately evaluate the reproducibility and predictive capacity of the GARDpotency and of a combined approach using GARDskin or another *in vitro* assay for initial identification of sensitisers together with GARDpotency for Cat. 1A/Cat. 1B classification, additional chemicals must be tested in the GARDpotency assay in at least three laboratories. While a post-hoc power analysis was provided by the test developer to justify the number of chemicals used to evaluate GARDpotency, the study design, which required a positive GARDskin test to move a chemical into the GARDpotency assay, led to an incomplete data matrix for final evaluation of reproducibility and accuracy of GARDpotency. Furthermore, no rationale was provided for the number of chemicals used to evaluate the GARDskin + GARDpotency combined approach to predict three skin sensitisation categories (versus two predicted categories by GARDskin and by GARDpotency). Thus, the ESAC WG recommends the use of an a priori power analysis to ensure optimal study design specifically for GARDpotency and for the GARDskin + GARDpotency combined approach.

16.2 Specific recommendations (e.g., concerning improvement of SOPs)

Recommendations on SOP

The ESAC WG recommends that the SOP be revised to clarify the experimental protocol, particularly with regard to the procedures for dealing with samples that fail viability controls. The terms “stimulations” and “biological replicates” should be clearly defined.

Note that, at present, the prediction model of the GARDpotency is described in Appendix 7 of the GARDpotency submission in the file “Amendment to the GARD assay SOP v.06.01: Predicting skin sensitiser potency using the GARD Data Analysis Application”. The ESAC WG recommends that this Appendix is incorporated in a new version of SOP to avoid having multiple documents.

Recommendations on SVM algorithms

Additional safety check on the correctness of the annotation file is recommended. The annotation file is a key file linking the raw data to the chemical (and concentration, in the case of GARDpotency) tested, which is manually filled in. Typos or mistakes in this file may affect the derivation of the Decision Value and therefore the final prediction. Furthermore, code and data should be provided in a format that would enable easy verification.

Even though the current GARDskin SVM algorithm is considered appropriate for its purpose, the ESAC WG considers it to be overly complex and that it could benefit from simplification (see Appendix II to this report). A simpler model would be cheaper to conduct and easier to assess and understand.

17. References

- Basketter D.A., Alépée N., Ashikaga T., Barroso J., Gilmour N., Goebel C., Hibatallah J., Hoffmann S., Kern P., Martinozzi-Teissier S., Maxell G., Reisinger K., Sakaguchi H., Schepky A., Tailhardat M., Templier, M. (2014) Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis* 25(1):11-21. doi: 10.1097/DER.0000000000000003.
- Casati S., Aeby P., Kimber I., Maxwell G., Ovigne J.M., Roggen E., Rovida C., Tosti L., Basketter D. (2009) Selection of chemicals for the development and evaluation of *in vitro* methods for skin sensitisation testing. *Altern. Lab. Anim.* 37(3):305-12. doi: 10.1177/026119290903700313.
- Diaz-Uriarte R. (2007) GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics* 8:328. doi: 10.1186/1471-2105-8-328.
- EC (2006) Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *OJ* L 396, 30.12.2006, p. 1–850 (Current consolidated version: 05/07/2021).
- EC (2008) Regulation (EC) No. 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006 (Text with EEA relevance). *OJ* L 353, 31.12.2008, p. 1–1355 (Current consolidated version: 10/05/2021).
- EC (2009a) Regulation (EC) No. 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products (Text with EEA relevance). *OJ* L 342, 22.12.2009, p. 59–209 (Current consolidated version: 26/05/2021).
- EC (2009b) Regulation (EC) No. 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. *OJ* L 309, 24.11.2009, p. 1–50 (Current consolidated version: 27/03/2021).
- EU (2012) Regulation (EU) No. 528/2012 of the European Parliament and of the Council of 22 May 2012 concerning the making available on the market and use of biocidal products (Text with EEA relevance). *OJ* L 167, 27.6.2012, p. 1–123 (Current consolidated version: 29/03/2021).
- EURL ECVAM (2012) Direct Peptide Reactivity Assay (DPRA): ECVAM Validation Study Report Available at: https://tsar.jrc.ec.europa.eu/test-method/tm2009-06 (accessed on 06/07/2021).
- Forreryd A., Zeller K.S., Lindberg T., Johansson H., Lindstedt M. (2016) From genome-wide arrays to tailor-made biomarker readout – Progress towards routine analysis of skin sensitizing chemicals with GARD. *Toxicol. In Vitro* 37:178-188. doi: 10.1016/j.tiv.2016.09.013.
- Gradin R., Lindstedt M., Johansson H. (2019) Batch adjustment by reference alignment (BARA): Improved prediction performance in biological test sets with batch effects. *PLoS One* 14(2): e0212669. doi: 10.1371/journal.pone.0212669.

- Gradin R., Johansson A., Forreryd A., Aaltonen E., Jerre A., Larne O., Mattson U., Johansson H. (2020) The GARDpotency assay for potency-associated subclassification of chemical skin sensitizers – Rationale, method development, and ring trial results of predictive performance and reproducibility. *Toxicol. Sci.* 176(2):423-432. doi: 10.1093/toxsci/kfaa068.
- Hartung T., Bremer S., Casati S., Coecke S., Corvi R., Fortaner S., Gribaldo L., Halder M., Hoffmann S., Roi A.J., Prieto P., Sabbioni E., Scott L., Worth A., Zuang V. (2004) A modular approach to the ECVAM principles on test validity. *Altern. Lab. Anim.* 32(5):467-472. doi: 10.1177/026119290403200503.
- Johansson H., Lindstedt M., Albrekt A.S., Borrebaeck C.A. (2011) A genomic biomarker signature can predict skin sensitizers using a cell-based *in vitro* alternative to animal tests. *BMC Genomics* 12:399. doi: 10.1186/1471-2164-12-399.
- Johansson H., Rydnert F., Kühnl J., Schepky A., Borrebaeck C., Lindstedt M. (2014) Genomic allergen rapide detection in-house validation – A proof of concept. *Toxicol. Sci.* 139(2):362-370. doi: 10.1093/toxsci/kfu046.
- Johansson H., Gradin R., Forreryd A., Agemark M., Zeller K., Johansson A., Larne O., van Vliet E., Borrebaeck C., Lindstedt M. (2017) Evaluation of the GARD assay in a blind Cosmetics Europe study. *ALTEX* 34(4):515-523. doi: 10.14573/altex.1701121.
- Natsch A., Ryan C.A., Foertsch L., Emter R., Jaworska J, Gerberick F., Kern P. (2013) A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. *J. Appl. Toxicol.* 33(11):1337-1352. doi: 10.1002/jat.2868.
- NCSS (2017) PASS 15 Power Analysis and Sample Size Software. NCSS, LLC. Kaysville, Utah, USA. ncss.com/software/pass.
- OECD (2005) Guidance Document on the validation and international acceptance of new or updated test methods for hazard assessment. OECD Environment, Health and Safety Publications; Series on Testing and Assessment, No. 34. Organisation for Economic Co-operation and Development, Paris. ENV/JM/MONO(2005)14.
- OECD (2007) Guidance Document on the validation on the validation of (Quantitative)Structure-Activity Relationships [(Q)SAR] models. OECD Environment, Health and Safety Publications; Series on Testing and Assessment, No. 69. Organisation for Economic Co-operation and Development, Paris. ENV/JM/MONO(2007)2.
- OECD (2010a) Test Guideline No. 429 - Skin Sensitization: Local Lymph Node Assay. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.
- OECD (2010b) Test Guideline No. 442A - Skin Sensitization: Local Lymph Node Assay: DA. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.
- OECD (2012) The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. OECD Environment, Health and Safety Publications; Series on Testing and Assessment, No. 168. Organisation for Economic Co-operation and Development, Paris. ENV/JM/MONO(2012)10.
- OECD (2015) Performance standards for assessment of proposed similar or modified *in vitro* skin sensitisation ARE-NRF2 luciferase test methods. OECD Environment, Health and Safety Publications; Series on Testing and Assessment, No. 213. Organisation for Economic Co-operation and Development, Paris. ENV/JM/MONO(2015)6.
- OECD (2016a) Guidance Document on the reporting of Defined Approaches to be used within Integrated Approaches to Testing and Assessment. OECD Environment, Health and

Safety Publications; Series on Testing and Assessment, No. 255. Organisation for Economic Co-operation and Development, Paris. ENV/JM/MONO(2016)28.

- OECD (2016b) Guidance Document on the reporting of Defined Approaches and individual information sources to be used within Integrated Approaches to Testing and Assessment (IATA) for skin sensitisation. OECD Environment, Health and Safety Publications; Series on Testing and Assessment, No. 256. Organisation for Economic Co-operation and Development, Paris. ENV/JM/MONO(2016)29.
- OECD (2017) Revised performance standards for the assessment of proposed similar or modified *in vitro* Reconstructed human Cornea-like Epithelium (RhCE) test methods for eye hazard. OECD Environment, Health and Safety Publications; Series on Testing and Assessment, No. 216. Organisation for Economic Co-operation and Development, Paris. ENV/JM/MONO(2015)23.
- OECD (2018a) Test Guideline No. 442D - Key Event-Based Test Guideline for *in vitro* skin sensitisation assays addressing the AOP Key Event on Keratinocyte Activation. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.
- OECD (2018b) Test Guideline No. 442E - Key Event-Based Test Guideline for *in vitro* skin sensitisation assays addressing the Key Event on Activation of Dendritic Cells on the Adverse Outcome Pathway for skin sensitisation. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.
- OECD (2018c) Test Guideline No. 442B - Local lymph node assay: BRDU-ELISA or -FCM. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.
- OECD (2021a) Test Guideline No. 442C - Key Event-Based Test Guideline for *in chemico* skin sensitisation assays addressing the Adverse Outcome Pathway Key Event on Covalent Binding to Proteins. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.
- OECD (2021b) Test Guideline No. 406 - Skin Sensitisation Guinea Pig Maximisation Test and Buehler Test. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.
- OECD (2021c) Guideline No. 497 - Guideline on Defined Approaches for Skin Sensitisation. OECD Guidelines for the Testing of Chemicals, Section 4, Health effects. Organisation for Economic Co-operation and Development, Paris.
- SCCS (Scientific Committee on Consumer Safety) (2019) (SCCS) Opinion on salicylic acid (CAS 69-72-7). Submission I, SCCS/1601/18, preliminary version of 10 September 2018, final version of 21 December 2018, Corrigendum on 20-21 June 2019. Available at: https://op.europa.eu/en/publication-detail/-/publication/75be19bf-5390-11ea-aece-01aa75ed71a1 (accessed on 06/07/2021).
- UN (2021) Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Ninth revised edition, United Nations, New York and Geneva. ST/SG/AC.10/30/Rev.9.
- Zeller K.S., Forreryd A., Lindberg T., Gradin R., Chawade A., Lindstedt M. (2017) The GARD platform for potency assessment of skin sensitizing chemicals. *ALTEX* 34(4):539-559. doi: 10.14573/altex.1701101.

Appendix I. Verification of the GARDskin and GARDpotency models by the ESAC

One step of the peer-review of the validation study of the GARD platform consisted of the replication of the Decision Values (DV) that were reported for each compound of the ring trial in the file *"GARDskin and GARDpotency data.xlsx"*.

This verification process mainly consisted of:

- implementation of the GARDskin and GARDpotency models offline
    - training of the GARDskin and GARDpotency Support Vector Machine (SVM) models (generation of SVM weights)
- processing of the nanoString raw data
- prediction of the DV for each laboratory and experiment

The raw data used for the process and the DV used as reference were those submitted to EURL ECVAM as annexes to the GARD submission.

1 Implementation of GARDskin and GARDpotency

The GARDskin and GARDpotency models are implemented in the GARD Data Analysis and Application (GDAA) cloud platform, which consists of an online shinyapps.io application coded in R. This app uses the gene transcripts generated with the nanoString platform for each test chemical to predict its skin sensitisation hazard and/or potency. The prediction step is carried out by applying the pre-trained SVM model on the normalised and quality-checked input data. Therefore, before the application of the SVM model, the app carries out a quality check on the input data in order to identify invalid runs that may lead to low quality predictions. The samples that pass the quality check steps undergo subsequent normalisation to adequate the data for its use in the SVM model. These normalisation steps reduce variability within genes and possible batch-to-batch effects (Gradin et al., 2019). A representation of the GARD workflow is shown in Figure 1.

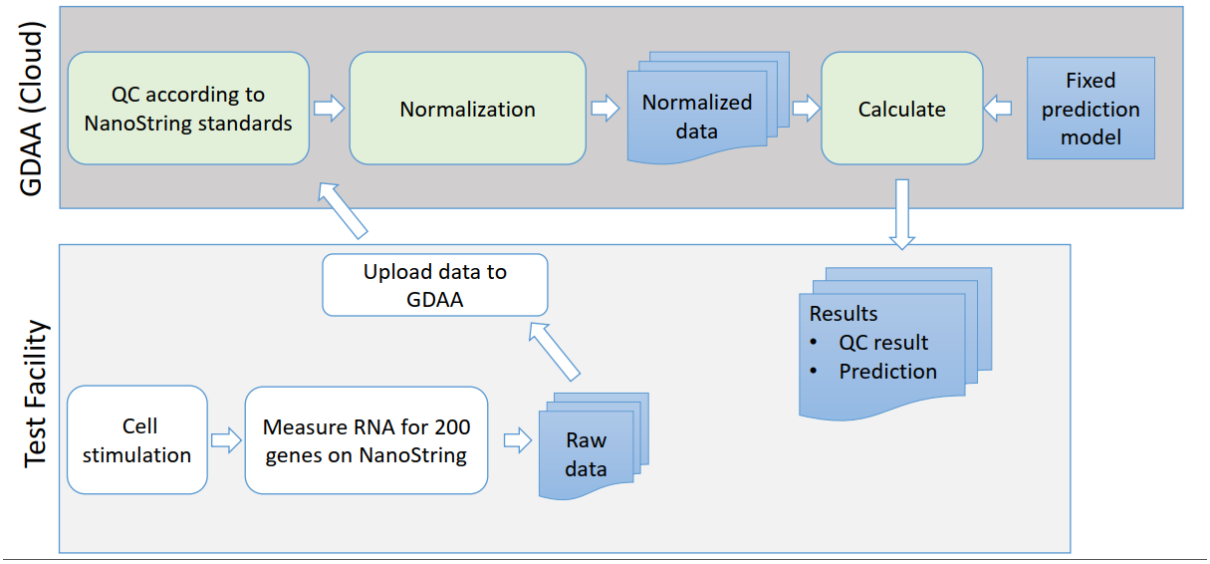


Figure 1. Representation of the GARD workflow.

The ESAC was given access to the GDAA platform but this does not allow the verification of the code that runs the application as it is not accessible to the user and is IP protected. In order to perform a deeper evaluation, the ESAC requested to SenzaGen the training data and the R code that were used to train the SVM models (GARDskin and GARDpotency) as well as the parameters (weights) of the trained models.

SenzaGen kindly provided the training set files necessary to train the model, the R code to normalise and train the model, and accompanying documentation with detailed information on how to perform each of the steps, which also included the corresponding mathematical equations. In addition, the parameters of the trained SVM models as well as a script to compare these to the SVM parameters obtained by the ESAC was provided.

By using the code provided, the accompanying data, and thanks to some clarifications answered by SenzaGen, the ESAC was capable of replicating:

- SVM model parameters for GARDskin
- SVM model parameters for GARDpotency

With this exercise, the ESAC considered as verified the implementation of the GARDskin and GARDpotency SVM models.

2 Verification of the ring trial results

In order to determine the within- and between- laboratory reproducibility of GARDskin and GARDpotency, a ring trial exercise with 28 blinded chemicals and 3 independent laboratories was organised (Johansson et al. 2019). The ring trial exercise was carried out by the laboratories SenzaGen, Burleson (BRT), and Eurofins. Skin sensitisation hazard or potency predictions for a chemical are made based on the results of at least two (typically three) independent "main stimulations" (i.e., biological replicates using different batches of cells). The group of main stimulations performed to obtain one prediction is called a GARD campaign or experiment. Each laboratory performed three GARD campaigns/experiments for each test chemical in order to evaluate the GARDskin and GARDpotency reproducibility. Each of these campaigns should contain valid results for two or three main stimulations (a maximum of 5 main stimulations was allowed as per the validation study to obtain 3 valid results). This means that each chemical was tested between 9 and 15 times by each laboratory. In the ring-trial, the data for each main stimulation (biological replicate) were used as input for the GDAA platform to derive the DV by using the trained SVM models (see Forreryd et al., 2016; Gradin et al., 2020; Johansson et al., 2011; Zeller et al., 2017). In the GARD protocols version 5.01 and newer, the final prediction for each GARD campaign/experiment is obtained by combining the DVs of its biological replicates into a single value. This "combined" single value is used to determine the final outcome (Sensitiser/Non-Sensitisers for GARDskin and Cat. 1A/Cat. 1B for GARDpotency). According to protocol v5.01 and newer, the mean of the biological replicates is used in the GARDskin, whereas for the GARDpotency the median is used. Mean DV ≥ 0 correspond to skin sensitisers in the GARDskin, and median DV ≥ 0 correspond to Cat. 1A sensitisers in the GARDpotency. To replicate the model, the ESAC calculated the DV of each replicate for GARDskin and GARDpotency as well as their means and medians, respectively. The results obtained are shown below for each laboratory and experiment.

It is important to note that in order to use the GARDskin or GARDpotency models, it is not enough to have the SVM model and the raw data of the test chemical. This is because the normalisation step, named batch adjustment reference alignment (BARA), which eliminates batch-to-batch effects (see Gradin et al. 2019), uses data of the training set that were used to train the SVM. Therefore, it is also necessary to have the training set data in order to generate GARD predictions.

Unlike the GDAA platform, the GARDskin and GARDpotency replication performed by the ESAC did not initially include a quality check step, as it was assumed that the data submitted contained only the files of tests that had passed the quality check.

The presence of some errors and typos in some files submitted to the ESAC were identified during this verification process. These are explained below together with the results of the verification process for each laboratory and campaign/experiment.

2.1 GARDskin

The mean DV for each main stimulation calculated by the ESAC are shown below in scatter plots that compare those calculated values to the mean DV provided in the submission.

