



Is a combination of assays really needed for non-animal prediction of skin sensitization potential? Performance of the GARD™ (Genomic Allergen Rapid Detection) assay in comparison with OECD guideline assays alone and in combination

David W. Roberts

School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, L3 3AF, United Kingdom

ARTICLE INFO

Keywords:

Skin sensitization potential
Non-animal assays
Integrated testing strategies
GARD™

ABSTRACT

To meet regulatory requirements, and avoid or minimize animal testing, there is a need for non-animal methods to assess the potential of chemicals to cause skin sensitization. It is widely assumed that no one test will be sufficient and that combined data from several assays spanning key events from the adverse outcome pathway will be required. This paper challenges that assumption. The predictive performance of a single assay, the Genomic Allergen Rapid Detection (GARD™) assay, was compared with the performance, singly and in combination, of three formally validated non-animal approaches that appear as OECD test guidelines: the direct peptide reactivity assay (DPRA), the ARE-Nrf2 luciferase test method, and the human cell line activation test (h-CLAT).

It is shown here that GARD™ alone outperforms each of DPRA, ARE-Nrf2 luciferase or h-CLAT, alone or in any combination as a 2 out of 3 strategy, in terms of sensitivity, specificity and accuracy.

Based on the datasets analysed here, the sensitivity and specificity of GARD™ alone are 90–92% and 79–84% (“2 out of 3”, 86% and 76%). Thus, in any situation where the 2 out of 3 strategy is considered adequate, GARD™ alone could be used with equal or better performance.

1. Introduction

To meet regulatory requirements, particularly in Europe, and to avoid or minimize animal testing, there is a need for non-animal methods to assess the potential of chemicals to cause skin sensitization (EC, 2009). It is widely assumed by the scientific community in this area that no single test is ever going to be sufficient, and that a combination of data from several assays that span key events (KEs) from the skin sensitization adverse outcome pathway (AOP) will be needed. Much work has been done on developing test methods, and in evaluating their validity for use (Adler et al., 2011; Goebel et al., 2012; Reisinger et al., 2015; Zuang et al., 2015). The AOP for skin sensitization has been used as a framework to put into context the KE information being generated by any of the test methods (OECD, 2012; Reisinger et al., 2015). Of the many non-animal methods in development, several approaches have undergone formal validation and appear as OECD test guidelines. Three in particular have been extensively investigated in combination as an integrated approach for prediction of skin sensitization potential (Urbisch et al., 2015, 2016). These are the

direct peptide reactivity assay (DPRA) (OECD TG442C), the ARE-Nrf2 luciferase method (OECD TG442D), and the human cell line activation test (h-CLAT) (OECD TG 442E) (OECD, 2015a,b, 2016a). These three methods are usually considered as being representative of Key Event (KE)1 (haptening of skin protein), KE2 (generation of danger signals), and KE3 (activation of dendritic cells; DC), respectively. However, the extent to which these assays are genuinely representative of the KEs has been brought into question (Roberts and Patlewicz, 2018), and the argument that a combination of assays, that are not 100% successful on a stand-alone basis, must be better than a single assay has been shown not to be well founded in logic (Johansson and Gradin, 2017).

The issue can be viewed from a different perspective, as follows. The potential of a chemical to sensitize depends on its ability, when applied to the skin, to covalently modify the appropriate proteins on AND NOTHING ELSE (Roberts and Aptula, 2008). This corresponds to KE1 of the AOP. Everything that follows in the AOP occurs as a consequence of KE1, and cannot currently, nor in the foreseeable future, be measured independently of KE1. [Logically to model KE2 one would have to treat a “KE2 culture” with the modified proteins or cells carrying those proteins, from KE1].

E-mail address: d.w.roberts@ljmu.ac.uk.

<https://doi.org/10.1016/j.yrtph.2018.07.014>

Received 2 March 2018; Received in revised form 13 July 2018; Accepted 22 July 2018

Available online 23 July 2018

0273-2300/© 2018 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

In reality the outputs of all the assays, including the LLNA itself, can be interpreted as indices of the ability to covalently modify the appropriate proteins. Although there is a great deal more that needs to happen for sensitization to develop, the extent to which all of this great deal more happens is dependent on the nature of the chemical only insofar as it is able to produce the covalent modification at KE1 which triggers the cascade of cellular and biochemical events that finally lead to the sensitized state. This ability depends on the chemical's: reactivity towards the appropriate protein (s); ability to access the appropriate protein(s) - in competition with other processes not leading to sensitization; ability to be transformed in cutaneo to protein-reactive species (for chemicals not directly reactive). The precise identities of the appropriate proteins or their precise locations in the skin are not known. Here the position that a combination of assays based on the AOP is necessary is challenged by analysing previously published LLNA (murine local lymph node assay) and non-animal assay data, so as to compare the performance of a single assay, the Genomic Allergen Rapid Detection (GARD™) assay, with the three formally validated assays, alone and in combination, for the binary classification of chemicals into skin sensitizers and non-sensitizers. GARD™ is based on global transcriptomic analysis of differential gene expression in a human myeloid cell line, serving as a surrogate for dendritic cells (DCs), induced by skin sensitizing chemicals in comparison with non-sensitizing controls. The resulting biomarker signature, the GARD™ prediction signature (GPS), consists of 200 transcripts, which are used as input into a support vector machine (SVM) model trained on a set of reference chemicals (Johansson et al., 2011). The changes in transcription can be linked to the maturation and activation of DC (KE3) during skin sensitization (Johansson et al., 2014).

2. Materials and methods

2.1. Data

Data analyses were conducted using two sets of non-animal test results for chemicals which had previously been tested in the murine local lymph node assay (LLNA). Dataset 1 (DS1), was provided by SenzaGen, and consists of GARD™ results (expressed as positive or negative) for 130 chemicals with matched LLNA data (also expressed as positive or negative). GARD™ is based on changes of gene expression in a myeloid cell line on exposure to the test chemical, and its output is based on a combination, derived by machine learning, of ca. 200 genomic markers. Although there can be a risk of overfitting when designing models of this type, the GARD™ assay used cross-validation during model establishment and subsequent validation using a true external test dataset after model establishment, and has been shown to be highly functional in a broad chemical space outside of the training data (Johansson et al., 2014, 2017). The chemicals in DS1, with their GARD™ and LLNA results, are listed in Table 1. Sources of the LLNA classifications (skin sensitizer or non-sensitizer) are indicated in Table 1: most were taken from the compilations by Gerberick et al. (2005), Kern et al. (2010) and Haneke et al. (2001). GARD™ predictions were provided by SenzaGen and have mostly been published by Forreryd et al. (2016) and Johannsson et al. (2017). Dataset 2 (DS2) was that used by Roberts and Patlewicz (2018), originally compiled by Asturiol et al. (2016), and consists of DPRA, KeratinoSens™ (ARE-Nrf2 luciferase), and h-CLAT results for 269 chemicals with LLNA results. (All results in DS2 were expressed as positive or negative). Not all of the chemicals in DS2 had been tested in all of the three non-animal assays. Five chemicals (coumarin, benzalkonium chloride, benzocaine, citronellol and amyl cinnamic alcohol) are listed as LLNA positive in DS1, but as LLNA negative in DS2. For that reason these were excluded from the DS2 analysis.

2.2. Data analyses

First the two datasets were compared manually to identify the chemicals in DS1 (GARD™ results) also present in DS2 (DPRA, ARE-Nrf2

Table 1
GARD™ and LLNA data (Dataset 1). S = predicted sensitizer; NS = predicted non-sensitizer.

