ALTEX, accepted manuscript
published July 8, 2021
doi:10.14573/altex.2105111

Review Article

A Triangular Approach for the Validation of New Approach Methods for Skin Sensitization

Andreas Natsch[1], Robert Landsiedel[2] and Susanne N. Kolle[2]
[1]Fragrances S&T, Ingredients Research, Givaudan Schweiz AG, Kemptthal, Switzerland; [2]BASF SE Experimental Toxicology and Ecology, Ludwigshafen, Germany

Abstract
The availability of reference data is a key requirement for the development of New Approach Methods (NAM), i.e. *in vitro, in chemico* and *in silico* methods and Integrated Approaches, like Defined Approaches (DA), combining these data sources. Reference data are of even higher importance for regulatory acceptance. In contrast to most other adverse effects, human skin sensitization data on many chemicals are available, next to data from animal studies, such as the Local Lymph Node Assay (LLNA). Skin sensitization NAM data can therefore be compared to different reference datasets. Recent publications and validation at the OECD focused on human and LLNA reference data. The "2 out of 3" DA (2o3 DA) is the first DA for skin sensitization solely based on experimental data from validated tests and was recently adopted as an OECD test guideline. Here we review the predictivity of the 2o3 DA on multiple human and LLNA reference datasets. Concomitantly, we compare the predictivity of the LLNA for human data within the same datasets. Comparing predictivity of methods not only bilaterally (NAM or DA vs. animal method) but including human data in a triangle "NAM data – animal data – human data" offers a comprehensive assessment of NAM's and DA's predictivity. In all these assessments the 2o3 DA was superior to the LLNA in predicting human skin sensitization hazard. This highlights the importance of a holistic view on reference data instead of limiting validation of NAM and DA to data from a single animal test only.

1 Introduction

The development of New Approach Methods (NAM), i.e., *in vitro, in chemico* and *in silico* methods, has become a key focus in toxicology. In order to develop hypothesis-driven, mechanistically based new tests, a limited and discrete set of reference substances with well-defined *in vivo* reference data is often sufficient, and for skin sensitization such a small set had been proposed early on (Casati et al., 2009). Larger datasets are usually needed to train empirical models based on a large number of input variables such as models based on genomic data or *in silico* models (Dimitrov et al., 2005; Johansson et al., 2013). Nevertheless, when it comes to method validation and regulatory acceptance, assessment of predictivity of the test method (or approaches combining multiple methods) becomes a key aspect and this can only be achieved with datasets of sufficient size and with high quality *in vivo* data (Kolle et al., 2019). Toxicological endpoints lacking such high-quality data may face a significant delay in acceptance of NAM. Thus, to name an example, for acute respiratory toxicity no method has gone into full validation despite decade long research and even though multiple models, including commercial 3D lung tissue models, had been developed (Lacroix et al., 2018). One may suspect that this lack of a validated replacement could be closely linked to the intrinsic noise of *in vivo* data in that systemic endpoint.

For ingredients in consumer products with topical exposure, the skin sensitization endpoint is one of the most important aspects in the safety evaluation of new and existing chemicals. The replacement of skin sensitization testing by non-animal methods has thus been a strong research focus, accelerated by the ban of animal testing for cosmetic ingredients in the European Union in 2013 (EC, 2009, 2013). This research focus led to the rapid development of multiple NAMs, by both academic and industrial laboratories (Ezendam et al., 2016). At the time of writing, eight of these methods addressing the endpoint skin sensitization were fully validated in multi-laboratory ring-trials and implemented as protocols in OECD TG 442C, 442D and 442E (OECD, 2018a,b, 2020). Relatively rapid adoption of these methods by the OECD may have been facilitated by the fact that all these methods were specifically labeled not to be used as stand-alone methods, especially not as stand-alone methods to rate a chemical as non-sensitizer (they could be used, though, for a positive labeling). Thus, some limitations in predictivity were not of major concern in the validation and adoption of the test guidelines, while technical

Received May 11, 2021; Accepted July 7, 2021;
Epub July 8, 2021; © The Authors, 2021.

ALTEX 38(#), ###-###. doi:10.14573/altex.2105111

Correspondence: Andreas Natsch, PhD
Kemptpark 50, CH-8310 Kemptthal
(andreas.natsch@givaudan.com)

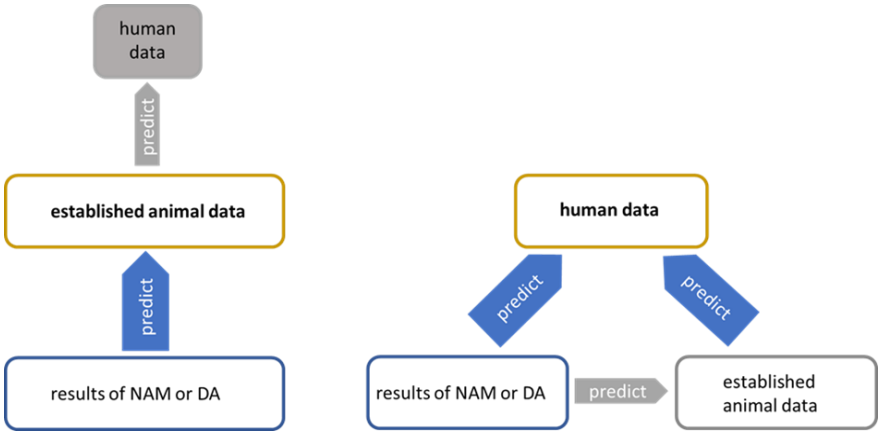

Fig. 1: Traditional bilateral and new triangular evaluation of predictive capacities of NAM and DA

reproducibility and protocol standardization were of major importance. It was always assumed, that for final prediction of the skin sensitization endpoint, the information of multiple tests would be integrated in integrated approaches (Jowsey et al., 2006; OECD, 2021a).