2.1.1 GARDskin SenzaGen – Exp1 assessment

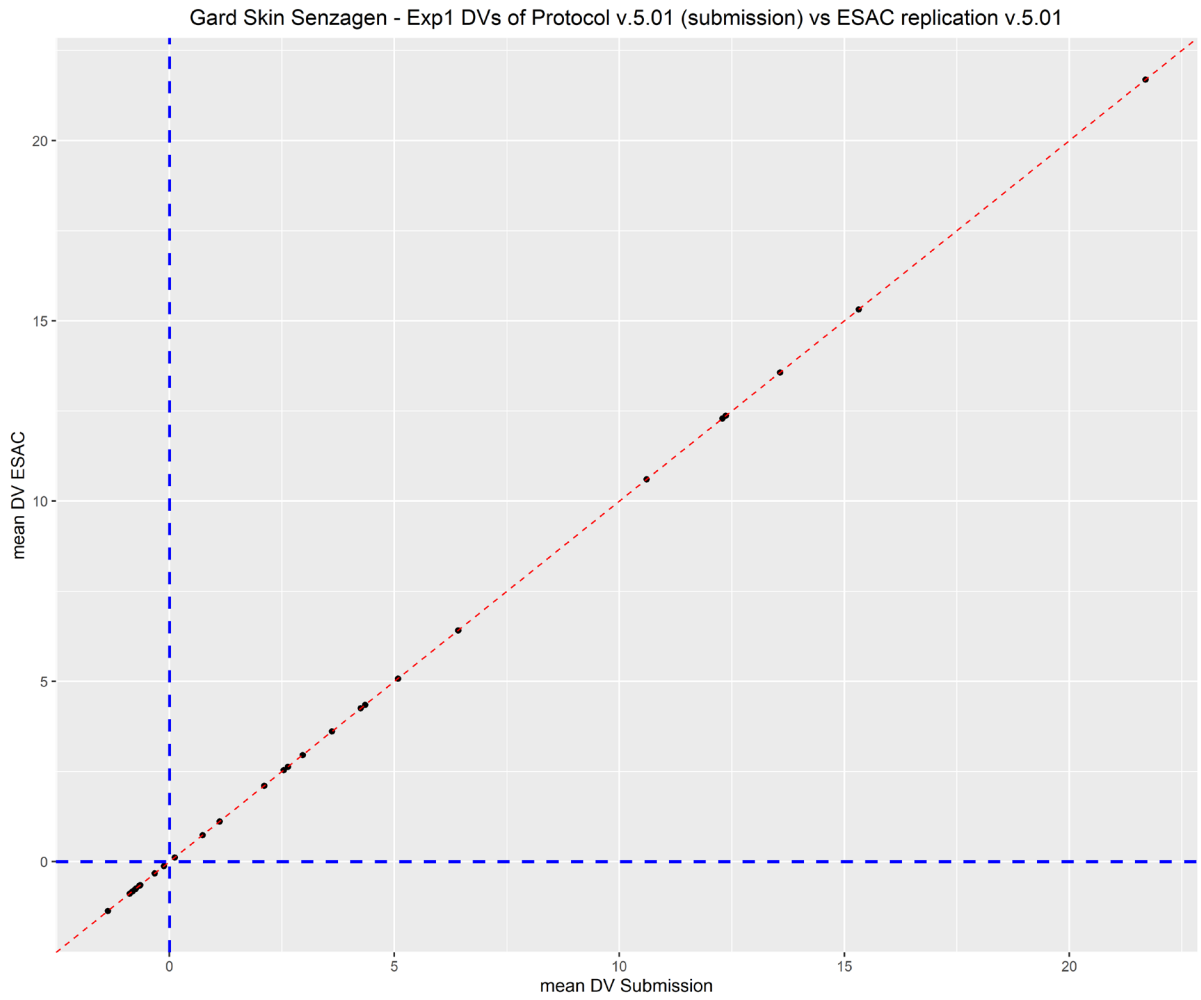


Figure 2. Scatter plot of the mean DV calculated by the ESAC for SenzaGen Experiment 1 data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The mean DV for all the chemicals could be reproduced with a precision of at least 10E-4.

2.1.2 GARDskin SenzaGen – Exp2 assessment

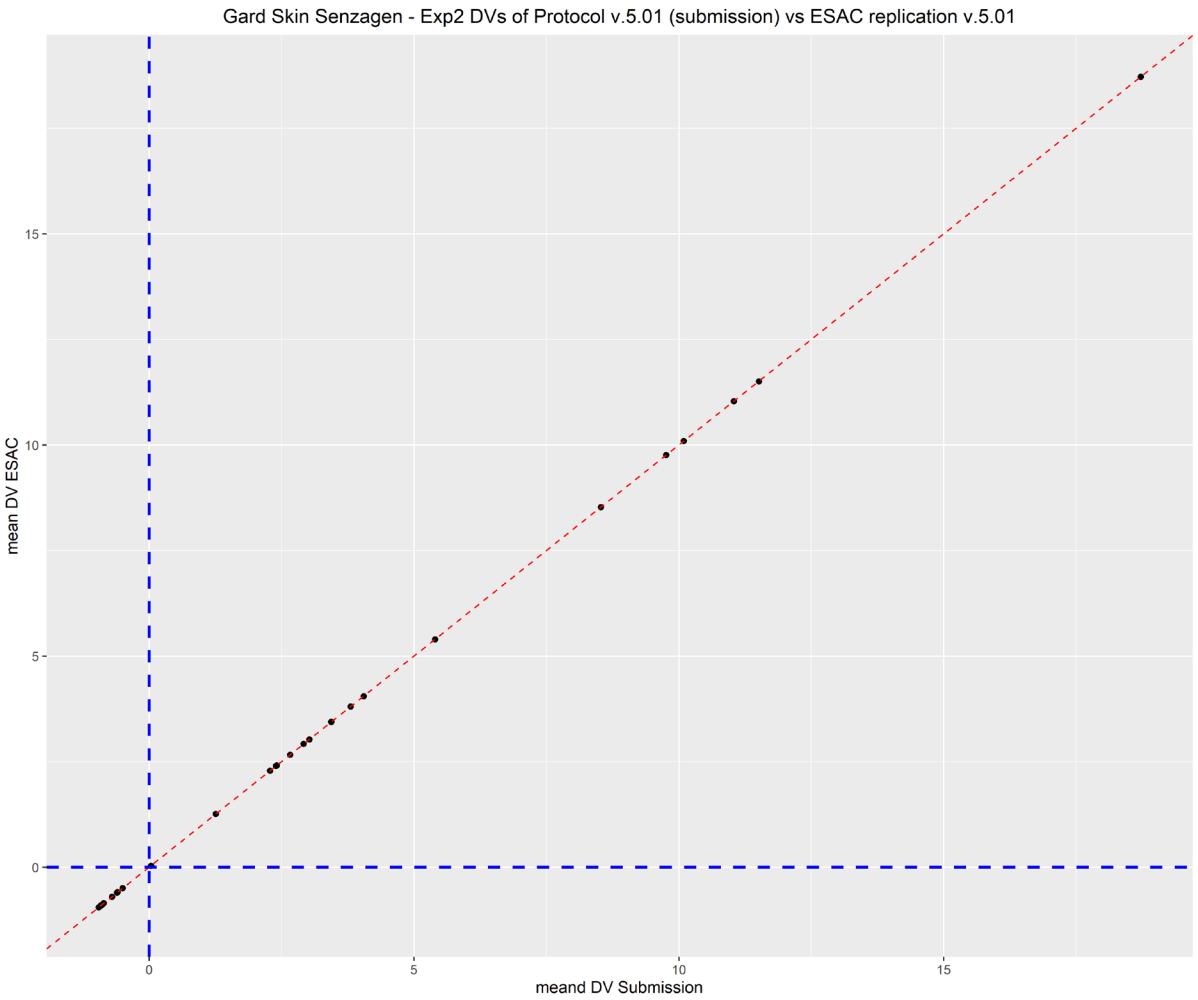


Figure 3. Scatter plot of the mean DV calculated by the ESAC for SenzaGen Experiment 2 data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

All chemicals could be reproduced with a precision of at least 10E-4.

2.1.3 GARDskin SenzaGen – Exp3 assessment

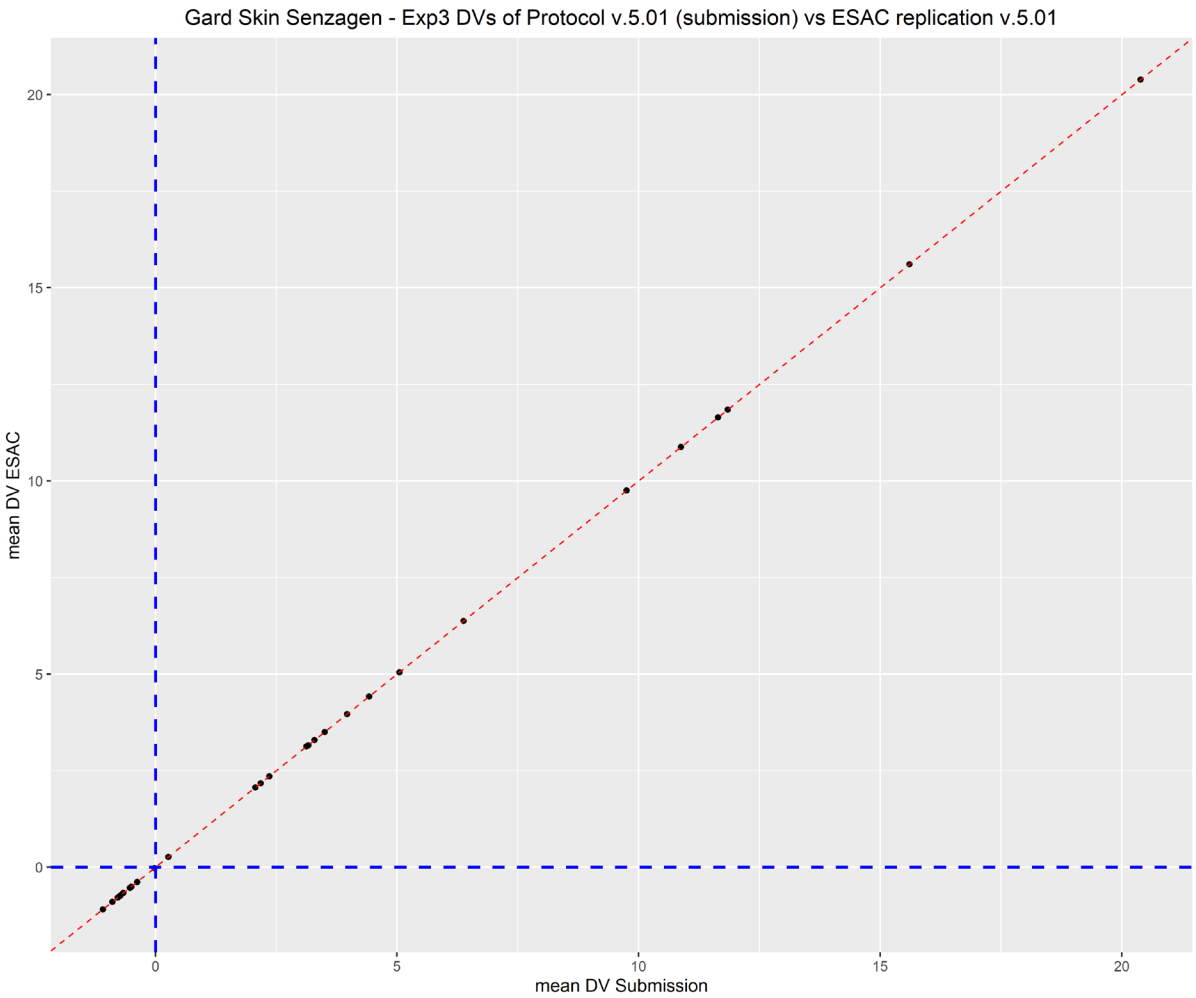


Figure 4. Scatter plot of the mean DV calculated by the ESAC for SenzaGen Experiment 3 data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

All chemicals could be reproduced with a precision of at least 10E-4.

2.1.4 GARDskin SenzaGen Extra – ExpX assessment

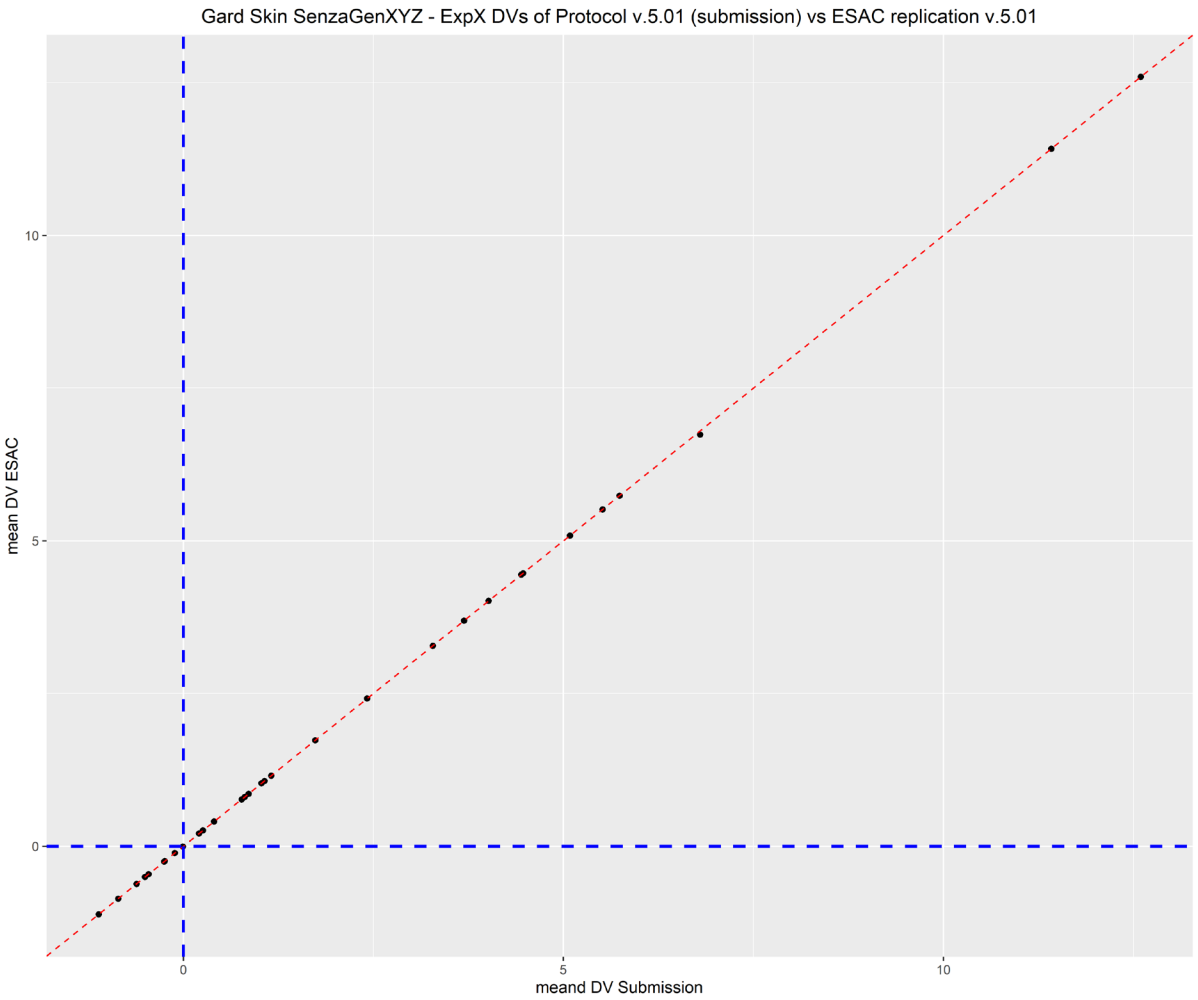


Figure 5. Scatter plot of the mean DV calculated by the ESAC for SenzaGen Extra Experiment X data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

Only one chemical, p-benzoquinone (X-132) could not be reproduced with the desired accuracy (this cannot be observed in the plot). However, it is explained in the Excel file with the data that a third main stimulation for this chemical was performed together with the Y experiments, and that the final mean DV was obtained from these 3 values. Once the third value is added, the reported mean DV can be replicated with the desired precision.

2.1.5 GARDskin SenzaGen Extra – ExpY assessment

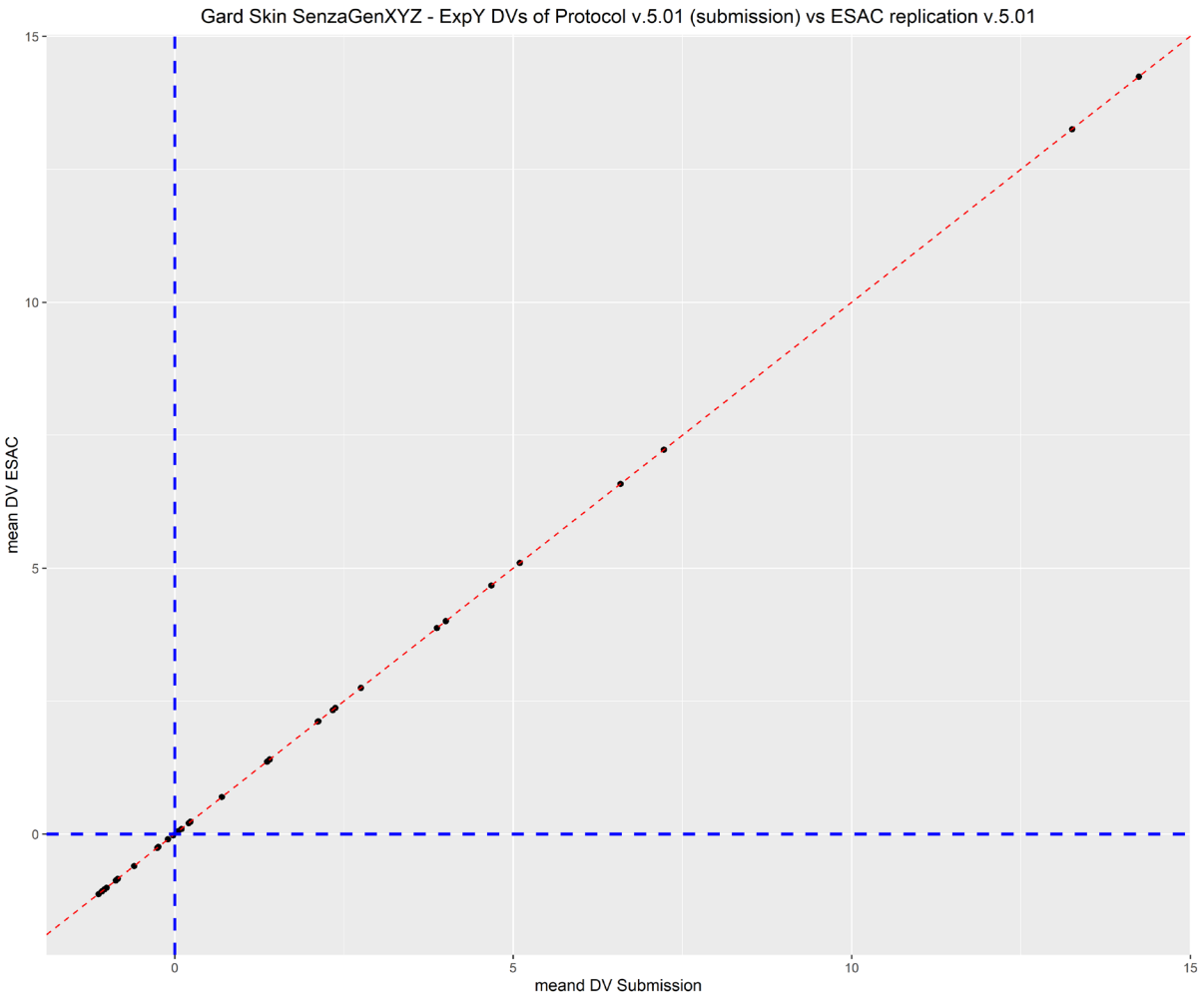


Figure 6. Scatter plot of the mean DV calculated by the ESAC for SenzaGen Extra Experiment Y data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

All chemicals could be reproduced with a precision of at least 10E-4.

