Skin Sensitization Testing: The Ascendancy of Non-Animal Methods

David A. Basketter 1,* and George F. Gerberick 2

1 DABMEB Consultancy Ltd., Kingswood GL12 8RN, UK
2 GF3 Consultancy LLC, Cincinnati, OH 45069-4381, USA; gf3consultancy@gmail.com
* Correspondence: dabmebconsultancy ltd@me.com

Abstract: A century ago, toxicology was an empirical science identifying substance hazards in surrogate mammalian models. Over several decades, these models improved, evolved to reduce animal usage, and recently have begun the process of dispensing with animals entirely. However, despite good hazard identification, the translation of hazards into adequately assessed risks to human health often has presented challenges. Unfortunately, many skin sensitizers known to produce contact allergy in humans, despite being readily identified as such in the predictive assays, continue to cause this adverse health effect. Increasing the rigour of hazard identification is inappropriate. Regulatory action has only proven effective via complete bans of individual substances. Since the problem applies to a broad range of substances and industry categories, and since generic banning of skin sensitizers would be an economic catastrophe, the solution is surprisingly simple—they should be subject to rigorous safety assessment, with the risks thereby managed accordingly. The ascendancy of non-animal methods in skin sensitization is giving unparalleled opportunities in which toxicologists, risk assessors, and regulators can work in concert to achieve a better outcome for the protection of human health than has been delivered by the in vivo methods and associated regulations that they are replacing.

Keywords: skin sensitization; predictive testing; non-animal methods; hazard identification; risk assessment; defined approaches

1. Skin Sensitization Testing—A Short History

It is fair to assert that the identification of the clinical problem caused by skin sensitizers—allergic contact dermatitis (ACD)—preceded the advent of predictive tests by about half a century. Diagnostic patch testing traces its roots to the end of the 19th century [1]. The first definitive predictive test in the guinea pig was promulgated much later [2]. Whereas the skin and eye irritation tests found in this paper achieved a substantial foothold in regulatory toxicology (despite widespread scientific dissatisfaction with almost every aspect of their performance), application of the Draize skin sensitization test gained less long-term traction. A short review of the method noted in particular its lack of sensitivity [3]. The same review records describe several modifications designed to enhance sensitivity (and in our view also specificity)—for example, in the assessment of fragrance allergens [4].

The concerns regarding the Draize skin sensitization test prompted the development of a variety of alternative guinea pig assays. It is not necessary to review all of these, not least since an earlier substantial review is available, alongside more recent perspectives [5–7]. Many test protocols were used only within individual companies or industries. Ultimately, the guinea pig methods that survived into the latter part of the 20th century ultimately came down to the Buehler occluded patch test (BT) and the guinea pig maximization test (GPMT) [8,9]. Aspects of their performance are discussed later, but our view is that the main characteristic that differentiates these methods is that the GPMT with its use of adjuvant and intradermal injection emphasised sensitivity, whereas the BT favoured specificity.
Towards the end of the 20th century, with clear benefits regarding animal welfare, but particularly driven by the wish to have an objective, quantitative endpoint for skin sensitization, the local lymph node assay (LLNA) was developed [10,11]. It became the first alternative procedure in toxicology to receive formal validation approval [12,13]. In terms of this review, further development of the LLNA beyond binary hazard identification became possible. Interpretation of the dose–response curve was seen as a route to characterisation of the potency of an identified skin sensitizing substance [14]. This opportunity was steadily embraced by the wider toxicological community and endorsed by expert review [15]. Ultimately, the relative skin sensitizing potency measurement in the LLNA, known as the EC3 value, was recognised as a central element of a quantitative risk assessment process [16–20].

It is relevant to note that, alongside the in vivo tests mentioned above, predictive human testing of substances also developed and evolved. Although human volunteers were no doubt involved in research before this time, in the 1950s, the human repeated insult patch test (HRIPT) was described for the first time [21]. Greater characterisation of the HRIPT was pursued in the 1970s [22,23]. Definitive guidance on the protocol and interpretation of results was promulgated [24,25]. Additionally, in the 1960s, a huge series of human studies on many of the variables associated with the induction of skin sensitization led to the development of the human maximisation test (HMT) [26,27]. Just as happened with the BT/GPMT, the aim of the HMT was to achieve an optimal level of sensitivity compared with the HRIPT. However, the major criticism of these assays is not so much a debate on sensitivity versus specificity but rather whether in the 21st century it can be regarded as ethical to carry out human sensitization tests for the purposes of hazard identification [28,29]. To be clear, it is the view of the authors of this paper that conduct of the HRIPT as a predictive tool for substances/formulations of unknown skin sensitizing potency is unacceptable.

