



LJMU Research Online

Johnson, C, Ahlberg, E, Anger, LT, Beilke, L, Benigni, R, Bercu, J, Bobst, S, Bower, D, Brigo, A, Campbell, S, Cronin, MTD, Crooks, I, Cross, CP, Doktorova, T, Exner, T, Faulkner, D, Fearon, IM, Fehr, M, Gad, SC, Gervais, V, Giddings, A, Glowienke, S, Hardy, B, Hasselgren, C, Hillegass, J, Jolly, R, Krupp, E, Lomnitski, L, Magby, J, Mestres, J, Milchak, L, Miller, S, Muster, W, Neilson, L, Parakhia, R, Parenty, A, Parris, P, Paulino, A, Paulino, AT, Roberts, DW, Schlecker, H, Stidl, R, Suarez-Rodriguez, D, Szabo, DT, Tice, RR, Urbisch, D, Vuorinen, A, Wall, B, Weiler, T, White, AT, Whritenour, J, Wichard, J, Woolley, D, Zwickl, C and Myatt, GJ

Skin sensitization in silico protocol

<http://researchonline.ljmu.ac.uk/id/eprint/13223/>

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

Johnson, C, Ahlberg, E, Anger, LT, Beilke, L, Benigni, R, Bercu, J, Bobst, S, Bower, D, Brigo, A, Campbell, S, Cronin, MTD, Crooks, I, Cross, CP, Doktorova, T, Exner, T, Faulkner, D, Fearon, IM, Fehr, M, Gad, SC, Gervais, V, Giddings, A, Glowienke, S, Hardy, B, Hasselgren, C, Hillegass, J, Jolly, R.

LJMU has developed **LJMU Research Online** for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

<http://researchonline.ljmu.ac.uk/>

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

Journal Pre-proof

Skin sensitization *in silico* protocol

Candice Johnson, Ernst Ahlberg, Lennart T. Anger, Lisa Beilke, Romualdo Benigni, Joel Bercu, Sol Bobst, David Bower, Alessandro Brigo, Sarah Campbell, Mark T.D. Cronin, Ian Crooks, Kevin P. Cross, Tatyana Doktorova, Thomas Exner, David Faulkner, Ian M. Fearon, Markus Fehr, Shayne C. Gad, Véronique Gervais, Amanda Giddings, Susanne Glowienke, Barry Hardy, Catrin Hasselgren, Jedd Hillegass, Robert Jolly, Eckart Krupp, Liat Lomnitski, Jason Magby, Jordi Mestres, Lawrence Milchak, Scott Miller, Wolfgang Muster, Louise Neilson, Rahul Parakhia, Alexis Parenty, Patricia Parris, Alexandre Paulino, Ana Theresa Paulino, David W. Roberts, Harald Schlecker, Reinhard Stidl, Diana Suarez-Rodriguez, David T. Szabo, Raymond R. Tice, Daniel Urbisch, Anna Vuorinen, Brian Wall, Thibaud Weiler, Angela T. White, Jessica Whritenour, Joerg Wichard, David Woolley, Craig Zwickl, Glenn J. Myatt



PII: S0273-2300(20)30114-8

DOI: <https://doi.org/10.1016/j.yrtph.2020.104688>

Reference: YRTPH 104688

To appear in: *Regulatory Toxicology and Pharmacology*

Received Date: 24 February 2020

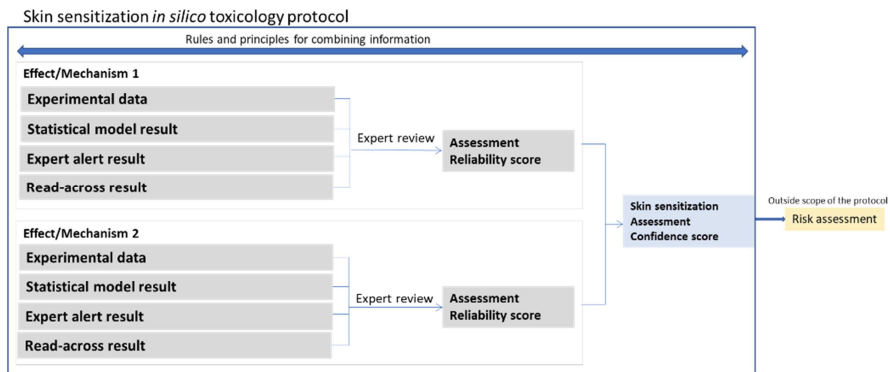
Revised Date: 18 May 2020

Accepted Date: 21 May 2020

Please cite this article as: Johnson, C., Ahlberg, E., Anger, L.T., Beilke, L., Benigni, R., Bercu, J., Bobst, S., Bower, D., Brigo, A., Campbell, S., Cronin, M.T.D., Crooks, I., Cross, K.P., Doktorova, T., Exner, T., Faulkner, D., Fearon, I.M., Fehr, M., Gad, S.C., Gervais, Vé., Giddings, A., Glowienke, S., Hardy, B., Hasselgren, C., Hillegass, J., Jolly, R., Krupp, E., Lomnitski, L., Magby, J., Mestres, J., Milchak, L., Miller, S., Muster, W., Neilson, L., Parakhia, R., Parenty, A., Parris, P., Paulino, A., Paulino, A.T., Roberts, D.W., Schlecker, H., Stidl, R., Suarez-Rodriguez, D., Szabo, D.T., Tice, R.R., Urbisch, D., Vuorinen, A., Wall, B., Weiler, T., White, A.T., Whritenour, J., Wichard, J., Woolley, D., Zwickl, C., Myatt, G.J., Skin sensitization *in silico* protocol, *Regulatory Toxicology and Pharmacology* (2020), doi: <https://doi.org/10.1016/j.yrtph.2020.104688>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 Published by Elsevier Inc.

Graphical abstract

1 Skin sensitization *in silico* protocol

2 Candice Johnson^{a*}, Ernst Ahlberg^b, Lennart T. Anger^c, Lisa Beilke^d, Romualdo Benigni^e, Joel Bercu^f, Sol
3 Bobst^g, David Bower^a, Alessandro Brigo^h, Sarah Campbellⁱ, Mark T.D. Cronin^j, Ian Crooks^k, Kevin P. Cross^a,
4 Tatyana Doktorova^l, Thomas Exner^l, David Faulkner^m, Ian M. Fearonⁿ, Markus Fehr^o, Shayne C Gad^p,
5 Véronique Gervais^q, Amanda Giddings^r, Susanne Glowienke^s, Barry Hardy^l, Catrin Hasselgren^c, Jedd
6 Hillegass^t, Robert Jolly^u, Eckart Krupp^v, Liat Lomnitski^w, Jason Magby^x, Jordi Mestres^y, Lawrence Milchak^z,
7 Scott Miller^a, Wolfgang Muster^h, Louise Neilson^{aa}, Rahul Parakhia^{bb}, Alexis Parenty^s, Patricia Parris^{cc},
8 Alexandre Paulino^{dd}, Ana Theresa Paulino^{dd}, David W. Roberts^j, Harald Schlecker^{ee}, Reinhard Stidl^{ff},
9 Diana Suarez-Rodriguez^{gg}, David T. Szabo^{hh}, Raymond R. Ticeⁱⁱ, Daniel Urbisch^{jj}, Anna Vuorinen^o, Brian
10 Wall^x, Thibaud Weiler^q, Angela T. White^r, Jessica Whritenour^{kk}, Joerg Wichard^{ee}, David Woolley^{ll}, Craig
11 Zwickl^{mm}, Glenn J. Myatt^a

- 12 a) Leadscope, Inc. 1393 Dublin Rd, Columbus, OH 43215, USA
13 b) Bioinformatics Department, The University of Uppsala, 752 36 Uppsala, Sweden
14 c) Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA
15 d) Toxicology Solutions Inc., San Diego, CA, USA
16 e) Alpha-PreTox, via G.Pascoli 1, 00184 Roma, Italy
17 f) Gilead Sciences, 333 Lakeside Drive, Foster City, CA, USA
18 g) Toxsci Advisors LLC, 2016 Main Suite 1901 Houston TX, USA
19 h) Roche Pharmaceutical Research & Early Development, Pharmaceutical Sciences, Roche
20 Innovation Center Basel, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel,
21 Switzerland
22 i) Nelson Laboratories, LLC, 6280 South Redwood Road, Salt Lake City, UT 84123
23 j) School of Pharmacy and Biomolecular Sciences ,Liverpool John Moores University, Liverpool, L3
24 3AF, UK
25 k) British American Tobacco, Research and Development, Regents Park Road, Southampton,
26 Hampshire SO15 8TL, UK
27 l) Edelweiss Connect GmbH, Technology Park Basel, Hochbergerstrasse 60C, CH-4057 Basel /
28 Basel-Stadt, Switzerland
29 m) Chemical Sciences Division, Lawrence Berkeley National Lab
30 n) whatIF? Consulting Ltd. The Crispin Burr Street, Harwell OX11 0DT, U.K.
31 o) DSM Nutritional Products, Kaiseraugst, Switzerland

- 32 p) Gad Consulting Services, 4008 Barrett Drive, Suite 201, Raleigh, NC 27609, USA
- 33 q) Servier Group, 905 route de Saran, 45520 Gidy, France
- 34 r) GlaxoSmithKline, Park Road, Ware, Hertfordshire, SG12 0DP, United Kingdom
- 35 s) Novartis Pharma AG, Pre-Clinical Safety, Werk Klybeck, CH-4057, Basel, Switzerland
- 36 t) Bristol-Myers Squibb, Drug Safety Evaluation, 1 Squibb Dr, New Brunswick, NJ 08903, USA
- 37 u) Toxicology Division, Eli Lilly and Company, Indianapolis, IN, USA
- 38 v) Sanofi, Corporate HSE, Global Product Stewardship, Industriepark Hoechst, D-65926 Frankfurt
- 39 am Main, Germany
- 40 w) Perrigo Israel Pharmaceuticals Ltd. Shoham Israel
- 41 x) Colgate-Palmolive Technology Center, 909 River Road, Piscataway NJ 08855 USA
- 42 y) IMIM Hospital del Mar Institute of Medical Research and University Pompeu Fabra, Doctor
- 43 Aiguader 88, Parc de Recerca Biomèdica, 08003 Barcelona, Spain; and Chemotargets SL, Baldiri
- 44 Reixac 4, Parc Científic de Barcelona, 08028 Barcelona, Spain
- 45 z) 3M Company, St. Paul, MN
- 46 aa) Broughton Nicotine Services, Oak Tree House, West Craven Drive, Earby, Lancashire. BB18 6JZ
- 47 UK
- 48 bb) Church & Dwight Co., Inc. 469 North Harrison Street, Princeton, NJ 08543
- 49 cc) Pfizer Worldwide Research and Development, Sandwich, UK
- 50 dd) ORO AGRI Europe, S.A. (Palmela - Portugal)
- 51 ee) Bayer AG, Research & Development, Pharmaceuticals, Industrial Chemicals Toxicology & Genetic
- 52 Toxicology, 42096 Wuppertal, Germany
- 53 ff) Safetree Consulting e.U., Vienna, Austria
- 54 gg) FStox consulting LTD, 2 Brooks Road Raunds Wellingborough NN9 6NS
- 55 hh) PPG Industries, Pittsburgh, PA 15146, USA
- 56 ii) RTice Consulting, Hillsborough, NC 27278, USA
- 57 jj) BASF SE , product safety, Carl-Bosch-Strasse 38, 67056 Ludwigshafen am Rhein, Germany
- 58 kk) Pfizer Inc., Drug Safety Research and Development, Eastern Point Road, Groton, CT 06340
- 59 ll) ForthTox Limited, PO Box 13550, Linlithgow, EH49 7YU, UK
- 60 mm) Transendix LLC, 1407 Moores Manor, Indianapolis, IN 46229, USA
- 61

62 *Corresponding author. E-mail address: cjohnson@leadscope.com (C. Johnson)

63 **Glossary of acronyms**

Ac	Acylation
ACD	Allergic Contact Dermatitis
ADRA	Amino Acid Derivative Reactivity Assay
AOP	Adverse Outcome Pathway
ARE	Antioxidant/electrophile response element
BT	Buehler test
CD54	Cluster of Differentiation 54, a co-stimulatory adhesion molecule that is expressed in dendritic cells
CD86	Cluster of Differentiation 86, a co-stimulatory adhesion molecule that is expressed in dendritic cells
CV ₇₀	Concentration of test chemical yielding a cell viability of 70% in the U-SENS™ method
DA	Defined Approach
DC	Dendritic cells
DIP	Data interpretation procedure
DPRA	Direct Peptide Reactivity Assay
DSA ₀₅	Dose per skin area that produced a positive response in 5% of the tested population
EC _{1.5}	Lowest concentration inducing a 1.5-fold change in luciferase activity in the assays measuring KE2
EC ₁₅₀	Effective concentrations yielding a relative fluorescence intensity [RFI] of 150% for CD86 in the h-CLAT test
EC ₂₀₀	Effective concentrations yielding a relative fluorescence intensity [RFI] of 200% for CD54 in the h-CLAT test
EC3	Effective concentration of a test chemical that gives a stimulation index with a three-fold increase over the vehicle control in the LLNA
EC ₃	Concentration with 3 fold luciferase induction in the KeratinoSens™ test
GARD	Genomic allergen rapid detection
GPMT	Guinea Pig Maximization test
GST	Glutathione S-transferase
HAF	Hazard assessment framework
h-CLAT	Human Cell Line Activation test
HMT	Human Maximization Test
HRIPT	Human Repeat Insult Patch Test
hTCPA	human T cell priming assay
IATA	Integrated approach to testing and assessment
IC	Induction concentration in GPMT
IC ₅₀	Concentration for 50% reduction of viability in KeratinoSens™ test
IL-18	Interleukin-18
IL-8	Interleukin-8
IL-8 Luc	Interleukin-8 Reporter Gene Assay
KE	Key Event
KE1	Key event 1: Covalent interaction with skin proteins
KE2	Key event 2: Events in keratinocytes

KE3	Key event 3: Events in dendritic cells
KE4	Key event 4: Events in lymphocytes
Keap1	Kelch-like ECH-associated protein 1
LLNA	Local Lymph Node Assay
LOEL	Lowest observed effect level
Log K _{ow}	n-octanol/water partition coefficient
MA	Michael addition
MHC	Major histo-compatibility complex
MIE	Molecular initiating event
NOEL	No Observed Effect Level
NQO1	NADPH-quinone oxidoreductase 1
Nrf2	Nuclear factor (erythroid-derived 2)-like 2
OECD	Organization for Economic Co-operation and Development
QMM	Quantitative Mechanistic Models
(Q)SAR	(Quantitative) Structure-Activity Relationship
RFI	Relative fluorescence intensity
SB	Schiff base formation
SI	Stimulation index
SLS	Sodium lauryl sulfate
SM	Supplementary material
SN1	Unimolecular nucleophilic substitution
SN2	Bimolecular nucleophilic substitution
SNAr	Nucleophilic aromatic substitution
STS	Sequential Testing Strategy
U-SENS™	U937 cell line activation Test

64

65

66

67

68

69

70

71

72

73

74 **Abstract**

75

76 The assessment of skin sensitization has evolved over the past few years to include *in vitro* assessments
77 of key events along the adverse outcome pathway and opportunistically capitalize on the strengths of *in*
78 *silico* methods to support a weight of evidence assessment without conducting a test in animals. While
79 *in silico* methods vary greatly in their purpose and format; there is a need to standardize the underlying
80 principles on which such models are developed and to make transparent the implications for the
81 uncertainty in the overall assessment. In this contribution, the relationship of skin sensitization relevant
82 effects, mechanisms, and endpoints are built into a hazard assessment framework. Based on the
83 relevance of the mechanisms and effects as well as the strengths and limitations of the experimental
84 systems used to identify them, rules and principles are defined for deriving skin sensitization *in silico*
85 assessments. Further, the assignments of reliability and confidence scores that reflect the overall
86 strength of the assessment are discussed. This skin sensitization protocol supports the implementation
87 and acceptance of *in silico* approaches for the prediction of skin sensitization.

88 **Keywords:** *In silico*, *in silico* toxicology, computational toxicology, computational toxicology protocols,
89 (Q)SAR, expert alerts, expert review, skin sensitization, defined approach, integrated approaches to
90 testing and assessment (IATA), extractables and leachables.

91

92	Contents	
93	Skin sensitization <i>in silico</i> protocol.....	1
94	Glossary of acronyms	3
95	Abstract	5
96	1. Introduction	8
97	1.1 Hazard Assessment Framework (HAF).....	9
98	1.1.1 Key Event (KE) 1: Molecular Initiating Event (MIE) – covalent interaction with skin proteins	10
99	1.1.2 Key Event (KE) 2: Events in keratinocytes	11
100	1.1.3 Key Event 3: Events in dendritic cells.....	11
101	1.1.4 Key Event 4: Events in lymphocytes.....	11
102	1.2 Integrated approach to testing and assessment (IATA).....	12
103	1.3 Defined Approaches	13
104	2. <i>In silico</i> methodologies and models.....	13
105	2.1 Covalent interaction with skin proteins, KE1	15
106	2.1.1 Dermal Metabolism	15
107	2.1.2 Reaction Domain	16
108	2.1.3 Protein Reactivity	17
109	2.2 Events in keratinocytes, KE2	17
110	2.3 Events in dendritic cells, KE3.....	18
111	2.4 Events in human lymphocytes, KE4	19
112	2.5 Events in rodent lymphocytes, KE4	19
113	2.6 Skin sensitization in rodents.....	21
114	2.7 Skin sensitization in humans.....	21
115	3. Endpoint assessment and confidence	21
116	3.1 Covalent interaction with skin proteins assessment.....	21
117	3.2 Events in keratinocytes.....	23
118	3.3 Events in dendritic cells	24
119	3.4 Skin sensitization <i>in vitro</i>	24
120	3.5 Skin sensitization <i>in vitro</i> to skin sensitization in human extrapolation.....	26
121	3.6 Skin sensitization in rodent lymphocytes	26
122	3.7 Skin sensitization in rodents.....	27

123	3.8 Skin sensitization in rodents to skin sensitization in human extrapolation.....	28
124	3.9 Skin sensitization in humans	28
125	4. Case Studies	30
126	4.1. Case 1a: Compound with conflicting data (“Skin Sensitization <i>in vitro</i> ” endpoint determination)	30
127	4.2 Case 1b: Compound with conflicting data (‘Skin Sensitization in Humans’ endpoint determination).	31
128	4.3 Case 2a: Pro/pre-hapten assessment	32
129	4.4 Case 2b: Pro/pre-hapten assessment Example 2	33
130	5. Reporting	33
131	6. Conclusion.....	34
132	7. Acknowledgements	34
133	Tables.....	35
134	Figures	35
135	References	38

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151 **1. Introduction**

152 Allergic contact dermatitis (ACD) is a common skin condition that results from the induction of a dermal
153 immunological response after repeated exposure to a skin-sensitizing substance. ACD poses a significant
154 public and occupational health concern, and much effort has been dedicated to the identification and
155 classification of skin sensitizers. Historically, assessors have relied on human (Human repeat insult patch
156 tests (HRIPT) and Human maximization tests (HMT)) or animal testing, the latter commonly using guinea
157 pig (Guinea pig maximization (GPMT) and Buehler tests(BT))(Organisation for Economic Co-operation
158 and Development (OECD), 1992) and mouse models (Local lymph node assay (LLNA))(OECD 2010a) to
159 identify potential skin sensitizers. The guiding principles of the “3Rs” (replacement, reduction, and
160 refinement) as applied to animal research(RUSSELL and BURCH 1959) have influenced the
161 implementation of regulations, such as the 7th amendment of the Cosmetic Directive (Council Directive
162 76/768/EEC of 1976-07-27; Cosmetics Regulation: REGULATION (EC) No. 1223/2009), European
163 substances legislation No. 1907/2006 (Registration, Evaluation, Authorization and Restriction of
164 Chemicals (REACH)) in the European Union; and Section 4(h) (Reduction of Testing in Vertebrates) of the
165 Toxic Substances Control Act (TSCA) in the United States. These regulations either prohibit the use of
166 animal testing or only allow animal testing if results obtained by alternative methods are not sufficient
167 to assess the sensitizing potential of a chemical. The “3Rs” together with the need for higher throughput
168 and more mechanistically informative methods, continue to drive the development of non-animal
169 methods. In this regard, *in silico*, *in chemico*, and *in vitro* methods in concert play an integral role in the
170 hazard assessment of skin sensitization.