2.1.6 GARDskin SenzaGen Extra – ExpZ assessment

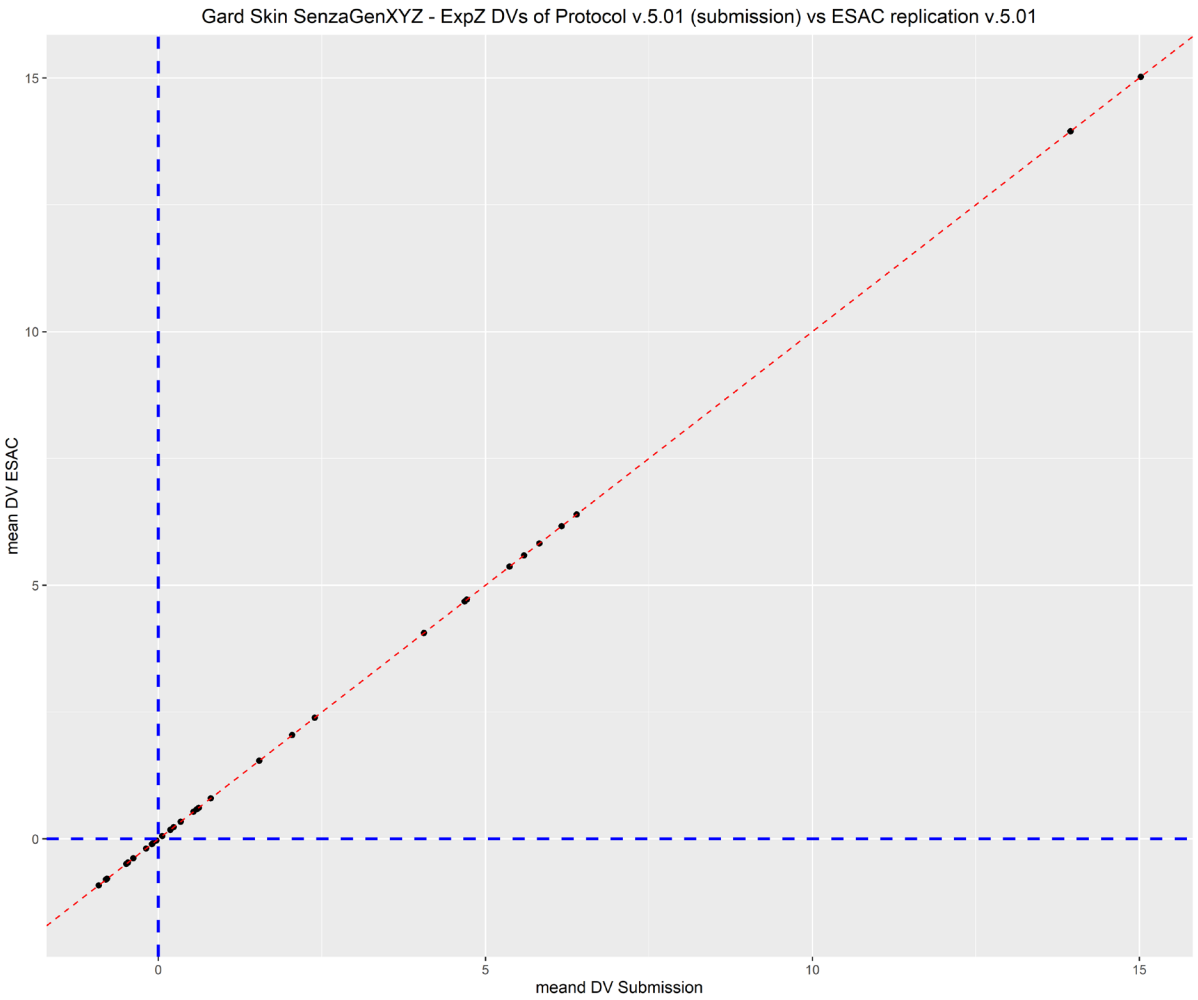


Figure 7. Scatter plot of the mean DV calculated by the ESAC for SenzaGen Extra Experiment Z data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

All chemicals could be reproduced with a precision of at least 10E-4.

2.1.7 GARDskin BRT – Exp1 assessment

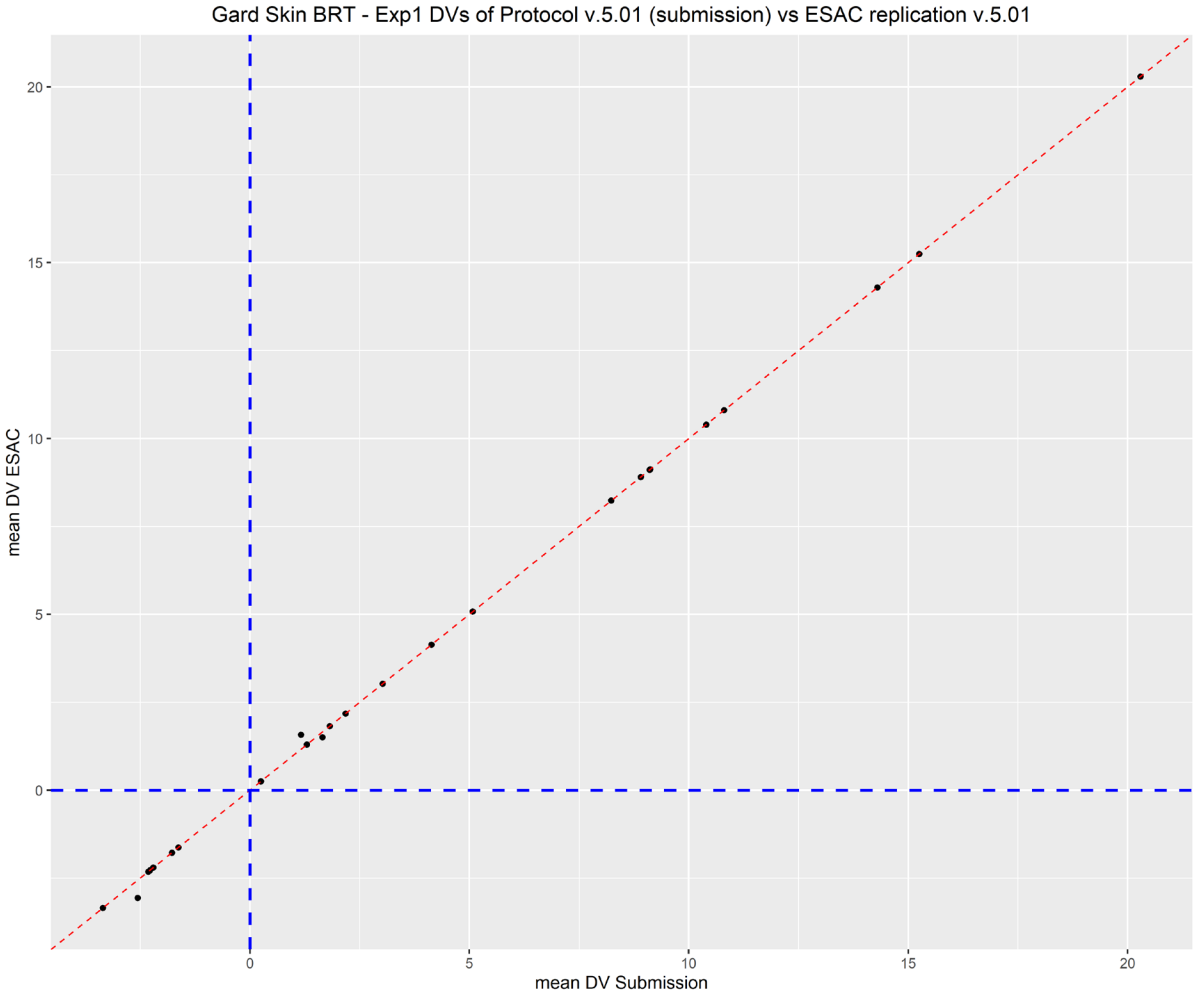


Figure 8. Scatter plot of the mean DV calculated by the ESAC for BRT Experiment 1 data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

Three chemicals could not be reproduced with the desired accuracy. These chemicals can be observed in the off-diagonal of the scatter plot figure. The exact values obtained are shown in Table 1 below.

Table 1. Summary table of chemicals for which the mean DV provided in the submission could not be reproduced with the data submitted as such.

Chemical name	Substance ID	Submitted mean DV	ESAC mean DV
Kanamycin	B165	-2.558523	-3.059923
Propylene glycol	B176	1.652483	1.506228
Toluene diamine sulphate	B24	1.166126	1.57311

After further analysis, these 3 chemicals were found to contain main stimulations that did not pass the QC in the GDAA. Once removed, the calculated mean DV matched those reported (see figure below).

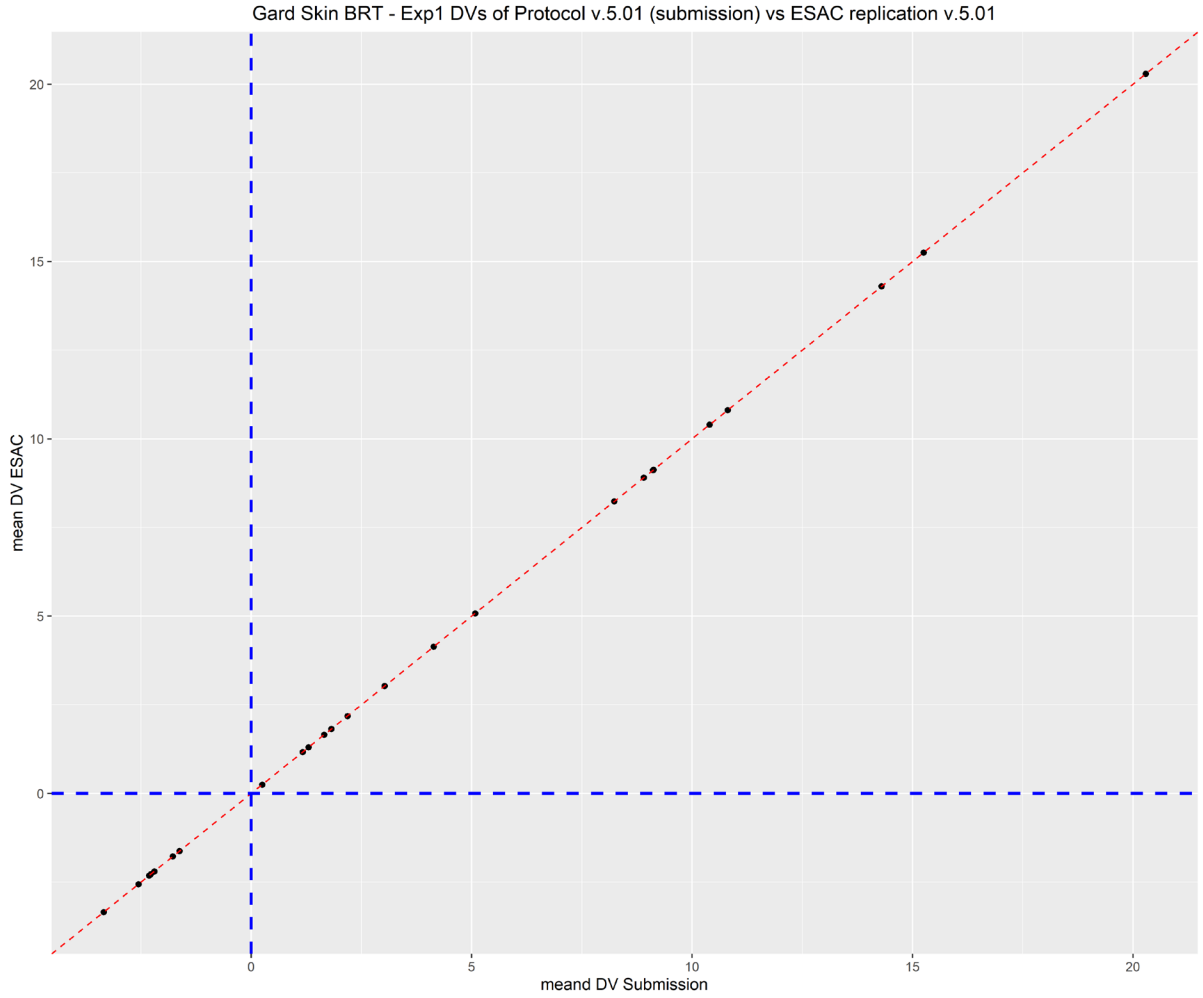


Figure 9. Scatter plot of the mean DV calculated by the ESAC for BRT Experiment 1 data after removal of the samples that did not pass the quality check vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

2.1.7.1 Other observations

Only one main stimulation for Methylisothiazolinone (B93) was provided in the submission. The calculated DV corresponded to the one reported. According to the acceptance criteria described in the Standard Operating Procedure (SOP), this prediction should not be valid as at least 2 valid main stimulations are needed for a prediction to be valid.

4-(Methylamino)phenol sulphate (B183) was reported to have 2 valid QC main stimulations but only one was found in the files provided. The calculated DV for this single file corresponded to the DV reported. As in the previous case, this prediction should not be valid as at least 2 valid main stimulations are needed for a prediction to be valid.

2.1.8 GARDskin BRT – Exp2 assessment

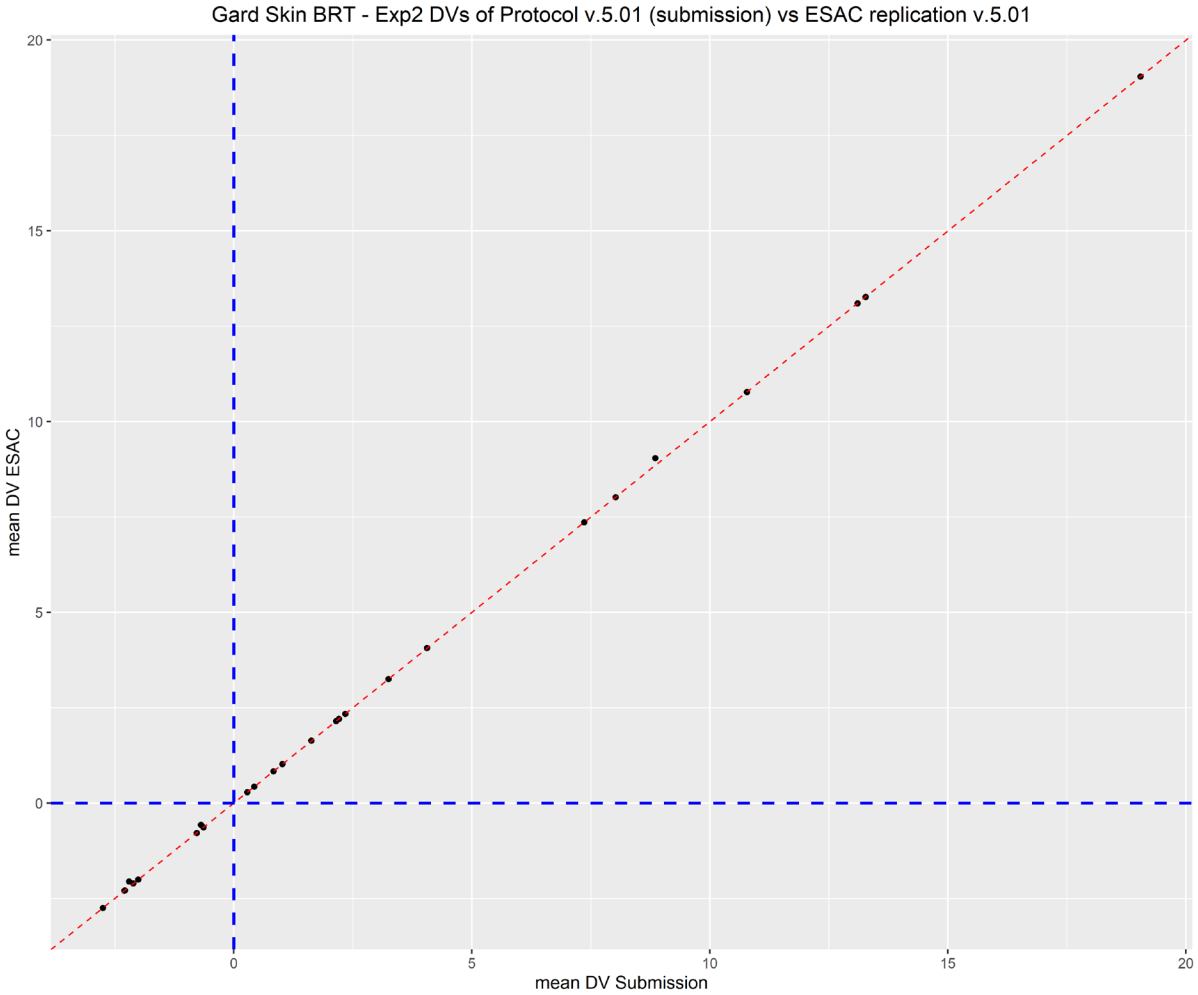


Figure 10. Scatter plot of the mean DV calculated by the ESAC for BRT Experiment 2 data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

Twenty-six chemicals could not be reproduced with the desired accuracy because some of the reported DVs had been truncated/rounded to 2 decimal points. Nevertheless, out of these 26 chemicals, only 3 had differences >10E-1, which did not affect the classification of the substances.

As in the previous case, the 3 chemicals reported in Table 2 were found to contain main stimulations that did not pass the QC in the GDAA. Once removed, the DV match was perfect.

Table 2. Summary table of chemicals for which the mean DV provided in the submission could not be reproduced with the data submitted as such.

