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-led validation study on two *in vitro* hepatic human-derived test methods for assessing induction of Cytochrome P450 enzymes (CYPs)

Title page information	
File name	**ESAC-WG_CYP REPORT.doc**
Abbreviated title of ESAC request	Validation project on CYP induction assay for the assessment of human metabolic competent hepatic test systems
Relating to ESAC REQUEST Nr.	2013-01
Request discussed through	ESAC 39, March 2014
Report to be handed over to ESAC Chair and EURL ECVAM Coordinator by	**Q2/Q3 2014**

Version tracking

Date	**Version**	**Author(s)**	**Description**
19 / 9 /2014	Final for ESAC 40	HC, WG, edits CG	Final version for discussion at ESAC 40. WLR and BLR re-calculations described. Re-calculations had been requested by WG in April 2014 (WG meeting)

TABLE OF CONTENTS

ESAC Working Group

This report was prepared by the "ESAC Working Group on CYP Induction" (ESAC WG), charged with conducting a detailed scientific peer review of the ECVAM-led validation study on two *in vitro* hepatic human-derived test methods for assessing induction of Cytochrome P450 enzymes (CYPs).

The ESAC WG had been set up by the ESAC following its meeting in June 2013. Agreement on the WG composition was reached by written procedure. Basis for the scientific review was the ECVAM request to ESAC concerning the scientific review (ESAC request Nr. 2013-01).

The ESAC WG conducted the peer review from April 2014 to September 2014. This report was endorsed by the ESAC WG on 19. September 2014 and represents the consensus view of the ESAC WG.

This ESAC WG peer review consensus report was endorsed by the ESAC on 21/22 October 2014.

The ESAC WG had the following members:

- ESAC members:
  - Harvey Clewell (WG Chair)
  - Claus-Michael Lehr
  - Jürgen Borlak
  - Annette Kopp-Schneider (focusing on statistical aspects)
- External experts:
  - Emanuela Testai, ISS, Rome, Italy (proposed by WG Chair)
  - Stephen Ferguson (nominated by NICEATM/ICCVAM and supported by ECVAM).

ESAC Coordination:

Dr. Claudius Griesinger (ESAC Coordinator)

Dr. Michael Schäffer

ABBREVIATIONS USED IN THE DOCUMENT

Formatting Examples below

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 validation study on two human-derived hepatic metabolically competent test methods for cytochrome P450 induction measurement, an ESAC Working Group (ESAC WG) was set up by ESAC. The ESAC WG was charged with conducting a detailed scientific peer review of this study which had addressed the reliability and predictive capacity of the methods for distinguishing potential inducers from non-inducers of three CYP isoforms.

The ESAC WG met in person at ECVAM in April 2014 and communicated further by email and teleconferences. The ESAC WG reviewed the validation study reports, statistical reports and all relevant documentation. The ESAC WG considered the scientific work presented was of excellent quality, despite some weaknesses in the characterization of the predictive capacity.

The main objective of this study was to assess the transferability, the reproducibility (within and between laboratories) and the predictive capacity of two Cytochrome P450 induction in vitro methods, each of which evaluated the induction of enzymatic activity of four CYP enzymes (CYP1A2, CYP2B6, CYP2C9, and CYP3A4).

The two CYP induction *in vitro* methods used different metabolically competent *in vitro* test systems: (a) cryopreserved human HepaRG cells and (b) cryopreserved human primary hepatocytes. Predictive capacity was assessed using exclusively human CYP induction reference data relating to chemicals used for pharmaceutical purposes. The WG agrees with the VSR assessment that the study findings satisfy the requirements for test definition, within laboratory reproducibility, transferability, and between laboratory reproducibility, but only partially satisfy the requirements for assessment of predictive capacity.

However, the WG believes that the VSR overstates the readiness of the test for regulatory use. In particular the WG recommends that additional CYP induction studies be conducted with rodent hepatocytes to further investigate the applicability domain and predictive power of the assay for environmental chemicals that are not pharmaceuticals. Nevertheless, the WG agrees that there may be a potential role for a human CYP induction assay as a marker of possible receptor activation as part of an integrated testing strategy for a particular Adverse Outcome Pathway. The assay also has potential use for evaluation of the human relevance of animal test results, whether *in vivo* or *in vitro*, that suggest activation of a receptor is a key element in a toxicity pathway.

The WG feels that the current study is a good reflection of expectations based on other existing information regarding hepatocyte assays, and that it provides evidence that reliable hepatocyte assays for other important purposes, including identification of metabolites and quantification of metabolic clearance, are feasible.

The WG strongly encourages ECVAM to continue to conduct studies with human hepatic models to develop methods for characterization of other kinetic data, including clearance, metabolic profiling, and inhibition. In this context, the importance of developing *in vitro* to *in vivo* extrapolation methods cannot be overemphasized.

1. Study objective and design

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

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

The main objective of this study was to assess the transferability, the reproducibility (within and between laboratories) and the predictive capacity of two Cytochrome P450 induction in vitro methods, each of which evaluated the induction of enzymatic activity of four CYP enzymes (CYP1A2, CYP2B6, CYP2C9, and CYP3A4). The two CYP induction in vitro methods used different metabolically competent in vitro Test Systems (TS): cryopreserved human HepaRG cells and cryopreserved human primary hepatocytes. Predictive capacity was assessed using exclusively human CYP induction in vivo reference data.

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

The primary objective of this study is clear: determining the reproducibility / predictive capacity for prediction of CYP induction activity.

