PreSS/MD: Predictor of Skin Sensitization Caused by Chemicals Leaching from Medical Devices

Joyce V. B. Borba,[a,Ψ] Vinicius M. Alves,[a,Ψ] Rodolpho C. Braga,[b] Daniel R. Korn,[a]

Nicole Kleinstreuer,[c] Kevin Causey,[d] Alexander Tropsha,[a,d]

Diego Rua,[e,*] and Eugene N. Muratov[a,*].

[a] Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA.

[b] InsilicAll, Sao Paulo, SP, Brazil.

[c] National Toxicology Program, Interagency Center for the Evaluation of Alternative Toxicological Methods, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA.

[d] Predictive, LLC, Research Triangle Park, NC, USA.

[e] Division of Biology, Chemistry, and Materials Science, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland 20993, USA.

[Ψ] These authors contributed equally.

***Corresponding Authors:** U.S. Food & Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD, 20993, USA; Telephone: (240) 402-2454; E-mail: diego.rua@fda.hhs.gov; 301 Beard Hall, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA; Telephone: (919) 966-3459; FAX: (919) 966-0204; E-mail: murik@email.unc.edu.

Abstract

Safety evaluation for medical devices includes the toxicity assessment of chemicals used in device manufacturing, cleansing and/or sterilization that may leach into a patient. According to international standards on biocompatibility assessments (ISO 10993), chemicals that could be released from medical devices should be evaluated for their potential to induce skin sensitization/allergenicity, and one of the commonly used approaches is the guinea pig maximization test (GPMT). However, there is growing trend in regulatory science to move away from costly animal assays to employing New Approach Methodologies including computational methods. Herein, we developed a new computational tool for rapid and accurate prediction of the GPMT outcome that we named PreSS/MD (Predictor of Skin Sensitization for Medical Devices). To enable model development, we (i) collected, curated, and integrated the largest publicly available dataset for GPMT; (ii) succeeded in developing externally predictive (balanced accuracy of 70-74% as evaluated by both 5-fold external cross-validation and testing of novel compounds) Quantitative Structure-Activity Relationships (QSAR) models for GPMT using machine learning algorithms, including Deep Learning; and (iii) developed a publicly accessible web portal integrating PreSS/MD models that enables the prediction of GPMT outcomes for any molecules using. We expect that PreSS/MD will be used by both researchers and regulatory agencies to support safety assessment for medical devices and help replace, reduce or refine the use of animals in toxicity testing. PreSS/MD is freely available at https://pressmd.mml.unc.edu/.

Keywords: sensitization, GPMT, QSAR, deep learning,

Introduction

Sensitization is a toxicological endpoint associated with the ability of an offending chemical to cause or elicit an allergic response in some people following repeated exposures to the allergen.[1,2] Traditionally, assessing the sensitization potential for a chemical or material has relied on the use of animal models. The guinea pig maximization test (GPMT) of Magnusson and Kligman[3] and the Buehler test[4] have been predominantly used methods for more than five-decades since their original development.[3,4] Alternative assays, such as the murine Local Lymph Node Assay (LLNA), have been employed for assessing skin sensitization as well. However, more recently, regulatory agencies have been supporting the development of alternative *in vitro* and *in chemico* methods that could help reduce, refine or replace testing in animals without compromising the acceptable standards for the identification of sensitizers.[5,6]

Medical devices encompass a vast array of products intended to treat patients or diagnose diseases or other health-compromising conditions.[7] For marketing in the United States, the Food and Drug Administration (FDA) has set the definition of a medical device in Section 201(h) of the Food, Drug, and Cosmetic Act.[8] Medical devices require a pre-market biocompatibility assessment described in *Guidance for Industry and FDA Staff on Use of International Standard ISO 10993-1, Biological evaluation of medical devices - Part 1: Evaluation and testing within a risk management process.*[9] Many medical devices, such as implants and glucose meters, contain chemicals that may leach and cause toxicity.[10–12] Depending on the type and the duration of the contact with the body, a device may be evaluated for its biocompatibility, including the potential to produce localized sensitization responses.[13] Pre-market submissions for medical devices address sensitization potential with data gathered primarily with the GPMT or Buehler tests as

recommended by the International Organization for Standardization (ISO) standard 10993 Part 10.[9]

In the last several years, both our[14–16] and other[17,18] groups have developed computational models for predicting the sensitizing activity of chemicals in LLNA. In an effort to modernize the evaluation of medical devices potential for causing skin sensitization and help reduce *in vivo* animal testing, we embarked on the development of a unique open-source computational tool and web app that we named PreSS/MD (Predictor of Skin Sensitization caused by Medical Devices). We envisioned a context of use where this tool can be employed to assess the skin sensitization potential of medical devices, to supplement and, potentially, replace the experimental assessments such as animal-based tests currently accepted for regulatory submissions of medical devices. To achieve this goal, we (i) collected, curated, and integrated the largest publicly available dataset for GPMT; (ii) developed and externally validated QSAR models to predict GPMT; and (iii) incorporated GMPT models into the PreSS/MD web portal to help evaluate the skin sensitization potential for medical devices.