Most recently, the direction of travel for skin sensitization testing has been towards complete avoidance of the use of animal testing. This has led to the development, evaluation, validation, and introduction into regulation of a suite of largely in vitro tests that align with the key events (KEs) of the adverse outcome pathway (AOP) for skin sensitization [30]. An overview of these non-animal methods is presented in Table 1.

Table 1. Available New Approach Methods for Assessing Skin Sensitization Potential.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DPRA</td>
<td>442C, 2021</td>
<td>Key Event 1, peptide/protein binding</td>
<td>Positive/Negative on KE1</td>
</tr>
<tr>
<td>ADPRA</td>
<td>442C, 2021</td>
<td>Key Event 1, peptide/protein binding</td>
<td>Positive/Negative on KE1</td>
</tr>
<tr>
<td>kDPRA</td>
<td>442C, 2021</td>
<td>Key Event 1, peptide/protein binding</td>
<td>Positive/Negative on KE1 and quantitative information for Cat 1A or Cat 1B/NS</td>
</tr>
<tr>
<td>KeratinoSens™</td>
<td>442D, 2018</td>
<td>Key Event 2, keratinocyte response</td>
<td>Positive/Negative on KE2</td>
</tr>
<tr>
<td>LuSens</td>
<td>442D, 2018</td>
<td>Key Event 2, keratinocyte response</td>
<td>Positive/Negative on KE2</td>
</tr>
<tr>
<td>h-CLAT</td>
<td>442E, 2018</td>
<td>Key Event 3, Monocyte/dendritic cell response</td>
<td>Positive/Negative on KE3</td>
</tr>
<tr>
<td>U-SENS™</td>
<td>442E, 2018</td>
<td>Key Event 3, Monocyte/dendritic cell response</td>
<td>Positive/Negative on KE3</td>
</tr>
<tr>
<td>IL-8 Luc</td>
<td>442E, 2018</td>
<td>Key Event 3, Monocyte/dendritic cell response</td>
<td>Positive/Negative on KE3</td>
</tr>
<tr>
<td>2 out of 3 DA</td>
<td>497, 2021</td>
<td>Combining Key Events 1, 2 and 3</td>
<td>Positive/Negative for sensitizer</td>
</tr>
<tr>
<td>ITS v1 and v2 DA</td>
<td>497, 2021</td>
<td>Combining Key Events 1 and 3</td>
<td>Positive/Negative for sensitizer and information for Cat 1A or Cat 1B/NS</td>
</tr>
</tbody>
</table>
2. The Guinea Pig Era

Stretching across several decades during the 20th century, predictive test methods based on a wide variety of induction and challenge protocols were deployed. None of these methods were ever validated; indeed, when they were introduced and for years afterwards, datasets which supported their utility were largely absent. Only with the 1985 publication already mentioned and some later work did substantial sets of results with these methods become available [5,35]. For the two methods which survived longest (GPMT and BT), retrospective assessment using current validation principles would readily indicate that the assays were relevant but would fall far short of demonstrating reliability [36]. Furthermore, the originator of the GPMT long recognised that his assay had delivered high sensitivity at the expense of specificity [37].

A key strength of these guinea pig tests was that, once sensitized to a substance, then it was possible to evaluate cross reactivity, such as between similar preservatives or hair dyes [38–40]. However, that option appears to have received limited use in practice and of course it remains unknown whether the cross-reactivity seen in guinea pigs reflects what happens in humans (where the study of cross reactivity is always confounded by the challenge of understanding prior exposures—a topic beyond the scope of this article).

As already mentioned, the guinea pig tests were just as likely as later methods to deliver false positive and/or false negative results [41].

3. The Mouse Era

Largely in the 1990s, there was a successful attempt to overcome a range of scientific and animal welfare issues presented by the guinea pig methods, which was the advent of the local lymph node assay (LLNA) in the mouse. Much has been published regarding this assay and its validation, so the detailed history need not be repeated here [12]. The LLNA delivered both refinement and reduction in animal usage but also employed an objective, quantitative readout and a dose–response element. These latter benefits have been exploited with considerable success to permit the definition of the relative skin sensitizing potency of positive substances [42]. From this came the elaboration of an approach to safety evaluation known as quantitative risk assessment (QRA) [17,20,43].