Chemical name	Substance ID	Submitted mean DV	ESAC mean DV
Propylene glycol	B336W	-0.69	-0.5743
Lactic acid	B347W	-2.20	-2.0574
Diethyl maleate	B486W	8.85	9.0447

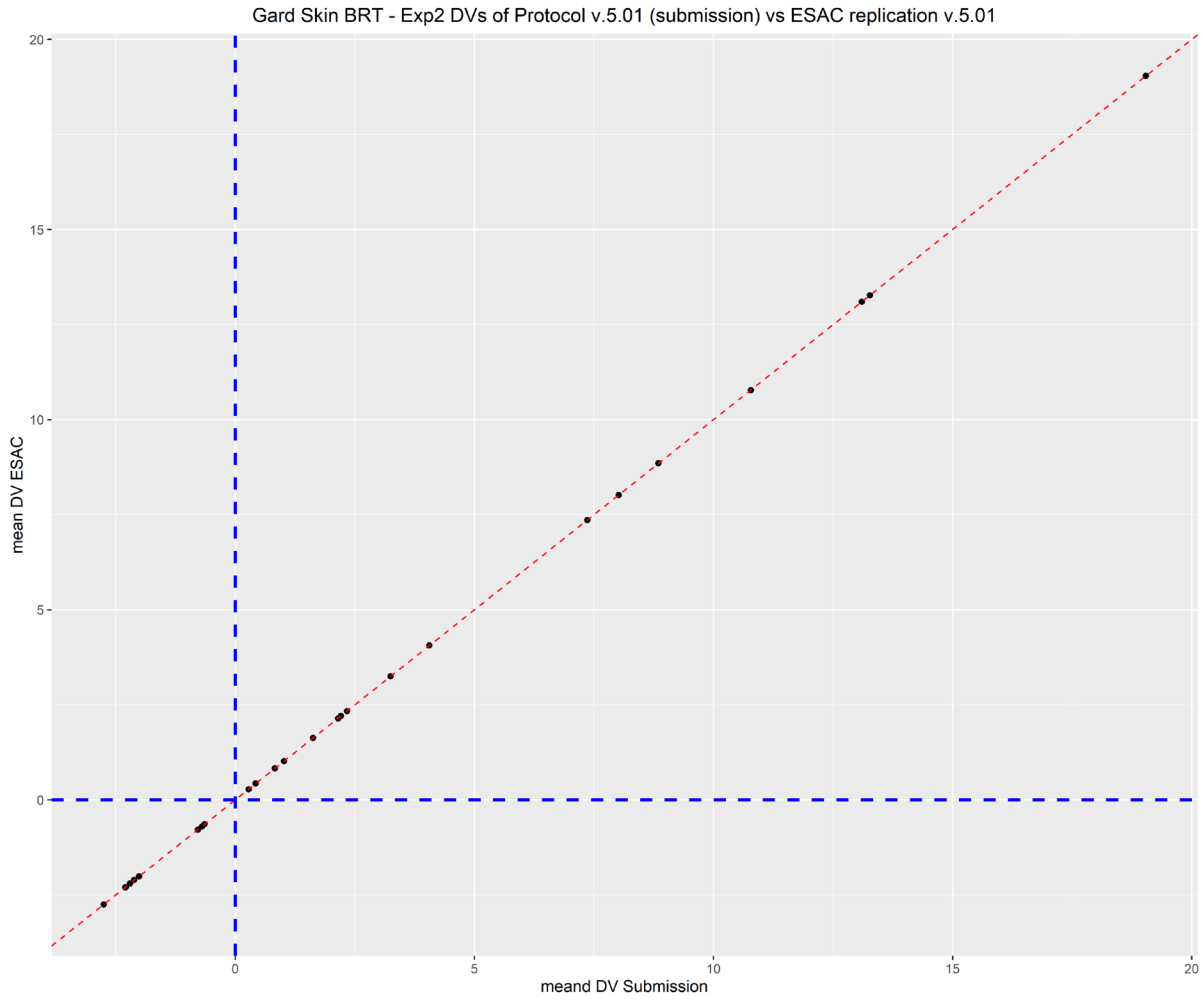


Figure 11. Scatter plot of the mean DV calculated by the ESAC for BRT Experiment 2 data after removal of the samples that did not pass the quality check vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

2.1.8.1 Other observations

As reported in the submission file, only one sample for 2-Bromo-2-glutaronitrile passed the QC and consequently only one main stimulation was provided. The calculated DV corresponded to the one reported. According to the SOP, this prediction should not be valid as at least 2 valid main stimulations are needed for a valid prediction.

2.1.9 GARDskin BRT – Exp3 assessment

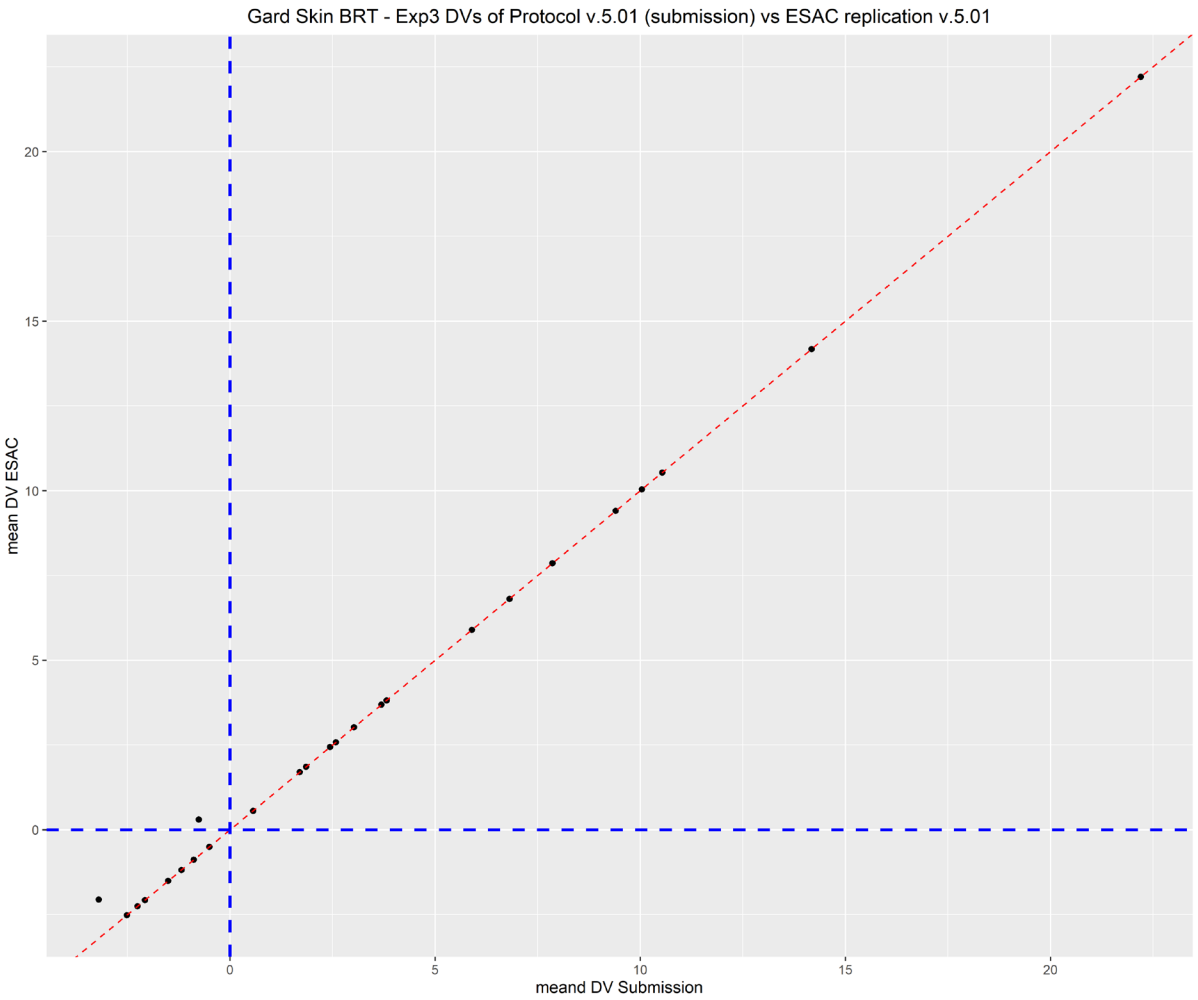


Figure 12. Scatter plot of the mean DV calculated by the ESAC for BRT Experiment 3 data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

Three chemicals could not be reproduced with the desired accuracy. These chemicals can be observed in the off-diagonal of the scatter plot figure. The exact values obtained are shown in Table 3 below.

Table 3. Summary table of chemicals for which the mean DV provided in the submission could not be reproduced with the data submitted as such.

Chemical name	Substance ID	Submitted mean DV	ESAC mean DV
Vanillin	B641	-3.2015304	-2.0550288
Propylene glycol	B738W	-0.7605858	0.3050136
Kanamycin	B871W	-2.5115809	-2.5175854

Propylene glycol is the most worrying case of all as the mean DV obtained by the ESAC corresponds to a sensitiser, while the mean DV submitted to the ESAC corresponds to a non-sensitiser. However, as in the other cases, these 3 substances have main stimulations that did not pass the QC in the GDAA. Once removed, the match was perfect as can be observed below.

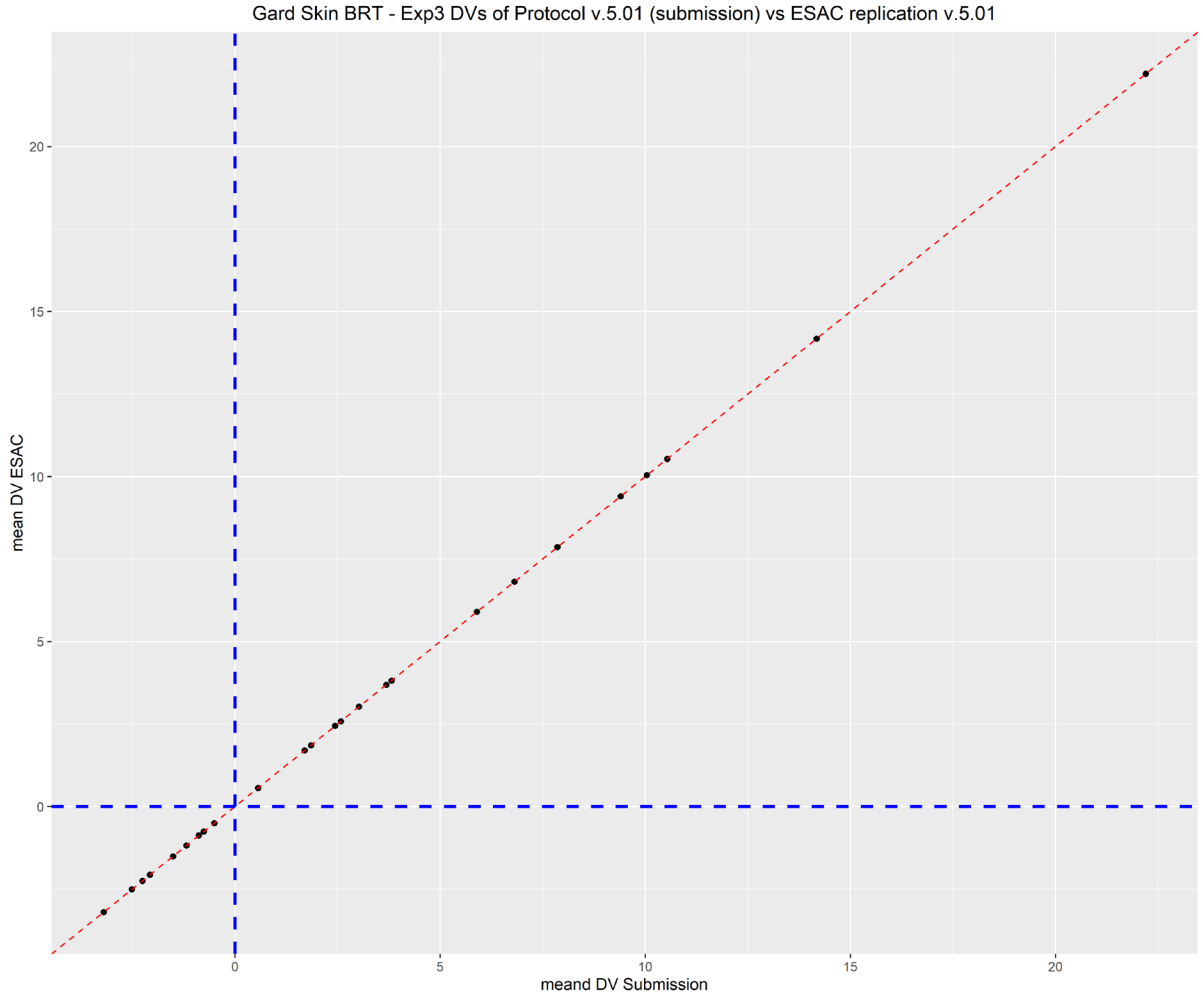


Figure 13. Scatter plot of the mean DV calculated by the ESAC for BRT Experiment 3 data after removal of the samples that did not pass the quality check vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

2.1.10 GARDskin Eurofins – Exp1 assessment

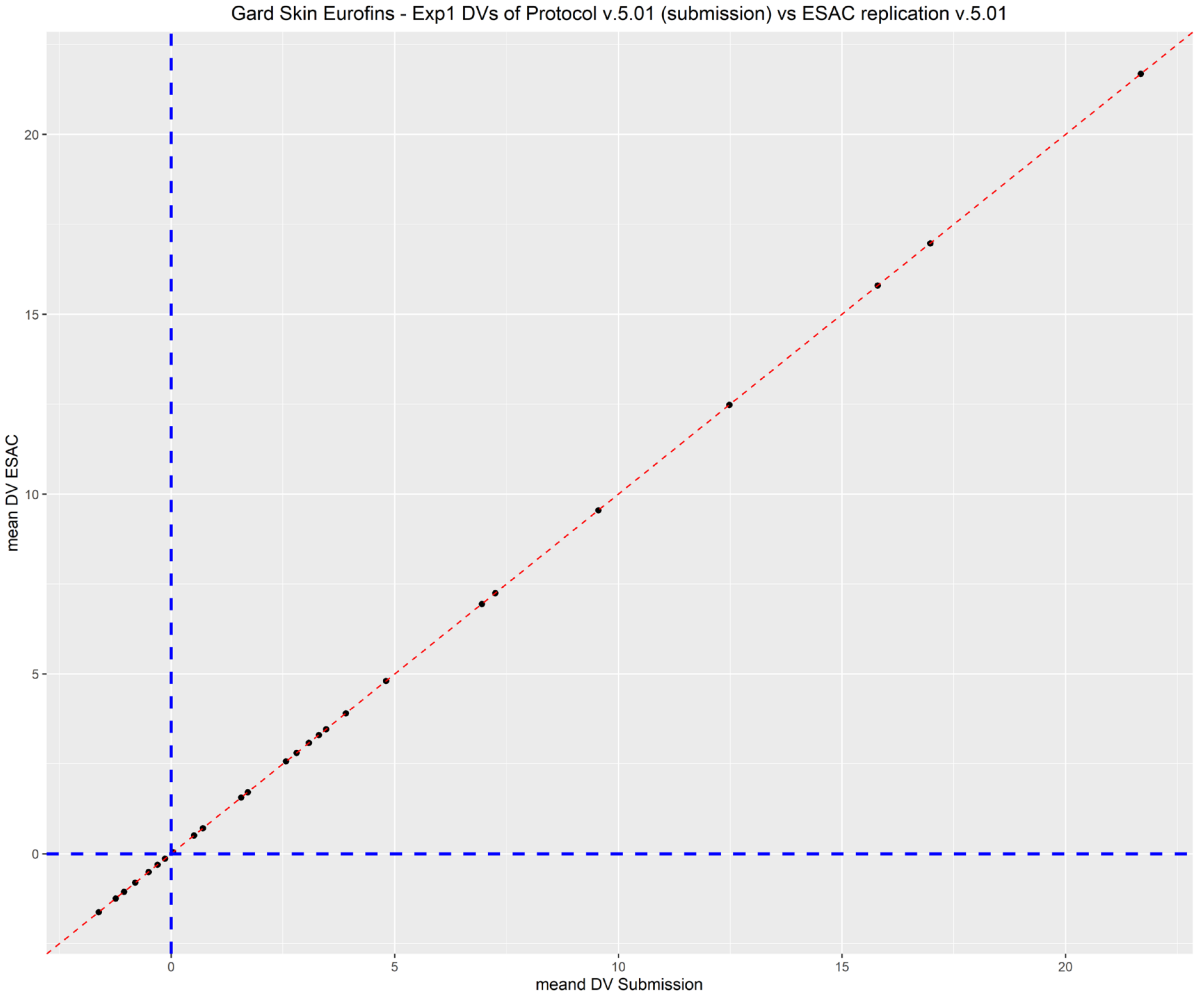


Figure 14. Scatter plot of the mean DV calculated by the ESAC for Eurofins Experiment 1 data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

All chemicals could be reproduced with a precision of at least 10E-4.

2.1.10.1 Other observations

The data provided for Vanillin had a repeated test (i.e., same sample id with two different files obtained in different days, files 20170630_sg9_Eurofins1-67_12.RCC and 20170628_sg5_Eurofins1-67_12.RCC). Vanillin turned out to be a particularly difficult chemical as there were 4 biological replicates (main stimulations), 2 with negative DVs and 2 with positive DVs but one of those was a "repetition" according to GDAA. It is not clear why there was a repetition for this chemical as the two files with the same id mentioned above passed the QC check in the GDAA. In fact, the submitted value of -0.30399 could only be replicated if these two repetitions were kept.

2.1.11 GARDskin Eurofins – Exp2 assessment

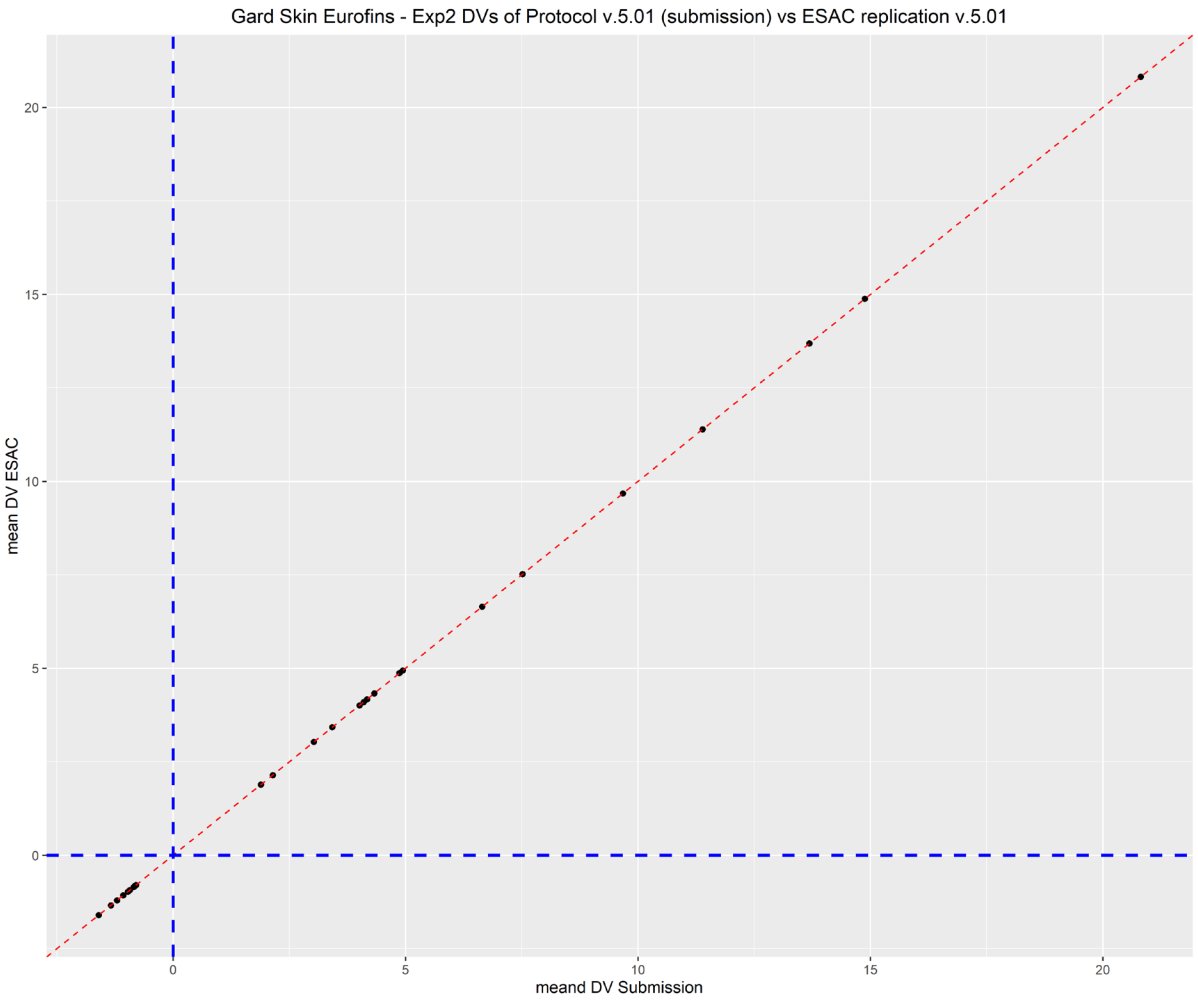


Figure 15. Scatter plot of the mean DV calculated by the ESAC for Eurofins Experiment 2 data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

All chemicals could be reproduced with a precision of at least 10E-4.