To standardize the interpretation of the aggregated information from multiple tests and information sources, Defined Approaches (DA) were proposed as a way forward. DAs are fixed workflows allowing an assessment of multiple inputs with a fixed data interpretation procedure (DIP) without expert judgment. DAs therefore should deliver an unambiguous rating and can thus be subject to Mutual Acceptance of Data (MAD) by OECD member countries. To put this concept into practice, the OECD has developed a new guideline (No. 497) which describes two simple DA for the skin sensitization endpoint (OECD, 2021a) and after three years of in-depth discussions by an expert group convening in multiple meetings, this guideline was finally adopted in June 2021.

To accept a DA or Integrated Approach to fully replace the endpoint of concern, in this case skin sensitization, predictivity becomes center stage and the detailed discussions within the OECD expert group mainly focused on this aspect. Hence reference data were of major interest, and a detailed review of the reference data was made by the OECD expert group (OECD, 2021b). However, predictivity of the Defined Approaches was already assessed repeatedly prior to OECD submission, and it thus becomes possible to compare the published predictivity of the DA on different reference datasets.

The "2 out of 3" approach (2o3 DA) (Bauch et al., 2012; Urbisch et al., 2015) described in the new OECD guideline allows hazard identification based on two concordant, non-borderline outcomes based on the validated prediction models from OECD test guideline 442D, DPRA (442C) and h-CLAT (442E). It is the only DA fully based on validated experimental methods and prediction models. Next to the 2o3 DA, the recently published OECD guideline on DAs for skin sensitization (OECD, 2021a) also includes a modified version of the Integrated Testing Strategy (ITS) (Nukada et al., 2013), which combines scores derived from the *in vitro* data (based on data obtained from DPRA and h-CLAT) and an *in* silico prediction. Here we review the predictivity of the 2o3 DA in different literature sources and for the OECD reference database used to evaluate the DAs and discuss the underlying reference datasets. By comparing predictivity in the "2o3 DA – animal data/LLNA – human data" triangle as shown in Figure 1, a detailed assessment of predictivity is possible, and predictivity of both the DA and the LLNA against human data can be compared to each other.

2 Description of the reference data sets and comparison of human and LLNA reference data

The predictivity of the 2o3 DA and of the individual underlying test methods had repeatedly been assessed against different databases prior to the assessment by the OECD (Bauch et al., 2012; Kleinstreuer et al., 2018; Natsch et al., 2013; Urbisch et al., 2015). These datasets were always significantly overlapping, but different authors had continuously added chemicals based on the evaluation target which was set differently in different studies:

The first evaluation of the 2o3 DA was performed by Bauch *et al.* when the approach was first described (Bauch et al., 2012). This was based on a set of 54 chemicals with LLNA data, 50 of which also had human evidence available. The human data were mainly retrieved from Basketter *et al.* (Basketter et al., 1999) among few other sources. In the dataset of Bauch *et al.*, the LLNA and human outcome are largely congruent. Thus, balanced accuracy (BA) of the LLNA vs. human data is 88% (Table 1). This value is affected by the fact that the underlying dataset (Basketter et al., 1999) was compiled for a retrospective statistical evaluation of the optimal threshold for LLNA positivity, and it contained largely congruent LLNA and human data. Furthermore, the selection of chemicals made by Bauch *et al.* was partly based on availability of KeratinoSens® data (since the same publication compared predictivity of LuSens and KeratinoSens®, the two assays described in OECD TG 442D), and part of the data were retrieved from the original KeratinoSens® publication (Emter et al., 2010). The dataset in Emter *et al.* was compiled from (i) the publication of Casati *et al.* (Casati et al., 2009) (ii) the LLNA performance standards (ICCVAM, 2009) and (iii) the initial list of *bona fide* reference chemicals for the European Sens-it-iv project (Rovida et al., 2007). Especially the former two lists were specifically made to contain chemicals with congruent LLNA and human/guinea pig data. These chemicals (n = 36) were combined with additional chemicals (n = 31) with congruent LLNA and guinea pig / human data to make up the so-called "Silver List" (n = 67). The name "Silver List" was used as we considered it the best possible list we could come up containing such congruent data from multiple *in vivo* sources for the validation of a method, anticipating that a "Gold List" with more in depth data curation would one day be generated. The chemical set in Bauch *et al.* has an overlap of 31 chemicals with the Silver List (Emter et al., 2010).

Tab. 1: Predictivity of the LLNA vs. human reference data in different studies[a]

	Sensitivity [%]	Specificity [%]	Balanced accuracy [%]	N
Bauch et al., 2012	96	81	88	50
Urbisch et al., 2015	91	64	77.5	111
Hoffmann / Kleinstreuer et al., 2018	85.2	50.0	67.6	128
OECD LLNA database[b] vs. Basketter et al. 2014 human data	99	39	69	96
OECD database[b]	94	22	58	56

[a]For transparency, values are provided with the same number of decimals as reported in the respective original publications.
[b](OECD, 2021a; OECD, 2021b)

Thus, in all these initial reference lists to validate non-animal methods to address skin sensitization, strong emphasis was put on selection of reference chemicals that had LLNA data available, but also other, *congruent* clear evidence for the sensitization risk and therefore a strong alignment esp. between LLNA and human data was intended. This was considered important, as the LLNA itself is an alternative (or in exact terms a "refinement") method, which does not directly measure skin sensitization and which was itself validated based on reference data (Kolle et al., 2020).