Materials and Methods

The workflow employed in the study is depicted in Figure 1.

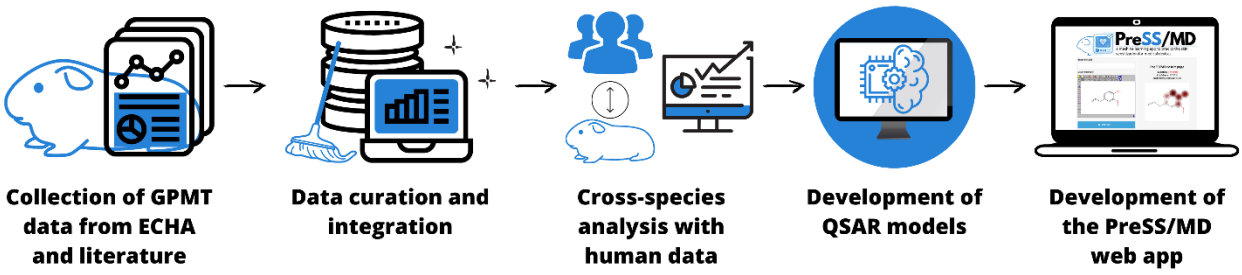

Figure 1. Key elements of the study design. See text for detailed description of each step of the workflow.

Data collection and curation

European Chemical Agency (ECHA) dataset

Experimental animal data on skin sensitization evaluated with the Guinea Pig Maximization Test (GPMT) were retrieved from the ECHA study results database (https://iuclid6.echa.europa.eu/reach-study-results). Unfortunately, there were numerous problems with the collected raw data. For instance, many numerical data were represented as string variables, the units of measurements were not standardized through the datasets, and there were many "free text" data. Therefore, we extensively cleaned and standardized all the data and converted measurements to the same units in each dataset. We also used regex expressions to find essential features for the database that were described in text format; this was key to classifying endpoints into GHS hazard categories. Following this laborious data preparation and standardization, we performed both chemical and biological data curation. After removing inconsistent data and non-modelable compounds (see Data Curation section), 1,023 out of the original 5,727 data points were kept. Among 23 duplicate chemical pairs in the dataset, biological annotations for 20 of them were concordant and for three, were discordant, *i.e.*, duplicative compounds had different annotated classifications (sensitizer *vs.* non-sensitizer). All the discordant replicates and one of each concordant replicate were removed. The final dataset comprised 995 unique chemical compounds, including 247 sensitizers and 748 non-sensitizers.

Literature

We also collected GPMT skin sensitization experimental data from the scientific literature.[19–23] After removing mixtures, inorganics, and counter ions, 701 out of the original 745 data points were kept. Only one pair of duplicates showed biological annotation disagreement

among 221 chemicals with more than one data point in the dataset. The discordant replicates were removed and only one data point for each concordant replicate was kept. Thus, the final dataset had 374 unique chemical compounds, including 173 sensitizers and 201 non-sensitizers.

Combined GPMT data from ECHA and the literature

We merged the curated data from ECHA and the research literature and examined the content of this combined data. There were 41 pairs of replicates between these two data sets, and the sensitization potential of only six of these pairs was annotated differently. These discordant records were removed, and only one record for each concordant pair of duplicates was kept. The merged data set had 1322 unique compounds including 432 sensitizers and 890 non-sensitizers, i.e., it was imbalanced with the ratio of sensitizers to non-sensitizers of approximately 1:2.

Case studies sets

An additional literature search executed identified nine new compounds with GPMT data that were not part of the training set used for model development. These compounds were standardized and used as an additional validation set. We also collected the 474 compounds available in the Extractables and Leachables Safety Information Exchange (ELSIE) Database[24] After the removal of inorganics, mixtures, and duplicates, 415 compounds remained. We found that 102 compounds were present on our GPMT list and 313 unique compounds were kept for model evaluation.

Data curation

Datasets were thoroughly curated following the workflows developed by us earlier.[25]. First, we performed chemical structure curation and removed mixtures, inorganics, and organometallic compounds, cleaned and neutralized salts, normalized the specific chemotypes, and applied the special treatment to chemicals with multiple replicated records as follows: (i) when replicated records presented the same binary outcome, only one record was kept; (ii) when a majority of replicated records presented the same binary outcome and one had different binary outcome, only one record with the agreeing binary outcome was kept, (iii) when replicated records had different binary outcomes, all of them were removed. All the curated data are available in Supplementary Material.