General Observations

a) The WG believes the value of the information generated by the test / potential usefulness of the information (e.g., for predicting CYP induction) may be overestimated in the VSR. The approach actually characterizes metabolic activity based on increases in metabolite production from probe substrates, which is not necessarily predictive for receptor pathway induction. The use of mRNA measurements would be a more proximal endpoint to receptor activation and complementary for this purpose.

b) CYP induction is described as providing evidence of biotransformation. However, induction and metabolite production are different processes that may not be causally related. There are many compounds that are substrates for a particular enzyme but do not lead to induction of the activity of that enzyme (e.g., acetaminophen, caffeine). Conversely, there are compounds that can induce a particular enzyme despite the fact that they are not effectively metabolized by it (e.g., dioxins and CYP1A2, perfluorinated alkyl acids and CYP4A). Thus a CYP induction assay should not be considered a method for evaluating the likelihood that a compound is significantly metabolized.

c) Decreases in probe substrate metabolite production is suggested to indicate an inhibition of metabolism by the tested chemicals, but this conclusion could be confounded by cytotoxicity or suppression of metabolizing enzyme expression.

d) CYP induction is also mentioned as a potential key event in a toxicity pathway; however, ESAC WG members are concerned that the evidence for a link between enzyme induction and human toxicity from low concentration exposures is limited at best. Nevertheless, the WG members agree that there may be a potential role for a CYP induction assay as a marker of possible receptor activation as part of an integrated testing strategy for a particular Adverse Outcome Pathway.

e) The WG believes that in order to be able to provide pertinent mode of action information regarding adverse outcomes observed in animal testing, CYP induction assays should also be assessed in rat hepatocytes. This would also allow for a greater variety of in vivo induction data for determination of the applicability domain for a CYP induction assay across a wider range of physicochemical properties and receptor pathways than could be accomplished with the human hepatocyte assays (c.f. section 13.4 which provides more details on this proposal of the WG).

f) Finally, the VSR suggests that another potential value of the CYP induction assay may be to identify possible mixture or substance / substance interactions; however, this is not sufficiently addressed to permit evaluation.

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.

As outlined in section 4.1 (page 42) of the VSR, the overall purpose of the test method is for regulatory testing, in particular to identify compounds that induce metabolism associated with activation of potential toxicity pathways, including CAR, PXR, and AhR, with possible application for elucidating mode of actions, predicting toxicity/metabolic activation, and providing a biomarker of exposure.

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

NOTE: Is the scientific rationale of the test method AND (consequently) for executing the study clearly explained? How does the test method contribute to
(a) the scientific understanding of the specified health/environmental effect or aspects of it?, i.e. does it provide relevant mechanistic information (toxicity pathways, key physiological events leading to toxicity)?
(b) the prediction of the specified health/environmental effect or aspects of it?
Does the VSR make sufficient reference to the relevant body of scientific literature?

The scientific rationale for the assay is stated in the context of the study objectives (section 3.1), as well as under Module 1 (section 4.1, which provides a short description of the intended purpose). The rationale is described in more detail under the secondary objectives, which state that the assay is intended to contribute to knowledge regarding:

- **CYP induction as a toxicity event (page 13, section 3.1.2)**
- **Elucidation of MoA**

- **Biomarkers of exposure to chemicals (if CYPs are induced – there may have been significant exposure)**
- **Potential for effects on mixture toxicity**
- **Indication of a role of metabolism in a compound's toxicity**

The WG believes that the clarity of the section on the scientific rationale would have benefitted from restricting the explanations to what the test method measures, i.e. identification of chemicals that lead to induction of four selected CYP molecules in human hepatic test systems. The WG feels that the text on the potential contribution of the assay for TK distracts from the explanations on the scientific rationale.

General Observations

a) In future, it should be clarified that while part of the purpose of the assay may be to identify potential effects of induction by the studied chemical on the metabolism of other xenobiotics and endogenous compounds to which an individual may be co-exposed, the assay does not characterize metabolism of the studied xenobiotic itself.

b) There is sufficient background/literature discussed/presented, however most of this background is linked to drugs/pharma (4.2 and 4.6) rather than to environmental compounds. The justification for the use of human test systems for induction are well presented, but the resulting limitations in the evaluation of the assay are not adequately discussed (see next section). Application of this assay to different classes of chemicals is suggested but not supported, and indeed is problematic (e.g., volatiles, lipophilics).

Specific Observations

a) The Cyp assay does not allow separate identification of inhibitory effects that could affect the assay outcome.

b) AhR is not a nuclear receptor.

c) It is not possible to "avoid" interspecies uncertainty, but it can be reduced.

d) It is important to distinguish receptor activation vs. receptor binding.

(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?

A regulatory rationale is provided but it is linked primarily to drugs (page 53, section 4.6). In order to facilitate inclusion in a OECD PBTG and support EU legislations (REACH, Cosmetics, Animal Welfare), as stated in section 4.2, the regulatory rationale/usefulness should have been expanded also to other chemicals of concern for environmental/cosmetic exposure. Limitations in the scientific and regulatory usefulness of the assay are not sufficiently addressed, e.g., the spectrum of isoforms is limited to those affecting pharmaceuticals and the variability of CYP induction profiles in different tissues/cells is not addressed. The latter point is important since exposure routes to environmental chemicals are not only oral (where hepatic induction is very relevant), but can also be inhalation or dermal.

General Observations

The assay also has potential use for evaluation of the human relevance of animal test results, whether in vivo or in vitro, that suggest activation of a receptor is an key element in a toxicity pathway.

The justification of the potential use of the assay for evaluating the impact of CYP induction on toxicity from exposure to other well characterized chemicals was inadequate.

In future descriptions of this assay, care should be taken not to imply that CYP induction, per se, is directly linked to toxicity or the potential for metabolite formation

1.3 Appraisal of the appropriateness of the study design

NOTE: This includes an analysis of the number of laboratories involved in the study, the organisation of study management, the statistical analysis and may include more technical aspects such as (a) a brief appraisal of the nature and number of test items used (details however to be provided in section 6, test items), retesting in case of unqualified tests and others.