A larger set of chemicals (n = 145) was later presented to combine all chemicals with available data for KeratinoSens®, DPRA and dendritic cell activation test (the U-937 test, an early modification of the U-Sens protocol was used) (Natsch et al., 2013). This set was developed to *specifically predict the LLNA outcome* based on a Bayesian net approach (Jaworska et al., 2013), and in this case human references were neither available nor required, thus this reference data set had a clearly different focus.

Cosmetics Europe then developed a completely different reference list, with the main goal to categorize chemicals *regarding their human sensitization potential* (Basketter et al., 2014). This list (n = 128) categorizes chemicals in 6 classes, whereby classes 5 and 6 would be categorized as human non-sensitizers for classification purposes according to the authors. No direct comparison to the LLNA was made, except a graphical representation. Cosmetics Europe then filled all the data gaps for the *in vitro* data for this list of chemicals for some guideline methods (KeratinoSens, DPRA, h-CLAT and U-Sens) and the back then emerging method SENS-IS (Hoffmann et al., 2018). This analysis was also complemented with a detailed review of all available LLNA data, often from multiple sources. In this analysis, the LLNA is still a very sensitive method when compared to human data (sensitivity of 85.2%), but it had a clearly lower specificity (50.0%) to predict the human sensitization risk (Table 1).

In parallel to the human data collection of Basketter *et al.*, a large compilation of data was published by Urbisch *et al.* (Urbisch et al., 2015). This list contains 180 chemicals with LLNA and *in vitro* data (KeratinoSens/ LuSens, DPRA and h-CLAT[1]), and a subset (n = 103) which additionally contains human data. The human data were aggregated from the RIFM database and from the earlier publications by (Basketter et al., 2014; Bauch et al., 2012). Also, in this evaluation, predictivity of the LLNA vs. human data was analyzed. Again, a good sensitivity (91%) and a limited specificity (64%) of the LLNA to predict human reference data were reported (Table 1).

Thus in summary, published reference datasets were made either to contain (i) congruent LLNA and human/other evidence (Bauch et al., 2012; Casati et al., 2009; Emter et al., 2010), (ii) LLNA only data for a project focused on the LLNA prediction (Natsch et al., 2013), (iii) a primary focus on human data (which were complemented with LLNA and *in vitro* data later) (Basketter et al., 2014; Hoffmann et al., 2018) or all chemicals with available *in vitro* data (Urbisch et al., 2015).

The human datasets used in these studies were not assessed based on uniform evaluation criteria and involved significant expert judgment. This is due to the fact, that human predictive tests have never been standardized, are based on multiple protocols, and in some cases also clinical data and data on safe use were included in the weight-of-evidence (WoE) assessment (Basketter et al., 2014). Furthermore, the rationale for the expert judgment and WoE analysis was often not fully transparent on a chemical-per-chemical basis in these published datasets. Regarding the LLNA data, negative calls were largely made according to the practice during the LLNA validation, when negative chemicals were hardly ever tested at concentrations >25% (Kolle et al., 2020). Thus, to name an example, in the Silver List (Emter et al., 2010), chemicals were rated negative if they were negative up to a maximum test concentration of 20% in line with how the LLNA was validated.

When the OECD expert group on Defined Approaches for Skin Sensitization assessed performance of DAs, the group decided that less expert judgment should be involved and that both the LLNA data and the human data needed a thorough data curation based on fixed criteria. The details are described elsewhere (OECD, 2021c,d). Briefly, for human data, only results from human repeat insult patch test (HRIPT) and Human Maximization Tests (HMT) were considered, and negative ratings were only made in case chemicals were tested up to at least 25 % test concentration and if not a single positive reaction was recorded at this or higher concentrations. For LLNA data, the maximal test concentration was required to be

[1] The 2o3 had originally been described both using KeratinoSens™ and LuSens test methods interchangeably to address the key event keratinocyte activation with similar predictive capacity (Bauch et al., 2012). Equivalent predictive capacity using both tests to address the key event keratinocyte activation had later been shown by Urbisch et al. (2015). The LuSens is a me-too assay, which had been developed based on essential test method components as described in OECD Guidance Document No 213 (OECD, 2015 and Ramirez et al., 2014). It has been validated according to performance standards of OECD TG 442D as laid down in Guidance document 213 (OECD, 2015, Ramirez et al., 2016).

Likewise, the 2o3 had originally been described both using the h-CLAT, mMUSST and U-937 test methods interchangeably to address the key event dendritic cell activation with similar predictive capacity (Urbisch et al., 2015).