171 *In silico* models, along with *in vitro* tests, have been and continue to be developed for predicting the
172 outcome of the four key events (KEs) described in the OECD adverse outcome pathway (AOP) for skin
173 sensitization (OECD 2014). It is generally accepted that the skin sensitizing hazard of a chemical can be
174 effectively assessed through the integration of non-animal approaches (Kleinstreuer et al., 2018; OECD,
175 2017). However, there may be data gaps that are generated through the exclusion of chemicals that do
176 not meet the physicochemical property requirements for the *in vitro* tests, and *in silico* methods that
177 could be used to fill such gaps may lack transparency as they are sometimes viewed as “black box” tools.
178 There is also no consensus on how to integrate *in vitro* data and/or *in silico* predictions for these events
179 with existing *in vivo* data.

180 The protocol detailed in this publication outlines a framework in which *in silico* methods could be
181 applied and integrated with existing *in vivo* and *in vitro* experimental data to identify potential skin

182 sensitizers, and to provide consensus on the development of models and the interpretation of model
183 results. *In silico* methods are likely to play an important role in understanding the hazard and risk
184 associated with chemicals (Myatt et al. 2018). Assessing sensitization is a necessary component of
185 classification and labelling, workers' safety and occupational health (where ~20-30% of compounds may
186 be sensitizers), regulation of cosmetics and other industrial chemicals as well as product discovery.
187 Previous studies have evaluated the potential use of *in silico* tools to predict sensitization hazard or
188 potential (Roberts and Aptula 2014; Roberts, Aptula, and Patlewicz 2006). However, there remains a
189 need for *in silico* guidelines and the definition of principles and procedures that are specific to the
190 prediction of skin sensitization relevant mechanisms. To this end, this skin sensitization protocol has
191 been developed based on the experience of a cross-industry consortium comprising 39 different
192 organizations and represents a consensus of how to use *in silico* methods to predict skin sensitization.

193 **1.1 Hazard Assessment Framework (HAF)**

194 Figure 1 provides a representation of a generic hazard assessment framework. The hazard assessment
195 framework defines the relationship between mechanisms and effects that are relevant for the
196 prediction of skin sensitization. The mechanisms and effects are molecular perturbations and
197 manifestations, respectively, that lead to the adverse outcome and are reflected in the AOP for skin
198 sensitization (Myatt et al. 2018). The mechanisms and effects are assessed based on *in silico* or existing
199 experimental data. Each mechanism/effect assessment is assigned a reliability score which reflects the
200 inherent quality of the assessment (Section 4). The relevance (scientific predictivity) of the
201 effect/mechanism is also assessed. Rules and principles are used to combine the mechanisms/effects to
202 derive an assessment of non-apical endpoints (i.e., endpoint 1 and 2 in figure 1) that are relevant for
203 sensitization. The non-apical endpoint assessment is assigned a confidence score, which is a reflection of
204 the reliability, relevance, and completeness of the assessment. Non-apical endpoints are combined via
205 rules and principles to derive an overall assessment for skin sensitization (the apical endpoint) with an
206 associated confidence score. The framework is designed to derive an assessment for hazard, with risk
207 being outside the scope of the protocol. Figure 2 shows the hazard assessment framework for
208 sensitization and the relationships between the following endpoints:

- 209 • Covalent interaction with skin proteins
- 210 • Events in keratinocytes
- 211 • Events in dendritic cells
- 212 • Skin sensitization *in vitro* (defined approach)

- 213 • Skin sensitization in rodent lymphocytes
- 214 • Skin sensitization in rodents
- 215 • Skin sensitization in humans (weight of evidence)

216 A comprehensive and mechanistic assessment for skin sensitization includes the four KEs described in
217 the AOP as well as available *in vivo* data and other supporting elements (OECD 2014). A mechanistic
218 understanding of the sensitizing process is detailed within the AOP for skin sensitization and becomes
219 necessary in the development of this framework. In order for a chemical to exert a sensitizing effect, a
220 series of well-defined stages/events occur that lead to the development of effector T cells (as opposed
221 to regulatory T cells, which lead to tolerance(OECD 2014). A chemical's ability to induce each KE is
222 critical information that is used in the development of the HAF. Sensitization is acquired through two
223 distinct phases. During the initial induction phase, the immune system is primed through dendritic cell
224 presentation of the sensitizing chemical to naïve T-cells. The induction phase occurs upon first contact
225 with the sensitizer and a physiological response is typically mild or absent. Upon re-exposure to the
226 same sensitizer, the primed immune system is activated and an inflammatory response occurs. This
227 phase is called the elicitation or challenge phase and results in the manifestation of the symptoms
228 associated with ACD: the appearance of rashes, blisters, and welts. A comprehensive assessment of the
229 skin sensitization potential of a chemical includes the four KEs that are described in the induction phase
230 (OECD 2014).

231 **1.1.1 Key Event (KE) 1: Molecular Initiating Event (MIE) – covalent interaction with skin** 232 **proteins**

233 The MIE for acquiring skin sensitization is the covalent binding of an electrophilic chemical to a
234 nucleophilic protein, typically the thiol group of cysteine or the primary amine group of lysine (Figure 3).
235 The interaction of the sensitizer (hapten) with the protein leads to the formation of a stable hapten-
236 protein conjugate. While a hapten-bound protein may result from direct interaction of the protein with
237 an electrophile, some chemicals require either metabolic (pro-haptens), or abiotic transformation
238 through oxidation (pre-haptens) prior to complexing with dermal proteins. The hapten-protein
239 interaction depends on the number of available nucleophilic target residues, steric considerations
240 (targets on the surface of a protein are more easily accessible than those in folds), and the
241 microenvironment (hydrophilic or hydrophobic)(OECD 2014). The formation of this complex is critical for
242 the activation of the immunological cells that are responsible for sensitization.

243 **1.1.2 Key Event (KE) 2: Events in keratinocytes**

244 It is accepted that interactions with the hapten lead to the modulation of inflammation-related
245 pathways and oxidative stress response pathways in keratinocytes (OECD 2014)(Figure 3).

246 Nuclear factor (erythroid-derived 2)-like 2 (Nrf2) is a transcription factor that trans-locates into the
247 nucleus of keratinocytes and binds to antioxidant/electrophile response elements (ARE). This in turn,
248 initiates the transcription of genes related to oxidative stress responses, such as NADPH-quinone
249 oxidoreductase 1 (NQO1) and glutathione S-transferase (GST). Nrf2 is repressed and controlled by the
250 Kelch-like ECH-associated protein 1 (Keap1), which facilitates the ubiquitination and degradation of
251 Nrf2. Keap1 is a cysteine (thiol) rich protein which can be modified by electrophiles (haptens) and
252 oxidants. This modification to Keap1 induces conformational changes in the protein that releases bound
253 Nrf2, allowing it to bind AREs and promote the expression of cyto-protective mechanisms (OECD, 2012).
254 In addition, interaction of the hapten with keratinocytes stimulates the production of pro-inflammatory
255 cytokines such as IL-18 (Natsch 2010). The release of cytokines by keratinocytes (among other factors)
256 plays a role in stimulating the maturation of dendritic cells (Sumpter, Balmert, and Kaplan 2019)

257 **1.1.3 Key Event 3: Events in dendritic cells**

258 Langerhans cells and dermal DCs are responsible for the presentation of the protein-hapten complex to
259 naïve T-cells in the lymph node during the induction phase (Figure 3). Following the uptake of the
260 protein-hapten conjugate, DCs process and present these peptide fragments in the context of major
261 histocompatibility complex (MHC) molecules to naïve T cells. Matured DCs migrate to the dermis and to
262 the lymph node under the influence of cytokines and chemokines that are secreted by keratinocytes and
263 fibroblast in the dermis (OECD 2014; Sumpter et al. 2019). During maturation, cell surface markers,
264 adhesion molecules, cytokines, and chemokines are upregulated. The upregulation of co-stimulatory
265 adhesion molecules (e.g., CD54, CD86) ensures that professional antigen presenting cells develop and
266 initiate an immune response. When there is a lack of co-stimulation, T-cell anergy (a state in which the
267 lymphocytes remain hypo-responsive after encounter with antigen) and a lack of sensitization may
268 result (OECD 2014; Vocanson et al. 2009)

269 **1.1.4 Key Event 4: Events in lymphocytes**

270 Presentation of the fragmented peptide complex within the MHC to naïve T-cells results in their
271 activation. This leads to the differentiation and proliferation of memory T-cells. Memory T-cells migrate
272 to the dermis and also circulate throughout the body. Upon re-exposure to the same hapten, the

273 memory T-cells are activated (elicitation phase) and the immune response is triggered; the result is the
274 manifestation of ACD, an irreversible immunologic response (OECD 2014).

275 KE 1-4 can be used to assess the 'skin sensitization *in vitro* endpoint', which in turn can be extrapolated
276 to the 'skin sensitization in humans' endpoint as shown in Figure 2. These *in vitro* endpoints can also be
277 predicted by *in silico* models as outlined in the HAF (Figure 2) and described in Section 2.

278 The availability of *in vivo* (usually rodent) data is relevant to the overall assessment of 'skin sensitization
279 in humans' and facilitates the development of *in silico* methods to predict the results. KE 4 (lymphocyte
280 activation and proliferation) can be measured with an *in vivo* mouse model and the adverse outcome
281 (e.g., erythema) can be assessed in guinea pigs. The events in lymphocytes (when assessed in mice) and
282 the guinea pig assessments can be combined to provide an overall assessment of 'Skin sensitization in
283 rodents'. Skin irritation may be a confounding factor and so is also considered at this point. An overall
284 assessment of 'skin sensitization in humans' can be determined through the integration of the 'skin
285 sensitization *in vitro*' and 'skin sensitization in rodents' endpoints. Historical human test data may also
286 be available and *in silico* models can be developed to facilitate its prediction. This information also
287 propagates into the 'skin sensitization in humans' endpoint.

288 The HAF consists of evaluation of KE1-4 via *in vitro* or *in vivo* testing, physio-chemical properties, and
289 human data (Figure 2). The assumption is made that all chemicals are capable of dermal penetration as
290 a conservative measure (Fitzpatrick, Roberts, and Patlewicz 2017). The endpoints in the framework may
291 be informed through available data, *in silico* predictions, or data acquired through conducting a test. The
292 protocol defines general rules and principles for integrating data towards an overall prediction of the
293 adverse outcome in humans. The incorporation of lines of evidence that may not directly relate to
294 sensitization; such as skin irritation, means that the protocol takes the form of an integrated approach
295 to testing and assessment (IATA).

296 **1.2 Integrated approach to testing and assessment (IATA)**

297 Given the definition of an AOP for skin sensitization and the availability of historical data, the endpoint is
298 effectively predicted using an IATA. Limited data for the KEs along the AOP have restricted the
299 development and applicability of *in silico* models to predict these endpoints while *in vitro* testing is
300 mainly used to derive an assessment of the activation of KEs along the AOP pathways. This may change
301 in the future, as more data become available and more robust *in silico* models can be developed.
302 Nonetheless, through an integrated scheme, the overall endpoint of 'skin sensitization in humans' is

303 assessed as a function of the activity at each KE, with additional evidence from either existing data or *in*
304 *silico* predictions of *in vivo* responses and metabolic biotransformation. Previous research has focused
305 on developing such schemes and these non-animal integrated strategies are receiving interest from
306 regulatory authorities. The publication of the 'Interim Science Policy: Use of Alternative Approaches for
307 Skin Sensitization as a Replacement for Laboratory Animal Testing' is an example of regulators adopting
308 this more integrated approach(EPA 2018). Additional non-animal assessment strategies are currently
309 being developed and validated, and more approaches may be adopted for regulatory purposes in the
310 future (Kleinstreuer et al. 2018). While several integrated approaches invoke the AOP and integrate the
311 KEs to derive an overall assessment of skin sensitization, it has been argued that failure or ability to
312 sensitize could be explained by (in)sufficient activity in the 'covalent interaction with skin proteins'
313 endpoint, and the evaluation of subsequent KEs is less important (Roberts and Aptula 2008). To this end,
314 the authors believe that a HAF that can facilitate multiple approaches is necessary. The ideal framework
315 should be generic enough to facilitate possible variations in analysis while maintaining a high level of
316 reproducibility and transparency. Rules and principles for combining results for each endpoint are
317 defined in this protocol. These rules will set the foundation for the reproducibility and flexibility of the
318 framework presented here.

319 **1.3 Defined Approaches**

320
321 Previous approaches have incorporated rules that connect various aspects of the toxicological pathway
322 to skin sensitization. The "2 out of 3" integrated testing strategy approach to skin sensitization hazard
323 identification proposed by BASF uses a data interpretation procedure (DIP) that labels a chemical as a
324 sensitizer or non-sensitizer based on the concordant reactivity of the chemical in two *in vitro* tests for
325 KE1 - KE3 (Urbisch et al. 2015). Several other integrated strategies have been developed to assess either
326 hazard or potency (Section 1 of the supplementary materials and described in detail elsewhere (OECD
327 2017)). Each approach addresses particular elements of the AOP. At the time of this manuscript, no
328 single approach is viewed as being superior to the others and selected approaches vary based on the
329 availability of computational tools and data.

330 **1. *In silico* methodologies and models**

331 Historically, *in silico* models have focused on the prediction of animal data (particularly the LLNA), and
332 few have considered the rest of the mechanisms established in the AOP. Therefore, it is necessary to
333 examine how *in silico* tools could be developed to model mechanisms related to the KEs described
334 earlier.

335 Depending on the availability of high-quantity data, different types of *in silico* models can be developed.
336 Table 1 provides a list of data sources. Larger amounts of data, preferably with a strong mechanistic
337 understanding of a specific toxicological process, can support many different types of models. Datasets
338 that cover a broad chemical space can support the development of global Quantitative Structure-
339 Activity Relationship ((Q)SAR) models, provided that the descriptors are relevant and mechanistically-
340 related to the endpoint that is being predicted (Roberts et al. 2007). Where data are sparse, generated
341 with different protocols, or generated through multiple mechanistic pathways (as may be the case in
342 human studies), methods such as expert-alerts or read-across may be more appropriate.¹ Statistical
343 models may also be developed; however, these models are potentially limited by a smaller applicability
344 domain. On the other hand, the mechanistic understanding and classification of chemicals into a
345 mechanistic domain means that local QSAR modeling may be a feasible approach for assessing events
346 related to the sensitizing endpoint. One of the earliest attempts to develop a local mechanism-based
347 QSAR model to predict EC3 concentrations in the LLNA, used the Relative Alkylation Index (RAI, a
348 function of electrophilic reactivity, lipophilicity, and dose)(Roberts et al. 1991; Roberts and Williams
349 1982). Subsequently, several Quantitative Mechanistic Models (QMM) have been developed with the
350 goal of identifying physicochemical and other descriptors that contribute to a mechanistic
351 understanding of an endpoint of interest (Aptula and Roberts 2006; Roberts and Aptula 2014; Roberts,
352 Aptula, and Patlewicz 2011; Roberts and Natsch 2009). The rest of Section 2 discusses the mechanisms
353 or effects that could be predicted and which types of *in silico* methodologies could facilitate the
354 predictions. On a general note, *in silico* methods typically derive structure activity relationships (SAR) for
355 organic salts by using the structure of the freebase. In cases where a metallic fragment will be removed
356 in the generation of the freebase to derive the SAR form of the structure, the potential hazard posed by
357 the metal should be considered. In the area of skin sensitization, removing nickel fragments may lead to
358 an underestimation of hazard for structures that contain them. To more accurately facilitate predictions
359 in these cases, the metal may be attached to the ligand, or the metal may be kept unattached in the
360 training set. The model builder may also decide to remove the salt structure entirely from the training
361 set; thereby, excluding the metal from the applicability domain of the model.
362 The following sections describe general considerations for building *in silico* models based on the
363 available chemistry, biology, and testing data. Section 1 of the supplementary material provides a
364 detailed description of the experimental data that are relevant for assessing skin sensitization. Methods

¹ The reader is referred to (Myatt et al., 2018) for a more general discussion on these methods

365 to assess the reliability of the data as well as *in silico* predictions have been previously described by
366 (Myatt et al. 2018) and are summarized in Section 2 of the supplemental material.

367 **2.1 Covalent interaction with skin proteins, KE1**

368 *In silico* and or experimental assessments for whether a given compound will participate in covalent
369 interactions with skin proteins are primarily generated based on understanding of metabolism, reaction
370 domain assignment and protein reactivity.

371 **2.1.1 Dermal Metabolism**

372 The allergenic potential of a chemical may be increased or decreased through metabolic pathways or
373 abiotic oxidation; these factors are important for predicting a chemical's potential to induce dermal
374 sensitization. Metabolic detoxification takes place in two phases, which may or may not occur
375 simultaneously. Phase II metabolism appears to be more abundant and active in the skin than in the
376 liver, although Phase I enzymes – though not dominant – are inducible in the skin (Dumont et al. 2015).
377 Given differences in expression profiles between the liver and skin, the potential use of liver metabolic
378 data to predict metabolites in the skin will necessitate strategies for accounting for the differences in
379 the expression of isoenzymes between liver and skin (Madden et al. 2017).² One strategy for predicting
380 metabolic activation towards sensitization in dermal tissues is to derive alerts to indicate if a chemical
381 may be a pro-hapten. This approach is currently limited by the size of the databases of pro-haptens and
382 a general lack of skin specific data (although knowledge has been gained through experience over the
383 years). Currently, it appears that the range of structural features that are activated towards sensitization
384 via metabolic pathways is small. Given the absence of skin-specific metabolic data, it is challenging to
385 definitively conclude on the topic. (Natsch and Haupt 2013) investigated the activation of pro-haptens
386 by rat liver S9 fractions in the KeratinoSensTM assay, and identified phenolic and alkoxy groups attached
387 to a benzene ring, some aromatic amines, and conjugated dienes in or in conjunction with six-
388 membered ring as structural features that may require pro-activation to behave as haptens in the assay
389 (Natsch & Haupt, 2013; Bergström, et al., 2006; Bergström, et al., 2006). The features identified do not
390 represent a comprehensive and thoroughly defined list of features that undergo metabolic
391 transformation leading to sensitization.

² The supplemental material provides a brief summary of the differences between skin and liver metabolic enzymes with relevance to humans

392 2.1.2 Reaction Domain

393 Existing mechanistic information on hapten-protein interactions has been used to construct *in silico*
394 models for predicting sensitization potential based on a compound's structure and known – or predicted
395 – reaction chemistry. The mechanisms for forming protein-hapten complexes involve the interaction
396 between an electrophilic chemical (hapten) and the nucleophilic moiety on a skin protein (generally thiol
397 or primary amine groups). Common mechanisms by which the sensitizer (hapten) may bind to the
398 protein are: Michael addition, acylation, Schiff base formation, unimolecular nucleophilic substitution
399 (SN1), bimolecular nucleophilic substitution (SN2), or nucleophilic aromatic substitution (SNAr). Within
400 each of these mechanistic domains, there are mechanistic alerts and structural alerts. Structural alerts
401 are defined as molecular substructures that can activate the toxicological effect or mechanism (Myatt et
402 al. 2018). Structural alerts that are characterized by a common reaction site are defined as mechanistic
403 alerts (Aptula and Roberts 2006; Enoch, Madden, and Cronin 2008; Roberts et al. 2015). Structural and
404 mechanistic data do not always suggest a toxic effect, however – some structural features, such as steric
405 hinderance, have been found to mitigate toxicity by decreasing the ability of the hapten to covalently
406 bind to proteins – and these features may improve an *in silico* model by providing this additional
407 information.