2.1.12 GARDskin Eurofins – Exp3 assessment

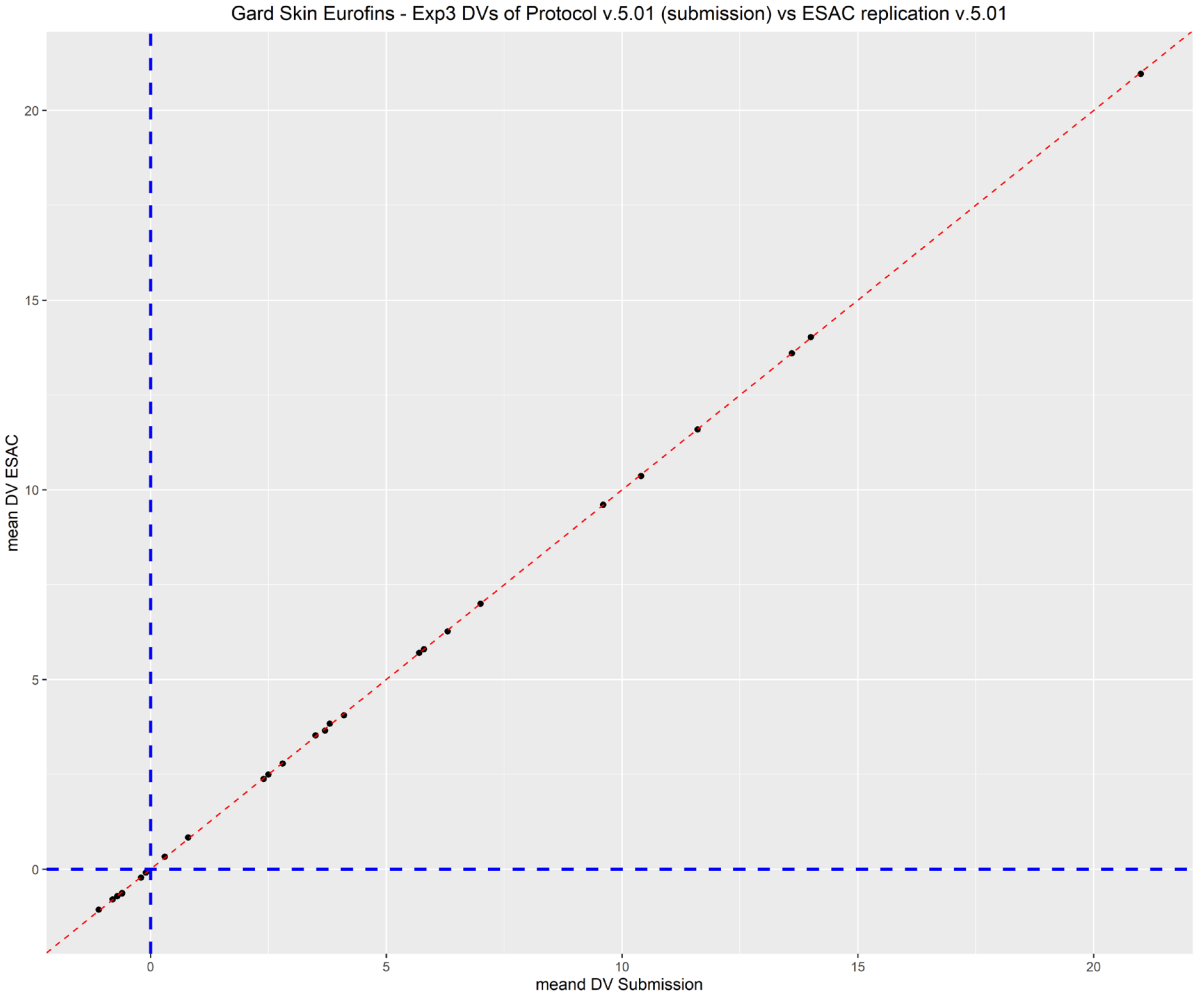


Figure 16. Scatter plot of the mean DV calculated by the ESAC for Eurofins Experiment 3 data vs mean DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

Twenty-seven chemicals could not be reproduced with the desired accuracy because some of the reported DVs had been truncated/rounded to 2 decimal points. Nevertheless, none of them had differences >10E-1.

2.2 GARDpotency

In this section, the verification of the GARDpotency model is described. As with the GARDskin, the ESAC tried to replicate the DVs for each laboratory and campaign/experiment. However, the SOP indicates that the predictions for GARDpotency are obtained using the median of the main stimulations (biological replicates) instead of the mean that was used in GARDskin. Therefore, the medians of the DVs obtained for each biological replicate are shown below in plots comparing the median DVs calculated by the ESAC to the ones provided in the submission.

The GARDpotency uses a different set of genes than the GARDskin and also uses the test concentration as input parameter. This parameter is very important in GARDpotency as it has one of the highest weights in the SVM GARDpotency model. Therefore, it must be reported accurately and with the right units. The reporting of the concentration is done in the Annotation file, which relates the sample IDs with the chemical names, including the reference set chemicals, i.e., positive control, negative control, and unstimulated controls (benchmark controls).

It is important to note that the benchmark chemicals used for the batch-to-batch removal step (i.e., BARA normalisation) for GARDpotency are not the “unstimulated ctrl” that are used in the GARDskin. This is because the potency classification is only carried out for sensitisers, which stimulate the cells and therefore, the benchmark is done with chemicals that can stimulate the cells. For this reason, in the GARDpotency, the negative controls are used as benchmark controls. In order to avoid modification of the code, SenzaGen opted for modifying the annotation file by naming the negative controls of the annotation file as “unstim ctrl”. As a consequence, in the GARDpotency reference set there are only outputs for the positive controls, as the negative controls are used as benchmark controls, which are only used for the normalisation but not calculated. Please note that the DV of the positive and negative controls have not been reported in the GARDskin section and are not reported here either.

2.2.1 GARDpotency SenzaGen – Exp1 assessment

The median DV calculated by the ESAC are shown in the plot below against the reported median DV.

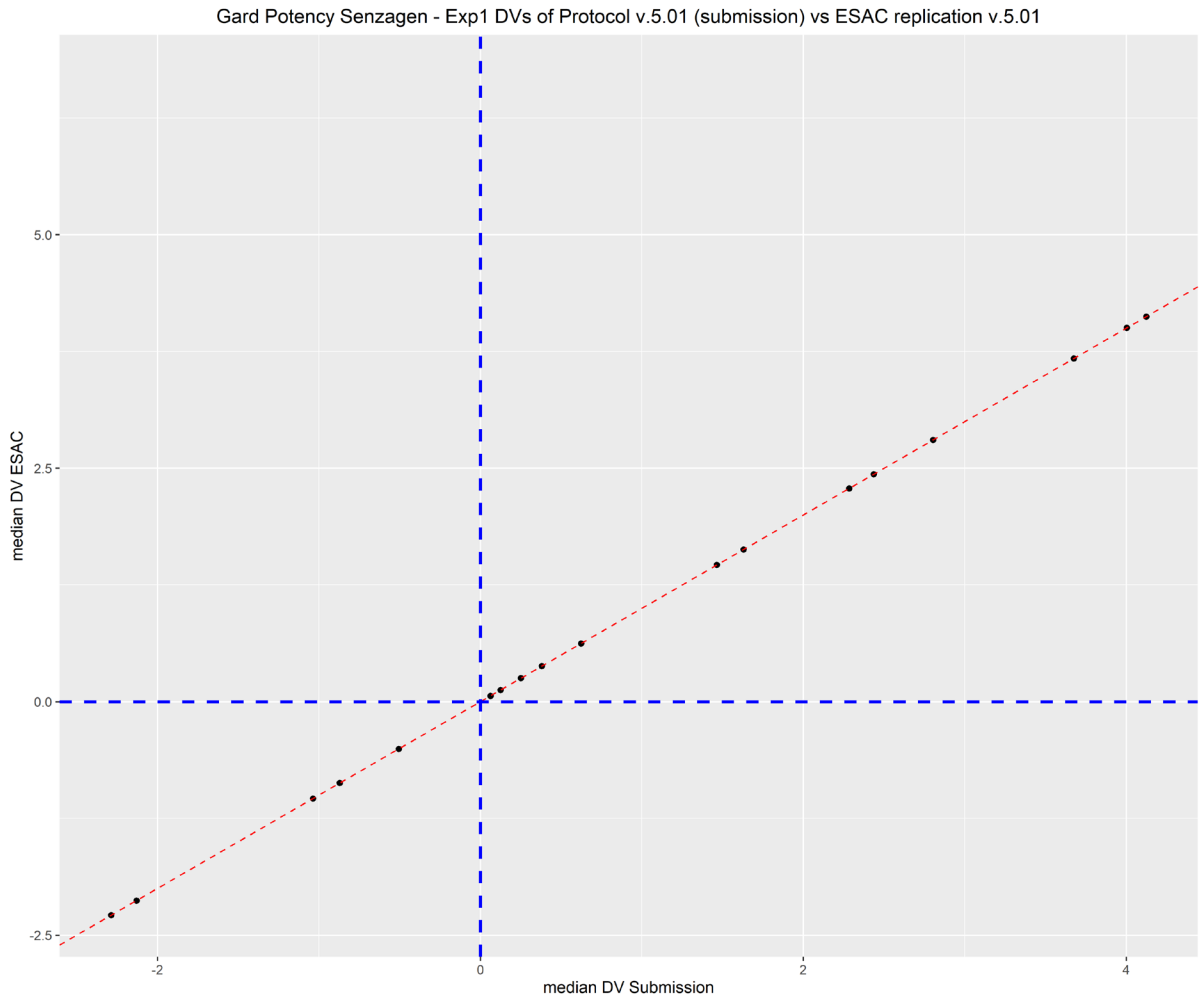


Figure 17. Scatter plot of the median DV calculated by the ESAC for SenzaGen Experiment 1 data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The median DV for all the chemicals could be reproduced with a precision of at least 10E-4.

2.2.2 GARDpotency SenzaGen – Exp2 assessment

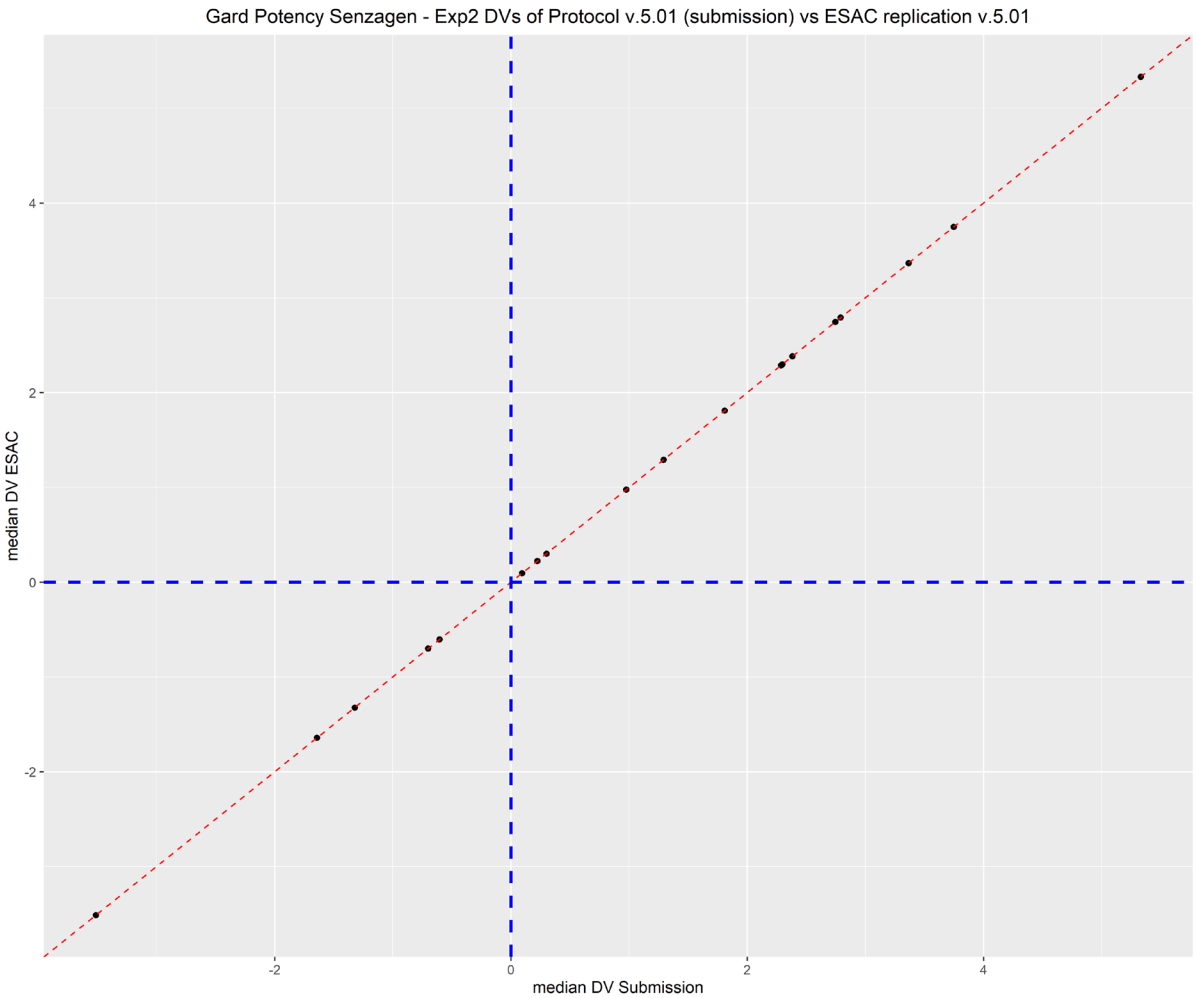


Figure 18. Scatter plot of the median DV calculated by the ESAC for SenzaGen Experiment 2 data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The median DV for all the chemicals could be reproduced with a precision of at least 10E-4.

2.2.3 GARDpotency SenzaGen – Exp3 assessment

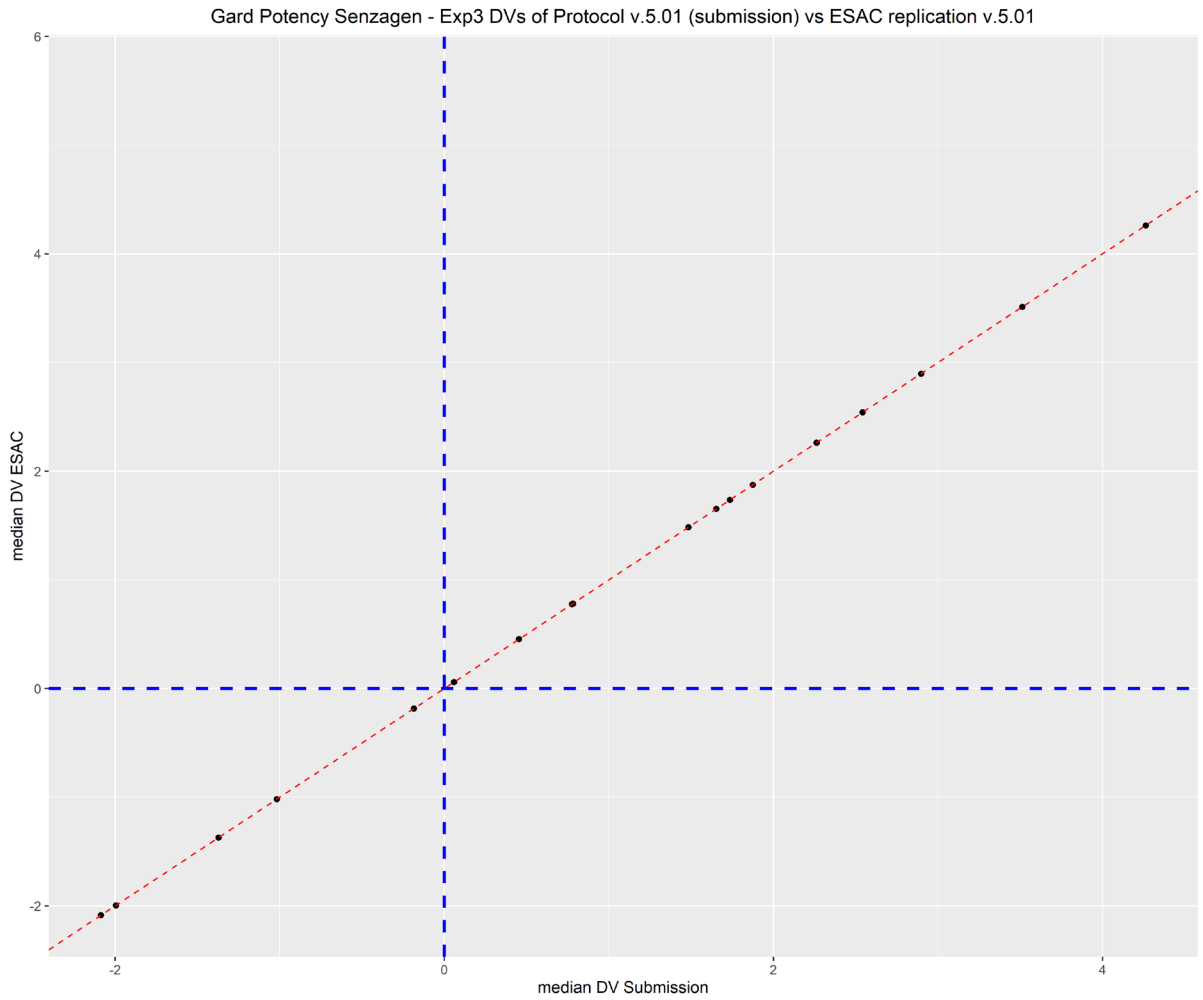


Figure 19. Scatter plot of the median DV calculated by the ESAC for SenzaGen Experiment 3 data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The median DV for all the chemicals could be reproduced with a precision of at least 10E-4.