The number of laboratories in this study is considered appropriate and the number of replicates (3) is considered adequate. The number of substances is considered to be too small for a confirmative data analysis of predictive capacity. A sensitivity, e.g., of 5/5 (Table M5.1) corresponds to a point estimate of 100%, but an exact 95%-Confidence interval ranging from 48% to 100%. This shows that proof of good predictive capacity can only be achieved by a strong increase in number of tested compounds. However, the WG also acknowledges that it is not always feasible in the context of validation studies to assess a sufficiently high number of chemicals. Practical constraints such as the availability of good reference data (and hence test chemicals with accompanying high quality data), the cost and time factors to be considered when organising practical testing as well as other factors, may impact on the final sample size that, realistically, can be studied. The WG considers that these factors (in particular availability of reference data) impacted on the sample size used in the CYP induction validation study. The WG nevertheless holds that the sample size, despite being insufficient for confirmative analysis, provides a good indication on the suitability of the chosen test systems for studying induction of CYP enzymes.

General Observations

Reference chemicals are well chosen, but it would have been valuable to include more compounds (non prototypical, weaker inducers), with particular emphasis on representative chemicals from different classes with non-druglike properties (e.g., persistent / highly lipophilic compounds, rapidly metabolized compounds, poorly soluble compounds). A broader diversity of chemical properties would increase the level of understanding of the applicability domain for this assay. The WG recognizes that finding human in vivo data for such a wide variety of chemicals would be problematic and therefore suggests consideration of a rat hepatocyte CYP induction assay to further explore the applicability domain (c.f. section 13.4).

Specific Observations

a) Comparison between the two test systems is problematic (i.e. primary cells and immortalized cell lines are too different to be able to be compared in terms of performance in a meaningful way). Thus some conclusions are potentially flawed (e.g., HepaRG is more reproducible: clearly a clone should be more reproducible than cells from individual donors with differing factors such as age, health condition, cause of death, medical background).

b) mRNA should be considered as a parallel endpoint together with activity and cytotoxicity to assess the induction potential of the compounds, in order to address the potential for inhibition, suppression, or toxicity that may confound induction assessments based on enzymatic activity. In addition, mRNA is more proximal to receptor activation and is required for drug submissions by regulatory agencies.

c) While the use of hepatocytes from three individual donors is consistent with current practice, it would be preferable (when it becomes practicable) to have a larger number of individual donor preparations or develop approaches that use pooled cells from a larger number (e.g., 10) of donors.

d) The concentration of phenobarbital (that gives a 2-fold induction) used was insufficient to saturate induction (to get an accurate EC50). The WG is not comfortable that the lot selection and test acceptance criteria were adequate, and feels that a greater fold-change for positive controls may have been preferable. Also, acceptance should be based on both fold-change in positive controls and metabolite generation in controls for cocktail exposures.

e) The Alamar Blue assay is primarily for mitochondrial activity, which can be perturbed by inducers without toxicity. Other biomarkers (LDH leakage, AST, ALT) would provide better markers for toxicity in hepatic cells.

f) The choice of IC30 as the cutoff for toxicity is not considered appropriate by the WG. It is probably responsible for the non-monotonic dose response behaviors observed. The WG would suggest no more than an IC10 as point of departure.

g) The WG recommends that ECVAM consider re-analysing the activity data without normalization for protein content (only normalize on plated cell number) as this can add additional variability due in part to dying cells and the use of an additional measure with additional sources of variation.

1.4 Appropriateness of the statistical evaluation

NOTE: Are the statistical methods used for evaluating the study data appropriate. Is the choice of methods sufficiently justified? Was the statistician independent from the test method submitter/developer?

The WG agrees with the use of a factor of 2 to identify efficacious inducers. The decision rule for a substance to be called an inducer is that at least in one concentration the factor of 2 is exceeded. In addition, we suggest in future to use ANOVA with post-hoc Dunnett test for multiple comparison to vehicle control and to exclude non-significant findings. WLR and BLR should be

evaluated on the call for a single curve (at least 1 concentration with >2 and Dunnett test significant), not on the individual concentrations.

General Observations

State exact 95%-confidence intervals for sensitivity and specificity.

2. Collection of existing data

NOTE: (Pre)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, (pre)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?

Human clinical data from pharma were used (Table 02), but the selection criteria were not provided.

2.2 Existing data used as testing data

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

Not relevant to this study.

2.3 Search strategy for retrieving existing data

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

The WG could not find a discussion of the search strategy.

2.4 Selection criteria applied to existing data

NOTE: Have consistent evaluation/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?

Documentation of the search strategy and selection criteria was not found in the report. However, the WG is satisfied with the compounds selected for validation of a human cell assay. The acceptance criteria for determining the specific studies to be used to characterize the in vivo induction are critical to assure that the points of comparison with the in vitro assay are correct, so in future these criteria should be documented.

3. Quality aspects relating to data generated during the study

3.1 Quality assurance systems used when generating the data

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

The WG is satisfied that quality systems were used. Pharmacelsus GmbH, Janssen Pharmaceutica and EURL ECVAM test facilities are certified as compliant to the GLP OECD principle. It is worth of note that EURL-ECVAM is certified for the 'validation of in vitro methods' (included in OECD category 9): although the topic is out of the scope of the EU regulation for the application of GLP, it was included as an area of interest by the GLP Italian Monitoring Authority, since the request came from a DG of the Commission. Indeed, there is no clear written regulation about the need for validation studies being conducted in full compliance with GLP. Therefore it is fully correct, that the test facilities carried out the studies 'according to' the GLP principles. For the analytical part performed at EURL-ECVAM, although not fully compliant to GLP, it can be considered again carried out 'according to the principle'; in addition the control/ maintenance of the instrument was under the ISO 17025 accreditation, which is considered sufficiently reliable. For the non ‐ GLP laboratories participating in the validation project, the minimum set of quality assurance requirements was considered appropriate. Regarding SOPs, please see section 5.2. Overall, the WG is comfortable that quality assurance was acceptable. Once the method is further validated and adopted as a guideline for a test to be used in the regulatory frame of safety assessment of chemicals or drugs, the test could be carried out in compliance with GLP principle (OECD category 2: toxicity testing).