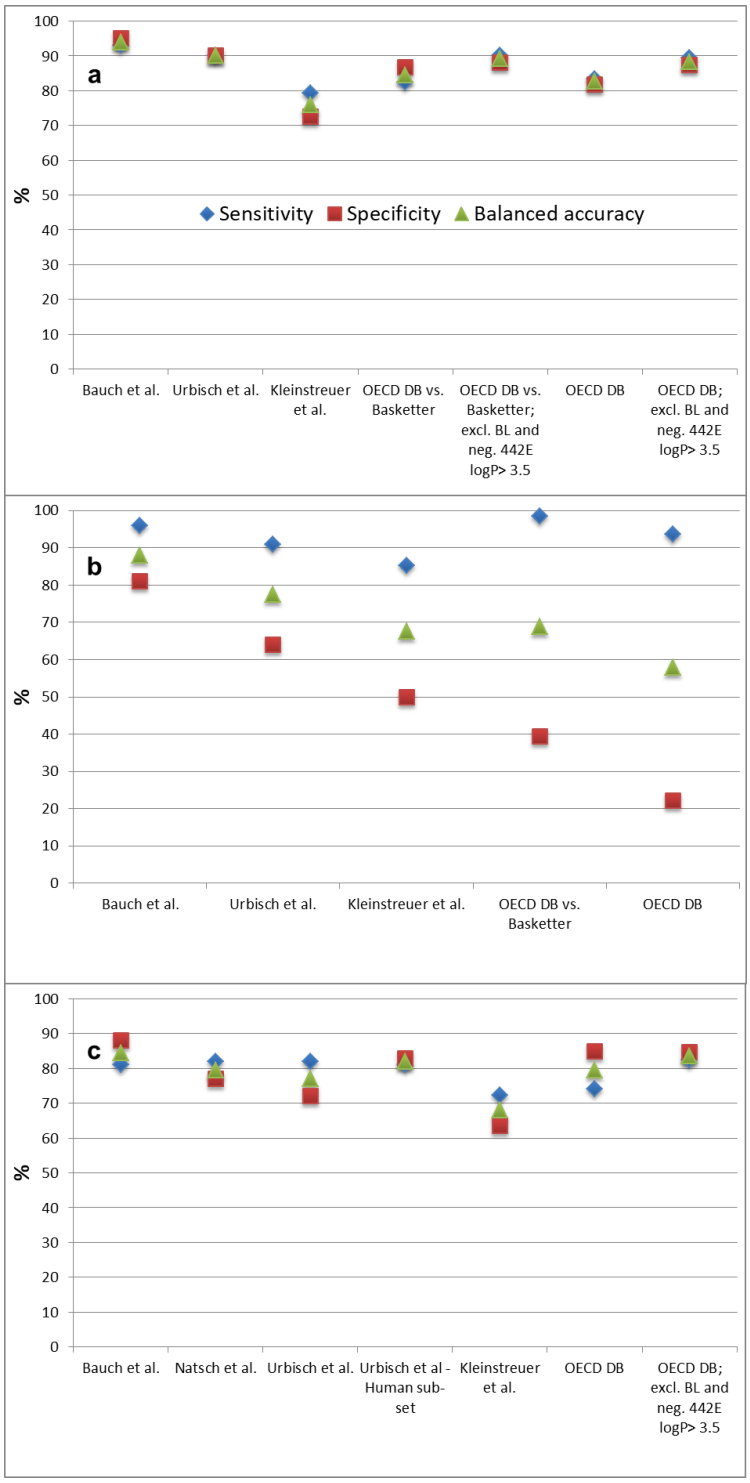

Fig. 2: Predictivity of a) 2o3 DA vs. human, b) LLNA vs. human and c) 2o3 DA vs. LLNA reference data in different evaluations
The different evaluations (x-axis) are ordered in the order of the publication date of the differently curated data sets with the reference data (please refer to text for details). Blue diamonds: Sensitivity, Red squares: Specificity; Green triangles: Balanced accuracy

at least 50%, and in case of multiple test results, all tests needed to be negative in order to accept a negative call. These stringent requirements led to a much smaller database, especially for human or LLNA negative calls. This database of highly curated data may now be considered the "Gold List", but it was not compiled to include congruent data from different *in vivo* sources, differently to the "Silver List" discussed above. Thus, evaluation of the performance of the LLNA to predict the human outcome in this database again indicates a high sensitivity (94%, n = 47) of the LLNA for the human sensitization risk but a poor specificity (22%), although the latter is based on a low number of chemicals (n = 9). The LLNA data curated by the OECD can also be compared with the human data compilation from Basketter *et al.* (Basketter et al., 2014) giving a higher number of comparisons (n=96), in this case a high sensitivity (99%) and better, but still poor, specificity is observed for the LLNA vs. human data (39%, Table 1).

In summary, the curation of data over time led a continuously reduced specificity of the LLNA vs. human data as summarized in Table 1 and Figure 2. This appears at least partly due to stringent requirements to accept negative results, which is partly based on guidance in OECD TG 429 (i.e. to test chemicals up to 100% concentration) but which actually had never been validated (Kolle et al., 2020), and this has to be kept in mind when evaluating an integrated approach or DA for skin sensitization.

3 Predictivity of 2o3 DA vs. human data

Table 2 summarizes the predictivity of the 2o3 DA approach vs. human data evaluated on the different datasets summarized above. The initial analysis of Bauch *et al.* (Bauch et al., 2012), which actually led to the definition of the 2o3 DA, had found a high predictivity (94% balanced accuracy, BA), which was above the also high BA of the LLNA (88%) for the presented dataset. This good predictivity was confirmed by the analysis of Urbisch *et al.* (Urbisch et al., 2015), with a BA of 90%. This latter analysis included the Bauch *et al.* subset but extended the list from 50 to 101 chemicals. Again, the 2o3 DA was more predictive as compared to the LLNA (BA 77.5%), which in this analysis had a lower specificity. For the Cosmetics Europe database, Kleinstreuer et al. (2018) found a lower predictivity for the 2o3 DA (Sensitivity 79.3%, Specificity 72.5%, BA 75.9%; n = 127), but this was again higher as compared to the LLNA (BA 67.6%).

Interestingly, on the reduced subset of chemicals from the Cosmetics Europe database which was retained after data review in the final OECD database (n = 104), but again comparing vs. the Cosmetics Europe human data (Basketter et al., 2014), the 2o3 DA has a higher BA (85%) than on the full Cosmetics Europe database, and again a higher predictivity as compared to the LLNA (BA 69%).