QSAR modeling

The modelability index (MODI)[26] was calculated to estimate the feasibility of obtaining predictive QSAR models. We developed our models following the best practices of QSAR modeling.[27] The models were developed using open-source chemical descriptors based on ECFP4-like circular fingerprints with 2048 bits and an atom radius of 2 (Morgan2) calculated in RDKit[28]. Machine learning approaches included Support Vector Machine (SVM)[29], Random Forest (RF),[30] and Light Gradient Boosting Machines (lightGBM) algorithms implemented in Scikit-learn.[31] All models were optimized using a Bayesian approach implemented in Scikit-Optimize v.0.7.4.[32] The details of hyperparameters explored in this work are available in the Supporting Information. The Bayesian optimization may be defined as follows (Equation 1):

$$P(f|D_{1:t}) \propto P(D_{1:t}|f)P(f) \quad (1)$$

where, x_i is the ith sample, and $f(x_i)$ is the observation of the objective function at x_i. The observations $D_{1:t} = \{x_{1:t}, f(x_{1:t})\}$ are accumulated. The prior distribution is combined with the likelihood function $P(D_{1:t}|f)$ of overserving $D_{1:t}$ given model f multiplied by the prior probability of $P(f)$. In doing so, Bayesian optimization finds hyperparameters that maximize the objective function (G-mean score) by building a surrogate function (probabilistic model) based on past evaluation hyperparameters of the objective.[32,33] The geometric (G)-mean was selected as the scorer since it measures the balance between classification performances on both the majority (non-toxic) and minority (toxic) classes.

The QSAR models employing deep learning were developed using Keras (https://keras.io/), a deep learning library, and Tensorflow (www.tensorflow.org), a flexible architecture that allows the deployment of calculations to desktops or servers, as backend. In addition, the following parameters of the deep learning method were optimized before model training: layer type (dense), hidden layers (3), activation function (ReLU), output layer function (sigmoid), model optimizer (Adam), loss function (binary cross-entropy). Balanced accuracy (BA) was used as a parameter to judge the performance of the models. The following hyperparameters were utilized for further deep learning training: epochs (5, 10, 50, 100) and batch size (10, 20, 40, 60, 80, 100).

The predictivity of the models were assessed by the Equations 2-7:

Balanced accuracy:

$$Balanced\ Accuracy = \frac{(sensitivity+specificity)}{2} \quad (2)$$

Sensitivity:

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

Specificity:

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

Positive Predictive Value (PPV):

$$PPV = \frac{TP}{TP+FP} \quad (5)$$

Negative Predictive Value (NPV):

$$NPV = \frac{TN}{TN+FN} \quad (6)$$

Kappa

$$Kappa = \frac{2\times(TP\times TN-FN\times FP)}{(TP+FP)\times(FP+TN)+(TP+FN)+(FN+TN)} \quad (7)$$

where TP are the true positives, FP are the false positives, TN are the true negatives, and FN are the false negatives.

Mechanistic interpretation of QSAR models

Maps of predicted fragment contribution[34,35] were generated from the QSAR models to help identify and visualize the substructure(s) predicted to provide significant contribution to the skin sensitization potential. Here, the contribution of an atom is estimated by a contribution difference obtained when the associated bits in the fingerprint corresponding to the atom are removed. Then, the normalized contributions were used to color-code the atoms in a topography-like map, in which green indicates negative contribution for toxicity (i.e., skin sensitization reduces when the atom is absent), and magenta indicating a positive contribution for toxicity (i.e., skin sensitization increases when the atom is present).[35]

Model implementation

The PreSS/MD web app was implemented on an Ubuntu Server. The app is coded using Flask (http://flask.pocoo.org), uWSGI (https://uwsgi-docs.readthedocs.org), Nginx (http://nginx.org), Python (https://www.python.org), RDKit (http://www.rdkit.org), scikit-learn (http://scikit-learn.org), and JavaScript (http://www.ecma-international.org). PreSS/MD also includes the JSME molecule editor written in JavaScript,[36] supported by the most popular web browsers. Java or Flash plugins are not required to use the app.

Results and discussions

QSAR models for predicting skin sensitization using GPMT data

High values of MODI (≥0.7) allowed us to expect that robust and predictive QSAR models could be developed for this dataset. The statistical characteristics of the skin sensitization models built and validated using GPMT data are shown in Table 3. The machine learning models built using RF, SVM, lightGBM, and Deep Learning were able to predict the external set with balanced accuracy of 73%, 74%, 70%, and 72%, respectively.

Table 1. Statistical Characteristics of QSAR Models developed for GPMT estimated on the external set.