2.2.4 GARDpotency SenzaGen Extra – ExpX assessment

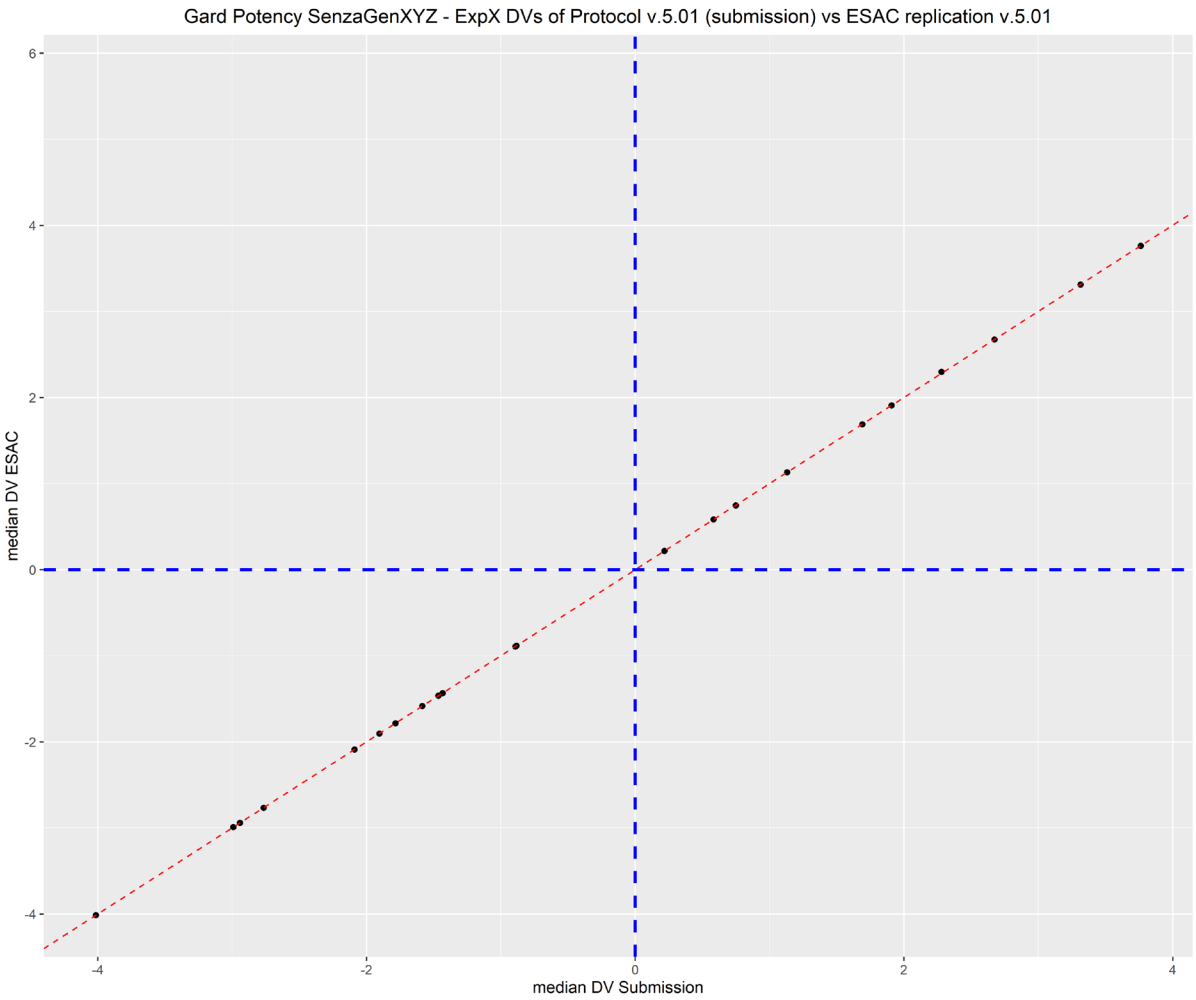


Figure 30. Scatter plot of the median DV calculated by the ESAC for SenzaGen Extra Experiment X data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

As in the case of GARDskin, only p-benzoquinone (X-132) could not be reproduced with the desired accuracy (this cannot be observed in the plot). However, it is explained in the Excel file with the data that a third main stimulation for this chemical was performed together with the Y experiments, and that the final median DV was obtained from these 3 values. Once the third value is added, the reported median DV can be replicated with the desired precision.

2.2.5 GARDpotency SenzaGen Extra – ExpY assessment

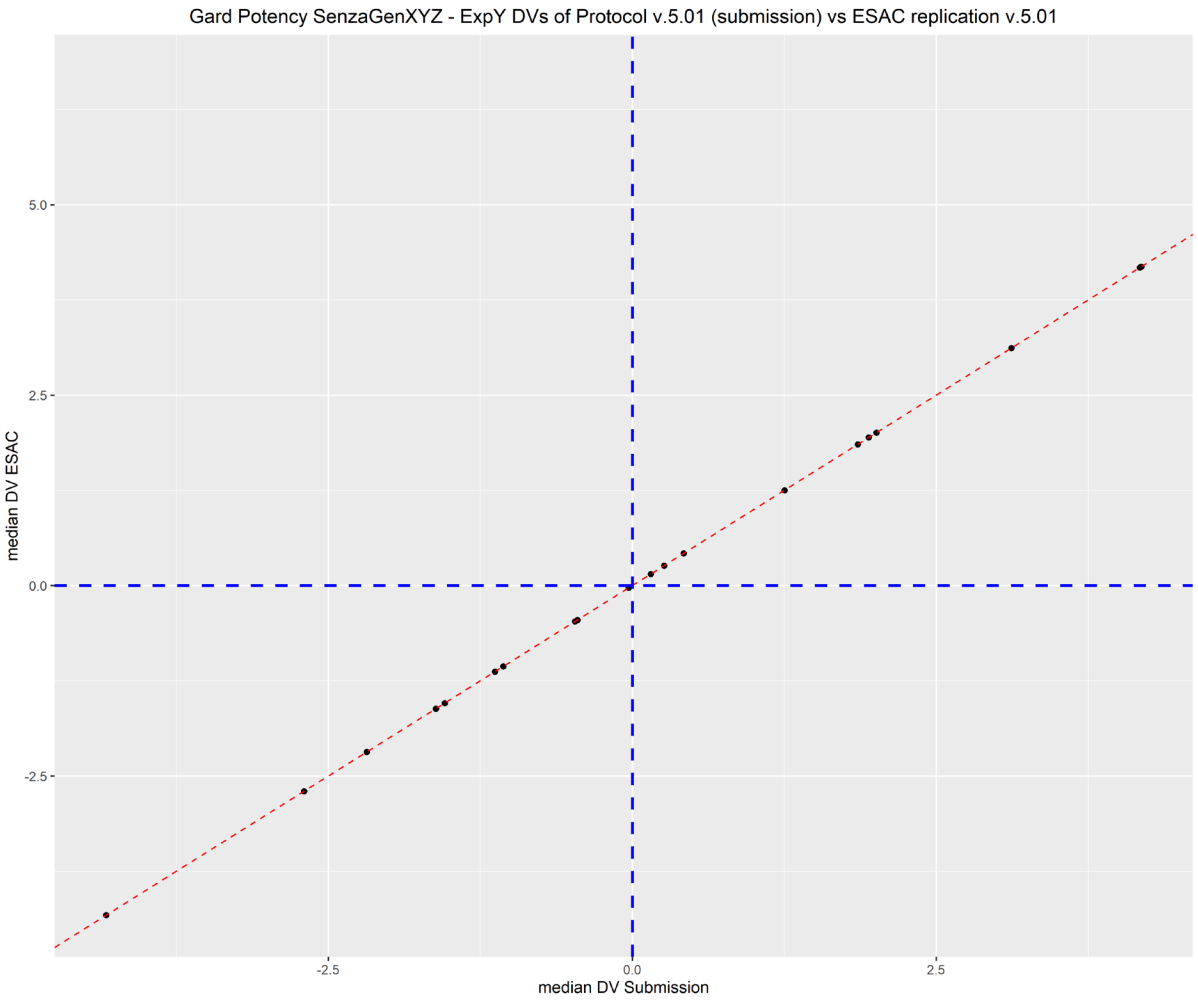


Figure 41. Scatter plot of the median DV calculated by the ESAC for SenzaGen Extra Experiment Y data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The median DV for all the chemicals could be reproduced with a precision of at least 10E-4.

2.2.6 GARDpotency SenzaGen Extra – ExpZ assessment

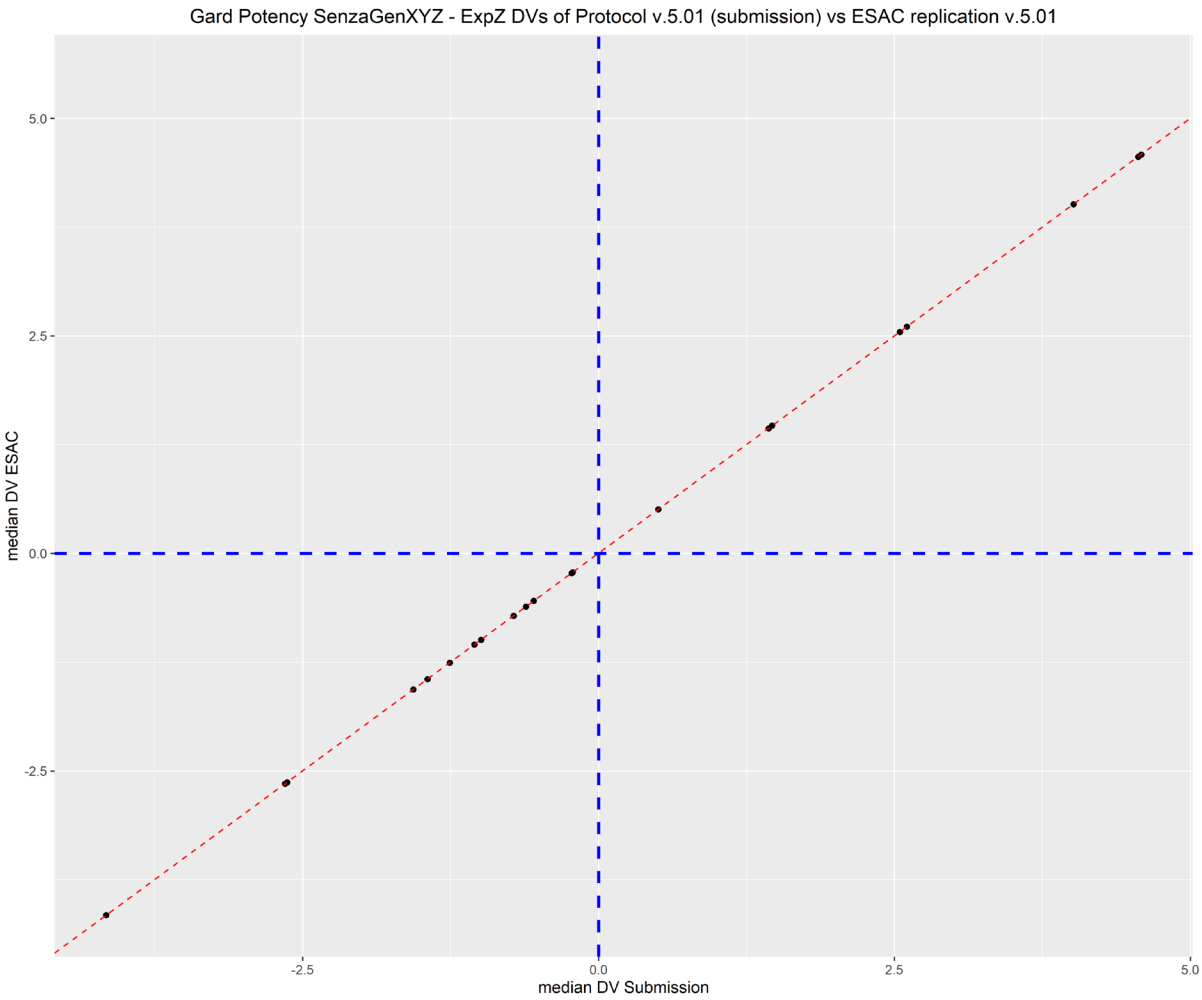


Figure 52. Scatter plot of the median DV calculated by the ESAC for SenzaGen Extra Experiment Z data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The median DV for all the chemicals could be reproduced with a precision of at least 10E-4.

2.2.7 GARDpotency BRT – Exp1 assessment

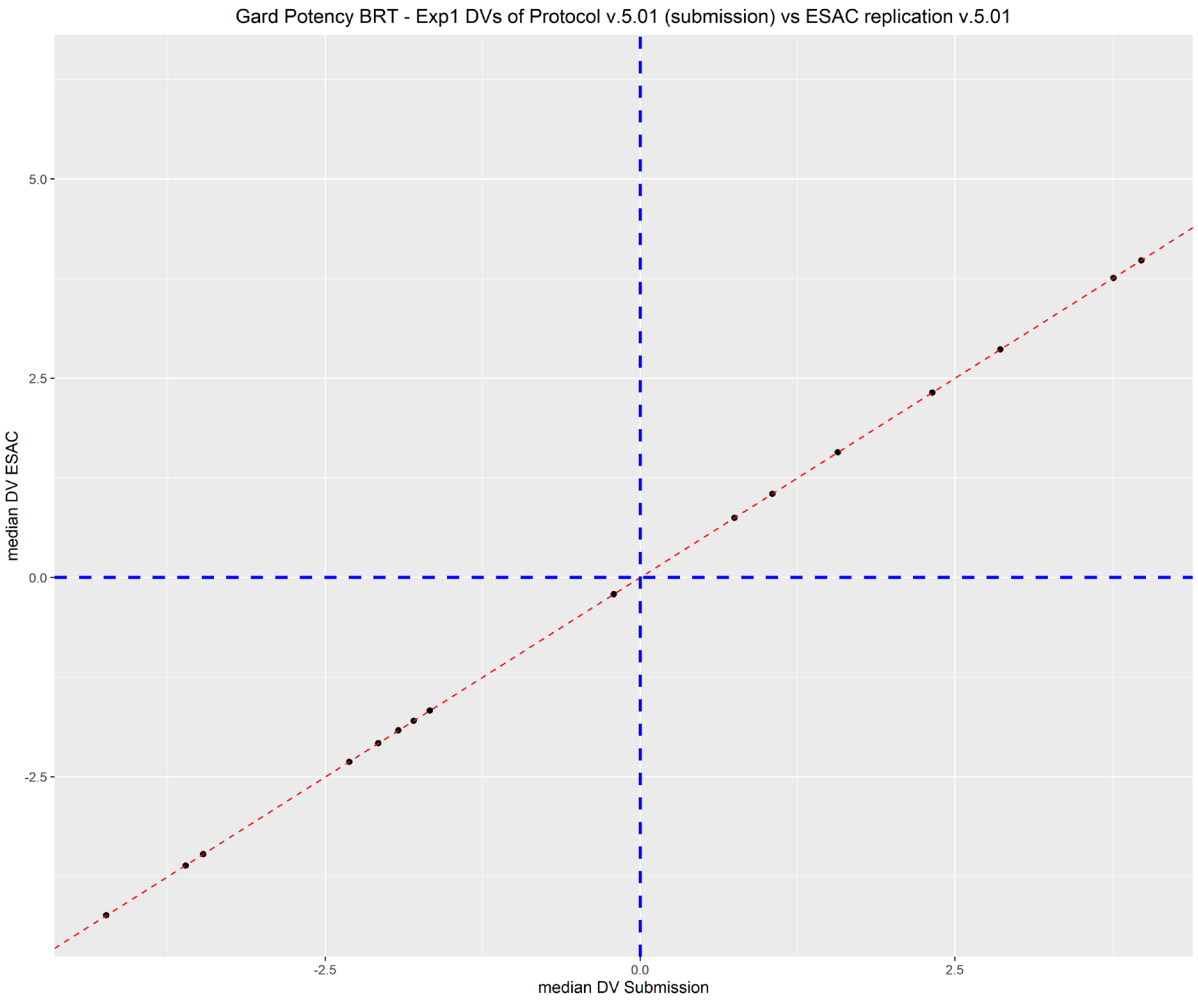


Figure 63. Scatter plot of the median DV calculated by the ESAC for BRT Experiment 1 data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The reported median DV for all the chemicals were provided rounded/truncated to 10E-2. The median DV calculated by the ESAC could reproduce the reported values with a precision of 10E-2 for all chemicals.

2.2.8 GARDpotency BRT – Exp2 assessment

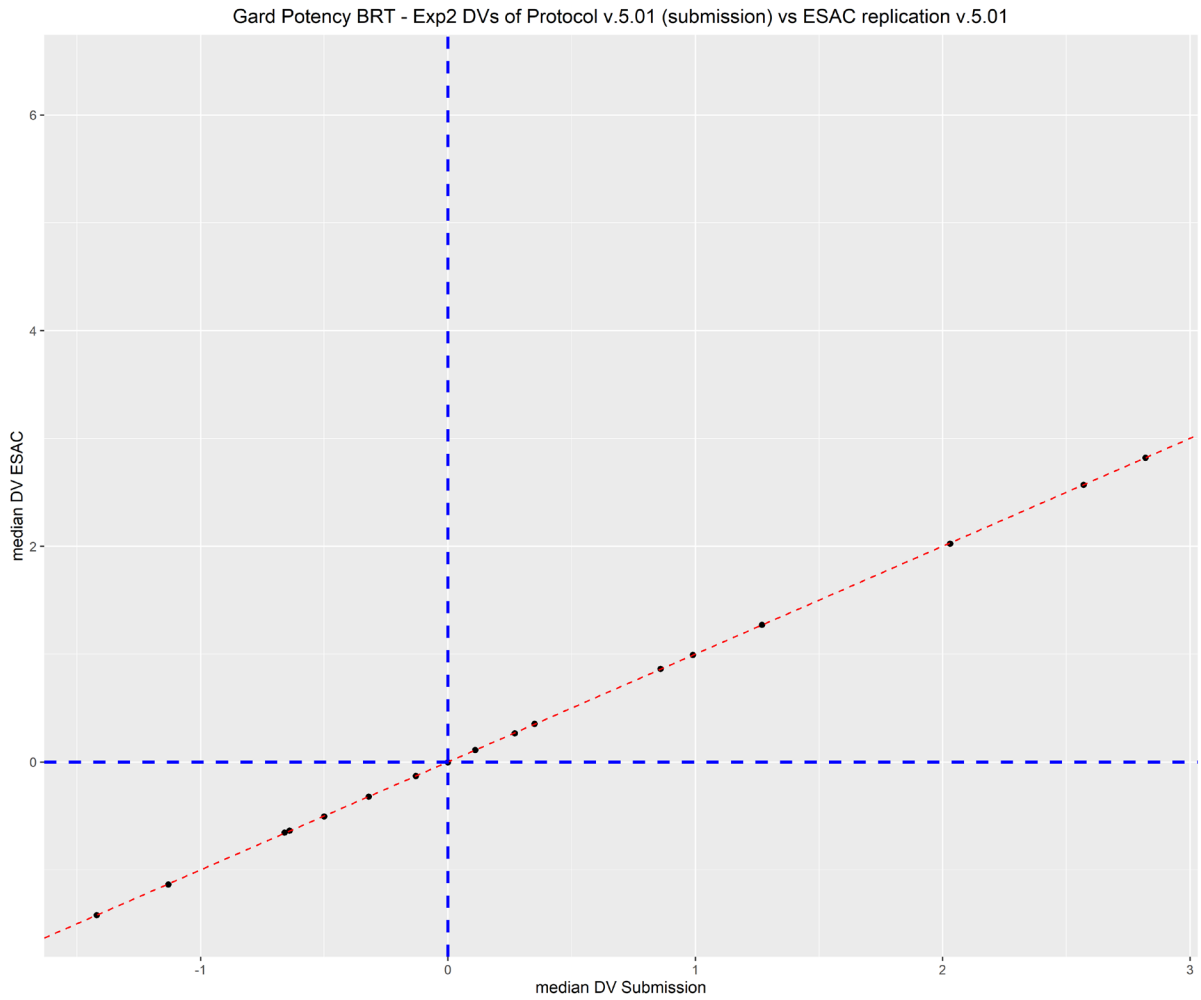


Figure 74. Scatter plot of the median DV calculated by the ESAC for BRT Experiment 2 data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The reported median DV for all the chemicals were provided rounded/truncated to 10E-2. The median DV calculated by the ESAC could reproduce the reported values with a precision of 10E-2 for all chemicals.

2.2.8.1 Other observations

One of the provided main stimulations, US-1655553-A01, did not pass the GDAA QC but since it was not included in the Annotation file, it had no effect on the calculated median DV.

Not all chemicals had 3 or more valid main stimulations. The chemicals with less than 3 valid main stimulations are listed below:

Table 4. List of compounds with less than 3 valid samples.

