method because respective toxicological studies are independent variables. The authors should comment on this concern as it directly questions the approach used to assess the accuracy of model predictions.

Also, there are 2 additional considerations that pertain to the modeling and model validation. First, it is unclear whether the authors have removed descriptors with missing values. Imputation is often performed for the datasets with missing variables; however, it is not clear if this was the case. Second, Y-randomization is a standard protocol used in the validation of QSAR models wherein models are built from randomly permuted data to test their validity and guard against chance descriptor correlations (Tropsha, 2010). It is not clear if this procedure was done by (Luechtefeld *et al*., 2018).

In conclusion, we are in full support of developing and using computational models as an alternative to animal testing (Bell *et al*., 2017; Hartung and Hoffmann, 2009). It is exciting that the issue of alternatives to animal testing receives high level of attention in both print and social media. However, because of heightened attention to the reproducibility in biomedical research (Collins and Tabak, 2014; Miller, 2014), it is important to ensure that bold claims are well justified, supported by carefully vetted data, and follow best scientific practices. Our conclusion is that it is difficult to accept RASAR model accuracy as stated in (Luechtefeld *et al*., 2018). Accurate prediction of adverse chemical effects is a critically important challenge; therefore, we hope that the authors would clarify ambiguities of their study highlighted in this letter.

SUPPLEMENTARY DATA

Supplementary data are available at *Toxicological Sciences* online.

FUNDING

National Institute of Environmental Health Sciences, (Grant/Award Number: P42 ES027704).

REFERENCES

Available as a Supplementary File.

Vinicius M. Alves,[*] Joyce Borba,[*,†] Stephen J. Capuzzi,[*] Eugene Muratov,[*,‡] Carolina H. Andrade,[†] Ivan Rusyn,[§,1] and Alexander Tropsha[*,1]

[*]UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599; [†]Faculty of Pharmacy, Federal University of Goias, Goiania, Goias 74605-170, Brazil; [‡]Odessa National Polytechnic University, Odessa 65000, Ukraine; and [§]College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, Texas 77843

[1]To whom correspondence should be addressed at College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, 4458 TAMU, College Station, TX 77843. E-mail: irusyn@tamu.edu and UNC Eshelman School of Pharmacy, University of North Carolina, 100K Beard Hall, Chapel Hill, NC 27599. E-mail: alex_tropsha@unc.edu.

doi: 10.1093/toxsci/kfy286
Advance Access Publication Date: November 30, 2018

Letter to the Editor

Missing the Difference Between Big Data and Artificial Intelligence in RASAR Versus Traditional QSAR

The letter to the editor by Alves *et al*. (2018a) was a surprise to the authors of "*Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility*" (Luechtefeld *et al*., 2018a). The letter challenges the approach as one would challenge a traditional QSAR, by which it ignores many attributes and consequences of the RASARs construction and performance as an implementation of big data and artificial intelligence (machine learning) (Hartung, 2016; Luechtefeld and Hartung, 2017).

To state it simply: the RASAR models are not traditional QSARs, wherein a highly curated, small training dataset is used to predict a single property based on chemical descriptors, ie, classifications per hazard. The published model uses data on 100 000+ chemical structures, calculates 5 billion+ similarities, and simultaneously makes 190 000 predictions for nine hazards of toxic properties of chemicals: 87% are correct, which should raise the question what we got right, not what we got wrong?

Regulators are already considering how best to integrate these new kinds of machine learning models into their decisions. We refer also to our attempts to carry out now an external validation by the United States validation body Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM). Notably, the corresponding author held earlier responsibilities for such validations in Europe including *in silico* approaches (Worth *et al*., 2004) and has continuously contributed to QSAR validation (Hartung *et al*., 2004).

WHAT IS SIGNIFICANT ABOUT THE ACCURACIES DEMONSTRATED IN OUR MODEL?

The RASAR model is exciting because of the ability for one "composite" model to estimate many endpoints and integrate datasets. Similar work in machine learning has demonstrated the benefit of simultaneously modeling multiple endpoints via "transfer learning". In practice, this may enable modeling of data-sparse endpoints (a critical need in toxicology).

None of our individual published metrics far exceed previous demonstrations. The Alves *et al*. letter authors themselves published models with comparable accuracies (95% correlation on aquatic endpoints, 82% balanced accuracy for skin sensitization, 88%–92% accuracy on Ames test). The recent NTP acute oral toxicity event received an NIH National Center for Advancing

Translational Sciences (NCATS) submission with 90%+ balanced accuracy (https://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/acute-systemic-tox/models/index.html, last accessed December 10, 2018). RASAR received an 85% balanced accuracy. What is remarkable of our model is the 100% coverage for nine hazards with throughout high ballanced accuracies.