Model	Balanced accuracy	Sensitivity	Specificity	PPV	NPV	Kappa
RF	0.73	0.84	0.63	0.53	0.89	0.41
SVM	0.74	0.70	0.79	0.62	0.84	0.47
lightGBM	0.70	0.66	0.75	0.56	0.82	0.39
Deep Learning	0.72	0.62	0.81	0.62	0.81	0.44

PreSS/MD usability

PreSS/MD has an intuitive user interface (**Figure 2**). The user may draw a molecule of interest or directly paste the query chemical structure's SMILES string in the "*molecular editor*"

box. After hitting the "Predict" button, the user will receive the predicted skin sensitization potential. These predictions are followed by the prediction's confidence, which is estimated by the ratio of predictions made by internal models,[30] the applicability domain (AD), and the maps of predicted fragment contribution.

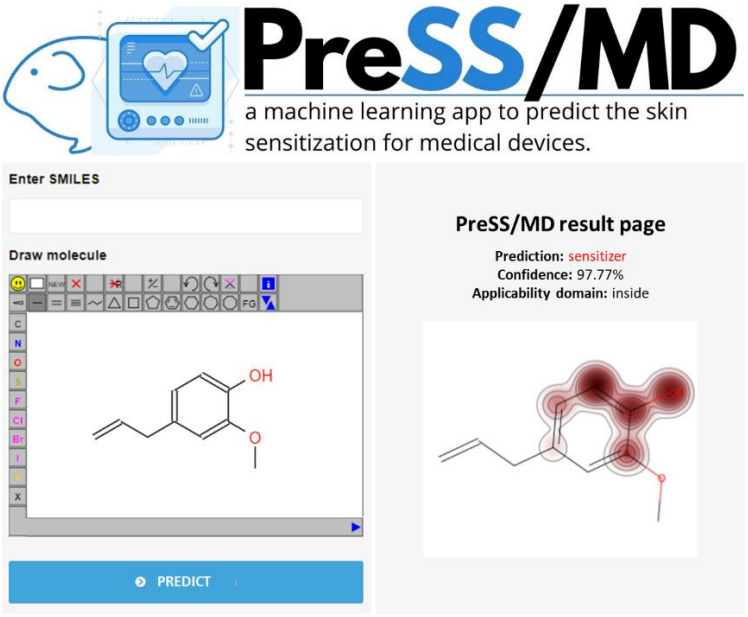

Figure 2. General capabilities of the PreSS/MD web portal, which is available at https://pressmd.mml.unc.edu/.

Case studies

As an example of a practical application, we tested PreSS/MD by employing it to predict the skin sensitization potential of nine medical device ingredients identified internally at the FDA with discordant data between GPMT and human clinical data. We compared this list with the dataset used to build our models and found that all these compounds were new and were not included in the original dataset. Therefore, we performed a blind prediction using the PreSS/MD to predict the skin sensitization potential of these nine compounds. The predicted results are shown in Table 4. PreSS/MD correctly predicted six out of nine compounds (balanced accuracy of 65%,

sensitivity of 80%, specificity of 50%, PPV of 66% and NPV of 66%). Although the evaluation of these nine compounds presented low specificity, the NPV indicates the probability of predicted non-sensitizer being truly non-sensitizers is high as 66%.

Table 2. Experimental activity and predictions for case study chemicals.

Ingredient	GPMT	Press/MD prediction
Abietic Acid	Sensitizer	Sensitizer
Ethanol	Sensitizer	Non-sensitizer
Eugenol	Sensitizer	Sensitizer
Geraniol	Non-sensitizer	Non-sensitizer
Methylparaben	Non-sensitizer	Non-sensitizer
Sulfanilic Acid	Sensitizer	Sensitizer
1,2-Dibromo-2,4-dicyanobutane	Non-sensitizer	Sensitizer
2-Methyl-3(2H)-isothiazolone	Sensitizer	Sensitizer
4,5-Dichloro-2-methyl-4-isothiazolin-3-one	Non-sensitizer	Sensitizer

In addition to these nine compounds with GPMT data, we exploited our models to predict a list of 474 chemicals known to leach from MD. After the removal of inorganics, mixtures, and duplicates, 415 compounds remained, and we found that 102 compounds were present in our curated GPMT list. Out of the 313 remaining compounds, our models predicted 98 compounds as sensitizers in the GPMT assay and 215 as non-sensitizers. We analyzed this list's overlap with the expanded skin sensitization dataset of human, LLNA, and three non-animal assays (DPRA, KeratinoSens, and h-CLAT) data described in our previous paper.[14] Out of 313 chemicals, we found that 34 had experimental data in one of the skin sensitization assays. **Table 3** shows the concordance of the predicted values using PreSS/MD and the skin sensitization potential available from experimental assays. Although the pool of compounds was small, the results show a high concordance with all assays. This high concordance suggests that integration of PreSS/MD models with non-animal methods, such as DPRA, KeratinoSens, and h-CLAT may be complementary to assess skin sensitization.