Lab-Experiment	**Compound**	**Code**	**Issue**
BRT-Exp2	2-Mercaptobenzothiazole	B432	Only 2 valid main stimulations were provided
BRT-Exp2	4-Nitrobenzyl bromide	B495W	Only 2 valid main stimulations were provided
BRT-Exp2	2-Bromo-2-glutaronitrile	B250	Only 1 valid main stimulation was provided

Chemicals with 2 valid main stimulations are accepted as valid predictions according to the SOP, but only 1 valid main stimulation is not accepted as a valid prediction.

2.2.9 GARDpotency BRT – Exp3 assessment

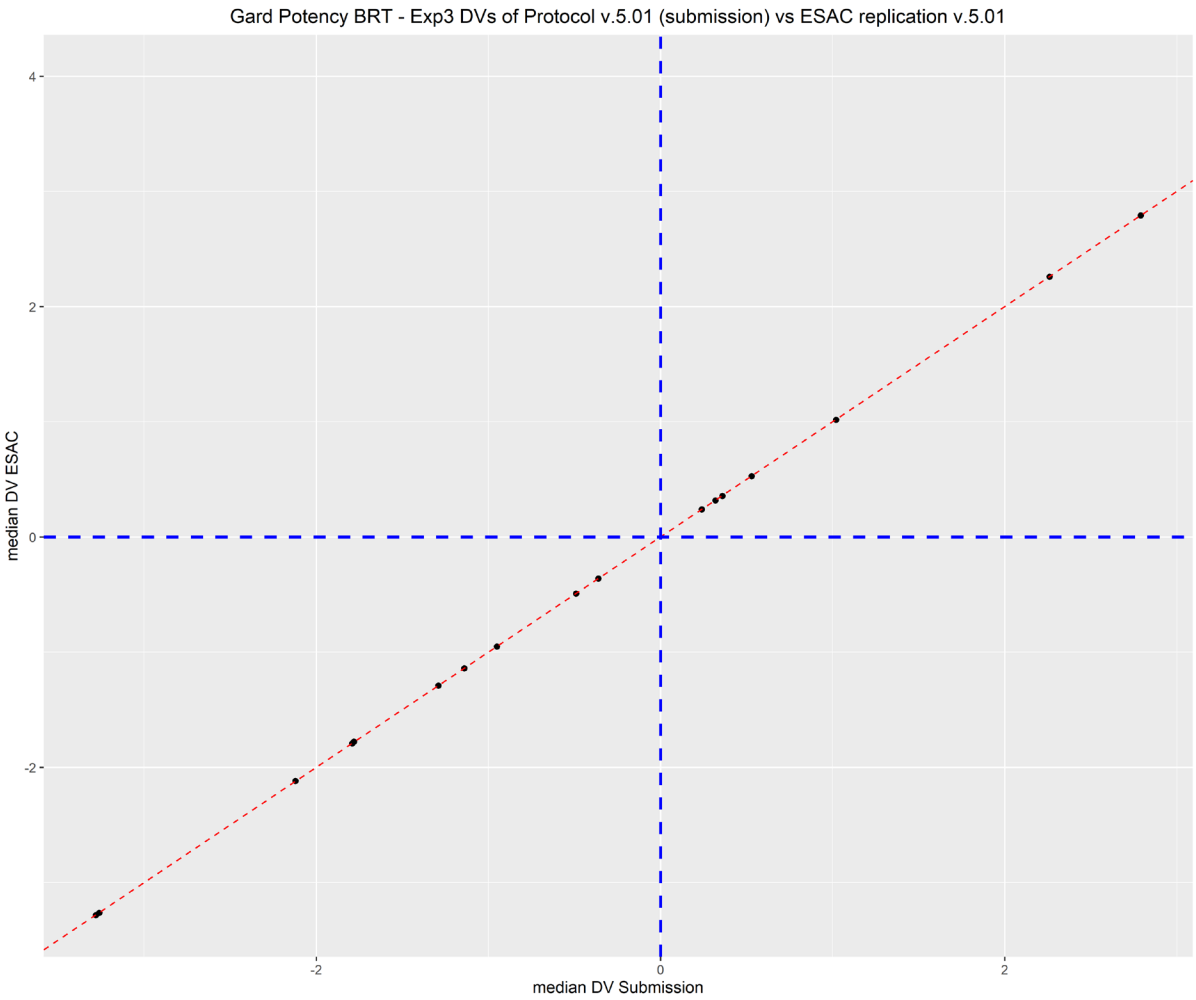


Figure 85. Scatter plot of the median DV calculated by the ESAC for BRT Experiment 3 data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The reported median DV for all the chemicals were provided rounded/truncated to 10E-2. The median DV calculated by the ESAC could be reproduced with a precision of 10E-2 for all chemicals.

2.2.9.1 Other observations

Two of the provided main stimulations, US-1663653-A01 and US-1663638-A01 (B743W), did not pass the GDAA QC. The former was not included in the Annotation file and had no effect on the calculated median DV, whereas the latter, corresponding to Benzyl benzoate (B743W), was. After being removed, the median DV calculated by the ESAC matched the reported one.

Not all chemicals had 3 or more valid main stimulations. The chemicals with less than 3 valid main stimulations are listed below:

Table 5. List of compounds with less than 3 valid samples.

Lab-Experiment	**Compound**	**Code**	**Issue**
BRT-Exp3	Benzyl benzoate	B743W	Only 2 valid main stimulations were provided (3 files were provided but 1 did not pass the GDAA QC)
BRT-Exp3	Citral	B877W	Only 2 valid main stimulations were provided

Chemicals with 2 valid main stimulations are accepted as valid predictions according to the SOP.

2.2.10 GARDpotency Eurofins – Exp1 assessment

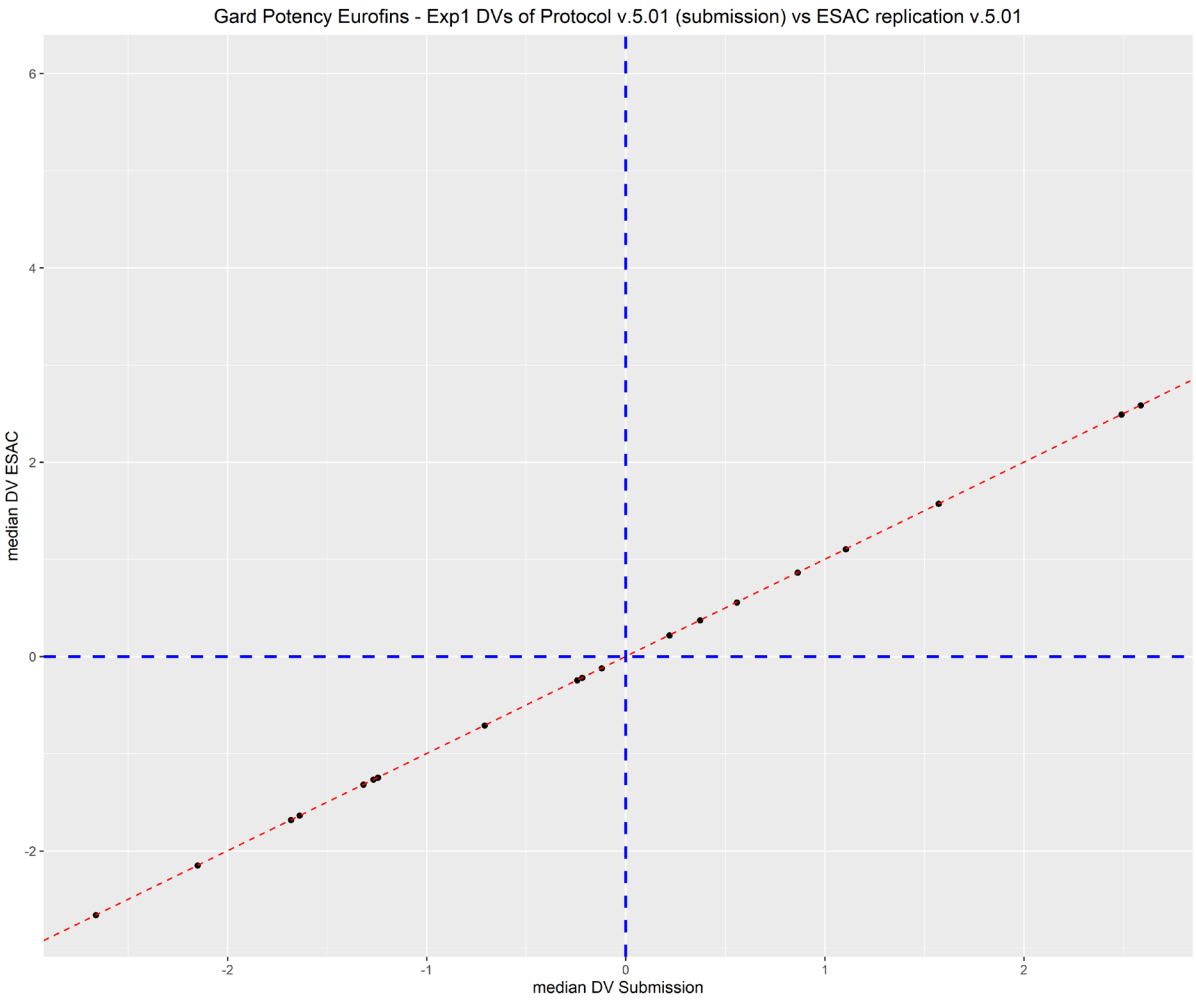


Figure 96. Scatter plot of the median DV calculated by the ESAC for Eurofins Experiment 1 data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The median DV for all the chemicals could be reproduced with a precision of at least 10E-4.

2.2.10.1 Other observations

Some chemicals had more than 3 valid main stimulations. The chemicals with more than 3 valid main stimulations are listed below:

Table 6. List of compounds with more than 3 valid samples.

Lab-Experiment	Compound	Code	Issue
Eurofins-Exp1	Formaldehyde	E 28	4 valid main stimulations provided
Eurofins-Exp1	2-Bromo-2-glutaronitrile	E 13	4 valid main stimulations provided
Eurofins-Exp1	4-Nitrobenzyl bromide	E 93	4 valid main stimulations provided

According to the SOP, it should not be possible to obtain 4 valid main stimulations, as a 4th main stimulation is only performed if one of the previous ones is considered invalid. However, the Validation Management Group agreed to allow performing up to 5 main stimulations for each chemical. For this reason, there are chemicals with more than 3 valid main stimulations.

2.2.11 GARDpotency Eurofins – Exp2 assessment

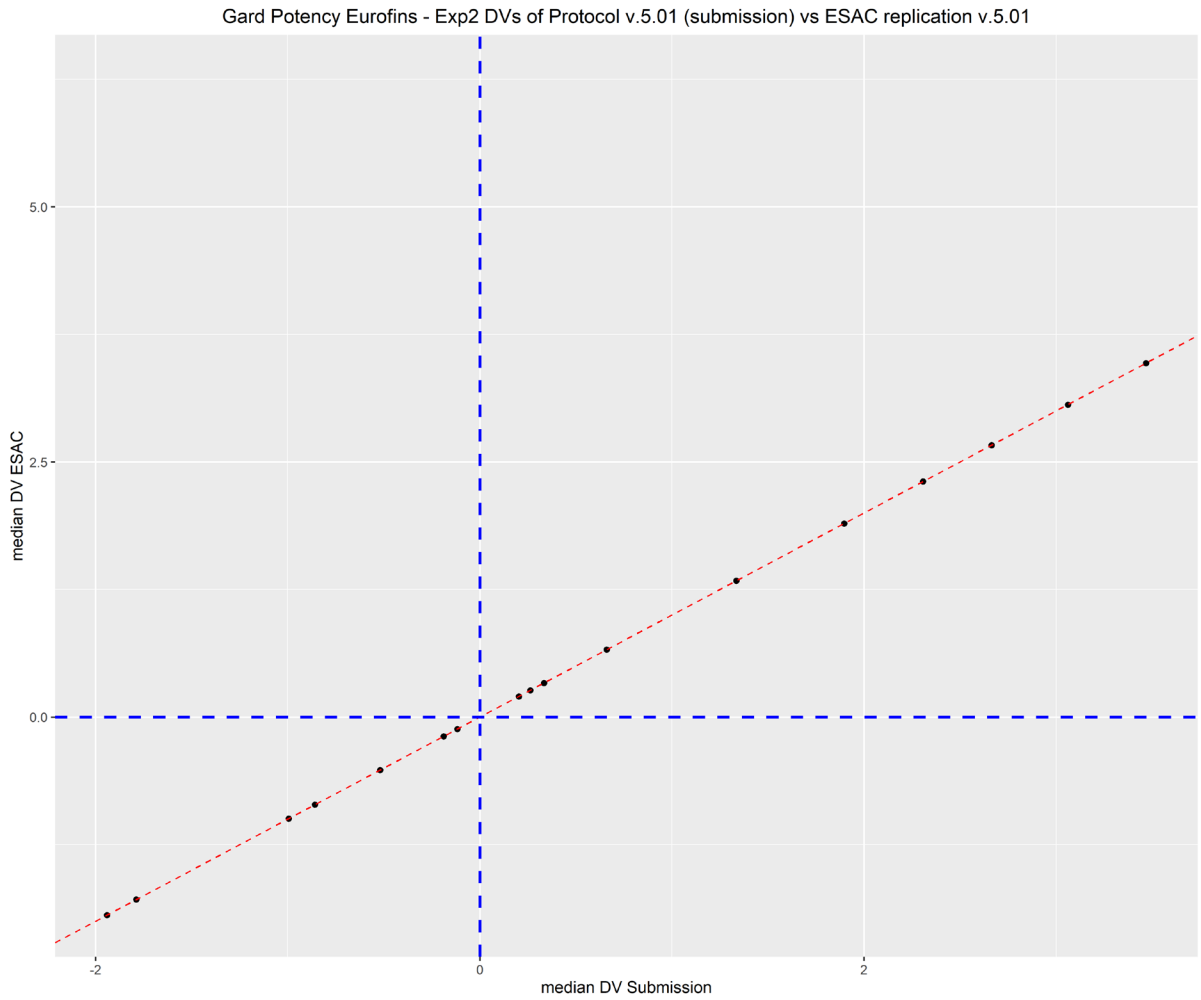


Figure 107. Scatter plot of the median DV calculated by the ESAC for Eurofins Experiment 2 data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The median DV for all the chemicals could be reproduced with a precision of at least 10E-4.

2.2.12 GARDpotency Eurofins – Exp3 assessment

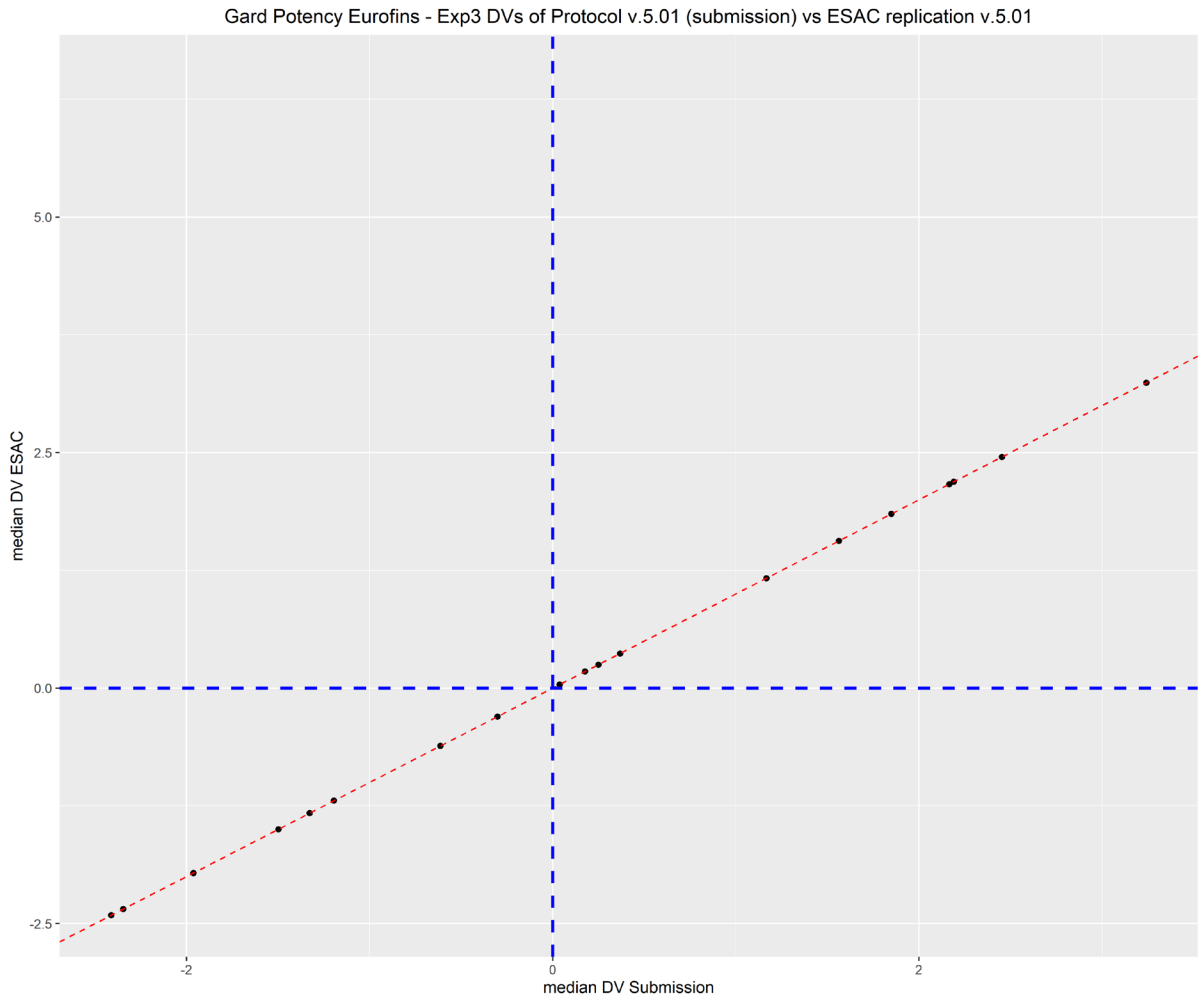


Figure 118. Scatter plot of the median DV calculated by the ESAC for Eurofins Experiment 3 data vs median DV provided in the test submission in file GARDskin and GARDpotency data.xlsx.

The median DV for all the chemicals could be reproduced with a precision of at least 10E-4.

2.3 Other observations

Some observations not directly related to the replication of DVs that were noted during the verification exercise are summarised and described below.

2.3.1 Insufficient number of valid main stimulations/missing files

Some chemicals do not have the sufficient number of valid main stimulations for their prediction to be considered valid. The experiments with less than 3 valid main stimulations are summarised below:

Table 7. List of compounds with not enough valid runs.

Model	Lab - Experiment	Compound	Code	Issue
GARDskin	BRT-Exp1	Methylisothiazolinone	B93	Only 1 valid main stimulation was provided
GARDskin	BRT-Exp1*	4-(Methylamino)phenol sulphate	B183	2 main stimulations with passed QC reported but only 1 file provided
GARDskin	BRT-Exp2	2-Bromo-2-glutaronitrile	B373	Only 1 valid main stimulation was provided
GARDskin	BRT-Exp3*	2-Mercaptobenzothiazole	B612	2 main stimulations with passed QC reported but only 1 file provided
GARDskin	BRT-Exp3*	Ethylene glycol dimethacrylate	B663	2 main stimulations with passed QC reported but only 1 file provided
GARDskin	BRT-Exp3*	Formaldehyde	B737W	2 main stimulations with passed QC reported but only 1 file provided
GARDpotency	BRT-Exp2	2-Bromo-2-glutaronitrile	B250	Only 1 valid main stimulation was provided
GARDpotency	BRT-Exp3	Benzyl benzoate	B743W	Only 2 valid main stimulations (3 files were provided but 1 did not pass the GDAA QC)

* The mean DVs reported for these chemicals correspond to those calculated with the single main stimulation. Therefore, it seems there has been an error and that those main stimulations should not have been used in the ring trial as at least 2 valid main stimulations (biological replicates) are needed for a prediction to be considered valid. B183 was not used in GARDpotency although it is 'positive' in GARDskin. B612, B663, B737W were used in GARDpotency although only 1 valid main stimulation was available from GARDskin.

Chemicals with 2 valid main stimulations are accepted as valid predictions according to the SOP, but only 1 valid main stimulation is not accepted as a valid prediction.

2.3.2 Repeated samples and in wrong folder

The folder that contained GARDskin Eurofins Exp3 files also contained files from Exp1 for methylisothiazolinone (Eurofins1-61) and glycerol (Eurofins1-90), and from Exp2 for 3-Dimethylaminopropylamine (Euforins2-162). All these main stimulations passed the GDAA QC but were not used in this exercise. These main stimulations were repetitions as there were already other main stimulations with ID Eurofins1-61 and Eurofins1-90 in Exp1 folder, and Eurofins2-162 in Exp2 folder. The compounds to which these main stimulations correspond to were reported to have 3 biological replicates, but if these extra main stimulations were counted, it would make a total of 4 for each of them. Thus, these are extra repetitions for those substances that were never used in the ring trial, possibly because they were in the wrong folder.

2.3.3 Typos in the annotation file

The annotation file of GARDskin Eurofins Exp3 contained a typo for E 662 (Toluene diamine sulphate sample, Eurofins3-69), i.e., “Eurrofins” vs “Eurofins”. Despite the typo on the file name, no effect was observed on the mean DV as the sample of the typo was not included in the folder. However, shall this happen to another file present in the folder, it would be missed and could lead to a potential error in the calculation of the mean DV and ultimately affect the classification of the compound.

2.3.4 Wrong number of files reported

The compounds listed below were reported to have only two main stimulations that passed the QC, but instead they had three that passed the QC. This was interpreted as a reporting error as the values calculated by the ESAC with three samples matched those reported by the test developer.

Table 8. List of compounds with wrongly reported number of passed QC samples.

Lab-Experiment	**Compound**	**Code**	**Issue**
BRT-Exp1	Cinnamyl alcohol	B178	2 main stimulations with passed QC reported but 3 provided
BRT-Exp2	Toluene diamine sulphate	B373	2 main stimulations with passed QC reported but 3 provided

2.3.5 Wrong formatting of Annotation file entries

The annotation file of GARDpotency for Eurofins Experiment 3 (Eurofins Exp3 Annotation file II.csv) had some wrongly formatted data. Some of the numeric entries (e.g., 8.75 and 6.25) had been transformed into dates, (e.g., “Aug 75” and “Jun 25”). This is a typical problem when opening .csv with Microsoft Excel. Fortunately, those entries did not correspond to valid samples and, therefore, the change of format did not affect the calculation of the median DV. However, it is worth noting that since the Annotation file is processed by the user, this can be a potential source of errors.

3 Rate of experiment failure

The rate of main stimulation failure was not provided in the submission report, but the submitted data contained all the experiments and main stimulations performed by the SenzaGen laboratory. The main stimulations included their metadata as well as the quality check evaluation, which indicated the reason for the failure e.g., cell viability too low or too high (>95% or <80%), low RNA concentration, “other”. These data were used to estimate the rate of failure of GARD, which is shown below.

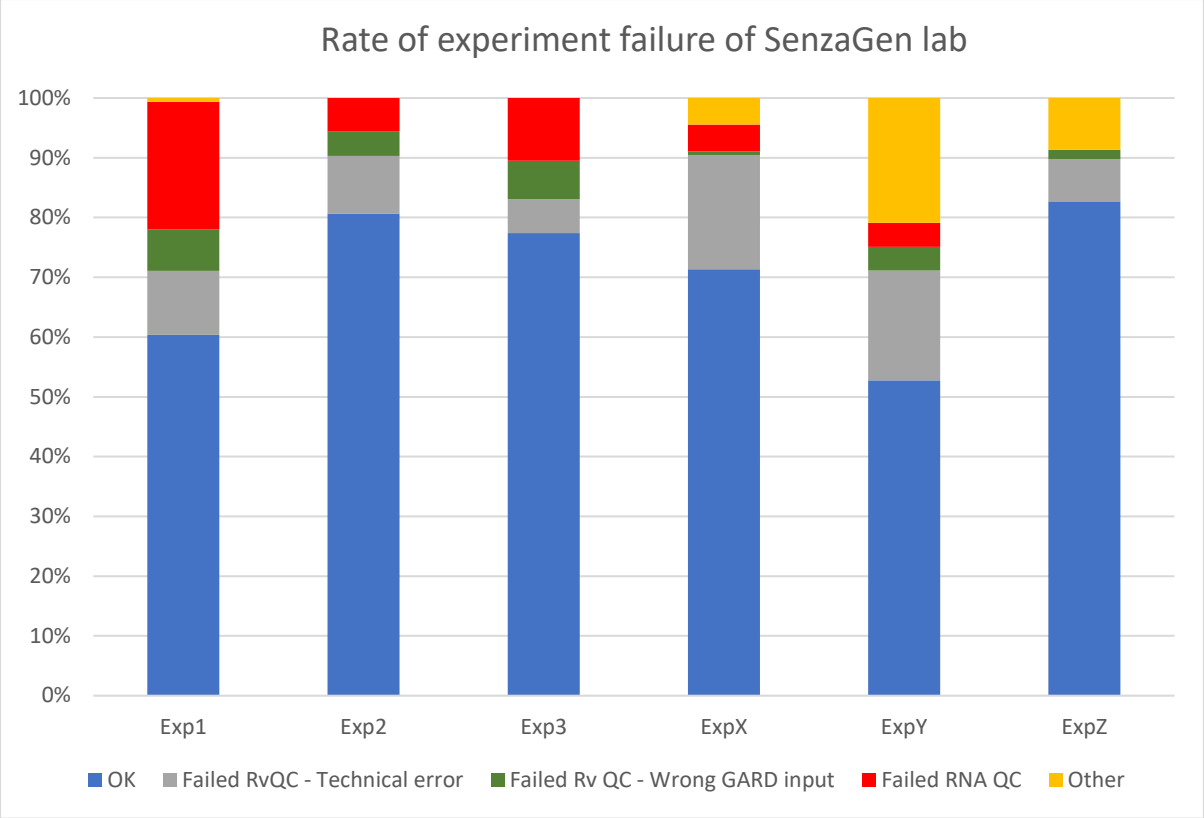


Figure 29. Stacked bar chart representing the rate of experiment failure of SenzaGen. The bars show the percentage of each type of experiment (success, failure due to RNA, failure due to Rv Quality Check – Technical error, failure due to RV Quality Check – Wrong GARD input, failure due to other reasons (e.g., discarded data, overlooked, lack of triplicates, etc.)).

The data provided in the excel files named *Els_SenzaGen ValExp XX GARDpotency CASI DMSO norm.xlsx* were summarised into successful main stimulations (OK) and those that did not pass the quality check. The latter were grouped by reason of failure, i.e., technical error or wrong GARD input.

The failure reasons have different implications in cost and stage of experiment in which the failure takes place. Failures in the phase of GARD input concentration definition have a low cost, as they can be easily repeated. Instead, failures due to RNA QC imply that several tests have already been carried out and the time and cost is higher. Some of the main stimulations in ExpX, ExpY, and ExpZ are considered failure because of "other" reason. Some of these reasons are: a main stimulation was discarded because 3 main stimulations were already available, the main stimulation was overlooked, missing replicates of that main stimulation, etc.

Experiment 1 shows a relatively high number of failed main stimulations (~40%), mainly due to RNA problems. This fact was explained by SenzaGen that it was due to a technical problem. Therefore, in order to have a good estimate of the "real" failure rate it is preferable to use the other data. Experiment 1 data are shown for transparency. The reasons for failure are variable and dependent on the experiment, therefore it is difficult to describe a trend. The only information that can be easily extracted is that the rate of failure oscillates between 17%-47% of the main stimulations (with an average of 27%, and a median of 23%).

4 Bibliography

Forreryd A., Zeller K.S., Lindberg T., Johansson H., Lindstedt M. (2016) From genome-wide arrays to tailor-made biomarker readout – Progress towards routine analysis of skin sensitizing chemicals with GARD. *Toxicol. In Vitro* 37:178-188. doi: 10.1016/j.tiv.2016.09.013.

Gradin R., Lindstedt M., Johansson H. (2019) Batch adjustment by reference alignment (BARA): Improved prediction performance in biological test sets with batch effects. *PLoS One* 14(2): e0212669. doi: 10.1371/journal.pone.0212669.

Gradin R., Johansson A., Forreryd A., Aaltonen E., Jerre A., Larne O., Mattson U., Johansson H. (2020) The GARDpotency assay for potency-associated subclassification of chemical skin sensitizers – Rationale, method development, and ring trial results of predictive performance and reproducibility. *Toxicol. Sci.* 176(2):423-432. doi: 10.1093/toxsci/kfaa068.

Johansson H., Lindstedt M., Albrekt A.S., Borrebaeck C.A. (2011) A genomic biomarker signature can predict skin sensitizers using a cell-based *in vitro* alternative to animal tests. *BMC Genomics* 12:399. doi: 10.1186/1471-2164-12-399.

Zeller K.S., Forreryd A., Lindberg T., Gradin R., Chawade A., Lindstedt M. (2017) The GARD platform for potency assessment of skin sensitizing chemicals. *ALTEX* 34(4):539-559. doi: 10.14573/altex.1701101.

Appendix II. Analysis of the GARDskin Support Vector Machine (SVM) model by the ESAC

1 Summary of the findings of this study

- The GARDskin SVM is clearly an overly complex model. SVM models with more than 25 genes do not significantly improve performance.
- A new SVM model trained with only the 10 genes gives similar performance against the external test set of the validation study than the actual GARDskin using 196 genes.

2 Development of alternative SVM models with less genes

Models in general, and especially machine learning models, have the risk of being overfitted to the data that was used to train them and, consequently, not being able to predict new data with the same level of performance. This usually happens when the models are too complex. Overfitted models usually show considerably better statistics when predicting the training set than the external test set. The rule of thumb for avoiding overfitting is given by the Topliss ratio, which states that models should use a number of features lower than 5 times the number of instances used to train the model.

$$Topliss\ ratio\ of\ overfitted\ models\colon\ \frac{n_{instances}}{n_{features}} < 5$$

In the case of GARD, the Topliss ratio shows a clear signal of overfitting, as $\frac{n_{instances}}{n_{features}} = \frac{40}{196} = 0.2$, which is clearly below 5. Even if the triplicates for each chemical were considered as independent instances ($n_{instances}$=120), the ratio would still be <5.

A simple exercise that can shed light on the possibility of overfitting is to retrain the model using fewer features. In the case of GARDskin, the original model was trained with 196 gene transcripts for the nanoString platform. In order to assess possible overfitting, alternative SVM models to GARDskin were trained using fewer features, starting from a single gene up to the 196 genes used in GARDskin. The models were trained using the same chemicals that had been used to train the original GARDskin model, and their performance was evaluated using the chemicals of the validation study and extra chemicals tested by SenzaGen that had not been used to train the model, i.e. external test set. The features available at every step were selected randomly, and in order to avoid bias, every step was repeated 100 times. The results are shown in Figure 2.

Figure 2 shows that there is no gain in balanced-accuracy, sensitivity, or specificity for models with more than 25 features/genes, and the gain for models with more than 3 features/genes is minimal. This information could have been used to build a simpler model. These results show that GARDskin with 196 genes is overly complex and could easily benefit from simplification.

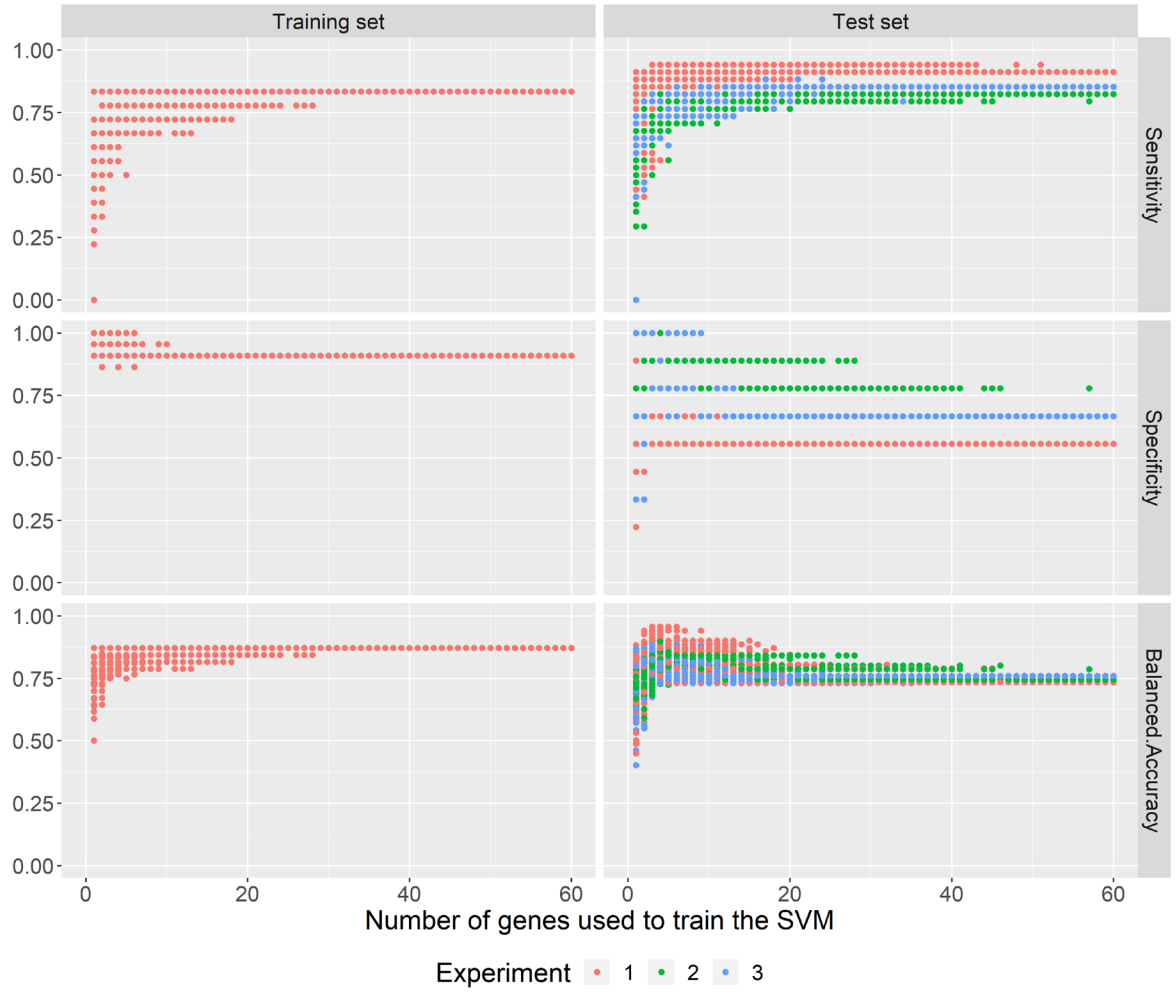


Figure 2. Dotplots showing sensitivity, specificity and balanced accuracy of the training and test sets of SVM models trained with increasing number of genes. Multiple values are shown for each number of genes as 100 different SVM with randomly selected genes were trained at each step. Values for models built with >60 features are not shown as their variability is minimal. The results of the test set are colour-coded per experiment.

3 Performance of a new SVM model trained only with the 10 genes

A new SVM model was trained using the GARDskin training set but using only 10 genes. The selected genes were: SNORA45, FTH1P2, RPSA SNORA62 SNORA6, HIST2H2AA3 HIST2H2AA4_x2, C20orf24, YBX1P1, TXNRD1, HIST2H2AA3 HIST2H2AA4_x1, HIST1H1C, and RP11.267J23.4.

This exercise shows that it is possible to develop a much simpler SVM model with similar performance to the original GARDskin SVM (see Tables 2 and 3). Unless there are justified mechanistic reasons to include a larger number of genes in the model, the preferable model should be the simpler one.

Table 2. Counts and performance metrics of an SVM trained with the 10 genes of GARD. The training set corresponds to the performanceof the model for the training set, while the test set corresponds to the performance of the model for the validation and extra chemicals that were tested by SenzaGen but were not used to train the model. Since the test chemicals were tested in triplicate, the results have also been averaged for easier comparison.

Performance of an alternative SVM model using the 10 most important genes of GARDskin		**Training Set**	**Test Set (average of Exp1, Exp2, Exp3)**	**Test Set Exp1**	**Test Set Exp2**	**Test Set Exp3**
Counts	**True Negatives**	20	6	5	7	6
	False Positives	2	3	4	2	3
	False Negatives	3	5	2	7	6
	True Positives	15	29	32	27	28
Performance metrics	**Sensitivity**	83%	85%	94%	79%	82%
	Specificity	91%	67%	56%	78%	67%
	Accuracy	88%	81%	86%	79%	79%
	Balanced Accuracy	87%	76%	75%	79%	75%
	Positive predicted value	88%	91%	89%	93%	90%
	Negative predicted value	87%	57%	71%	50%	50%
	Precision	88%	91%	89%	93%	90%
	Recall	83%	85%	94%	79%	82%
	F1	86%	88%	91%	86%	86%
	Prevalence of class 1 (sensitisers)	45%	79%	79%	79%	79%

Table 3. Counts and performance metrics of GARDskin. The training set corresponds to the performance of GARDskin for the training set, while the test set corresponds to the performance of GARDskin for the validation and extra chemicals that were tested by SenzaGen but had not been used to train the model. Since the test chemicals were tested in triplicate, the results have also been averaged for easier comparison.

Performance of the GARDskin SVM model using 196 genes		**Training Set**	**Test Set (average of Exp1, Exp2, Exp3)**	**Test Set Exp1**	**Test Set Exp2**	**Test Set Exp3**
Counts	**True Negatives**	20	5.67	5	6	6
	False Positives	2	3.33	4	3	3
	False Negatives	3	4.67	3	6	5
	True Positives	15	29.33	31	28	29
Performance metrics	**Sensitivity**	83%	86%	91%	82%	85%
	Specificity	91%	63%	56%	67%	67%
	Accuracy	88%	81%	84%	79%	81%
	Balanced Accuracy	87%	75%	73%	75%	76%
	Positive predicted value	88%	90%	89%	90%	91%
	Negative predicted value	87%	56%	63%	50%	55%
	Precision	88%	90%	89%	90%	91%
	Recall	83%	86%	91%	82%	85%
	F1	86%	88%	90%	86%	88%
	Prevalence of class 1 (sensitisers)	45%	79%	79%	79%	79%

This page intentionally left blank

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-06-21-007-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/626728

ISBN 978-92-76-40345-6