408 Classification of mechanistic and structural alerts within mechanistic domains allows for local QSAR
409 modelling within each domain (OECD 2011), provided that one has the relevant quantitative information
410 describing the protein-hapten bond. To this end, the following physical-chemical property descriptors
411 are commonly used to predict interactions between haptens and proteins: Molecular weight (MW), Log
412 P, solubility, rotational bonds, electronic and topological descriptors (e.g., quantum mechanics
413 calculations), or chemical structure-based descriptors (e.g., the presence or absence of different
414 functional groups) (OECD 2011). The factors constituting an acceptable and validated model have been
415 described in previous work (Myatt et al. 2018). However, it must be noted that due to the expert nature
416 of deriving structural alerts based on reaction chemistry, existing *in silico* tools can only incorporate our
417 current knowledge of protein-hapten reaction chemistry (rather than the quantification of a physical or
418 biological process), and that future models could be improved as we increase our mechanistic
419 understanding of these processes. QSARs on the other hand, are not limited by current knowledge of
420 mechanistic processes and the combined use of structural alerts and QSARs may add value to the
421 analysis.

422 **2.1.3 Protein Reactivity**

423 Protein reactivity has been studied using model nucleophiles to assess protein-chemical interaction in *in*
424 *chemico* assays. While the binding mechanism between the protein and the chemical could be described
425 based on reaction chemistry as discussed in the previous section, any *in silico* tools (either statistical or
426 expert rule-based) developed based on *in chemico* assay data will be limited in their ability to predict
427 sensitization due to pro-activation. To overcome this limitation, predictions based on reaction
428 chemistry, protein reactivity, and dermal metabolism should be considered in concert to generate an
429 overall assessment (described in Section 3.1).

430 While protein reactivity measurement is feasible across all reaction domains described in section 2.1.2,
431 experimental results show that within the domain of Schiff base formers there is a lower correlation
432 between the *in chemico*-based DPRA model and *in vivo* and human data (Urbisch et al. 2015). While
433 Schiff base formation may be theoretically feasible, the abundance of water within the peptide
434 reactivity testing environment may limit some reactions. As such, peptide reactivity was found to
435 correlate poorly with the potency of aldehydes, as Schiff base formation may be limited under testing
436 conditions in the DPRA (Natsch et al. 2015). Further analysis revealed that more potent Schiff base
437 formers (atranol, chloratranol, and salicylaldehyde) are reactive under physiological conditions (Natsch
438 et al. 2012). However, the LLNA EC3 values of Schiff base formers are well correlated ($R^2 = 0.95$) with a
439 combination of logP and a reactivity parameter based on substituent constants (Roberts et al. 2006).
440 Differential reactivity within a mechanistic domain is an issue that could become relevant in the
441 development of *in silico* models, and particularly in those that use read-across. Such instances may not
442 be unique to the protein reactivity mechanism but may require examination across all toxicological
443 endpoints.

444 **2.2 Events in keratinocytes, KE2**

445 A comprehensive prediction of keratinocyte activation covers events on several levels of biological
446 organization and includes the expression of biochemical, genomic, and proteomic pathways, and
447 quantifies the release of pro-inflammatory mediators that stimulate dendritic cells in KE3 (OECD 2014).
448 Validated protocols are established for assessing the induction of ARE dependent pathways, and, as
449 such, the development of *in silico* models can be considered for this assessment. However, the breadth
450 of information and data describing other pathways could be informative and may drive the development
451 of *in silico* models to predict additional pathways in the future.

452 Statistical modelling is feasible; however, the availability of data is a critical factor influencing the
453 success of measures to implement models based on AOP *in vitro* tests. Descriptors relating to the
454 covalent modification of the cysteine-rich Keap1 protein could be used to develop mechanistically-
455 relevant QSAR models. There may be limitations in predicting compounds which preferentially bind
456 hard nucleophiles such as lysine since the *in vitro* tests predicting KE2 rely on the cysteine-dependent
457 modification of Keap1. Therefore, false negative predictions may be more common for compounds that
458 react via acyl transfer, within the domain of Schiff base formers, including short chain aldehydes, and
459 longer chain saturated alkanals. Other electrophiles that prefer hard nucleophiles may also produce
460 false negative predictions (Urbisch et al. 2015). This could be a potential issue in read-across analysis
461 and should be addressed during an expert review.

462 *In silico* prediction of KeratinoSens™ and LuSens (*in vitro* test methods for assessing ARE activation in
463 keratinocytes) data yields dichotomous (either positive or negative) test results (OECD 2018b). However,
464 integrated assessments of potency may require continuous data input such as EC_{1.5} (the lowest
465 concentration inducing a 1.5-fold change in luciferase activity), IC₅₀ (concentration for 50% reduction of
466 viability) and EC₃ values (concentration with 3 fold luciferase induction) (Natsch et al. 2015).

467 **2.3 Events in dendritic cells, KE3**

468 Dendritic cell activation is similar to keratinocyte activation in that predictions can be made on the levels
469 of protein and gene expression. Methods have been validated for measuring the expression of specific
470 cell surface markers which contribute to T cell activation and proliferation. Published databases may
471 contain data for dendritic cell gene expression of co-stimulatory and adhesion molecules (cell surface
472 markers: CD54 and CD86) and Interleukin-8 (IL-8) (Nukada et al. 2011; Urbisch et al. 2015).

473 As noted for the KE2 endpoint, care must be taken when integrating testing data from the various *in*
474 *vitro* assays into KE3 *in silico* models due to differences in the types of data that may be produced by
475 different assays. The continuous data outcomes predicted for these assays, such as the EC₁₅₀ and EC₂₀₀
476 values from the h-CLAT assay; the CV₇₀ and the EC₁₅₀ in the U-SENS™ assay could be used in integrated
477 strategies to predict potency. These and other *in silico* predictions of the Ind-IL8LA (induced interleukin-
478 8 luciferase activity) could be used to support the hazard assessment; however, since a statistically-
479 derived experimental variable (confidence interval) is needed to determine a positive call, a more
480 practical approach may be to dichotomize the assay results and make binary predictions.

481 Often, it is helpful to build models that use threshold values to convert continuous data into
482 dichotomous (yes or no) values. For any of the *in vitro* or *in chemico* test methods that are used to
483 assess a KE along the AOP, using threshold values, *in silico* predictions could generate dichotomous
484 predictions of KE activity using these *in vitro* or *in chemico* test endpoints.

485 **2.4 Events in human lymphocytes, KE4**

486 The lack of standardized data makes *in silico* predictions of *in vitro* T cell activation and proliferation
487 challenging. A paucity of data for this endpoint is not surprising, however, as the value of predicting this
488 key event remains in question, and the significance of an *in vitro* estimate of KE4 can only be speculated
489 at this time. It is possible that the magnitude of the T cell responses at KE4 may be the key event that
490 allows us to make distinctions between different potency classes *in vitro* (OECD 2014), but the issue has
491 not been settled. Consequently, only the *in vivo* Local Lymph Node Assay has been accepted as a
492 standardized method for assessing this endpoint.

493 **2.5 Events in rodent lymphocytes, KE4**

494 The LLNA is the only standardized *in vivo* method used to measure the proliferation of lymphocytes in
495 response to immune system priming by a test chemical as well as the potency of the chemical as a skin
496 sensitizer. The results of the assay are reported as the concentration of the chemical needed to induce
497 T-cell proliferation by a pre-chosen factor (usually 3, 1.6, or 1.8 times the baseline amount as assessed
498 by the stimulation index (SI))(OECD 2010b, 2010a, 2018a). The LLNA has been used extensively, and it is
499 quite feasible to build *in silico* models using statistical and rule-based methods due to the ready
500 availability of data, although, the majority of such data is proprietary. While the publicly-available LLNA
501 data could facilitate statistical modeling, the model coverage may be reduced for industrial applications.
502 However, the combined use of statistical modeling and structural alert definitions could be a strategy to
503 overcome this limitation.

504 The irritation potential of a chemical could be a confounding factor in the experimental LLNA, and the
505 issue of irritation translates into *in silico* assessments. Training set examples and analogs under
506 consideration for read-across should be examined for their irritation potential. Studies indicate that non-
507 sensitizing irritants (such as surfactants) could be overestimated by the LLNA, leading to false positive
508 results (Ball et al. 2011; OECD 2010a). While this is certainly the case for sodium lauryl sulfate (SLS),
509 chloroform/methanol, Triton X-100, oxalic acid, methyl salicylate, and nonanoic acid, analysis of
510 chemicals known to be skin irritants has not validated this generalization across the entire class of non-
511 sensitizing irritants(Ball et al. 2011). Most non-sensitizing irritants are negative in the LLNA and those

512 that are positive may produce borderline results (with few exceptions). For example, the sensitization
513 hazard of SLS is derived from a clear dose-response curve that is indicative of a positive LLNA result;
514 however, when a weight-of-evidence (WoE) approach is used, the interpretation of the LLNA results may
515 be reversed. There is no evidence that SLS is a skin sensitizer in humans despite exposure; albeit limited,
516 it lacks a structural alert for sensitization and is a strong irritant (Basketter et al. 2009). Hence, Basketter
517 et al. 2009 have suggested that for the SI results obtained for SLS in the LLNA (SI_{SLS}), a WoE approach
518 could be developed around the false positive result to implement this approach in a general sense.
519 Using SLS as reference for a test chemical with unknown skin sensitization hazard, irritant potential and
520 SI predictions (SI_{test}); if the $SI_{test} < SI_{SLS}$ and no structural alert exists of sensitization, then the LLNA
521 prediction could be a suspected false positive and confidence in a positive prediction of the “skin
522 sensitization in humans” endpoint is low. The reverse may also be considered: If $SI_{test} > SI_{SLS}$ and an
523 alerting structure exists for sensitization; then the chemical may be suspected to be a true positive
524 (Basketter et al. 2009). The confidence could be adjusted accordingly based on the weight of evidence
525 presented. This sort of analysis would be considered with a low reliability LLNA study which may have
526 been conducted at irritant concentrations. Generally, the LLNA test is preceded by dose finding range
527 studies and minimally irritating to not irritating concentrations are tested.

528 Some LLNA protocols (LLNA-DA, and LLNA-BrdU-ELISA) use non-radioactive methods to quantify
529 lymphocyte proliferation. Results from these protocols could be combined in training sets that would
530 facilitate binary level predictions; however, varying criteria for predicting a positive call may complicate
531 the prediction of a meaningful continuous SI or EC_x value (where x is 3, 1.6, or 1.8 depending on the
532 LLNA protocol used) from such a dataset and would require a valid strategy for integrating the data.
533 Another relevant issue with LLNA datasets that arises in the curation process is the comparison and
534 combination of SI and EC_3 values for tests conducted in different vehicles. While it seems logical that
535 vehicle effects are normalized in the derivation of the SI and EC_3 values, there are mechanisms that
536 could lead to enhanced bioavailability depending on the choice of vehicle. The rapid evaporation of
537 acetone, for example, may result in volatilization of the test chemical and decreased bioavailability;
538 whereas dimethyl sulfoxide (DMSO) could potentially enhance penetration. Differing results may be
539 obtained between two LLNA tests using different vehicles and this could influence hazard assessment
540 (Hoffmann 2015). In some cases, vehicle effects may lead to the assignment of a chemical to two
541 neighboring potency classes (Anderson, Siegel, and Meade 2011; Basketter, Gerberick, and Kimber 2001;
542 Dumont et al. 2016; Hoffmann 2015). This inherent variability in the LLNA data (not exclusively caused
543 by different vehicles) is translated to *in silico* predictions. When combining multiple data sources, the

544 most conservative SI and EC_x values could be adopted, unless there is compelling evidence that the
545 vehicle is potentiating or attenuating the effect of the test chemical. A less conservative, but valid,
546 approach is to use the mean, or median values, among other valid approaches (Hoffmann et al. 2018).

547 **2.6 Skin sensitization in rodents**

548 The skin sensitization in rodent endpoint is evaluated through the use of the GPMT and the BT method.
549 Guinea pigs were historically used to assess skin sensitization. Similar to the LLNA, while public data are
550 available, much of the GPMT and BT data are proprietary. The data that exists could facilitate statistical
551 modeling, the derivation of expert alerts, and read-across.

552 **2.7 Skin sensitization in humans**

553 Historical data exist for this endpoint and, based on data quantity, expert-alert derivation and read-
554 across may be preferable to statistical methods. *In silico* predictions could be useful for the prediction of
555 dichotomized results of positive/negative. Potency predictions could be challenging based on data
556 availability. Evidence to support human predictions includes clinical data (DPT) and usage/occupational
557 exposure data (Api et al. 2017). Further, the integration of the 'skin sensitization *in vitro*' and the 'skin
558 sensitization in rodents' endpoints, along with any direct human evidence, are considered together as
559 weight of evidence for the prediction of the 'skin sensitization in humans' endpoint.

560 **3. Endpoint assessment and confidence**

561 The protocol details the integration of data with different reliabilities and relevance. Further, there may
562 be cases in which information that is critical to an assessment is missing. This section outlines the
563 rules/principles that could be applied when deriving an assessment and its associated confidence based
564 on the totality of evidence presented. Figure 4 shows the hazard assessment framework annotated with
565 references to where each of the following sections applies.

566 **3.1 Covalent interaction with skin proteins assessment**

567 Assessment of the 'covalent interaction with skin proteins' endpoint includes consideration of metabolic
568 transformation, reaction chemistry, and DPRA/ADRA predictions. Figure 5 shows how rules could be
569 made around the available information to derive an overall prediction of hazard. If an experimental
570 result is positive for the methods assessing KE1 (DPRA/ADRA), then a positive assessment of the
571 'covalent interaction with skin proteins' is warranted. However, the reliability of the prediction, as
572 assessed by the scheme presented in Table 6 of the supplementary material and described in (Myatt et
573 al. 2018), varies depending on the quality of the information presented and this has an influence on the

574 confidence score. The quality and reliability of an *in silico* DPRA/ADRA prediction could be assessed
575 according to the expert review criteria described in (Myatt et al. 2018). Additional considerations for
576 both experimental (test article) and *in silico* (training set examples and analogues) results include
577 situations in which DPRA/ADRA could lead to a false positive result due to oxidizing properties of the
578 test chemical, which can lead to peptide dimerization. An expert review could inform on whether or not
579 this is likely and if the assessment and confidence score need adjustment. Assessments of negative
580 DPRA/ADRA results vary based on consideration of the metabolic potential of the chemical together
581 with knowledge of reaction chemistry. In general, when the chemical is expected to be out of the
582 metabolic domain of the DPRA/ADRA then precedence is given to clearly-defined knowledge of reaction
583 chemistry (including mitigating factors, such as sterics) in the overall assessment of the 'covalent
584 interaction with skin proteins' endpoint. If the reaction chemistry indicates a mechanism leading to
585 sensitization; particularly if the mechanism requires pro-activation then the overall assessment of the
586 'covalent interaction with skin proteins' is positive based on reaction chemistry knowledge, but the
587 confidence is medium. If the test article is out of the metabolic domain, negative in DPRA/ADRA and no
588 mechanistic alert could be identified in the structure of the test chemical based on reaction chemistry,
589 then the DPRA/ADRA result is inconclusive as it cannot be said that the overall assessment is either
590 negative or positive. However, if metabolism is not predicted to occur and the chemical is considered
591 within the metabolic domain of the DPRA/ADRA, then the negative result should be given consideration
592 in the overall assessment. A negative DPRA/ADRA prediction (within the DPRA/ADRA metabolic domain)
593 and a positive mechanistic alert lead to a negative overall assessment, with a medium confidence level,
594 given that the DPRA/ADRA result is experimental and the positive mechanistic alert introduces some
595 uncertainty. An expert review would consider whether or not the test chemical is within the Schiff base
596 reaction domain. In these cases a negative DPRA/ADRA result may be mechanistically justifiable due to
597 the protein-hapten interaction being unfavorable under the test conditions as a result of the abundance
598 of water; particularly for chemicals that are indicated as less potent sensitizers by other methods. In this
599 case, the overall assessment could be considered positive (after expert review) with a low confidence.
600 This positive result is based on giving greater precedence to the mechanistic alert within this domain,
601 and the decreased relevance of the DPRA/ADRA due to the differential reactivity of chemicals within the
602 Schiff base domain. Further, co-elution of the test article with the model nucleophile may lead to false
603 negative predictions, although this occurs to a lesser extent in the ADRA than in the DPRA (Fujita et al.
604 2019).

605 In cases where the DPRA/ADRA result is positive, but no mechanistic alert can be assigned, it is worth
606 considering whether mechanistic knowledge could be provided by the protein reactivity results
607 particularly when close analogs point to the same structure-activity relationship. Figure 5 shows the
608 'covalent interaction with skin proteins' endpoint and the confidence score decision tree based on RS1
609 data. The confidence scores are expected to vary based on reliability and relevance; as such, there are
610 several possible permutations of the decision tree. These general "rules" are expanded to provide a
611 sense of the confidence assigned to assessments with varying reliabilities and relevance, Supplementary
612 Material, section 4 (SM 4).

613

614 **3.2 Events in keratinocytes**

615 The confidence score obtained for the activation of the events in keratinocytes towards skin
616 sensitization varies based on the Log K_{ow} of the chemical. If there is a positive prediction (RS1,
617 experimental) and the Log K_{ow} is <5 , then the result is assigned a high confidence. If the Log K_{ow} is
618 greater than 5, then the confidence is medium for a positive result and low for a negative prediction,
619 since limited information is available for such chemicals (OECD 2018b). Regardless of Log K_{ow} values,
620 negative results could be further assessed based on the occurrence of metabolism and the chemical
621 mechanism of action.

622 A metabolic alert (indicative of an expected metabolic transformation) along with a negative RS1/2
623 experimental or RS3 *in silico* result, could indicate reduced relevance of the *in vitro* assays predicting KE2
624 in this case – possibly because limited metabolic competency of the cells used in the assay are
625 responsible for a false negative. Therefore, the overall assessment would be negative but with a low
626 confidence score. If there is no biochemical transformation predicted, then the chemical mechanism of
627 action could be considered. A negative assessment for a chemical within the acyl transfer domain and
628 Schiff Base domain is conservatively assigned a low confidence score based on the preference of
629 chemicals within these domains for the lysine instead of the cysteine moiety (representing decreased
630 relevance). It is worth mentioning that some chemicals within these domains are accurately predicted as
631 true negatives and a review of the relevance is necessary to assign a higher confidence. Such a review
632 might include an examination of close analogs (or the test structure if data is available) for their
633 assessment in the DPRA/ADRA and or an animal model. If close analogs are positive in the DPRA/ADRA
634 and the lysine moiety; but not cysteine, is implicated for covalent modification then the relevance of the
635 KE2 assays for predicting the test structure may be challenged. However, if cysteine modification is
636 apparent in the DPRA/ADRA (positive for covalent interaction with skin proteins), it is more difficult to

637 challenge the relevance of the KE2 assays on that basis and conflicting information is presented by the
638 two KEs. The analogs may be further assessed and screened for existing animal data and/or *in silico*
639 predictions of the LLNA or GPMT. This serves the purpose to assess the likelihood of a false negative
640 prediction of the test structure by the KE2 assays. Where a false negative seems likely, the low
641 confidence is appropriate. In cases where the analogs are true negatives, the confidence score could be
642 increased to a medium level and this reflects that while uncertainty is somewhat reduced, there is not
643 absolute certainty in the assessment. Within any other domain, a negative KE2 prediction is considered
644 with high confidence, given RS1/2 data. Varying reliabilities of the data could change the confidence
645 scores in figures 6A and B (see SM Table 9).

646 **3.3 Events in dendritic cells**

647 An overall assessment of the events in dendritic cells could be made based on the h-CLAT (Figure 7), U-
648 SENS™ or IL-8 Luc assays (Figure 8). A positive response from these assays typically translates to a
649 positive overall call for the events in dendritic cells with high confidence in the activation of the
650 dendritic cells towards sensitization, but an expert reviewer would be needed to adjust overall calls and
651 confidence scores for certain chemical classes, structural features, and physical-chemical properties. For
652 example: some chemical classes, such as surfactants, may lead to false positive results in the U-SENS™,
653 and a negative result for a chemical that has a Log K_{ow} greater than 3.5 is considered inconclusive for the
654 h-CLAT. The pro/pre-hapten status of the test chemical is also relevant in each of the three assays.
655 Negative results for structures in which a site of metabolism leading to sensitization has been identified
656 are accepted with a medium level confidence from the h-CLAT, U-SENS™ and IL-8 assays. In cases where
657 there are no additional parameters confounding the prediction, then the confidence level is high for the
658 negative predictions from the h-CLAT, U-SENS™, and IL-8 Luc assays.