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, the WG consider that the Quality Assurance system adopted during the study is acceptable, although it should have been more completely described in the report. On the other hand the reporting of data could be improved by standardizing it with a clear description in the SOPs, which is missing. The CYP induction SOPs contained a set of acceptance criteria for the evaluation of runs to determine whether the obtained results are valid. However, despite the length of the SOPs, no indication about the reporting of data is given.

The study is described as being carried out 'according to GLP principles' in test facilities certified as compliant to GLP principles; therefore, it is expected that the personnel of the QAU were independent from the laboratory staff generating the data (as well as from the Study director). It is stated in the minimum requirements for non GLP test facilities: 'Quality Assurance should be performed in accordance with the principles of GLP (for GLP compliant laboratories)'. From the parenthetical text, it is not clear whether this bullet point applied to the non GLP test facilities as well. Although the Quality Assurance responsibilities are also described in the Study Plan for non

GLP test facilities, the WG could not determine whether the QA activity was performed by a staff member independent from the laboratory staff generating the data.

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 validation studies).

The WG is generally satisfied with the quality of the data collected in this study. There are concerns regarding the large error bars in some tests with the HepaRG cells (statistical report figs 9-12). Variability in phase II metabolism could be a contributing factor in the overall variability of fold-change induction. With a modified study design addressing the suggestions above, the utility of the data collected and resulting ability to assess the overall quality of the data would be improved.

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?*

Unfortunately, the use of human cells and resulting focus on human in vivo reference data limits the value of this assay due to the narrow range of physico-chemical properties associated with drugs (high water solubility and bioavailability, low metabolism and lipophilicity). To fully assess the applicability domain for a CYP induction assay would require more chemicals, with particular emphasis on representative chemicals from different classes with non-druglike properties (e.g., persistent/highly lipophilic compounds, rapidly metabolized compounds, poorly soluble compounds); such a study would probably have to be conducted with rodent hepatocytes (c.f. section 13.4).

General Observations

The general approach of focusing on inducers with clinical relationships is logical and provides a framework for evaluating system effectiveness and relevance to humans. However, recent efforts in the field to provide more quantitative relationships between in vitro induction responses and predicted changes in pharmacokinetics (e.g., AUC) of probe drugs could be incorporated into the study designs for evaluation of the effectiveness of data generated within this 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."

A fuller characterization of the criteria for selection of studies on in vivo induction should have been provided. The WG concern is that specific CYP induction may be dose-dependent due to the variation in Km's and Vmax's for different isozymes.

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

NOTE: Are the data and their quality sufficient in view of the stated objective of the study?

In most aspects, the data collected in this study support the primary objective of the study (WLR, BLR, and predictive power for inducer/non-inducer) and for comparison of the two test systems (cryo Hep and cryoHepaRG), although the number of compounds is too small for reliable estimates of predictive capacity. Given the variability of analytical SOPs across laboratories, the comparability of their results is surprisingly good. However, the WG is concerned that mRNA measurement was not performed in order to be able to have confidence that the increased metabolism represents transcriptional induction.

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.

Although the WG agrees with the use of a factor of 2 to identify efficacious inducers as a general rule, there was some concern regarding the potential impact of this choice on the sensitivity and specificity of the assay (c.f. section 7.2).

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?

The generated SOPs (provided not as originals, without signatures and date) describing the method as a whole (including activities to be performed in different laboratories) are sufficiently detailed. However, they cannot be used as SOPs as intended in a GLP environment. They are not user-friendly: a document of more than 80 pages does not help consultation for the operators and it is very likely that other working documents were generated (indicated also in the study plan as 'home documents') and used in the daily laboratory work. Therefore to name them as SOPs could generate confusion, which is not compliant with a quality system. The system of drafting and managing SOPs could be improved. In addition the SOPs failed in describing a harmonized format for data reporting (see section 3.2).

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 tested during the study 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?

The appropriateness of the sample size has been discussed in section 1.3 (study design). The sample size is too small for a confirmative analysis of a dichotomous predictive capacity. However practical constraints also need to be considered: A much larger study with human cells would not be practical due to lack of clinical data for comparisons; it would have to be performed in rats (c.f. section 13.4).

6.2 Representativeness of the test items with respect to applicability

NOTE: Analysis of how well the test items were selected in order to gain – through empirical testing during the study – insight into the applicability domain / limitations of the test method OR analysis to which extent the test items used during the study map an applicability domain already known.

It is appreciated that the main focus of this study was on pharmaceuticals. However, an expansion to environmental chemicals, which could be evaluated in animals (e.g. rats), would be valuable for validating a test method for informing animal mode of action. The WG suggests that there may be value in considering another study using in vitro rat hepatocytes to study CYP induction for a number of CYPs (e.g., 1A2, 2B, 2E, 3A, 4A) , and comparing these data to in vivo (reference) data from animal studies on general chemicals with a wider range of properties (lipophilicity, water solubility, vapour pressure, etc.). The combination of rat and human cell assays may provide a richer source of information on a greater range of chemicals than a human cell assay alone.

7. Within-laboratory reproducibility (Module 2)

7.1 Assessment of repeatability and reproducibility in the same laboratory

NOTE: How were repeatability and reproducibility assessed? Are the conclusions justified by the data as evaluated?