During development of the Defined Approach guideline, the analysis of borderline outcomes from *in vitro* data (Gabbert et al., 2020; Leontaridou et al., 2017) was introduced to assess certainty of the outcome as described elsewhere (Kolle et al., in press). Excluding 15 borderline calls according to this analysis, the BA for the Cosmetics Europe human data retained in OECD database raises to 89%. In a strict interpretation of the 442E guideline, negative calls for chemicals with log P > 3.5 in the h-CLAT are not accepted as negative (but rated "inconclusive"). Translating this limitation into the 2o3 as currently implemented also in the OECD DASS guideline leads to three more inconclusive chemicals but does not improve predictivity for human data (BA 88%).

Finally, and most importantly, predictivity was also assessed vs. the curated OECD human dataset (OECD, 2021a), although this set is significantly smaller and sensitizers are largely overrepresented (n=54 sensitizers; n=11 non-sensitizers). This dataset may be viewed as rather too small for firm conclusions to be drawn. However, it should be noted that the number of human non-sensitizers in the data set used to evaluate the DAs was considerably higher than in the data set used to validate the LLNA (containing 68 human sensitizers and 6 human non-sensitizers (Haneke et al., 2001; ICCVAM, 1999). Most interestingly the predictivity values for the 2o3 DA are almost identical to those when assessing against the Cosmetics Europe human data, although those underwent much less curation. This is the case for all three analyses (all chemicals: BA = 83%; borderlines excluded: BA = 88%; h-CLAT negatives log P > 3.5 additionally excluded: BA = 88%). Thus, even if this dataset is relatively small – which may raise criticism concerning statistical power - the fact that it completely confirms results on the previous evaluation vs. the Cosmetics Europe / Basketter *et al.* human dataset should give this analysis a high credibility. Thus, analysis on larger number of less curated data gave the same outcome as analysis of lower number of more curated data – this helped to overcome the criticism that either data are not curated or that the numbers are too low, as it is highly unlikely that both types of analysis by chance came to congruent conclusions.

Interestingly, when looking at the global picture of Table 2 and Figure 2a, the values, for a given evaluation, are always very similar for sensitivity and specificity vs. human data. This indicates that 2o3 DA offers a very balanced predictivity of the human sensitization hazard, and this seems to be superior to the situation summarized in Table 1 for the LLNA, with a predictivity which tends to be increasingly skewed towards sensitivity.

Tab. 2: Predictivity of the 2o3 DA vs. human reference data in different studies

	Sensitivity [%]	Specificity [%]	Balanced Accuracy [%]	N
Bauch et al., 2012[a]	93	95	94	50
Urbisch et al., 2015	90	90	90	101
Kleinstreuer et al. 2018	79.3	72.5	75.9	127
OECD database vs Basketter et al., 2014[b]	82	87	85	104
OECD database vs Basketter et al., 2014; excluding borderlines [b,c)]	89	89	89	89 (15 inconcl.)
OECD database vs Basketter et al., 2014; excluding borderlines and h-CLAT negatives with log P > 3.5[b-d]	90	88	89	86 (18 inconcl.)
OECD database (all chemicals with human data)	83	82	83	65
OECD database (all chemicals with human data) excluding borderlines[c]	89	88	88	55 (10 inconcl.)
OECD database (all chemicals with human data) excluding borderlines and h-CLAT negatives with log P > 3.5[d]	89	88	88	55 (10 inconcl.)

[a] U-937 test instead of h-CLAT

[b] The subset (n = 104) of the Kleinstreuer et al., 2018 data set (n = 127) remaining after the curation process for LLNA data of the OECD expert group.

[c] Chemicals with predictions in the statistically derived borderline range around the prediction threshold (Kolle et al., submitted) are considered inconclusive. Two congruent, conclusive results are needed for a conclusive 2o3 prediction.

[d] Negative h-CLAT results for chemicals with log P > 3.5 are considered inconclusive. Two concordant and conclusive negative results from 442D and DPRA are needed for a conclusive negative 2o3 prediction for these chemicals according current OECD guideline draft (OECD, 2021a).

4 Predictivity of 2o3 DA vs LLNA data

Since the LLNA has been the method of choice for the sensitization endpoint of industrial chemicals in the last two decades, and since a DA should fully replace the LLNA as stand-alone method for hazard identification, predictivity for the LLNA has been emphasized in most studies and evaluations, and it is of prime concern to regulators currently assessing chemicals based on LLNA results. Of course – with the limitations of the LLNA in the different datasets for predicting human sensitization (Table 1), a perfect predictivity for the LLNA cannot and should not be expected, as then a NAM might replicate the LLNA including all of its identified weaknesses.

Predictivity of the 2o3 DA vs. the LLNA is summarized in Table 3 and was high (Sens. = 81%; Spec. = 88; %; BA = 84.5%, n= 54) in the first assessment of the 2o3 DA (Bauch et al., 2012). This is not surprising, as this dataset had overall a good alignment between LLNA and human data as discussed above. A decent predictivity (Sens. = 82%; Spec. = 77; %; BA = 79.5%, n= 145) on a much larger dataset was also reported in the LLNA focused evaluation (Natsch et al., 2013). Sensitivity remained similar while specificity decreased in the larger set by (Urbisch et al., 2015) (Sens. = 82%; Spec. = 72; %; BA = 77%, n= 180). The poorest predictivity for LLNA was reported on the Cosmetics Europe database (Sens. = 72.3%; Spec. = 63.6; %; BA = 68.5%, n= 127), which had been put together for the assessment vs. human data. After OECD data review and on the larger database put together by the OECD expert group, a higher predictivity (Sens. =74%; Spec. = 85; %; BA = 79%, n= 168) was again found also for LLNA data. This is further increased when excluding the 29 chemicals with borderline results (Sens. = 79%; Spec. = 85; %; BA = 82%, n= 139, 29 inconclusive calls in the DA). By further excluding chemicals with a log P > 3.5 rated negative in the h-CLAT, sensitivity is further increased (Sens. = 82%; Spec. = 85; %; BA = 84%, n= 134, 34 inconclusive calls in the DA).