The underlying principles of QRA parallel those of many other types of systemic toxicology risk assessment: in effect, a threshold is identified in a predictive model (in the case of the LLNA, the EC3 value), to which are applied various safety factors, leading to a value which can then be considered with respect to anticipated human exposure. It is worth noting that one of the advantages of QRA is that—as the whole process is transparent—it can be subjected to critical review and if appropriate, consequent updating, which is a key difference from safety evaluation processes based on the older guinea pig methods [36]. The most recent iteration, QRA2, has updated the safety factors and takes into account human exposure from multiple sources [20,44]. This is of particular importance as the notable outbreaks of ACD to a fragrance (Lyril) and preservatives (methylidibromoglutaronitrile—MDGN and methylisothiazolinone—MIT) occurred as a result of exposures significantly exceeding those that were built into the safety evaluation processes.

4. New Approach Methods for Skin Sensitization Testing

Regulations for skin sensitization data are different worldwide depending on the type of substance and its use category. The European ban on animal testing for new cosmetic ingredients was implemented in 2009 (Cosmetic legislation, Regulation (EC) No. 1223/2009). In 2013, the European Union’s 7th Amendment of the Cosmetic Directive placed a complete ban on animal testing for all cosmetics ingredients. Furthermore, the European Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) regulation calls for the use of non-animal methods. While REACH mandates that animal tests should be used as a last resort, a 2017 update of the regulation requires the uses of in vitro and in silico methods as the first choice for skin sensitization and permits animal testing under exceptional circumstances, at times including potency assessment [45,46].
Before and after the European ban on animal testing, scientists from government, industry, and academia actively started developing New Approach Methods (NAM), which are non-animal-based approaches, for assessing the skin sensitization potential of ingredients [47]. The work was advanced more quickly because there was a deep understanding of the underlying molecular and cellular mechanisms of ACD. This mechanistic understanding was used to develop and describe the key events of skin sensitization (e.g., the binding of hapten to proteins of the skin) within “The Adverse Outcome Pathway (AOP) for Skin Sensitization” [30]. The AOP for skin sensitization has become fundamental for applying NAMs [48]. A Long-Range Science Strategy (LRSS) has been implemented by Cosmetics Europe to aid the development of NAMs in human health risk assessments and to support regulatory decision making [49].

The Scientific Committee on Consumer Safety (SCCS) has provided guidance on using NAMs in a Next-Generation Risk Assessment (NGRA) framework [50]. For example, this involves how to use NAM data for determining a Point of Departure (PoD) termed No Expected Sensitization Induction Level (NESIL) in QRA, which have relied on animal data in the past [17,19,20]. The skin sensitization testing needs and data used by US regulatory and research agencies have been under review for several years [51,52]. To better protect human health and the environment, EPA’s Office of Pesticide Programs (OPP) is developing and evaluating new NAMs in molecular, cellular, and computational sciences to aid in the replacement of animal test methods. In 2020, the EPA released their New Approach Work Plan, which describes how the agency will develop, test, and apply chemical safety testing approaches that will reduce or replace animals (https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/strategic-vision-adopting-new-approach, accessed on 26 March 2022).

The challenge for industries developing new ingredients is that the regulations in different world regions are not consistent. For example, some countries may require animal tests for their approval process. Fortunately, progress is being made and the future looks to have more consistency in what NAM data are needed to support the skin sensitization safety of new ingredients or formulations.

Three OECD guidelines have been published that cover mechanistic key events (covalent binding to protein, keratinocyte activation, and dendritic cell activation) [30]. There are eight non-animal test methods approved in OECD TG that address these key events (Direct Peptide Reactivity Assay, DPRA; Amino Acid Derivative Reactivity Assay, ADRA; kDPRA; ARE-Nrf2 luciferase assay KeratinoSens, KS; ARE-Nrf2 luciferase assay LuSens; human Cell Line Activation Test, h-CLAT; U937 cell line activation test, U-SENS; and Interleukin-8 Reporter Gene Assay, IL-8 Luc Assay) [31–34]. For the most part, these test methods have been developed to identify skin sensitization hazards but are now being assessed for their ability to decide potency information for risk assessment determinations (Table 1) [47,50,53–56]. In addition to information from NAM predictions, data from in silico tools (e.g., Derek Nexus, TIMES-SS, ToxTree, and OECD toolbox) are also used to inform on skin sensitization hazards and potency [57–62].

