Integrated hazard identification of chemical sensitizers using *in vitro* and *in silico* readouts – A comparative evaluation of predictive performance

Donna Macmillan[1], Henrik Johansson[2], Olivia Larne[2], Malin Lindstedt[3]

1. Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS
2. SenzaGen, Lund, Sweden
3. Lund University, Lund, Sweden

Abstract no. 322

Introduction

There has been a significant drive to reduce, refine and replace animal models for the prediction of skin sensitization. This is in part due to the implementation of EU regulation 1223/2009[1] which prohibits the sale and marketing of any cosmetics and cosmetic ingredients which have been tested on animals, alongside REACH[2] and CLP[3] regulations which state that non-animal methods must be exhausted prior to considering the use of animal tests. The use and availability of non-animal methods is ever-increasing and 3 assays have been validated by the OECD thus far; the *in chemico* DPRA, the *in vitro* KeratinoSens™ and the *in vitro* h-CLAT. A number of other assays are undergoing OECD validation, including the GARDskin assay (Genomic Allergen Rapid Detection), a dendritic cell-based assay which identifies skin sensitizers from 200 genomic biomarkers[4]. However, it is generally accepted that no single non-animal method can be used as a standalone approach to replace animal models such as the murine local lymph node assay (LLNA). The focus has instead turned to combining multiple *in chemico/in vitro/in silico* assays and/or molecular descriptors to derive a more accurate assessment of hazard or risk, known as integrated testing strategies (ITS)[5]. The GARDskin assay has demonstrated high predictivity and has been reported as ready to use in an ITS[6], therefore, it was decided to investigate the effect on performance when GARD was used in combination with Derek Nexus - and to compare these results against Derek with the DPRA, KeratinoSens™ and h-CLAT.

How does the integration of Derek affect *in chemico/in vitro* assay performance against the LLNA?

Performance using *in chemico/in vitro* assays or *in silico* methods alone - The predictive performance is largely comparable across the four *in chemico/in vitro* assays being evaluated in this study and Derek - although h-CLAT, GARDskin and Derek demonstrate slightly superior results to DPRA and KeratinoSens™ (Figure 1). Derek correctly predicts non-sensitizers more often than the *in chemico/in vitro* assays. The number of false negatives i.e. sensitizers incorrectly classified as non-sensitizers, is of particular importance where alternative methods are used in lieu of animal tests for establishing human safety, as the consequences (consumers being exposed to sensitizers) are severe. Accordingly, Derek predictions were used together with the *in chemico/in vitro* assay result to try and increase negative predictivity and improve overall performance.

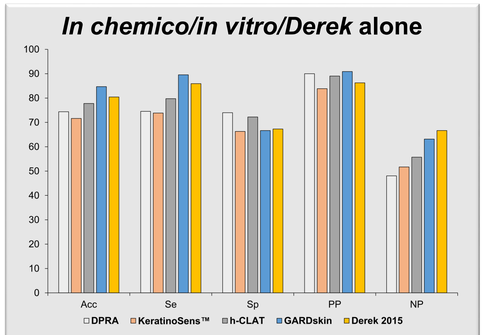

Figure 1. Performance metrics for DPRA, KeratinoSens™, h-CLAT, GARDskin and Derek against the LLNA.

Performance using Derek and *in chemico/in vitro* assays - Derek predictions were incorporated with the assay results using a conservative call approach - if either the Derek prediction or the assay result were positive then the chemical was classified as a sensitizer. If both Derek and the assay were negative then the chemical was classified as a non-sensitizer. Combining a Derek prediction with one assay has a mostly beneficial effect on performance (Figure 2) especially accuracy, sensitivity and negative predictivity, compared to using *in chemico/in vitro* assays alone (Figure 1). This has been at the expense of specificity which has fallen by 20-25% (Table 1) for most assay combinations (Derek + DPRA, KeratinoSens™, or h-CLAT) - however, an increase in false positives, and concomitant drop in specificity and positive predictivity is not unexpected from a conservative approach. For this dataset, using Derek with GARDskin demonstrates the best overall results (Figure 2 and Table 1).

Affect on number of mispredictions - Combining Derek and assay results can reduce the number of overall mispredictions (mispredictions = sensitizers classified as non-sensitizers and vice versa) significantly. DPRA mispredictions decrease by 40%; KeratinoSens™, 36%; h-CLAT, 26%; and GARDskin, 46% (Figure 3).

1 out of 2 conservative call using Derek and *in chemico/in vitro* assays

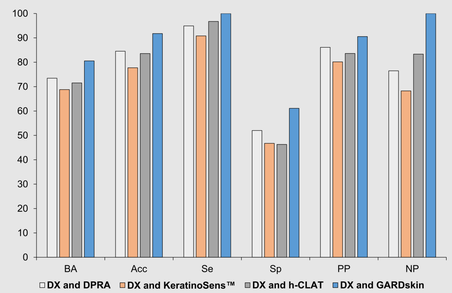

Figure 2. Performance metrics when using a 1 out of 2 approach with Derek (DX) and DPRA; KeratinoSens™; h-CLAT; or GARDskin against the LLNA.