Sensitizers in the LLNA	CAS number	LLNA source	Prediction GARD™
1,2-Benzisothiazolin-3-one	2634-33-5	G	S
2,4-Dinitrochlorobenzene	97-00-7	G	S
2,4-Dinitrofluorobenzene	70-34-8	H	S
2-Aminophenol	95-55-6	G	S
2-Ethylhexyl acrylate	103-11-7	K	S
2-Hydroxyethyl acrylate	818-61-1	G	S
2-Mercaptobenzothiazole	149-30-4	G	S
2-Methyl-4isothiazolin-3-one	2682-20-4	G	S
2-Nitro-1,4-phenylenediamine	5307-14-2	G	S
3-Dimethylaminopropylamine	109-55-7	G	S
3-Methylcatechol	488-17-5	H	S
4-Methylaminophenol sulphate (Metol)	55-55-0	G	S
4-Nitrobenzyl bromide	100-11-8	G	S
5-Methyl-2,3-hexanedione	13706-86-0	G	S
7-Hydroxycitronellal	107-75-5	G	S
Abietic acid	514-10-3	G	S
Amyl cinnamic aldehyde	122-40-7	G	S
Amylcinnamyl alcohol	101-85-9	K	S
Anethole	104-46-1	G	S
Aniline	62-53-3	G	S
Anisyl alcohol	105-13-5	K	S
Benzalkonium chloride	63449-41-2	H	S
Benzocaine	94-09-7	G	S
Benzoyl peroxide	94-36-0	H	NS
Benzyl benzoate	120-51-4	G	S
Benzyl salicylate	118-58-1	K	S
Bisphenol A glycerolate dimethacrylate	1565-94-2	P	S
Bisphenol A-diglycidyl ether	1675-54-3	G	S
Butyl acrylate	141-32-2	K	S
Butyl glycidyl ether	2426-08-6	G	S
CAPE	104594-70-9	J1	S
Carvone	99-49-0	K	S
Chlorpromazine	50-53-3	H	S
Cinnamaldehyde	104-55-2	G	S
Cinnamic alcohol	104-54-1	G	S
Citral	5392-40-5	G	S
Citronellol	106-22-9	R	S
Coumarin	91-64-5	G	S
Diethanolamine	111-42-2	J2	S
Diethyl maleate	141-05-9	G	S
Diethylthiocarbamyl benzothiazole sulfide	95-30-7	F	S
Dimethylthiocarbamyl benzothiazole sulfide	3432-25-5	F	S
Dimethyl fumarate	624-49-7	Ho	S
Diphenylcyclopropanone	886-38-4	J1	S
Ethyl acrylate	140-88-5	G	S
Tocopherol	59-02-9	J2	S
Ethylene diamine	107-15-3	G	S
Ethylene glycol dimethacrylate	97-90-5	G	S
Eugenol	97-53-0	G	S
Farnesol	4602-84-0	G	S
Formaldehyde	50-00-0	G	S
Geraniol	106-24-1	G	S
Glutaraldehyde	111-30-8	G	S
Glyceryl monothioglycolate	30618-84-9	J2	S
Glyoxal	107-22-2	G	S
Hexyl salicylate	6259-76-3	K	NS
Hexylcinnamic aldehyde	101-86-0	G	S
Hydroquinone	123-31-9	G	S
Imidazolidinyl urea	39236-46-9	G	S
Iodopropynyl butylcarbamate	55406-53-6	K	S
Isoeugenol	97-54-1	G	S
Kanamycin	8063-7-8	G	S
Kathon CG (MC/MCI)	26172-55-4 & 2682-20-4	F	S
Lauryl gallate	1166-52-5	G	NS
Lillial	80-54-6	J2	S
Limonene (not oxidised)	5989-27-5	G	S
Linalool	78-70-6	G	S
Lylal	31906-04-4	J2	S
Methyl acrylate	96-33-3	K	S
Methyl heptene carbonate	111-12-6	K	S
Methyldibromoglutaronitrile	35691-65-7	G	S

(continued on next page)

Table 1 (continued)

Sensitizers in the LLNA	CAS number	LLNA source	Prediction GARD™
Methylmethacrylate	80-62-6	K	NS
Neomycin sulphate	1404-04-2	H	S
Penicillin G	61-33-6	G	NS
Pentachlorophenol	87-86-5	G	S
Phenyl benzoate	93-99-2	G	S
Phenylacetaldehyde	122-78-1	G	S
PHMG-HCl	57028-96-3	J1	S
Pinanediol	22422-34-0	J1	S
Potassium dichromate	7778-50-9	G	S
PPG-6-decyltetradeceth-30	72484-69-6	J1	S
p-Phenylenediamine	106-50-3	G	S
Propyl gallate	121-79-9	H	S
Pyridine	110-86-1	G	S
Resorcinol	108-46-3	G	S
Taxifolin	480-18-2	J1	S
Tetrachlorosalicylanilide	1154-59-2	G	S
Tetramethylthiuram disulfide	137-26-8	G	S
Thioglycerol	96-27-5	H	NS
Toluene diamine sulphate	615-50-9	G	S

Non-Sensitizers in the LLNA	CAS number	Prediction GARD
1-Bromobutane	109-65-9	G S
1-Butanol	71-36-3	G NS
4-Aminobenzoic acid	150-13-0	H NS
Alpha-glucosylrutin	130603-71-3	J1 NS
Benzaldehyde	100-52-7	G NS
Benzoic acid	65-85-0	Z NS
Chlorobenzene	108-90-7	G NS
Citric acid	77-92-9	Z NS
Dextran	9004-54-0	H NS
Diethyl phthalate	84-66-2	G S
Diethyl toluamide	94271-03-1	J2 S
Dimethyl CAPE	14551-14-0	J1 NS
Dimethyl formamide	68-12-2	F2 NS
Dimethyl isophthalate	1459-93-4	H NS
Dimethyl sulfoxide	67-68-5	G NS
Ethyl vanillin	121-32-4	G NS
Glycerol	56-81-5	G NS
Hexane	110-54-3	G NS
Hydrocortisone	50-23-7	H S
Isopropanol	67-63-0	G NS
Lactic acid	50-21-5	G NS
Menthol	89-78-1	J1 NS
Methyl salicylate	119-36-8	G NS
Octanoic acid	124-07-2	G NS
Phenol	108-95-2	H NS
Phenoxyethanol	122-99-6	J2 NS
p-Hydroxybenzoic acid	99-96-7	G NS
Potassium permanganate	7722-64-7	G NS
Propyl paraben	94-13-3	G S
Propylene glycol	57-55-6	G NS
Sacharin	81-07-2	G NS
Salicylic acid	69-72-7	G NS
Sodium dodecyl sulphate	151-21-3	G S
Triethanolamine	102-71-6	J2 S
Tween 80	9005-65-6	H S
Vanillin	121-33-5	G NS
Xylene	1330-20-7	H NS
Zinc sulphate	7733-02-0	H NS
4-Chloroaniline	106-47-8	H NS

G = Gerberick et al., 2005.

K = Kern et al., 2010.

H = Haneke et al., 2001.

Ho = Hoffmann et al., 2018.

J1 = Johansson et al., 2014.

J2 = Johansson et al., 2017.

F = Forreryd et al., 2016.

Z = Zeller et al., 2017.

P = Proprietary information supplied by SenzaGen.

luciferase, and h-CLAT results). The GARD™ results for these chemicals (94 in total) were added to DS2. GARD™ performance in isolation was then assessed both for all chemicals in DS1, and for the subset of chemicals with GARD™ results in the augmented DS2. The performance indices calculated for each assay, and for the combination strategy, are as follows:

Sensitivity - % of skin sensitizers predicted positive

Specificity - % of non-sensitizers predicted negative

Accuracy - % of predictions correct

Probability of positive prediction being correct (PPV) - % of predicted positives that are true

Probability of negative prediction being correct (NPV) - % of predicted negatives that are true

3. Results

Table 2 compares the performance of GARD™ alone with the best performing of DPRA, ARE-Nrf2 luciferase, and h-CLAT, for each of the performance criteria.

On the basis of these data, GARD™ is the best performing of the four single assays by all criteria.

Table 3 compares the performance of GARD™ alone against the performance of the binary combination of DPRA + h-CLAT, which was found previously by Roberts and Patlewicz (2018) to be the best performing binary combination from among DPRA, ARE-Nrf2 luciferase, and h-CLAT, and better in terms of sensitivity than the “2 out of 3” strategy based on “majority voting” by DPRA, ARE-Nrf2 luciferase, and h-CLAT. For the binary combination, the rule is that positive in either or both assays constitutes a ‘Skin Sensitizer’ prediction and negative in both constitutes a ‘Non-sensitizer’ prediction. Table 3 also shows the performance of the “2 out of 3” strategy. Also shown are performance figures reported by Urbisch et al. (2015) for the “2 out of 3” strategy applied to LLNA data in their dataset, which substantially overlaps but is not identical to the Asturiol et al. (2016) dataset.