A few NAMs have been designed to predict potency, including the SENS-IS [63,64], the Genomic Allergen Rapid Detection [65,66], and the kDPRA [67]. Using data from multiple NAMs, including in silico, in chemico and in vitro data, have been suggested for predicting potency, including a Bayesian network approach [68]; regression models with KS and peptide reactivity data [69]; an artificial neural network model [70,71]; and integrated use of h-CLAT, DPRA and DEREK data [72]. The critical need of these models is to deliver an accurate PoD value for conducting sound risk assessments that will meet the acceptance of risk assessors and regulators.

Defined approaches for skin sensitization contain fixed data interpretation procedures on how to assemble data obtained from different in chemico, in vitro, and in silico methods to determine whether a new chemical, without existing data, is a skin sensitizer and, if so, its potency [50,51,68]. Current work is now focused on combining NAM data to generate integrated approaches to testing and assessment (IATA) or defined approaches
Two DAs for determining skin sensitization have been published in a new OECD guideline [73]. OECD Guideline 497 includes the 2o3 DA and an integrated testing strategy (ITS v1 and ITS v2) DA. The 2o3 DA provides a hazard assessment based on two concordant, non-borderline results from DPRA, KS, and h-CLAT [74–76]. Skin sensitization potency information is not provided by the 2o3 DA. The ITS v1 and v2 DA integrates h-CLAT and DPRA data along with an in silico prediction. The ITS DA informs on both hazard and on potency using the UN Global Harmonized System (GHS) classification.

The GHS for skin sensitization potency is a binary subclassification of sensitizers into 1A (strong sensitizers) and 1B (other sensitizers). The kDPRA assay, which has been recently added to OECD guideline 442c, is a standalone assay for the application of Sub-category 1A [77–79]. An assessment of potency on a more granular scale is needed for next-generation risk assessment of new chemical entities. Thus, it is advantageous for risk assessors to have available approaches that can provide continuous PoD values so that more quantitative assessments can be made to help protect workers and consumers.

Linear regression models using KS and kinetic peptide reactivity data have been proposed to provide a PoD value in the form of a predicted EC3 value in the local lymph node assay (LLNA) [69,80]. Recently, updated quantitative regression models using input data from the kDPRA, KS, and h-CLAT were generated to calculate a PoD (Natsch and Gerberick, submitted). The predictivity of the models was characterized by comparing a comprehensive historical database vs. the curated dataset provided by the OECD working group on DA [81]. The predicted PoD was within or close to the variation in the historical LLNA data for most of the cases studies. Overall, the models predict the in vivo value with a median fold-misprediction factor of around 2.5. These updated regression models offer risk assessors flexibility in the choice of tests. A PoD value can still be determined when there are compatibility issues or chemicals outside an individual assay’s chemical domain.

Predicting an EC3 value offers the advantage of generating continuous potency values compared with predicting a chemical potency class [64,65,68,82]. It also provides the opportunity to manage uncertainty using statistical tools based on knowledge of the accuracy of the prediction. Such uncertainty could be factored in to refine the PoD value for conducting a skin sensitization risk assessment. The determination of potency has been primarily dependent on the use of the LLNA [83,84]. The LLNA has long been considered the “gold standard” for potency assessment because it yields quantitative data suitable for a dose–response evaluation. An alternative, non-animal approach for generating a PoD is urgently needed.

For risk assessors responsible for assessing the skin sensitization risk of new chemical entities or chemicals lacking sufficient data, it is critical to have tools available that are principally dependent on NAM data and read-across information. The PoD value obtained from these ITS models or existing DAs can assist in conducting skin sensitization risk assessments to help assure safety for consumers and workers. However, there remain challenges in incorporating NAMs and DAs into next-generation risk assessment approaches. Progress for assessing the hazard and potency of individual substances has been significant, but much more work is needed for evaluating complex mixtures and formulations [54]. Therefore, an effort is underway to develop next-generation risk assessment approaches for skin sensitization that do not rely on the need for new animal test data [85,86].