Table 3. Confusion matrices comparing PreSS/MD predictions and the experimental data from other assays for the 34 compounds from the leachable medical device list.

Experimental data		PreSS/MD predictions *Sensitizer*	*Non-sensitizer*	*Total*
Human	***Sensitizer***	3	1	4
	Non-sensitizer	1	3	4
LLNA	***Sensitizer***	6	3	9
	Non-sensitizer	6	18	24
DPRA	***Sensitizer***	3	1	4
	Non-sensitizer	1	4	5
KeratinoSens	***Sensitizer***	4	2	6
	Non-sensitizer	2	3	5
h-CLAT	***Sensitizer***	5	2	7
	Non-sensitizer	1	1	2

The use of GPMT to predict human skin sensitization.

Previously, we analyzed the correlation of LLNA and Human skin sensitization data to understand how valuable the animal model is for determining risk assessment.[37] As GPMT is still being used to check the sensitization potential of leachable chemicals from medical devices,[9] we decided to conduct a similar analysis we reported before, when comparing LLNA vs. human data.[37] Here we compared the overlap between the 1322 compounds with GPMT data and the 138 compounds with human data we previously reported elsewhere.[14] As seen in **Table 4**, 109 compounds were both tested in GPMT and had human clinical data. In total, 46 compounds were sensitizers in both tests and 41 compounds were classified as non-sensitizers in both tests, while 22 disagreed in classification. Therefore, our analysis has shown that the accuracy of using GPMT to predict human skin sensitization is estimated to have the balanced accuracy of 80%, sensitivity of 85%, PPV of 77%, specificity of 74%, and NPV of 84%. Out of the 112 compounds shown in Table 4, 14 compounds were labeled to be leaching from medical devices in the ELSIE dataset. Of these there were 9 sensitizers and 5 non-sensitizers with human data. All the non-sensitizers in

humans were also non-sensitizers in GPMT and only one sensitizer in humans was labeled as a non-sensitizer in GPMT.

Given the small number of compounds with known experimental values from both GPMT and humans, we decided to apply our previously developed QSAR models of human data[16] to the remaining 1210 compounds with GPMT data lacking human data. The use of QSAR-imputed human data allowed us to examine the possible relationships between the two endpoints for a much larger set of compounds.

Table 4. Comparison of Skin Sensitization profile of GPMT and human clinical data.

	Human		
GPMT	**Sensitizer**	**Non-Sensitizer**	**Total**
Sensitizer	46	14	60
Non-sensitizer	8	41	49
Total	54	55	**109**

A previous analysis published by Haneke et al.[38] found that GPMT had sensitivity of 70% and specificity of 100%. However, the data analyzed was much smaller, with 57 chemicals and only 3 non-sensitizers. Variability of the GPMT has been documented as dependent on the total number of animals, dosage, and grade patterns of the sensitization response considered in the test.[39] Within the extensive data collected in this work, GPMT data showed high reproducibility. In the ECHA dataset, only three pairs of compounds out of 23 duplicate chemicals had discordant annotations. The data collected from the literature had only one pair of duplicates with discordant annotations among 221 chemicals. Finally, there were 41 pairs of replicates between these two data sets, and the sensitization potential was different for only six of these pairs. Conversely,

human tests show high inter-individual variability, especially for compounds tested at a high dose, which can show weak sensitization rates in the tested populations.[40]

In our previous analysis,[14] we found the accuracy of LLNA to predict Human skin sensitization was estimated to have a balanced accuracy of 68%, sensitivity of 84%, and specificity of 52%. The low specificity means that LLNA is oversensitive to predict human skin sensitization, *i.e.*, more compounds tend to be skin sensitizers in mice than in humans. Conversely, GMPT showed higher concordance with human data, with specificity as high as 75%.