The performance figures in Tables 2 and 3 are based on using the maximum number of chemicals to assess each assay or combination, and consequently are not based on identical datasets for each assay or assay combination. For example not all of the chemicals with GARD™ data in DS2 have sufficient data for the 2 out of 3 approach to be assessed, and not all those for which the 2 out of 3 approach can be assessed have GARD™ data. The justification for making comparisons of the different approaches based on different (although substantially overlapping) sets of chemicals is as follows. An implicit assumption underlying any assay or combination of assays approved for use in predicting any biological activity is that the set of chemicals it has been evaluated on, for which the biological data already exist, is representative of the wider universe of chemicals for which biological activity data are not available. Thus, if an approach gives 80% accuracy for the chemicals in one dataset, the expectation is that for a new chemical the prediction of that approach is about 80% likely to be correct. And for 200 or so new chemicals, the performance statistics should be very similar.

As a further check, comparisons were also made on the basis of

Table 2

Performance of single assays.

Performance index (%)	GARD™, 94 cpds in DS2	GARD™, 129 cpds in DS1	Best performing assay, other than GARD™, in DS2	
Sensitivity	90	93	DPRA	82
Specificity	84	79	DPRA	75
Accuracy	88	89	DPRA	80
PPV	94	91	DPRA	90
NPV	75	84	ARE-Nrf2	70

Table 3

Comparison of GARD™ alone with the binary DPRA + h-CLAT combination and with the 2 out of 3 strategy (based on DPRA, ARE-Nrf2, h-CLAT).

Performance index (%)	GARD™, 94 cpds in DS2	GARD™, 129 cpds in DS1	DPRA + h-CLAT in DS2	2 out of 3 (DPRA, ARE-Nrf2, h-CLAT) in DS2 ^a
Sensitivity	90	93	97	86 (82)
Specificity	84	79	52	76 (72)
Accuracy	88	89	86	83 (79)
PPV	94	91	86	89
NPV	75	84	84	69

^a Figures in brackets are the Sensitivity, Specificity and Accuracy values reported by Urbisch et al. (2015) for performance of the 2 out of 3 strategy against LLNA data in their dataset.

identical sets of chemicals for GARD™ performance against DPRA + h-CLAT performance (Table 4) and against “2 out of 3” performance (Table 5), for the subsets of DS2 for which both GARD™ data and sufficient data to make either a DPRA + h-CLAT or a “2 out of 3” prediction were available. Of the 94 chemicals in DS2 with GARD™ results available, 81 also have assay data enabling the DPRA + h-CLAT combination to be assessed, and 80 have assay data enabling the 2 out of 3 approach to be assessed.

The overall accuracy of GARD™ alone is similar to (and marginally better than) that of the binary DPRA + h-CLAT combination, and its specificity is very clearly and substantially better (i.e. it identifies correctly a higher proportion of the non-sensitizers). The binary DPRA + h-CLAT has a higher sensitivity than GARD™ alone (i.e. it produces fewer false negatives) but its specificity is inferior (i.e. it produces more false positives). This applies irrespective of whether the comparison is made on the basis of the maximum number of chemicals with data analysable for each approach (Table 3) or on the basis of comparing performance for identical sets of chemicals (Table 4).

GARD™ alone performs better than “2 out of 3” on all performance criteria, and again this applies irrespective of whether the comparison is made on the basis of the maximum number of chemicals with data analysable for each approach (Table 3) or on the basis of comparing performance for identical sets of chemicals (Table 5). The differences are not large, but they are consistent.

4. Discussion

The analyses presented here are based entirely on LLNA data. Although human data are included in the Asturiol et al. dataset, the number of chemicals for which GARD data are also available are insufficient for a proper analysis, particularly since many of the chemicals given in Asturiol et al. as negative in humans are positive in the LLNA and arguably may in reality be weak sensitizers in humans. A realistic analysis of GARD™ performance against human data would require a larger number of chemicals, particularly more that are definitely non-sensitizers. One of the problems with the human data is that classification as S or NS is much more of a subjective expert judgement than

Table 4

Comparison of GARD™ and DPRA + h-CLAT for the exact same set of 81 chemicals.

Performance index (%)	GARD™	DPRA + h-CLAT
Sensitivity	91	95
Specificity	89	58
Accuracy	91	86
PPV	97	88
NPV	73	79

Table 5

Comparison of GARD™ and “2 out of 3” for the exact same set of 80 chemicals.

Performance index (%)	GARD™	2 out of 3
Sensitivity	91	88
Specificity	91	83
Accuracy	91	86
PPV	96	93
NPV	80	73

the LLNA where there is a consistent criterion. Nevertheless, the “2 out of 3” approach, for which there are sufficient non-animal data to make an assessment, performs as well, if not better, against the human data in the Urbisch et al. dataset as it does against LLNA data (Urbisch et al., 2015). It is the author's intention to follow up the present paper with a thorough discussion of the apparent false negatives and false positives, in the in vitro assays and also in human and LLNA data. This will provide a better basis to evaluate the non-animal assays against human data. In the meantime, it should be noted that recent publications have reported a strong correlation between LLNA potency and human potency, with some outliers having largely predictable special chemistry (Api et al., 2015; Basketter et al., 2018; Roberts and Api, 2018). These publications confirm the value of LLNA data, and consequently support the potential value of non-animal approaches to predict LLNA results.

GARD™ as a standalone assay performs better than any of DPRA, ARE-Nrf2-luciferase, or h-CLAT on their own, in terms of sensitivity, specificity, accuracy, probability of a positive prediction being correct and probability of a negative prediction being correct. It also performs better than the “2 out of 3” strategy in terms of all these criteria. This applies irrespective of whether the comparison is made on the basis of the maximum number of chemicals with data analysable for each approach (Table 3) or on the basis of comparing performance for identical sets of chemicals (Table 5). Thus, in any situation where the “2 out of 3” strategy is considered adequate, GARD™ could be used instead on a standalone basis. Based on the findings summarised in Table 3, the sensitivity of GARD™ is 90–92% (compared with “2 out of 3”, 86%), i.e. on the basis of the datasets analysed here, it can identify 90–92% of skin sensitizers as such. This high sensitivity is coupled with high specificity (79–84%, compared with “2 out of 3”, 76%).

At first sight it may seem surprising that a stand-alone assay has advantages over strategies based on more than one method that must be deployed in combination, particularly since it is widely considered that an integrated testing strategy (ITS) using a battery of assays representing the key events of the adverse outcome pathway will be necessary to distinguish adequately between skin sensitizers and non-sensitizers by non-animal methods (e.g. OECD, 2016b). The three OECD guideline assays DPRA, ARE-Nrf2 luciferase and h-CLAT are considered to represent KEs 1, 2 and 3, respectively, of the AOP. If these assays truly reflect these KEs, then following the logic that all three of these KEs are required for skin sensitization to develop, only chemicals that are positive in all three assays should be skin sensitizers, and a chemical that is negative in any one of these three assays should be classed as a non-sensitizer. This can easily be tested by analysis of data from DS2, whereby only chemicals testing positive in all three of the above OECD guideline assays are predicted as sensitizers and chemicals testing negative in any of these three assays are predicted not to be skin sensitizers. Results of this analysis, based on AOP-logic with the assumption that the assays truly model their assigned KEs, are shown in Table 6.

Clearly, the ITS based on AOP-logic with the three OECD guideline assays is inadequate. It gives more false negatives than true positives, i.e. less than half of the skin sensitizers are predicted as such. This failure of the ITS based on AOP-logic does not mean that the AOP is wrong. What it does indicate is that the three OECD Guideline assays do not truly or adequately reflect the KEs that they are assumed to model.

It has been pointed out that the ARE-Nrf2 luciferase approach and

Table 6

Predictive performance of ITS based on AOP-logic with OECD guideline assays (LLNA data from DS2).

Prediction, true/false	Number of chemicals so predicted
True positives	51
True negatives	76
False positives	3
False negatives	57
Performance index	%
Sensitivity	47
Specificity	96
Accuracy	68
PPV	94
NPV	57

the h-CLAT are both dependent on protein being covalently modified by the test chemical, and that consequently all three of the OECD guideline assays are in reality reactivity assays, with largely, but not completely, overlapping domains of applicability and inapplicability (Roberts and Patlewicz, 2018; Johansson and Gradin, 2017). Taking this further, based on the argument that having sufficient ability to covalently modify the relevant skin proteins *in cutaneo* is a mandatory requirement for a chemical to be a skin sensitizer (Roberts and Aptula, 2008), it follows that, as discussed in more detail by Roberts and Patlewicz (2018), non-animal test methods for skin sensitization are best considered as stand-alone assays and will be most effectively used either in isolation or in combinations that are carefully chosen to minimize shared inapplicability domains. This applies not only to the three assays mentioned above, but also to non-animal skin sensitization assays in general.

Viewed in this context, GARD™ may be considered as an assay that, via the GARD prediction signature (GPS), models the ability of chemicals to covalently modify the relevant skin proteins whose modification leads to skin sensitization. This involves not only the chemical's reactivity *per se*, but also its ability to gain access to the relevant skin proteins, and the ability of the assay to represent the activation of skin sensitizers that are not directly protein reactive. The present analyses indicate that the inapplicability domains of GARD™ are narrower than those of any of the three OECD guideline assays, and hence GARD™ outperforms all of these assays, alone and in combination.

5. Conclusions

As a new generation standalone assay, GARD™ can distinguish between skin sensitizers and non-sensitizers with greater accuracy, sensitivity and specificity than any of the OECD test guideline assays; DPRA, ARE-Nrf2 luciferase, or h-CLAT. GARD™ as a standalone assay also outperforms a “2 out of 3” strategy based on the majority prediction by DPRA, ARE-Nrf2 luciferase, and h-CLAT. It would be desirable in due course to confirm that this finding applies to an expanded range of data.

There may be scope for further improvement of GARD™ by inclusion of chemistry and SAR principles to identify ambiguous cases where supplementary *in vitro* data could be used to arrive at a more confident conclusion. However, this is beyond the scope of the present article and will be explored in a separate paper (in preparation) that will discuss the mis-predicted chemicals from GARD™, the OECD assays and their combinations, from the perspective of their chemistry (aiming at better definition of applicability domains) and their borderline/non-borderline status in the *in vivo* assays (aiming at more precise evaluation of performance of the assays and combinations). It should also be noted that the performance data employed here are based on uncritical

acceptance of all the LLNA data, on a binary positive-negative basis, as presented in the source references. In reality, some of the GARD false negatives are very weak sensitizers in the LLNA and some of the GARD™ false positives may have a weak sensitization potential. Similarly, all of the assays have defined borderline values for their outputs to separate predicted positives from predicted negatives, giving potential for misclassification of chemicals giving outputs within the measurement error limits of the borderline output, as discussed by Leontaridou et al. (2017). Consequently, the true predictive performance of GARD™ may be even better than implied by the figures in Table 3. This also applies to the OECD guideline assay combinations (Roberts and Patlewicz, 2018).

Notwithstanding these considerations, and opportunities for further refinement, the overall conclusion drawn is that GARD™ outperforms the 3 non-animal OECD guideline assays, either alone or in combination. In any situation where the “2 out of 3” strategy is considered adequate, GARD™ could be used instead on a standalone basis.

More generally, the hitherto widely assumed necessity for an ITS combining data from several assays that span KEs from the skin sensitization AOP is not supported by the evidence. The evidence is consistent with the different perspective outlined in the Introduction, that the potential of a chemical to sensitize depends only on its ability, when applied to the skin, to covalently modify the appropriate proteins, corresponding to KE1 of the AOP. From this perspective it can clearly be understood why:

1. The LLNA works very well - it uses all the appropriate proteins, even though their identity is not known, in all the appropriate locations, and it has the appropriate range and degree of capability to activate chemicals that are not directly reactive.
2. It has proved possible to derive very good Quantitative Mechanistic Models (QMMs) correlating potency (in the LLNA, and earlier in GP assays) with chemical reactivity and partitioning parameters, covering individual chemical reaction domains (discussed by Roberts and Aptula, 2008; Roberts et al., 2016), but it has so far proved impossible to get good global QSARs for skin sensitization. This is because within a chemical reaction mechanistic domain, but not across different domains, reactivity towards one model (e.g. peptide, butylamine, aniline ... which can be used to measure a reactivity value) correlates with reactivity in other models (the proteins whose identity are not known). Also different reaction mechanistic domains require different models for ability to access the relevant proteins.
3. The DPRA alone performs quite well, but far from perfectly. Lack of perfection is in part because it is unsuitable for some reactive chemicals (e.g. those that are easily hydrolysed, those whose reaction products are easily hydrolysed, those that are water-insoluble) and in part because it cannot adequately model activation processes for chemicals that are not directly reactive. The latter deficiency is partly overcome by combinations of DPRA with ARE-Nrf2 luciferase, and/or h-CLAT (Urbisch et al., 2016; Roberts and Patlewicz, 2018), these cell-based assays having metabolic capability (Fabian et al., 2013).
4. Likewise, ARE-Nrf2 luciferase, and h-CLAT perform quite well but far from perfectly. As discussed earlier, combinations work better than the individual assays because they to some extent make up for each others' deficiencies.
5. GARD outperforms DPRA, ARE-Nrf2 luciferase, and h-CLAT, and their combinations, arguably because: a) the range of proteins is more representative of the *in cutaneo* range in terms of their chemical nature and their accessibility (applying particularly to the DPRA); b) the assay models the activation processes better than the other assays.

Funding

This work was supported by SenzaGen AB.

Acknowledgement

To Prof. Ian Kimber for many helpful and insightful comments during preparation of this paper.

Transparency document

Transparency document related to this article can be found online at <https://doi.org/10.1016/j.yrtph.2018.07.014>.

References

- Adler, S., Basketter, D., Creton, S., Pelkonen, O., van Benthem, J., Zuang, V., Andersen, K.E., Angers-Loustau, A., Aptula, A., Bal-Price, A., et al., 2011. Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. *Arch. Toxicol.* 85, 367–485. <https://doi.org/10.1007/s00204-011-0693-2>.
- Api, A.M., Basketter, D., Lalko, J., 2015. Correlation between experimental human and murine skin sensitization induction thresholds. *Cutan. Ocul. Toxicol.* 34 (4), 298–302.
- Asturiol, D., Casati, S., Worth, A., 2016. Consensus of classification trees for skin sensitisation hazard prediction. *Toxicol. Vitro*. <https://doi.org/10.1016/j.tiv.2016.07.014>. pii: S0887-S2333 (16)30143-6.
- Basketter, D.A., Natsch, A., Ellis, G., Api, A.M., Irizar, A., Safford, B., Ryan, C., Kern, P., 2018. Interspecies Assessment Factors and Skin Sensitization Risk Assessment. <https://doi.org/10.1016/j.yrtph.2018.06.018>. [Epub ahead of print] (RTP 18–60).
- EC, 2009. Regulation (EC) No 1223/2009 of the European parliament and the Council of 30 November 2009 on cosmetic products. *Off. J. Eur. Union* L342, 59–209.
- Fabian, E., Vogel, D., Blatz, V., Ramirez, T., Kolle, S., Eltze, T., van Ravenzwaay, B., Oesch, F., Landsiedel, R., 2013. Xenobiotic metabolizing enzyme activities in cells used for testing skin sensitization in vitro. *Arch. Toxicol.* 87, 1683–1696.
- Forreryd, A., Zeller, K.S., Lindberg, T., Johansson, H., Lindstedt, M., 2016. From genome-wide arrays to tailor-made biomarker readout – progress towards routine analysis of skin sensitizing chemicals with GARD. *Toxicol. Vitro* 37, 178–188. <https://doi.org/10.1016/j.tiv.2016.09.013>.
- Gerberick, G.F., Ryan, C.A., Kern, P.S., 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* 16, 157–202.
- Goebel, C., Aeby, P., Ade, N., Alépée, N., Aptula, A., Araki, D., Dufour, E., Gilmour, N., Hibatallah, J., Keller, D., et al., 2012. Guiding principles for the implementation of non-animal safety assessment approaches for cosmetics: skin sensitisation. *Regul. Toxicol. Pharmacol.* 63, 40–52. <https://doi.org/10.1016/j.yrtph.2012.02.007>.
- Haneke, K.E., Tice, R.R., Carson, B.L., Margolin, B.H., Stokes, W.S., 2001. ICCVAM evaluation of the murine local lymph node assay. Data analyses completed by the national toxicology program interagency center for the evaluation of alternative toxicological methods. *Regul. Toxicol. Pharmacol.* 34, 274–286. <https://doi.org/10.1006/rtp.2001.1498>.
- Hoffmann, S., Kleinstreuer, N., Nathalie, Alépée, Allen, D., Api, A.M., Ashikaga, T., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Goebel, C., Kern, P.S., Klaric, M., Kühnl, J., Lalko, J.F., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Parakhia, R., van Vliet, E., Zang, Q., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database. *Crit. Rev. Toxicol.* 48 (5), 344–358. <https://doi.org/10.1080/10408444.2018.1429385> (2018).
- Johansson, H., Gradin, R., 2017. Skin sensitization: challenging the conventional thinking—a case against 2 out of 3 as integrated testing strategy. *Tox. I Sci.* 2017, 1–3. <https://doi.org/10.1093/toxsci/kfx115>.
- Johansson, H., Gradin, R., Forreryd, A., Agemark, M., Zeller, K., Johansson, A., Larne, O., van Vliet, E., Borrebaeck, C., Lindstedt, M., 2017. Evaluation of the GARD assay in a blind cosmetics Europe study. *ALTEX*. <https://doi.org/10.14573/altex.1701121>.
- Johansson, H., Lindstedt, M., Albrekt, A.-S., Borrebaeck, C., 2011. A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. *Genomics* 12, 39.
- Johansson, H., Rydnert, F., Kühnl, J., Schepky, J., Borrebaeck, C., Lindstedt, M., 2014. Genomic allergen rapid detection in-house validation—a proof of concept. *Toxicol. Sci.* 139, 362–370. <https://doi.org/10.1093/toxsci/kfu046>.
- 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. <https://doi.org/10.2310/6620.2009.09038>.
- Leontaridou, M., Urbisch, D., Kolle, S.N., Ott, K., Mulliner, D.S., Gabbert, S., Landsiedel, R., 2017. The borderline range of toxicological methods: quantitation and implications for evaluating precision. *ALTEX* 34, 525–538.
- OECD, 2012. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins Part 1: Scientific Evidence. Series on Testing and Assessment No. 168 ENV/JM/MONO(2012)10/PART1. .
- OECD, 2015a. Organisation for Economic Cooperation and Development. Test Guideline 442C: in Chemico, Skin Sensitisation, Direct Peptide Reactivity Assay (DPRA). http://www.oecd-ilibrary.org/environment/test-no-442c-in-chemicoskin-sensitisation_9789264229709-en.
- OECD, 2015b. Organisation for Economic Cooperation and Development. Test Guideline 442E: in Vitro Skin Sensitisation, human Cell Line Activation Test (h-CLAT). <http://www.oecd-ilibrary.org/content/book/9789264229822-en>.
- OECD, 2016a. Organisation for Economic Cooperation and Development. Test Guideline 442D: in Vitro Skin Sensitisation, ARE-Nrf2 Luciferase Test Method. <https://doi.org/10.1787/9789264264359-en>.
- OECD, 2016b. Guidance Document on the Reporting of Defined Approaches and Individual Information Sources to Be Used within Integrated Approaches to Testing and Assessment (IATA) for Skin Sensitization. Series on Testing and Assessment No. 256 ENV/JM/MONO(2016)29. .
- Reisinger, K., Hoffmann, S., Alépée, N., Ashikaga, T., Barroso, J., Elcombe, C., Gellatly, N., Galbiati, V., Gibbs, S., Groux, H., et al., 2015. Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. *Toxicol. Vitro* 29, 259–270.
- Roberts, D.W., Aptula, A.O., 2008. Determinants of skin sensitisation potential. *J. Appl. Toxicol.* 28, 377–387.
- Roberts, D.W., Api, A.M., Patlewicz, G.Y., Schultz, T.W., 2016. Chemical applicability domain of the Local Lymph Node Assay (LLNA) for skin sensitisation potency. Part 1. Underlying physical organic chemistry principles and the extent to which they are represented in the LLNA validation dataset. *Regul. Toxicol. Pharmacol.* 80, 247–254. <https://doi.org/10.1016/j.yrtph.2016.07.007>.
- Roberts, D.W., Patlewicz, G., 2018. Non-animal assessment of skin sensitization hazard: is an integrated testing strategy needed, and if so what should be integrated? *J. Appl. Toxicol.* 38, 41–50. <https://doi.org/10.1002/jat.3479>.
- Roberts, D.W., Api, A.M., 2018. Chemical applicability domain of the local lymph node assay (LLNA) for skin sensitisation potency. Part 4. Quantitative correlation of LLNA potency with human potency. *Regul. Toxicol. Pharmacol.* 96, 76–84.
- Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., Landsiedel, R., Jaworska, J., Kern, P.S., Gerberick, F., Natsch, A., Emter, R., Ashikaga, T., Miyazawa, M., Sakaguchi, H., 2015. Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul. Toxicol. Pharmacol.* 71, 337–351.
- Urbisch, D., Becker, M., Honarvar, N., Kolle, S.N., Mehling, A., Teubner, W., Wareing, B., Landsiedel, R., 2016. Assessment of pre- and pro-haptens using nonanimal test methods for skin sensitization. *Chem. Res. Toxicol.* 29, 901–913. <https://doi.org/10.1021/acs.chemrestox.6b00055>.
- Zeller, K.S., Forreryd, A., Lindberg, T., Gradin, R., Chawade, A., Lindstedt, M., 2017. The GARD Platform for Potency Assessment of Skin Sensitizing Chemicals. *ALTEX Online* first published April 12, 2017. <https://doi.org/10.14573/altex.1701101>, version 2.
- Zuang, V., Desprez, B., Barroso, J., Belz, S., Berggren, E., Bernasconi, C., Bessems, J., Bopp, S., Casati, S., Coecke, S., et al., 2015. EURL ECVAM Status Report on the Development, Validation and Regulatory Acceptance of Alternative Methods and Approaches. EUR 27474 EN Available at: <http://publications.jrc.ec.europa.eu/repository/bitstream/JRC97811/lbna27474enn.pdf>.



Letter to the editor regarding the article by Roberts, 2018

To the editor

In a recent contribution by D. Roberts, funded by the developers of the GARD™ assay, the question was raised: „Is a combination of assays really needed for non-animal prediction of skin sensitization potential?” (Roberts, 2018) and it was proposed that, based on the superior predictivity on a set of chemicals tested by the assay developers, the transcriptomic GARD™ assay as stand-alone assay may be a better option as the currently widely discussed proposal of using multiple test covering different steps of the Adverse Outcome Pathway (AOP) (OECD, 2012). A strong argument for a stand-alone assay instead of a combination of tests, namely a ‘2 out of 3’ approach, had been made by the GARD™ development team in an earlier review, too (Johansson and Gradin, 2017). This proposal and the development of the GARD™ assay challenges a number of conventions in *in vitro* toxicology, challenges we think should be carefully discussed.

1. Challenging the AOP approach

The author makes the correct observation, that strict ‘AOP logic’ – describing the sensitization phase as a linear process covered by mechanistic steps – would require a series of *in vitro* tests, and all of those would need to be tested positive for a positive final outcome (Roberts, 2018). This linear view is actually the proposal made for the first time in the seminal paper by Jowsey et al. (2006), which predates the AOP discussion but which is very much recapitulated in the way the skin sensitization AOP was drawn by the OECD (OECD, 2012). This approach would always lead to very high specificity, but low sensitivity. However, it is widely acknowledged, that AOPs should not be viewed in such a linear way. Using a suite of mechanistic tests within an AOP may also be viewed as collecting mechanistically relevant data on defined molecular endpoints to finally make a ‘weight of evidence’ judgement. For skin sensitization, it is widely acknowledged that this WoE can be derived from chemical information, reactivity measurements describing the molecular initiating event (MiE) combined with cell based assays.

The assays and *in silico* tools developed to test chemicals for skin sensitization potential are all based on a good, albeit not complete, understanding of the molecular mechanisms underlying skin sensitization. The development of the AOP approach in toxicology in general is based on the assumption that increased mechanistic understanding and testing of defined molecular endpoints will lead to continuously improved prediction of apical endpoints.

The GARD™ assay is only using a single cell-type and (like some related assays based on a pan-transcriptomic analysis) is based on an unsupervised transcriptomic approach not taking any *a priori* mechanistic knowledge into account but defining the measured endpoints

(transcriptomic signature) completely based on statistical evaluation of gene expression changes induced by a set of sensitizers and non-sensitizers. On the positive side, by measuring multiple transcripts not previously assessed, information content is greatly increased and any bias from previous research is removed – however, using this assay as stand-alone assay as proposed would also ignore all previous mechanistic learnings in skin sensitization and statistics would dominate mechanistic knowledge – which of course is fine if indeed our mechanistic understanding is inferior to a purely statistical evaluation of ‘big data’. However, to miss the opportunity to carefully characterize the MiE with *in chemico* methods and trust that chemicals will react with all relevant proteins in the cells and then trigger a most relevant gene induction as proposed by Roberts is a true paradigm shift and also contradicting all previous teaching by the same author.

2. Stable biological endpoint

Currently, one key requirement for validation of *in vitro* methods is mechanistic relevance of reproducible *in vitro* endpoints. Two observations from the literature of the GARD™ developers should be discussed in this respect. The original assay was developed with the Affymetrix platform (Johansson et al., 2013), later the approach was tested on different platforms (Forreryd et al., 2014) and finally it was transferred to the nanoString® nCounter technology (Forreryd et al., 2016). It was observed that this change of technology platform led to only moderate to poor correlation of the original transcriptomic signature to the new platform (Forreryd et al., 2014), yet the new platform again gave a similarly high predictivity. Thus the measured signature seems not to be a fixed biological endpoint measurable with different technical means – but it is partly dependent on detection technology. This is compensated for, obviously, by partial redundancy in information content in the gene signature which keeps reported predictivity high, independent of the exact molecular endpoints measured (Forreryd et al., 2016). However, it indicates that we do not deal here with a completely defined molecular endpoint. Secondly, it is described that there are batch effects in the GARD™ assay, necessitating that each test series includes a benchmarking set of 10 chemicals for fine-tuning of the prediction signature to correct for batch effects (Forreryd et al., 2016). This is different from other *in vitro* assays, where the prediction model is fixed and positive controls are only run for assay performance check.

3. Predictivity

The paper by Roberts makes strong arguments based on the superior predictivity of the published dataset vs. the OECD validated assays and

DOI of original article: <https://doi.org/10.1016/j.yrtph.2019.01.006>

<https://doi.org/10.1016/j.yrtph.2019.01.005>

Received 29 October 2018; Accepted 2 January 2019

Available online 07 January 2019

0273-2300/ © 2019 Elsevier Inc. All rights reserved.

vs. the simplest WoE approach, namely the ‘2 out of 3’ approach which is listed as an OECD case study (Guidance Document on the, 2016). One may, from a regulatory toxicology viewpoint, indeed claim that ‘all is well, that predicts well’. However this will require more independent evaluation. First, it should be noted, that the analysis of Roberts is comparing *in vitro* data to the LLNA and not to human data (albeit these are available for skin sensitization).

Comparing GARD™ results to LLNA data, Roberts is overlooking uncertainties arising from the choice of the data base and the limited precision of the assay. In an independent Cosmetic Europe evaluation with blind coded chemicals (Johansson et al., 2017), the balanced accuracy was at 75%, while it was 88% in the in-house validation for the GARD assay. This lower balanced accuracy may be partly due to a bias of a low number of non-sensitizers tested in the Cosmetics Europe study. Nevertheless, a similar effect was noted by Cosmetic Europe when evaluating another assay based on multiple transcripts, where the balanced accuracy for the independent evaluation was 19% lower than the one reported from an in-house dataset, an effect which was not seen to that extend for the OECD validated assays. In addition to the chosen dataset, also limited precision of testing methods results in uncertainties of their predictive accuracy metrics. Hence the comparison of distinct numbers – Roberts compares 88% accuracy of the GARD™ assay to 83% of the “2 out of 3” approach – is neglecting this uncertainty. The 95% confidence interval of the accuracy of the “2 out of 3” approach is 74–97%; uncertainty information on the GARD™ assay is missing, but there is clearly an overlap rather than a significant difference in the accuracy measures.

4. Lack of transparency

It was often stated that the Tox21 approach will lead us into a mechanistic toxicology, which brings us away from the ‘black box’ animal test. The purely statistical approach of a transcriptomic analysis on > 200 transcripts (which are dependent on the detection methodology), selected from a pan-genomic analysis, however leads to a black-box outcome again, and there is no full transparency on what is determining the predictions being made, although some interrogation of relevant pathways was made.

5. Gard potency – predicting potency without dose-response information

The mechanistically most challenging proposal from the GARD platform is the recently published potency assay (Zeller et al., 2017), which claims a good discrimination of GHS sub-categories 1A and 1B by differential gene expression. The GARD assay does make a concentration range finder, but then test chemicals at only one defined concentration at 90% cellular viability. As described in the paper, the selected test concentration does not form part of the prediction model, which is entirely based on a gene signature and does not include dose-response information. It is of note, that these two GHS sub-categories (GHS sub-category 1A assigned for high incidence in humans or high potency as indicated by LLNA EC₃ < 2%) are based on animal test dose-response data and the threshold of EC₃ < 2% is purely a regulatory decision without additional biological meaning. It would be logical, that this arbitrary threshold in a continuous dose-dependent variable in an animal test would best be modeled based on dose-response

information from an *in vitro* test or a battery of *in vitro* tests. Although the authors have done a careful discrimination between training and test set, it is intriguing and biologically not intuitive that it is possible to classify chemicals into two regulatory defined potency sub-categories based on a gene signature not taking dose-response information into account. It will be very important to safeguard against overfit when making such proposals.

In conclusion, the strong proposals made in the paper by Roberts and by the GARD developers raise several questions which need careful discussion by the toxicological community before we can confirm the superior predictivity of the statistics-derived single cell-type method and move away from the AOP paradigm. Statistics-derived methods in toxicology may harness ‘big data’ and hence be information-rich but also lose on the transparency and biological as well as on the mechanistic side.

Transparency document

Transparency document related to this article can be found online at <https://doi.org/10.1016/j.yrtph.2019.01.005>.

References

- Forreryd, A., Johansson, H., Albrekt, A.S., Lindstedt, M., 2014. Evaluation of high throughput gene expression platforms using a genomic biomarker signature for prediction of skin sensitization. *BMC Genomics* 15, 379.
- Forreryd, A., Zeller, K.S., Lindberg, T., Johansson, H., Lindstedt, M., 2016. From genome-wide arrays to tailor-made biomarker readout - progress towards routine analysis of skin sensitizing chemicals with GARD. *Toxicol. Vitro* 37, 178–188.
- Guidance Document on the Reporting of Defined Approaches and Individual Information Sources to Be Used within Integrated Approaches to Testing and Assessment (IATA) for Skin Sensitisation. ENV/JM/MONO, pp. 29. <https://doi.org/10.1787/9789264279285-en>.
- Johansson, H., Gradin, R., 2017. Skin sensitization: challenging the conventional thinking-A case against 2 out of 3 as integrated testing strategy. *Toxicol. Sci.* 159, 3–5.
- Johansson, H., Albrekt, A.S., Borrebaeck, C.A., Lindstedt, M., 2013 Apr. The GARD assay for assessment of chemical skin sensitizers. *Toxicol. Vitro* 27 (3). <https://doi.org/10.1016/j.tiv.2012.05.019>.
- Johansson, H., Gradin, R., Forreryd, A., Agemark, M., Zeller, K., Johansson, A., Larne, O., van Vliet, E., Borrebaeck, C., Lindstedt, M., 2017. Evaluation of the GARD assay in a blind Cosmetics Europe study. *ALTEX* 34, 515–523.
- Jowsey, I.R., Basketter, D.A., Westmoreland, C., Kimber, I., 2006. A future approach to measuring relative skin sensitising potency: a proposal. *J. Appl. Toxicol.* 26, 341–350.
- OECD, 2012. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins, Part 1: Scientific Evidence. OECD ENVIRONMENT, HEALTH AND SAFETY PUBLICATIONS SERIES ON TESTING AND ASSESSMENT NO. 168.
- Roberts, D.W., 2018. Is a combination of assays really needed for non-animal prediction of skin sensitization potential? Performance of the GARD (Genomic Allergen Rapid Detection) assay in comparison with OECD guideline assays alone and in combination. *Regul. Toxicol. Pharmacol.* 98, 155–160.
- Zeller, K.S., Forreryd, A., Lindberg, T., Gradin, R., Chawade, A., Lindstedt, M., 2017. The GARD platform for potency assessment of skin sensitizing chemicals. *ALTEX* 34, 539–559.

Andreas Natsch*

Givaudan Schweiz AG, Ueberlandstrasse 138, CH-8600, Duebendorf, Switzerland

E-mail address: andreas.natsch@givaudan.com.

Robert Landsiedel

BASF SE, RB/TB – Z470, D-67056, Ludwigshafen, Germany

* Corresponding author.



Reply to Letter to the Editor

Response to the comments of Natsch and Landsiedel on my paper “Is a combination of assays really needed for non-animal prediction of skin sensitization potential? Performance of the GARD (genomic allergen rapid detection) assay in comparison with OECD guideline assays alone and in combination.”



ARTICLE INFO

Keywords:

Skin sensitization

Adverse outcome pathway

GARD assay

1. Introduction

I am grateful to Natsch and Landsiedel (2019) for raising some important issues in their letter regarding my recent paper (Roberts, 2018). Some of the points they raise concern the development and details of the GARD assay, which my paper does not discuss. In my responses I have focussed on comments relating to the content of my own paper.

2. Challenging the AOP approach

Natsch and Landsiedel, referring to my argument that a stand-alone assay can match or perform better than multiple tests based on the AOP, state that “... This proposal ... challenges a number of conventions in *in vitro* toxicology ...”

As applied to skin sensitization, I would regard the field of *in vitro* toxicology as too new to have established conventions, unless one includes the structure-activity work starting with Landsteiner and Jacobs (1936) under the term *in vitro* toxicology. I assume that the main “convention” implied by Natsch and Landsiedel concerns the AOP.

While I acknowledge that AOPs can provide a useful lexicon to describe mechanisms and modes of action in toxicology, I do challenge the widely perceived necessity to invoke the AOP in all discussions of predictive models for skin sensitization. Mechanistic toxicology pre-dates the advent of AOPs by many decades.

Indeed, the AOP for skin sensitization was built upon already existing mechanistic understanding; the AOP did not *per se* deliver new mechanistic understanding. This point can be illustrated by two cases, at opposite ends of the AOP but pre-dating the AOP concept:

Development of the LLNA (Kimber et al., 1994). This was based on the then current (late 1980s-early 1990s) understanding of the biological mechanisms, in particular the insight that the acquisition of sensitization involves clonal expansion of T-cells that have recognised presented antigen, in the lymph node near to the site of exposure. Based on the recognition that this process of clonal expansion would involve an

increase in lymph node metabolic activity, which could be represented by the increase in thymidine uptake, expressed as the stimulation index SI, the degree of sensitization produced by the dose of chemical applied could be modelled. The EC₃, used as a quantitative index of potency, is the dose required to produce an SI value of 3. The LLNA is in essence a quantitative model of the final event in the AOP, without the need to consider explicitly the preceding events.

Structure-Activity Relationships (SARs) and Quantitative Mechanistic Models (QMMs). The first SAR model for skin sensitization was published in 1936 by Landsteiner and Jacobs, who reported what would now be called a discriminant analysis to assign halogen- and nitro-substituted benzene derivatives (chemicals belonging to what is now known as the S_NAr electrophile domain) as skin sensitizers or non-sensitizers, depending on whether or not they reacted with aniline in a simple chemical assay. At that time much less was known than at present about the biological pathway (AOP), and the SAR was based simply on the recognition that covalent reaction with protein(s) leading to formation of antigens is a *sine qua non* for sensitization by non-macromolecular chemicals. Subsequently the Landsteiner and Jacobs SAR for the S_NAr domain was extended and has been put on a quantitative basis using calculated (Roberts and Aptula, 2014) or experimental (Natsch et al., 2011) reactivity parameters. For other groups of chemicals, related by common chemical reaction mechanisms but not necessarily by similarity in structure, QMMs have been developed correlating sensitization potency with reactivity (in some cases measured, in other cases calculated) and (some but not all cases) hydrophobicity. These SARs and QMMs are all based on modelling the protein modification step (KE1 of the AOP) and take the subsequent KEs for granted. The good fit of these models to the available potency data provides strong support for the argument that whether or not a chemical can cause KE1 to occur to sufficient extent is the predominant determinant of whether it is a skin sensitizer or not.

I therefore stand by my argument that for skin sensitization there is no need to invoke the AOP. All of the assays should be seen for what they really are: stand-alone assays, none of which is perfectly

DOI of original article: <https://doi.org/10.1016/j.yrtph.2019.01.005>

<https://doi.org/10.1016/j.yrtph.2019.01.006>

Received 4 December 2018; Accepted 2 January 2019

Available online 07 January 2019

0273-2300/© 2019 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

predictive. Using them in combination (e.g. as in the 2 out of 3 approach) can, but need not necessarily, improve predictive performance. Importantly, this potential improvement is not because they represent different key events in the AOP, but because their inapplicability domains are different. Their lack of 100% perfection is not because of deficiencies in their modelling of their assigned KEs, but because of deficiencies in their modelling of the protein modification step.

Agreeing with my observation that strict AOP logic would require a series of *in vitro* tests, and all of those would need to be tested positive for a positive final outcome, Natsch and Landsiedel state “... it is widely acknowledged, that AOPs should not be viewed in such a linear way. Using a suite of mechanistic tests within an AOP may also be viewed as collecting mechanistically relevant data on defined molecular endpoints to finally make a ‘weight of evidence’ judgement”. From my reading of the literature it is not widely acknowledged that AOPs should “not be viewed in a ... linear way”. Rather, the contradiction between the K (key) of KE and not requiring a full set of positives in the assays purported to represent the KEs is rarely mentioned. In any event, I would question whether there is any choice in how the AOP should be viewed. It is what it is.

Natsch and Landsiedel continue “The development of the AOP approach in toxicology in general is based on the assumption that increased mechanistic understanding and testing of defined molecular endpoints will lead to continuously improved prediction of apical endpoints.” It is not clear what the authors mean by “defined molecular endpoints”. For skin sensitization KE1, covalent modification of the relevant protein(s), is the only KE that I would consider to involve a molecular mechanism. After KE1 the acquisition of skin sensitization is complicated. Many cells and many types of biomolecules are required to initiate the development of sensitization, and their interactions are all closely regulated in time and space.