5. Summary

The journey in the development of NAMs has been long and challenging at times. Of course, hindsight on how to do things differently is always easier. However, one specific area that might have helped move the development of NAMs and DA along faster and more effectively would have been the establishment of a highly curated dataset of non-sensitizers and sensitizers of different potencies. There are numerous examples of publications proposing skin sensitization datasets for the evaluation of new methodologies [55,75,87–93]. What would have been helpful would have been the establishment of highly curated datasets of animal and human data that involved all stakeholders, includ-
ing government regulators and researchers, dermatologists, academic researchers, and industry researchers. Recently, an OECD working group has made a highly curated human dataset available [81]. Moreover, it has been suggested that a triangular approach involving NAM, animal, and human data be used for comparing the predictivity of skin sensitization test methods [76]. The authors show that relying too much on comparison with animal data (e.g., LLNA) or human data (e.g., diagnostic patch or repeat insult patch test data) without using uniform evaluation criteria or significant expert judgment is problematic. The result of not having datasets carefully reviewed can lead to misleading interpretations of specific assays or approaches. Recently, it has been proposed that a way forward is to use all available data, including chemistry knowledge from in silico models, to assign skin sensitization hazard and potency to skin allergens [89]. Thus, individual test data should not be used for evaluating alternatives, but a more holistic approach should be undertaken to establish the most relevant reference dataset for evaluating NAMs and DAs. Despite this ongoing challenge, significant progress has been made in eliminating animal testing for skin sensitization assessments. It is certainly worth noting that for other endpoints that are not nearly as far along as skin sensitization (e.g., respiratory sensitization, respiratory toxicity, phototoxicity, and photoallergy), a process for establishing criteria for selection of “gold” datasets should be undertaken that involves all key stakeholders.

Taking a longer-term view, it seems that skin sensitization has led the way in formal validation of refined/reduced alternatives (the LLNA) and in hazard characterisation (the EC3 value) and has taken a leading role in the adoption of transparent risk assessment (via QRA) and so continues to lead in the development of fully integrated safety evaluation strategies using NAMs. As with all innovations, there will be issues of one kind or another, but these should be seen as learning opportunities on our road to progress.

Author Contributions: Writing—original draft preparation, D.A.B. and G.F.G.; writing—review and editing, D.A.B. and G.F.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

4. Sharp, D.W. The sensitization potential of some perfume ingredients tested using a modified draize procedure. Toxicology 1978, 9, 261–271. [CrossRef]
6. Thyssen, J.P.; Giménez-Arnau, E.; Lepoittevin, J.P.; Menné, T.; Roman, A.; Schnuch, A. The critical review of methodologies and approaches to assess the inherent skin sensitization potential (skin allergies) of chemicals. Part II. Contact Dermat. 2012, 66 (Suppl. 1), 25–52. [CrossRef]
7. Thyssen, J.P.; McFadden, J.P.; Kimber, I. The multiple factors affecting the association between atopic dermatitis and contact sensitization. Allergy 2014, 69, 28–36. [CrossRef]
11. Kimber, I.; Dearman, R.J.; Scholes, E.W.; Basketter, D.A. The local lymph node assay: Developments and applications. Toxicology 1994, 93, 13–31. [CrossRef]


27. Kligman, A.M.; Epstein, W. Updating the maximization test for identifying contact allergens. *Contact Dermat.* 1975, 1, 231–239. [CrossRef]


30. OECD. The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins; OECD: Paris, France, 2014.


33. OECD. Test No. 442D: In Vitro Skin Sensitisation; OECD: Paris, France, 2018.

34. OECD. Test No. 442C: In Chemico Skin Sensitisation; OECD: Paris, France, 2021.


40. Basketter, D.A.; Goodwin, B.F. Investigation of the prohapten concept. Cross reactions between 1,4-substituted benzene derivatives in the guinea pig. *Contact Dermat.* 1988, 19, 248–253. [CrossRef]


42. Basketter, D.A.; Gerberick, F.; Kimber, I. The local lymph node assay and the assessment of relative potency: Status of validation. *Contact Dermat.* 2007, 57, 70–75. [CrossRef]


57. Macmillan, D.S.; Chilton, M.L. A defined approach for predicting skin sensitisation hazard and potency based on the guided integration of in silico, in chemico and in vitro data using exclusion criteria. Regul. Toxicol. Pharm. 2019, 101, 35–47. [CrossRef]


64. Cottrez, F.; Boitel, E.; Auriault, C.; Aeyb, P.; Groux, H. Genes specifically modulated in sensitized skins allow the detection of sensitizers in a reconstructed human skin model. Development of the SENS-IS assay. Toxicol. Vitro 2015, 29, 787–802. [CrossRef]


77. OECD. Test Guideline No. 442C: In Chemico Skin Sensitisation Assays addressing the Adverse Outcome Pathway, key event on covalent binding to proteins. In *OECD Guidelines for the Testing of Chemicals, Section 4; OECD Publishing*: Paris, France, 2021. [CrossRef]