659 **3.4 Skin sensitization *in vitro***

660 Integrating data to derive an overall assessment for the 'skin sensitization *in vitro*' endpoint that
661 correlates with the *in vivo* endpoint is an active area of research. A number of defined approaches (DA)
662 which use varying DIPs have been developed to determine an overall assessment of skin sensitization
663 using non-animal/*in-vitro/in silico* models. Any of the DAs described in Section 1 may be adopted here.
664 There has been regulatory acceptance of the "AOP 2 out of 3" approach and the KE3/1 sequential
665 testing strategy (STS) as alternatives to the LLNA for regulatory submission to the United States
666 Environmental Protection Agency (US EPA) (EPA 2018). Here, we discuss how to derive an overall

667 assessment and confidence when the “AOP 2 out of 3” approach is used within the framework
668 presented in this protocol.

669 The “AOP 2 out of 3” uses the outcome of three individual assays that map to three KEs to derive a final
670 assessment; however, within the framework presented the assay results are integrated and propagated
671 to the three endpoints related to each key event. The difference between the “AOP 2 out of 3” and the
672 approach used in the framework is subtle, but is worth mention. The “AOP 2 out of 3” approach
673 considers the outcome of the experimental systems – DPRA, KeratinoSensTM, and h-CLAT – but within
674 the framework presented, the: ‘Covalent interaction with skin proteins’, ‘Events in keratinocytes’, and
675 ‘Events in dendritic cells’ (KEs in the AOP) are considered. These KEs are assessed based on knowledge
676 of reaction chemistry and mechanistic understanding that is not explicitly considered within the “AOP 2
677 out of 3” approach. Similar to the “AOP 2 out of 3,” an overall assessment of hazard for the ‘skin
678 sensitization *in vitro*’ endpoint is determined based on a 2 out of 3 consensus among the endpoints. If
679 outcomes (*in silico*/experimental) are available for only two endpoints, and they have aligned outcomes,
680 the overall assessment of the endpoint is based on the concordant assessments and the lower
681 confidence score propagates. The adoption of the lower confidence score reflects a conservative view of
682 the assessment at this stage of the analysis. However, if the confidence scores have the same value for
683 non-concordant assessments, then the overall prediction for the ‘skin sensitization *in vitro*’ endpoint is
684 inconclusive. Where there are two concordant assessments, and the non-concordant assessment occurs
685 with high confidence, then the overall confidence could be lowered by one level. Table 2 provides
686 examples showing the derivation of the overall assessment and the rationale for the final confidence
687 score. An alternative point of view suggests that the assays that predict the ‘events in keratinocytes’,
688 and ‘events in dendritic cells’, are dependent on the ability of the test chemical to bind protein and
689 therefore point to the activation of the molecular initiating event, ‘covalent interaction with skin
690 proteins’. In this point of view, any improvement in predictive performance that results from
691 integrating the KEs across the AOP is a result of reducing the influence of technical limitations of each of
692 the assays (Roberts 2018; Roberts and Patlewicz 2018).

693

694

695

696

697

698

699

700

701
702
703 The discussion thus far has focused on assessing hazard from *in vitro* data, but there are also existing
704 strategies for predicting potency in humans from *in vitro* data based on the DAs described in Section 1
705 and reviewed in (Kleinstreuer et al. 2018). The Artificial Neural Network Model for Predicting LLNA EC3
706 (Shiseido); Bayesian Network DIP (BN-ITS-3) for Hazard and Potency Identification of Skin Sensitizers
707 (P&G); Sequential Testing Strategy (STS) for Sensitizing Potency Classification Based on *in Chemico* and
708 *In Vitro* Data (Kao); and ITS for Sensitizing Potency Classification Based on *In Silico*; *In Chemico*, and *In*
709 *Vitro* Data (Kao) were found to predict potency class equally well, or better than the LLNA. Similar to the
710 earlier discussion on hazard, the DAs for assessing potency use biological assay outcomes
711 (mechanisms/effects assessment within the HAF e.g. DPRA, KeratinoSens™, h-CLAT) as endpoints and
712 may integrate the information with *in silico* methods to determine a potency class. Within the HAF
713 presented, the assay outcomes (*in vitro/in silico* effects/mechanisms assessment) are interpreted in the
714 context of their toxicological significance and integrated to determine a toxicological endpoint according
715 to the rules and principles outlined in previous sections. The overall assessments of the KE endpoints
716 may substitute for the outcome of the individual test methods in data interpretation procedures.

717 **3.5 Skin sensitization *in vitro* to skin sensitization in human extrapolation**

718 Extrapolation of *in vitro* skin sensitization results to human skin sensitization predictions is necessary to
719 satisfy the European Union's 7th Amendment of the Cosmetic Directive and REACH regulations which
720 require and prefer the use of non-animal test methods for assessing the human skin sensitization
721 endpoint. The definition of the AOP and the mechanistic information provided by the assays that map to
722 the AOP allow the human hazard identified for the 'skin sensitization *in vitro*' outcome to be propagated
723 to the human endpoint. The relevance of the integrated *in vitro* battery of tests is equally weighted with
724 the *in vivo* studies except in unique cases; for example, when metabolism is thought to influence the
725 outcome. As such, no change in confidence (reliability and relevance of the prediction) is expected due
726 to the extrapolation of *in vitro* hazard.

727 **3.6 Skin sensitization in rodent lymphocytes**

728 A negative result in the LLNA is propagated to the skin sensitization in rodent lymphocytes endpoint
729 with high confidence. A weak sensitizer may require investigation of the skin irritation potential of the
730 chemical, particularly if the result is derived from a lower-reliability study that may not have considered
731 irritation prior to designing the test. The skin irritation potential will be determined through a HAF that
732 will be published in a separate protocol. Positive results due to confounding factors from irritants usually

733 result in a low-level increase in lymphocytes which could be misinterpreted as a weak sensitizing
734 response. In cases where a chemical is found to have a strong skin irritation potential and is a weak
735 sensitizer and the influence of irritation cannot be ruled out, a positive assessment with low confidence
736 could be assigned to the 'Events in rodent lymphocytes' endpoint (Figure 9).

737 **3.7 Skin sensitization in rodents**

738 This endpoint integrates guinea pig (GPMT and BT) and mouse (LLNA) data. In the absence of LLNA data,
739 the endpoint could be determined through the scheme shown in Figure 10. If guinea pig tests are not
740 conducted according to standard protocols, irritation could become a confounding factor in the
741 interpretation of the guinea pig test results and influence the relevance of the study (OECD 1992).
742 Freund's complete adjuvant (FCA) is used to maximize the guinea pig response; however, FCA may also
743 lower the irritation threshold. The implication is that concentrations that were identified as non-
744 irritating and suitable for the challenge reaction might in fact produce an irritant response. Further, a
745 hyperirritable state may be induced by the test article during the induction phase that is not
746 represented in the control, unless a suitably irritating surrogate is used to induce the hyperirritable state
747 in the controls (Kligman and Basketter 1995; OECD 1992). An irritant effect cannot be distinguished from
748 an allergic response by visual examination. As such, post challenge examination is helpful in
749 distinguishing a sensitization response from an irritant effect. Chemicals that are identified as irritants
750 could be confidently predicted as non-sensitizers if observations of erythema dissipate within one day of
751 challenge and/or there is a negative re-challenge test one week after the initial challenge (Kligman and
752 Basketter 1995). A positive result for a chemical that is irritating but predicted to be a weak sensitizer is
753 afforded a low confidence score if deviations from OECD 1996 result in decreased reliability and
754 relevance of the study as discussed above.

755
756 When both guinea pig and mouse data are available and are concordant, then the result is translated to
757 the 'skin sensitization in rodent' endpoint with exact or higher confidence scores being adopted. For
758 example, if the LLNA is positive with medium confidence and the GPMT/BT is positive with low
759 confidence, then the skin sensitization in rodent endpoint is assessed as positive with medium
760 confidence. In cases where the data are discordant, the strategy for deriving an overall assessment may
761 vary case-to-case. A high reliability guinea pig test has an advantage over the LLNA because it includes
762 both induction and challenge phases, and is as such, more representative of the entire sensitization
763 process. However, in contrast to the LLNA, the guinea pig test results are based on a qualitative measure
764 and a subjective endpoint. Potency is better assessed through the LLNA since it is derived from dose-

765 response relationships and the read-out is quantitative; nonetheless, some chemical classes are over-
766 classified in the LLNA. It is valuable to consider how the challenge reaction affects interpretation of an
767 assessment. It could be argued that the LLNA is an assay and non-specific reactions can occur that may
768 or may not relate to allergenic potential (respiratory sensitizers test positive in the LLNA, for example)
769 while the dermal challenge in the guinea pig tests lends more confidence that any observations of
770 sensitization are specific to the skin. A default principle that could be adopted is to evaluate the 'skin
771 sensitization in rodent' endpoint based on either the LLNA or GPMT/BT assessment with the higher
772 confidence score and conservatively decrease the score by one level to reflect any uncertainty. For
773 example, an LLNA that is assessed as positive with medium confidence, and a GP test that is negative
774 with low confidence, would lead to a 'skin sensitization in rodent' assessment as positive with low
775 confidence. In these circumstances, a review of the predictions is prudent and the assessment and
776 confidence scores may be adjusted based on the review.

777 **3.8 Skin sensitization in rodents to skin sensitization in human extrapolation**

778 There are two schools of thought on rodent-to-human extrapolation that draw from a two different
779 perspectives on risk assessment: one is that LLNA potency categories and EC3 values correlate well with
780 human potency categories and NOEL values, and could therefore be used as a surrogate for the NOEL
781 and for direct prediction of human potency class (Basketter et al. 2005). Alternatively, a safety factor
782 may be incorporated based on the interspecies variation that may occur between the mouse and
783 humans; although, this factor could be lowered in cases where a better correlation may be expected
784 (e.g., based on existing human data for a close analogue) (Roberts and Api 2018). Roberts and Api 2018,
785 have defined alerts for cases where the LLNA is not a good predictor of human potency. Guinea pig tests
786 also provide relevant information on hazard and potency. However, tests that use adjuvant and
787 intradermal routes of exposure (GPMT) present a challenge for interpreting human potency, and in
788 those situations potency estimation via the BT may be more relevant. The data however, could serve in
789 a weight-of-evidence case for potency determination through interpretation and comparison of
790 different test results and also with known benchmark chemicals (Kimber et al. 2001).

791 **3.9 Skin sensitization in humans**

792 The 'skin sensitization in humans' endpoint could be evaluated through several other endpoints such as
793 the 'skin sensitization *in vitro*' endpoint (section 3.5), the 'skin sensitization in rodent endpoints' (section
794 3.8), or through the integration of the 'skin sensitization *in vitro*', 'skin sensitization in rodents', and
795 human assessments, combined with supporting data from non-standard endpoints such as photoallergy.

796 A positive HMT/HRIPT is indicative of adverse outcome in humans and can potentially be used to assign
797 a potency class. In the absence of reliable studies, other sources of evidence may be sought. The first
798 line of evidence arises from the toxicological relationships that could be drawn from the chemical's
799 structure. The presence of a structural alert for sensitization in humans provides evidence for the
800 elicitation of the adverse outcome. Structural alerts and diagnostic patch testing with positive incidences
801 in greater than 1% of the population (considered to be high incidence) in relation to low usage volume
802 (a measure of exposure) provides evidence for the skin sensitization potential of a chemical, although it
803 does not provide a definite assessment (Api et al. 2017). If a compound has no structural alerts and
804 diagnostic patch testing data indicate < 1% frequency, the overall evidence may together indicate a
805 negative assessment, especially if the use volume is high. It is important to note that the indication of a
806 1% incidence rate is based on expert opinion and as such is not meant to represent a rule that requires
807 strict compliance. Many combinations of scenarios are possible.

808 In cases where human and *in vitro/in vivo* sensitization assessments do not align, additional information
809 could be gathered from the 'skin sensitization *in vitro*' and/or 'skin sensitization in rodents' endpoints to
810 build a weight of evidence case. There are many permutations of assay results at this level but some
811 general guidance can be provided to the evaluator towards an overall assessment. It is generally
812 recommended that the assessments that are assigned more frequently should be propagated to the
813 overall human endpoint. However, if reliable human data (RS1/2) is available, then the assessment of
814 this data is given priority in the decision-making process. Table 13 of the supplementary material
815 expands on the principles to derive an overall assessment given *in vitro* and rodent evidence. Due to
816 ethical concerns, human testing is no longer considered appropriate for most compounds, so much of
817 the human data is older, or based on clinical reports, and may therefore lack information to assess its
818 quality, necessitating the filter of expert opinion. Careful consideration is required in assessing
819 confidence of the HMT and HRIPTs. For the HMT and, especially for the HRIPT as used by the fragrance
820 industry, low doses are often tested as the goal is to corroborate an animal study while trying to avoid
821 sensitizing the subjects. Therefore, there can be quite a bit of uncertainty in a negative result because a
822 higher test concentration could potentially produce a positive result in humans.

823 Table 3 shows factors to consider in assigning confidence to a human study in general. There are
824 however some specific exceptions to these criteria when assessing the HMT and HRIPTs. The exposure
825 scenarios in the HMT and HRIPT may not represent real-world exposure because the test chemical is
826 applied under occlusive conditions and the outcomes can be viewed as subjective because an observer
827 grades the skin reaction.

828

829 4. Case Studies

830 The case studies demonstrate the interpretation of results when a series of statistical models ((Q)SARs),
831 structural alerts, or read-across are used to fill data gaps for effects and mechanisms that are included in
832 the hazard assessment framework. The studies demonstrate how aspects of the rules and principles are
833 implemented to derive an assessment, reliability score and confidence score, when the assessment is
834 made using either or both existing experimental data or *in silico* methods.

835 4.1. Case 1a: Compound with conflicting data ("Skin Sensitization *in vitro*" endpoint 836 determination)

837 An assessor needed to determine the hazard associated with a compound. The compound was
838 predicted to be reactive towards proteins via an Acyl or SN2 reaction, and could be assigned to a
839 reaction domain based on reaction chemistry alerts. Data that was generated based on OECD TG 442C
840 (DPRA) was available for the compound. The data indicated that the compound was negative for protein
841 reactivity. Based on adherence to the test guideline, a reliability score of RS1 was assigned to the study.
842 *In silico* tools (statistical results (QSAR) and alerts) were available for the DPRA prediction, and these
843 predictions were also negative. The statistical model and the alerts both had a reliability score of RS5. *In*
844 *silico* assessments of dermal metabolism were negative after an expert review. The review increased the
845 reliability of the dermal metabolism alert from RS5 to RS3. The overall assessment for 'covalent
846 interaction with skin proteins' was negative; however, the confidence was assigned as medium, based
847 on the conflicting mechanistic/reaction chemistry alert for protein reactivity, Figure 11a.

848

849 There is experimental data for the KeratinoSens™ assay which is afforded a positive assessment with a
850 reliability score of RS1 (the study adhered to OECD TG 442D), so the overall assessment for the 'Events
851 in Keratinocytes' KE is positive with high confidence. Experimental data is not available for the 'Events in
852 Dendritic Cells' KE. The assessor would like to use the "2 out of 3" approach and is faced with two
853 conflicting assessments based on *in vitro* data. A statistical model (QSAR) was used to predict the results
854 of the h-CLAT assay and the assessment is negative with a reliability score of RS3, after an expert review.
855 The overall assessment of the 'Events in Dendritic Cells' KE is negative with a medium confidence. Based
856 on the two concordant assessments with aligned confidence scores (Negative, Medium confidence), and
857 a third assessment that is conflicting with high confidence (Positive, High confidence), the overall
858 assessment of *in vitro* skin sensitization endpoint is negative with low confidence.

859 **4.2 Case 1b: Compound with conflicting data ('Skin Sensitization in Humans' endpoint**
860 **determination)**

861 A further assessment was completed for the same compound as in Case 1a. This assessor has LLNA and
862 GPMT data with conflicting assessments. The LLNA data is positive with an EC3 (%) value that indicates
863 weak sensitization. The study is assigned the lowest reliability score of 5 based on significant deviations
864 from OECD Test No. 429 that could alter both the reliability and relevance of the study. *In silico*
865 assessments using expert alerts and statistical models are both negative. The weak sensitizing effect and
866 the mis-aligned *in silico* results prompt the assessor to consider the irritation potential of the chemical.
867 Experimental data is available for the *in vitro* skin irritation test using the Reconstructed Human
868 Epidermis (RHE) test method. The assessment of skin irritation is positive with a score of RS1. The
869 assessor conducts an expert review of the LLNA and suspects a false positive LLNA result. The 'Events in
870 Rodent Lymphocytes' endpoint could be assigned as positive with low confidence; however, the
871 negative *in silico* results are more reliable and relevant in this situation and the negative assessment
872 carries over to the 'Events in Rodent Lymphocytes' endpoint with medium confidence. The GPMT data is
873 negative with a reliability of RS1 since the study adhered to OECD 406 and the irritant effect was
874 considered in the study design and interpretation of results. *In silico* models agree with the experimental
875 GPMT result. The overall assessment of 'Skin sensitization in rodents' is negative with a high confidence,
876 Figure 11b.

877 To further investigate the outcome in humans, the assessor conducted an *in silico* assessment using a
878 set of alerts that were developed using HMT and HRIPT data as a reference database and no alerting
879 structure were found. No human study data were available; however, DPT data were available and
880 consecutive patients showed frequencies of 0% in a study. The absence of positive DPT results are
881 indicative of no sensitization in humans, although a conclusion cannot be made from DPT data alone.

882
883 Given the weight of evidence presented in Case 1a and 1b a final determination of the 'skin sensitization
884 in humans' can be made. In this case, a well conducted GPMT carried significant weight towards the
885 negative sensitization assessment with high confidence; reflecting the high reliability and relevance of
886 the information. Other evidence supporting a negative assessment included a negative protein binding
887 test which was reinforced by negative *in silico* models predictions of protein binding; and negative
888 (Q)SARs predicting the 'Events in dendritic cells', and LLNA. The conflicting piece of information
889 presented by the LLNA study was viewed as less reliable and relevant information due primarily to
890 confounding irritant effects in the study. A second piece of conflicting information was presented by the
891 KeratinoSensTM experimental study. While no specific explanation for this false positive was

892 determined, the body of negative evidence for the 'skin sensitization *in vitro*' endpoint supports the
893 negative assessment and the low confidence reflects any uncertainty in the assessment of that
894 endpoint. However, the 'sensitization *in vitro*' assessment does not discredit the 'skin sensitization in
895 rodents' assessment. Since the *in vitro* and rodent endpoints are both equally relevant when the *in vitro*
896 endpoint is derived through a defined approach, the endpoint that contains more reliable information
897 contributes more to the overall confidence. The *in vitro* endpoint does not introduce any uncertainty in
898 the GPMT experimental findings, and taken together with the DPT data, the final confidence score is
899 high in this negative case, Figure 11c. There may be instances where a higher level of conservatism is
900 necessary than presented. In such instances, the confidence score could be reduced to medium,
901 although a change in the assessment might be difficult to justify.