It is important to note, that the limitation in sensitivity of the h-CLAT for chemicals with a log P > 3.5 had been found when evaluating predictivity against LLNA data only (Takenouchi et al., 2013). As shown in the Supporting document to the DA guideline (Annex 6) (OECD, 2021e)and as will be reported in detail elsewhere, actually the LLNA has a high false discovery rate (FDR) vs. human data for chemicals in this physiochemical range – thus it appears that the LLNA generates an increased rate of false-positives for lipophilic chemicals, rather than the h-CLAT (and other NAMs) generating a particularly high rate of false-negatives. Thus, the limitation introduced into the OECD guideline 442E, and now also translated into the DA guideline, specifically optimizes predictivity for LLNA data, trying to replicate a potential weak spot of the LLNA. As shown in Table 2, this modification does not improve predictivity for human data and therefore it is questionable whether such a limitation just duplicating mistakes of the animal test should be carried along rather than being corrected based on learnings from analyzing human data.

Tab. 3: Predictivity of the 2o3 DA vs. LLNA reference data in different studies

	Sensitivity [%]	Specificity [%]	Balanced Accuracy [%]	N
Bauch et al., (Bauch et al., 2012)	81	88	84.5	54
Natsch *et al. (*Natsch et al., 2013)[a]	82	77	79.5	145
Urbisch et al. (Urbisch et al., 2015) – all data	82	72	77	180
Urbisch et al. (Urbisch et al., 2015) - Human sub-set[b]	81	83	82	103
Kleinstreuer et al. (Kleinstreuer et al., 2018)	72.3	63.6	68	127
OECD database (all chemicals with LLNA data)	74	85	79	168
OECD database (all chemicals with LLNA data) excluding borderlines[c]	79	85	82	139 (29 inconcl.)
OECD database (all chemicals with LLNA data) excluding borderlines and h-CLAT negatives with log P > 3.5[d]	82	85	84	134 (34 inconcl.)

[a] U-937 test instead of h-CLAT

[b] Evaluation vs. those chemicals for which human data were available.

[c] Chemicals with predictions in the statistically derived borderline range around the prediction threshold (Kolle et al., submitted) are considered inconclusive. Two concordant, conclusive results are needed for a conclusive 2o3 prediction.

[d] Negative h-CLAT results for chemicals with log P > 3.5 are considered inconclusive. Two concordant and conclusive negative results from 442D and DPRA are needed for a conclusive negative 2o3 prediction for these chemicals according current OECD guideline draft (http://www.oecd.org/chemicalsafety/testing/draft-guideline-for-defined-approaches-for-skin-sensitisation.pdf).

5 Conclusions

The predictive performances of NAMs and DA are key criteria for their regulatory acceptance and hence the replacement of animal tests. The predictive performance depends, however, not only on the performance of the method but also on the quality and comprehensiveness of the reference data (Kolle et al., 2019).

While evaluating the DA for skin sensitization, the OECD conducted the probably most in-depth curation effort of reference data ever, furthering an already thorough analysis made before (Hoffmann et al., 2018). Predicting skin sensitization, we have the unique possibility to not only compare to animal (LLNA) data but also to human data. Traditionally and following the example of other areas of toxicology, where human data are sparse, there is a tendency to attribute more weight to the animal data, which in this case is the LLNA. This may be due to (i) two decades of regulatory practice (ii) the LLNA being the current OECD standard, (iii) the LLNA providing standardized data (unlike human data), and (iv) due to the high sensitivity of the LLNA and overlooking low specificity. The very simple prediction model (a single value – the estimated

concentration leading to a stimulation index of three) might be an additional reason for making LLNA data attractive as a sole reference for a validation. However, since the LLNA does not measure skin sensitization and resulting contact allergy – but only a surrogate (cell proliferation as an important step in the induction phase), there is an intrinsic risk in focusing on LLNA solely, as potential limitations and blind spots of the LLNA may be translated to animal-free testing, such as the log P > 3.5 limitation for the h-CLAT.

The early observation that the 2o3 DA (and other DA such as the ITS (Kleinstreuer et al., 2018)) is abler to predict the human sensitization outcome than the LLNA had been questioned, and a detailed data review was undertaken to scrutinize this finding within the OECD expert group. Interestingly, using these more refined data to evaluate the LLNA, an ever-decreasing specificity of the LLNA vs. human data (Figure 2) was found, which may be due to stringent, but not validated, criteria to accept negative results (Kolle et al., 2020). On the other hand, for the 2o3 DA, the high balanced accuracy (based on a balanced sensitivity and specificity) vs. human data could be confirmed by all analyses (Figure 2). This triangular evaluation included animal and human data and utilized different, partly overlapping, datasets. It is obviously superior to validation solely based on LLNA data. This triangular analysis of predictivity should build trust in using the 2o3 DA in regulatory settings and encourage the analysis of other DA by the same approach. The triangular evaluation of predictivity underlies the notion that animal data from a single test alone should not always be the gold standard when evaluating alternatives, but that a more holistic view including more, more refined, and more relevant reference data shall be preferred.

The triangular approach including human data is only possible, for the few endpoints where human data are available. For other endpoints, it will still be valuable to collect multiple data on the individual chemicals, ideally from two different animal tests. This was for example recently done to evaluate *in vitro* models for androgen antagonists (Gray et al., 2020). If such data are available, reference lists can be constructed in two ways: (i) either based on congruent calls from multiple sources, as we had done with the "Silver List" – predictivity of NAMs vs. such a consensus list then provides for a best estimate of true predictivity for the endpoint of interest. (ii) On the other hand, a similar triangular approach using a NAM compared to two animal tests will indicate to which extend prediction uncertainty exist in the data by judging how well the two animal tests predict each other and compare this uncertainty to the prediction of the endpoint of interest by the NAM. This approach was used in the recent study by Gray *et al.*, in that case this evaluation could show that the *in vitro* approach does not yet offer sufficient predictivity. In any way, carefully curating and referencing the original data is a key step to improve such evaluations, and in this regard, the recent OECD data curation effort represents a further step forward.

References

Basketter, D. A., Alepee, N., Ashikaga, T., Barroso, J. et al. (2014). Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis* 25, 11-21. doi:10.1097/der.0000000000000003

Basketter, D. A., Lea, L. J., Cooper, K., Stocks, J. et al. (1999). Threshold for classification as a skin sensitizer in the local lymph node assay: A statistical evaluation. *Food and Chemical Toxicology* 37**,** 1167-1174. doi:10.1016/s0278-6915(99)00112-x

Bauch, C., Kolle, S. N., Ramirez, T., Eltze, T. et al. (2012). Putting the parts together: Combining in vitro methods to test for skin sensitizing potentials. *Regul Toxicol Pharmacol* 63**,** 489-504. doi:10.1016/j.yrtph.2012.05.013; Corrigendum: doi:10.1016/j.yrtph.2012.08.014

Casati, S., Aeby, P., Kimber, I., Maxwell, G. et al. (2009). Selection of chemicals for the development and evaluation of in vitro methods for skin sensitisation testing. *Altern Lab Anim* 37**,** 305-12. doi:10.1177/026119290903700313

Dimitrov, S. D., Low, L. K., Patlewicz, G. Y., Kern, P. S. et al. (2005). Skin sensitization: Modeling based on skin metabolism simulation and formation of protein conjugates. *International Journal of Toxicology* 24**,** 189-204. doi:10.1080/10915810591000631

EC - European Commission (2009). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. *Official Journal of the European Union* https://eur-lex.europa.eu/eli/reg/2009/1223/oj/eng.

EC (2013). Full EU ban on animal testing for cosmetics enters into force. *https://ec.europa.eu/commission/presscorner/detail/en/IP_13_210*.

Emter, R., Ellis, G. and Natsch, A. (2010). Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. *Toxicol. Appl. Pharmacol.* 245**,** 281-290. doi:10.1016/j.taap.2010.03.009

Ezendam, J., Braakhuis, H. M. and Vandebriel, R. J. (2016). State of the art in non-animal approaches for skin sensitization testing: from individual test methods towards testing strategies. *Arch Toxicol*. doi:10.1007/s00204-016-1842-4

Gabbert, S., Mathea, M., Kolle, S. N. and Landsiedel, R. (2020). Accounting for Precision Uncertainty of Toxicity Testing: Methods to Define Borderline Ranges and Implications for Hazard Assessment of Chemicals. *Risk Anal*. doi:10.1111/risa.13648

Gray, L. E., Jr., Furr, J. R., Lambright, C. S., Evans, N. et al. (2020). Quantification of the uncertainties in extrapolating from in vitro androgen receptor antagonism to in vivo hershberger assay endpoints and adverse reproductive development in male rats. *Toxicological Sciences* 176**,** 297-311. doi:10.1093/toxsci/kfaa067

Haneke, K. E., Tice, R. R., Carson, B. L., Margolin, B. H. et al. (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**,** 274-86. doi:10.1006/rtph.2001.1498

Hoffmann, S., Kleinstreuer, N., Alepee, N., Allen, D. et al. (2018). Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database(). *Crit. Rev. Toxicol.* in press**,** 1-15. doi:10.1080/10408444.2018.1429385

ICCVAM (1999). The Murine Local Lymph Node Assay: A Test Method for Assessing the Allergic Contact Dermatitis Potential of Chemicals/Compounds. The Results of an Independent Peer Review Evaluation Coordinated by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and the National Toxicology Program Center for the Evaluation of Alternative Toxicological Methods (NICEATM). *NIH Publication No. 99-4494*. doi:https://ntp.niehs.nih.gov/iccvam/docs/immunotox_docs/llna/llnarep.pdf

ICCVAM (2009). Recommended Performance Standards: Murine Local Lymph Node Assay *ICCVAM report* NIH Publication No. 09-7357

Jaworska, J., Dancik, Y., Kern, P., Gerberick, F. et al. (2013). Bayesian integrated testing strategy to assess skin sensitization potency: From theory to practice. *J. Appl. Toxicol.* 33**,** 1353-1364. doi:10.1002/jat.2869

Johansson, H., Albrekt, A. S., Borrebaeck, C. A. and Lindstedt, M. (2013). The GARD assay for assessment of chemical skin sensitizers. *Toxicol In Vitro* 27**,** 1163-9. doi:10.1016/j.tiv.2012.05.019