An alternative to animal testing for skin sensitization for medical devices

The GPMT was first published in 1969[3] and was considered the preferred animal method to assess skin sensitization caused by chemicals for decades. In 1989, the LLNA was first described.[41] Since then, it underwent multiple evaluations and refinements, becoming the preferred animal testing for skin sensitization after the publication of the Organisation for Economic Co-operation and Development (OECD) Testing Guideline No. 429.[42] However, international standards (ISO 10993)[43] still recommend the evaluation of chemicals released from MD for skin sensitization/allergenicity potential using the Guinea Pig Maximization Test (GPMT).[6]

Recently, Svobodová et al.[44] evaluated the sensitization potential of chemicals present in MD using a combination of *in chemico* (DPRA) and *in vitro* (LuSens) methods in comparison with the LLNA method and suggested a testing strategy for the safety assessment of medical device extracts. The authors reported an overall concordance of 63.9-82.5% between LLNA and DPRA and 80-85.4% between LLNA and LuSens. Unfortunately, no sensitivity and specificity were reported. The results shown in **Table 4** of this study reveal that there is a high concordance between GPMT and human data, which is in contrast with our previous findings showing that LLNA tends

to be oversensitive as compared to the human response.[14,37] Although GPMT shows a higher concordance with human data than the LLNA, it is important to note that GPMT requires the sacrifice of several animals[45] for each tested chemicals and, therefore, better approaches need to become available soon. Recently, the *Interagency Coordinating Committee on the Validation of Alternative Methods* (ICCVAM) published a Strategic Roadmap,[1] calling for the development of alternative approaches to reduce animal testing of chemical and medical agents. Thus, there is an expressed need to modernize the safety evaluation of MD using alternative methods, shorten the regulatory review time, and ultimately bring safer devices to the market faster.

QSAR models developed in this study and implemented in the PreSS/MD web app showed balanced accuracy of 70-74%. Although our analysis of replicates identified only six out of 41 replicated entries to disagree, a previous study has shown that dose, number of animals, and response pattern may influence in the outcome, which is evaluated by a specialist. Therefore, considering the absence of state-of-the-art predictors of GPMT as well as the variability of the assay, we suggest these models can be used to reduce the use of GPMT when used within integrated testing strategies. Moreover, since GPMT has shown higher concordance to human data than the LLNA, we suggest that QSAR models based on GPMT are more appropriate than running GPMT to assess the response to chemicals in humans.

Discussion and conclusions.

Previously, our group has developed the first QSAR models for skin sensitization based on human data.[37] Later, we employed an innovative approach using human, LLNA, and three validated non-animal assays within a Bayesian model to predict the human response.[14] This model showed higher accuracy in predicting the human response than the model built using only human

data.[14] These models were implemented in a newer version of the Pred-Skin web app.[16] Since the publication of the OECD Testing Guideline No. 429,[42] LLNA has been regarded as the preferred animal test for evaluating skin sensitization. However, the GPMT is still required for the approval of MD. For this reason, we decided to develop a separate skin sensitization web application focusing on the safety evaluation of these devices.

In order to apply *in silico* methods to predict the toxicity of MD, it is essential to note that a cornerstone in any safety evaluation of FDA-regulated products is an exposure assessment focused on actual conditions of use. Traditional methods to estimate exposure do not apply to all MD. Consequently, the medical device regulatory framework has implemented a chemical characterization and subsequent toxicological risk assessment approach. The chemical characterization involves identifying the device's component or determining chemicals that might leach into a patient during use and corresponding quantities.[46] Toxicologists use this information to conduct a risk assessment to ascertain whether any of the leachable chemicals might pose a health risk to patients at the doses quantitated. Both the chemical characterization and toxicological risk assessment for MD are generally done as recommended by the ISO standard 10993 Parts 18 and 17, respectively. PreSS/MD can predict potential leachable compounds submitted for regulatory pre-market consideration.

In summary, in this contribution we described the development of PreSS/MD, a web application to predict the skin sensitization potential of chemicals based on GPMT. This tool is the first publicly available tool based on this assay. Although non-animal assays have been explored to evaluate the potential skin sensitization effects of chemical hazards,[2] animals are still required by regulatory agencies to evaluate MD. Our results here show that GPMT has a good correlation with human data, which is higher than the murine LLNA. However, although the use of guinea

pigs is justified as their response to various skin sensitizers is similar to humans, interpretation of these assays' results requires unique expertise.[47] Moreover, the use of guinea pigs raises moral and ethical concerns, defying the principle of the 3Rs – Replacement, Refinement, and Reduction – whose goal is to identify alternative methods that utilize phylogenetically lower species, reduce the number, and refine the use of animals to lessen pain and distress.[1,48] Therefore, there is an imperative need to replace these assays. Our results show that the historical and publicly available GPMT data is sufficient to generate predictive and robust in silico models using machine learning approaches. The PreSS/MD web application fulfills an unmet need to help modernize the evaluation of skin sensitization for MD to reduce the need for animal testing. These models can be employed within integrated testing strategies to provide a weight of evidence of the sensitization potential of chemicals leaching from MD without requiring further animal tests. Moreover, we expect that the models developed in this study are applicable to estimate the toxicity of other industrial chemicals.[49] The PreSS/MD web application is publicly available at https://pressmd.mml.unc.edu/.