902 **4.3 Case 2a: Pro/pre-hapten assessment**

903 Figure 12a details the assessment for the mechanisms/effects that were considered in the case. A
904 chemical is being screened for possible use in the cosmetics industry. It is expected to undergo
905 metabolic transformation leading to the formation of quinones, which have a high probability to react
906 via Michael addition (MA). There are positive alerts for dermal metabolism and the site of metabolism
907 coincides with a pro-MA reactivity alert. Negative DPRA data are available and the DPRA study is
908 assigned a reliability score of 1 based on adherence to OECD TG 442C. However, based on knowledge
909 that the compound contains a pro-reactive feature that coincides with a site of metabolism, the
910 relevance of the DPRA for testing the compound is challenged since any activity that results from a
911 metabolic transformation may be missed. The DPRA test is considered not relevant for the compound
912 tested, and the assessment of the 'Covalent interaction with skin proteins' is based on the assignment of
913 a pro-reactive domain. Although there may be cases where a pro-reactive domain assignment does not
914 lead to protein interaction due to deactivating features, a conservative approach to assessing the
915 endpoint given a pro-reactive feature is to assign a positive assessment with a lowered confidence. No
916 other *in vitro* data are available for the compound. A (Q)SAR was developed based on proprietary data
917 for the KeratinoSensTM assay. The test compound is assessed as positive in the (Q)SAR, with the pro-MA
918 feature identified as significant by the model. After a review of the (Q)SAR prediction, the 'Events in
919 Keratinocytes' is assessed as positive with medium confidence. No data or models were available for the
920 'Events in dendritic cells' endpoint. Given the positive assessment for 'Covalent interaction with skin
921 proteins' and the 'Events in Keratinocytes', the overall assessment for the 'Skin Sensitization *in vitro*'
922 endpoint is made using the "2 out of 3" approach. The overall assessment of the 'Skin Sensitization *in*
923 *vitro*' endpoint is positive with low confidence based on the two aligned positive assessments and the

924 lower confidence score propagating to the endpoint, Figure 12a. It is possible to extrapolate the existing
925 hazard information to the 'Skin sensitization in humans' endpoint and assess it as positive with low
926 confidence.

927 **4.4 Case 2b: Pro/pre-hapten assessment Example 2**

928 Consider an extension of the case presented in Section 6.3. LLNA data are not available for the test
929 compound but are available for close analogs. In addition there is a low quality guinea pig test for the
930 test compound that indicates a positive sensitization response. Read-across is performed using the
931 LLNA data for the analogs. The analogs all contained the pro-reactive feature and formed a congeneric
932 series that allowed interpolation of the LLNA EC3 value. The EC3 value was predicted to be 3.2%,
933 indicative of a moderate sensitizer. The 'Events in rodent lymphocytes' endpoint was assessed as
934 positive with medium confidence based on the read-across result. The guinea pig test is assigned a
935 reliability score of RS5 based on deviations from OECD 406. A review of the study showed that for an
936 induction concentration of 1%, the sensitization incidence is 100% suggesting that the compound could
937 be classified as a Category 1A sensitizer. After an expert review of the study, the reliability score is
938 increased to RS3. The overall assessment of the 'Skin Sensitization in Rodents' endpoint is assessed as
939 positive, with medium confidence based on the weight of evidence presented by the LLNA read-across
940 and guinea pig study.

941 The 'Skin sensitization in rodents' and the 'Skin sensitization *in vitro*' endpoints both support that
942 assignment of a positive hazard for the 'skin sensitization in humans' with medium confidence. Figure
943 12b shows the flow of information within the hazard assessment framework.

944 **5. Reporting**

945 An important consideration towards *in silico* standardization, reproducibility and transparency is a
946 consistent reporting format (Myatt et al. 2018). The general protocol (Myatt et al. 2018) describes a
947 proposed reporting format that includes the elements that provide completeness of information. The
948 report format is reproduced in Table 4 with a minor modification for the skin sensitization endpoint. In
949 addition to the description of models, databases, and tools that were used, it is also recommended to
950 describe any IATAs, DIPs or DAs that were used in deriving the overall assessment. The details that are
951 suggested should allow another expert to repeat the process and achieve the same results. Further, the
952 standardized report enables streamlined and consistent review of regulatory submissions across
953 industries and endpoints. Section 5 of the Supplementary Material (SM5) provides an example of a
954 report for sensitization hazard.

955 **6. Conclusion**

956 The skin sensitization *in silico* protocol presented here is the first publication to outline a systematic
957 assessment of skin sensitization based on both experimental data and *in silico* predictions. It includes a
958 HAF and provides general rules for the *in silico* toxicological assessment of chemicals within the
959 framework. The framework is transparent and flexible as it does not require the generation of all
960 endpoints to derive an overall assessment of 'skin sensitization in humans' and can accommodate
961 quantitative and qualitative predictions and/or experimental results. There are cases where
962 extrapolation to the human endpoint is possible and this has been described. The corresponding
963 assessment of the confidence for all endpoints allows the protocol to be used in a variety of use cases.
964 For example, assessments with low confidence scores may still have practical usage in screening or
965 prioritization use cases. In addition, the protocol highlights experimental approaches or *in silico* models
966 that could be incorporated into the HAF in the near future. Expert review is a critical element in any such
967 procedure and items to consider as part of this review are listed to support a more consistent
968 assessment. The standardization of the HAF for performing *in silico* methods is designed to support
969 increased use and acceptance of *in silico* tools among regulatory agencies and industries alike.

970 **7. Acknowledgements**

971 Research reported in this publication was supported by the National Institute of Environmental Health
972 Sciences of the National Institutes of Health under Award Number R43ES026909. The content is solely
973 the responsibility of the authors and does not necessarily represent the official views of the National
974 Institutes of Health.

975

976

977

978

979

980

981

982

983

984 **Tables**985 Table 1. Sources of data for the development of *in silico* methods986 Table 2. Examples of deriving an overall assessment and confidence for the “skin sensitization *in vitro*”
987 endpoint using the “AOP 2 out of 3” approach

988

989 Table 3: Factors increasing and decreasing confidence in a human study (Schulz, Altman, and Moher
990 2010; Sibbald and Roland 1998)991 Table 4: Elements of an *in silico* toxicology report992 **Figures**993 Figure 1. A generic hazard assessment framework that shows the relationship between the key
994 components of the protocol995 Figure 2. The hazard assessment framework describing the *in silico* components relevant for skin
996 sensitization. *In silico* models could be developed for any effect or mechanism within grey boxes.

997

998 Figure 3. Adverse Outcome Pathway (AOP) for skin sensitization. MIE- molecular initiating event, KE (1-4)
999 - Key Events 1-4.1000 Figure 4. The hazard assessment framework annotated with sections that discuss the assessment and
1001 confidence score of each endpoint.

1002

1003 Figure 5. Decision tree showing how an overall assessment and confidence score could be derived for
1004 the covalent interaction of skin proteins. The confidence scores are based on RS1 experimental data:
1005 assuming relevant data and high reliability, and, in practice, confidence scores may need to be adjusted
1006 based on reliability scores, SM Table 8. *If a pro-reactivity domain is assigned and the metabolic site
1007 (determined using structural alerts for skin metabolism) coincides with the pro-reactivity domain center
1008 then the reversal in assessment occurs. If the metabolic site and the reactivity domain center do not
1009 align then the assessment is inconclusive. [§]The inconclusive result is applicable in situations where
1010 structural alerts could be used to determine if a structure is expected to undergo metabolism but not
1011 identify the metabolites. In this case, since the reactivity of the metabolite cannot be confirmed, a
1012 conclusion cannot be made on the assessment. If the reactivity of the metabolites could be predicted
1013 then the final assessment depends on the metabolite reactivity.

1014 Figures 6A. Decision trees showing how an overall assessment and confidence score could be derived for
1015 the 'events in keratinocytes'. The confidence scores here are based on RS1 experimental data: assuming
1016 relevant data and high reliability, and, in practice, confidence scores may need to be adjusted based on
1017 reliability scores.

1018

1019 Figures 6B. Decision trees showing how an overall assessment and confidence score could be derived for
1020 the 'events in keratinocytes'. The confidence scores here are based on RS1 experimental data: assuming
1021 relevant data and high reliability, and, in practice, confidence scores may need to be adjusted based on
1022 reliability scores.

1023

1024 Figure 7. Decision tree showing how an overall assessment and confidence score could be derived for
1025 the 'events in dendritic cells' based on the h-CLAT assay. The confidence scores here are based on RS1
1026 experimental data: assuming relevant data and high reliability, and, in practice, confidence scores may
1027 need to be adjusted based on reliability scores.

1028

1029 Figure 8. Decision tree showing how an overall assessment and confidence score could be derived for
1030 the 'events in dendritic cells' based on the U-SENS™ and IL-8 Luc assay data. The confidence scores here
1031 are based on RS1 experimental data: assuming relevant data and high reliability, and, in practice,
1032 confidence scores may need to be adjusted based on reliability scores, SM Table 10.

1033

1034 Figure 9. Decision tree showing how an overall assessment and confidence score could be derived for the "Events
1035 in rodent lymphocytes" based on the LLNA. The confidence scores here are based on RS1/2 experimental data
1036 (except in the case of *): assuming relevant data and high reliability, and, in practice, confidence scores may need
1037 to be adjusted based on reliability scores. *Concentrations tested in the LLNA are either non-irritating or mildly
1038 irritating. The low confidence score reflects the non-specific increase in lymphocyte proliferation that could occur
1039 with irritants.

1040 Figure 10. Decision tree showing how an overall assessment and confidence score could be derived for
1041 the 'skin sensitization in rodents' endpoint based on guinea pig tests. The confidence scores here are
1042 based on RS1 experimental data (except in the case of *): assuming relevant data and high reliability,
1043 and, in practice, confidence scores may need to be adjusted based on reliability scores. *GPMT/BT

1044 challenge concentrations are non-irritating; however, deviations from OECD 406 may reduce the
1045 relevance of the study and decrease the confidence in the endpoint.

1046

1047 Figure 11a. Derivation of the 'skin sensitization *in vitro*' endpoint using the "AOP 2 out of 3" approach
1048 (Case 1a)

1049 Figure 11b. Derivation of the 'Skin Sensitization in Rodents' endpoint

1050 Figure 11c. Derivation of the 'skin Sensitization in Humans' endpoint from the weight of evidence
1051 presented from the 'Skin Sensitization skin *in vitro*' and 'Skin Sensitization in Rodents' endpoints. DPT
1052 data is also used to support the overall assessment.

1053 Figure 12a. Derivation of the 'skin sensitization *in vitro*' endpoint using the "AOP 2 out of 3" approach
1054 (Case 2)

1055 Figure 12b. Derivation of the 'Skin Sensitization in Humans' using the "AOP 2 out of 3" approach (Case 2)

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070 **References**

- 1071 Anderson, Stacey E., Paul D. Siegel, and B. J. Meade. 2011. "The LLNA: A Brief Review of Recent
1072 Advances and Limitations." *Journal of Allergy* 2011:424203.
- 1073 Api, Anne Marie, Rahul Parakhia, Devin O'Brien, and David A. Basketter. 2017. "Fragrances Categorized
1074 According to Relative Human Skin Sensitization Potency." *Dermatitis: Contact, Atopic,
1075 Occupational, Drug* 28(5):299–307.
- 1076 Aptula, Aynur O. and David W. Roberts. 2006. "Mechanistic Applicability Domains for Nonanimal-Based
1077 Prediction of Toxicological End Points: General Principles and Application to Reactive Toxicity."
1078 *Chemical Research in Toxicology* 19(8):1097–1105.
- 1079 Ball, Nicholas, Stuart Cagen, Juan-Carlos Carrillo, Hans Certa, Dorothea Eigler, Roger Emter, Frank
1080 Faulhammer, Christine Garcia, Cynthia Graham, Carl Haux, Susanne N. Kolle, Reinhard Kreiling,
1081 Andreas Natsch, and Annette Mehling. 2011. "Evaluating the Sensitization Potential of Surfactants:
1082 Integrating Data from the Local Lymph Node Assay, Guinea Pig Maximization Test, and in Vitro
1083 Methods in a Weight-of-Evidence Approach." *Regulatory Toxicology and Pharmacology* 60(3):389–
1084 400.
- 1085 Basketter, D. A., G. F. Gerberick, and I. Kimber. 2001. "Skin Sensitisation, Vehicle Effects and the Local
1086 Lymph Node Assay." *Food and Chemical Toxicology* 39(6):621–27.
- 1087 Basketter, David A., Catherine Clapp, Donna Jefferies, Bob Safford, Cindy A. Ryan, Frank Gerberick,
1088 Rebecca J. Dearman, and Ian Kimber. 2005. "Predictive Identification of Human Skin Sensitization
1089 Thresholds." *Contact Dermatitis* 53(5):260–67.
- 1090 Basketter, David A., John F. McFadden, Frank Gerberick, Amanda Cockshott, and Ian Kimber. 2009.
1091 "Nothing Is Perfect, Not Even the Local Lymph Node Assay: A Commentary and the Implications for
1092 REACH." *Contact Dermatitis* 60(2):65–69.
- 1093 Dumont, Coralie, João Barroso, Izabela Matys, Andrew Worth, and Silvia Casati. 2016. "Analysis of the
1094 Local Lymph Node Assay (LLNA) Variability for Assessing the Prediction of Skin Sensitisation
1095 Potential and Potency of Chemicals with Non-Animal Approaches." *Toxicology in Vitro* 34:220–28.
- 1096 Dumont, Coralie, Pilar Prieto, David Asturiol, and Andrew Worth. 2015. "Review of the Availability of In
1097 Vitro and In Silico Methods for Assessing Dermal Bioavailability." *Applied In Vitro Toxicology*
1098 1(2):147–64.
- 1099 Enoch, S. J., J. C. Madden, and M. T. D. Cronin. 2008. "Identification of Mechanisms of Toxic Action for
1100 Skin Sensitisation Using a SMARTS Pattern Based Approach." *SAR and QSAR in Environmental
1101 Research* 19(5–6):555–78.
- 1102 EPA. 2018. *Draft Interim Science Policy: Use of Alternative Approaches for Skin Sensitization as a
1103 Replacement for Laboratory Animal Testing*.
- 1104 Fitzpatrick, Jeremy M., David W. Roberts, and Grace Patlewicz. 2017. "What Determines Skin
1105 Sensitization Potency: Myths, Maybes and Realities. The 500 Molecular Weight Cut-off: An
1106 Updated Analysis." *Journal of Applied Toxicology* 37(1):105–16.
- 1107 Fujita, Masaharu, Yusuke Yamamoto, Sayaka Wanibuchi, Yasuhiro Katsuoka, and Toshihiko Kasahara.
1108 2019. "The Underlying Factors That Explain Why Nucleophilic Reagents Rarely Co-Elute with Test

- 1109 Chemicals in the ADRA." *Journal of Pharmacological and Toxicological Methods* 96:95–105.
- 1110 Hoffmann, Sebastian. 2015. "LLNA Variability: An Essential Ingredient for a Comprehensive Assessment
1111 of Non-Animal Skin Sensitization Test Methods and Strategies." *ALTEX - Alternatives to Animal
1112 Experimentation* 32(4 SE-Short communications).
- 1113 Hoffmann, Sebastian, Nicole Kleinstreuer, Nathalie Alépée, David Allen, Anne Marie Api, Takao Ashikaga,
1114 Elodie Clouet, Magalie Cluzel, Bertrand Desprez, Nichola Gellatly, Carsten Goebel, Petra S. Kern,
1115 Martina Klaric, Jochen Kühnl, Jon F. Lalko, Silvia Martinozzi-Teissier, Karsten Mewes, Masaaki
1116 Miyazawa, Rahul Parakhia, Erwin van Vliet, Qingda Zang, and Dirk Petersohn. 2018. "Non-Animal
1117 Methods to Predict Skin Sensitization (I): The Cosmetics Europe Database." *Critical Reviews in
1118 Toxicology* 48(5):344–58.
- 1119 Kimber, I., D. A. Basketter, K. Berthold, M. Butler, J. L. Garrigue, L. Lea, C. Newsome, R. Roggeband, W.
1120 Steiling, G. Stropp, S. Waterman, and C. Wiemann. 2001. "Skin Sensitization Testing in Potency and
1121 Risk Assessment." *Toxicological Sciences* 59(2):198–208.
- 1122 Kleinstreuer, Nicole C., Sebastian Hoffmann, Nathalie Alépée, David Allen, Takao Ashikaga, Warren
1123 Casey, Elodie Clouet, Magalie Cluzel, Bertrand Desprez, Nichola Gellatly, Carsten Göbel, Petra S.
1124 Kern, Martina Klaric, Jochen Kühnl, Silvia Martinozzi-Teissier, Karsten Mewes, Masaaki Miyazawa,
1125 Judy Strickland, Erwin van Vliet, Qingda Zang, and Dirk Petersohn. 2018. "Non-Animal Methods to
1126 Predict Skin Sensitization (II): An Assessment of Defined Approaches." *Critical Reviews in Toxicology*
1127 48(5):359–74.
- 1128 Kligman, A. M. and D. A. Basketter. 1995. "A Critical Commentary and Updating of the Guinea Pig
1129 Maximization Test." *Contact Dermatitis* 32(3):129–34.
- 1130 Madden, J. C., S. Webb, S. J. Enoch, H. E. Colley, C. Murdoch, R. Shipley, P. Sharma, C. Yang, and M. T. D.
1131 Cronin. 2017. "In Silico Prediction of Skin Metabolism and Its Implication in Toxicity Assessment."
1132 *Computational Toxicology* 3:44–57.
- 1133 Myatt, G. J., E. Ahlberg, Y. Akahori, D. Allen, A. Amberg, L. T. Anger, A. Aptula, S. Auerbach, L. Beilke, P.
1134 Bellion, R. Benigni, J. Bercu, E. D. Booth, D. Bower, A. Brigo, N. Burden, Z. Cammerer, M. T. D.
1135 Cronin, K. P. Cross, L. Custer, M. Dettwiler, K. Dobo, K. A. Ford, M. C. Fortin, S. E. Gad-McDonald, N.
1136 Gellatly, V. Gervais, K. P. Glover, S. Glowienke, J. Van Gompel, S. Gutsell, B. Hardy, J. S. Harvey, J.
1137 Hillegass, M. Honma, J. H. Hsieh, C. W. Hsu, K. Hughes, C. Johnson, R. Jolly, D. Jones, R. Kemper, M.
1138 O. Kenyon, M. T. Kim, N. L. Kruhlak, S. A. Kulkarni, K. Kümmerer, P. Leavitt, B. Majer, S. Masten, S.
1139 Miller, J. Moser, M. Mumtaz, W. Muster, L. Neilson, T. I. Oprea, G. Patlewicz, A. Paulino, E. Lo
1140 Piparo, M. Powley, D. P. Quigley, M. V. Reddy, A. N. Richarz, P. Ruiz, B. Schilter, R. Serafimova, W.
1141 Simpson, L. Stavitskaya, R. Stidl, D. Suarez-Rodriguez, D. T. Szabo, A. Teasdale, A. Trejo-Martin, J. P.
1142 Valentin, A. Vuorinen, B. A. Wall, P. Watts, A. T. White, J. Wichard, K. L. Witt, A. Woolley, D.
1143 Woolley, C. Zwickl, and C. Hasselgren. 2018. "In Silico Toxicology Protocols." *Regulatory Toxicology
1144 and Pharmacology* 96.
- 1145 Natsch, Andreas. 2010. "The Nrf2-Keap1-ARE Toxicity Pathway as a Cellular Sensor for Skin Sensitizers—
1146 Functional Relevance and a Hypothesis on Innate Reactions to Skin Sensitizers." *Toxicological
1147 Sciences* 113(2):284–92.
- 1148 Natsch, Andreas, Roger Emter, Hans Gfeller, Tina Haupt, and Graham Ellis. 2015. "Predicting Skin
1149 Sensitizer Potency Based on In Vitro Data from KeratinoSens and Kinetic Peptide Binding: Global