% change to metrics after introduction of Derek predictions						
	Acc	**Se**	**Sp**	**PP**	**NP**	***n***
DPRA	+10.1	+20.4	-22.0	-3.9	+28.4	207
KeratinoSens™	+6.1	+17.0	-19.6	-3.7	+16.6	310
h-CLAT	+5.8	+17.0	-25.9	-5.4	+27.6	207
GARDskin	+7.1	+10.4	-5.6	-0.4	+36.8	85

Table 1. Performance metrics when using a 1 out of 2 approach with Derek and DPRA, KeratinoSens™, h-CLAT or GARDskin against the LLNA.

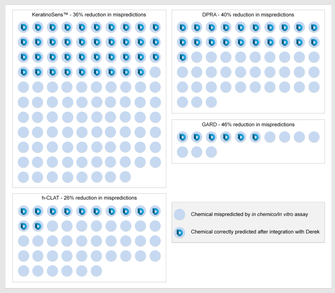

Figure 3. Pictogram illustrating the number of mispredictions when using one assay vs using Derek and one assay against the LLNA.

Method

Derek Nexus: Derek Nexus v5.0.2 and knowledge base 2015.2.0 was used to assess skin sensitization potential. Alerting chemicals with a likelihood of equivocal or above were assigned as sensitizers. Chemicals with a likelihood of improbable or non-alerting chemicals were assigned as non-sensitizers.

Data set and analysis: An in-house dataset of 373 chemicals with LLNA data and one or more of the following data (DPRA, *n* = 207; KeratinoSens™, *n* = 310; h-CLAT, *n* = 207; GARDskin, *n* = 85, and human, *n* = 113) was used. The following metrics were calculated: sensitivity (Se), specificity (Sp), positive predictivity (PP), negative predictivity (NP), and accuracy (Acc).

Does GARDskin predict well against human data?

GARDskin (*n* = 57) displays a similar predictivity performance towards human sensitization data as both Derek (*n* = 113) and the LLNA (*n* = 105) (Figure 4). Direct comparison is complicated by differences in data size and composition e.g. the GARDskin dataset has very few non-sensitizers which significantly impacts the specificity. Nevertheless, addition of Derek to GARDskin improves performance by correctly predicting sensitizers benzoyl peroxide, lauryl gallate, thioglycerol and methymethacrylate, all incorrectly predicted as non-sensitizers (Table 2).

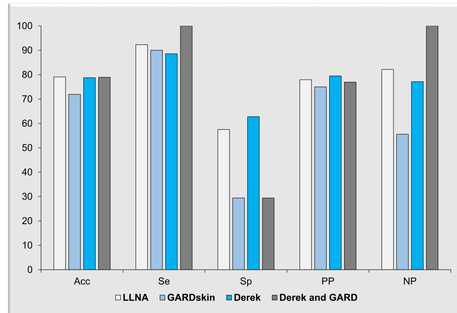

Figure 4. Performance for the LLNA, Derek, GARDskin, and Derek + GARDskin against human data.

Name	Derek	GARD	Human	LLNA
Benzoyl peroxide	1	0	1	1
Lauryl gallate	1	0	1	1
Thio-glycerol	1	0	1	1
Methyl-methacrylate	1	0	1	1

Table 2. Chemicals mispredicted by GARDskin but correctly predicted when a Derek (DX) prediction is added.

Conclusion

Using Derek skin sensitization predictions in combination with *in chemico/in vitro* assay results has a beneficial effect when predicting the LLNA outcome. GARDskin in particular performs extremely well when used with Derek in a conservative call approach. Human sensitization is more challenging to predict and GARDskin performs less well for this compared to predicting the LLNA - attributed to the small number of chemicals with both GARDskin and human data (*n* = 57), in addition to the positive bias in the GARD dataset (70%). However, the addition of Derek predictions clearly improve assay performance. Future work will focus on repeating this analysis on a larger, more balanced dataset.

References: (1) European Union. (2013) *Off. J. Eur. Union* 56, 34–66. (2) http://www.hse.gov.uk/reach/ (3) https://echa.europa.eu/testing-clp (4) Johansson et al (2011) *BMC Genomics* 12, 399. (5) Ezendam et al (2016) *Arch. Toxicol. 90*, 2861–2883. (6) Forreryd et al (2016) *Toxicol. Vitr.* 37, 178–188. (7) Macmillan et al (2016) *Regul. Toxicol. Pharmacol.* 76, 30-38.