HOW CAN OUR MODEL OUTPERFORM INDIVIDUAL OUTCOMES OF EXPERIMENT?

We model chemical hazard classification labels, not the observed adverse effects in toxicological tests. Weight of evidence (WoE) is a commonly used approach for labeling chemicals by weighing and combining test outcomes. This simple approach outperforms experimental uncertainty at the cost of multiple tests. RASAR models improve on WoE via machine learning and the use of more features.

WHAT MAKES CONDITIONAL PROBABILITY USEFUL FOR MEASURING EXPERIMENTAL REPEATABILITY?

Measuring experimental repeatability is a rater reliability question. We used the simple method of joint probability of agreement. The 1960 paper by Cohen (30.527 citations) and another by Fleiss (1971) describe related approaches in more detail. The RASAR publication focused on modeling and does not give a complete statistical treatment of repeatability.

The Alves letter asks how tests done by independent labs can be conditionally dependent. Repeat tests of the same chemical are conditionally dependent because they share a common cause in the property being tested, and the property being tested is unobserved.

As a last note, average repeatability fails to account for dependence of test repeatability on the chemicals being tested. A future paper may evaluate animal test repeatability in greater detail, the RASAR paper is complex enough without a complete discussion of repeatability.

WOULD OUR MODEL VALIDATION BENEFIT FROM Y-RANDOMIZATION?

RASAR models train on Amazon computing clusters. Each training results in greater cloud computing cost. Although valuable, Y-randomization requires additional runs and would increase cost. Recent updates to the RASAR model increase training efficiency through GPU processing (https://developer.nvidia.com/cudnn). Future work should be able to integrate this method.

ON THE USE OF CLASSIFICATION AND LABELING DATA

The Alves *et al.* letter suggests that models should "*only [use] experiment-derived results*". The RASAR publication reports on models of ECHA classification and labeling. C&L are property assignments to chemicals based on criteria defined by the REACH legislation and ECHA. A large repository of regulatory chemical classifications certainly makes a valuable modeling endpoint. This data is sometimes noisy, but remains an excellent choice for toxicological modeling. Modeling experimentally derived results is an important and different exercise.

In their editorial letter, our colleagues state "*Unfortunately, neither the exact modeling dataset, nor the descriptors used by* (Luechtefeld *et al.*, 2018a) *have been made publicly available, precluding independent evaluation of the quality of both the data and the models*". The practical truth is that sometimes data cannot be shared. Unfortunately, ECHA prohibits the publication of the core dataset (Luechtefeld *et al.*, 2016). We think the modeling community can understand restrictions on data sharing, especially when the restrictions are imposed by the regulatory or industrial entity, which owns the data, as is the case here.

FINAL NOTE

The authors are happy to discuss these methods with any interested researcher. Yet, the critiques contained in the Alves *et al.* letter were made without beta-testing or test outcomes, and are not an evidence-based assessment. We have a history of sharing relevant information and collaborating with our fellow researchers. In fact, the ECHA dataset referenced in the original RASAR publication had been shared with the authoring UNC group 2 years ago leading to a joint publication (Alves *et al.*, 2018a). Our offer to test the software made independently to both senior authors before their editorial letter was not taken up. Moreover, Underwriter's Laboratories, who supports these models, has an open policy for beta-testing.

Finally, the editorial letter states an urgent need to address concerns based on the use of models in decision-making. In our experience, regulators are extremely careful in the use of new methods. Moreover, we are very much aware of the complex process for model validation and regulatory acceptance (eg, ICCVAM, the Toxic Substances Control Act (TSCA)). We appreciate many of the sentiments that authors of the letter to the editor shared. We agree that it is important to ensure claims in scientific publications are well justified and follow best modeling practices. Such validation is necessary for the proper growth and use of computational methodologies within toxicology and chemistry. We invite any researcher or regulatory entity to contact us with any similar or new questions, concerns, or ideas. Continuing the dialog will only further advance the field.

FUNDING

European Commission (Grant/Award Number: 681002).

REFERENCES

Available as a Supplementary File.

Thomas Luechtefeld,* Dan Marsh,* and Thomas Hartung[†,‡,1]

*ToxTrack, Baltimore, Maryland 21209; [†]Center for Alternatives to Animal Testing (CAAT), Johns Hopkins University, Bloomberg School of Public Health, Baltimore, Maryland 21205; and [‡]CAAT-Europe, University of Konstanz, Konstanz 78464, Germany

[1]To whom correspondence should be addressed. Fax: (410) 614-2871. E-mail: thartung@jhu.edu.

doi: 10.1093/toxsci/kfy287
Advance Access Publication Date: November 30, 2018