- 1150 Versus Domain-Based Assessment." *Toxicological Sciences* 143(2):319–32.
- 1151 Natsch, Andreas, Hans Gfeller, Tina Haupt, and Gerhard Brunner. 2012. "Chemical Reactivity and Skin
1152 Sensitization Potential for Benzaldehydes: Can Schiff Base Formation Explain Everything?"
1153 *Chemical Research in Toxicology* 25(10):2203–15.
- 1154 Natsch, Andreas and Tina Haupt. 2013. "Utility of Rat Liver S9 Fractions to Study Skin-Sensitizing
1155 Prohaptens in a Modified KeratinoSens Assay." *Toxicological Sciences* 135(2):356–68.
- 1156 Nukada, Yuko, Takao Ashikaga, Hitoshi Sakaguchi, Sakiko Sono, Nanae Mugita, Morihiko Hirota, Masaaki
1157 Miyazawa, Yuichi Ito, Hitoshi Sasa, and Naohiro Nishiyama. 2011. "Predictive Performance for
1158 Human Skin Sensitizing Potential of the Human Cell Line Activation Test (h-CLAT)." *Contact*
1159 *Dermatitis* 65(6):343–53.
- 1160 OECD. 1992. "Test No. 406: Skin Sensitisation: OECD Guidelines for the Testing of Chemicals, Section 4."
1161 *OECD Publishing, Paris*.
- 1162 OECD. 2010a. "Test No. 429: Skin Sensitisation: Local Lymph Node Assay, OECD Guidelines for the
1163 Testing of Chemicals, Section 4." *OECD Publishing, Paris*.
- 1164 OECD. 2010b. "Test No. 442A: Skin Sensitization: Skin Sensitization: Local Lymph Node Assay: DA, OECD
1165 Guidelines for the Testing of Chemicals, Section 4." *OECD Publishing, Paris*.
- 1166 OECD. 2011. "Report of the Expert Consultation on Scientific and Regulatory Evaluation of Organic
1167 Chemistry Mechanism-Based Structural Alerts for the Identification of Protein-Binding Chemicals."
1168 *Series on Testing and Assessment No. 139*.
- 1169 OECD. 2014. *The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to*
1170 *Proteins*. OECD.
- 1171 OECD. 2017. *Guidance Document on the Reporting of Defined Approaches and Individual Information*
1172 *Sources to Be Used within Integrated Approaches to Testing and Assessment (IATA) for Skin*
1173 *Sensitisation*. OECD.
- 1174 OECD. 2018a. "Test No. 442B: Skin Sensitization: Local Lymph Node Assay: BrdU-ELISA or –FCM, OECD
1175 Guidelines for the Testing of Chemicals, Section 4." *OECD Publishing, Paris*.
- 1176 OECD. 2018b. "Test No. 442D: In Vitro Skin Sensitisation: ARE-Nrf2 Luciferase Test Method, OECD
1177 Guidelines for the Testing of Chemicals, Section 4." *OECD Publishing, Paris*.
- 1178 Roberts, D. W. and A. O. Aptula. 2014. "Electrophilic Reactivity and Skin Sensitization Potency of S N Ar
1179 Electrophiles." *Chemical Research in Toxicology* 27(2):240–46.
- 1180 Roberts, D. W., A. O. Aptula, M. T. D. Cronin, E. Hulzebos, and G. Patlewicz. 2007. "Global (Q)SARs for
1181 Skin Sensitisation—Assessment against OECD Principles." *SAR and QSAR in Environmental Research*
1182 18(3–4):343–65.
- 1183 Roberts, D. W., A. O. Aptula, and G. Y. Patlewicz. 2011. "Chemistry-Based Risk Assessment for Skin
1184 Sensitization: Quantitative Mechanistic Modeling for the S N Ar Domain." *Chemical Research in*
1185 *Toxicology* 24(7):1003–11.
- 1186 Roberts, D. W., R. Fraginals, J. P. Lepoittevin, and C. Benezra. 1991. "Refinement of the Relative

- 1187 Alkylation Index (RAI) Model for Skin Sensitization and Application to Mouse and Guinea-Pig Test
1188 Data for Alkyl Alkanesulphonates." *Archives of Dermatological Research* 283(6):387–94.
- 1189 Roberts, D. W. and D. L. Williams. 1982. "The Derivation of Quantitative Correlations between Skin
1190 Sensitisation and Physio-Chemical Parameters for Alkylating Agents, and Their Application to
1191 Experimental Data for Sultones." *Journal of Theoretical Biology* 99(4):807–25.
- 1192 Roberts, David W. 2018. "Is a Combination of Assays Really Needed for Non-Animal Prediction of Skin
1193 Sensitization Potential? Performance of the GARD™ (Genomic Allergen Rapid Detection) Assay in
1194 Comparison with OECD Guideline Assays Alone and in Combination." *Regulatory Toxicology and
1195 Pharmacology* 98:155–60.
- 1196 Roberts, David W. and Anne Marie Api. 2018. "Chemical Applicability Domain of the Local Lymph Node
1197 Assay (LLNA) for Skin Sensitisation Potency. Part 4. Quantitative Correlation of LLNA Potency with
1198 Human Potency." *Regulatory Toxicology and Pharmacology* 96:76–84.
- 1199 Roberts, David W., Anne Marie Api, Robert J. Safford, and Jon F. Lalko. 2015. "Principles for
1200 Identification of High Potency Category Chemicals for Which the Dermal Sensitisation Threshold
1201 (DST) Approach Should Not Be Applied." *Regulatory Toxicology and Pharmacology* 72(3):683–93.
- 1202 Roberts, David W. and Aynur O. Aptula. 2008. "Determinants of Skin Sensitisation Potential." *Journal of
1203 Applied Toxicology* 28(3):377–87.
- 1204 Roberts, David W., Aynur O. Aptula, and Grace Patlewicz. 2006. "Mechanistic Applicability Domains for
1205 Non-Animal Based Prediction of Toxicological Endpoints. QSAR Analysis of the Schiff Base
1206 Applicability Domain for Skin Sensitization." *Chemical Research in Toxicology* 19(9):1228–33.
- 1207 Roberts, David W. and Andreas Natsch. 2009. "High Throughput Kinetic Profiling Approach for Covalent
1208 Binding to Peptides: Application to Skin Sensitization Potency of Michael Acceptor Electrophiles."
1209 *Chemical Research in Toxicology* 22(3):592–603.
- 1210 Roberts, David W. and Grace Patlewicz. 2018. "Non-Animal Assessment of Skin Sensitization Hazard: Is
1211 an Integrated Testing Strategy Needed, and If so What Should Be Integrated?" *Journal of Applied
1212 Toxicology* 38(1):41–50.
- 1213 RUSSELL, W. M. S. and R. L. BURCH. 1959. *The Principles of Humane Experimental Technique*. London:
1214 Methuen & Co. Ltd.
- 1215 Schulz, K. F., D. G. Altman, and D. Moher. 2010. "CONSORT 2010 Statement: Updated Guidelines for
1216 Reporting Parallel Group Randomised Trials." *BMJ* 340(mar23 1):c332–c332.
- 1217 Sibbald, B. and M. Roland. 1998. "Understanding Controlled Trials. Why Are Randomised Controlled
1218 Trials Important?" *BMJ (Clinical Research Ed.)* 316(7126):201.
- 1219 Sumpter, Tina L., Stephen C. Balmert, and Daniel H. Kaplan. 2019. "Cutaneous Immune Responses
1220 Mediated by Dendritic Cells and Mast Cells." *JCI Insight* 4(1).
- 1221 Urbisch, Daniel, Annette Mehling, Katharina Guth, Tzutzuy Ramirez, Naveed Honarvar, Susanne Kolle,
1222 Robert Landsiedel, Joanna Jaworska, Petra S. Kern, Frank Gerberick, Andreas Natsch, Roger Emter,
1223 Takao Ashikaga, Masaaki Miyazawa, and Hitoshi Sakaguchi. 2015. "Assessing Skin Sensitization
1224 Hazard in Mice and Men Using Non-Animal Test Methods." *Regulatory Toxicology and*

- 1225 *Pharmacology* 71(2):337–51.
- 1226 Vocanson, M., A. Hennino, A. Rozières, G. Poyet, and J. F. Nicolas. 2009. “Effector and Regulatory
1227 Mechanisms in Allergic Contact Dermatitis.” *Allergy* 64(12):1699–1714.
- 1228 Wang, Chia Chi, Ying Chi Lin, Shan Shan Wang, Chieh Shih, Yi Hui Lin, and Chun Wei Tung. 2017.
1229 “SkinSensDB: A Curated Database for Skin Sensitization Assays.” *Journal of Cheminformatics*.
- 1230

Journal Pre-proof

TablesTable 1. Sources of data for the development of *in silico* methods

Database	Description
NTP-ICE	Integrated Chemical Environment (ICE), an open access database with results from NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM)
SkinSensDB	SkinSensDB is a collection of data from published literature to facilitate the development of AOP-based computational prediction methods(Wang et al. 2017)
ECHA-CHEM	European Chemicals Agency (ECHA) database is an open access database containing data for chemicals manufactured and imported in Europe. Although the summaries are publicly available, extracting data in large amounts requires special consideration as the studies are proprietary
TOXNET-HSDB	Hazardous Substances Data Bank (HSDB) is an open source database that provides information on human exposure to potentially hazardous chemicals
EURL-ECVAM-DB-ALM	The European Union Reference Laboratory for alternatives to animal testing database service on alternative methods to animal experimentation is an open access database, containing information on percutaneous absorption
CosIng	European Commission database of current and historical data for cosmetic substances and ingredients
RIFM	The Research Institute For Fragrance Materials (RIFM) monographs contain human health and toxicological data for fragrance and flavor raw materials.
Proprietary	Databases generated within a specific institution. Structure activity relationship (SAR) fingerprints
Literature	Manual curation of peer-reviewed articles and published training sets such as (MTD Cronin, 1994)

Table 2. Examples of deriving an overall assessment and confidence for the “skin sensitization *in vitro*” endpoint using the “AOP 2 out of 3” approach

Assessment and confidence scores			Overall Assessment	Explanation
Covalent Interaction with Skin Proteins	Events in Keratinocytes	Events in Dendritic cells	Skin sensitization <i>in vitro</i>	
Positive, high confidence	Positive, high confidence	Negative, low confidence	Positive, high confidence	Positive, high confidence is the majority assessment
Negative, high confidence	Negative, high confidence	Positive, high confidence	Negative, medium confidence	Negative is the majority assessment, the confidence score is lowered based on a consideration of a third high confidence result
Positive, high confidence	Positive, medium confidence	Positive, low confidence	Positive, medium confidence	Positive is the majority assessment, the confidence score is medium based on three aligned calls with different confidence scores
Negative, high confidence	Negative, medium confidence	Positive, medium confidence	Negative, medium confidence	Negative is the majority assessment, the lower confidence score of the two aligned calls propagates to the overall assessment
Negative, Low confidence		Positive, Low confidence	Inconclusive	The assessments are non-concordant and the confidence scores are aligned

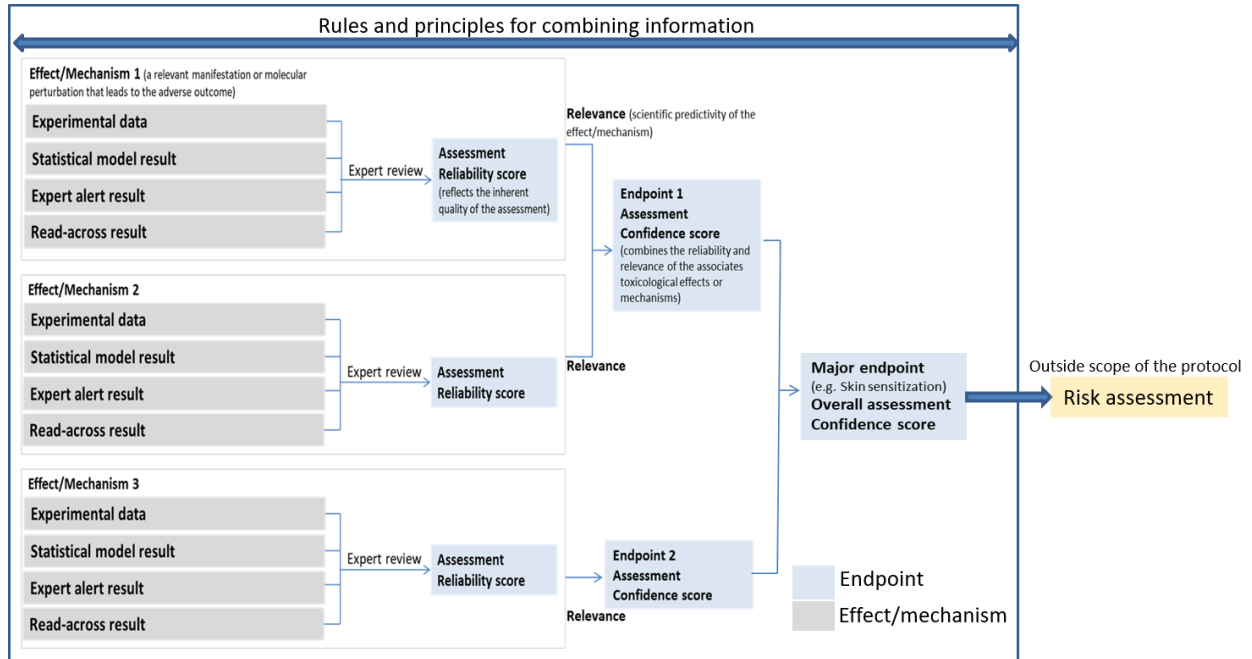
Table 3: Factors increasing and decreasing confidence in a human study (Schulz, Altman, and Moher 2010; Sibbald and Roland 1998)

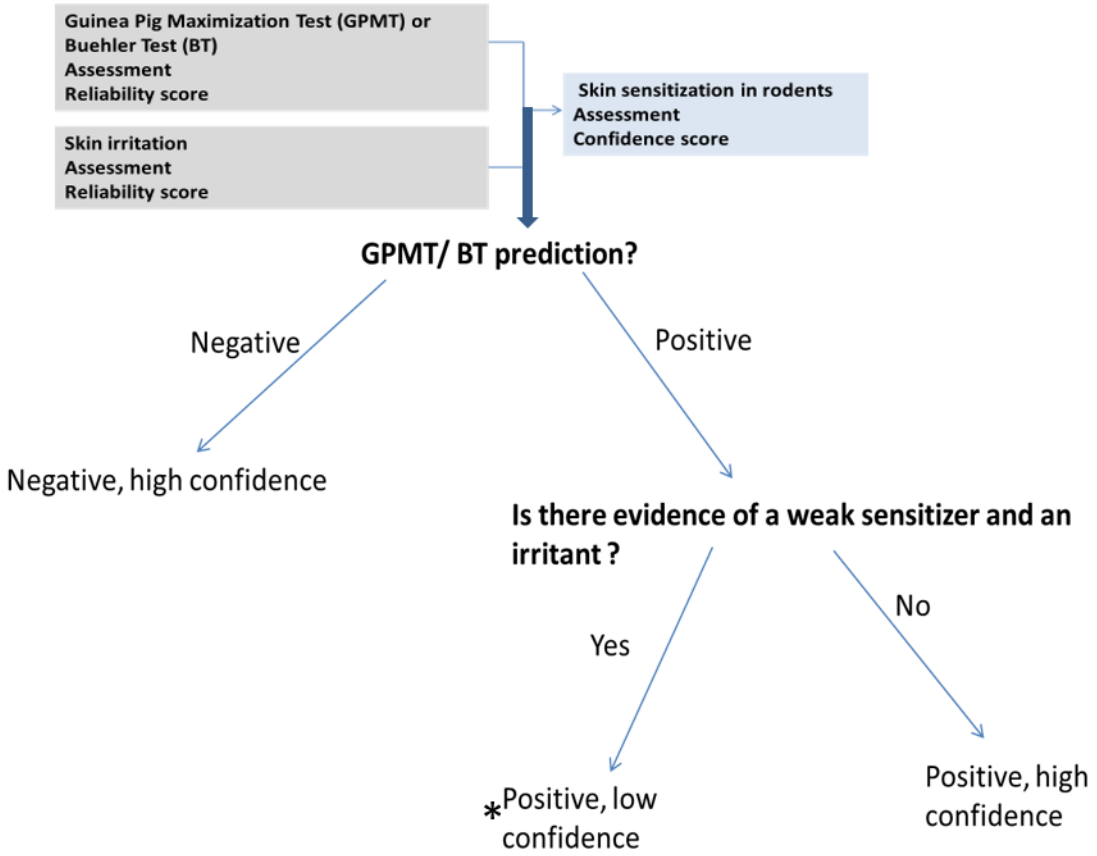
Factors increasing confidence	Factors decreasing confidence
Objective clearly stated and linked to measured outcome	Ambiguous objective, poorly linked to measured outcome
Randomized controlled study Randomized double-blind study	Uncontrolled and not randomized (or case report) No blinded control in study
Study conducted long enough to observe the effect	Study duration too short to observe the effect
Control substance application matches test substance application and represents the real-world exposure	Control substance application does not match test substance application or does not represent the real-world exposure scenario
Outcome clearly defined and measured through a quantitative endpoint	Subjective outcome based on perception
Statistical rationale behind determination of sample size	No rationale behind sample size selection
Description of study population available for review	No description of study population available

Table 4: Elements of an *in silico* toxicology report

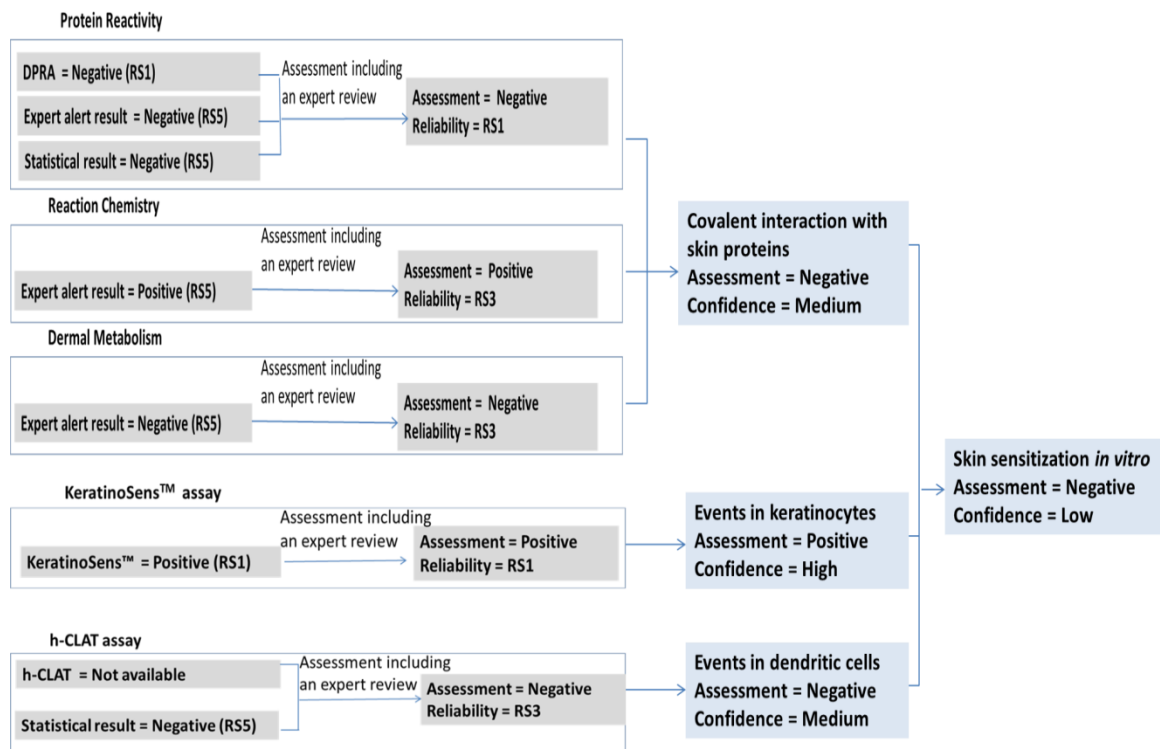
Section	Content
Title page	<ul style="list-style-type: none"> - Title (including information on the decision context) - Who generated the report and from which organization - Who performed the <i>in silico</i> analysis and/or expert review, including their organization - Date when this analysis was performed - Who the analysis was conducted for
Executive summary	<ul style="list-style-type: none"> - Provide a summary of the study - Describe the toxicity or properties being predicted - Include a table or summary showing the following: <ul style="list-style-type: none"> o The chemical(s) analyzed o Summary of <i>in silico</i> results, reviewed experimental data and overall assessment for each toxicological effect or mechanism o Summary of toxicological endpoint assessment and confidence o Summary of supporting information
Purpose	<ul style="list-style-type: none"> - Specification of the problem formulation
Materials and methods	<ul style="list-style-type: none"> - QSAR model(s), expert alerts, and other models used with version number(s) and any parameters set as part of the prediction (e.g., QMRF¹ format) - Databases searched with version number(s) - Description of any IATAs, DIPs, DAs used - Tools used as part of any read-across with version number(s)
Results of Analysis	<ul style="list-style-type: none"> - Details of the results and expert review of the <i>in silico</i> models and any experimental data, including results of the applicability domain analysis - Report of any read-across analysis, including source analogs and read-across justifications
Conclusion	<ul style="list-style-type: none"> - Summarize the overall analysis including experimental data, <i>in silico</i> methods and expert review - Final prediction that is based on expert judgment
References	<ul style="list-style-type: none"> - Complete bibliographic information or links to this information, including test guidelines referred to in the experimental data, etc.
Appendices (optional)	<ul style="list-style-type: none"> - Full (or summary) study reports used or links to the report, detailed (or summary) <i>in silico</i> reports, reports on the models used (e.g., QMRF reports)