Data availability

All curated datasets in SDF format and the results for virtual screening of the ELSIE dataset are freely available at https://doi.org/10.6084/m9.figshare.17708714.v1.

Acknowledgments

This study was supported by NIH grant 1R43ES032371. JB thanks FAPESP for the postdoctoral fellowship and CNPq's Science without Borders program for the financial support of her visit to the University of North Carolina at Chapel Hill. VMA thanks the Lush Prize.

Competing Interests

AT, ENM, KC, and VMA are co-founders of Predictive, LLC, which develops computational methodologies and software for toxicity prediction. RCB is the CTO of InsilicAll. DRK was working on this contract as unpaid volunteer. All the other authors declare no conflicts.

References

1. ICCVAM. A Strategic Roadmap for Establishing New Approaches to Evaluate the Safety of Chemicals and Medical Products in the United States. https://ntp.niehs.nih.gov/pubhealth/evalatm/natl-strategy/index.html (2018).
2. Grundström, G. & Borrebaeck, C. Skin Sensitization Testing—What's Next? *Int. J. Mol. Sci.* **20**, 666 (2019).
3. Magnusson, B. & Kligman, A. M. The identification of contact allergens by animal assay. The guinea pig maximization test. *J. Invest. Dermatol.* **52**, 268–276 (1969).
4. Buehler, E. V. Delayed contact hypersensitivity in the guinea pig. *Arch. Dermatol.* **91**, 171–7 (1965).
5. Casati, S. *et al.* Standardisation of defined approaches for skin sensitisation testing to support regulatory use and international adoption: position of the International Cooperation on Alternative Test Methods. *Arch. Toxicol.* **92**, 611–617 (2018).
6. Daniel, A. B. *et al.* International regulatory requirements for skin sensitization testing. *Regul. Toxicol. Pharmacol.* **95**, 52–65 (2018).
7. Bronzino, J. D. *Medical Devices and Systems*. (CRC Press, 2006).
8. Kramer, D. B., Xu, S. & Kesselheim, A. S. Regulation of medical devices in the United States and European union. in *The Ethical Challenges of Emerging Medical Technologies*

41–49 (Taylor and Francis, 2020). doi:10.4324/9781003074984-3.

9. FDA. Use of International Standard ISO 10993-1, ‘Biological evaluation of medical devices - Part 1: Evaluation and testing within a risk management process’. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-international-standard-iso-10993-1-biological-evaluation-medical-devices-part-1-evaluation-and (2020).

10. Herman, A., Darrigade, A.-S., de Montjoye, L. & Baeck, M. Contact dermatitis caused by glucose sensors in diabetic children. *Contact Dermatitis* (2019) doi:10.1111/cod.13429.

11. Hansel, K. *et al.* Contact sensitivity to electrocardiogram electrodes due to acrylic acid: A rare cause of medical device allergy. *Contact Dermatitis* (2019) doi:10.1111/cod.13403.

12. Herman, A. *et al.* Allergic contact dermatitis caused by isobornyl acrylate in the Enlite glucose sensor and the Paradigm MiniMed Quick-set insulin infusion set. *Contact Dermatitis* **81**, 432–437 (2019).

13. Reeve, L. & Baldrick, P. Biocompatibility assessments for medical devices – evolving regulatory considerations. *Expert Rev. Med. Devices* **14**, 161–167 (2017).

14. Alves, V. M. *et al.* A Perspective and a New Integrated Computational Strategy for Skin Sensitization Assessment. *ACS Sustain. Chem. Eng.* **6**, 2845–2859 (2018).

15. Braga, R. C. *et al.* Pred-Skin: A Fast and Reliable Web Application to Assess Skin Sensitization Effect of Chemicals. *J. Chem. Inf. Model.* **57**, 1013–1017 (2017).

16. Borba, J. V. B. *et al.* Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. *Chem. Res. Toxicol.* **34**, 258–267 (2021).

17. Roberts, D. W., Aptula, A. & Api, A. M. Structure–Potency Relationships for Epoxides in Allergic Contact Dermatitis. *Chem. Res. Toxicol.* **30**, 524–531 (2017).