So I do not share the assumption that increased mechanistic understanding and testing of defined molecular endpoints will lead to continuously improved prediction of apical endpoints. Quantitative mechanistic modelling (QMM) of the MIE, where this is possible (and to a large though not complete extent it IS possible in skin sensitization) is in my experience sufficient in many cases. Where it is not (inadequate understanding of the chemistry, unavailability of expertise) then “chemistry-blind” approaches are available and can be developed further. If these are good, then SARs and QMMs relating the predictions of these assays to chemistry can be explored. The problem with testing in multiple assays is that they don't always agree with each other, and from a chemistry-blind perspective it is difficult to know which result to believe. So rather than real integration based on mechanistic understanding we end up with majority voting and the “weight of evidence” is based on the unstated simplifying assumption that all units of evidence have the same weight, which works reasonably well but not perfectly.

3. The GARD assay

Much of what Natsch and Landsiedel write about the GARD assay is focused on details of the assay and on publications by SenzaGen authors rather than on the content of my paper. As regards the latter:

Natsch and Landsiedel state, referring to the statistical evaluation of gene expression changes so as to arrive at the GARD prediction “... to miss the opportunity to carefully characterize the MIE with *in chemico* methods and trust that chemicals will react with all relevant proteins in the cells and then trigger a most relevant gene induction as proposed by Roberts is a true paradigm shift and also contradicting all previous teaching by the same author.” I don't think the paper was proposing “to miss the opportunity to [apply *in chemico* methods]”. But not all organisations who need to assess chemicals for sensitization have chemistry expertise. As regards my “previous teaching” which I presume refers to my published work on chemistry-based SARs, QSARs and QMMs in skin sensitization, I would summarise it as follows:

All assays, including the LLNA, can be regarded as measures of how good the chemical is at modifying the appropriate proteins *in cutaneo* (bearing in mind that we don't know which are the appropriate proteins and how they are distributed *in cutaneo*). As I see it, we have the animal assays (LLNA, guinea pig ...) that could tell people - non-chemists and chemists alike - how potent (if at all) a chemical is as a sensitizer. If we want to predict what a chemical would do without animal tests, we can try to use chemistry mechanistic insights or we can try to develop a non-animal “surrogate LLNA”. If the latter approach is really successful, chemists could apply chemistry to predict and interpret the “surrogate LLNA” in the same way as we do now in SAR and QMM work with LLNA data. I suspect that however successful chemistry-based prediction may become, for regulatory purposes there will always be a need for an assay or ITS that can be used without needing to know any chemistry. Even the DPRA is basically a simple chemistry assay that can be done and interpreted by non-chemists. I don't know if chemistry will ever be able to predict sensitization as successfully as the LLNA can, but even as it stands at present we can use chemistry to assess applicability domain boundaries, question and explore whether positives in animal or non-animal assays reflect intrinsic potency, presence of impurities, false positive mechanisms ... And of course chemistry (based on doing experiments such as kinetics and reaction mechanism studies, or based on evaluation by human experts or based on expert systems) is very useful for product developers.

I state in my paper that I interpret the good performance of GARD on the presumption that it must be modelling the ability of chemicals to access the relevant proteins, to react with them and, where necessary, to be metabolically activated. In my analysis of the data, I treat GARD output ($\pm$) in exactly the same way as I would treat LLNA output on a $\pm$ basis. In effect, assessing GARD as a “surrogate LLNA” on a $\pm$ basis. I don't see any contradiction with my “previous teaching”.

4. Predictivity

Natsch and Landsiedel state “it should be noted, that the analysis of Roberts is comparing *in vitro* data to the LLNA and not to human data (albeit these are available for skin sensitization).” I'm not sure if I say this in the paper, but I've certainly said it in presentations - for evaluating *in vitro* assays, SARs and QMMs, LLNA data are more suitable than human data. Even LLNA data aren't always perfect, but compared with human data they have a more precisely defined endpoint and there is more control over test sample purity. Several compounds have a history of being defined as human sensitizers alternating with being defined as human non-sensitizers.

Regarding predictive performance Natsch and Landsiedel state “In an independent Cosmetics Europe evaluation with blind coded chemicals [their ref. 9], the balanced accuracy was at 75%, while it was 88% in the in-house validation for the GARD assay. This lower balanced accuracy may be partly due to a bias of a low number of non-sensitizers tested in the Cosmetics Europe study.” I have studied reference 9, and in line with what Natsch and Landsiedel suggest, the lower accuracy is the result of a rather low number of LLNA non-sensitizers – 18 are listed, of which 8 are positive in the GARD assay. In fact one of the listed “LLNA non-sensitizers, tocopherol, is positive in the LLNA (EC₃ = 7.4%), so it should have been listed as a true positive prediction. This still leaves 7 out of 17 LLNA negatives predicted as positive (59% specificity). Of these some (e.g. diethyl phthalate and diethyl toluamide) lack any structural alerts for reactivity or potential to be activated to reactive derivatives, and are probably true non-sensitizers. For others (e.g. hydrocortisone, propyl paraben) weak sensitization has been reported in other assays and they are probably borderline cases. The larger datasets that I analysed in my paper contain all of these false positives in addition to further LLNA negatives which are mostly predicted correctly by the GARD assay. Most if not all of these additional negatives also feature in datasets that have been used to evaluate other assays and combinations of them. (I have not made a detailed check but I recognise

many of these LLNA negatives from previous data analyses that I have done). So it seems reasonable to consider the larger dataset that I analysed to be more representative than the smaller dataset in reference 9.

I agree that the performances of “2 out of 3” and GARD, as found in my paper, are not greatly different. It will be interesting to see how they compare on larger datasets and in light of the further analysis currently in progress for the mis-predicted chemicals.

5. Lack of transparency

In biology, and in many other fields of science and technology, degree of transparency is highly subjective. Natsch and Landsiedel write “It was often stated that the Tox21 approach will lead us into a mechanistic toxicology, which brings us away from the ‘black box’ animal test.” Fine words, but do they really mean anything? Is the LLNA really a “black box”? Are the guinea pig assays such as the Buehler test and the GPMT to be considered as “black boxes”? In the same way as I can use a mobile phone without needing to understand how it works, risk assessors can use the outputs of *in vivo*, *in vitro* and *in chemico* assays without needing to understand the mechanistic basis of these assays.

From my perspective GARD is no different in this respect. It gives outputs which can be distinguished from each other, depending whether the test chemical is a sensitizer or a non-sensitizer, as do the other assays. With all of these, the user can just accept the result, but if the expertise is available the result can always be interrogated from a mechanistic standpoint.

6. In conclusion

I appreciate the initiative by Natsch and Landsiedel in commenting on my paper, and am grateful to the editor for the opportunity to respond. It has long been my view that science progresses more by disagreement and debate than by consensus.

Funding

This work was supported by SenzaGen AB.

References

- Kimber, I., Dearman, R.J., Scholes, E.W., Basketter, D.A., 1994. The local lymph node assay: developments and applications. *Toxicology* 93, 13–31.
- Landsteiner, K., Jacobs, J.L., 1936. Studies on the sensitisation of animals with simple chemicals III. *J. Exp. Med.* 64, 625–639.
- Natsch, A., Haupt, T., Laue, H., 2011. Relating skin sensitizing potency to chemical reactivity: reactive Michael acceptors inhibit NF- κ B signaling and are less sensitizing than S_NAr- and S_N2- reactive chemicals. *Chem. Res. Toxicol.* 24, 2018–2027.
- Natsch, A., Landsiedel, R., 2019. RTP-18-448 Roberts, D. W. 2018. Is a combination of assays really needed for non-animal prediction of skin sensitization potential? Performance of the GARD (Genomic Allergen Rapid Detection) assay in comparison with OECD guideline assays alone and in combination. *Regul. Toxicol. Pharmacol.* 98, 155–160.
- Roberts, D.W., Aptula, A.O., 2014. Electrophilic reactivity and skin sensitization potency of S_NAr electrophiles. *Chem. Res. Toxicol.* 27, 240–246.

David W. Roberts
School of Pharmacy and Biomolecular Sciences, Liverpool John Moores
University, Liverpool, L3 3AF, United Kingdom
E-mail address: d.w.roberts@ljmu.ac.uk.