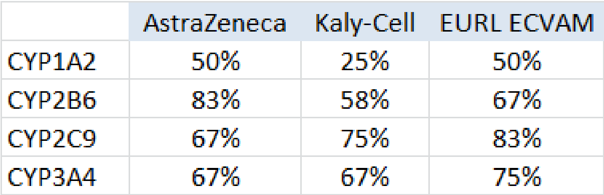
In the validation study, within-laboratory reproducibility (WLR) had been analysed by counting the frequency of greater two-fold induction obtained for all **tested batches** in a given laboratory and for a given CYP isoform. WLR analysis thus focused on Between-Batch Reproducibility (BBR) in each participating laboratory. Moreover, for BBR analysis, induction values measured for all **six different exposure concentrations** were taken into account. As outlined in the statistical report (VSR Appendix XIII, page 9): *"BBR-lab is represented by frequency of n-fold induction being > 2. Frequency is taken over three batches (for a given laboratory, concentration and enzyme). Max Frequency is 3, i.e. for all three batches n-fold induction > 2. Min Frequency is 0, i.e. for all three batches is n-fold induction ≤ 2."*

After discussion between the ESAC working group and the ECVAM scientists involved in the study, the WLR was re-evaluated using a revised definition for a positive finding of a batch/laboratory: **at least one observation of a greater than two-fold induction at any of the six concentrations used.**

Using this prediction model of greater two-fold induction as a cut-off, observed at any concentration, WLR calculations were obtained by analysing the **concordance of dichotomous predictions obtained** (1=inducer; 0=non-inducer) between the three batches used in each laboratory. This was done separately for all four CYP isoforms tested. The WG is comfortable with the results of this re-evaluation shown in tables a) and b) below and summarised for the two test systems in table c).

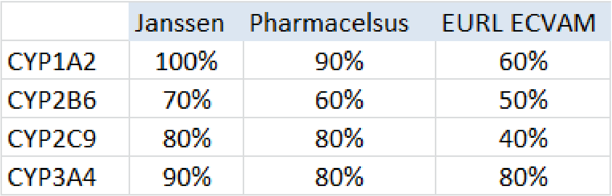
Tables a) b): Re-evaluation of WLR in the participating laboratories based on concordance of predictions obtained for three experiments (runs) based on three cell batches: a) cryopreserved hepatocytes; b) HepaRGs. Each single percentage value represents the concordance of predictions between three batches obtained in each laboratory and based on twelve chemicals. The observation of one single occurrence of a greater two-fold induction at any of the six concentrations measured was considered sufficient for concluding on "positive".

Table a) WLR for cryopreserved hepatocytes.



	AstraZeneca	Kaly-Cell	EURL ECVAM
CYP1A2	50%	25%	50%
CYP2B6	83%	58%	67%
CYP2C9	67%	75%	83%
CYP3A4	67%	67%	75%

Table b) WLR for HepaRGs



	Janssen	Pharmacelsus	EURL ECVAM
CYP1A2	100%	90%	60%
CYP2B6	70%	60%	50%
CYP2C9	80%	80%	40%
CYP3A4	90%	80%	80%

Table c) Ranges of observed WLR (% of predictions between batches within each laboratory) for the two test systems studied.

	cryoHep	**cryoHepaRG**
CYP1A2	**25-50%**	**60-100%**
CYP2B6	**58-83%**	**50-70%**
CYP2C9	**66-83%**	**40-80%**
CYP3A4	**66-75%**	**80-90%**

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

NOTE: How was within-laboratory reproducibility assessed? Are the conclusions justified by the data as evaluated?

The WG considers the WLR to be acceptable for all CYPs. The low reproducibility for CYP1A2 with hepatocytes is unsurprising given its high variation in expression across individuals and the use of a 2-fold cutoff to define induction. Use of a higher fold cutoff (e.g., 5-fold) would decrease sensitivity to background noise and probably increase the reproducibility for this enzyme.

8. Transferability (Module 3)

8.1 Quality of design and analysis of the transfer phase

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

The WG believes that the transfer phase was conducted in a reasonable fashion.

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?

There is good evidence from the WLR results that the transferability of this test method to naïve laboratories is acceptable.

9. Between-laboratory reproducibility (Module 4)

9.1 Assessment of reproducibility in different laboratories

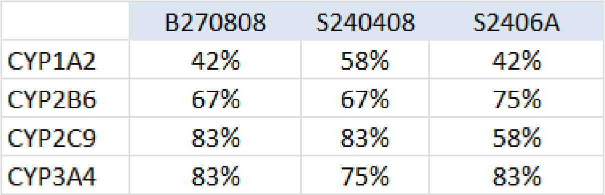
NOTE: How was reproducibility between laboratories assessed?

Similar to the approach chosen for WLR assessment (c.f. section 7. of this report), BLR assessment was done by counting the frequency of n-fold induction being > 2. As outlined by the Statistical Report (VSR Appendix XIII, page 9): *"Frequency is taken over three labs (for a given batch, concentration and enzyme). Max Frequency is 3, i.e. in all three labs n-fold induction > 2. Min Frequency is 0, i.e. in all three labs is n-fold induction ≤ 2."*

After discussion between the ESAC working group and the ECVAM scientists, the BLR was re-evaluated using the revised definition for a positive finding (see section on WLR): **at least one observation of a greater than two-fold induction at any of the six concentrations used** for a given batch. The concordance of predictions **obtained for a given batch** was then compared between laboratories to arrive at a measure of BLR. Thus **separate BLR** values are obtained for the three batches used. The batches were, in case of the cryopreserved hepatocytes, B270808, S240408, S2406A and, for HepaRGs, 16020, 16035,16036. The values are shown in tables e) and f). The data are further summarised in table g) by showing the ranges of BLR obtained for the two test systems.

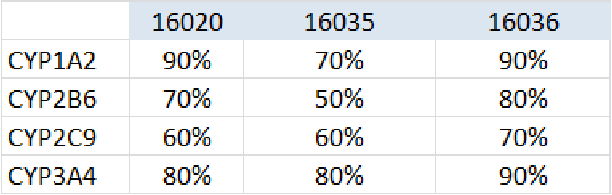
Tables c) d): Re-evaluation of BLR in the participating laboratories based on concordance of predictions obtained for one particular batch across the three laboratories and for twelve chemicals.

Table c) BLR values for the three batches of cryopreserved hepatocytes used



	B270808	S240408	S2406A
CYP1A2	42%	58%	42%
CYP2B6	67%	67%	75%
CYP2C9	83%	83%	58%
CYP3A4	83%	75%	83%

Table d) BLR values for the three batches of HepaRGs used



	16020	16035	16036
CYP1A2	90%	70%	90%
CYP2B6	70%	50%	80%
CYP2C9	60%	60%	70%
CYP3A4	80%	80%	90%

Table e) Ranges of BLR values obtained for the two test systems and four CYP isoforms.

	Human in vitro test system	
CYP enzyme isoform	**cryoHep**	**cryoHepaRG**
CYP1A2	**42-58%**	**70-90%**
CYP2B6	**67-75%**	**50-80%**
CYP2C9	**58-83%**	**60-70%**
CYP3A4	**75-83%**	**80-90%**

9.2 Conclusion on reproducibility as assessed by the study

NOTE: Are the conclusions justified by the data generated?

The WG considers the BLR to be acceptable for all CYPs, given inherent variability in the functionality of hepatocyte cultures. Although the reproducibility is lower for the cryohepatocytes, a reproducibility of about 60% or above is considered reasonable for primary cells from multiple individuals.

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?

Table 5.1 contains the predictions from the in vitro assays. A table with the in vivo isoform induction for each of the chemicals would have been useful. The WG does not feel the tripartite analysis (strong, weak, and non-inducers) is appropriate because the compounds used as positive controls do not provide an adequate basis for such an evaluation.

General Observations

The analysis shown in Tables 5M.1-6 is for the most part appropriate. For the small number of compounds considered, it demonstrates reasonably good predictive capacity. Due to the small number of compounds, however, the observed sensitivities and specificities are associated with large variation. A sensitivity, e.g., of 5/5 (Table M5.1) corresponds to a point estimate of 100%, but an exact 95%-Confidence interval ranging from 48% to 100%. This shows that proof of good predictive capacity can only be achieved by a strong increase in number of tested compounds.

The WG would recommend, however, not to use the comparison with in vivo Cmax to change the in vitro call (e.g., for omeprazole and artemisinin) since in the application of this test, these data will not be available on the test chemicals unless they have been clinically studied. The assay as described in the VSR is intended to identify chemicals with the potential to cause induction and is not described as a predictor of induction at environmentally relevant concentrations.

Incorporation of predicted AUC changes from activity and mRNA concentration-response data would improve our ability to assess the predictivity of observed results for human or rodent models.

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 test method is biologically relevant for the endpoint described in the report: induction of CYP activity for CYPs that may be associated with receptor-activated pathways. This information could be integrated into a test strategy that could help to assign a chemical to a particular AOP. The use of a minimum 2-fold increase in metabolite production as the definition of induction is consistent with current practice, but a more robust definition that encompasses both potency and efficacy would have greater biological relevance.

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)?

The applicability domain is uncertain since all the test chemicals were pharmaceuticals with similar properties (low volatility/metabolism/lipophilicity, high bioavailability/solubility). The application of this test to environmental chemicals with a much wider range of physical-chemical and other properties remains to be established, including, importantly, persistent and bioaccumulative substances for which CYP induction data may be useful but which may be challenging to test in *in vitro* systems. Ideally, additional receptor-related CYPs would also be included, such as CYP4A, and several other CYP2 isozymes such as 2E1 and 2C19 (c.f. section 13.4).

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 document describes the fact that the available reference chemical data is limited to pharmaceuticals but appears over-optimistic concerning the potential for broad application to environmental compounds without additional studies (c.f. section 13.4).

12. Performance standards (Module 7)

12.1 Adequacy of the proposed Essential Test Method Components

NOTE: Are the proposed Essential Test Method Components adequate with respect to the key elements of the validated method as evidenced by existing information and testing data generated during the study?

With the exception of the lack of mRNA assays, the essential test method components are adequate. However, apart from ECVAM the study laboratories assessed solubility visually, which is not a reliable approach.

12.2 Adequacy of the Reference Chemicals

NOTE: Are the Reference Chemicals adequately mapping the accuracy values of the validated method? Do they provide a representative range of the applicability domain of the test substances used during validation? Do they map an appropriate range of toxicity effects of the particular health endpoint in question? Are they commercially available?

Reference chemicals are well chosen, but need other chemicals, with particular emphasis on representative chemicals from different classes with non-druglike properties (e.g., persistent / highly lipophilic compounds, rapidly metabolized compounds, poorly soluble compounds). A broader diversity of chemical properties would increase the level of understanding of the applicability domain for this assay . The WG recognizes that finding human in vivo data for such a wide variety of chemicals would be problematic and therefore suggests consideration of a rat hepatocyte CYP induction assay to further explore the applicability domain (c.f. section 13.4).

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?

The potential applications of this test for regulatory purposes remain uncertain. As risk assessment moves toward greater use of quantitative kinetic data, this test will be of increasing value. Before it could be used, however, additional studies are needed to characterize the applicability domain, as discussed in section 13.4.

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).

This test should eventually be useful as part of an integrated testing strategy to assign a chemical to a receptor mediated AOP. However, additional studies are needed to characterize the applicability domain, as discussed elsewhere. The BLR and WLR results in this study could serve as a reference point for defining performance criteria for routine assessment, i.e., for developing a Performance-Based Test Guideline (PBTG) for assays based on measurement of metabolism by CYPs 1A2, 2B6, 2C9, and 3A4.

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)?

For the use of primary human hepatocytes, the key issues are donor variability (which is both an opportunity and a challenge), technical challenges associated with culturing, and finite lot sizes. Pooling of multiple donors for hepatocyte cultures is not yet practicable. For the use of HepaRG cells, the key issues are high cost, reduced metabolic capacity, and heterogeneity of the co-culture that varies from lot to lot. If the test were to be expanded to other receptor pathways, the HepaRGs may lack relevance. In addition, the limited enzyme activity in HepaRGs hampers their use in parallel for clearance/metabolite ID. Acceptability criteria for cell lots and for assay performance will be needed. Hepatocellular phenotype is strongly dependent on culture

conditions. Another limitation is the need to use LC-MS, which is an expensive and not a generally available technology.

13.4 Gap analysis

NOTE: Identify, if appropriate, gaps in the study design and/or execution that impact on the stated study objective or the conclusions drawn.

One of the primary issues highlighted by the ESAC WG was the need to determine the broader applicability of these induction models and assay systems to environmental chemicals. However, this issue would be difficult to assess with the relatively small number of chemicals that could be examined using human data. Part of the rationale for the selection of a relatively small set of reference compounds was the limited number of compounds, mostly pharmaceuticals, for which there are in vivo data on induction of human xenobiotic metabolizing enzymes. To more broadly evaluate the application of these human assays for environmental chemicals could be particularly challenging since human in vivo induction data is unlikely to be available for non-pharmaceuticals. The WG suggests that a possible path forward for broadening the applicability domain of this type of approach to environmental chemicals would be to conduct a similar study using primary rat hepatocyte culture models or rat hepatocyte cell lines, and to include induction assays for key enzymes reflective of AhR (e.g. CYP1A2), CAR (e.g. CYP2B1/2B2), PXR (e.g. CYP3A1/3A23), while possibly adding PPARα (CYP4A1), which is known to be linked to rodent carcinogenicity, as well as CYP2E1 (although it is not related to activation of a pathway). This would allow a wider range of environmental chemicals to be studied in concentration response and afford the opportunity to make use of rodent in vivo induction data for evaluation of predictions. While these results would be less directly translatable to human health, they could avoid the practical impossibility of intentionally exposing humans to higher concentrations of environmental chemicals to assess induction potential. Such a rat study would support assessment of limitations of the applicability domain associated with chemical properties. In addition, the rodent hepatocyte induction assay could provide useful information for interpreting toxicity test results from in vitro or in vivo rodent studies and could contribute to a mechanistic understanding of the toxicity observed towards the elucidation of an AOP. The human hepatocyte induction assay could then be used in parallel for evaluation of human relevance.

General Observations

Addition of mRNA measurement would greatly improve the interpretability of the assay. The choice of Alamar Blue as a cytotoxicity assay is questionable, as is the choice of an IC30 as a cutoff for toxicity. The WG suggests future consideration of other biomarkers (ALT, AST, LDH) and a cutoff of no more than the IC10. For an adequate validation, more than 3 hepatocyte donor preparations (e.g., 10) would be preferred. Ideally, a larger number of test items (e.g., n=20 positives and n=20 negatives) would be needed to demonstrate the robustness of the test (Fahmi et al, DMD 2010, v38(9):1605-1611).

14. Other considerations

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

The WG feels that a test for CYP induction may be of only limited value, especially when considering the intended purpose of the test method as suggested in the VSR (i.e. use of in vitro CYP induction test within integrated approaches for biotransformation and toxicological Mode of Action studies of substances; VSR page 42, section 4.1). Induction as a marker for receptor activation is likely to contribute to the weight of evidence only for a small fraction of possible modes of action or AOPs.

There is little evidence that induction of metabolism could affect the toxicity of other compounds at environmental exposure levels. The more likely interaction at these low concentrations is inhibitory. The exception to this would be persistent, bioaccumulative compounds, which could in principal accumulate to inductive concentrations. However, the current study does not provide any evidence that the proposed test method is applicable to such compounds, which can have very challenging properties for in vitro testing.

The WG strongly encourages ECVAM to continue to conduct studies with human hepatic models to develop methods for characterization of other kinetic data, including clearance, metabolic profiling, and inhibition. The importance of developing in vitro to in vivo extrapolation methods cannot be overemphasized. The study reported in the VSR specifically addresses the availability, transferability and reliability of metabolically competent hepatocellular test systems for in vitro testing, using induction of metabolism as a case study. As such, the reproducibility in this study could be a useful indicator of the potential for other uses. The current study design could be adapted for these purposes while concurrently addressing the primary objectives of this study, for example induction and mode of action, as discussed above in section 13.4.

15. Conclusions on the study

NOTE: This section should present a brief overview of the study results and conclusions as described in the VSR (section 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 ESAC WG summary of the results and conclusions of the study

The VSR concludes that, with the exception of batch dependence of results in the human cryoHep method, the study findings satisfy the requirements for test definition, within laboratory reproducibility, transferability, and between laboratory reproducibility) and contribute to assessment of predictive capacity. The VSR concludes that the information generated in the study shows that the human in vitro CYP induction method is robust, reliable and relevant. Therefore, the VSR supports the use of the human in vitro CYP induction method in a weight ‐ of ‐ evidence approach to support regulatory decision making.

The VSR cautions that the CYP induction method relies on a complex experimental setup and thus requires a skilled and analytically well ‐ resourced biochemical and cell biological laboratory. Frequent occurrences of irregularities in concentration response curves and uncertainties in their interpretations suggested that there are a number of critical points to be taken into consideration in the design and execution of the experiments, such as the selection of concentration range and delineation of solubility limit and potential cytotoxicity range of an unknown compound.

The VSR also suggests that the CYP induction method deserves further evaluation as part of an integrated testing strategy for the role it might play in the determination of xenobiotic exposure and potency predictions and its role in alternatives for systemic toxicity hazard identification. The CYP induction in vitro method can be considered as a candidate regulatory in vitro test method to gain insight in the toxicological MoA of substances in the context of the new safety assessment paradigm using exclusively in vitro approaches based on human cells and tissues in combination with the appropriate in silico approaches and overall systems biology knowledge.

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

The WG agrees with the VSR assessment that the study findings satisfy the requirements for test definition, within laboratory reproducibility, transferability, and between laboratory reproducibility, but only partially satisfy the requirements for assessment of predictive capacity. However, the WG believes that the VSR overstates the readiness of the test for regulatory use. In particular the CYP induction assay cannot be referred to as robust until additional studies are conducted with rodent hepatocytes to further investigate the applicability and predictive power of the assay for chemicals that are not pharmaceuticals, as discussed above.

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

The WG feels the current study is a good reflection of expectations based on other existing information regarding hepatocyte assays, and provides evidence that reliable hepatocyte assays for other important purposes, including identification of metabolites and quantification of metabolic clearance, are feasible.

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 use of human cells and resulting focus on human in vivo reference data, while enhancing human relevance of the method, limits the value of this assay due to the narrow range of physico-chemical properties associated with drugs (high water solubility and bioavailability, low metabolism and lipophilicity). To fully assess the applicability domain for a CYP induction assay would require more chemicals, with particular emphasis on representative chemicals from different classes with non-druglike properties (e.g., persistent / highly lipophilic compounds, rapidly metabolized compounds, poorly soluble compounds); such a study would probably have to be conducted with rodent hepatocytes. The WG suggests that there may be value in considering another study using in vitro rat hepatocytes to study CYP induction for a number of CYPs (e.g., 1A2, 2B, 2E, 3A, 4A) and comparing these data to in vivo (reference) data from animal studies on general chemicals with a wider range of properties (lipophilicity, water solubility, vapour pressure, etc.). The combination of rat and human cell assays may provide a richer source of information on a greater range of chemicals than a human cell assay alone.

The WG believes that CYP induction assays assessed in rat hepatocytes could also provide pertinent mode of action information regarding adverse outcomes observed in animal testing. This would also allow for a greater variety of in vivo induction data across a wider range of receptor pathways than could be accomplished with the human hepatocyte assays. These assays should eventually be useful as part of an integrated testing strategy to assign a chemical to a receptor mediated AOP. The BLR and WLR results in this study could serve as a reference point for defining performance criteria for routine assessment, i.e., for developing a Performance-Based Test Guideline (PBTG) for assays based on measurement of metabolism by CYPs 1A2, 2B6, 2C9, and 3A4.

The WG strongly encourages ECVAM to continue to conduct studies with human hepatic models to develop methods for characterization of other kinetic data, including clearance, metabolic profiling, and inhibition. The importance of developing in vitro to in vivo extrapolation methods cannot be overemphasized.

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

a) mRNA should be considered as a parallel endpoint together with activity and cytotoxicity to assess the induction potential of the compounds, in order to address the potential for inhibition, suppression, or toxicity that may confound induction assessments based on enzymatic activity. In addition, mRNA is more proximal to receptor activation and is required for drug submissions by regulatory agencies.

b) While the use of hepatocytes from three individual donors is consistent with current practice, it would be preferable (when it becomes practicable) to have a larger number of individual donor preparations or develop approaches that use pooled cells from a larger number (e.g., 10) of donors.

c) The concentration of phenobarbital (that gives a 2-fold induction) was insufficient to saturate induction (to get an accurate EC50) of tests. The WG is not comfortable that the lot selection and test acceptance criteria were adequate, and feels that a greater fold-change for positive controls may have been preferable. Also, acceptance should be based on both fold-change in positive controls and metabolite generation in controls for cocktail exposures.

d) The Alamar Blue assay is primarily for mitochondrial activity, which can be perturbed by inducers without toxicity. Other biomarkers (LDH leakage, AST, ALT) would provide better markers for toxicity in hepatic models.

e) The choice of IC30 as the cutoff for toxicity is not considered appropriate by the WG. It is probably responsible for the non-monotonic dose response behaviors observed. The WG would suggest no more than an IC10 as point of departure.

f) The WG recommends that ECVAM consider re-analysing the activity data without normalization for protein content (only normalize on plated cell number) as this can add additional variability due in part to dying cells and the use of an additional measure with additional sources of variation.

g) In future studies, use of a higher fold cutoff (e.g., 5-fold) would decrease sensitivity to background noise and probably increase the reproducibility of the assay, especially for some CYPs with highly variable expression levels (e.g. CYP1A2).

h) The WG recommends that in future evaluations the comparison with in vivo Cmax not be used to change the in vitro call (e.g., for omeprazole and artemisinin in the current study) since in the application of this test, these data will not be available on the test chemicals unless they have been studied clinically. The assay as described in the VSR is intended to identify chemicals with the potential to cause induction and is not described as a predictor of induction at environmentally relevant concentrations.

i) The WG suggests in future to use ANOVA with post-hoc Dunnett test for multiple comparison to vehicle control and to exclude non-significant findings. WLR and BLR should be evaluated on the call for a single curve (at least 1 concentration with >2 and Dunnett test significant), not on the individual concentrations. Exact 95%-confidence intervals for sensitivity and specificity should be provided.

j) The acceptance criteria for determining the specific studies to be used to characterize the in vivo induction are critical to assure that the points of comparison with the in vitro assay are correct, so in future these criteria should be documented.

17. References