¹QMRF – QSAR Model Reporting Format

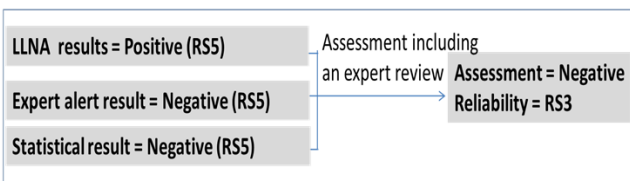




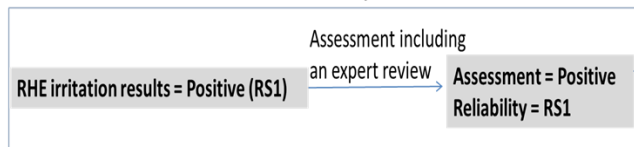
Journal



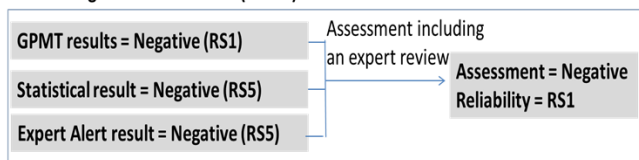
Local Lymph Node Assay (LLNA)



In Vitro Skin Irritation: Reconstructed Human Epidermis Test Method



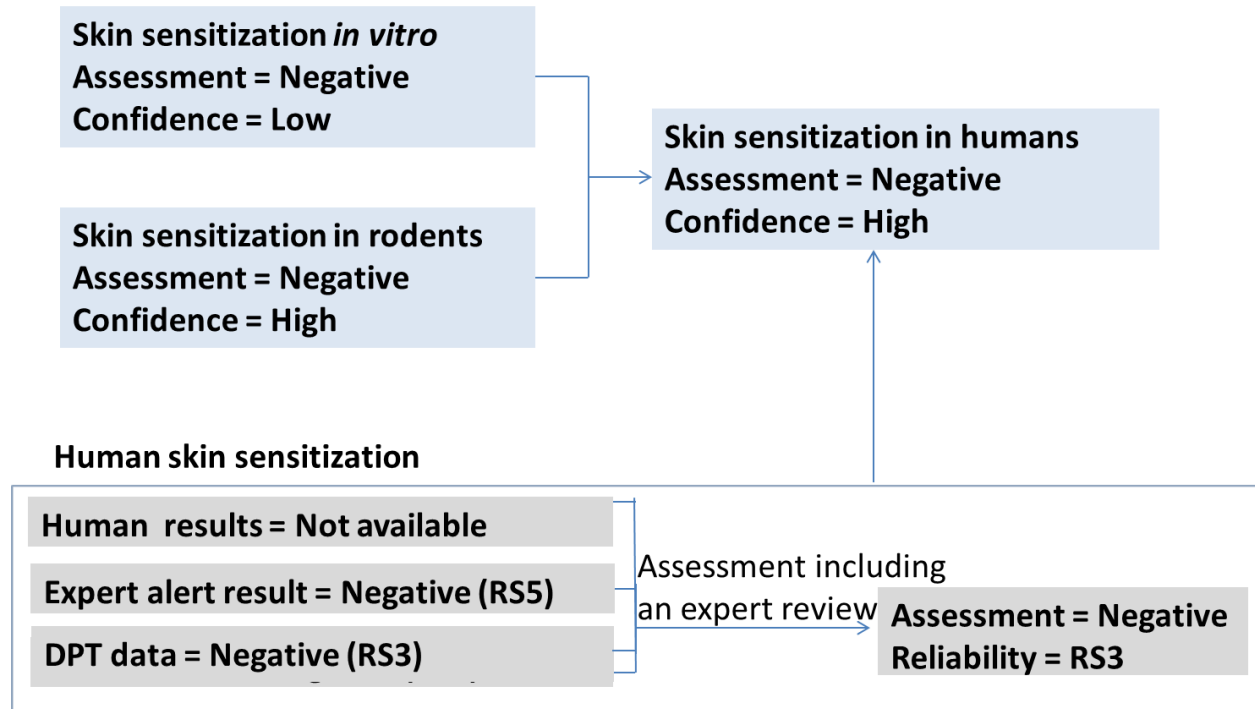
Guinea Pig Maximization Test (GPMT)

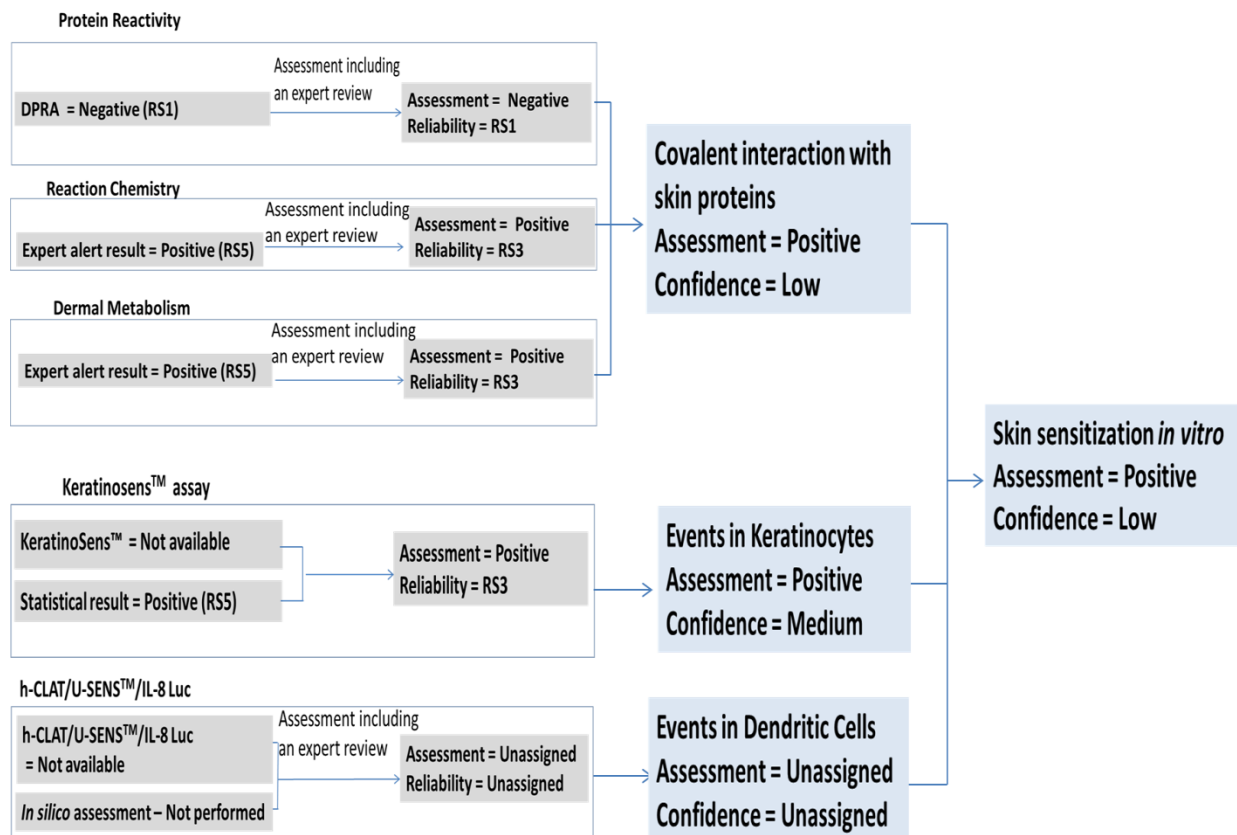


Events in rodent lymphocytes
Assessment = Negative
Confidence = Medium

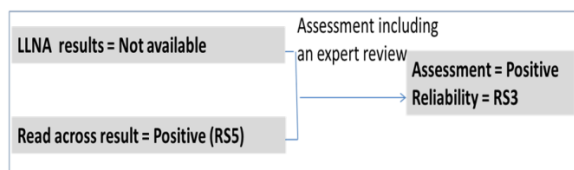
Skin sensitization in
rodents
Assessment = Negative
Confidence = High

Journal Pre-proof



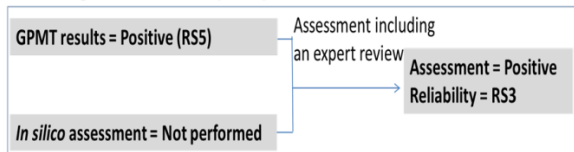


Local Lymph Node Assay (LLNA)



Events in rodent lymphocytes
Assessment = Positive
Confidence = Medium

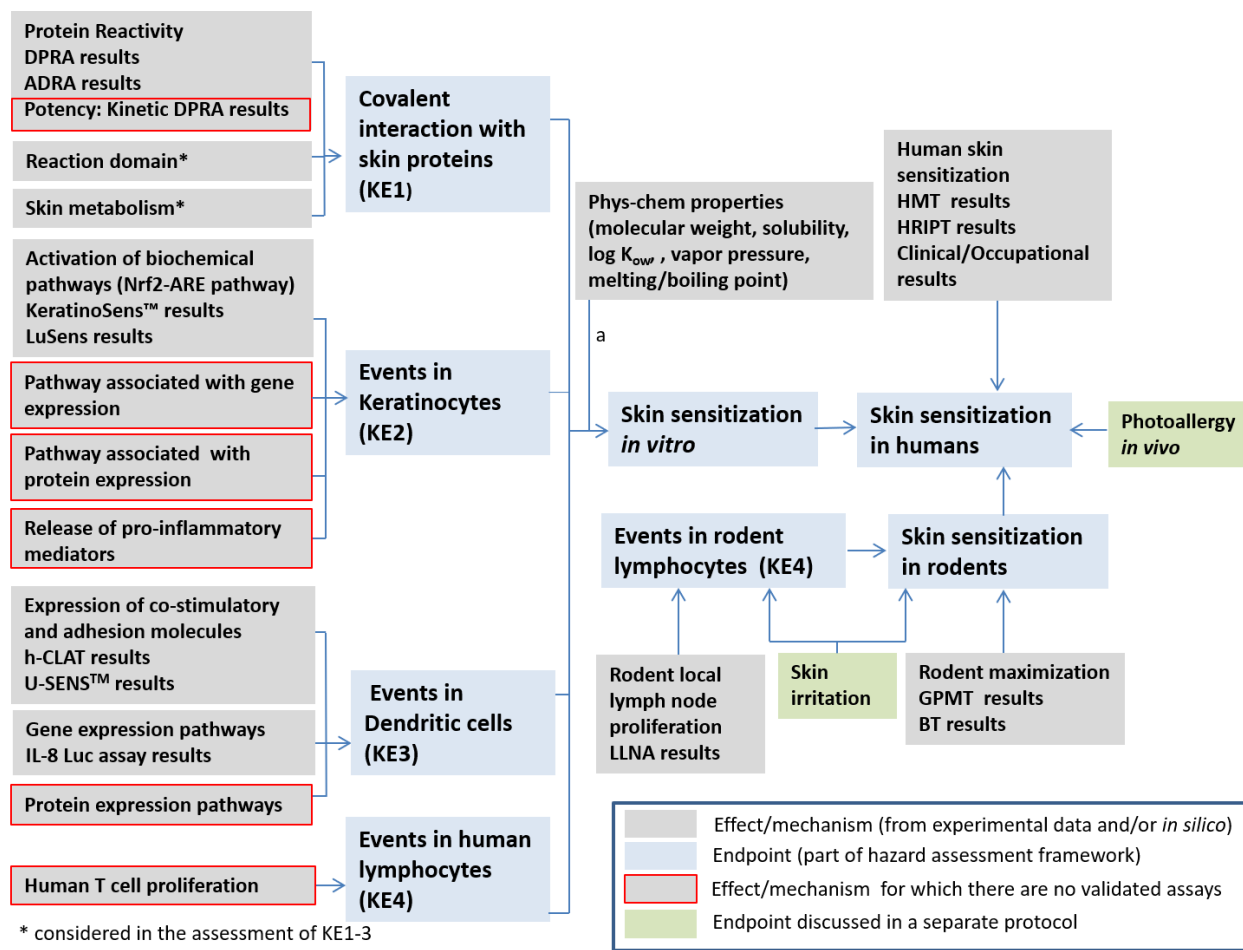
Guinea Pig Maximization Test (GPMT)

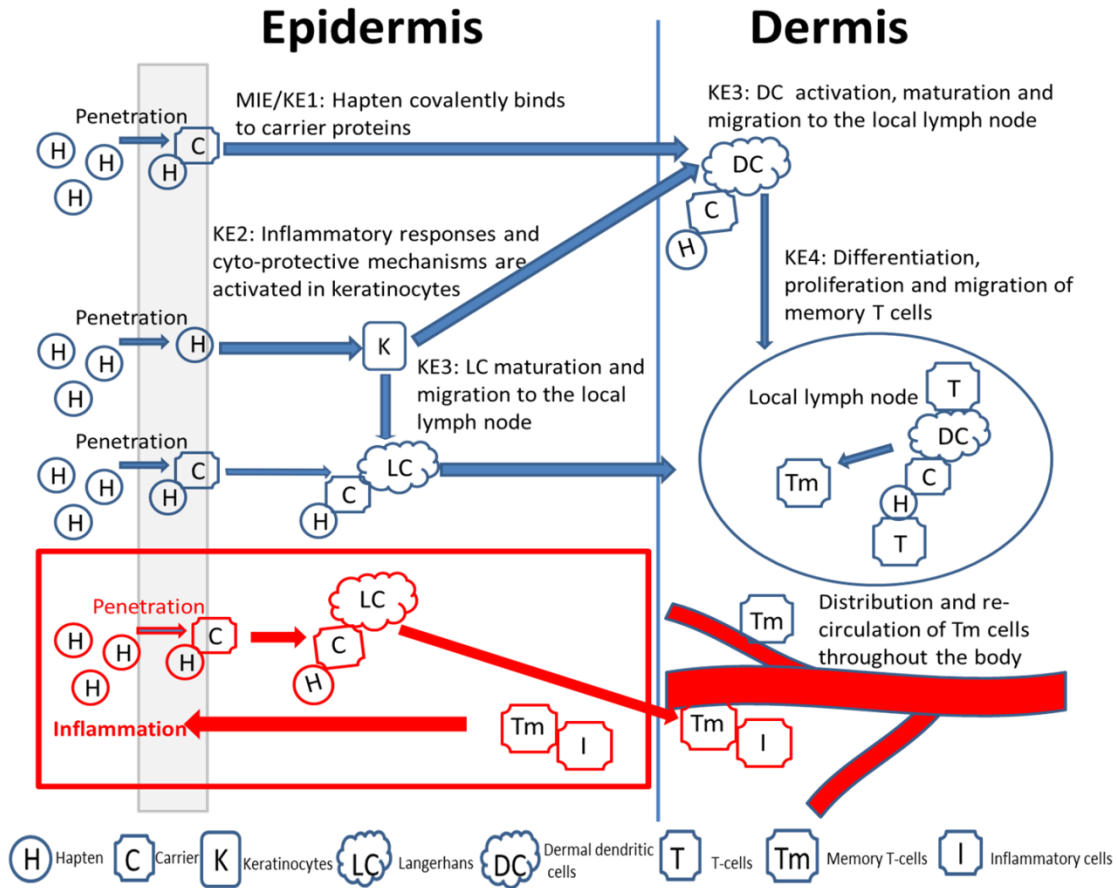


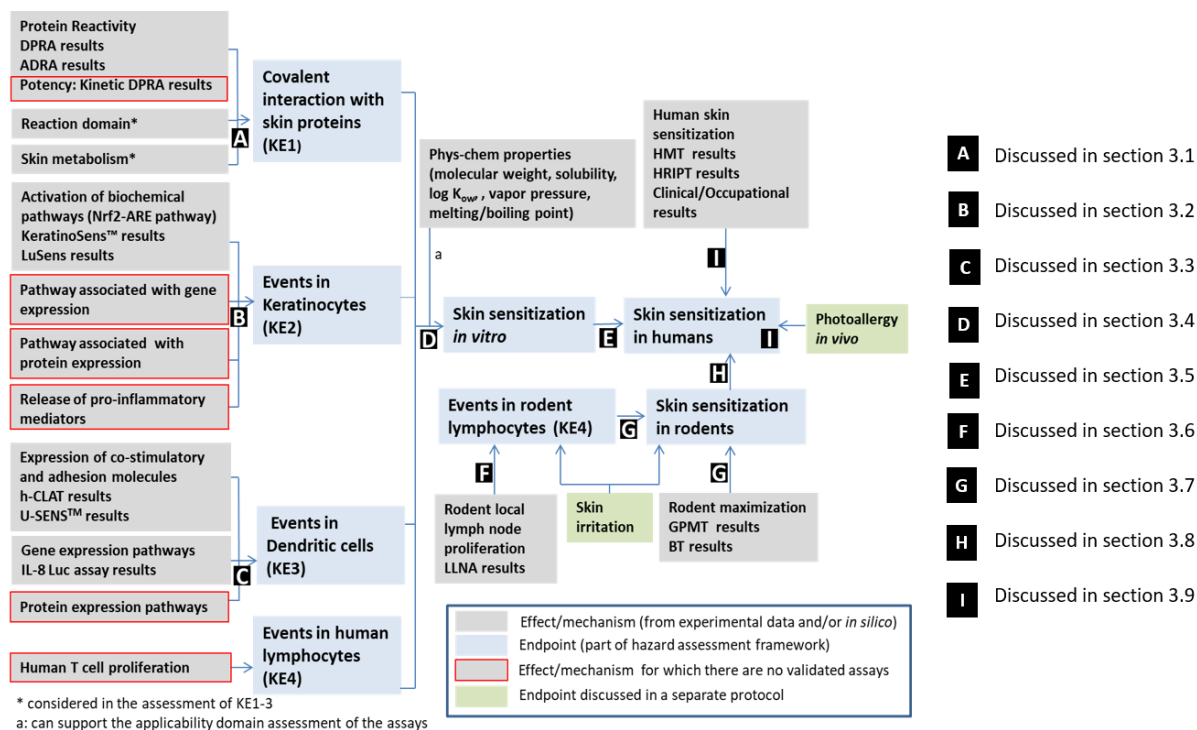
Skin sensitization in rodents
Assessment = Positive
Confidence = Medium