Jowsey, I. R., Basketter, D. A., Westmoreland, C. and Kimber, I. (2006). A future approach to measuring relative skin sensitising potency: A proposal. *Journal of Applied Toxicology* 26**,** 341-350. doi:10.1002/jat.1146

Kleinstreuer, N. C., Hoffmann, S., Alepee, N., Allen, D. et al. (2018). Non-animal methods to predict skin sensitization (II): an assessment of defined approaches (*). *Crit Rev Toxicol***,** 1-16. doi:10.1080/10408444.2018.1429386

Kolle, S. N., Hill, E., Raabe, H., Landsiedel, R. et al. (2019). Regarding the references for reference chemicals of alternative methods. *Toxicology in Vitro* 57**,** 48-53. doi:10.1016/j.tiv.2019.02.007

Kolle, S. N., Landsiedel, R. and Natsch, A. (2020). Replacing the refinement for skin sensitization testing: Considerations to the implementation of adverse outcome pathway (AOP)-based defined approaches (DA) in OECD guidelines. *Regul Toxicol Pharmacol* 115**,** 104713. doi:10.1016/j.yrtph.2020.104713

Kolle, S. N., Mathea, M., Natsch A., and Landsiedel R. (in press). Assessing Experimental Uncertainty in Defined Approaches: Borderline Ranges for In Chemico and In Vitro Skin Sensitization Methods Determined from Ring Trial Data. *Appl In Vitro Toxicol*. doi:10.1089/aivt.2021.0003

Lacroix, G., Koch, W., Ritter, D., Gutleb, A. C. et al. (2018). Air-Liquid Interface In Vitro Models for Respiratory Toxicology Research: Consensus Workshop and Recommendations. *Appl In Vitro Toxicol* 4**,** 91-106. doi:10.1089/aivt.2017.0034

Leontaridou, M., Urbisch, D., Kolle, S. N., Ott, K. et al. (2017). The borderline range of toxicological methods: Quantification and implications for evaluating precision. *ALTEX 34*, 525-538. doi:10.14573/altex.1606271

Natsch, A., Ryan, C. A., Foertsch, L., Emter, R. et al. (2013). A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. *J Appl Toxicol* 33**,** 1337-52. doi:10.1002/jat.2868

Nukada, Y., Miyazawa, M., Kazutoshi, S., Sakaguchi, H. et al. (2013). Data integration of non-animal tests for the development of a test battery to predict the skin sensitizing potential and potency of chemicals. *Toxicology in Vitro* 27**,** 609-618. doi:10.1016/j.tiv.2012.11.006

OECD (2018a). In vitro skin sensitisation assays addressing the key event on activation of dendritic cells on the adverse outcome pathway for skin sensitisation. *OECD testing guidelines* 442e. doi:10.1787/9789264264359-en

OECD (2018b). Test No. 442D: In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method. *OECD Guidelines for the Testing of Chemicals* No. 442D. doi:10.1787/9789264229822-en

OECD (2020). In Chemico Skin Sensitisation Assays addressing the Adverse Outcome Pathway, key event on covalent binding to proteins. *OECD testing guidelines* 442c. doi:10.1787/9789264229709-en

OECD (2021a). Guideline No. 497: Defined Approaches on Skin Sensitisation. *OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing, Paris* https://www.oecd-ilibrary.org/environment/guideline-no-497-defined-approaches-on-skin-sensitisation_b92879a4-en. doi:10.1177/026119290703500311

OECD (2021b). Series on Testing and Assessment No. 336: Supporting document to the Guideline (GL) on Defined Approaches (DAs) for Skin Sensitisation- Annex 2. Organisation for Economic Cooperation and Development, Paris.

OECD (2021c). Series on Testing and Assessment No. 336: Supporting document to the Guideline (GL) on Defined Approaches (DAs) for Skin Sensitisation- Annex 3. Organisation for Economic Cooperation and Development, Paris.

OECD (2021d). Series on Testing and Assessment No. 336: Supporting document to the Guideline (GL) on Defined Approaches (DAs) for Skin Sensitisation- Annex 4. Organisation for Economic Cooperation and Development, Paris.

OECD (2021e). Series on Testing and Assessment No. 336: Supporting document to the Guideline (GL) on Defined Approaches (DAs) for Skin Sensitisation- Annex 6. Organisation for Economic Cooperation and Development, Paris.

Rovida, C., Basketter, D., Casati, S., de Silva, O. et al. (2007). Management of an integrated project (Sens-it-iv) to develop in vitro tests to assess sensitisation. *Altern. Lab. Anim.* 35**,** 317-22. doi:10.1177/026119290703500311

Takenouchi, O., Miyazawa, M., Saito, K., Ashikaga, T. et al. (2013). Predictive performance of the human Cell Line Activation Test (h-CLAT) for lipophilic chemicals with high octanol-water partition coefficients. *J Toxicol Sci* 38**,** 599-609. doi:10.2131/jts.38.599

Urbisch, D., Mehling, A., Guth, K., Ramirez, T. et al. (2015). Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul. Toxicol. Pharmacol.* 71**,** 337-51. doi:10.1016/j.yrtph.2014.12.008

Acknowledgements

This work was entirely funded by Givaudan Schweiz SA and BASF SE. This work heavily relies on the work done by the large international expert group which worked on the OECD Guideline on DA for skin sensitization. The authors A.N. and S.N.K. were members and contributors to this expert group. The tremendous work on the database and data review for LLNA and human reference data made by sub-groups of that working group is published elsewhere and is highly acknowledged here.