18. Toropova, A. P. & Toropov, A. A. Hybrid optimal descriptors as a tool to predict skin sensitization in accordance to OECD principles. *Toxicol. Lett.* **275**, 57–66 (2017).
19. Fedorowicz, A., Singh, H., Soderholm, S. & Demchuk, E. Structure-activity models for contact sensitization. *Chem. Res. Toxicol.* **18**, 954–69 (2005).
20. Golla, S., Madihally, S., Robinson, R. L. & Gasem, K. A. M. Quantitative structure–property relationship modeling of skin sensitization: A quantitative prediction. *Toxicol. Vitr.* **23**, 454–465 (2009).
21. Devillers, J. A neural network SAR model for allergic contact dermatitis. *Toxicol. methods* **10**, 181–193 (2000).
22. Tomlinson, S. M. *et al.* Structure-based discovery of dengue virus protease inhibitors. *Antiviral Res.* **82**, 110–4 (2009).
23. Yuan, H., Huang, J. & Cao, C. Prediction of skin sensitization with a particle swarm optimized support vector machine. *Int. J. Mol. Sci.* **10**, 3237–54 (2009).
24. ELSIE Consortium. ELSIE Database. https://www.elsiedata.org/.
25. Fourches, D., Muratov, E. & Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **56**, 1243–52 (2016).
26. Golbraikh, A., Muratov, E., Fourches, D. & Tropsha, A. Data set modelability by QSAR. *J. Chem. Inf. Model.* **54**, 1–4 (2014).
27. Cherkasov, A. *et al.* QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).
28. Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org/ http://www.rdkit.org/ (2020).
29. Vapnik, V. *The Nature of Statistical Learning Theory*. (Springer, 2000).

30. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

31. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2012).

32. Wu, J. *et al.* Hyperparameter optimization for machine learning models based on Bayesian optimization. *J. Electron. Sci. Technol.* **17**, 26–40 (2019).

33. Jiang, D. *et al.* ADMET evaluation in drug discovery. 20. Prediction of breast cancer resistance protein inhibition through machine learning. *J. Cheminform.* **12**, 16 (2020).

34. Neves, B. J. *et al.* Deep Learning-driven research for drug discovery: Tackling Malaria. *PLOS Comput. Biol.* **16**, e1007025 (2020).

35. Riniker, S. & Landrum, G. a. Similarity maps - A visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminform.* **5**, 43 (2013).

36. Bienfait, B. & Ertl, P. JSME: a free molecule editor in JavaScript. *J. Cheminform.* **5**, 24 (2013).

37. Alves, V. M. *et al.* QSAR models of human data can enrich or replace LLNA testing for human skin sensitization. *Green Chem.* **18**, 6501–6515 (2016).

38. Haneke, K. E., Tice, R. R., Carson, B. L., Margolin, B. H. & Stokes, W. S. ICCVAM evaluation of the murine local lymph node assay. Data analyses completed by the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods. *Regul. Toxicol. Pharmacol.* **34**, 274–86 (2001).

39. Andersen, K. E., Vølund, A. & Frankild, S. The guinea pig maximization test--with a multiple dose design. *Acta Derm. Venereol.* **75**, 463–9 (1995).

40. Basketter, D. & Safford, B. Skin sensitization quantitative risk assessment: A review of underlying assumptions. *Regul. Toxicol. Pharmacol.* **74**, 105–116 (2016).

41. Kimber, I. & Weisenberger, C. A murine local lymph node assay for the identification of contact allergens. *Arch. Toxicol.* **63**, 274–282 (1989).

42. OECD. Test No. 429 Skin Sensitization: Local Lymph Node Assay. *OECD Guidelines for the Testing of Chemicals* Section 4 https://www.oecd-ilibrary.org/environment/test-no-429-skin-sensitisation_9789264071100-en (2010) doi:10.1787/9789264071100-en.

43. ISO. 10993-10:2010 - Biological evaluation of medical devices — Part 10: Tests for irritation and skin sensitization. https://www.iso.org/standard/40884.html (2010).

44. Svobodová, L. *et al.* Sensitization potential of medical devices detected by in vitro and in vivo methods. *ALTEX - Altern. to Anim. Exp.* **38**, 419–430 (2021).

45. Coleman, K. P. *et al.* Evaluation of an In Vitro Human Dermal Sensitization Test for Use with Medical Device Extracts. *Appl. Vitr. Toxicol.* **1**, 118–130 (2015).

46. Albert, D. E. A practical approach to analytical chemistry of medical devices. in *Biocompatibility and Performance of Medical Devices* (ed. Boutrand, J.-P.) 49–100 (Elsevier, 2019). doi:10.1533/9780857096456.2.63.

47. ICCVAM & NICEATM. The murine local lymph node assay: a test method for assessing the allergic contact dermatitis potential of chemicals/compounds. http://ntp.niehs.nih.gov/iccvam/docs/immunotox_docs/llna/llnarep.pdf (1999).

48. Tannenbaum, J. & Bennett, B. T. Russell and Burch's 3Rs then and now: The need for clarity in definition and purpose. *J. Am. Assoc. Lab. Anim. Sci.* **54**, 120–132 (2015).

49. Alves, V. M. *et al.* Chemical toxicity prediction for major classes of industrial chemicals: Is it possible to develop universal models covering cosmetics, drugs, and pesticides? *Food Chem. Toxicol.* **112**, 526–534 (2018).