EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection
European Centre for the Validation of Alternative Methods (ECVAM)

ECVAM
SCIENTIFIC
ADVISORY
COMMITTEE
(ESAC)

ESAC Working Group Peer Review Consensus Report

on

a Givaudan-coordinated study transferability and reliability of the KeratinoSens assay for skin sensitisation testing.

Title page information	
File name	**TEMPLATE_ESAC-WG_REPORT-v3.doc**
Abbreviated title of ESAC request	ECVAM study concerning the transferability and reliability of the KeratinoSens assay
Relating to ESAC REQUEST Nr.	ESAC request ER2011-04
Request discussed through	The ESAC WG conducted the peer review from December 2011 to September 2012. Two face-to-face meetings were organized (December 2011, and February 2012), followed by two telephone conferences (February and April 2012) and finalized by written procedure in November 2012.

TABLE OF CONTENTS

ESAC Working Group

This report was prepared by the "ESAC Working Group (ESAC WG) for Sensitization/KeratinoSens, charged with conducting a detailed scientific peer review of the ECVAM study concerning the transferability and reliability of the KeratinoSens assay.

The ESAC WG had been set up by the ESAC during its meeting on March 2011 (ESAC 34). Basis for the scientific review was the ECVAM request to ESAC concerning a scientific review (ESAC request ER2011-04).

This report was endorsed by the ESAC WG on 22.10.2012 and represents the consensus view of the ESAC WG.

This ESAC WG peer review consensus report was endorsed by the ESAC on 17.12.2012.

The ESAC WG had the following members:

- Dr. Erwin Roggen (ESAC member, Chair of ESAC WG and rapporteur)
- Prof. Walter Pfaller (ESAC member, ESAC Vice Chair)
- Prof. Wallace Hayes (ESAC member)
- Dr. Maja Alecsic (external expert)
- Dr. Emanuela Corsini (external expert)
- Dr. David Lovell (external expert)
- Dr. Michael Woolhiser (external expert)
- Prof. Yong Heo (external expert)

ESAC Coordination / Scientific Secretariat:

- Dr. Claudius Griesinger (ESAC Coordination)
- Dr. Alexandre Angers (specific support)

NOTE ON THIS REPORTING TEMPLATE

The template follows the ECVAM modular approach and allows at the same time for the description of the analysis and conclusions concerning more specific questions. The template was approved by the ESAC through written procedure on 29 October 2010.

The template can be used for various types of validation studies (*e.g.* prospective full studies, retrospective studies, performance-based studies and prevalidation studies).

Depending on the study type and the objective of the study, not all sections may be applicable. However, for reasons of consistency and to clearly identify which information requirements have not been sufficiently addressed by a specific study, this template is uniformly used for the evaluation of validation studies.

- **Explanatory notes to the paragraph titles (in green)** have been added on 17 November 2010. These notes provide guidance on the type of information / analysis expected under each section. Depending on the purpose and scope of the study to be reviewed, some of the aspects mentioned in the explanatory notes may not be applicable or only be applicable to some extent. Moreover, the explanatory notes are not intended to represent an exhaustive list of possible issues to be addressed under the respective heading, but are thought to provide some guidance with respect to the considerations typically expected.
- For all of the template's numbered sections **the summary view of ESAC WG is given in bold** followed by more detailed comments ("general observations" and "specific observations").

ABBREVIATIONS USED IN THE DOCUMENT

- **BLR** Between-laboratory reproducibility
- **ECVAM** European Centre for the Validation of Alternative Methods
- **ESAC** ECVAM Scientific Advisory Committee
- **ESAC WG** ESAC Working Group
- **GCCP** Good Cell Culture Practice
- **GLP** Good Laboratory Practice
- **PC** Positive Control
- **SOP** Standard Operating Procedure (used here as equivalent to 'protocol')
- **VC** Vehicle Control
- **VMT** Validation Management Team
- **WLR** Within-laboratory reproducibility

Executive summary

Following a request from ECVAM to ESAC for peer review of and scientific advice on an ECVAM-coordinated prevalidation study concerning the KeratinoSens assay, an ESAC Working Group (ESAC WG) was set up by ESAC. The ESAC WG was charged with conducting a detailed scientific peer review of the ECVAM study concerning the transferability and reliability of the KeratinoSens assay.

The ESAC WG had been set up by the ESAC during its meeting on March 2011 (ESAC 34). Basis for the scientific review was the ECVAM request to ESAC concerning a scientific review (ESAC request ER2011-04).

The date for the opinion was set to be 4-5 October 2011 (ESAC 35). However, ambiguities and inconsistencies in the report required clarification by the test submitter. Two requests regarding clarification were sent from the WG (via the ECVAM Coordinator) to Givaudan: 16.12.2011 and 08.02.2012. These extra steps resulted in substantial delays.

The ESAC WG conducted the peer review from December 2011 to September 2012. Two face-to-face meetings were organized (December 2011, and February 2012), followed by two telephone conferences (February and April 2012) and finalized by written procedure in November 2012.

The WG was presented with a wealth of information about the test chemicals, and the assessment of WLR, transferability, BLR and predictive capacity of the test. Also the applicability domain of the test was addressed in detail.

It was obvious from the submitted material that this study had not been under EURL-ECVAM supervision. The data and the flow of events would have been more transparent if the report had followed the lay-out of this ESAC report. It would have been very helpful if the test submitters had formulated their own conclusions/opinions when referring to any of the numerous attachments that had followed the report. By referral to the attachments, the WG had to work out for itself what was meant and how the data had been interpreted by the submitter.

The WG identified a number of ambiguities and inconsistencies, without explanations being provided, which added hurtles to the evaluation of the report.

1. Ambiguities:

   - It was not clear why the statistical approach applied was chosen for the evaluation of the test results.
   - The test design was not clear.

2. Inconsistencies:

   - Data analysis apparently moved from a test result oriented (e.g. Imax, EC1.5) to a prediction (S/NS) oriented approach.
   - Test acceptance criteria changed over time without explanation as to why this occurred.

- Acceptance criteria were not consistently applied.
- Chemicals that were used for test development and refinement were inappropriately included in the assessment of the BLR and the predicitive capacity.

The WG addressed these issues by requesting additional information and re-analysis of the data from the test submitter (See Annexes).

The information provided did not provide clarification about the statistical approach applied in the study. The WG decide not to go into further discussion, and to focus on the outcome of the prediction model (S/NS).

The test design was sufficiently clarified, and the data were re-analysed on the basis of the various identified test acceptance criteria. This allowed the WG to assess properly the reproducibility, transferability and predictive capacity.

The WG attempted to recalculate the predictive capacity of the KeratinoSens based upon the chemicals that had not been included in test development and refinement. Since the number of well-characterized non-sensitizers (i.e. chemicals with negative LLNA outcome) among the eligible chemicals was considered too low, therefore the WG requested data on more negative compounds.

On the basis of this requested additional information the WG came to the following conclusions:

<u>Test chemicals:</u>

The 113 selected chemicals represented a sufficient number of materials, reasonable structural diversity and a variety of sensitising potency classes. Pre- and pro-haptens were included. Therefore, the selection of chemicals was considered sufficient to gain information on the applicability domain and the limitations of the test method.

The small number of non-sensitizers (N=4) in the list of chemicals (N=46) considered eligible for assessing the predictive capacity of the test was supplemented with 80 chemicals with negative LLNA.

<u>WLR (14 chemicals):</u>

The WG considers the concordance reported acceptable and in agreement with target value (85%) for WLR performance standards as published in international accepted guidelines (Performance standards of TG439 in vitro skin irritation).

<u>Transferability (7 chemicals):</u>

The conclusion on transferability was justified on the basis of concordant predictions (S/NS) between the lead laboratory and the naive laboratories. The WG endorses the conclusion that the test method can be transferred to naive laboratories that are experienced with cell culture techniques.

BLR (21 chemicals):

The S/NS prediction gave comparible results for the majority of chemicals (85.7 – 90.5%) among laboratories, taking into consideration the explanations given for the outliers.

Predictive capacity:

The conclusions regarding the predictivity are sound given the overall value of 76.6%. The key point here is that the 'weight of evidence' data were considered for comparison as opposed to a single assay outcome.

Compiling all the reliable data provided by the test submitter (N = 213), the KeratinoSens revealed an acceptable sensitivity, specificity and accuracy of 79.3%, 84.5% and 81.7%, respectively.

Negative results cannot, however, exclude the sensitization potential as weak and low moderate sensitizers are likely to be missed.

Applicability domain:

The applicability domain is less clearly defined with this data set and it is prudent to assess this further by testing additional sets of chemicals that are not obviously part of the applicability domain. It is clear that specific amine reactivity and requirement for some form of activation are not the only issues that may need to be addressed.

The WG made the following recommendations:

The test method can be used for S/NS identification of chemicals. Therefore, the test was considered ready for the next steps in the ECVAM process. A Validation study should however include more well-defined non-sensitizing compounds. Furthermore, a consistent use of acceptance criteria (Annex 4) should be assured.

Since the test revealed issues around weak and low to moderate sensitizers, negative results cannot rule out a sensitization potential. This problem should be clearly flagged and/or addressed to be solved.

At the SOP level, the test submitters were recommended to modify the 96-well plate design, which currently is prone to bias.

Integration of this assay with other predictive tests as they emerge needs to be based on a better defined applicability domain.

Eventual combination of the KeratinoSens assay with a reactivity based approach needs to include unambiguous identification of reactivity and any specificity associated with it.

Training should be considered for laboratories with no experience with this test.

1. Data collection – information / data sources used

1.1 Existing data used as reference data

General comments:

The chemicals were selected from the ECVAM, ICCVAM and Sens-it-iv databases (DB). Three different sets of chemicals were distinguished (Table 1).

From the submitted information it was not very clear to the WG how the reference data were used to identify for the purpose of this study credible reference sensitizers (S) and non-sensitizers (NS).

Specific comments:

- The 'Silver List' (N=67, Attachment 11) was derived from the ICCVAM, ECVAM, Sens-it-iv lists. It was unclear for the WG how the data were integrated with result the classification of the compounds as sensitizers and non-sensitizers.

  Additional input requested: By telephone the test submitter (Andres Natsch) explained that classification of the selected compounds was based upon concordant results for LLNA and Guinea pig (GP) results.

- Additional chemicals sent by ECVAM (N=8) originated from the ECVAM Coordinative Prevalidation Study. The 'weight-of-evidence' (WoE) approach used for selecting these chemicals needed further explanation.

  Additional input requested: By telephone the test submitter (Andres Natsch) explained that classification was based upon concordant LLNA and GP data, with Ni-Chloride and Xylene as additional compounds.

- The chemicals in the Extended list (N=46, Attachment 12c), were selected from the ICCVAM Validation paper and database The WoE approach used for selecting these chemicals needs further explanation.

  Additional input requested: By telephone the test submitter (Andres Natsch) explained that classification was based upon concordant LLNA and GP results, or LLNA and human maximization test.

Table 1:

Chemical sets	Derivation	Source
Silver List (N = 67)	Concordant results for LLNA and GP data	ICCVAM, ECVAM, Sens-it-iv
ECVAM list (N = 8)	Concordant results for LLNA, GP, + Ni chloride and Xylene	Published data (ECVAM, ICCVAM DB)
Extended list (N = 46)	Concordant results for LLNA and GP data, OR LLNA and human maximization test	ICCVAM validation paper, DB

1.2 Existing data used as testing data

General comments:

As with the reference data, the information submitted did not provide a clarification of how the existing data were used to identify sensitizers (S) and non-sensitizers (NS).

Specific comments:

Additional input requested: See 'Specific Comments' and Table 1 in section 1.1

The Silver List was used for test development, refinement and evaluation. Data obtained from this list were also included in the assessment of the predictive capacity of the test. A bias may have been introduced in the assessment of the predictive capacity of the test, because the test was optimized to detect these chemicals.

1.3 Search strategy

General comments:

The test submitters extracted their reference and test data from established highly credible sources (ECVAM, ICCVAM and Sens-it-iv databases).

1.4 Selection criteria applied to the available information

General comments:

The VMG relied fully on the quality of the data provided in the various existing databases.

2. Study objective and design

2.1 Clarity of the definition of the study objective

General comments:

The objective of this study is defined comprehensibly in two specific test reports (attachments. 4c and 12c) and as attachment 17e: "a screening method to test for the potential of chemicals to be skin sensitizing--The main focus is the testing of chemicals evaluated under REACH and under the Cosmetic Directive."

The study has been designed to generate information on the test methods' transferability and reproducibility to allow recommendations to be made on these two aspects in view of the future use of this test method in an integrated approach for the full replacement of the currently used regulatory animal tests. In addition the data generated in this study will inform possible future evaluations of the test methods' predictive capacity.

Specific comment:

It would have been helpful if the objectives had been specifically formulated in the body of the submission.

2.2 Analysis of the scientific rationale provided

General comments:

The KeratinoSens assay addresses indirectly steps 2 (haptenization) and step 3 (epidermal inflammation) in the mechanism of skin sensitization induction (Adler et al., 2011). The relevance of Nrf2-Keap1 pathway to skin sensitisation is explained by the direct reactivity of sensitising materials to key cysteine residues of Keap1 (Nrf2 repressor protein).

The available evidence for this links to processes of sensitization via sensitizer-induced up-regulation of genes with ARE in their promoter sequence. The question of direct reactivity with Keap1 cysteine residues remains unanswered.

Not all skin sensitisers appear to up-regulate the Nrf2 pathway. This can be explained by either a need for metabolic transformation or an exclusive reactivity of some test chemicals with lysine.

It may be possible that additional or alternative pathways can be activated via modification of cysteine residues on Keap1 protein.

Specific comments:

The Nrf2 regulatory pathway, comprising of the repressor protein Keap1 (kelch-like ECH-associated protein 1), the transcription factor Nrf2 and the antioxidant response element (ARE), has been shown to play a critical role in protecting a variety of tissues from a wide array of toxic insults. It is emerging as one of the key toxicity pathways induced by skin sensitisers. The transcription factor Nrf2 regulates the battery of genes that are induced from the ARE in response to an electrophilic insult to coordinate cytoprotective response. The Nrf2-ARE system regulates expression of numerous cytoprotective enzymes and under non-stimulated conditions Keap1 negatively regulates nuclear translocation of Nrf2 and facilitates degradation of Nrf2 via proteasome. Upon exposure to electrophilic chemical Nrf2 is liberated from Keap1-dependent degradation and accumulates in the nucleus (Maruyama and Itoh, 2005).

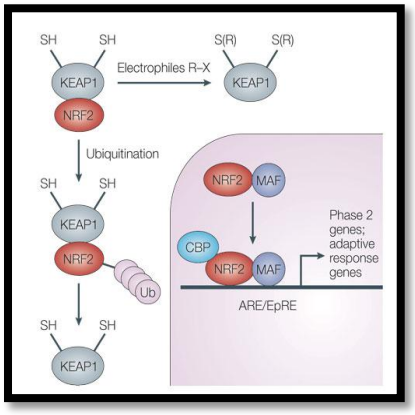


Figure 1: Mode of action

It has been shown that Keap1 interacts with some electrophilic chemicals, leading to activation of ARE-dependent genes, however the precise nature of this interaction is not clear. It is postulated that skin sensitising electrophiles covalently modify the key cysteine residues on Keap1 which sets off the described sequence of events. Natsch and Emter (2008) investigated ARE-dependent gene induction by skin sensitising chemicals and reported that majority of these chemicals do induce genes

regulated by this element (Fig. 1).

In itself, this interaction does not constitute functional *in vivo* relevance. Natsch et al (2010) reviewed recent literature in the field allowing insights into *in vivo* relevance and identifying some missing links for further research. The integrated hypothesis they put forward suggests that there are two separate signalling cascades for cysteine and lysine reactive chemicals, ultimately leading to Th1 and Th2 responses respectively, with modification of the key cysteine residues on Keap1 being the first step in the Th1 pathway. Several lines of evidence exist that indicate that Nrf2-Keap1-ARE regulatory pathway is activated by cysteine-reactive skin sensitizers:

- The sensor protein Keap1 contains highly reactive cysteine residues and covalent modification of these leads to dislocation from Nrf2, which then accumulates and activates genes with ARE promoter sequence
- Natsch and Emter (2008) used Hepa1C1C7 murine hepatoma cells and measured induction of ARE-regulated quinone reductase in response to 96 skin sensitising chemicals. In parallel, a number of chemicals were tested in the reporter cell line AREc32 (containing an 8-fold repeat of the ARE sequence upstream of the luciferase gene. 14/15 of strong/extreme sensitisers and 30/34 moderate sensitisers induced ARE-dependent luciferase activity and in many cases this was paralleled by an induction of quinine reductase activity in murine hepatoma cell line. Subsequently, Emter et al (2010) have used a human relevant cell line with luciferase reporter gene under control of a single copy of the ARE element of the human AKR1C2 gene stably inserted into HaCaT keratinocytes. The different steps in the signalling cascade (Keap1 binding, nuclear accumulation of Nrf2 and binding to the consensus ARE sequence) have not been investigated separately. Other investigators showed Nrf2 increase in dendritic cells (DC) upon treatment with skin sensitizers (Ade et al 2009 and Megherbi et al 2009). Similarly, Python et al (2009) showed up-regulation of two key Nrf2-regulated genes, CES1 and NQO1 in cinnamic aldehyde treated DCs and DC cell line MUTZ3, indicating an up-regulation if the Nrf2 pathway in both cell types. However, not all skin sensitizers appear to up-regulate the Nrf2 pathway and part of the explanation for some of these materials is either a need for metabolic transformation (generally poor in cell lines) or exclusive lysine reactivity of some of the test chemicals.
- Evidence for the *in vivo* relevance of the Nrf2 pathway induction in skin sensitisation comes from studies on the effects of DNCB and oxazolone on the Nrf2 -/- knockout mice (ref). These animals show a reduced reaction to sensitizers, with Th1 cytokine IFNγ response completely abolished, and Th2 cytokine IL-4 unaffected. These data indicate that the induction of Nrf2 pathway is essential for sensitisation to occur.
- Further *in vivo* studies reveal up-regulation of IFNγ and IFNγ-regulated genes by sensitizers (DNCB, oxazolone and toluene-2,4-diisocyanate) but not irritants (croton oil and nonanoic acid) (Ku et al., 2008). All three sensitizers have intrinsic reactivity to both cysteine and lysine. Boverhof et al (2009) tested lysine reactive chemicals and did not observed up-regulation of the IFNγ gene.
- A large body of evidence exists for differential cytokine induction in the lymph nodes by either respiratory or skin sensitizers (anhydrides and isocyanates Vs DNCB), which led to a conclusion that respiratory senitizers preferentially induce Th2 cytokines and skin sensitizers induce Th1 cytokines. Natsch et al suggest that this difference is due to selective reactivity to lysine of respiratory sensitisers, compared to preferential cysteine reactivity or mixed reactivity to cysteine and lysine for skin sensitizers. The authors argue that solely lysine reactive sensitizers are rare amongst sensitizers, however peptide reactivity assays show that there are a number of solely lysine reactive chemicals which are known sensitizers.

Furthermore, human proteins are very rich in amine-type nucleophiles and cysteine residues are relatively rare and most often unavailable in the secreted proteins as they form disulphide bridges.

The relevance of this pathway to skin sensitisation is explained by the direct reactivity of sensitising materials to key cysteine residues of Keap1, however there is no direct evidence of this, rather an indirect evidence via up-regulation of genes with ARE in their promoter sequence. This event is then exploited in the KeratinoSens assay, but the question of direct reactivity with Keap1 cysteine residues remains unanswered in this case.

According to Dinkova-Kostova et al (2005), the chemicals that induce phase 2 enzymes via Nrf2 pathway are structurally dissimilar and from a variety of structural classes but all appear to be able to react with sulfhydryl groups via oxido-reduction, alkylation or disulphide interchange. Natsch et al state that this interaction is covalent in nature (i.e. solely alkylation). The fact that some solely lysine reactive chemicals appear to induce Nrf2 pathway in KeratinoSens assay is probably due to the ability of these chemicals to oxidise cysteine residues on Keap1 rather than covalently modify them. Indeed, reactivity studies show that chemicals (in particular sensitising aldehydes) conjugate to lysine residues but can oxidise cysteine residues without actually generating adducts. It is therefore logical that this pathway may also be activated by simple (chemically driven) oxidation of key cysteine residues on Keap1. For example, benzaldehyde or phenylacetaldehyde are not directly reactive with cysteine, but both strongly oxidise thiols, both react via Schiff base formation to amine based nucleophiles, yet both are positive in KeratinoSens assay.

Specificity of the interaction with key cyssteine residues has been studied extensively with a variety of inducers of this pathway (reviewed by Dinkova-Kostova et al., 2005). The consensus of such studies suggests that only certain set of key cysteine residues (four key residues are often mentioned out of the 27 residues in human Keap1) are likely to lead to the Nrf2 pathway induction, and that some reactive chemicals might not induce the Nrf2 if they modify cysteine residues that are not from this key set. Others argue that there are more than the four exclusive Cys residues which mediate electrophile/oxidative stress (Holland et al 2008 and others).

It is likely that this system is much more finely tuned *in vivo*, where the concentrations of Keap1 and a reactive chemical are much lower, and the modification sites vary between the chemicals. It is also likely that different reactions will modify different cysteine residues and all of this flexibility in the system allows for the variety of situations which can trigger the phase 2 response in cells.

The interaction of Keap1 cysteine residues with GSH/GSSG has also been a subject of study (Holland et al 2008). This mechanism of Nrf2 pathway induction involves glutathionylation of cysteine residues on Keap1 (as a result of oxidative stress and disturbance of the cellular GSH/GSSG balance) inducing formation of both type 1 and type 2 disulphides. However, only about half of Keap1 cysteine residues are subject to this type of interaction. There is a suggestion, but no direct evidence of this happening *in vivo*, although the above study worked with physiologically relevant GSH/GSSG ratios. It is possible that this pathway may also be activated by simple (chemically driven) oxidation of key cysteine residues on Keap1.

Kobayashi et al (2009) show some evidence that suggests that targeting particular set of cysteine residues will dictate what type of response (distinct biological effect) will be resulting from a chemical treatment ('cysteine code'). There is also evidence that Keap1 inhibits the NF-κB signalling pathway via induction (through a direct interaction) of IKKβ degradation (Lee et al., 2009). Similar to

the Nrf2 pathway discussed above, it appears that Keap1 specifically interacts with IKKβ, a kinase which regulates the ubiquitination of IkB protein (inhibitor of kB). Upon stimuli (which can presumably include modification of Keap1 protein), IKKβ is released from its complex with Keap1 and is able to catalyse the phosphorylation of 3 serine residues on IkB, thus tagging this protein for ubiquitination and proteosomal degradation. This releases NF-kB from its complex with IkB and allows its nuclear translocation where it binds the response elements (RE) on the relevant parts of DNA, resulting in the upregulation of proteins and change in cell function. Thus it is possible that additional or alternative pathways can be activated via modification of Cys residues on Keap1 protein.

On this background, it can be concluded that the applicability domain for the KeratinoSens may not be as clear as suggested by the test submitters. Further research should be undertaken to acquire a better understanding of the mechanisms driving the Keap1-Nrf2 pathway.

2.3 Analysis of the regulatory rationale provided

General comments:

The regulatory objective of the study is clearly stated and comprehensibly defined: a screening method to test for the potential of chemicals to be skin sensitizing.

The main focus is the testing of chemicals evaluated under REACH and under the Cosmetic Directive. The test is proposed as a stand-alone method for classification and labelling (Regulation No 1272/2008 on classification, labelling and packaging of substances and mixtures) or as part of an integrated testing strategy and combined with read-across for a basic risk assessment and safety prediction.

Specific comments:

The integrated testing strategy for skin sensitisation has not yet been identified or proposed in any detail. The test submitters suggest that in combination with other assays and read-across a basic risk assessment may be achieved. It is difficult to envisage the exact role KeratinoSens assay would play in as yet undefined strategy.

It was stated that for classification and labelling the assay could be used as a stand-alone method, however, there are clear limitations with the applicability domain and there may be a requirement for additional reactivity testing:

- limitations with respect to scoring moderate and weak sensitizers (See Module 5).
- potential limitations with the applicability domain (Module 6) imposing additional reactivity testing for some chemicals which would fail to be identified as sensitizers in KeratinoSens assay.

2.4 Appropriateness of the study design

General comments:

An overall study design was not provided, but Modules 1-6 were carefully described. The number of test items ranged from 7 chemicals to a larger set of 67 chemicals (44 sensitizers and 23 non-sensitizers).

The selection of the test items (with a good spread between potency categories) as well as the number of test items (both sensitizers and non-sensitizers) is appropriate for the purpose of the study. The number of laboratories involved in the ring trial was sufficient.

Some of the inter-laboratory studies were run with coded samples. None of the studies were under GLP. The overall technical aspects of the various studies were conducted in a quality fashion.

Test acceptance criteria were established, but not subsequently adhered to in the final studies.

Specific comments:

The Silver List of chemicals (Natsch and Emter, 2008) was used for test development, refinement and evaluation. Results obtained from this list were also included in the assessment of the predictive capacity of the test. By including test chemicals previously used for training the test, a bias might have been introduced in the assessment of the predictive capacity of the test.

The test chemicals cover:

a) a molecular weight range from 30 to 388 daltons,
b) a cLogP range from -4.8 to 5.2,
c) a range of skin sensitizer classes (from no sensitizer tor mild to extreme sensitisers),
d) all key reference lists of chemicals which have been published on skin sensitization.

The test method (SOPs) was transferred to 5 laboratories and a pre-validation study conducted which comprised 6 of the 7 modules required by ECVAM for test validation.

There was no retesting for unqualified tests, nor was there a strategy described to do so.

It would have been useful to report reactivity data (actual or predicted) of the chosen chemical dataset.

2.5 Appropriateness of the statistical evaluation

General comments:

Analysis of the appropriateness of the statistical evaluation indentified a number of issues needing clarification (See 'Specific comments', pp 17-19).

Additional input requested: The WG requested additional information by letter (16.12.2011 and 08.02.2012) to clarify issues related to:

- Test design
  - It was unclear how the KeratinoSens assay was carried out to derive a final prediction (S/NS)
- Acceptance criteria
  - Why acceptance criteria were not consistently applied, and varied between different sections of the document;
  - Concerns that these inconsistencies affected WLR and BLR calculations
- Statistical methods
  - Apparent inconsistency between the statistical analyses reported (attachments 7a-10e, and Natsch et al. (2011)).

The additional information received back was analysed by the WG, and the following conclusions were drawn:

- Test design
  - It was agreed that the schematic outline, updated by Givaudan by adding a MTT parallel plate to assess cytotoxicity (condition 4 of the Test Acceptance Criteria), helped in understanding how a KeratinoSens test is carried out to achieve a final prediction (Fig. 2).
- Acceptance criteria
  - The WG established that the criteria sets found and described by the ESAC WG corresponded to the criteria sets communicated by Givaudan in the resubmission ("clarification") as shown in Table 2. The ESAC WG agreed that the question of which criteria had been used to analyse in particular the ring trial data was now sufficiently clear. Moreover, it was now clear which criteria Givaudan recommends for future use of the assay (Annex 3, Section 3.2).
- Statistical methods
  - The WG did not consider that the additional information sent by Givaudan regarding the statistical analyses provided the clarification they were requesting. In any case, the WG decided to focus on the analysis of the within laboratory reproducibility (WLR) in terms of concordance of the predictions obtained within the same laboratory. This means that the only study that can provide this information is in attachment 4c, the evaluation of 14 chemicals at Givaudan, performed in three complete experiments.

- Concerning the occurrence and impact of the invalid runs, the WG requested Givaudan to provide a final analysis of the results in the submission, taking into account three different scenarios. On the basis of this new information, the WG came to the following conclusions:
  - WLR
    - The ESAC WG agreed that the resubmitted re-analysis data were satisfactory with regard to answering the question to what extent the non-qualified test results might have influenced the WLR analysis**.** The impact was felt to be negligible because in case 2 (the most stringent criteria) only 3 individual laboratory predictions had not qualified. The data were found sufficient to judge WLR.
  - BLR
    - Agreement was reached regarding the deletion of the fourth column in the Word document as well as the corresponding columns in the excel files. These columns were intended to allow analysis of concordance of predictions on the basis of four laboratories only instead of the five laboratories that had participated in the ring trial. The rationale for this column was that with an increasing number of labs it may be more likely to get non-concordant results. Givaudan, when planning the ring trial had been unaware that it is common practice in the context of validation to use three laboratories within a ring trial for assessing transferability and between-lab reproducibility (BLR). However, as this analysis had not been properly conducted in the resubmitted data package, the ESAC WG felt that this approach should not be followed-up when finalising the ESAC review.
    - No agreement was reached with respect to the question whether or not the data on BLR (when taking non-qualified laboratory predictions into account: case 2 and 3) were sufficient to judge reproducibility between laboratories. The reason for this uncertainty lies with one of the principal flaws of the planning and execution of the study, i.e. that the Test Acceptance Criteria (TACs) provided to the participating laboratories during the ring trial (a) had not been applied when analysing the data (a variation of the initial TACs had been employed) as they were found too stringent, (b) that no provisions for re-testing had been made in case test results would not meet the TACs and (c) that, as a consequence, the final data matrix contained an appreciable number of non-qualified test results that were included in the analysis.

Figure 2: Test design.

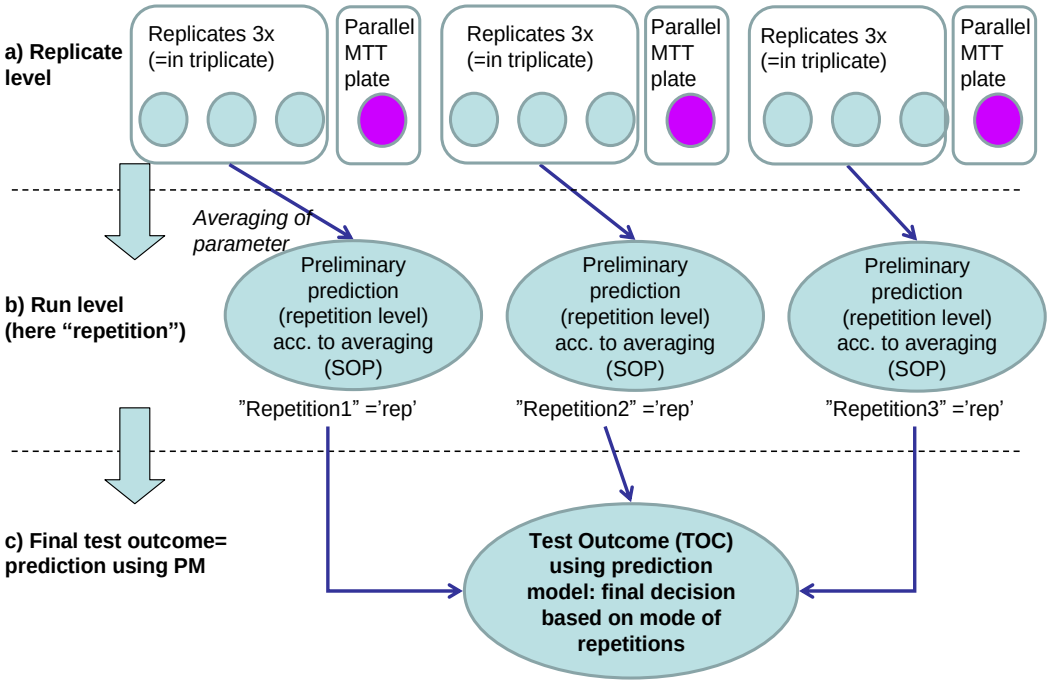


Table 2: Acceptance criteria

Criteria Set found by WG and communicated in the document "Final Clarifications…" to Givaudan by the ESAC Sec. (sent to Givaudan 8.2.2012)	**Criteria Set as described in the resubmission from Givaudan (received by ECVAM 14.3.2012)**
Criteria Set 1 (page 7 of submission).	***Criteria Set 1*** (page 7 of submission) Additional clarification (in blue) reg. condition 1: CA positive in the range 4-64uM. *NOTE ESAC SEC: These criteria (and not Criteria 2 communicated in SOP of ring trial) were used to analyse the ring trial data. However, some ring trial data were invalid when applying these criteria. As no rules for retesting in case of unfulfilled TACs had been stipulated beforehand, this led to some non-qualified test results being included in the analysis contained in the full submission to ECVAM. Having realised that non-qualified test results had been included in the analysis on reproducibility and predictive capacity, the ESAC WG requested (on 8/2/2012) a reanalysis of the data using (a) concordance of predictions as a measure for reliability and (b) performing this analysis applying or not-applying the specified test acceptance criteria.*
Criteria Set 2 (Att. 7a SOP of ring trial): contained in the ring trial SOP but neither used by participating labs nor by lead lab for final analysis. This set should be ignored for the purposed of the review.	
Criteria Set 3 (page 9 of submission)	***Criteria Set 2*** Additional clarification (in blue) reg. condition 1: CA positive in the range 4-64uM *NOTE ESAC SEC: These are the test acceptance criteria as currently used by Givaudan and as recommended for future use.*
Criteria Set 4 (page 17 of Invittox protocol): This set is identical to set 3, apart from a minor typo (7.5 instead of 7.0 uM in condition 2). This set should be ignored for the purposes of the review.	

For more detailed information, please consult Annexes 1-4.

Specific comments:

The methodology for assessing the within and between laboratory variability from a quantitative perspective is difficult to follow. There are two separate sections dealing with apparently different statistical analyses of the same data. Attachments 10.a. to 10.f. There is also a statistical analysis of data in Attachment 4c plus graphs in 4c.

It seems that ECVAM's internal statistician had expressed difficulties in understanding the analyses provided using the geometric means for assessing within and between laboratory variability using geometric means and requested a review of the analysis by the test developer. One issue was about the appropriateness of the use of the geometric standard deviation for the assessment. It seems that the developer responded by providing another analysis of the data set using a different methodology. There is very little discussion of the comparison between the two methods but the developer suggests that the two analyses provide compatible conclusions (and seems to confirm this in a separate note sent after a teleconference with the panel.)

Each experiment consists of three repeats of three replicates. The description of the hierarchical design and interpretation of sources of variability in the attachment 10.f is difficult to follow or to see how it is used.

It's not completely clear if the n's for means and, particularly, SDs are based upon 3 or 9 measures. They are probably based upon 3 values. This remains to be clarified.

The discussion about the estimates of variability measured by SDs especially by the use of geometric means and log2 transformations is difficult to follow. It is not clear completely clear what specific points are being made or how the values are interpreted.

It is not completely clear why the analysis of variability is based upon data transformed to the log of 2. The choice of a log transformation to the base 2 seems to be to try to relate the variability seen to the log2 dose spacing on the plates. It is not completely obvious that this transformation can be justified on statistical grounds. (This was, presumably, one of the reasons the ECVAM statistician wanted an 'independent' review of the data.).

The relationship described between the geometric standard deviation and the data expressed as log to base2 and to a 95% confidence interval needs to be clarified and expressed more clearly.

Is there evidence that the log transformation is an appropriate one? Is there a possibility that the transformation is too 'fierce'? The untransformed data may result in higher variability being associated with higher means (i.e. means and SDs positively correlated). The log transformation may make the lower means appear to be more variable (i.e. means and SDs negatively correlated). The figure at the end of attachment 3 shows the untransformed data are approximately normal but with a long right-hand tail. The transformed data, although symmetrical, had 'fat' tails with an appreciable number of very low and very high values.

Some of the summary statistics have been checked by hand by the panel and the relationship between the test results and the summary statistics described in the text have been confirmed. However, the description of the precise methods used to produce the table in attachment 4b are currently very difficult to follow and it is hard to relate the tables of the test results with the table of the summary statistics.

When the term % variance is used it presumably relates to the coefficient of variance (CV) measured as (SD/Mean) x 100.

It was not completely clear if CVs based upon logarithmically transformed data were produced. It is possibly that this was not done: but the text is unclear. This approach is, though, not usually considered an appropriate way of handling data. Problems arise if the log transformation results in a zero or negative value. (For log10 then log(X) =0 when X =1 and is negative when X is between 0 and 1 so with transformed data the CV can apparently be negative).

Attachment 10f

Attachment 10f discusses the statistical analysis of the within and between laboratory variability. There is mention of analysis of variance. It is not clear if this related directly to an ANOVA carried out on the data or a more general discussion on how to handle variability.

It appears to be based upon some sort of ANOVA methodology. It is not clear how the data were analysed. Were transformed data used? (In a separate note the reviewer says no.) Variance components are specifically mentioned which seems to imply that some sort of ANOVA is used. The specific ANOVA and software should be provided.

It is not clear what the term in the report 'parameter' refers to.

It is not clear how the variances excluding a laboratory are compared with the variances including all laboratories. What statistical test is used to show the derived ratio is statistically significant ($P<0.0001$) which leads to Laboratory 1 being termed the 'worst performing laboratory'.

The case is made that laboratory 1 had a poorer quantitative (but not qualitative) performance (i.e. more variable than the other laboratories) due to less opportunity to run the method.

How do the two tables in 10f relate to one another? What are the numbers in these tables (there are no legends)?

Statistical analysis of single plate data

The statistical methods that are used to identify statistically significant differences are not clearly defined. It appears that t-tests have been carried out for different concentrations against the negative control wells.

The decision criteria for a positive result (a significant difference in a t-test AND a fold change) is based upon relative (fold) difference rather than an absolute (because of variability in the Imax value) difference. Could this be an issue? The 1.5-fold concentration is derived by 'linear interpolation of the value above and below the threshold'.

Diagnostic statistics

Cooper statistics (standard statistics used in diagnostic tests) are presented. In practice, these convey limited information given the relatively small n values. Good statistical practice would have them presented with confidence intervals.

Experimental design

The 96 plate design is susceptible to biases in the allocation of the test articles. For instance, there may also be edge effects or other localized effects on the plates. The current design may introduce bias and increase variability.

It is suggested to look at Nature Biotechnology paper on plate design (Nathalie Malo, James A Hanley, Sonia Cerquozzi, Jerry Pelletier & Robert Nadon (2006) Statistical practice in high-throughput screening data analysis. Nature Biotechnology 24, 167 – 175.

Handling of censored data

The method for handling censored data (e.g. >2000) seems to be inconsistent. In some cases the mean is calculated including the censored data, in some cases excluding it. Including or excluding

censored data will affect the size of the standard deviation and consequently the CV. The choice of inclusion or exclusion seems to be subjective.

Negative control data

1.8 Variability of 20% presumably relates to a CV of 20%.

Other points

The data set for 2-EHA is interesting in terms of reproducibility, effect of cytotoxicity. Rep2 for experiment 1 just reaches a fold increase of 1.5. Visually, it is the most different pattern of the 9 replicate experiments. Four out of 9 experiments seem to be reported as negative because of cytotoxicity.

The reviewing statistician had not been involved in study design and original study evaluation, but was from Givaudan the developer of the test method and is therefore not independent.

3. Test definition (Module 1)

3.1 Quality and completeness of the overall test definition

General comments:

Overall , the test definition is complete and clearly outlined:

- the scientific background on which the test procedure is based;
- the intended purpose of the test as well as the need for the suggested test;
- the technical details of test procedure:
    - description of the test system;
    - parameters and endpoints measured;
    - quality criteria;
    - definition of positive and negative controls as well as benchmarks;
    - acceptance criteria applied to the results.

Specific comments:

An important limitation in the test definition is that during the prevalidation process acceptance criteria drifted, resulting in 3 sets of data (See section 2.5, p14).

In parallel cytotoxicity was assessed and expressed as an IC 50 value (MTT assay), but this was not specifically mentioned (See section 2.5, p14).

As the transfected cell line expresses a certain luciferase luminescence the fold increase over this background was assessed. This is determined for the full dose response curve ranging from 0,98 µM to 2000µM. From this curve an EC1.5 value was determined (gene induction reaching 50% over the background).

Induction relative to DMSO background in each well was measured. The experiment was accepted when the coefficient of variation for the DMSO background calculated from 3 triplicate plates in each test (18 x 96 well plates) for repeated measurements was below 20%.

Cinnamic aldehyde also must be positive in each accepted test.

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

General comments:

The provided background information is adequate, clearly formulated and justified by the dossier (see Module 1: Test definition) (See also sections 2.2 and 2.3 of this report).

3.3 Quality of the documentation and completeness of SOPs and prediction models

General comments:

As submitted, the documentation (including the SOP) contained a number of ambiguities with respect to how the KeratinoSens test was to be performed and how the final prediction was derived.

The WG requested more information from Givaudan for clarification of the observed issues (For a list, see section 2.5, p14). On the basis of the new information and after update of the submitted material by Givaudan, the WG concluded that the schematic outline (Fig. 2) was clear, and that the it was transparent which acceptance criteria Givaudan recommends for future use of the assay.

Specific comments:

The raw data entering the processing with the Excel template should have been added for all measurements not just for 2 examples, so that a complete reconstruction of data provided was possible.

4. Data quality

4.1 Overall quality of the evaluated data

General comments:

The data quality was deemed adequate for the purpose of this study. However, a number of limitations should be listed:

- The participating laboratories were not given a list with requirements defining the acceptability of the data produced .
- There was no quality check of the incoming data, except in the transfer phase, and the check differed from those in the SOP.
- Test data for which the test acceptance criteria were not met were included. According to the submitter this was for the purpose of assessing the rigidity of the test acceptance criteria. For the purpose of demonstrating reproducibility these data should have been excluded and/or repeated.

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

General comments:

The quality of the data provided was deemed sufficient to judge WLR. The WG did not reach an agreement regarding the question of whether or not the data on BLR (when taking non-qualified laboratory predictions into account: case 2 and 3) were sufficient to judge reproducibility between laboratories. The reason for this lack of agreement is explained in section 2.5 (p. 15).

4.3 Quality of the reference data for evaluating reliability and relevance[1]

General comments:

The quality of the reference data is considered to be sound (see also section 1). The reference data selected only allowed assessment of sensitizers versus non-sensitizers (S/NS) not an assessment of potency (quantitative responses). The latter is based upon the observations that the test underscores moderate and weak sensitizers (See Module 5).

[1] OECD guidance document Nr. 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."

5. Test materials

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

General comments:

Overall, 114 chemicals were tested. These represented a good number of materials, reasonable structural diversity and a variety of sensitising potency classes.

The number of test items was considered adequate to draw conclusions about the transferability (N=7) and reproducibility (N=21) of the test.

The small number of non-sensitizers (N=4) in the extended list of chemicals (N=48) for assessing the predictive capacity of the test was considered too low. The 67 chemicals used for development, refinement and evaluation of the test were not taken into consideration in assessing the predictive capacity by the WG.

5.2 Representativeness of the test items with respect to applicability

General comments:

The chemicals tested span a range of molecular weights (30-388 Da), and cLogP (-4.8-5.2), cover the full range of skin sensitizer potency (weak-extreme) and included a wide array of structural classes. Pre- and pro-haptens were included. Therefore, the selection of chemicals was sufficient to gain information on the applicability domain and limitation of the test method.

The results indicate that the KeratinoSens assay frequently underscores moderate as well as weak sensitizers, i.e. there is a high false negative rate (Tables 3-5).

Specific comments:

It would have been very useful to have added reactivity data to this data set for all the chemicals (including the origin of the data). This is mainly to substantiate claims about the applicability domain being specific to cysteine reactivity of the sensitzsers. For example, there are materials in this data set which are positive in the KeratinoSens assay are lysine but not cysteine reactive (phenylacetaldehyde, dihydroeugenol, hexylcinnamic aldehyde, hydroxycitronellal, imidazolidinyl urea, benzaldehyde). Furthermore, eugenol and phenyl benzoate are cysteine reactive but negative in the KeratinoSens assay.

Table 3:
Summary results of KERATINOSENS **Extended List** by chemicals of known skin sensitization potency

SENSITIZATION CLASS[a]	NUMBER OF CHEMICALS TESTED	Luciferase EC1.5	No induction	% false negative
Extreme/strong	27	23	4	14.4
Moderate	30	21	9	30.0
Weak	26	20	6	23.1
Non classified	3	2	1	
				% false positive
Non sensitizers	28	5	23	21.7
Total sensitizers	83 (+ 3 non classified)			
Total non sensitizers	28			
Total compounds tested	114			

[a]Sensitization class is based on LLNA EC3 values.

Table 4:
Summary results of KERATINOSENS **Silver list** by chemicals of known skin sensitization potency

SENSITIZATION CLASSa	NUMBER OF CHEMICALS TESTED	Luciferase EC1.5	No induction	% false negative
Extreme/strong	12	11	1	8.3
Moderate	16	13	3	18.8
Weak	16	14	2	12.5
				% false positive
Non sensitizers	23	4	19	17.4
Total sensitizers	44			
Total non sensitizers	23			
Total compounds tested	67			

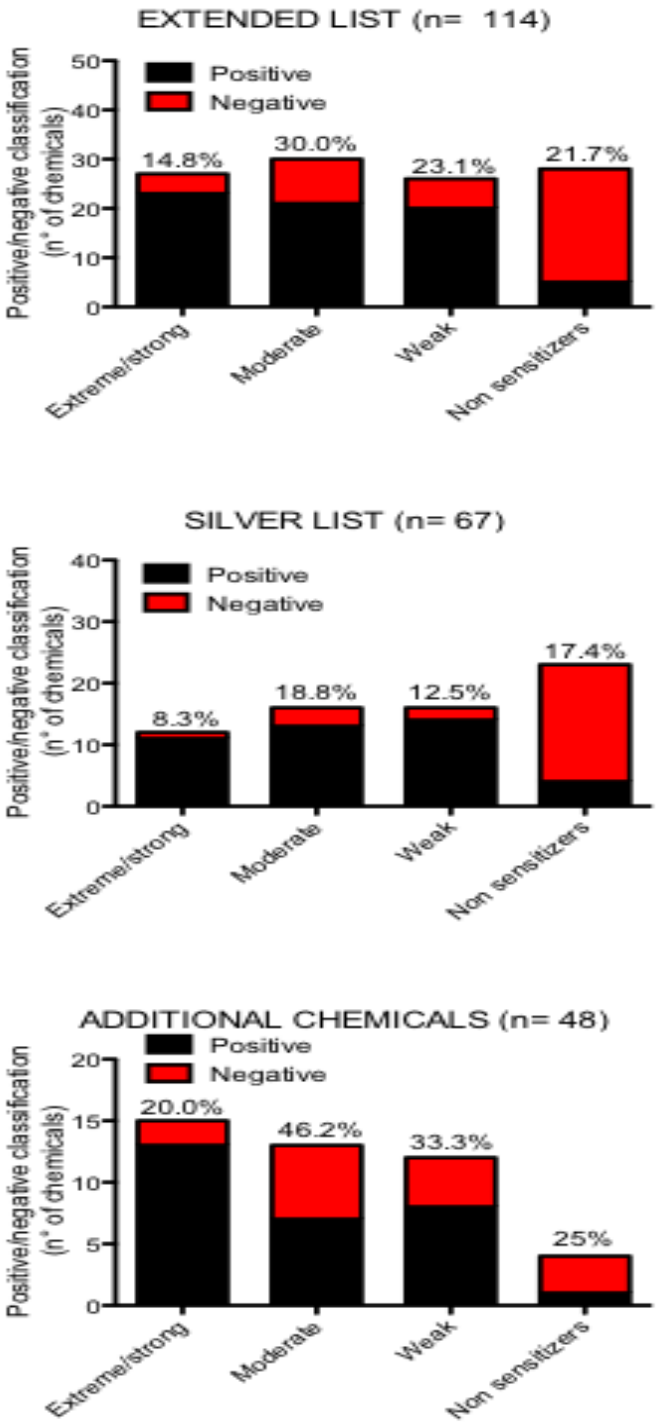
[a]Sensitization class is based on LLNA EC3 values.

Table 5.
Summary results of KERATINOSENS **Additional chemicals** by chemicals of known skin sensitization potency

SENSITIZATION CLASSa	NUMBER OF CHEMICALS TESTED	Luciferase EC1.5	No induction	% false negative
Extreme/strong	15	13	2	20.0
Moderate	13	7	6	46.2
Weak	12	8	4	33.3
Non classified	3	2	1	
				% false positive
Non sensitizers	5	1	4	20
Total sensitizers	40 (+ 3 non classified)			
Total non sensitizers	4			
Total compounds tested	48			

[a]Sensitization class is based on LLNA EC3 values.

Figure 3: Representation of potency classes in the various studies.

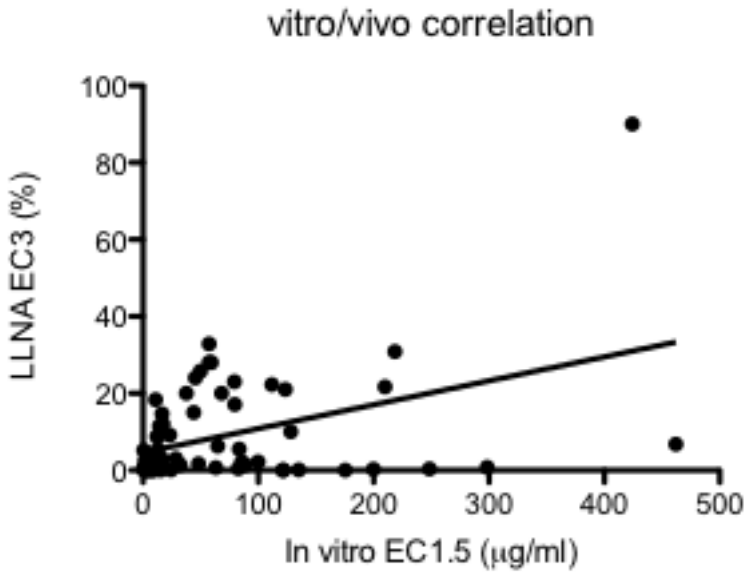


In the following page a graphic representation of the three tables is presented (Fig, 3). For sensitizers the red bar represents chemicals misclassified. The percent of chemicals misclassified (false negative for sensitizers, false positive for non sensitizers) is reported on the top of each bar.

Regarding the correlation between in vitro EC1.5 and *in vivo* LLNA EC3, the following values are reported (Fig. 4):

Figure 4: Correlation between *in vitro* EC1.5 and *in vivo* LLNA EC3.

	In vivo LLNA EC3
Pearson r	0.4226
95% confidence interval	0.1992 to 0.6043
P value (two-tailed)	0.0005
P value summary	***



6. Within-laboratory reproducibility (Module 2)

6.1 Assessment of repeatability and reproducibility in the same laboratory

General comments:

The WLR assessment was presented by the submitters using two studies and types of analyses:

1) Analysis of variability of the measured parameters (e.g. EC1.5, IC50, Imax) (N=28)

The within-laboratory reproducibility (WLR) was initially assessed by running a ring-trial using 28 chemicals selected from the same list (Silver list, N=67) that had provided the chemicals for optimising the test (e.g. the prediction model) by the lead laboratory (Casati et al.).

With regard to this initial assessment, the WG believes that it was the intention of the test submitters to analyse WLR on the basis of the test values including EC15, IC50 and Imax (see section 2.5). As chemicals were tested only once (=1 experiment) in triplicates, rather than in three consecutive experiments, and considerable variation between the triplicate values within an experiment were observed, the test values could not, and were not, used to assess test reproducibility.

The correlation between in vitro EC1.5 and in vivo LLNA EC3 was o.43 (CI 0.20-0.60) was reported (Fig. 4).

The WG compared the predictions obtained by the lead laboratory during this ring trial with historical data on these 28 chemicals generated by the lead laboratory (when testing 67 chemicals of the silver list (attachment 2)). When comparing these sets of predictions the apparent concordance in predictivity for the 28 chemicals was 96%. It should be noted though that such comparative analysis of data is not good practice for the purpose of determining WLR within the context of a (pre)validation study.

2) Analysis of concordance of prediction (N=14)

An additional series of 14 chemicals was run in the lead laboratory, including eight coded chemicals supplied by ECVAM (double-blind) and six additional chemicals selected by Givaudan. These six chemicals had not been previously tested and were blinded only to the operator. The WG considers this study as the primary source for assessing reproducibility of the test method.

When considering all available data, concordant predictions were obtained for 12 out of 14 substances over 3 consecutive experiments (85.7% concordance). Non concordant predictions were obtained for R(+)Limonene and 2-Ethylhexyl acrylate.

The defined acceptance criteria were not met for three substances due to the positive control (cinnamic aldehyde) being slightly out of the specified range. These substances were Beryllium sulphate, R(+)Limonene and 4-amino benzoic acid.

Additional input requested: To understand better the impact of unqualified data on the prediction in terms of S/NS, the test submitter was requested (in writing) to re-assess the reproducibility of the test by redoing the calculations only on the basis of test values that qualified according to either set of criteria (Annex 4, p60). The updated calculations were submitted (Annex 4, p61 (C. 1)).

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

General comments:

Concordant results were obtained for 12/14 chemicals (85.7%) when all available data were included. The WG endorsed the conclusion of the VMG that the test is reproducible within laboratories. WG considered this concordance in agreement with target values (85%) for WLR performance standards as published in international accepted guidelines (e.g. Performance standards of TG439 *in vitro* skin irritation).

The ESAC WG agreed that the re-analysis that was submitted upon request (see section 6.1) satisfactory answered the question to which extent non-qualified test results might have influenced the WLR analysis**.** The impact on WLR was considered negligible as even under the most stringent criteria (set 2 in Annex 4, p60) only 3 individual laboratory predictions had not qualified.

7. Transferability (Module 3)

7.1 Quality of design and analysis of the transfer phase

General comments:

In general, the transferability assessment of the test was well-designed, e.g. preparation of an SOP, use of the same batches of seven chemicals, a rationale for chemical selection, and assurance criteria of data generated from external laboratories.

Even though potential problems affecting test performance were anticipated, no face-to-face training was performed as the test submitter did not deem this necessary.

Specific comments:

The SOP (Attachment 7a, 2009-17-07 version) used for transferability is a different version, even though very similar to the INVITTOX protocol KeratinoSens (last up-dated at 2010-20-08; Attachment 1). There is no description of the differences included in the report, nor is it explained why a new version was preferred.

Data quality (Attachment 8a & 8b) from external laboratories was assessed by the lead laboratory. Assessment was on the basis of the 'DMSO wells' variability, dose-response curve reproducibility within the laboratory, and significant induction by the positive control cinnamic aldehyde. Even though the criteria for acceptance of 'external' data were explained, no detailed assessment procedure (e.g. statistical analyses) was given.

Additional input requested: To understand better the impact of unqualified data on the prediction in terms of S/NS, the test submitter was requested (in writing) to re-assess the reproducibility of the test by redoing the calculations on the basis of test values that qualified according to either set of criteria (Annex 4, p60). The updated calculations were submitted (Annex 4, p62 (C. 3)).

7.2 Conclusion on transferability to a second laboratory as assessed by the study

General comments:

The conclusion on transferability was justified on the basis of concordant predictions (S/NS) between the lead laboratory and the naive laboratories. The WG endorses the conclusion that the test method can be transferred to naive laboratories that are experienced with cell culture techniques.

Specific comments:

As the SOP (INVITTOX protocol KeratinoSens: Attachment 1) was suggested to be sufficiently detailed to perform the test, no face-to-face training was organized before transfering the test to naive laboratories. However, the WG raised concerns about the reliability of luciferase measurements for transferability. Differences in the brand of luminometer or substrate were demonstrated by the test submitters not to affect the liability of the luminescence measurement. Based on this fact, it seems obvious to the WG that the variation observed in luminescence measurements between laboratories is due to lack of experience, stressing the necessity of operating a number of training experiments in the naïve laboratory before the test method can be used to identify skin sensitizers.

Variability among the laboratories was observed in the dose-response curve or EC1.5 (Attachment 8a & 8b) for cinnamic aldehyde and ethylene glycol dimethacrylate. No further explanation was however given whether these variabilities originated from the chemicals' own physico-chemical characteristics or from luminescence measurement issues.

8. Between-laboratory reproducibility (Module 4)

8.1 Assessment of reproducibility in different laboratories

General comments:

As for the WLR, the BLR was assessed on the basis of concordance between results, accuracy and data dispersion by comparing geometric standard deviation of both EC1.5 and IC50 values. In total 21 chemicals were tested.

The BRL was assessed by a ring-trial involving 4 naive laboratories and the lead laboratory.

Additional input requested: To understand better the impact of unqualified data on the prediction in terms of S/NS, the test submitter was requested (in writing) to re-assess the reproducibility of the test by redoing the calculations only on the basis of test values that qualified according to either set of criteria (Annex 4, p60). The updated calculations were submitted (Annex 4, p61 (C. 2)).

Specific comments:

It was not clear to the WG how/where the coding was performed and how 'blinding' was assured.

If the question is one of criteria reproducibility, then this is not very good when looking at data from laboratories 2, 3 and 4. If the question focuses on concordance, reproducibility is better.

- Case 1 (all data included): the concordance is very good (85.7 %).
- Case 2 (only qualified predictions based on Criteria set 1): The reproducibility is not acceptable due to poor performance of laboratories 2, 3 and also 4. Considering only the substances for which at least 3 laboratories produce acceptable data (N = 17) the concordance is 88%.
- Case 3 (only qualified predictions based on Criteria set 2): The concordance is acceptable for laboratories 1, 4 and 5 (> 85%) (N = 21). Too many unqualified runs are observed with laboratories 2 and 3.

The above indicates that perhaps a formal training phase would have been useful after all (See also section 7.2).

Givaudan, when planning the ring trial had been unaware that it is common practice in the context of validation to use three laboratories within a ring trial for assessing transferability and between-lab reproducibility (BLR). While it was acknowledged by the WG that with an increasing number of laboratories it may be more likely to get non-concordant results, the WG agreed not to consider the

analysis by the test submitter of the concordance of predictions on the basis of four laboratories only instead of the five laboratories that had participated in the ring trial.

8.2 Conclusion on reproducibility as assessed by the study

General comments:

The S/NS prediction gave concordant results for the majority of chemicals (85.7 – 90.5%), taking into consideration the explanations give for the outliers, also between laboratories. (See section 6.1).

The test acceptance criteria provided to the participating laboratories during the ring trial had not been applied when analysing the data. The reason for this inconsistency was that the criteria were found to be too stringent. In contrast with WLR and transferability assessment, these non-qualified data had an effect on the concordance of predictions (Annex 4, p62 (C. 2)).

No provisions were made for re-testing in the case of nonqualified predictions.

Specific comments:

It was noted by the WG that an attempt was made to assess the BRL using 5 laboratories. It is acknowledged that ECVAM requires only 'at least' 3 laboratories participating in a prevalidation study. It is also acknowledged that the test developer involved 5 laboratories thereby raising the bar significantly. However, it is not appropriate to choose the best 3 of 5 where the data is best. For the purposes of reproducibility assessment one may consider all 5 labs or choose 3 laboratories upfront to compare.

Eleven out of 15 rated positive in all 5 laboratories, while 4 out of 6 non sensitizers were correctly classified. Contradictory results were obtained for Eugenol (S in 2 laboratories, and NS in 3 laboratories). The explanation given for the Eugenol reactivity should be reassessed.

The irritants diethylphthalate and SDS were positive in one laboratory. Phenylbenzoate was a clear false negative.

9. Predictive capacity (Module 5)

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

General comments:

The WG was impressed by the wealth of information provided by the test submitter on the 113 chemicals assessed in this study. Based upon the 113 chemicals included in the study, the predicitive capacity of the KeratinoSens assay was 78%.

The 114 chemicals include the 67 chemicals of the Silver list. Including chemicals that were used for development, refinement and evaluation of a test system might induce a bias in the assessment of the predictive capacity and was therefore considered by the WG as a limitation.

Considering only the new chemicals (43 sensitizers and 3 non-sensitizers), the calculations showed that the predictive capacity (69%) was considerably lower than the 78% presented by the submitter. It was noted that the number of new qualified non-sensitizers used in this study was considered insufficient (N = 3).

Additional input requested: The submitters were requested to submit additional data on chemicals with negative LLNA reference data. Such data were provided for an additional 80 chemicals.

Compiling all the data provided by the submitters (N = 220), the KeratinoSens assay revealed a sensitivity, specificity and accuracy of 79.3%, 79.8% and 79.5%. Omitting the reactive peptide alkylating chemicals, for which the LLNA data were not trusted and no human data were available, the remaining 213 chemicals resulted in a sensitivity, specificity and accuracy of 79.3%, 84.5% and 81.7% (Annex 4, p64 (C8)).

The WG observed a poor performance of the test with weak sensitizers. Based on the predictions using the 113 chemicals, 41% of the weak and 86% of the very weak sensitizers were missed (Table 6). Furthermore, the frequency false negative results were found to increase with decreasing potency of the test chemical. This limitation is not indicated clearly in the submission.

Table 6 - Distribution of the predictive capacity over the potency classes

SENSITIZATION CLASS	N° of chemicals	Luciferase induction (positive)	No luciferase induction (negative)	% false negative within each class
Extreme	6	6	0	0
Strong	12	10	2	17
Moderate	39	34	5	13
Weak	27	16	11	41
Very weak/none	7	1	6	86

Specific comments:

The test submitters assessed the predicitive capacity of the test in three steps. Based on the calculations the predictive capacity ranged between 96.7 and 78%. However, for each step careful considerations have to be made.

- Step 1:
  Predicitive capacity based on the screening in the lead laboratory of 67 chemicals (Silver list) was **85.1%**, but chemicals were used for development, refinement and evaluation of the test.

- Step 2:
  Predictive capacity based on the screening of the 28 chemicals of the Ring study was **85.4-96.7%**, but the 28 chemicals were a subset of the Silver list chemicals.

- Step 3:
  Predictive capacity based on the entire list of 114 chemicals was **78%**, but only 46 new chemicals were included.

It is interesting to note that similar figures can be obtained using the AREc32 cell line (a stable human breast carcinoma cell line transfected with an ARE element) (Natsch and Emter, Tox Sc 102, 2008).

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

General comments:

While the presented calculation of the predicitive capacity of the KeratinoSens was criticized, the WG is confident that the test can identify strong and moderate sensitizers. However, the WG is less confident that this test can identify weak sensitizers and possibly also moderate sensitizers at the lower end of the scale.

The accuracy, sensitvity and specificity appears acceptable. If we consider all the chemicals tested, the overall accuracy is 79.5%, with a sensitivity of 79.3% and a specificity of 79.8, which can be considered acceptable.

However, the criticism and the necessity to underlie limitations in term of identification of weak and moderate sensitizers still remain.

10. Applicability domain (Module 6)

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

General comments:

The applicability domain was described in the section 1.6 of KeratinoSens report. The authors stated a variety of chemical classes which were expected to be successfully tested in the KeratinoSens assay. The exclusions were mainly related to issues of solubility or stability in vehicle (e.g. interactions with the vehicle, such as hydrolysis).

The WG discussed this issue (See section 2.2) and came to the conclusion that there is indirect evidence that the applicability domain of the test may extent to chemicals that not (only) react with the cysteine residues of Keap1. Alternative mechanisms may lead to Nrf2 activiation.

The study design allowed testing of some of the limitations of the applicability domain.

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

General comments:

The anticipated applicability domain of the KeratinoSens assay is described in the section 1.6.3. (KeratinoSens report). The clearly explained limitations are for chemicals with extremes of cLogP which would not be testable in this assay due to poor solubility. Another clear limitation adequately explained is related to the stability of the chemical in the solvent system (DMSO, water), resulting in hydrolysis or interaction with DMSO, thus resulting in a change of the test material through preparation. These limitations are common to many in vitro/in chemico test methods.

The applicability domain which is less well described concerns chemical reactivity. The consensus opinion on reactivity is unclear. Reactivity is described in section 1.1.2 as covalent modification of the key cysteine residues on Keap1, resulting in activation of the Nrf2 signalling pathway,. From the literature it is clear however, that this pathway can be activated by other means – such as oxidation of the key Cys residues on Keap1 as well as other type of modification of Keap1 Cysteines (e.g. glutathionylation, see Holland et al 2008). This is potentially the reason why some of the exclusively Lys reactive residues can be positive in KeratinoSens assay: whilst such chemicals generate adducts with Lys and not Cys residue, they can potentially oxidise those cysteines generating the same effect. The argument about exclusivity of Lys reactivity of some chemicals is not always applicable, as some

of the materials quoted as lysine reactive are also cysteine reactive, and therefore should be positive in this assay (such as phenyl benzoate, see Aleksic et al 2009). There are also exclusively Lys reactive chemicals which are positive in this assay.

The suggested inclusion of the peptide reactivity assay with KeratinoSens in an integrated testing strategy is a good one, providing that the peptide reactivity does make a distinction between lysine and cysteine reactivity. The assay suggested in the attachment 12c of the TST does not provide this distinction.

The added value of the cell based assay over a reactivity assays would be a possibility of some metabolic activation being available to capture the true prohaptens, which would not be reactive in direct peptide reactivity assays. The activation mechanisms of this particular cell line appear to be poor.

In conclusion, the stated applicability domain is likely to be broader than that stated in the TST.

11. Performance standards (Module 7)

11.1 Adequacy of the proposed Essential Test Method Components

Not relevant

11.2 Adequacy of the Reference Chemicals

Not relevant

12. Readiness for standardised use

12.1 Assessment of the readiness for regulatory purposes

General comments:

The WG considered the test method sufficiently mature for classification and labelling of chemicals (relevant to Regulation EC N° 1271/2008).

Negative results however have to be considered with care as weak sensitizers (and possibly also moderate sensitizers at the lower end of the scale) will be probably missed (see section 9).

Unless this issue gets resolved, the KeratinoSens assay has to be seen as one brick in an intergrated testing strategy of weight-of-evidence approach. The consideration of the chemistry /reactivity must be included either by combination with a peptide reactivity test or predictive chemistry assessment. This reactivity assessment should include consideration concerning activating mechanism(s).

Specific comments:

The performance of the Keratinosens assay extended list is acceptable (Specificity = 82.1%; Sensitivity = 76.7%), if a specificity = 80% and sensitivity = 70% is the accepted criteria.

Furthermore, the possibility to use the EC1.5 values for sensitization potency classification is unlikely from the data presented.

It is likely that the combined analysis of this and/or other biomarkers rather than the analysis of a single biomarker will give even more satisfactory results. It is anticipated that the combination of different *in vitro* assays will increase the accuracy, i.e., in the KeratinoSens assay, the inclusion of the peptide reactivity data will increase accuracy from 85.1% to 89.6% (Emter et al., 2010). The h-CLAT has an accuracy of 75.9% with CD86 alone, while 93.1% in combination with CD54 (Sakaguchi et al., 2009).

It is, however, important to note that the combination of the KeratinoSens with the DPRA increases sensitivity (correct identification of positive compounds) to 84.9% but its specificity will decrease to 78.6%. The rate of false positive will increase.

12.2. Assessment of the readiness for other uses

General comments:

The KeratinoSens assay was considered useful for screening purposes, to identify molecular initiators and to gain mechanistic information on the role of e.g. oxidative stress in sensitization.

12.3 Critical aspects impacting on standardised use

General comments:

Since the applicability domain of the test is still not fully defined, and probably is not limited to direct reactivity of a chemical with key cysteine residues in Keap1, it can be anticipated that chemicals are excluded from testing for the wrong reason.

The dependence of the predictive capacity of the test on decreasing potency has to be considered with care (See section 9).

The WG was concerned about an intellectual property (IP) issue, related to the use of the luciferase gene from Promega. A license from Promega is required for commercial uses. This may have economic consequences (attachment 1, p22).

The use of the Promega luciferase gene in the KeratinoSens cell line is still linked to the assay substrate from Promega. A question is whether it is mandatory to use Promega's substrate, or whether one purchase a similar substrate from other companies?

The 96 plate design as submitted was susceptible to biases in the allocation of the test articles. There may be edge effects resulting in bias and increased variability.

12.4 Gap analysis

General comments:

Weak and low-moderate sensitizers, as well as pro-haptens performed poorly.

When considering cytotoxicity, more emphasis could have been given to the GSH status of the cells and their GSH regenerating capacity. This system may have an impact on the inherent chemical reactivity whether directly conjugating to GSH or oxidising it.

The data do not support the expectation that this test can be used as a stand-alone test (preliminary, waiting for PC and reproducibility assessment).

The correlation between *in vivo* and *in vitro* data is weak because there was a relative high variability among the *in vitro* scores of chemicals belonging to the same potency class (Natsch et al., 2009).

13. Other considerations

The report mentioned the high power of the study design. Some indication of the size of effects that can be reasonably detected would be useful.

The chemicals selected from the silver list (N=28) for reproducibility assessment were also used for test development, refinement and evaluation. This may induce a bias because the test may have been optimized for these chemicals.

14. Conclusions on the study

14.1 Summary of the results and conclusions of the study

14.1.1 Test items

Overall, 114 chemicals representing structural diversity and a variety of sensitising potency classes were tested. The chemicals tested span over a range of molecular weights (30-388 Da), of cLogP (-4.8-5.2), cover the full range of skin sensitizer potency (weak-extreme) and widely differing structural classes were tested. Pre- and pro-haptens were included as well.

14.1.2 Summary of study results

WLR (14 chemicals):

Concordant results were obtained for 12/14 chemicals (85.7%) when including the available data.

Transferability (7 chemicals):

The conclusion on transferability was that the test method can be transferred to naive laboratories that are experienced with cell culture techniques.

BLR (21 chemicals):

Among the 21 chemicals tested, 11 out of 15 were called positive in all 5 laboratories, while 4 out of 6 non sensitizers were correctly classified. Some irritants were called positive in one laboratory.

The VMG considered the test reproducible in the laboratories involved in the study.

Predictive capacity:

The test submitters assessed the predicitive capacity of the test in three steps. Based on the calculations the predictive capacity ranged between 78% and 96.7. However, for each step careful considerations have to be made.

- Step 1:
  Predicitive capacity based on the screening in the lead laboratory of 67 chemicals (Silver list) was **85.1%**, but these chemicals were used for the development, refinement and evaluation of the test.

- Step 2:
  Predictive capacity based on the screening of the 28 chemicals in the Ring study was **85.4-96.7%**, but these 28 chemicals were a subset of the Silver list chemicals.

- Step 3:
  Predictive capacity based on the entire list of 114 chemicals was **78%**, but only 46 new chemicals were included.

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

Test chemicals:

A total of 114 chemicals was tested which represented a good number of materials, reasonable structural diversity and a variety of sensitising potency classes. The chemicals tested extended over a range of molecular weights (30-388 Da), of cLogP (-4.8-5.2), cover the full range of skin sensitizer potency (weak to extreme) and differing structural classes. Pre- and pro-haptens were also included. Therefore, the selection of chemicals was considered sufficient to gain information on the applicability domain and limitation of the test method.

The number of test items was considered sufficient to draw conclusions about the transferability (N=7) and reproducibility (N=21) of the test.

The small number of non-sensitizers (N=4) in the extended list of chemicals (N=48) for assessing the predictive capacity of the test was considered too low. The 67 chemicals used for development, refinement and evaluation of the test were not taken into consideration for assessing the predictive capacity by the WG.

WLR (14 chemicals):

When considering all available data including tests that did not meet the specified test acceptance criteria (unqualified tests), concordant predictions were obtained for 12 out of 14 substances over 3 consecutive experiments (85.7% concordance). For three substances the defined acceptance criteria were not met. When excluding these non-qualified results from the data matrix, concordant predictions were obtained for 10 out of 14 chemicals resulting in a concordance of 71.0%.

Including all available data concordant results were obtained for 12/14 chemicals (85.7%). The WG considers this concordance acceptable and in agreement with target values (85%) for WLR performance standards as published in international accepted guidelines (e.g. Performance standards in TG439 in vitro skin irritation).

Transferability (7 chemicals):

In general, the transferability assessment of the test was well-designed, e.g. preparation of an SOP, use of the same batches of chemicals, rationale for chemical selection, and assurance criteria of data generated from external laboratories.

Even though potential problems affecting test performance were anticipated, no face-to-face training was performed as the test submitter did not deem this necessary.

The conclusion on transferability was justified on the basis of concordant predictions (S/NS) between the lead laboratory and the naive laboratories. The WG endorses the conclusion that the test method can be transferred to naive laboratories that are experienced with cell culture techniques.

BLR (21 chemicals):

Among the 21 chemicals tested, 11 of 15 rated positive in all the 5 laboratories, while 4 of 6 non-sensitizers were correctly classified. Some irritants were classed as positive in one laboratory.

The WG discussed in detail the question of whether or not the data on BLR were sufficient to judge reproducibility between laboratories. As explained in section 2.5, the reason for this discussion was the observation that the TACs provided to the participating laboratories during the ring trial (a) had not been applied when analysing the data as they were found too stringent, (b) that no provisions for re-testing had been made and (c) that the final data matrix contained an appreciable number of non-qualified test results that were included in the analysis.

Predictive capacity:

The conclusions regarding the predictivity are sound given the overall value of 76.6%, the key here is that weight of evidence data were used for comparison rather than to a single assay outcome (such as LLNA EC3% values).

Compiling all the data (including additional chemicals with negative LLNA) provided by the submitters (N = 220), the KeratinoSens revealed a sensitivity, specificity and accuracy of 79.3%, 79.8% and 79.5%. Omission of the reactive, peptide alkylating chemicals, for which the LLNA data were not reliable and where not supported by human data, the remaining 213 chemicals resulted in a sensitivity, specificity and accuracy of 79.3%, 84.5% and 81.7% (Annex 4, p64 (C8)).

Negative results cannot, however, exclude the sensitization potential as weak and low moderate sensitizers are likely to be missed.

Applicability domain is less clearly defined with this data set and it is prudent to assess this further by testing the additional set of chemicals with possible issues. It is clear that specific amine reactivity and requirement for some form of activation are not the only issues that may need to be addressed.

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

A large body of impressive experimental work was carried out using a robust test method. Based on the information supplied, the conclusions are plausible.

15. Recommendations

15.1 General recommendations

The test method can be used for S/NS identification of chemicals. Therefore, the test was considered ready for the next steps in the ECVAM process. A Validation study should however include more well-defined non-sensitizing compounds. Furthermore, a consistent use of proposed acceptance criteria is a necessary prerequisite.

Negative results cannot rule out a sensitization potential because the test revealed issues around weak and low moderate sensitizers. This problem needs to be clearly flagged.

On the basis of the data provided, the developer should be advised to evaluate critically the values obtained in laboratories 2 and 3, and justify any new criteria to be applied during validation of the test with the clear proviso that any further modification after this to the protocol would raise serious concerns about the protocol's 'general applicability' and the assay's transferability and reproducibility.

Integration of this assay with other predictive tests when they emerge needs to be based on a better defined applicability domain.

Combination of KeratinoSens with a reactivity based approach needs to include unambiguous identification of reactivity and any specificity associated with it. Furthermore, complex reactivity pathways need to be considered and studied in the context of this assay because some *in chemico* derived reactivity data may be an artefact of the test system rather than a true reflection of actual reactivity of the said chemicals *in vivo.* With specific reactivity in mind, consideration needs to be given to other effects which are reactivity-driven in a cell, such as the depletion of GSH via direct conjugation and chemically-driven GSH oxidation

Training should be considered for laboratories that have no experience with the KeratinoSens test.

15.2 Specific recommendations (e.g. concerning improvement of SOPs)

The 96 plate design is susceptible to biases because of the allocation of the test articles. The test submitters were recommended to look at Nature Biotechnology paper on plate design (Nathalie Malo, James A Hanley, Sonia Cerquozzi, Jerry Pelletier & Robert Nadon (2006) Statistical practice in high-throughput screening data analysis. Nature Biotechnology 24, 167 – 175.

16. References

Ade N., Leon F., Pallardy M., Peiffer JL, Kerdine-Romer S., Tissier M.H., Bonnet P.A., Fabre I. and Ourlin J.C. (2009) HMOX1 and NQO1 genes are upregulated in response to contact sensitisers in dendritic cells and THP-1 cell line: role of the Keap1/Nrf2 pathway. Toxicol Sci. 107(2): 451-460

Adler S, Baketter D, Creton S, et al. (2009) Alternative (non-animal) methods for cosmetic testing: current status and future prospects – 2010. Arch Toxicol DOI 10.1007/s00204-011-0693-2.

Aleksic M., Thain E., Roger D., Saib O., Davies M., Li J., Aptula A., Zazzeroni R. (2009) Reactivity Profiling: Covalent Modification of Single Nucleophile Peptides for Skin Sensitisation Risk Assessment. Toxicol Sci. 108(2), 401–411

Barrett JC and Ts'o PO (1978) Relationship between somatic mutation and neoplastic transformation. *Proc. Natl. Acad. Sci. USA*, **75**: 3297-301.

Boverhof D.R. et al (2009) Evaluation of a toxicogenomic approach to the local lymph node assay (LLNA). Toxicol. Sci. 107(2): 427-39

Casati, S.,Aeby, P.,Kimber, I.,Maxwell, G.,Ovigne, J. M.,Roggen, E.,Rovida, C.,Tosti, L.and Basketter, D., (2009) Selection of chemicals for the development and evaluation of in vitro methods for skin sensitisation testing. Altern Lab Anim 37, 305-12.

Combes R, Balls M, Curren R, Fischbach M, Fusenig N, Kirkland D, Lasne A, Landolph J, LeBoeuf R, Marquardt H, McCormick J, Mueller L, Rivedal E, Sabbioni E, Tanaka N, Vasseur P and Yamasaki H (1999) Cell transformation assay as predictors of human carcinogenicity. *Alter. Lab. Anim.,* **27**: 745-67.

Dinkova-Kostova A.T., Holtzclaw W.D. and Kensler T.W. (2005) The role of Keap1 in cellular protective responses. Chem. Res. Toxicol. 18: 1779-1791

Emter R., Ellis G. and Natsch A. (2010) Performance of the novel keratrinocyte-based reporter cell line to screen skin sensitisers *in vitro*. Toxicol. Appl. Pharmacol. 245, 281-290

Holland R., Hawkins A.E., Eggler A.L., Mesecar A.D., Fabris D., Fishbein J.C. (2008) Prospective Type 1 and Type 2 Disulfides of Keap1 Protein. Chem. Res. Toxicol. 2008, 21, 2051–2060

Kim, H. J.,Barajas, B.,Wang, M.and Nel, A. E., (2008) Nrf2 activation by sulforaphane restores the age-related decrease of T(H)1 immunity: role of dendritic cells. J Allergy Clin Immunol 121, 1255-1261 e7.

Kobayashi M., Li L., Iwamoto N., nakajima-Takagi Y., Kaneko H., Nakayama Y., Eguchi M., Wada Y., Kumagai Y., Yamamoto M. (2009) The antioxidant defense system Keap1-Nrf2 comprises a multiple sensing mechanism for responding to a wide range of chemical compounds. Mol Cell Biol 29(2): 493-502

Ku, H.O. et al (2008) Analysis of differential gene expression in auricular lymph nodes draining skin exposed to sensitisers and irritants. Toxicol Lett 177(1):1-9

Lee DF, Kuo HP, Liu M, Chou CK, Xia W, Du Y, Shen J, Chen CT, Huo L, Hsu MC, Li CW, Ding Q, Liao TL, Lai CC, Lin AC, Chang YH, Tsai SF, Li LY, Hung MC. (2009) KEAP1 E3 ligase-mediated downregulation of NF-kappaB signaling by targeting IKKbeta. Mol Cell. 2009 Oct 9;36(1):131-40.

Maruyama A. and Itoh K. (2005) The role of Nrf2 in the protection against inflammation and innate immunity. Hirosaki Med. J. 59: S167-171

Megherbi R., Kiorpelidou E., Foster B., Rowe C., Naisbitt D.J., Goldring C.E. and Park B.K. (2009) Role of protein haptenation intriggering maturation events in the dendritic cell surrogate cell line THP-1. Toxicol Appl Pharmacol 238(2): 120-32

Natsch A. (2010) the Nrf2-Keap1-ARE toxicity pathway as a cellular sensor for skin sensitisers – functional relevance and a hypothesis on innate reactions to skin sensitisers. Toxicol. Sci. 113 (2): 284-92

Natsch A. and Emter R. (2008) Skin sensitisers induce antioxidant response element dependent genes: application to the *in vitro* testing of the sensitisation potential of the chemicals. Toxicol. Sci. 102 (1): 110-9.

NTP, NTP website at http://www.ntp-server.niehs.nih.gov.

OECD (2005) Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment. *OECD Series on testing and assessment Nr. 34.*

Python F., Goebel C. and Aeby P. (2009) Comparative DNA microarray analysis of human monocyte derived dendritic cells and MUTZ-3 cells exposed to the moderate skin sensitiser cinnamaldehyde. Toxicol Appl Pharmacol 239 (3): 273-83

17. Annexes

Annex 1

Document date: 23.1.2012
Author: Andreas Natsch, Givaudan *in vitro* toxicology laboratory

KeratinoSens test submission

Response to questions raised by the peer-review panel and forwarded to Givaudan by the ESAC Coordinator on 16.12.2011:

The following request was obtained:
We would kindly like to request the clarifications listed below:

> **1)** a copy of the **full statistical report**, including the printouts of the specific analyses carried out, which was used to produce the Givaudan Statistical Report (*attachment 10f*). The availability of this full analysis may help with the interpretation of the analyses of within-and between-laboratory reproducibility. The WG is currently of the opinion that the results of this second analysis may not be completely consistent with those carried out in attachment 7a-10e and those reported in the paper by Natsch et al (2011) and would therefore like to comprehend the statistical analyses employed for the purpose of the submission package.
>
> **2) clarification** as to whether also non qualified tests (=those not meeting the performance criteria) have been used possibly also in the context of
> a) the analysis of the study "further evaluation of the intralaboratory reproducibility of the KeratinoSens assay to detect skin sensitizers" (*attachment 4c*).
> b) the analysis of the study "further evaluation of the predictivity of the KeratinoSens assay to detect skin sensitizers" (*attachment 12c*).
>
> **3) submission of amended tables that identify the occurrence of non-qualified tests**. This definitely concerns table 2 of the SOP (ring trial), but possibly also other studies/attachments (see point 2).
>
> **4) submission** of amended analyses of reproducibility and (preliminary) predictive capacity on the basis of **qualified test results** only.

Point 1)
The initial statistical analysis, which was also presented in the paper on the ring study, was made directly by the Givaudan *in vitro* toxicology laboratory. This is entirely based on descriptive statistics, as we felt this is sufficient to display the data and to describe the within and between laboratory variation of the data (Original attachments 10d and 10e).
Upon review of the data submission by the ECVAM statistician a more detailed statistical evaluation was requested. The data package was therefore submitted to Dr. Paul Talsma, the Givaudan statistician who was involved neither in the study design nor in the original study evaluation. He

proposed the additional evaluation based on a statistical analysis of the impact individual labs had on the overall data variance for the quantitative parameters. This independent approach was summarized in Attachment 10f and represents an alternative analysis of the quantitative parameters (EC1.5, IC50, I_{max}) and was thus included in the final submission.

As some inconsistencies were detected by the peer-review panel we have re-checked all the data for this latter analysis, and indeed found one mistake in the data-transfer to the statistical software for the parameter IC50, for which, for the lead lab two repetitions and the average were used instead of the three repetitions (shift in data source by one column). This mistake was corrected and the analysis re-run. This is updated in the revised file

Attachment10f_Statistical analysis BLR_revised.doc.

The impact of this mistake is small and only affects the IC50 parameter and not the EC1.5 and I_{max}. The conclusions were not changed as can be seen in this amended document.
As requested we also include the printout of the statistical analysis software of this revised analysis. This can be found in the file

Attachment10f_Statistical analysis BLR_SAS output.doc.

Point 2) to 4):
General consideration:
Certainly this additional requested analysis and clarifications on the effect of the runs not qualified by the quantitative criteria for cinnamic aldehyde are very important, in order to clearly evaluate how sensitive the assay and the conclusions are in regard to these quantitative criteria for the positive control.
However, we may have complicated things by defining these quantitative criteria for cinnamic aldehyde before knowing how important they really are. It may have been a better approach to only define the criterium: 'Cinnamic aldehyde must be positive in each valid run at the low subtoxic concentration selected in the SOP', and then do a post-hoc analysis how the quantitative values of the positive control may affect the results.
It is important to keep in mind, that the first criterium, namely that the positive control is positive at the low, non-cytotoxic range selected for this control (statistically significant induction above the 1.5-fold threshold between 4 and 64 µM) was fulfilled in each run used in the analysis in all studies. Thus in all the trials from the ring study laboratories and in all the trials summarized in attachment 4c, cinnamic aldehyde gave a positive result. In addition, as can be seen in attachment 17a, in 188 consecutive runs in the Lead lab, cinnamic aldehyde was always positive at the selected dose-range. We consider this an important indication of the stability of the assay over time and the robustness of the assay in general.

Point 2 a) (Attachment 4c):
18 runs were performed **for attachment 4c**. As indicated in the report, in two runs the criteria were not fulfilled, but they were just at the borderline:

- In Experiment 2, run 2, plate 2, the automatically calculated EC1.5 was at 30.08 instead of a maximum of 30
- In Experiment 3, run 3, plate 1, the induction at 64 µM was at 1.993 instead of 2.0.

Since these two runs were only in the second digit different from the target, they were not repeated and used for the analysis. Also the dose-response depicted below would clearly indicate these were valid runs. Nevertheless, we have re-analyzed the data now without these data and the updated test report is attached as the file:

Attachment 4b_second intralab study_wo_borderline-runs.pdf

In this revised report all calculations were made without these two borderline runs.
Omitting these data had no impact on the analysis of the intralaboratory variability of the EC1.5 and IC50 values. As an example, the geometric standard deviations of the EC 1.5 values with and without these two runs are shown in Table 1a and Table 1b below.
Omitting these two runs did affect the predictions for three chemicals in Experiment three, as for these three chemicals we then only have two repetitions and a conclusion cannot be made regarding their positive or negative rating in this third repetition. These chemicals include Limonene and Beryllium sulfate, which had already been discussed in the previous version of the report as borderline chemicals.

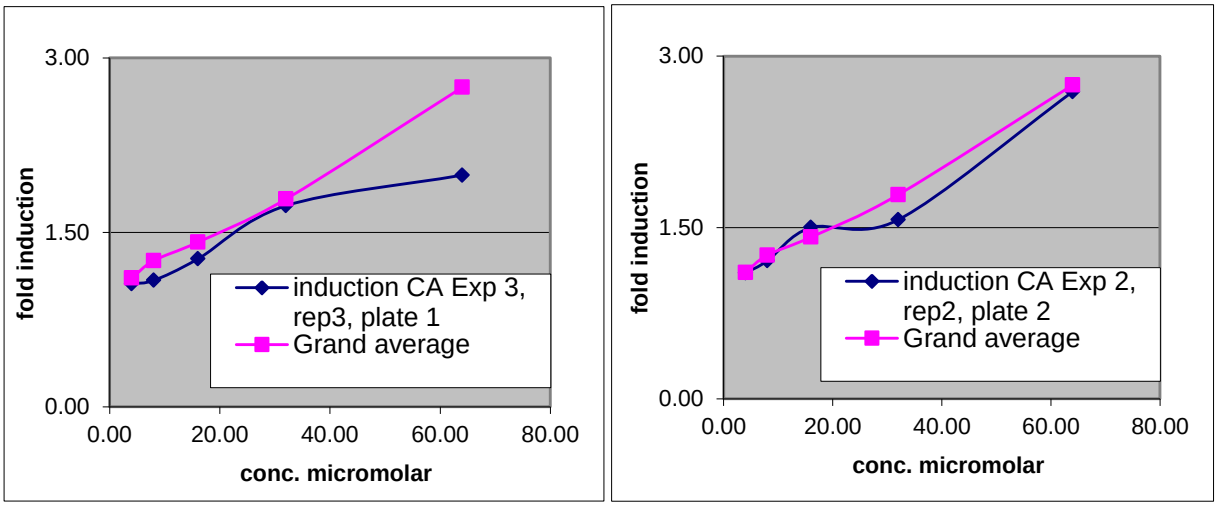


Figure 1. The positive control in the two runs, for which the quantitative criteria for cinnamic aldehyde were not met, compared to the grand average of the positive control in the second WLR study.

Table 1 a. Geometric standard deviation of the EC1.5 values for the positive chemicals including the two runs with borderline results for cinnamic aldehyde

	Experiment 1	Experiment 2	Experiment 3	Between experiments
GMP	1.65	1.57	1.13	1.66
GFN	1.25	1.08	1.07	1.13
GCS	1.08	1.02	1.16	1.05
GBT	1.21	1.41	1.45	1.56
DNBS	1.19	1.12	1.19	1.13
2-EHA	1.41	1.33	1.01	1.13
2-PPA	1.15	1.26	1.12	1.09
4-AmC	1.22	1.30	1.24	1.07
Average	**1.27**	**1.26**	**1.17**	**1.23**

[1] For each Experiment the geometric standard deviation of the three repetitions was calculated
[2] Indicates the geometric standard deviation over the geometric means of each experiment

Table 1 b. Geometric standard deviation of the EC1.5 values for the positive chemicals, excluding the two runs with borderline results for cinnamic aldehyde

	Experiment 1	Experiment 2	Experiment 3	Between experiments
GMP	1.65 [1]	1.57	1.11	1.67 [2)]
GFN	1.25	1.08	1.09	1.13
GCS	1.08	1.02	1.20	1.08
GBT	1.21	1.41	1.65	1.58
DNBS	1.19	1.08	1.19	1.16
2-EHA	1.41	1.42	1.01	1.15
2-PPA	1.15	1.25	1.12	1.14
4-AmC	1.22	1.43	1.24	1.05
		1.28	**1.20**	**1.25**
Average	**1.27**			

Point 2b): (Attachment 12c)

Attachment 12 contains data from the original screening on the 'silver list' – these chemicals were tested with the previous positive control tert-butyl-hydroquinone.

Note: Among these 67 chemicals, 28 chemicals were repeated in the ring study with the positive control cinnamic aldehyde, and the criteria were all fulfilled in all runs in the lead lab in the ring study. Only for one chemical (Hexylcinnamic aldehyde) a discordant result between these two studies performed with the different positive controls was recorded (see Attachment 4a). Since this chemical is borderline also in the BLR analysis, we consider that this change in the positive control had no effect on overall performance.

The additional chemicals were tested with the current positive control and these data were now re-evaluated to analyze the occurrence and impact of runs not meeting the quantitative criteria for cinnamic aldehyde. Runs which clearly did not meet the criteria had been excluded from analysis *ab initio* (such runs are marked yellow in attachment 17a). However some runs were used for the analysis in Attachment 12 c, for which criteria were slightly out of the range. This is summarized in the Table 2 below. In most cases one criterium was fulfilled. In particular, some results with an EC1.5 between 30 and 35 µM were used.

One critical run which had been included is highlighted in bold, in this case both criteria were not met, (fold induction 10.9, EC1.5 at 6.6 µM): However, the two chemicals in Attachment 12c with data from this run (3-aminophenol and 3-Dimethyl-amino-1-propylamine) were repeated in 4 or 6 runs, and thus reporting of this run did not affect the conclusion for these two chemicals in attachment 12c.

If we do analysis without the runs not meeting the quantitative criteria (see Table 2, bold column "reps. Positive w/o non-qual. Runs"), the rating does not change for all but two chemicals:

- Amylcinnamic aldehyde, similarly to hexyl cinnamic aldehyde, is borderline, sometimes positive and negative in some repetitions. The result with the excluded run changes from 3 of 6 positive to 2 of 5 positive, and we would be even less confident of a positive rating
- 2-hydroxypropyl methacrylate (see below)

Besides these two borderline chemicals we have a group of chemicals for which we are left with one repetition, if we exclude the runs not fulfilling the criteria. To be sure whether we had a correct rating in absence of these runs, we now repeated these chemicals in additional runs.

Table 3 shows the quantitative data and rating of these chemicals in the original submission compared with a new analysis including only runs with cinnamic aldehyde within the target range. For all these chemicals we reach the same yes/no rating and similar quantitative data with these new runs fulfilling the quantitative criteria, with the exception of 2-hydroxypropyl-methacrylate. This chemical had before an average EC1.5 at 1025µM just above the threshold, and it falls below the threshold of 1000 µM with the new data. It would therefore be rated positive with the new data, but in the rather weak range.

Note: This is the only chemical which was included as non-sensitizer in Attachment 12c due to negative rating in both the LLNA and guinea pig tests, but which still gives covalent adduct formation with peptides (our unpublished results in Attachment 12c) and (Gerberick et al., 2007; Aleksic et al., 2009). Interestingly 2-hydroxypropyl-methacrylate is known to cause relatively frequent positives human patch test reactions in occupation settings and due to artificial nails (Kanerva et al., 1997; Lazarov, 2007) and thus a weak human sensitization potential certainly exists.

Conclusion Point 2b) / Attachment 12c:

This analysis contained some runs not meeting the quantitative criteria for cinnamic aldehyde. While in some cases the conclusion was based on many repetitions and excluding these runs does not affect the overall conclusions drawn, there were some cases where exclusion of these runs would leave us with insufficient data. In these cases repetitions were run, and the same conclusions were reached as before. The only case where the final conclusions from above analysis is different is amylcinnamic aldehyde, which is similarly borderline as hexylcinnamic aldehyde, and 2-hydroxypropyl-methacrylate which just becomes positive in the revised analysis.

EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection
European Centre for the Validation of Alternative Methods (ECVAM)

Table 2. Chemicals which had been tested with cinnamic aldehyde as positive control included in Attachment 12c

			KeratinoSens result					reps. Positive w/o non-qual. Runs	Positive control cinnamic aldehyde
Name	**CAS-Number**	**LLNA EC3**	I_{max}	EC 1.5	IC50	Pos / Neg	reps. Positive		
Methylisoeugenol	93-16-3	Pos. [3)]	1.4	>2000	815.1	0	0 of 4	**0 of 4**	All within range
4-Methylcatechol	452-86-8	Pos. [3)]	8.1	19.2	71.7	1	2 of 2	**1 of 1**	1 run criteria 64µM at 1.7-fold
1-Bromododecane	143-15-7	Pos. 3)	2.2	44.0	98.0	1	5 of 6	**5 of 5**	1 run criteria 64µM at 1.7-fold
Diphenylmethane-4,4'-diisocyanate	101-68-8	Pos. [3)]	2.4	121.8	>2000	1	2 of 2	**2 of 2**	All within range
Dodecyl methanesulfonate	51323-71-8	Pos. [3)]	2.03	12.14	19.3	at cytotox	3 of 4 at cytotox	**3 of 4 at cytotox**	All within range
4-Nitrobenzyl chloride	100-14-1	Pos. [3)]	93.4	4.0	27.6	1	2 of 2	**2 of 2**	All within range
Chlorpromazine hydrochloride	69-09-0	Pos. [3)]	1.1	>2000	10.1	0	0 of 9	**0 of 8**	8 of 9 within range (attachment 4c)
Beryllium sulfate	7787-56-6	Pos. [3)]	5.7	15.4	51.3	at cytotox	9 of 9 at cytotox	**8 of 8 at cytotox**	8 of 9 within range (attachment 4c)
Methyl methacrylate	80-62-6	90	1.7	424.4	>2000	1	2 of 3	**2 of 3**	All within range
d-Limonene	5989-27-5	52.7	1.2	>2000	82.3	0	0 of 2	**0 of 2**	All within range
Penicillin G	61-33-6	30	10.7	1308.6	>2000	0	0 of 4	**0 of 4**	All within range
Methylhexanedione	13706-86-0	25.8	23.4	49.8	1431.9	1	2 of 2	**2 of 2**	All within range
Cyclamen aldehyde	103-95-7	22.3	3.1	111.9	190.8	1	3 of 4	**3 of 4**	All within range
Geraniol	106-24-1	21.74	2.0	209.8	722.0	1	2 of 2	**2 of 2**	All within range
Butyl acrylate	141-32-2	~20	6.9	37.7	200.9	1	2 of 2	**2 of 2**	All within range
Estragole	140-67-0	20.2	1.3	>2000	419.0	0	0 of 2	**0 of 2**	All within range
Lilial	80-54-6	18.7	1.1	>2000	94.5	0	0 of 2	**0 of 2**	All within range
Amylcinnamic aldehyde	122-40-7	11.5	1.56	14.4	46.8	1	3 of 6	**2 of 5**	1 run criteria EC1.5 at 32 µM
Bromohexane	111-25-1	10	2.0	128.1	391.9	1	2 of 2	**2 of 2**	All within range
Methylanisylidene acetone	104-27-8	9.3	835.8	14.8	159.3	1	2 of 2	**1 of 1**	1 run criteria EC1.5 at 34 µM
Phenylpropionaldehyde	93-53-8	6.3	9.1	64.8	195.1	1	2 of 2	**2 of 2**	All within range
Creosol	93-51-6	5.8	1.0	>2000	>2000	0	0 of 3	**0 of 3**	All within range
3,4-dihydrocoumarin	119-84-6	5.6	1.0	>2000	>2000	0	0 of 2	**0 of 2**	All within range
Farnesol	4602-84-0	5.5	1.6	13.0	23.3	1	2 of 2	**2 of 2**	All within range

trans-2-Hexenal	6728-26-3	5.5	85.4	83.4	802.8	1	4 of 4	**4 of 4**	All within range
Propylidene phthalide	17369-59-4	3.7	1.1	>2000	717.4	0	0 of 2	**0 of 2**	All within range
3-Aminophenol	591-27-5	3.2	1.4	>2000	>2000	0	1 of 6	**0 of 4**	**1 run both criteria not fulfilled**, 1 run EC1.5 at 31.6
3-Dimethyl-amino-1-propylamine	109-55-7	2.2	30.2	85.8	1337.9	1	4 of 4	**3 of 3**	**1 run both criteria not fulfilled**
Diethylmaleate	141-05-9	2.1	60.7	9.4	361.1	1	4 of 4	**4 of 4**	All within range
Methylisothiazolinone	2682-20-4	1.9	22.6	11.8	139.0	1	4 of 4	**4 of 4**	All within range
Trimellitic anhydride	552-30-7	1.42	1.1	>2000	>2000	0	0 of 2	**0 of 1**	1 run criteria EC1.5 at 32 µM
1,3-phenylenediamine	108-45-2	0.49	2.5	82.5	>2000	1	2 of 2	**2 of 2**	All within range
N,N-dimethyl-4-nitrosoaniline	138-89-6	0.48	8.2	0.5	15.1	1	2 of 2	**2 of 2**	All within range
Methyl 2-octynoate	111-12-6	0.45	46.6	2.5	87.6	1	2 of 2	**2 of 2**	All within range
2-amino-phenol	95-55-6	0.4	13.1	1.1	138.2	1	2 of 2	**1 of 1**	1 run criteria EC1.5 at 33 µM
Chloramine T	127-65-1	0.4	50.2	248.4	404.7	1	2 of 2	**2 of 2**	All within range
Propyl gallate	121-79-9	0.32	8.2	199.8	650.3	1	2 of 2	**1 of 1**	1 run criteria EC1.5 at 33 µM
Toluene 2,4-diisocyanate	584-84-9	0.11	4.6	135.0	359.0	1	4 of 4	**4 of 4**	All within range
1,4-Hydrochinone	123-31-9	0.1	16.4	9.8	130.7	1	2 of 2	**2 of 2**	All within range
2,4,6-Trinitrochlorobenzene	88-88-0	0.05	1.6	121.3	616.8	1	2 of 2	**1 of 1**	1 run criteria 64µM at 1.91-fold
2,4-Dinitrothiocyanatobenzene	1594-56-5	0.047	7.2	2.1	6.4	1	2 of 2	**1 of 1**	1 run criteria 64µM at 1.91-fold
Tetrachlorsalicylanilide	1154-59-2	0.04	4.9	<0.98	9.15	1	4 of 4	**4 of 4**	All within range
4-Aminobenzoic acid	150-13-0	>10	1.2	>2000	>2000	0	0 of 2	**0 of 2**	All within range
Benzalkonium chloride			1.5	>2000	4.0	0	1 of 4	**1 of 4**	All within range
Fumaric acid	110-17-8	>25	1.3	>2000	>2000	0	0 of 2	**0 of 1**	1 run criteria EC1.5 at 34 µM
2-hydroxypropyl methacrylate	923-26-2	>50	1.95	1025	>2000	0	1 of 4	**1 of 2**	2 runs criteria EC1.5 at 32/34 µM

Table 3. New analysis including new runs for chemicals included in attachment 12c for which the quantitative criteria for cinnamic aldehyde were not fully met according to Table 2 and for which the conclusion is not sufficiently substantiated after excluding these runs.

		Previous results					With new results, excluding the runs not meeting the quantitative criteria				
Name	**Cas-Nr.**	**Imax**	**EC 1.5**	**IC50**	**Pos / Neg**	**reps. Positive**	**Imax**	**EC 1.5**	**IC50**	**Pos / Neg**	**reps. Positive**
4-Methylcatechol	452-86-8	8.1	19.2	71.7	1	2 of 2	9.1	11.3	61.1	1	2 of 2
Methylanisylidene acetone	104-27-8	835.8	14.8	159.3	1	2 of 2	702.6	15.5	166.6	1	2 of 2
Trimellitic anhydride	552-30-7	1.1	>2000	>2000	0	0 of 2	1.0	>2000	>2000	0	0 of 2
2-amino-phenol	95-55-6	13.1	1.1	138.2	1	2 of 2	19.0	1.7	115.3	1	2 of 2
Propyl gallate	121-79-9	8.2	199.8	650.3	1	2 of 2	6.3	136.7	538.6	1	2 of 2
2,4,6-Trinitro-chlorobenzene	88-88-0	1.6	121.3	616.8	1	2 of 2	4.4	41.4	791.2	1	2 of 2
2,4-Dinitro-thiocyanatobenzene	1594-56-5	7.2	2.1	6.4	1	2 of 2	8.2	2.2	7.5	1	2 of 2
Fumaric acid	110-17-8	1.3	>2000	>2000	0	0 of 2	1.1	>2000	>2000	0	0 of 4
2-hydroxypropyl methacrylate	923-26-2	1.95	1025	>2000	0	1 of 4	2.0	410.8	>2000	1	4 of 5

Point 3) and 4) (Ring study Attachments 8b, 10b, 10c, 12b)

We had previously reported the overall occurrence of tests outside of the quantitative criteria for cinnamic aldehyde in the ring study in Attachment 17b and all the data and dose-response data for the positive control were reported in this attachment. So we were fully transparent on the occurrence of results not meeting these quantitative criteria and on our approach to accept the three consecutive runs performed by the laboratories.
We have now attributed the runs with the criteria outside the range to the individual data points, and this is shown in the revised attachment 8b and 10b, see files:

Attachment8b_Transferability_Table_wo_non-qualified.xls.

and

Attachment10b_BLR_Table_wo_non-qualified.xls

These revised attachments give colour codes for the individual repetitions outside the range, and they give the results for I_{max} (green and faint green), EC1.5 and IC50 (red and orange) with and without these runs included for each individual lab, and they also give the overall results from all laboratories, with these runs included or excluded. To see whether the inclusion of the runs affected the conclusions from the individual labs, it is best to directly compare the two columns for the individual labs in the Excel files of these attachments. In most cases inclusion or exclusion of these runs did not affect the quantitative results, but certainly the statistical power is reduced if we have only 1-2 repetitions in a lab.
To give an overall impression for the quantitative data, below in Figure 2 are shown the overall results for all chemicals from all labs with and without the runs outside of the quantitative criteria included.

Figure 2a. Average of the I_{max} of all five laboratories calculated for all original runs or only those runs with the quantitative criteria within the range.

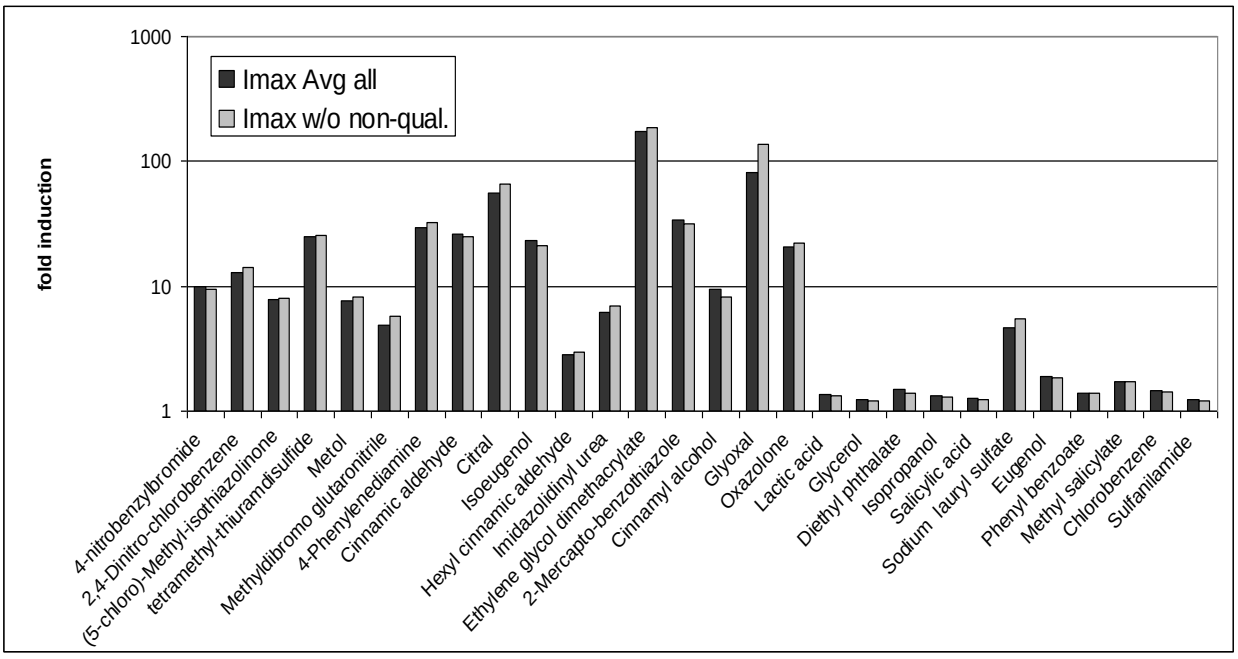


Figure 2b. Geometric Mean of the EC1.5 values from all five laboratories calculated for all original runs or only those runs with the quantitative criteria within the range. Chemicals with no significant gene induction above the 1.5 threshold are excluded from this graph.

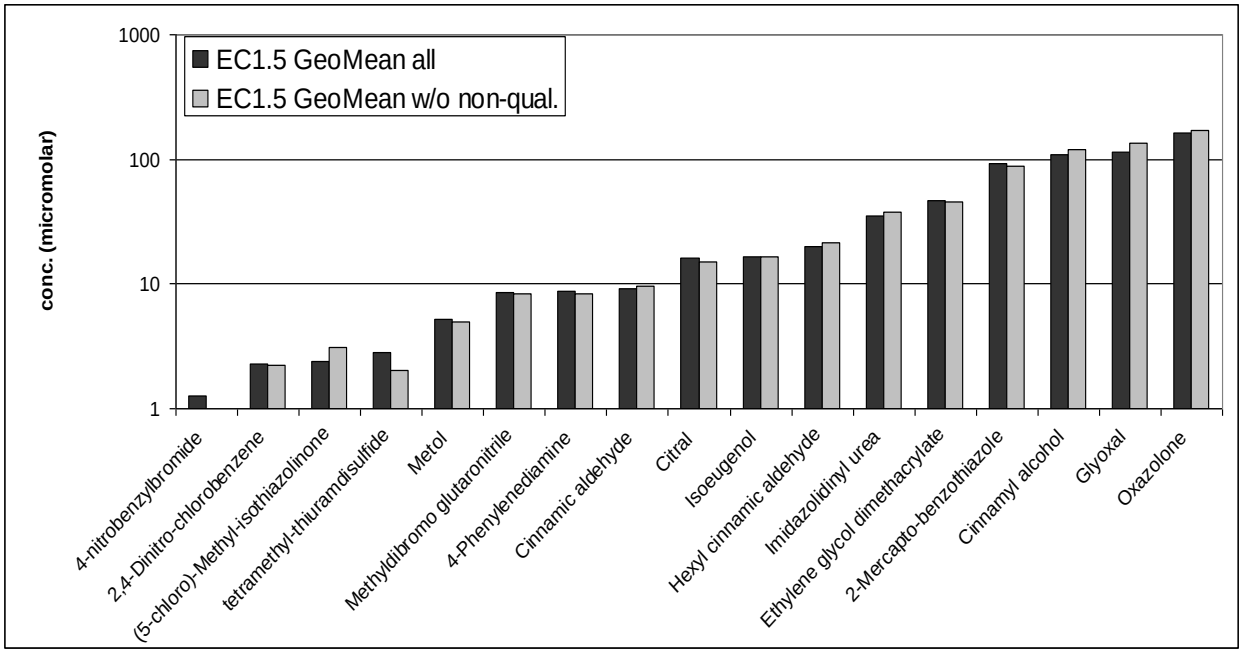
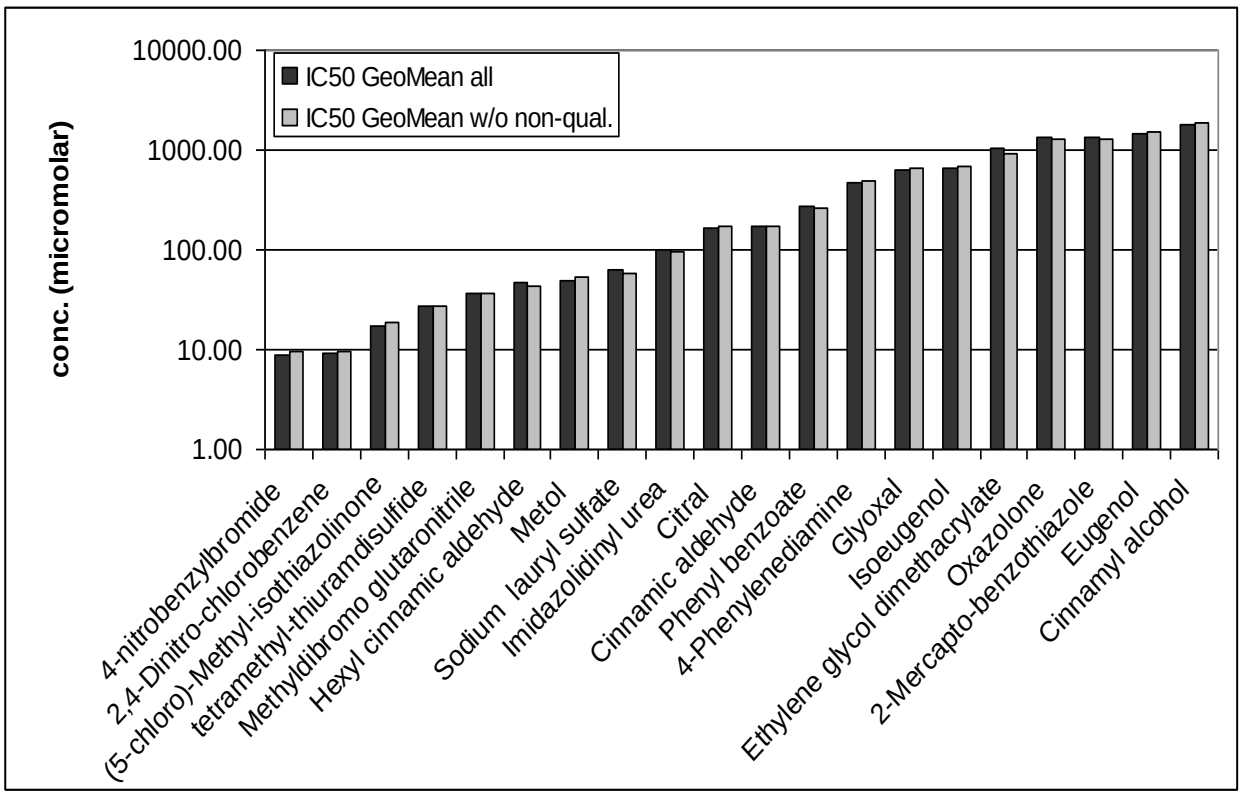


Figure 2c. Geometric Mean of the IC50 values from all five laboratories calculated for all original runs or only those runs with the quantitative criteria within the range. Chemicals with IC50 values over 2000 µM are excluded from this graph.



The revised Attachment 10 b *(Attachment10b_BLR_Table_wo_non-qualified.xls)* also contains a sheet with the predictivity Table (Table 2 in the SOP) excluding the runs outside the quantitative criteria. If we look at this analysis, we certainly obtain significant data-gaps by excluding these runs and the study partly becomes invalid due to lack of data. At this point it becomes very important to discuss the main question:

> ***Are our conclusions regarding transferability and between laboratory reproducibility affected by our approach of directly using the three consecutive repetitions performed by the laboratories, even if the quantitative criteria were not fulfilled?***

As shown above and in the revised attachments, the quantitative results are not affected by this approach, but if we exclude the runs we do have fewer repetitions / fewer laboratories depending on the chemical. Looking at the Cooper statistics table, we would need to exclude some labs due to data gaps.
The case of Lab 2 is most dramatic – in this case the data gaps would lead to complete exclusion of this laboratory from analysis. On the other hand we can specifically look at the result obtained from this lab in comparison to the lead lab, which has fulfilled the criteria in all runs:
Regarding the predictive capacity, this lab obtained the same rating for all chemicals as the lead lab with the exception of Eugenol, which was even correct-positive in Lab2. Based on the statistical analysis in Attachment 10f, the best performing external lab is lab 2, and indeed the quantitative data of Lab 2 are highly comparable to the lead lab as is illustrated in Figure 3.

Figure 3a. Average of the I_{max} of the lead lab with all repetition within the quantitative range for cinnamic aldehyde and of lab 2 (all results, the majority being outside of the quantitative range).

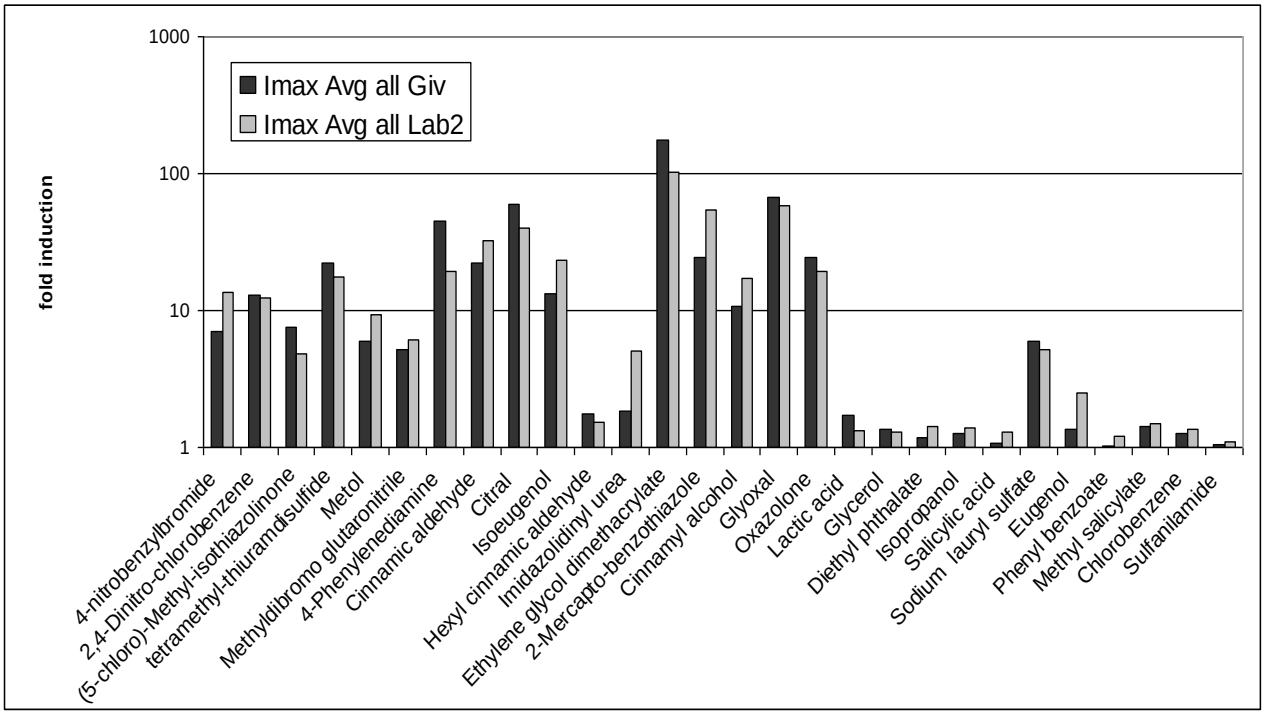
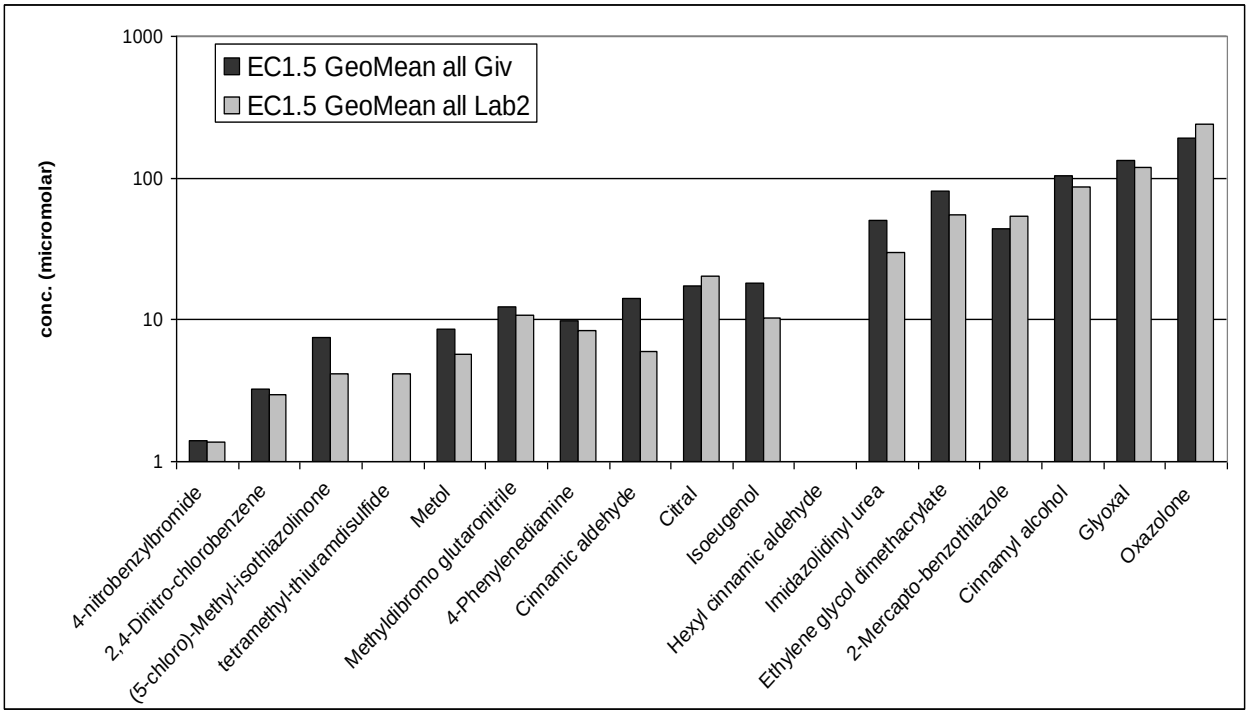


Figure 3b. Geometric Mean of the EC1.5 values from of the lead lab with all repetition within the quantitative range for cinnamic aldehyde and of lab 2 (all results, the majority being outside of the quantitative range). Note: tetramethylthiuramdisulfide in the lead lab is at <0.98 µM and could not be plotted.



Based on this specific analysis we would conclude that the assay and the quantitative data were transferable and reproducible even if the targeted quantitative range of the positive control was not met (as is especially the case in Lab2), and the assay thus is sufficiently robust to deliver reliable data, even if the quantitative performance of the positive control was not in the target range.

Overall conclusion to point 2) – 4)

We fully agree that our approach of setting a quantitative target, but then accepting the consecutive runs if the quantitative criteria were not met are unconventional, to say the least. Starting with the more conservative requirement of a statistically significant positive result for the positive control as the minimal target would have been better, and then only adding more stringent data requirements if needed based on experience at a later stage would be a more transparent and more logical approach.

Nevertheless, this detailed retrospective analysis looking at runs fulfilling the quantitative target and those not fulfilling the quantitative target indicates that the study conclusions appear not to be affected by our unconventional approach. This approach did avoid iterative testing and shows the level of reproducibility which can be achieved even without such iterative testing. It thereby gives an indication not only of transferability and reproducibility, but also certain indications of the robustness of the assay.

However, it is clear that going forward this should be handled more strictly. Based on all this experience we had recommended in the SOP that one of the two criteria must be fulfilled, and otherwise runs are discarded. See Page 17 of the submitted SOP where we indicated :

ACCEPTANCE CRITERIA

- Cinnamic aldehyde as positive control must be positive, thus the gene induction by this control must be statistically significant above the threshold of 1.5 in at least one dose.
- The Imax and the EC 1.5 for cinnamic aldehyde is calculated. The targets are: (i) Average induction in the three replicates for cinnamic aldehyde at 64 µM should be between 2 and 8, and (ii) the EC 1.5 value should be between 7.5 µM and 30 µM. At least one of these criteria must be met, otherwise the run is discarded. If only one criteria is fulfilled, it is recommended to carefully check the dose-response of cinnamic aldehyde in order to decide on acceptability
- For acceptance of the test for a given master plate in a given repetition, the average variability in the 3 × 6 solvent control wells for each master plate/repetition should be below 20%. If the variability is higher results are discarded.

References:

Aleksic, M., Thain, E., Roger, D., Saib, O., Davies, M., Li, J., Aptula, A.Zazzeroni, R., 2009. Reactivity profiling: Covalent modification of single nucleophile peptides for skin sensitization risk assessment. Toxicological Sciences 108, 401-411.

Gerberick, G. F., Vassallo, J. D., Foertsch, L. M., Price, B. B., Chaney, J. G.Lepoittevin, J. P., 2007. Quantification of chemical peptide reactivity for screening contact allergens: A classification tree model approach. Toxicological Sciences 97, 417-427.

Kanerva, L., Jolanki, R.Estlander, T., 1997. 10 years of patch testing with the (meth)acrylate series. Contact Dermatitis 37, 255-8.

Lazarov, A., 2007. Sensitization to acrylates is a common adverse reaction to artificial fingernails. J Eur Acad Dermatol Venereol 21, 169-74.

Annex 2

Meeting of the ESAC working group Sensitisation
1-3 February 2012
EC JRC, Ispra, Italy

Participating:

ESAC members: Erwin Roggen (Chair ESAC WG), Wally Hayes, Walter Pfaller

Experts: Maja Aleksic, Emanuela Corsini, Yong Heo, David Lovell, Michael Woolhiser

ECVAM: Claudius Griesinger (ESAC Coordination/Secretariat), Alexandre Angers (Scientific Support)

1) First discussions concerning the DPRA report.

Not reproduced in the present context (Keratinosens WG report).

2) Further discussions on the Keratinosens report

The WG did not consider that the additional information sent by Givaudan regarding the statistical analyses provided the clarification they were requesting. In any case, the WG decided to focus the analysis of the within laboratory reproducibility (WLR) in terms of concordance of the predictions obtained within the same laboratory. This means that the only study that can provide this information is in attachment 4c, the evaluation of 14 chemicals at Givaudan, performed in three complete experiments.

Concerning the occurrence and impact of the invalid runs, the WG will request from Givaudan a final analyses of the results in the submission, taking into account three different scenarios:

- The final results (in terms of WLR, BLR and Predictive Capacity), accepting all the experiments generated for the submission
- The final results, rejecting all the runs/experiments that did not meet the acceptance criteria as they were defined prior to the start of the ring trial
- The final results, rejecting all the runs/experiments that did not meet the acceptance criteria the test submitter would like to propose for future use based on the experience gained in the study.

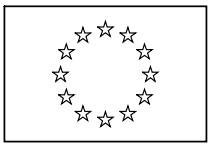
The WG would also ask Givaudan to clarify the exact details of these acceptance criteria, as they vary between the different sections of the document.

Documents will be prepared and sent to Andreas Natsch (AN) in order to make clear what the requests of the WG are. In addition, a teleconference will be set on the 8th of February (14.00 CET) with AN to explain these requests and the documents.

3) Dates for future meetings

Next meeting for DPRA review: 10/11 May

Annex 3



EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection
European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM)

ESAC Working Group Skin Sensitisation

Draft Minutes Teleconference
April 24th 2012
14:00 – 16:00 CET

Minutes: Claudius Griesinger

1. Participation

Participating:
ESAC members: Erwin Roggen (Chair ESAC WG), Walter Pfaller
Experts: Maja Aleksic, Emanuela Corsini, Michael Woolhiser
ECVAM: Claudius Griesinger (ESAC Coordination/Secretariat)
Excused:
David Lovell, Wally Hayes, Yong Heo

2. Agenda / purpose of TC

The Secretariat briefly revisited the agenda and purpose of the teleconference:

Agenda:

1. Discussion of the material forwarded by Givaudan on 14/3/2012 following request from ESAC WG forwarded on 8/2/2012.
   1.1 To which extent does this information (reproducibility expressed as concordance of predictions and taking different cases regarding fulfilment of test acceptance criteria into account) address the open questions of the WG and is it sufficient for addressing the main question of the review (reliability of the assay)?
   1.2 To which extent does the analysis forwarded allow addressing (preliminary) predictive capacity and inform on possible limitations (including the biological relevance: pathway) of the test method in view of defining the necessary follow-up work (gaps) required to sufficiently characterise the test method for possible use within a testing strategy?
2. Organisation/distribution of work - draft ESAC WG report and draft ESAC opinion.

Taking the purpose of the TC into account, it was agreed to discuss the submitted Word Document ("Final clarifications…") as well as the more detailed excel spreadsheets providing summary data of reliability and preliminary predictive capacity as requested by the ESAC WG.

3. Discussion of material resubmitted by Givaudan

3.1 Section A of the clarification document: Test design

It was agreed that the schematic outline of the test design helped in understanding how a Keratinosens test is carried out to arrive at a final prediction (Sensitizer / Non-Sensitizer). The

schematic had been initially provided by the ESAC WG/ECVAM and was now updated by Givaudan for the purpose of this resubmission. The update concerned the addition of the MTT parallel plate to assess cytotoxicity (condition 4 of the Test Acceptance Criteria). It was agreed to include this schematic figure in the ESAC WG report.

3.2 Section B: Test Acceptance Criteria

The WG established that the Criteria Sets found and described by the ESAC WG corresponded to the Criteria Sets communicated by Givaudan in the resubmission ("clarification") as shown in table 1. The ESAC WG agreed that the question of which criteria had been used to analyse in particular the ring trial data was now sufficiently clear. Moreover, it was transparent which criteria Givaudan recommends for future use of the assay.

Criteria Set found by WG and communicated in the document "Final Clarifications…" to Givaudan by the ESAC Sec. (sent to Givaudan 8.2.2012)	**Criteria Set as described in the resubmission from Givaudan (received by ECVAM 14.3.2012)**
Criteria Set 1 (page 7 of submission).	***Criteria Set 1*** (page 7 of submission) Additional clarification (in blue) reg. condition 1: CA positive in the range 4-64uM. *NOTE ESAC SEC: These criteria (and not Criteria 2 communicated in SOP of ring trial) were used to analyse the ring trial data. However, some ring trial data were invalid when applying these criteria. As no rules for retesting in case of unfulfilled TACs had been stipulated beforehand, this led to some non-qualified test results being included in the analysis contained in the full submission to ECVAM. Having realised that non-qualified test results had been included in the analysis on reproducibility and predictive capacity, the ESAC WG requested (on 8/2/2012) a reanalysis of the data using (a) concordance of predictions as a measure for reliability and (b) performing this analysis applying or not-applying the specified test acceptance criteria.*
Criteria Set 2 (Att. 7a SOP of ring trial): contained in the ring trial SOP but neither used by participating labs nor by lead lab for final analysis. This set should be ignored for the purposed of the review.	
Criteria Set 3 (page 9 of submission)	***Criteria Set 2*** Additional clarification (in blue) reg. condition 1: CA positive in the range 4-64uM *NOTE ESAC SEC: These are the test acceptance criteria as currently used by Givaudan and as recommended for future use.*
Criteria Set 4 (page 17 of Invittox protocol): This set is identical to set 3, apart from a minor typo (7.5 instead of 7.0 uM in condition 2). This set should be ignored for the purposes of the review.	

3.3 Section C: 1) Within Laboratory Reproducibility

The ESAC WG agreed that the re-analysis resubmitted was satisfying with regard to answering the question to which extent non-qualified test results might have influenced the WLR analysis. The impact was felt to be negligible as in case 2 (most stringent criteria) only 3 individual laboratory predictions had not qualified. The data were found sufficient to judge WLR.

3.4 Section C: 2) Between Laboratory Reproducibility

Agreement was reached regarding the deletion of the fourth column in the word document as well as the corresponding columns in the excel files. These columns were intended to allow analysis of

concordance of predictions on the basis of four laboratories only instead of the five laboratories that had participated in the ring trial. The rationale for this column was that with an increasing number of labs it may be more likely to get non-concordant results. Givaudan, when planning the ring trial had been unaware that it is common practice in the context of validation to use three laboratories within a ring trial for assessing transferability and between-lab reproducibility (BLR). However, as this analysis had not been properly conducted in the resubmitted data package, the ESAC WG felt that this approach should not be followed-up when finalising the ESAC WG / ESAC review.
No agreement was reached with respect to the question whether or not the data on BLR (when taking non-qualified laboratory predictions into account: case 2 and 3) were sufficient to judge reproducibility between laboratories. The reason for this uncertainty lies with one of the principal flaws of the planning and execution of the study, i.e. that the Test Acceptance Criteria (TACs) provided to the participating laboratories during the ring trial were (a) had not applied when analysing the data (a variation of the initial TACs had been employed) as they were found too stringent, (b) that no provisions for re-testing had been made in case test results would not meet the TACs and (c) that, as a consequence, the final data matrix contained an appreciable number of non-qualified test results that were included in the analysis.

- With regard to the analysis requested in the ESAC WG's letter to Givaudan requesting clarification/re-analysis of the data, the Chair (E. Roggen) noted that while the non-qualified test results had been deleted from the data matrix (case 2 and case 3), the concordance had nevertheless been calculated on the basis of the *remaining* laboratory predictions (which were in all cases at least 3 lab predictions per chemical). This issue was contentiously discussed in the group. Some members felt that chemicals for which some individual laboratory predictions had to be deleted should not be included in the analysis of concordance/reliability.
- M. Woolhiser remarked that the approach chosen by Givaudan seemed nevertheless reasonable as the common standard were 3 laboratory predictions per chemical in a validation ring trial and since for all chemicals at least 3 laboratory predictions were available after deletion of non-qualified results.
- The ESAC Secretariat remarked that out of 105 individual lab predictions (21 chemicals x 5 labs) only 14 were not qualified and had been deleted. Hence there seemed to be sufficient data for judging whether the test was reliable between laboratories especially when considering, as pointed out by M. Woolhiser, that in all cases a minimum of three lab predictions was available. The Secretariat also suggested to the WG to consider separating the issues of flawed planning/conduct (e.g. no provisions for retesting, criteria appropriately fixed before ring trial) from the issue of whether or not there were sufficient data for *judging* reliability.

It was finally agreed to postpone conclusion of this issue and allow some time to reflect on it. Members should forward to the Chair and Secretariat a short summary statement describing their view concerning BLR analysis and in particular whether there are sufficient data to conclude on BLR.

Due to the advanced time, the TC was closed at this point, although transferability as well as preliminary predictive capacity (including the extended non-sensitizer list: LLNA negatives) had not yet been discussed.

Finally, the ESAC Secretariat kindly requested that discussions at the next ESAC WG meeting (10/11 May) be limited to the DPRA. Moreover, the Secretariat kindly requested that the Keratinosens review be finalised as soon as possible, especially in the light of the fact that all critical clarifications had now been made available and the ESAC WG report was in an advanced state.

4. Actions

Item Nr.	Action	Actor	Timeline
1	Forward an **appraisal of the data on BLR** as contained in the resubmission to Chair and Secretariat	All ESAC WG members	asap, deadline is May 4 to 7
2	**Block date and time** for the next TC on Keratinosens: 22 MAY from 14:00 to 16:30. With this TC the ESAC WG should come to a conclusion on the main findings of the Keratinosens review. These should provide the essential points for the draft ESAC opinion.	All ESAC WG members	
3	Reformat the submitted spreadsheets (deleting 1 column referring to 4 out of 5 lab predictions).	ESAC Secretariat	asap

Annex 4

JOINT RESEARCH CENTRE
Institute for Health and Consumer Protection
EURL-ECVAM
The European Union Reference Laboratory for Alternative Methods to Animal Testing

ECVAM
SCIENTIFIC
ADVISORY
COMMITTEE
(**ESAC**)

ESAC Working Group Sensitisation:

Final clarifications regarding the Givaudan/Keratinosens submission

(ER 2011-04)

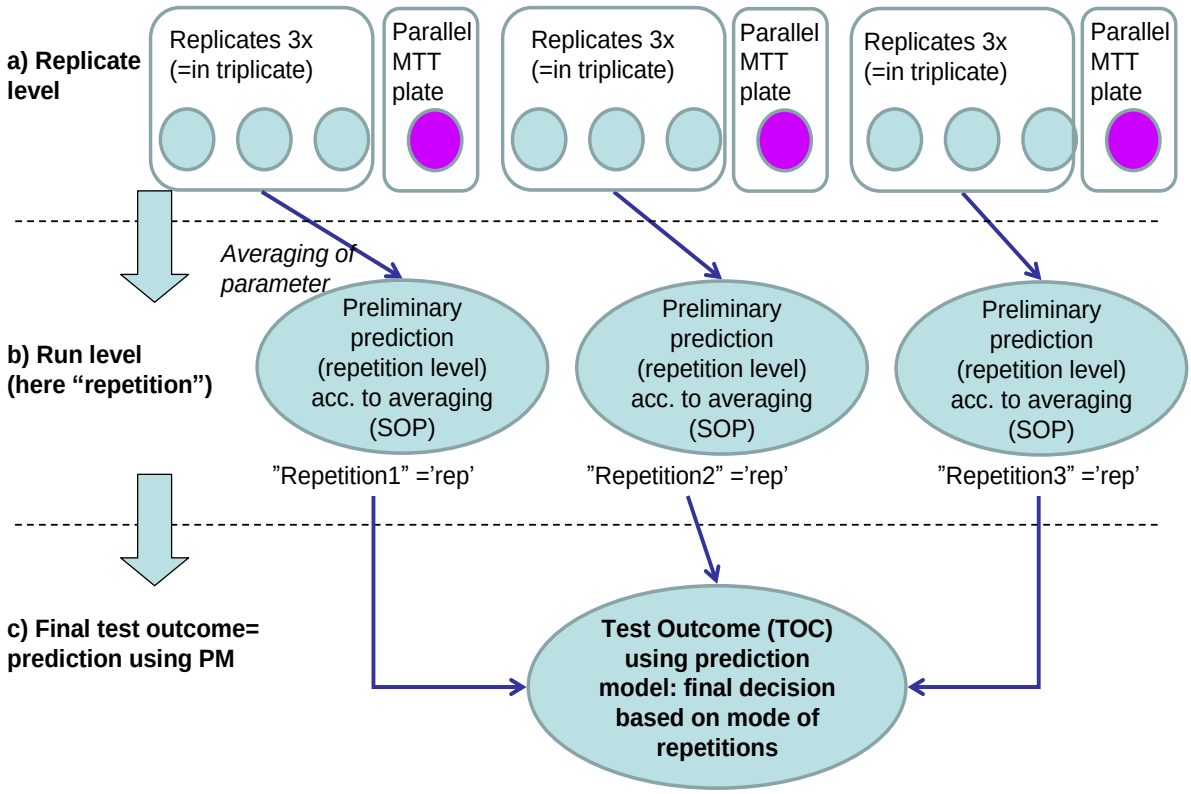
A. Test design

A complete test, i.e. experiment yielding the intended TEST OUTCOME of the KeratinoSens test method (i.e. a binary prediction: S=1 or NS=0) is based on the following design:

- ***a) Replicate level:*** Chemicals are tested in 3 replicates (in triplicate). The parameters measured of the individual replicates are averaged. One replicate parallel plate is run for the cytotoxicity determination.
- ***b) Repetition level:*** Using this average value, a preliminary prediction is derived by applying the prediction model (PM). This result constitutes a so-called "repetition" (consisting of three replicates for luciferase and one replicate for MTT).
- ***c) Test outcome level:*** According to the provisions of the PM, three repetitions are required to arrive at a final decision/"final call" (=the TEST OUTCOME as foreseen when applying the test). The final decision is made by using the mode of the predictions of the 3 repetitions allowing a unequivocal predictions to be made.

N.B: while the algorithm of the PM is applied to average values at run/repetition level to reach a prediction, this result is not equivalent to a final prediction (although being made on the basis of the PM), since the PM requires the two concordant results of three runs/repetitions to arrive at a final prediction (TEST OUTCOME, TOC). Three repetitions were always made for all the intra- and interlaboratory studies. As we need two concordant predictions in the PM, two unequivocal concordant predictions in the first two repetitions were considered sufficient for a TOC in the additional screenings. In cases of borderline results or non-concordance of the first two repetitions then 2 – 4 additional repetitions were run for a better assurance of the TOC.

Figure 1: *Schematic depiction of the design of a full TEST, leading to the predefined TEST OUTCOME = a prediction as to whether a chemical is a sensitizer or not.*



B. Clarification on the Test Acceptance Criteria used/suggested

CRITERIA SET 1			
Condition:	**QUALIFIED IF 1 AND 2 AND 3 AND 4 FULFILLED**		
	Criteria (page 7 of submission)	logical operator	
1	Cinnamic aldehyde positive in the range 4 - 64uM	**AND**	
2	EC1.5 CA 7-30uM	**AND**	
3	Induction at 64uM CA 2-8x	**AND**	
4	Variability in solvent wells 20%		
CRITERIA SET 2			
Condition:	**QUALIFIED IF 1 AND (2 AND/OR 3) AND 4 FULFILLED**		
	Criteria (page 9 of submission)	logical operator	
1	Cinnamic aldehyde positive in the range 4 - 64uM		**AND**
2	EC1.5 CA 7-30uM	**OR**	
3	Induction at 64uM CA 2-8x		**AND**
4	Variability in solvent wells <20%		

Criteria set 1 was defined as intralaboratory data on the positive control accumulated and it was used to evaluate the ring-study data against. However, these criteria were based on the responses observed experimentally from the positive control in the lead lab, and there was no indication concerning the importance of these ranges of responses on the validity of the runs, so it was decided to include all the runs in the ring study irrespective of whether or not they met these criteria.
Based on the experience of the ring study it was proposed in the original test submission to ECVAM to abandon condition 2 and 3 and only continue with condition 1 and 4. However, during internal ECVAM review, the ECVAM team recommended to keep minimal quantitative criteria and eventually modify conditions 2/3. Based on this recommendation and the overall experience, criteria set 2 was defined with the "OR" operator for condition 2 and 3 in the revised test submission which was then accepted for the peer review.
This criteria are set based on the accumulated experience, namely that this is a practical achievable criterium – and we thus decided to use it in our Lab on routine use and in the SOP for future use.

C. Performance values (reproducibility, predictive capacity)

1) WITHIN LABORATORY REPRODUCIBILITY SET (N=14)

CASE	EXPLANATION	Concordance of Predictions within laboratory	Sensitivity	Specificity	Accuracy
1	**ALL PREDICTIONS INCLUDED**	85.7	70.0	100.0	78.6
2	**ONLY QUALIFIED PREDICTIONS (a)**	85.7	70.0	100.0	78.6
3	**ONLY QUALIFIED PREDICTIONS (b)**	85.7	70.0	100.0	78.6

a. On the basis of predictions that fulfilled the Test Acceptance Criteria (TAC) as used for the ring trial (Criteria Set 1).
b. On the basis of predictions that fulfilled the Test Acceptance Criteria (TAC) as specified for future use of the test method (Criteria Set 2).

2) BETWEEN LABORATORY REPRODUCIBILITY SET (N=21)

CASE	EXPLANATION	Concordance of predictions between laboratories: 5/5 = concordant	Concordance of predictions between laboratories: 4/5 = concordant	Sensitivity	Specificity	Accuracy
1	**ALL PREDICTIONS INCLUDED**	85.7 (n=21)	95.2 (n=21)	92.9	100.0	95.0 (n=20) [1]
2	**ONLY QUALIFIED PREDICTIONS (a)**	90.5 [2)] (n=21)	Not applicable	86.7	100.0	90.5 (n=21)
3	**ONLY QUALIFIED PREDICTIONS (b)**	85.7 [2)] (n=21)	100.0 (n=11)	86.7	100.0	90.5 (n=21)

[1)] For one chemical (Eugenol) the status is 2 labs positive, 3 labs negative, thus final call cannot be decide and hence n=20

a. On the basis of predictions that fulfilled the Test Acceptance Criteria (TAC) as used for the ring trial (Criteria Set 1), [2)] concordant for all the labs for which a call could be made, in this case not always 5 labs, see excel file.
b. On the basis of predictions that fulfilled the Test Acceptance Criteria (TAC) as specified for future use of the test method (Criteria Set 2), [2)] concordant for all the labs for which a call could be made, in this case not always 5 labs, see excel file

3) TRANSFERABILITY SET (N=7)

CASE	EXPLANATION	Concordance of predictions between laboratories: 5/5 = concordant	Concordance of predictions between laboratories: 4/5 = concordant	Sensitivity	Specificity	Accuracy
1	**ALL PREDICTIONS INCLUDED**	85.7 (n=7)	85.7 (n=7)	100	100	100 (n = 6) [2)]
2	**ONLY QUALIFIED PREDICTIONS (a)**	85.7 [1)] (n=7)	Not applicable	100	100	100 (n = 6)
3	**ONLY QUALIFIED PREDICTIONS (b)**	85.7 [1)] (n=7)	Not applicable	100	100	100 (n = 6)

[1)] One Lab is excluded due to two repetitions outside of the Criteria Set 1, concordance for 4/4

[2)] For one chemical (Hexyl-cinnamic aldehyde) the status is 2 labs positive, 3 labs negative, thus final call cannot be decide and hence n=6, see excel file

a. On the basis of predictions that fulfilled the Test Acceptance Criteria (TAC) as used for the ring trial (Criteria Set 1).
b. On the basis of predictions that fulfilled the Test Acceptance Criteria (TAC) as specified for future use of the test method (Criteria Set 2).

4) PREDICTIVE CAPACITY SET BASED ON EXISTING DATA = SILVER LIST (N=67) [USING HISTORICAL DATA]

CASE	EXPLANATION	Sensitivity	Specificity	Accuracy
1	**ALL PREDICTIONS INCLUDED**	88.4	79.2	85.1

5) PREDICTIVE CAPACITY SET BASED ON NEWLY GENERATED DATA (N=47)

CASE	EXPLANATION	Sensitivity [1)]	Specificity	Accuracy
1	**ALL PREDICTIONS INCLUDED**	67.0	100.0	70.2 (n= 47)
2	**ONLY QUALIFIED PREDICTIONS (a)**	63.9	100.0	65.8 (n= 38)
3	**ONLY QUALIFIED PREDICTIONS (b)** [2)]	67.0	100.0	70.2 (n= 47)

[1)] We suggest that this Table should always be viewed along with the discussion and definition of the applicability domain discussed in detail on page 9-11 of Attachment 12c. This list contains a significant number of specifically amine-reactive chemicals and phenolic prohaptens, which were found to be outside of the applicability domain of the assay and which need to be detected by additional peptide reactivity assays and metabolic activation assays.

[2)] Note: The studies in section 5, 6 and 7 only contain data which fulfil the criteria set 2 (i.e. the acceptance criteria suggested for further use of the final SOP as used now in our laboratory, with one of the two quantitative criteria required positive). Hence in this and the following three Tables the numerical values for case 1 and case 3 are always identical. In the case 2, data were omitted for which one of the two quantitative criteria were not met).

6) PREDICTIVE CAPACITY FOR CHEMICALS WITH NEG. LLNA REFERENCE DATA (SUBMITTED IN JANUARY 2012)

CASE	EXPLANATION	Specificity
1	**ALL PREDICTIONS INCLUDED**	72.5 (n=80)
2	**ONLY QUALIFIED PREDICTIONS (a)**	74.0 (n=73)
3	**ONLY QUALIFIED PREDICTIONS (b)**	72.5 (n=80)

7) PREDICTIVE CAPACITY FOR CHEMICALS WITH NEG. LLNA REFERENCE DATA (SUBMITTED IN JANUARY 2012) AND WHEN CONSIDERING AVAILABLE HUMAN REFERENCE DATA

CASE	EXPLANATION	Specificity
1	**ALL PREDICTIONS INCLUDED**	76.4 (n=80)
2	**ONLY QUALIFIED PREDICTIONS (a)**	77.3 (n=73)
3	**ONLY QUALIFIED PREDICTIONS (b)**	76.4 (n=80)

Omitting the reactive, peptide alkylating chemicals, for which we do not trust the LLNA despite absence of human data:

CASE	EXPLANATION	Specificity
1	**ALL PREDICTIONS INCLUDED**	83.1 (n=73)
2	**ONLY QUALIFIED PREDICTIONS (a)**	83.3 (n=67)
3	**ONLY QUALIFIED PREDICTIONS (b)**	83.1 (n=73)

8) PREDICTIVE CAPACITY FOR ALL SUBMITTED DATA

Compiles all the data of all the submitted attachments (compiled in attached Excel file ”Keratinosens_SYN_MOD5_all.xls”). Human evidence was taken into account as in section 7 above. Only Case 1 applies, as it includes the historical silver list data.

CASE	EXPLANATION	Sensitivity	Specificity	Accuracy
1	**ALL PREDICTIONS INCLUDED**	79.3	79.8	79.5 (n=220)

Omitting the reactive, peptide alkylating chemicals, for which we do not trust the LLNA despite absence of human data

CASE	EXPLANATION	Sensitivity	Specificity	Accuracy
1	**ALL PREDICTIONS INCLUDED**	79.3	84.5	81.7 (n=213)

End of document