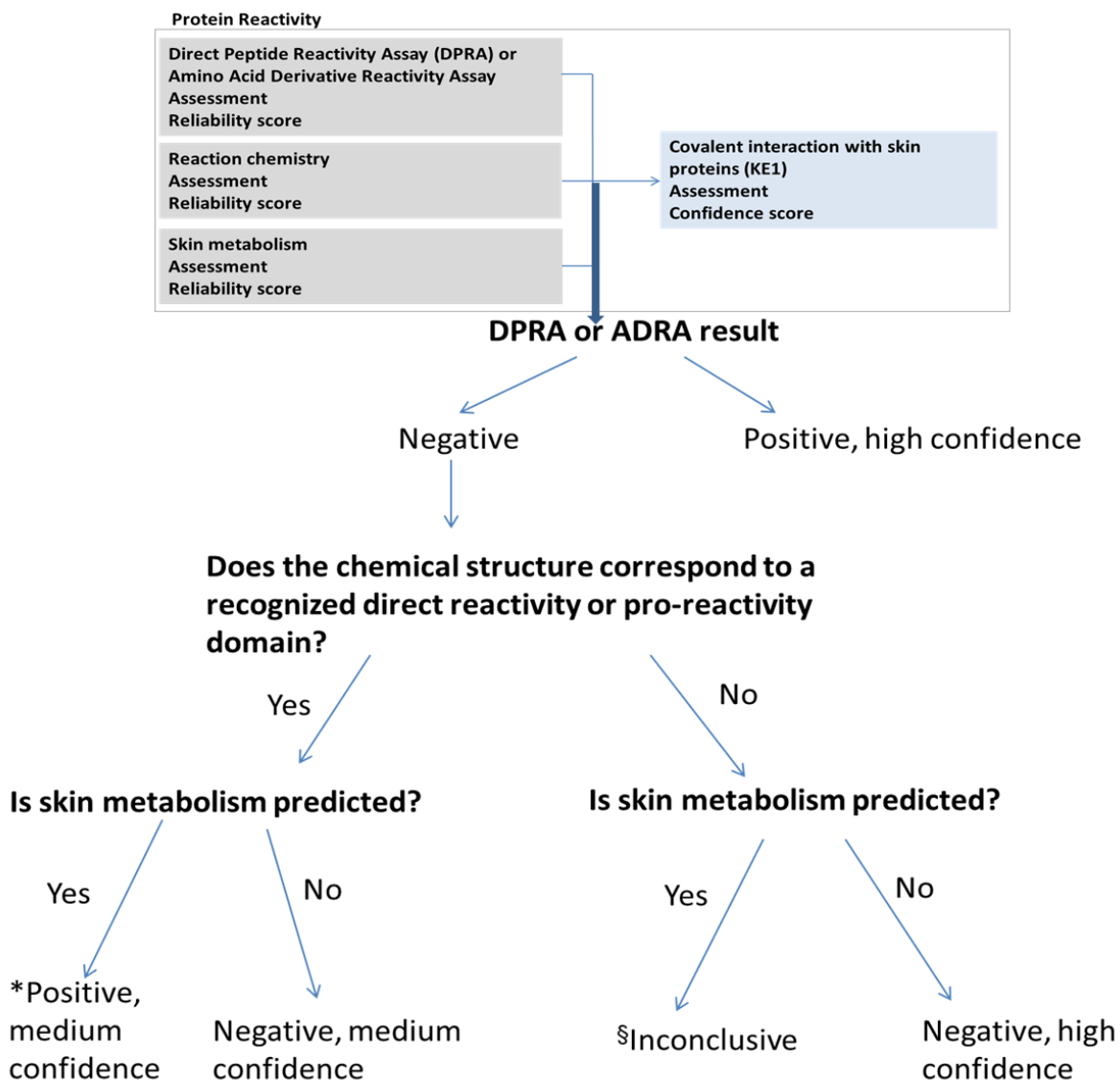
Skin sensitization in humans
Assessment = Positive
Confidence = Medium

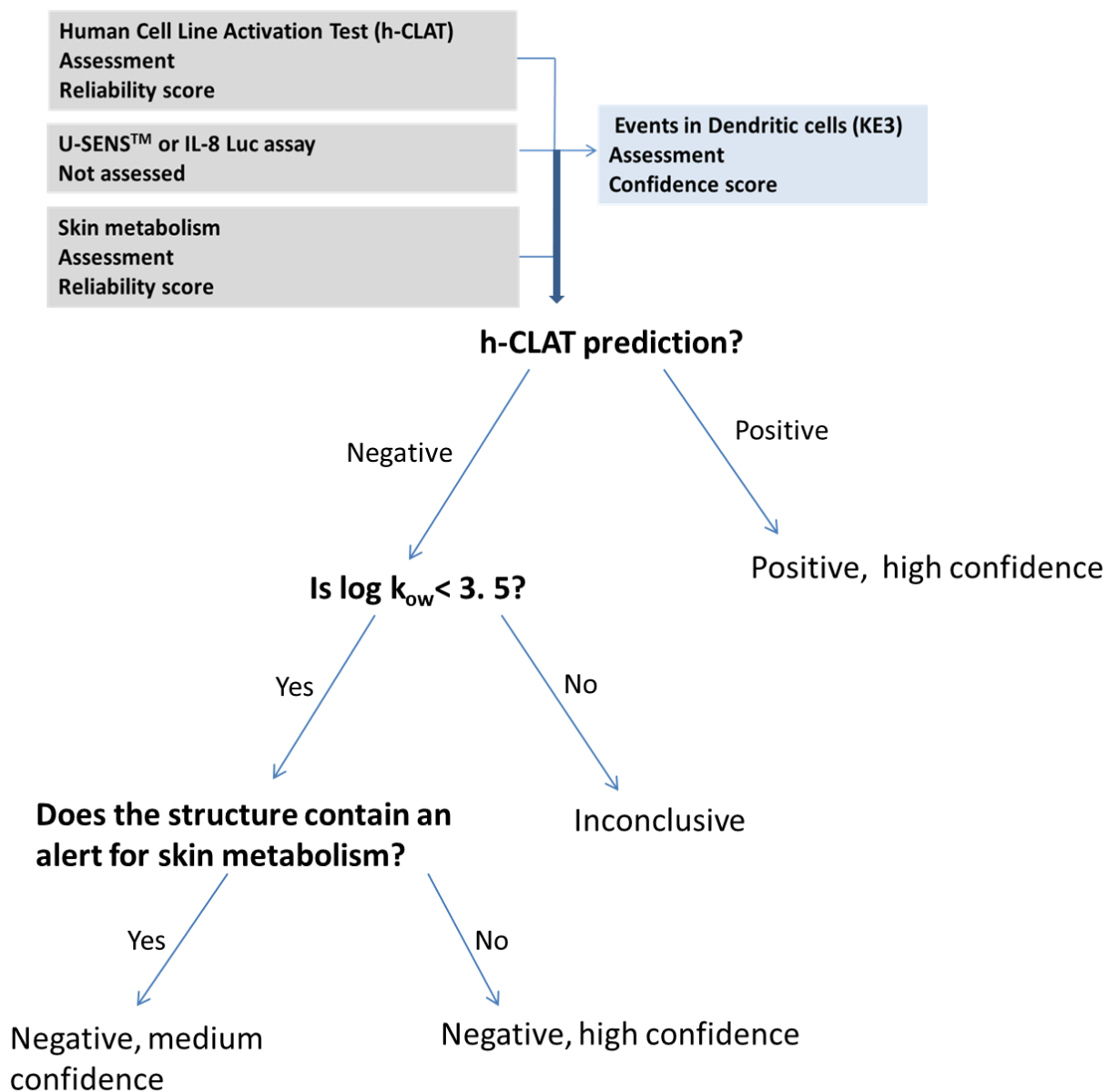
Journal Pre-proof

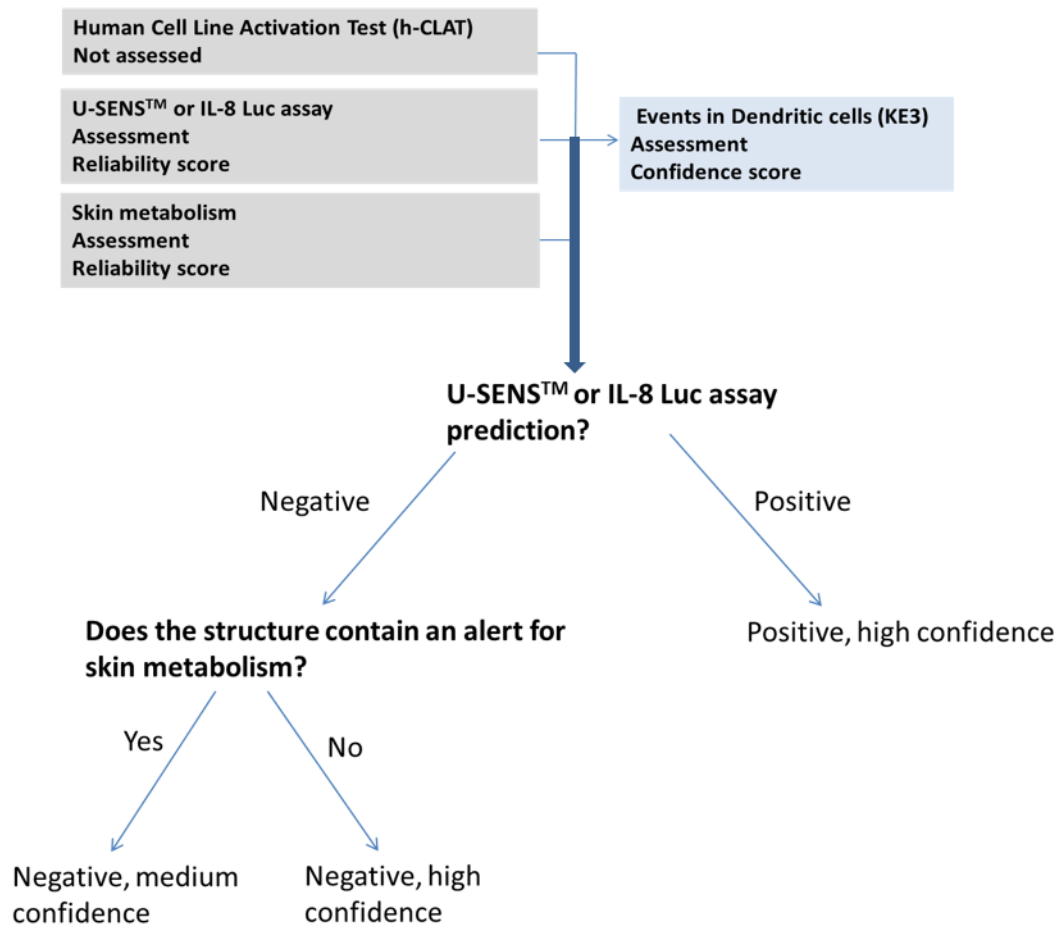


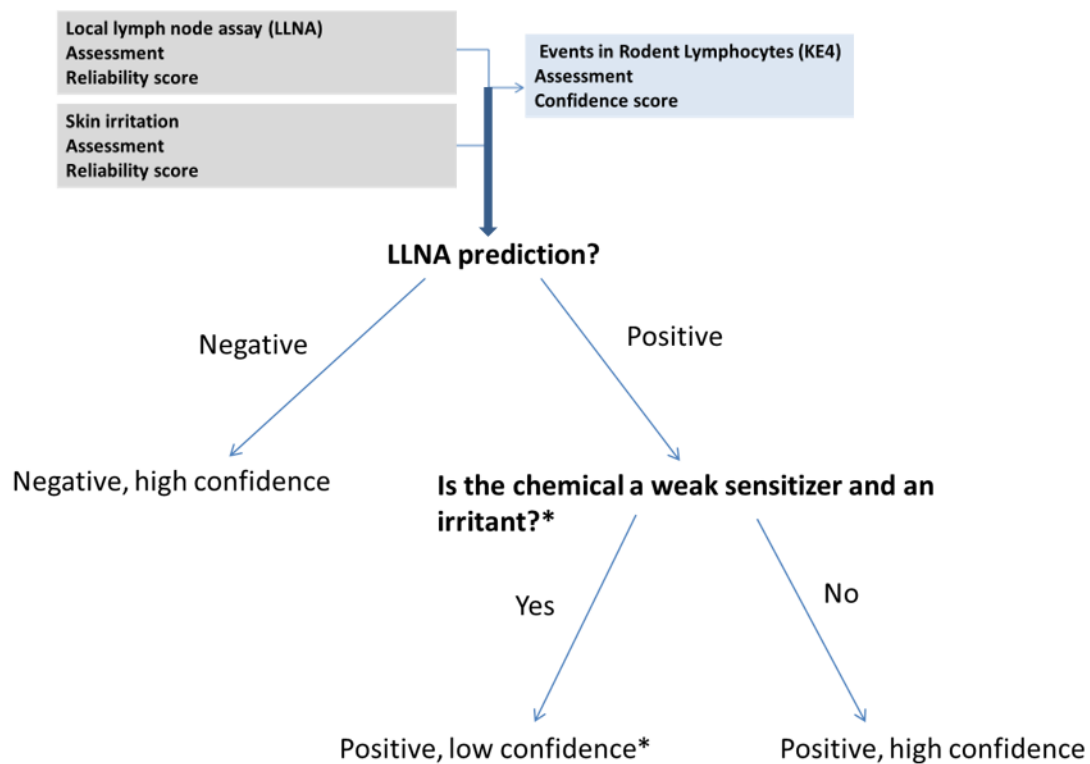




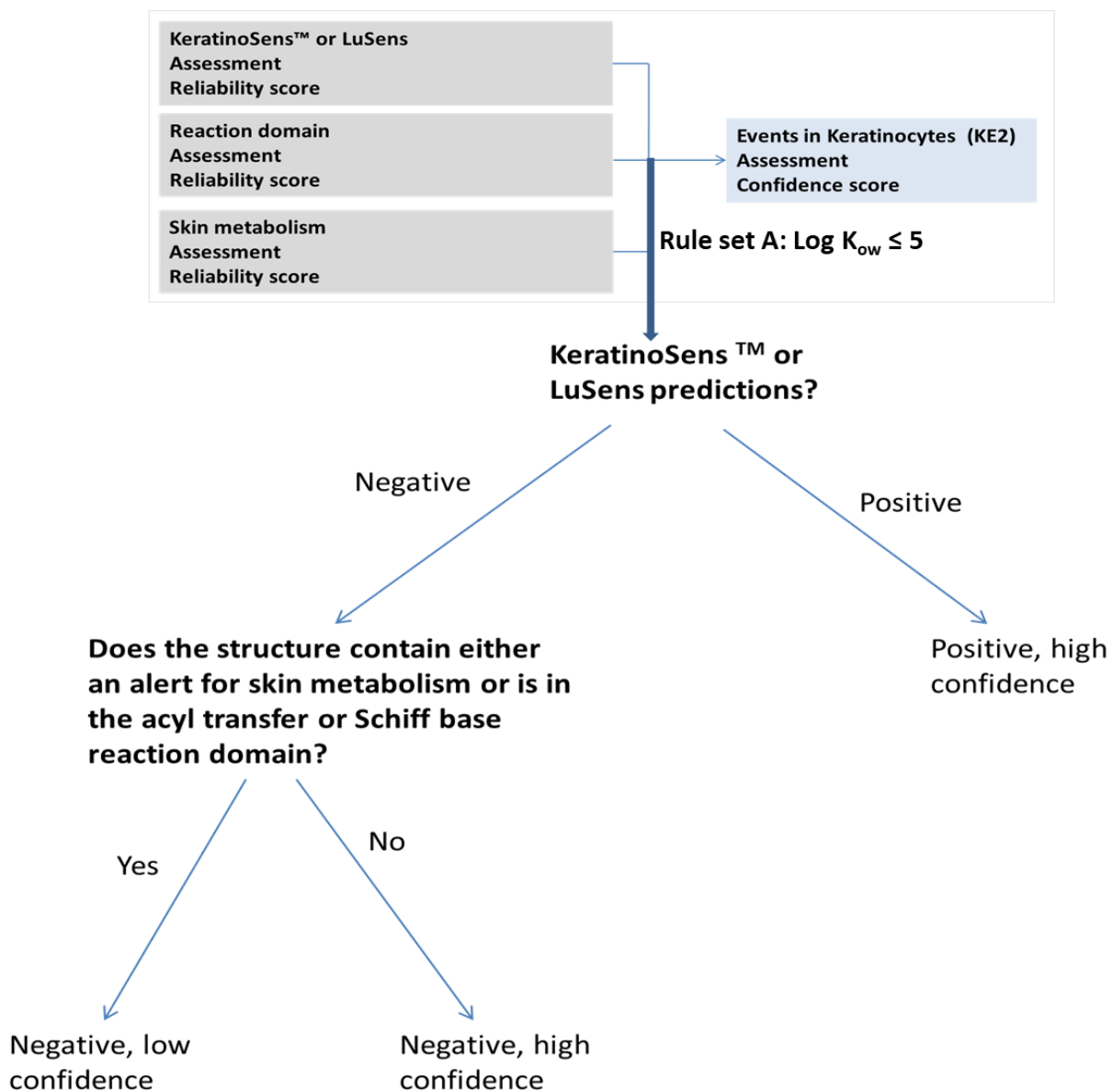


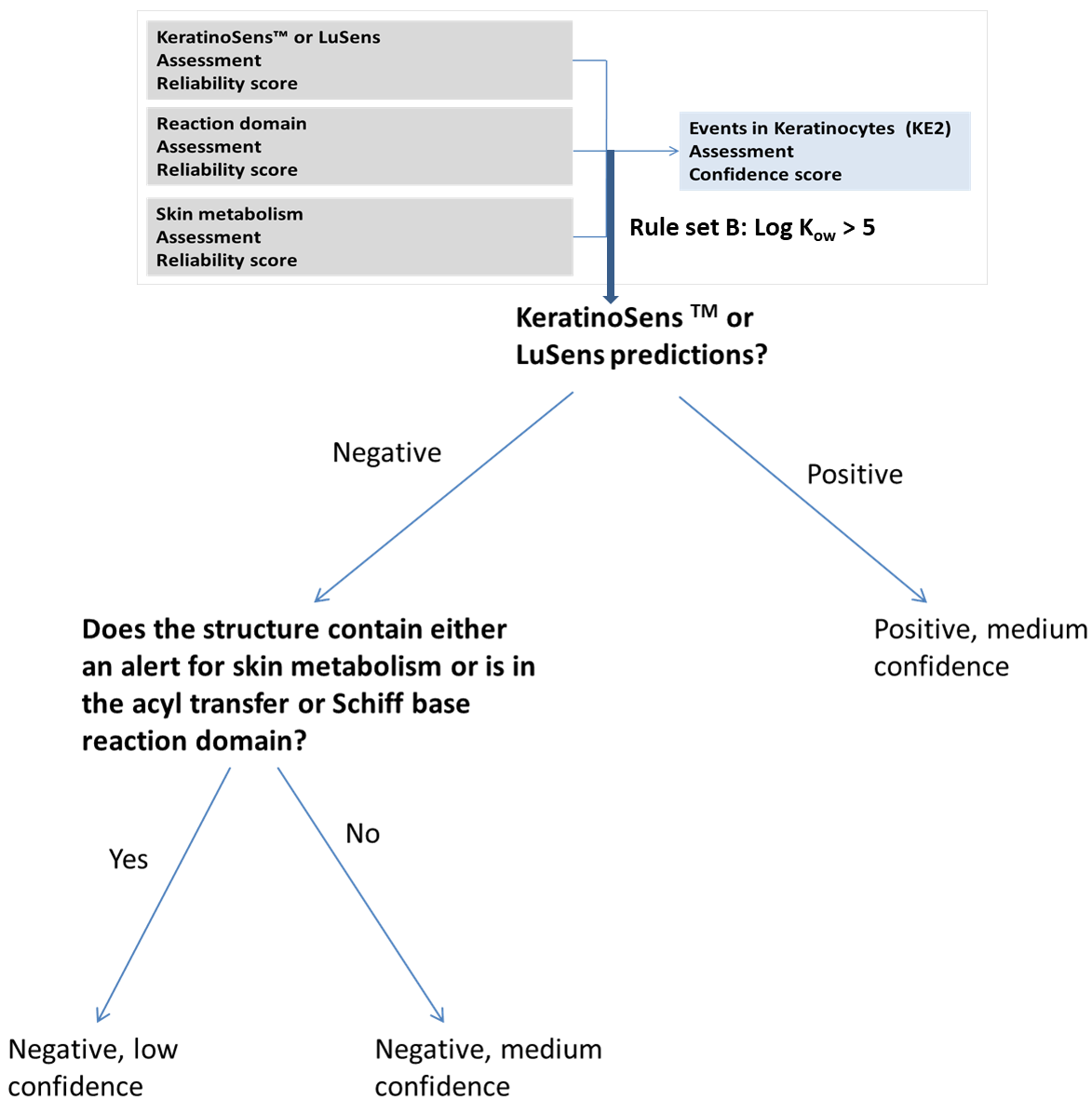






Journal Pre-proof





Highlights

1. Details a hazard assessment framework for skin sensitization that includes experimental data and *in silico* results
2. Defines rules and principles for deriving an assessment from the available information
3. Outlines criteria to be considered as part of an expert review of an assessment
4. A method for assigning confidence to skin sensitization assessments is proposed

Journal Pre-proof

Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number R44ES026909. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Journal Pre-proof

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: