

FORUM

Skin Sensitization: Challenging the Conventional Thinking—A Case Against 2 Out of 3 as Integrated Testing Strategy

Henrik Johansson¹ and Robin Gradin²

SenzaGen AB, Medicon Village, 22381 Lund, Sweden

¹To whom correspondence should be addressed at SenzaGen AB, Building 406, Medicon Village, Scheelevägen 2, 22381 Lund, Sweden. Fax: +46-704-492724. E-mail: henrik.johansson@senzagen.com.

²To whom correspondence should be addressed. Fax: +46-767-760123. E-mail: robin.gradin@senzagen.com.

Key words: skin sensitization; alternative methods; in vitro; integrated testing strategies.

BACKGROUND

Recent years have seen a surge in development of alternative methods for assessment of skin sensitizers. This has been in order to meet the regulatory and industrial demand for accurate safety assessment without the use of animal experimentation, and to address societal expectations. Today, 3 such assays are formally validated and have achieved OECD Guideline status (EC, 2013, 2014, 2015). However, partly due to insufficient predictive capacity, none of the validated assays have received a recommendation for use as a stand-alone method. In addition, there is a view that accurate hazard identification must be based upon an assessment of several key events in the relevant Adverse Outcome Pathway (AOP).

For these reasons it is proposed that tests are used in combinations, forming so-called Integrated Testing Strategies (ITSs), to address deficiencies in the performance of individual methods and to incorporate end points that reflect several key events in the AOP (Hartung *et al.*, 2013; Jaworska and Hoffmann, 2010; Rovida *et al.*, 2015). Although such ITSs may be configured in a number of ways, one heavily advocated approach relies on the majority vote of individual assays, often referred to as a 2 out of 3 ITS. That is, the identification of hazard is dependent upon the chemical eliciting a positive response in at least 2 of 3 test methods. This approach is based on the view that this will provide an increased weight of evidence and thereby greater confidence in hazard characterization and risk assessment.

Here, we present arguments opposing the claimed benefits of 2 out of 3 ITSs by examining the theoretical probability propagation of combined outcomes, simulated predictions generated

by hypothetical assays, and actual data obtained from recently published literature. We argue that the added value of combining the top-performing assay with other assays with less predictive accuracy is overestimated, if present, and may actually be detrimental.

THEORETICAL PROBABILITY PROPAGATION OF COMBINED OUTCOMES

With the assumption that predictions made by assays are independent random variables X_1 , X_2 , and X_3 with outcomes x_1 , x_2 , and x_3 , the probability that any given outcome is obtained is described by Equation (1).

$$P(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } X_3 = x_3) = P(X_1 = x_1) * P(X_2 = x_2) * P(X_3 = x_3) \quad (1)$$

Where $x_i \in \{\text{Positive, Negative}\}$, for $i = 1, 2, 3$.

Further, the probability that a prediction is correct is here assumed to be defined by the predictive accuracy of the assay.

$$P(X_i = x | \text{True Class} = x) = \text{Accuracy}_i \quad (2)$$

for $i = 1, 2, 3$ and $x \in \{\text{Positive, Negative}\}$

The prediction made by a 2 out of 3 ITS is based on voting, i.e. at least 2 predictions giving the same outcome gives the final prediction. A correct prediction is thus made if at least 2 assays in the ITS give correct predictions.

$$\text{Prediction}_{\text{ITS}} = \begin{cases} \text{correct if } \sum [X_i = \text{True Class}] \geq 2 \\ \text{wrong if } \sum [X_i = \text{True Class}] < 2 \end{cases}, \text{ for } i = 1, 2, 3 \quad (3)$$

The probability of a correct prediction by the ITS can thus be calculated based on the accuracy of the individual assays and the assumption of independence, by summarizing the probabilities for each of the 4 possible outcomes that would result in a correct classification.

$$\begin{aligned} P(\text{Prediction}_{\text{ITS}} = \text{correct}) &= P(X_1 = c) * P(X_2 = c) * P(X_3 = c) \\ &\quad + P(X_1 = c) * P(X_2 = c) * P(X_3 = w) \\ &\quad + P(X_1 = c) * P(X_2 = w) * P(X_3 = c) \\ &\quad + P(X_1 = w) * P(X_2 = c) * P(X_3 = c) \end{aligned} \quad (4)$$

Where *c* and *w* denotes correct and wrong predictions respectively, and $P(X_i = c)$ and $P(X_i = w)$ are shorter representations for $P(X_i = x | \text{True Class} = x)$ and $P(X_i \neq x | \text{True Class} = x)$ where $x \in \{\text{Positive}, \text{Negative}\}$.

SIMULATIONS OF A HYPOTHETICAL ITS

Using Equation (4) above, the expected theoretical predictive performance of any 2 out of 3 ITS can be evaluated if estimations of the predictive accuracies of the ITS constituents are available. It is easily demonstrated that if 3 equally good assay constituents are used, there is indeed an added value in combining the outputs into a 2 out of 3 ITS. Assuming an accuracy for each assay of 70%, the expected accuracy of an ITS would be 78.4%. However, if the predictive value of one of the assays is increased to an accuracy of 80%, the relative impact and added value of the remaining 2 assays is lowered, as the expected ITS accuracy is now calculated to be 82.6%. Taking this reasoning even further, if the accuracy of the first assay is increased to 90%, the expected accuracy of the ITS is calculated to 86.8%. Here, the ITS actually performs worse than the top performing assay constituent, and the consequence of combining this assay with less predictive assays is in fact unfavorable to the overall test accuracy. For the sake of completeness, it is of obvious interest to observe the expected ITS performance as a function of all possible combinations of ITS constituent performances. Such an exercise was carried out, with results demonstrated in Figure 1. Assuming a fixed predictive accuracy of the first assay at 90%, the ITS predictive accuracy is plotted over 2 varying complementary assays, as calculated by Equation (4). From Figure 1 it is clear that, assuming a high performing assay is available, the requirements of the complementary assay accuracies are quite stringent in order to achieve a substantial added value.

CONCORDANCE WITH PUBLISHED LITERATURE

Although the situations described above are based on theoretical estimations of expected ITS performances, it is of interest to see how these estimations compare with real data. In the case of chemical sensitization assessment, such data are readily available. Here, we will be studying 2 different datasets from recent literature. The first dataset, published by Natsch et al. (2013), includes predictions of 145 chemicals using 3 different *in vitro* assays. The second dataset, published by Urbisch et al. (2015), contains data of 213 chemicals from up to 6 different assays. The authors propose a 2 out of 3 ITS using 3 of these

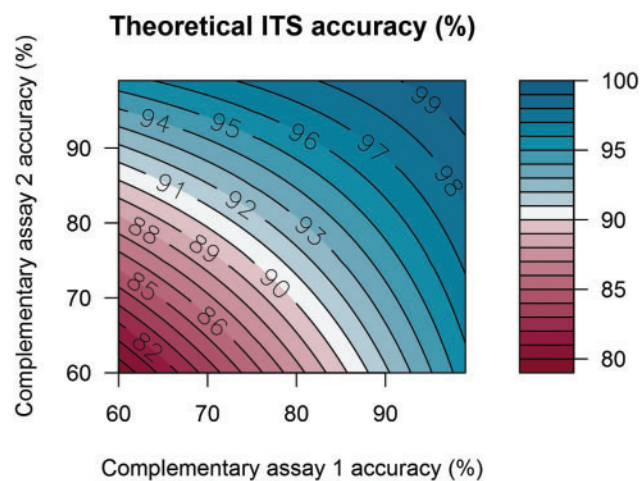


Figure 1. Theoretical ITS performance as a function of the accuracies of 3 ITS assay constituents. The first assay performance is fixed at an accuracy of 90%. The predictive accuracy of 2 complementary assays vary along the x- and y-axis, and the ITS performance is represented by the graph contours.

Table 1. Predictive Performances of Published Data

Assay ID	Natch et al. (2013)	Urbisch et al. (2015)
	Accuracy (%)	Accuracy (%)
DPRA	80	75
Keratinosens	77	73
MUSST (U-Sens)	71	—
h-CLAT	—	76
Observed (reported) ITS	81	79
Theoretical ITS	86	84

assays, which covers data for 180 chemicals. The details of both of these datasets are summarized in Table 1. The predictive performances of each individual assay, along with the performance of proposed 2 out of 3 ITSs, both as observed by the respective authors and as estimated by Equation (4), are presented.

The authors of both papers report an increase in accuracy when combining the outputs from 3 different assays, compared with using each assay as a stand-alone method. However, there is a consistent discrepancy between the ITS performance as observed by the respective authors, and what would be expected from the theoretical calculations using Equation (4). Why is that? Recall that equations (1–4) assume that the outcome of each random variable is independent. In reality, this is rarely the case. On the contrary, it is evident from available data that a number of chemicals are inherently hard to accurately predict. This may be due to physical characteristics of the chemical, such as poor solubility or other incompatibility with test systems. False positives are also prevalent among chemical irritants. False predictions may also be due to chemical domain issues, meaning that prediction models have been designed to recognize chemicals of certain molecular structures, but are unsuitable to make predictions on chemicals with vastly different molecular structures. Last, it is evident that weak sensitizers are harder to predict accurately, compared with their stronger counterparts. Taken together, with currently available assays, there is a negative bias in ITSs towards false predictions due to tendencies to misclassify troublesome chemicals with

more than 1 assay. This means that the added value of a 2 out of 3 ITS, compared with only using the top performing assay, is actually even smaller compared with what was expected from the theoretical discussion above, and as illustrated in Figure 1.

CONCLUSIONS

The idea of using the combined output of several assays in 2 out of 3 ITSs has become the most prevailing strategy to account for insufficient predictive performances among currently validated assays for assessment of skin sensitization (Ezendam *et al.*, 2016). Under the right circumstances, the idea is not flawed. Indeed, if a number of assays perform equally well, exhibiting comparable accuracies at moderate levels, the predictive value does increase when combining the outputs, allowing for more accurate safety assessment. However, if a high performing assay is available, the added value of combining it with less accurate assays diminishes and even disappears. In the reality of assessment of chemical sensitizers, the predictive performance of available assays does vary. Thus, we argue that the added value of constructing 2 out of 3 ITSs, compared with only using the top performing assay, is overestimated in the literature, among regulatory bodies, and in the scientific community.

Another argument for the value of ITSs has been that accurate safety assessment must rely on a number of assays that each monitor separate key events of the AOP. The reasoning has been that risks of losing valuable information when moving from a systemic response, as achieved by gold standard *in vivo* assays, is mitigated when using a range of *in vitro* methods, each contributing information from different key events. Again, the idea is not flawed. However, in the case of a 2 out of 3 ITS, information regarding the target key event of a certain assay is in no way passed on to the final prediction model, which consists solely of a majority vote. Thus, the desired positive effect is not achieved.

It should, however, be acknowledged that these arguments do not object to the concept of ITSs *per se*, but are applicable to ITSs based on a majority vote prediction model. More sophisticated ITSs have been proposed, utilizing eg, machine learning algorithms (Luechtefeld *et al.*, 2015; Tsujita-Inoue *et al.*, 2015), or weight of evidence assessment based on Bayesian statistics (Jaworska *et al.*, 2015).

However, such proposed ITSs are still very much in the conceptual phases, and are far from regulatory acceptance or widespread use. In addition, many variable inputs such as *in silico* parameters of chemico-physical properties require an expertise that is becoming more widespread among strategy developers, but is rarely seen among end users. Finally, the economic burden of using multiple *in vitro* endpoints together with the suggested *in silico* modeling expertise is substantial.

Thus, until feasible options are available, the added value of combining assay endpoints should be carefully evaluated and cost-benefit estimations should be taken into account. In situations where one sufficiently accurate method is available, the option of relying only on the top performing assay should not be strictly dismissed.

REFERENCES

- EC. (2013). EURL ECVAM Recommendation on the Direct Peptide Reactivity Assay (DPRA) for Skin Sensitisation Testing. <https://eurl-ecvam.jrc.ec.europa.eu/eurl-ecvam-recommendations/eurl-ecvam-recommendation-on-the-direct-peptide-reactivity-assay-dpra>. Accessed June 12, 2017.
- EC. (2014). EURL ECVAM Recommendation on the KeratinoSens™ assay for skin sensitization testing. <https://eurl-ecvam.jrc.ec.europa.eu/eurl-ecvam-recommendations/recommendation-keratinosens-skin-sensitisation>. Accessed June 12, 2017.
- EC. (2015). EURL ECVAM Recommendation on the human Cell Line Activation Test (h-CLAT) for skin sensitisation testing. <https://eurl-ecvam.jrc.ec.europa.eu/eurl-ecvam-recommendations/eurl-ecvam-recommendation-on-the-human-cell-line-activation-test-h-clat-for-skin-sensitisation-testing>. Accessed June 12, 2017.
- 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.* **90**, 2861–2883.
- Hartung, T., Luechtefeld, T., Maertens, A., and Kleensang, A. (2013). Integrated testing strategies for safety assessments. *Altex* **30**, 3–18.
- Jaworska, J., and Hoffmann, S. (2010). Integrated Testing Strategy (ITS) - Opportunities to better use existing data and guide future testing in toxicology. *Altex* **27**, 231–242.
- Jaworska, J. S., Natsch, A., Ryan, C., Strickland, J., Ashikaga, T., and Miyazawa, M. (2015). Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: A decision support system for quantitative weight of evidence and adaptive testing strategy. *Arch. Toxicol.* **89**, 2355–2383.
- Luechtefeld, T., Maertens, A., McKim, J. M., Hartung, T., Kleensang, A., and Sa-Rocha, V. (2015). Probabilistic hazard assessment for skin sensitization potency by dose-response modeling using feature elimination instead of quantitative structure-activity relationships. *J. Appl. Toxicol.* **35**, 1361–1371.
- Natsch, A., Ryan, C. A., Foertsch, L., Emter, R., Jaworska, J., Gerberick, F., and Kern, P. (2013). A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. *J. Appl. Toxicol.* **33**, 1337–1352.
- Rovida, C., Alepee, N., Api, A. M., Basketter, D. A., Bois, F. Y., Caloni, F., Corsini, E., Daneshian, M., Eskes, C., Ezendam, J., *et al.* (2015). Integrated testing strategies (ITS) for safety assessment. *Altex* **32**, 25–40.
- Tsujita-Inoue, K., Atobe, T., Hirota, M., Ashikaga, T., and Kouzuki, H. (2015). *In silico* risk assessment for skin sensitization using artificial neural network analysis. *J. Toxicol. Sci.* **40**, 193–209.
- Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., Landsiedel, R., Jaworska, J., Kern, P. S., Gerberick, F., *et al.* (2015). Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul. Toxicol. Pharmacol.* **71**, 337–351.