EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

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

ECVAM
SCIENTIFIC
ADVISORY
COMMITTEE
(ESAC)

ESAC Working Group Peer Review Consensus Report

on

an ECVAM-coordinated study concerning the transferability and reliability of the human Cell Line Activation Test (h-CLAT) for skin sensitisation testing

TABLE OF CONTENTS

ESAC Working Group

This report was prepared by the "ESAC Working Group (ESAC WG), charged with conducting a detailed scientific peer review of two ECVAM-coordinated studies on *in vitro* test methods for skin sensitisation testing: (1) the Direct Peptide Reactivity Assay, DPRA; (2) the human Cell Line Activation assay, h-CLAT as well as an external validation study on the KeratinSens assay for skin sensitisation, conducted by Givaudan and submitted to EURL ECVAM for evaluation and ESAC review. The ECVAM coordinated validation study on the MYELOID U937 Skin Sensitisation test method (MUSST) was also foreseen to be reviewed by the ESAC / ESAC WG. However, during validation it became apparent that the test method required further optimisation.

The ESAC WG had been set up by the ESAC during its meeting on March 2011 (ESAC 34). The primary goal of this WG was to evaluate the transferability and reliability of these assays in view of assessing their potential usability within a testing strategy for skin sensitisation. This WG report focuses on the part of the ECVAM study addressing the **human Cell Line Activation Test (h-CLAT)** for skin sensitisation testing. Basis for the scientific review was the ECVAM request to ESAC concerning the scientific peer review of the h-CLAT and the formulation of scientific advice based on three principal questions outlined by ECVAM (ESAC request 2013-023, see Annex 1):

1. Was the study conducted appropriately in view of the objective of the study?
2. Are the conclusions, as presented in the Validation Study Report, substantiated by the information generated in the study and are plausible with respect to existing information and current views?
3. What could be the suggested use of the test method?

The ESAC WG conducted the peer review from October 2013 to February 2014. The ESAC WG met once in person at ECVAM (01-02.10.2013) to discuss initial findings, and resolve contentious points. The first version of this ESAC WG report (06.11.2013) was circulated amongst the WG which was amended via written procedure following input from the ESAC WG, the ESAC Coordinator. Following two teleconferences on 5 and 6 December, a consolidated version was circulated on 7 February 2014. The final report was adopted by the ESAC WG on 17 February 2014.

This report was endorsed by the ESAC WG on 21 February 2014 and represents the consensus view of the ESAC WG.

This ESAC WG peer review consensus report was endorsed by the ESAC on 11. March 2014.

<u>The ESAC WG had the following members:</u>

- Dr. Erwin ROGGEN (ESAC member, Chair of ESAC WG and rapporteur)
- Dr. Ed CARNEY (ESAC member)
- Prof. A. Wallace HAYES (external expert)
- Dr. Maja ALEKSIC (external expert)
- Prof. Emanuela CORSINI (external expert)
- Dr. David LOVELL (external expert)
- Dr. Michael WOOLHISER (external expert)
- Prof. Yong HEO (external expert, ICATM nomination*)*

<u>ESAC Coordinator:</u>

- Dr. Claudius GRIESINGER *(EURL ECVAM Coordinator for ESAC peer reviews and EURL ECVAM recommendations)*

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
- **GPMT** Guinea Pig Maximization Test
- **LLNA** Local Lymph Node Assay
- **NS** Non-sensitizer
- **PC** Positive Control
- **S** Sensitizer
- **SOP** Standard Operating Procedure (used here as equivalent to 'protocol')
- **VC** Vehicle Control
- **VMT** Validation Management Team
- ***VSR*** Validation Study Report
- **WLR** Within-laboratory reproducibility

Executive summary

Following a request from ECVAM to ESAC for peer review of and scientific advice on an ECVAM-coordinated study focusing on the assessment of the transferability and reproducibility and reliability of the human Cell Line Activation Test (h-CLAT), ESAC, supported by the ESAC Coordinator, organized a meeting of the ESAC Working Group (ESAC WG) on Sensitization (established in March 2011) to prepare a detailed scientific peer review report and to draft an opinion for consideration by the ESAC.

The WG was specifically charged to assess the transferability and reproducibility (within- and between- laboratories) of the h-CLAT (primary objectives of the study) in view of its possible future use as part of a non-animal testing strategy for skin sensitization hazard assessment. As the study was also designed to provide *preliminary* information on a) the predictive capacity of the test method and b) its potential use for assessing the sensitization potency of substances (subcategorisation), the ESAC WG was requested to review the information on these two aspects. Finally, in view of the potential use of the h-CLAT within an Integrated Testing Strategy (ITS), the ESAC WG was requested to comment, if possible, on the particular strengths and weaknesses of the assay and its potential placing within such an ITS.

The key findings of the ESAC WG are summarized below:

A) Strengths of the study:

- The h-CLAT is a mechanistically based *in vitro* test method that addresses one of the key events of the induction of skin sensitisation: dendritic cell (DC) activation.
- The primary criterion for chemical selection was the availability of robust *in vivo* data (LLNA and GPMT primarily, but also human data allowing for adequate comparisons). The chemical selection included substances previously tested in the h-CLAT method and substances not tested in the h-CLAT assay.
- The study design was considered appropriate for the purpose of addressing the following objectives: Assessing transferability, WLR and BLR of the h-CLAT.
- The statistical approach chosen to analyse the data was considered appropriate.

B) Weaknesses of the study:

- The Test Definition of the h-CLAT assay (p. 25) would benefit from a more detailed rationale behind the selection of the THP-1 cell line, and CD86 and CD54 membrane markers; in particular as to why both of the markers are required. The biological and mechanistic relevance of the test is not sufficiently explained. There is ample evidence showing that CD86 and CD54 are generally up-regulated in response to challenges that cause cell damage, inflammation and cytotoxicity. There is a need to explain what special features of the test or the prediction model are making the test specific for sensitization.
- The WG had some concerns about the statistical calculations underlying the determination of an adequate sample size to analyse reproducibility as a primary study goal of the ECVAM validation study on the DPRA, the MUSST and h-CLAT. There is concern about the inadequate sample size used to analyse one of the primary outcomes, reproducibility, in the various statistical calculations.
- An average WLR of 80% (KAO (86.7%), Shiseido (80.0%), EURL-ECVAM (80.0%) and Bioassay (73.3%)) did not meet the 85% WLR target set by the VMG. It was observed that a subset of chemicals consistently drove discrepancies in reproducibility. To what extent these

discrepancies were related to issues of applicability domain or inherent sources of variability of the h-CLAT assay (e.g. the time course for expression of the cell surface markers; the state of cell differentiation) was not addressed in the document reviewed by the WG.

- The preliminary results on predictive capacity (secondary goal of the study) were considered promising by the VMT as far as they concerned the capacity of the assay to distinguish sensitisers from non-sensitisers. However, with regard to the capacity of the h-CLAT to provide reliable and relevant information on subcategorisation, the WG noted an appreciable difference between the results for subcategorization obtained for 100 historical substances (ca 72% accuracy) and the results obtained with the substances selected for the validation study (ca 57% accuracy). The WG further noted that these differences were not addressed or at least discussed by the VMT.

C) Conclusions of the study:

Overall, the conclusions made by the WG correspond with the conclusions formulated in the report by the VMG. The WG disagrees, however, with the VMG conclusion concerning the WLR.

- The targeted 85% WLR (defined by the VMG) of the test method with respect to concordance of classification (S/NS) was met only by one laboratory.
- The data were considered strong enough to support transferability of the test to properly equipped, trained and staffed laboratories with the appropriate cell culture and flow cytometry capabilities.
- The BLR (lowest: 79.2%) was considered to meet the target of 80% (as defined by the VMG) .
- In agreement with the VMG the number of chemicals (N=24) was considered insufficient to draw firm conclusions on the predictive capacity of the test method. It was observed that the accuracy (76%), sensitivity (81.3%) and specificity (65.6%) values of this study were lower than the historical values (84%, 87%, 75%, respectively) obtained with 100 chemicals. Due to this discrepancy the ESAC WG concluded that the number of substances and the information available for these substances was insufficient for allowing more than a purely preliminary indication on the predictive capacity in terms of S/NS.
- Subcategorization based upon the data from the 24 chemicals was 57% accurate, while the historical data (N=100) determined an accuracy of 72% (Ashikaga et al. 2010. ATLA 38; 275-284). As for the predictive capacity of the h-CLAT, this discrepancy made the ESAC WG conclude that the number of substances and the information available for these substances was insufficient for allowing more than a purely preliminary indication on the predictive capacity in terms of potency classification.
- Assessment of the applicability domain was not a primary objective of this study. Consequently, the number of chemicals tested was insufficient to draw a conclusion about the applicability domain of the test. The existing information tested in-house and submitted to EURL ECVAM had not been forwarded for ESAC review, although some information on historical chemicals was contained in Appendix 13, however without identifying the chemicals tested and thus not allowing a detailed review of the predictive capacity, in particular in relation to the applicability domain. The ESAC WG requested further information on these chemicasl (Appendix 13 of the study report) during the review.

The chemicals that drove discrepancies in the BLR study were the same as those driving discrepancies in the WLR study. The VMG concluded that these problem chemicals fall outside the applicability domain. The Working Group suspects that there may be other explanations for the discordant results that are due to inherent variability in the assay (e.g. the time course for expression of the cell surface markers and the state of cell differentiation could be sources of variability).

Empirically the applicability domain seems to exclude pre-/pro-haptens, autofluorescent compounds, and chemicals with limited solubility or stability in water, metal salts and volatile compounds. The WG observed that the pre- and pro-haptens included in the chemical list were consistently scored correctly in spite of THP-1 cells lacking detectable metabolic activity. The tested metal salts gave an inconsistent response. Both observations would warrant further investigation to better understand the mechanisms triggered by these compounds in THP-1 cells.

D) Recommendations:

The h-CLAT addresses a key mechanism (DC activation) in the development of sensitization. Overall the provided data support BLR and transferability of the test to qualified laboratories.

The WLR was unsatisfactory as three of the four laboratories did not meet the 85% target. The ESAC Working Group is concerned that there may be other inherent characteristics of the h-CLAT which could lead to appreciable variability (e.g. the time course for expression of the cell surface markers; the state of cell differentiation or simply the level of training required for this test method). The ESAC WG recommends that possible sources of variability be identified, and solutions should be provided because in general poor reproducibility of test methods may create more difficulties in interpreting data as part of an ITS due to conflicting results.

The WG recommends to better define (1) the predictive capacity and (2) the applicability domain of the h-CLAT (to eliminate the uncertainty currently associated with a negative result) either through further testing (i.e. prospective validation) or through restrospective analysis of existing information (retrospective validation: data grouping avoiding possible sources of bias / meta-analysis of the grouped data).

For hazard classification purposes, the SOP could be adapted to reduce resource costs by eliminating the need for a third evaluation run in case the first two runs are consistent.

Globally harmonized system (GHS) sub-categorisation of sensitisers should form part of a wider assessment, as on the basis of the submitted results it is envisaged that h-CLAT EC150 and EC200 values might provide useful information contributing to this purpose. In two recent papers, Nukada et al. (2012; 2013) addressed the possibility to use hCLAT data for allergic potency classification. Results hold promise, but further studies are needed to identify the best integrated testing strategy or strategies necessary to cover the different regulatory goals and risk assessments.

1. Study objective and design

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

NOTE: (a) please summarise briefly in your own words the study objective as outlined in the VSR and (b) provide an appraisal as to whether the study objective is clearly and comprehensibly defined in the VSR.

(a) ESAC WG summary of the study objective as outlined in the VSR

Primary objective:
Evaluating transferability and reliability (reproducibility within and between-laboratories) with a view to its future use in integrated non-animal approaches for replacing the currently used regulatory animal tests.
Secondary objectives:

a) Evaluation of the ability of the h-CLAT to reliably discriminate skin sensitising (S) from non-sensitising (NS) chemicals (Globally Harmonised System (GHS) of Classification and Labelling of Chemicals for Skin Sensitization (category 1; no category) and as implemented in the European Commission Regulation (EC) No 1272/2008 (EC, 2008) on classification, labelling and packaging (CLP) of substances and mixtures.
b) Assessment of the ability of the h-CLAT to contribute to sub-categorisation of skin sensitising chemicals, e.g. into Sub-category 1A and Sub-category 1B as adopted in the 3rd revised version of the GHS (UN, 2009).

(b) Appraisal of clarity of study objective as outlined in the VSR

Overall observations:

Study objectives are clearly formulated and comply with the ECVAM's modular approach to validation (Hartung et al., 2004) with focus on Modules 1 to 4.

Specific observations:

The concept of ‘’integrated non-animal approaches for replacing the currently used regulatory animal tests’’ is yet to be defined. It is therefore unclear how the applicants ‘view’ the use of the h-CLAT in this context. Two recently published papers showed two different possible uses of hCLAT in an integrated testing strategy (Bauch et al., 2012; Nukada et al., 2013), but further studies are necessary to identify the best integrated testing strategy or strategies necessary to cover the different regulatory goals and risk assessment needs. The recent EC-JRC proposal for an OECD Project on the development of Guidance Document on the Evaluation and Application of ITS for Skin Sensitisation will provide guidance on the evaluation and application of IST.

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

NOTE: What is, according to the VSR, the overall purpose of the test method? Examples are a) scientific use (e.g. basic/applied research, b) screening for product development c) regulatory testing etc.

(a) Analysis of the scientific rationale provided in the VSR

NOTE: Is the scientific rationale of the test method AND (consequently) for conducting the study clearly explained? Consider how the test method may contribute

(a) to the scientific understanding of the specified health/environmental effect or aspects of it?, i.e. does it provide relevant mechanistic information such as physiological pathways relevant for toxicity ("toxicity pathways") or other key physiological events leading to toxicity?

(b) to the prediction of the specified downstream health/environmental effect or aspects of it?

Moreover, does the VSR make sufficient reference to the relevant body of scientific literature?

Overall observations:

The description of the scientific rational for the test method and for conducting the study requires in depth knowledge with respect to context (Adverse Outcome Pathway (AOP)), mechanistic background and relevance of the parameters measured by the test. While it is beyond the scope of the report to provide a detailed description of the AOP, an outline of the key molecular events in the AOP and/or a figure depicting the AOP plus the position of the h-CLAT within the OECD AOP would provide useful context, especially for more general audiences.

Specific observations:

The background section contains the information useful for understanding the need to develop alternatives, however, it does not make an attempt to explain what particular part of the well described skin sensitisation AOP this method is trying to address. One minor point is that the AOP is described as a list of six steps. One could easily get the impression that all six steps are linear, with each step dependent on the previous one. Some additional clarity around the interdependence of the steps would be helpful.

There is a cryptic explanation about what the h-CLAT actually measures (p. 25), but how the measurements obtained help to distinguish between chemicals with or without sensitising potential is not explained. There are equally no references to the original published work on the h-CLAT method as such to help the reader rationalise the development of the assay and the need for method validation. The only reference included is the 'comprehensive review' by Adler et al 2010.

Important is the lack of explanation with respect to the specificity of maturation markers in general, and CD86 and CD54 in specific. There is ample evidence showing that these markers are generally up-regulated in response to challenges that cause inflammation and cytotoxicity. The test submitters could have elaborated more on the differences between cellular stress induced by allergens and by irritants, the window in which only sensitizers affect dendritic cells, and the history of the test assessed by this study.

(b) Analysis of the regulatory rationale provided in the VSR

NOTE: Is a regulatory rationale specified, i.e. a specific application of the test method for purposes of generating data with respect to regulatory requirements as specified in legislation or internationally agreed guidelines etc.? If so, how does the study and its objective and design relate to this regulatory rationale? Are the relevant regulatory documents appropriately referenced?

Overall observations:

The regulatory rationale is well covered in the Background Section. This section clearly explains the current tests (Table 1, p. 7) and the need to replace them with mechanistically based non-animal alternatives.

The report is clear that the h-CLAT is intended for regulatory use as part of an, as yet to be defined or emerging, integrated testing strategies.

1.3 Appraisal of the appropriateness of the study design

NOTE: Is the study design appropriate in view of the stated objective of the study? This includes an analysis of the number of laboratories involved in the study, the organisation of study management including chemical selection, quality check of data, and independence of statistical analysis, i.e. was the statistician independent from the test method submitter/developer and, depending on the study, from the VMG. More technical aspects can also be considered such as an appraisal of the nature and number of test items used (details however to be provided in section 6, test materials), retesting in case of unqualified tests, pre-defined test acceptance criteria etc.

Overall observations:

The study design was appropriate for the primary objective. The study was designed to generate data for modules 1-4 of the ECVAM modular approach. It is accepted that whilst the data generated can contribute to modules 5 and 6, the number of chemicals tested during the validation is not sufficient to address these modules fully.

Specific observations:

ICCVAM and JaCVAM participated in study design, chemical selection and SOP formulation before they were approved and issued by the VMT. The study set-up was adopted and implemented in the participating laboratories.

Statistical analysis related to the primary objective and the first secondary objective (S/NS) was performed by an external biostatistician, selected by an open call and paid by EuRL-ECVAM. Statistical analysis related to the second secondary objective (sub-categorization) was performed by two independent biostatisticians appointed by JaCVAM working on a voluntary basis.

The study was divided in two phases:
- Phase A: training and transfer

- Phase B: 24 coded chemicals were tested by each of the four laboratories to generate information on the BLR; a subset of 15 chemicals was tested two additional times in each laboratory for the evaluation of the WLR.

The study involved the two test developers (lead laboratories) and two naive laboratories: EURL-ECVAM and Bioassay. To assure reliability of the data from the other laboratories, the VMT predefined quality assurance requirements considered essential for the acceptance of information and data produced in the validation process (Section 3.1).

The chemical selection followed a predefined strategy as outlined in the document (Section 4, p. 14; App. 3 and 4). One third of the chemicals tested were non-sensitizers and two third were sensitizers. Approximately 50 % of the selected chemicals were liquids (44% of the S and 62% of the NS). A stratified random sampling of the 24 chemicals was applied to identify the 9 chemicals to be tested once and the 15 chemicals to be tested three times.

The study plan required that GLP-compliant laboratories conducted the pre-validation study in compliance with GLP Standards (OECD, 1999), while non-GLP compliant laboratories had to follow specific requirements believed essential for the mutual acceptance of information produced by the pre-validation process (See section 3).

1.4 Appropriateness of the statistical evaluation

NOTE: Consider whether the statistical approaches chosen are appropriate. This includes statistical calculations performed ex-ante such as sample size calculations as well as ex post statistical analysis of the data (e.g. for purposes of variability and predictive capacity). Is the choice of methods sufficiently justified?

Overall observations:

A detailed statistical analysis plan was produced and agreed to by the VMG before the start of the testing phase (App. 2). It was stipulated that only data from valid experiments would be considered for statistical analyses. Failed runs and/or experiments were reported in order to assess their occurrence and frequency.

Specific observations:

The biostatistician at ECVAM calculated the minimum number of test chemicals required for BLR (21) and WLR (13). As ECVAM was conducting validation on three *in vitro* assays concomitantly (DPRA, h-CLAT, MUSST), the VMG approved to slightly increase the number of test chemicals (24 for BLR and 15 for WLR) to allow for the possibility that some chemicals might fall out of the applicability domain of one or more of the *in vitro* assays.

Statistical analysis related to the primary objective and the first secondary objective (S/NS) was performed by an external biostatistician, selected by an open call and paid by EuRL-ECVAM. Statistical analysis related to the second secondary objective (sub-categorization) was performed by two independent biostatisticians appointed by JaCVAM and working on a voluntary basis on the basis of 100 historical (in-house) data as well as the 24 chemicals used during the study.

Data generated during the study were collected on reporting templates for cytotoxicity and CV75 values and evaluation runs and RFI values (App. 6). These templates were prepared by the lead laboratories.

The templates and formulae were tested and verified by the biostatistician at ECVAM before release to the participating laboratories. The data gathered by the test laboratories were subject to a quality check (App. 8) at ECVAM prior to analyses. The assessment criteria were defined in advance by the VMG and target performance set at 80% and 85% for BLR and WLR, respectively.

The statistical analysis on the test method's reproducibility focused on the concordance of classification, sensitizer (S) versus non-sensitizers (NS). Reproducibility was evaluated with respect to both WLR and BLR.

2. Collection of existing data

NOTE: Validation studies typically make use of existing data, e.g. either as reference data (prospective studies) OR as reference data and testing data (retrospective study). Moreover, validation studies may use other information such as data in the literature, data banks etc.

2.1 Existing data used as reference data

Which data sources were used for compiling the reference data associated with the test chemicals?

General observations:

Two recognised databases were used as a convenient source of authoritative peer-reviewed data for chemical selection: 1) the ICCVAM database containing information about 103 chemicals (NIH Publication number 11_7709 (http://iccvam.niehs.nih.gov/docs/immunotox_docs/LLNA-pot/TMER.pdf)), and 2) the LLNA database of 341 chemicals (Gerberick et al., 2005 and 2007, Kern et al, 2010). These databases include all the relevant reference data.

The primary criterion for the chemical selection was the availability of robust *in vivo* data (primarily LLNA and GPMT, but also human data) to allow for adequate comparisons to be made. Of the selected chemicals, approximately one third have already been tested using this method.

Specific observations:

The test chemicals for this validation study were selected by an independent Chemicals Selection Group (CSG) appointed by ECVAM and chaired by Dr. Thomas Cole (ECVAM).

2.2 Existing data used as testing data

Point 2.2 only concerns retrospective validation studies or modular studies combining existing and newly generated data to assess an assay. Which data sources were used to collect existing testing data?

Not applicable for main body of the study.
However, Appendix 13 presented a statistical analysis of 100 historical data (i.e. chemicals tested in house by the test method developer). Unfortunately, the chemicals tested had not been identified in this Appendix. Therefore, during review, the ESAC WG requested clarification on the chemical identities to be able to review the data on predictive capacity, also considering possible issues of applicability domain, when discussing false predictions.

2.3 Search strategy for retrieving existing data

*NOTE: Please describe and evaluate whether and how the search for existing data was planned, organised and conducted. In particular: has a **search strategy** been described and consistently applied?*

General observations:

The rationale, the strategy and the procedure followed for the chemicals selection and associated reference data are exhaustively described (pp. 14-21).

The two recognised databases (i.e., ICCVAM performance standards and P&G publications regarding the LLNA) provided a convenient source of authoritative data for selection of substances meeting criteria of being associated with reference testing data of sufficient quality. These datasets were considered reliable and sufficient for the purposes of the study design and objectives. A sufficient diversity of test substances to satisfy selection criteria was identified from these lists.

With regard to the selection of the historical data used to analyse subcategorisation (Appendix 13), **no** information had been provided with regard to the possible selection of this data set, i.e. does it constitute all data produced in house on subcategorisation? Or does it constitute a selection based on criteria? If so, which were the criteria?

2.4 Selection criteria applied to existing data

NOTE: Have consistent evaluation and decision criteria been pre-defined and applied in order to select the data and has the selection of data been explained in a transparent manner?

General observations:

The rationale for the selection of reference substances was adequately described under the criteria defining the data reliability.

Specific observations:

A primary eligibility criterion for the chemical selection was the availability of robust *in vivo* data to allow a proper comparative evaluation of the *in vitro* results. In particular, availability of both LLNA and GPMT *in vivo* data, with concordance of their corresponding skin sensitisation classification as an assurance of quality, formed the basis for short-listing candidate chemicals.

Availability of accepted human data was adopted as a secondary criterion, in cases where there would otherwise be an insufficient number of sub-sets of eligible chemicals as determined by the primary criterion.

Finally, the chemicals that were selected for the prospective part of the study (n=24) were part of chemical lists used for prevalidation of DPRA.

3. Quality aspects relating to data generated during the study

3.1 Quality assurance (QA) systems used when generating the data

NOTE: Have quality assurance systems such as GLP (Good Laboratory Practice) or GCCP (Good Cell Culture Practice) been used when generating the data?

Overall observations:

The study was not performed under GLP. The VMT defined and requested application of a minimum set of QA requirements, which were thought essential for the acceptance of the generated data. It was not clear how compliance with the SOPs was assured.

Specific observations:

The minimum set of requirements include (pp. 12-13):

- Qualified personnel, and appropriate facilities, equipment and materials;
- Records of the qualifications, training and experience, and a job description for each professional and technical individual;
- For each study, an individual with appropriate qualifications, training and experience shall be appointed to be responsible for its overall conduct and for any report issued;
- Instruments used for the generation of experimental data was inspected regularly, cleaned, maintained and calibrated according to established SOPs, if available, or to manufacturers' instructions. Records of these processes was kept, and made available for inspection on request;
- Reagents was labelled, as appropriate, to indicate their source, identity, concentration and stability. The labelling shall include the preparation and expiry dates, and specific storage conditions;
- All data generated during a study was recorded directly, promptly and legibly by the individual(s) responsible. These entries shall be attributable and dated;
- All changes to data was identified with the date and the identity of the individual responsible, and a reason for the change shall be documented and explained at the time.

The extent to which the study performance complied with GCCP was not described.

Quality of the test chemicals was provided by the supplier through the certificate of analysis. No further quality control was performed.

3.2 Quality check of the generated data prior to analysis

NOTE: Have the generated data been checked for quality including correct formatting (-> data reporting) prior to analysis. Has the quality check been performed by a staff member independent from the laboratory staff generating the data?

Overall observations:

It was not clear to the ESAC WG how the accuracy of the data was verified prior to the data being added to the templates for final data collection. Were the data verified prior to being analyzed either at ECVAM or at the testing laboratories?

Specific observations:

The results from Phase B experiments (WLR and BLR assessment) were submitted by the laboratories directly and exclusively to ECVAM by e-mail. The submitted templates were not protected/locked, and each completed template was formally quality controlled according to a checklist provided in App. 8. ECVAM was overall responsible for the data quality check once the raw data were received. The quality check focused on the acceptance criteria for the run and for each of the chemicals' data to ensure that the results were valid, and to confirm the correct selection of the CV75 concentration. Once completed, the checklists were scanned as a PDF file. The templates and checklists were then added to the official results folder of the study.

For the statistical analyses, a summary template for the relevant data produced by the laboratories was designed by the statistician contracted by ECVAM, and the results were transferred to this template by ECVAM. Preparation of this summary template contained internal checks based upon the performance of positive and negative controls that ensured that no transcription errors were made during the transfer of the results. As an additional check, the final conclusions/outcomes for each chemical were compared to the conclusions/outcomes in the reports sent by the laboratories.

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

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

*NOTE: Please describe the quality of the **testing data**. This may concern data newly generated in the context of the study and/or existing data (e.g. in case of retrospective or modular studies).*

General observations:

The h-CLAT SOP contained a set of acceptance criteria for the evaluation of runs and to determine whether the results obtained were valid or whether the run was to be discarded and the chemical retested. No modifications were made to this section of the SOP.

Specific observations:

The acceptance criteria were:

- Medium controls: viability should be > 90%.
- DMSO controls: viability should be > 90%.
- DNCB controls: viability should be > 50%
- CD54 and CD86 RFI values:
    - DMSO should be negative for both markers
    - DNCB should be positive for both markers

The proportion of invalid runs was generally low between laboratories (2.9 - 7.2%) (See table p.14). For the majority of these rejected runs, the acceptance criteria were not met for the positive control (DNCB): RFI values below the threshold (EC150, EC200).

There are some concerns regarding the reproducibility of (raw) data, even if the ultimate results are good.

Evaluation of surface markers expression versus viability (App. 12) presents some inconsistent data which may not be evident from just reviewing prediction results.

Relationship between reactivity, cytotoxicity and sensitization as demonstrated via surface marker expression? Seems somewhat too frequent that cytotoxicity and RFI change at similar concentrations.

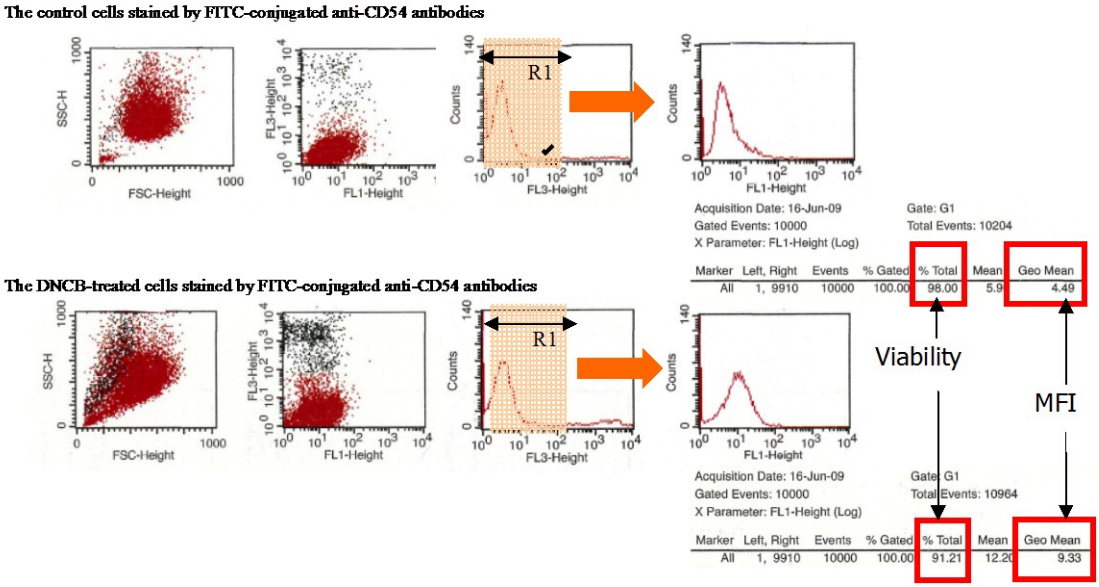
Despite apparent consistency in CV75 amongst laboratories, a good number of chemicals elicited viability approaching 50%. This was consistent between laboratories during testing.

In THP-1 cells, contact allergens are reported to induce phenotypic alteration including the production of pro-inflammatory cytokines and augmentation of cell surface molecules especially at sub-toxic doses (i.e. CV75). The ESAC WG was concerned by the induction of CD54 and/or CD86 by some chemicals occuring only at cytotoxic concentrations. Sakaguchi et al. (2009) demonstrated that the expression patterns of CD86/CD54 differ depending on the chemical challenge. For most

allergens, cytotoxicity (65-90% cell viability) is required for CD86/CD54 expression induction. Nukada et al. (2011) reported that cytotoxicity induced by contact allergens was caused by both apoptosis and necrosis. Apoptosis was preferentially induced by contact allergens, as the irritant sodium lauryl sulfate hardly induced apoptosis. Oxidative stress generated by contact allergens was found to be involved.

The other cytotoxicity-related concern of the ESAC WG was the possible interference with FACS analysis of autofluorescence associate with dying cells. As mentioned in the SOPs, CD54 and CD86 expression is evaluated after gating out dead cells. For expample as read in Annex 9 SOP vs 6:

Dead cells are gated-out by staining with PI (R1). Gating by FSC (forward scatter) and SSC (side scatter) is not done. A total of 10,000 living cells are analyzed. If cell viability is low, up to 30,000 cells including dead cells should be acquired. Alternatively, the acquisition can be finished 1 minute after starting of the acquisition.



The proportion of invalid runs was 3-7% for all laboratories, which seems a bit high. In general one would expect the percentage of invalid runs to be in the range of 1-3%.

Table 19: Overview of valid and invalid runs by laboratory.

Laboratory	Number of runs			Proportion invalid
	Total	Valid	Invalid	
Kao	69	65	4	5.8%
Shiseido	102	99	3	2.9%
Bioassay	174	162	12	6.9%
EURL ECVAM	69	64	5	7.2%

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

*NOTE: What is the quality of the **reference data** used? Are the data and their quality sufficient in view of the study objective? To which extent has the quality of the reference data impacted on the conclusions drawn reg. performance of the assay studied?*

General observations:

The reference data used to classification S/NS were considered reliable. The reference data were considered accurate in the characterisation of skin sensitisation classification, again based on previous expert-reviews.

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

NOTE: Having considered the quality of the testing data (section 4.1) and reference data (section 4.2), consider here whether the quality of the entire data set was sufficient to draw robust conclusions?

General observations:
The reference data on the selected test chemicals and their quality are sufficient in view of the stated objective of the study.

[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 definition (Module 1)

5.1 Quality and completeness of the overall test definition

NOTE: This included an analysis of the description of the (a) test system, (b) the protocol, (c) test acceptance criteria, (d) prediction models, (e) biological and/or mechanistic relevance of the test method for the target organ/species/system etc.

General observations:

The description of the h-CLAT assay (p. 25) would benefit from a more detailed rationale behind the selection of the THP-1 cell line, and the CD86 and CD54 membrane markers.

The test protocol developed throughout the study to result into version 7, suggested for future use of the h-CLAT assay. The acceptance criteria were clearly formulated.

The prediction model gives clear guidance as to when to define a compound as a sensitizers or a non-sensitizer.

The biological and mechanistic relevance of the test is not sufficiently explained (See section 1.2 (a)). There is ample evidence showing that CD86 and CD54 are generally up-regulated in response to challenges that cause cell damage, inflammation and cytotoxicity. There is a need for explaining what special features of the test or the prediction model are making the test specific for sensitization. See above comments.

Specific observations:

The protocol (Version 7) was prepared over the course of the validation program and was released by the VMG at the end of the validation study to include the provision and conditions to allow performing more than one run with the same chemical on the same day for the purpose of increasing throughput.

The test acceptance criteria were clearly formulated in all versions of the SOP (App. 9/10, pp. 24-25: qualifying experiment and data).

The prediction model gives clear guidance as to when to define a compound as a sensitizer or a non-sensitizer:

If the RFI of CD86 is ≥ 150 at any dose in at least 2 of 3 independent evaluation runs

AND/OR

If the RFI of CD54 is ≥ 200 at any dose in at least 2 of 3 independent evaluation runs,

THEN

The chemical is considered to be a sensitiser.

Otherwise, the chemical is considered to be a non-sensitiser.

Additional guidance was provided:

- The final maximum concentration of a test chemical should not exceed 5000 µg/mL (if the vehicle is saline) or 1000 µg/mL (if the vehicle is DMSO).
- When the chemical is tested at 5000 µg/mL in saline, 1000 µg/mL in DMSO, or at the highest soluble dose as the maximal test concentration instead of the CV75-based dose and does not meet the positive criteria as stated above without affecting cytotoxicity at all tested doses, it is labelled a non-sensitizer. It should be noted that recent results suggest that the h-CLAT tends to produce false-negative results with sensitising chemicals with log Kow value greater than 3.5. A negative result should be interpreted taking this into consideration.

There is a cryptic explanation about what the h-CLAT actually measures (p. 25), but how the measurements obtained help to distinguish between chemicals with or without sensitising potential is not explained. There is equally no reference to the original published work on the h-CLAT method to help the reader rationalise the development and the need for method validation. The only reference included here is the 'comprehensive review' by Adler et al 2010.

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

NOTE: Are the SOPs sufficiently detailed and complete? Are the prediction models sufficiently well explained to be applied in the correct manner?

General observations:

There are several modifications to the original protocol that resulted from a detailed examination of the SOP by ECVAM. Each SOP version was described in detail.

SOP version 3 was used for Phase A (transfer); SOP version 5 (see Table 5 (pp. 29-31) for changes in the protocol) was used in Phase B-1; SOP version 6 (see Table 5 (pp. 29-31) for changes in the protocol) was used in Phase B-2.

Version 7 (App. 10) is the version suggested by the VMG for future running of the method, issued at the end of the validation study (allowing more than one run with the same chemical on the same day).

The prediction model is clear and guides the user to a decision S/NS.

6. Test materials

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

NOTE: Is the number of test items sufficient in order to draw conclusions with respect to the objective of the study? If not, are there reasons for deviations and are these explained and justified?

General observations:

The number of evaluated chemicals (N=24) was sufficient in view of the primary study objective: WLR (n=15) and BLR (n=24).

Concerning the secondary objectives (S/NS classification and subcategorization) the following observations were made:

1. S/NS classification: 24 chemicals must be considered as insufficient for determining the potential of the test to accurately make this distinction. However, the obtained values for sensitivity, specificity and accuracy were similar to is in line with the published data (N=100). Subcategorization: The data obtained with the 24 chemicals tested in the current study support the concern raised by the VMT that additional evaluation with more chemicals is required.

6.2 Representativeness of the test items with respect to applicability

NOTE: Describe how suitable the selected test items are in order to gain – through empirical testing during the study – insight into the applicability domain / limitations of the test method.

Overall observations:

The main objective was to assess WLR and BLR, whereas defining the applicability domain of the test was not the purpose of this study.

The test chemicals (Table p.17) were chosen by experts and include a variety of chemicals covering various reactivity mechanisms as well as non-reactive species, variety of potency categories and solubility.

The ESAC WG considers the limited number of chemicals insufficient to address an applicability domain for this test. The existing results for 100 substances (Ashikaga et al., 2010) may help provide information about the applicability domain of the test.

Specific observations:

Among the 24 chemicals used in the study, about half have characteristics that can be considered as limiting:

- *Pre/pro haptens:*
On the background of the low CYP activity in THP-1 cells pro-haptens are considered outside the applicability domain of the h-CLAT. The chemical list of this validation study contains a well

characterised pro-hapten (Dihydroeugenol), which in principle should fall out of the applicability domain, but all three were correctly classified.

- *Auto-fluorescent chemicals:*

Since the h-CLAT uses a fluorescently labelled antibody, auto-fluorescence of test chemicals might interfere with the flow cytometry acquisition (e.g. Abietic acid).

- *Solubility and stability:*

In the h-CLAT test method the test substance needs to be dissolved in a solvent compatible with cell culture conditions (i.e. saline or DMSO). In the submission to ECVAM, chemicals reported wrongly classified because of solubility problems include: e.g. Abietic acid and Phthalic anhydride.

- *Metals may or may not be misclassified:*

Beryllium sulphate and nickel chloride were included in the study. Beryllium sulphate, but not nickel chloride, was found to give non-concordant results.

- *Volatile compounds:*

The discordant results obtained with benzyl salicylate, formaldehyde, limonene, methyl salicylate, and xylene suggest that volatile compounds may cause unreliable results in the h-CLAT as in most other assay. The other problem with volatile compounds is the 'contamination' of other wells; controls cells should be in a different plate.

- Weak sensitizers:

Weak sensitizers (e.g. Hexyl cinnamic aldehyde) may cause misleading negative results also in the h-CLAT (as in many other alternative test methods).

Table 2: List of the 24 chemicals selected for the coded testing phase.

	Chemical Name	CAS#	State	LLNA	LLNA potency category	GP	EC3	GHS potency category	DPRA R&D result	hCLAT R&D result	MUSST R&D result
SENSITISERS	Beryllium sulfate	7787-56-6	Solid	+	extreme	+	0.001	1A			
	Kathon CG (1.2% CMI)	26172-55-4	Liquid	+	extreme	+	0.009	1A	+	+	+
	Benzoquinone	106-51-4	Solid	+	extreme	+	0.0099	1A	+	+	+
	4-Phenylenediamine	106-50-3	Solid	+	strong	+	0.11	1A	+	+	+
	Chlorpromazine HCl	69-09-0	Solid	+	strong	+	0.14	1A			
	Chloramine T	127-65-1	Solid	+	strong	+	0.4	1A			
	Formaldehyde	50-00-0	Liquid	+	strong	+	0.61	1A	+	+	+
	2-Mercaptobenzothiazole	149-30-4	Solid	+	moderate	+	1.7	1A	+	+	
	Benzylsalicylate	118-58-1	Liquid	+	moderate	+	2.9	1B			
	1-Thioglycerol	96-27-5	Liquid	+	moderate	+	3.6	1B			
	Dihydroeugenol	2785-87-7	Liquid	+	moderate	+	6.8	1B			
	Nickel chloride	7718-54-9	Solid	-	no category[1]	+		1B			
	Benzylcinnamate	103-41-3	Solid	+	weak	+	18.4	1B			
	Imidazolidinylurea	39236-46-9	Solid	+	weak	+	24	1B	+	+	+
	R(+)-Limonene	5989-27-5	Liquid	+	weak	+	69	1B			
	Methylmethacrylate	80-62-6	Liquid	+	weak	+	90	1B			
NON-SENSITISERS	Glycerol	56-81-5	Liquid	-	no category	-		NC[3]	-	-	-
	2,4-Dichloronitrobenzene	611-06-3	Solid	-	no category	-		NC			
	Benzyl alcohol	100-51-6	Liquid	-	no category	-		NC			
	Methylsalicylate	119-36-8	Liquid	-	no category	-		NC	-	-	-
	Isopropanol	67-63-0	Liquid	-	no category	-		NC	-	-	-
	Dimethylisophthalate	1459-93-4	Solid	-	no category	-		NC			
	4-Aminobenzoic acid	150-13-0	Solid	-	no category	-		NC			
	Xylene	1330-20-7	Liquid	+	weak[2]		95.8	NC			

[1] False negative in the LLNA

[2] False positive in the LLNA

[3] NC: Not Classified

7. Within-laboratory reproducibility (Module 2)

7.1 Assessment of repeatability and reproducibility in the same laboratory

*NOTE: How was variability and reproducibility within laboratories assessed? Possible parameters to study are (a) intrinsic data variability e.g. between replicates or runs; (b) concordance in predictions between replicates or runs. Regarding point (b), consider whether reproducibility and repeatability have been assessed separately. [**repeatability** = agreement of test results (same substance, identical conditions, e.g. equipment, operator etc.) while **reproducibility** = agreement of test results (same substance, same protocol, but not under identical conditions, e.g. different operator).*

Overall observations:

Several aspects were assessed during this phase in all four participating laboratories using 15 coded chemicals. Each lab looked at the concordance of predictions (S/NS), reproducibility of CV75 and reproducibility of EC values for the two biomarkers.

Reproducibility primarily assessed using predictive results and not much consideration was given to WLR or BLR repeatability of independent runs (App. 12).

Each of the four labs ran three independent replicates for 15 chemicals. The *a priori* target for acceptability was 85%. Only one of the four labs (86.7%) met the target for WLR. Two labs were fairly close to the target (80%) and one fell considerably below target (73%).

Specific observations:

Tables 8 to 11 are supposed to summarize the reproducibility of the h-CLAT in the four laboratories. These table are not clear in terms of the message they want to bring and inconsistent scores were identified (e.g. Table 9, chemical 10 is scored P but should be N).

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

NOTE: Are the conclusions on within-laboratory variability and repeatability reproducibility justified by the data as evaluated?

The WLR was assessed at the level of concordance with a binary prediction (S/NS). An average reproducibility of 80% (KAO (86.7%), Shiseido (80.0%), EURL-ECVAM (80.0%) and Bioassay (73.3%)) did not meet the 85% reproducibility target set by the VMG. Actually, only one out of four participants met this target.

Despite missing the expected performance level, the VMG still concluded that the h-CLAT is a reproducible method.

This conclusion was partly based on the premise that a subset of chemicals consistently drove discrepancies in reproducibility, and some of these problem chemicals might fall outside the applicability domain. While the chemical limitations of the test are appreciated, the ESAC Working Group is concerned that there may be other inherent characteristics of the h-CLAT, which could be important sources of variability (e.g. the time course for expression of the cell surface markers; the state of cell differentiation be a source of variability).

The other reason given by the VMG to support their conclusion is that the h-CLAT assay is intended to be used as part of an ITS. With respect to the h-CLAT, this argument is weakened by the low reproducibility of the test. The Working Group is concerned that potential ITS building blocks with poor reproducibility might actually create more difficulties in interpreting data as part of an ITS due to conflicting results.

8. Transferability (Module 3)

8.1 Quality of design and analysis of the transfer phase

NOTE: Was the transfer phase appropriately planned, e.g. were there transfer instructions, training, minimum requirements, training SOP (if appropriate)? Were evaluation / decision criteria established beforehand defining successful transfer? If so, were these consistently applied during the analysis?

Overall observations:

The training and transfer phases of the validation study were meticulously planned and executed. All the stages appear well documented. It is clear from the transfer data that adopting this method in a laboratory requires sufficient experience in flow cytometry and cell culture.

Specific observations:

The transfer chemicals were tested uncoded, but Sigma catalogue numbers were provided to ensure the same chemicals were used in all laboratories, which is adequate. The transfer was run in stages with each stage reported to the lead labs for analyses and permission to proceed further based on the outcomes of the previous phase.

More stringent criteria were applied for run acceptance where both CD86 and CD54 had to be positive. This is probably due to the choice of chemicals, where the selective positive chemicals were meant to be very obviously positive.

The importance of the flow cytometry set up was highlighted during the transfer phase. In one laboratory, the auto-sampler setup on the instrument was successfully modified by keeping samples on ice in the dark, and testing one chemical per plate. Analogue and digital flow cytometry instruments were compared by the lead laboratories in a separate study but no additional differences were identified. The transfer phase was delayed due to these issues, but the delay was approved by the VMG.

Further issues were identified, in particular the need to carefully determine the CV75, as failure to determine this concentration can lead to problems. Also, it was decided to use lactic acid instead of SLS in the 'reactivity checks' as lactic acid gave a lower rate of false positive results. This choice was shown to be a correct one by the inconsistent results obtained with SLS throughout the study.

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

NOTE: Are the conclusions justified by the data generated? Have critical issues that may impact on transferability been identified?

The overall conclusions are justified by the generated data. Some key issues have been identified during the process of transfer to the naïve laboratories and effort has been put into identifying and solving these issues. These changes were taken up in SOP version 5 (used for transfer) and resulted in SOP versions 6 and 7.

Training and demonstration of competence in the conduct of the assay is, however, considered important. In particular, the chemicals used during this study's transfer phase should be considered for transfer in the future.

9. Between-laboratory reproducibility (Module 4)

9.1 Assessment of reproducibility in different laboratories

NOTE: How was variability and reproducibility between laboratories assessed? Possible parameters to study are (a) intrinsic data variability; (b) concordance in predictions between laboratories.

Overall observations:

Two BLR values were generated by testing 24 chemicals, one comparing the concordance of predictions obtained by the two naïve labs with the first lead lab and the second comparing them with the second lead lab. These BLR values (83.3% (vs KAO) and 79.2% (vs Shiseido)) are comparable, with the latter being slightly below the pre-set 80% target. It appears from table 13 (p. 56) that 5 chemicals were problematic.

It is worthwhile mentioning that the reproducibility between the experienced laboratories was 87. 5%, and thus above the 80% target set by the VMT.

Specific observations:

As shown in Section 6.2 in this report (Table 2), three chemicals (methyl methacrylate, DCNB and benzyl alcohol) were consistently, reproducibly wrongly classified.

Historical data revealed a BLR of 87% testing 15 chemicals (App. 2). Five of these 15 chemicals were common with perfect concordance in both studies (PPD, DNCB, kathon, glycerol, 2-mercaptobenzothiazole).

9.2 Conclusion on reproducibility as assessed by the study

NOTE: Are the conclusions justified by the data generated?

The conclusions in the report with respect to BLR seem reasonable in light of the marginal difference between the lowest BLR (79.2%) and expected performance of 80%.

The chemicals that drove discrepancies in the BLR study were the same as those driving discrepancies in the WLR study. Again, this led to the conclusion that these problem chemicals might fall outside the applicability domain. As discussed in Section 7.1, it appears unfortunate that that 5 of 24 expert-selected chemicals fall outside the applicability domain of the test. The Working Group suspects that there may be other explanations related to inherent biological variability for the discordant results.

10. Predictive capacity and overall relevance (Module 5)

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

NOTE: How was the predictive capacity assessed? Where the reference data used in an appropriate manner? Are the conclusions justified based on the data evaluated and in view of the test method's purpose?

The study was not intended to assess predictive capacity.

In agreement with the comments made in WLR and BLR sections, there are some concerns with regard to predictive capacity of this assay.

1. For S/NS classification, values are, overall, lower than the values resulting from the historical data on 100 chemicals (Ashikaga et al. 2010. ATLA 38; 275-284), which were submitted to EURL-ECVAM as part of the test submission.

	Number of compounds	Concordancy (%) (with LLNA)	Sensitivity (%)	Specificity (%)
Ashikaga et al. (2010)	100	84	87	75
Validation Study Report	24	76 (83.3 – 70.8)	81.3	65.6

2. Sub-categorization: From the data generated and statistically assigned cut-offs, it seems that a maximum of 58% accuracy can be reached with the 24 chemicals tested in the current study. The historical data (N=100) determined an accuracy of 72% (Ashikaga et al. 2010. ATLA 38: 275-284; Appendix 13).

In the set of 100 chemicals, 3 substances with reference values of 1A and 6 with reference values 1B were wrongly scored as non-sensitizers (false negatives). Eight 1A substances were underpredicted as 1B substances. Since the chemical identities were not identified, it was neither possible for the ESAC WG to identify the possible reasons for these misclassifications nor to understand how these misclassifications relate to the lower accuracy of sub-categorization reported for the submitted validation study.

Possible reasons for the observed discrepancy were discussed: i) updating of the SOP, ii) inclusion of substances previously used for test development and/or optimization, and iii) bias due to mixing of training and testing set.

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

NOTE: Are the conclusions reg. biological/mechanistic relevance and relevance in terms of making accurate predictions/measurements for the specific toxicity effect justified by the evaluated data?

The ESAC Working Group highlights below what it believes to be the key-conclusions of the VMT on module 5 (pp. 65 of the report):

1. The VMG concluded that the predictive accuracy of 76% is lower than previously published information (Ashikaga et al. 2010. ATLA 38; 275-284) on the predictive capacity of the h-CLAT.
2. Based on the accuracy of GHS sub-categorization of the proposal made by Dr. Omori and Dr. Yoshimura, the VMG also agreed that the initial results were encouraging, and that further evaluations will be necessary to determine how information generated by the h-CLAT can successfully contribute to potency sub-categorization.

The Working Group recognizes the fact that this study was not designed to address the predictive capacity of the h-CLAT due to the low number of chemicals. This also applies to the sub-categorization.

However, having recognized this, it was not clear for the WG why the VMG stated (VSR, p65) that the results on sub-categorization were "encouraging" based on the analysis of Drs. Omori and Yoshimura. The WG could not identify such qualifying statement in their statistical report.

In contrast to other tests having low accuracy with moderate and weak sensitizers, the main challenge for the h-CLAT seems to be the accurate identification of strong (1A) sensitizers. It can be speculated that this may be due to high cytotoxicity of strong sensitizers.

11. Applicability domain (Module 6)

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

NOTE: When considering the objective of the study, was the study designed in a way to enable conclusions on the applicability domain, the limitations and possible exclusions (e.g. technical incompatibility of the test method with specific chemicals)?

Assessment/description of the applicability domain was not the objective of this study. Consequently, the small number of chemicals used in the validation study, which was set to satisfy the primary goal of the study, is not sufficient on its own to draw robust conclusions on predictive capacity nor on applicability domain.

Empirically, the test method was set to have the following limitations with respect to the chemicals that can be tested:

- Lacking sufficient metabolic activity, the h-CLAT was considered not to be able to identify pro-haptens.
- Auto-fluorescent compounds can be anticipated to interfere with cell-sorting based upon fluorescent labels.
- Chemicals of low water solubility/stability can be expected to produce false negative results in a water (cell culture medium) environment.
- Metal salts (forming co-ordination bonds with specific amino acids) often perform badly in cell-based assay systems.
- Considering the incubation time at 37^{o}C, volatile compounds are likely to be underscored

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

NOTE: When considering the objective of the study and the data generated/analysed, have the applicability domain, the limitations and the exclusions of the method been sufficiently described?

The WG considered the provided information to be insufficient.

12. Performance standards (Module 7)

Section 12 Not applicable to this study.

13. Readiness for standardised use

13.1 Assessment of the readiness for regulatory purposes

NOTE: Is the test method ready for regulatory purposes? If yes, why? If no – what impediments currently exclude application for regulatory purposes?

For specific regulatory purposes (for example REACH), a positive h-CLAT result could be considered sufficient to classify a test material as a skin sensitizer (Sensitivity: 81.3%).

Given the low specificity of the h-CLAT (65.6%), the generated information should preferably be used in the context of a weight-of-evidence approach or ITS. It is important to use the test in a context that allows confident conclusions about the protein-reactivity of the chemical, especially when the chemical in question is negative in the DPRA. As such the method may be more helpful to address testing requirements of the REACH legislation and the 7th Amendment of the Cosmetic Directive as part of a ITS.

Since the predictive capacity of the test cannot be established from the submitted data, the h-CLAT cannot be used as a stand-alone assay in a regulatory context but should be considered for use in an Integrated Testing Strategy (ITS).

Regarding potency class, the data obtained did not support the use of the h-CLAT as a stand-alone assay for potency classification. This is in agreement with the statement of the VMG that the assay should be further evaluated for its capacity to "contribute" to a potency classification.

13.2. Assessment of the readiness for other uses

NOTE: Is the test method ready for other uses (e.g. screening purposes, testing to gain mechanistic insight, to generate supportive information for hazard/risk assessment).

The h-CLAT may be useful for a variety of possible screening purposes of chemicals expected to fall within the applicability domain of the test, once that domain has been determined. As yet, deciding whether or not a chemical falls within the applicability domain of the test is a challenge with regard to pro-haptens, metal (salts), chemicals with limited solubility/stability in water, volatile compounds and auto-fluorescent compounds.

13.3 Critical aspects impacting on standardised use

Note: What are the factors that may impact on standardised use (in regulatory or non-regulatory settings)?

Cell culture and flow cytometry skills and appropriate facilities are essential. .

Sensitizing chemicals falling outside the applicability domain of the test (e.g. metal salts) may or may not be identified by the h-CLAT.

The chemicals that drove discrepancies in the BLR study were the same as those driving discrepancies in the WLR study. It appears unfortunate that that 5 of 24 expert-selected chemicals fall outside the applicability domain of the test. The Working Group therefore suspects that there may be other explanations for the discordant results which are indicative of inherent biological variability (e.g. the time course for expression of the cell surface markers; the state of cell differentiation).

13.4 Gap analysis

NOTE: Identify, if appropriate, gaps in the study design and/or conduct that may have impacted on the stated study objective or the conclusions drawn.

With respect to the predictive capacity and potency class identification, the obtained values should not be considered as more then indicative.

14. Other considerations

NOTE: Please address any other consideration you might have in relation to the proposed approach under this section.

The other reason given by the VMG to support their conclusion is that the h-CLAT assay is intended to be used as part of an ITS. With respect to the h-CLAT, this argument is weakened by the low reproducibility of the test. The Working Group is concerned that the poor reproducibility might actually create more difficulties in interpreting data as part of an ITS due to conflicting results.

15. Conclusions on the study

NOTE: This section should present a brief summary of the study results and conclusions as described in the VSR (subsection 15.1), discuss to which extent the conclusions drawn in the study reports are justified by the study results on their own (subsection 15.2) and evaluate to which extent the conclusions are plausible with respect to other information (subsection 15.3).

15.1 Summary of the results and the VMG conclusions of the study

The **conclusions drawn by the VMG** as described in the VSR on the basis of the results shown in the report (pages 68 – 70 of Validation Study Report:

- *The WLR of the test method with respect to concordance of classification (S/NS) met the target of 85% and was considered sufficient for the purpose of this study.*
- *The data were considered strong enough to support transferability of the test to properly equipped, trained and staffed laboratories with the appropriate analytical capabilities.*
- *The BLR of the test method with respect to concordance of classification was considered to meet the 80% target.*
- *The number of chemicals (N=24) did not provide support for a firm conclusion about the predicative capacity of the test method.*
- *The number of chemicals did not allow drawing a conclusion about the applicability domain of the test. Empirically the applicability domain seems to exclude pro-haptens, auto-fluorescent compounds, chemicals with limited water solubility/stability, metal salts and volatile compounds.*

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

Having listed the VMG conclusions above (15.1), the ESAC WG summarises its appraisal as follows:

Overall, the test design and the quality of the selected chemicals (N=24) were considered appropriate for the purpose of addressing the first objective of the study: Assessing the WLR and BLR of the h-CLAT.

In agreement with the VMG statement, the number of chemicals was considered too small a sample size for allowing a firm conclusion about the predictive capacity (in terms of S/NS as well as potency classification) of the h-CLAT (secondary objective of the study).

Overall, the conclusions made by the WG correspond well with the conclusions drawn by the VMG as described in the VSR, indicating that these conclusions are supported by the results shown in the report (see Section 15.1). The ESAC WG disagreed, however, with the VMG conclusion concerning the WLR.

- The WLR was assessed using 15 chemicals in three independent experiments. The acceptance criteria were well described. An average reproducibility of 80% (KAO (86.7%), Shiseido (80.0%), EURL-ECVAM (80.0%) and Bioassay (73.3%)) did not meet the 85% reproducibility target set by the VMG. Actually, only one out of four participants met this target. Despite missing the expected performance level, the VMG still concluded that the h-CLAT is a reproducible method. This conclusion was partly based on the premise that a subset of chemicals consistently drove discrepancies in reproducibility, and some of these problem chemicals might fall outside the applicability domain. While the chemical limitations of the test are appreciated, the ESAC Working Group is concerned that there may be other inherent characteristics of the h-CLAT, which could be important sources of variability (e.g. the time course for expression of the cell surface markers; the state of cell differentiation be a source of variability). The other reason given by the VMG to support their conclusion is that the h-CLAT assay is intended to be used as part of an ITS. With respect to the h-CLAT, this argument is weakened by the low reproducibility of the test. The Working Group is concerned that the poor reproducibility might actually create more difficulties in interpreting data as part of an ITS due to conflicting results.
- The data were considered strong enough to support transferability of the test to properly equipped, trained and staffed laboratories with the appropriate analytical capabilities.
- Five of the 24 chemicals produced a discordant classification by the laboratories resulting in an average BLR reproducibility of 81.3%, meeting the target (80%).
- For S/NS classification, values (accuracy: 76%; sensitivity: 81.3%; specificity: 65.6%) are, overall, lower than the values (84%, 87%, 75%, respectively) resulting from the historical data on 100 chemicals (Ashikaga et al., 2010), which were submitted to EURL-ECVAM as part of the test submission.
- For sub-categorization, the data generated and statistically assigned cut-offs propose a maximum accuracy of 58% accuracy. The historical data (N=100) determined an accuracy of 72% (Ashikaga et al. 2010. ATLA 38; 275-284). These data support the statement by the VMT that 24 chemicals is not a sufficient number for assessing the potential of the test to subcategorize chemicals. Furthermore, the ESAC WG does not agree with the statement by the VMT that the results were promising. The mean reason for this disagreement was the observed and unexplained discrepancy between the historical data and the results obtained with the chemicals selected for this validation study and the lack of information furnished (chemical identities) that would have allowed an appraisal of the existing information on predictive capacity and subcategorisation.
- The number of chemicals did not allow drawing a conclusion about the applicability domain of the test (which, notably, was not one of the study objectives). Empirically the applicability domain seems to exclude pro-haptens, auto-fluorescent compounds, chemicals with limited water solubility/stability, metal salts and volatile compounds. However, pre-/pro-haptens were reported as correctly identified.

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

The h-CLAT was developed to represent one of the key events of sensitization using well characterized reference data. The test results are consistent with what is known about the reactivity and sensitization potential of test chemicals (Gerberick et al., 2005, 2007).

On this background the conclusions of the report are plausible in the context of the existing information.

16. Recommendations

Note: This section should provide recommendations on the test method (e.g. further work, possible use) and their constituting elements (e.g. test system, prediction model, SOP).

16.1 General recommendations

The h-CLAT addresses a key mechanism (DC activation) in the development of sensitization. Overall the provided data support BLR and transferability of the test to qualified laboratories.

According to the ESAC WG, the WLR was unsatisfactory as three of the four laboratories did not meet the 85% target. The ESAC Working Group is concerned that there may be other inherent characteristics of the h-CLAT, which could be important sources of variability (e.g. the time course for expression of the cell surface markers; the state of cell differentiation be a source of variability). The ESAC WG recommends that these sources be identified, and that solutions be provided because poor reproducibility may create more difficulties in interpreting data as part of an ITS due to conflicting results.

The ESAC WG recommends to better define (1) the predictive capacity and (2) the applicability domain of the h-CLAT (to eliminate the uncertainty currently associated with a negative result) either through further testing (i.e. prospective validation) or through retrospective analysis of existing information (retrospective validation: data grouping / meta-analysis).

The submitted SOP was amended during the study solely to provide clarifications on the procedure that was described. The WG recommends to reassess the SOP version 7 using existing/historical results with the purpose to re-evaluate the predictive capacity of this test method.

For hazard classification purposes, the SOP can be adapted to reduce resource costs by eliminating the need for a third evaluation run in case where the first two runs are consistent.

The submitted study does not provide strong evidence supporting the usefulness of the h-CLAT for GHS sub-categorisation of sensitizers. However, recent studies substantiate the preliminary data of the VSR.

Nukada et al. (2012) reported a correlation of h-CLAT data with LLNA EC3 and GHS sub-categorization. A statistically significant correlation was observed between h-CLAT concentration providing a cell viability of 75% (CV75), h-CLAT estimated concentration of RFI=150 for CD86 (EC150), and for CD54 (EC200) with LLNA's EC3. From EC150 and EC200, a minimum induction threshold (MIT) was determined as the smaller of either EC150 or EC200. MIT showed a correlation with EC3 (R=0.638) and approximate 80% accuracy for GHS sub-categories when a tentative threshold of 13 µg/mL was used (in the paper mentioned below they used 10 µg/mL). More recently, Nukada et al. (2013) reported a data integration strategy including HCLAT, DPRA and DEREK for the development of a test battery to predict the skin sensitizing potential and potency of chemicals. Using a dataset of 101 chemicals with LLNA, h-CLAT, DPRA and *in silico* prediction system, by converting the *in vitro* results into scores of 0–2, the sum of individual scores provided an accuracy of 85% and 71% for the potential and potency prediction, respectively, compared with LLNA. Similarly, using a tiered system of h-CLAT and DPRA an accuracy of 86% and 73% for the potential and potency prediction was

obtained. The tiered system showed a higher sensitivity (from 88 to 96%) compared with h-CLAT alone.

Further studies are needed to identify the best integrated testing strategy or strategies necessary to cover the different regulatory goals and risk assessments (hazard identification, classification, potency assessment, etc.).

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

The WG recommends including in the SOP practical procedures to deal with potential issues regarding differences in the sensitivity between flow cytometers.

17. References

Ashikaga, T., Sakaguchi, H., Sono, S., Kosaka, N., Ishikawa, M., Nukada, Y., Miyazawa, M., Ito, Y., Nishiyama, N., Itagaki, H. (2010)A comparative evaluation of in vitro skin sensitization tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). ATLA 38; 275-284.

Gerberick, G., F., Kern, P., S., Schlatter, H., Dearman, R., J., Kimber, I., Patlewicz, G., Y., Basketter, D., A. (2005) Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods. Dermatitis 2005: 16: 157-202.

Gerberick, G., F., Ryan, C., A., Dearman, R., J., Kimber, I. (2007) Local lymph node assay (LLNA) for detection of sensitization capacity of chemicals. Methods 41: 54-60.

Kern, P., S., Gerberick, G., F., Ryan, C., A., Kimber, I., Aptula, A., Basketter, D., A. (2010) Local lymph node data for the evaluation of skin sensitization alternatives: a second compilation. Dermatitis 21: 8-32.

Nukada, Y., Ashikaga, T., Miyazawa. M., Hirota. M., Sakaguchi. H., Sasa. H., Nishiyama, N. (2012) Prediction of skin sensitization potency of chemicals by human Cell Line Activation Test (h-CLAT) and an attempt at classifying skin sensitization potency. Toxicol. In Vitro 26: 1150-60.

Nukada, Y., Ito, Y., Miyazawa, M., Sakaguchi, H., Nishiyama, N. (2011) The relationship between CD86 and CD54 protein expression and cytotoxicity following stimulation with contact allergen in THP-1 cells. J. Toxicol. Sci. 36: 313-324.

Nukada, Y., Miyazawa, M., Kazutochi, S., Sakaguchi, H., Nishiyama, N. (2013) Data integration of non-animal tests for the development of a test battery to predict the skin sensitizing potential and potency of chemicals. Toxicol. In Vitro 27: 609-18.

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.

Sakaguchi, H., Ashikaga, T., Miyazawa, M., Kosaka, N., Ito, Y., Yoneyama, K., Sono, S., Itagaki, H., Toyoda, H., Suzuki, H. (2009) The relationship between CD86/CD54 expression and THP-1 cell viability in an in vitro skin sensitization test--human cell line activation test (h-CLAT). Cell. Biol. Toxicol. 25: 109-126.