SYSTEMATIC OMICS ANALYSIS REVIEW TOOL TO SUPPORT RISK ASSESSMENT

by

Emma R. McConnell

Dr. Richard Di Giulio, Adviser

May 2013

Masters project submitted in partial fulfillment of the requirements for the Master of Environmental Management degree in the Nicholas School of the Environment of Duke University

2013

ABSTRACT

Environmental health risk assessors are challenged to understand and incorporate new data streams as the field of toxicology continues to adopt new molecular and systems biology technologies. Systematic screening reviews can help risk assessors determine which studies to consider for inclusion in a human health assessment. A tool for systematic reviews should be standardized and transparent in order to consistently determine which studies meet minimum quality criteria prior to performing in-depth analyses of the data. The Systematic Omics Analysis Review (SOAR) tool is a spreadsheet of 35 objective questions developed by domain experts, focused on transcriptomic microarray studies, and including four main topics: test system, test substance, experimental design, and microarray data. The tool will be used as a guide to identify studies that meet basic published quality criteria, such as those defined by the Minimum Information About a Microarray Experiment (MIAME) standard and the Toxicological data Reliability assessment Tool (ToxR Tool). Seven scientists were recruited to test the tool by using it to rate 19 published manuscripts that study chemical exposures with microarrays. Using their feedback, questions were weighted based on importance of the information and a suitability cutoff was set for each of the four topic sections. The final validation resulted in 100% agreement between the users on four separate manuscripts, showing that the SOAR tool may be used to facilitate the standardized and transparent screening of microarray literature for environmental human health risk assessment.

Introduction

Government agencies and environmental consultants develop human health risk assessments to determine the potential exposure and toxicity risks of chemicals, a process which involves consideration of all of the available published scientific literature on that chemical. Experts evaluate and integrate the studies that are available, make judgments on the quality of the science, and choose appropriate studies to derive cancer or noncancer toxicity reference values. A National Academy of Science Committee reviewing the Environmental Protection Agency's draft Integrated Risk Information System (IRIS) Toxicological Review of Formaldehyde recommended that the IRIS Program develop "clear concise statements of criteria" when choosing studies to exclude or include for toxicity reference value calculations (NRC, 2011).

Significant work has been done by authors such as Fostel et al (2007) and Schneider et al (2009) to determine the criteria that are crucial for understanding the quality and reproducibility of toxicological studies in general. However, these criteria are not designed for use with transcriptomic studies, and are not adequate to provide an assessment of the entire study. Microarrays, one of many transcriptomic tools, are vastly different than the whole-animal toxicity studies that risk assessors are accustomed to evaluating. In acknowledgement of the complicated and varied procedures and analysis required to perform a microarray experiment, the gene expression microarray community created the "Minimum Information About a Microarray Experiment" (MIAME) (FGDS, 2010) standard, along with data reporting requirements that have been adopted by several journals. Though this is a community standard for transcriptomic microarrays, it does not specifically consider their application to toxicogenomic studies for the purpose of human health risk assessment.

One method of combining the need to consider next generation technology with systematic approaches and transparency is through the development of a tool for "systematic reviews" of microarray literature.

Systematic review methods are becoming increasingly more common, especially in medical and public health fields which involve a plethora of stakeholders and have wide-ranging human health implications (Institute of Medicine, 2011). A tool for performing such reviews would allow risk assessors to transparently apply standard criteria for judging the studies that they find in literature searches and include in their assessments. However, there are currently no systematic review tools focused on the applicability of toxicogenomic studies for use in human health risk assessment.

The Systematic Omics Analysis Review (SOAR) tool originated from interest in developing a distributable tool to facilitate the systematic screening of transcriptomics studies using existing community standards as criteria, so that such studies can become more widely applied to risk assessment. The Toxicological Reliability Assessment (ToxR) Tool (Schneider et al, 2009), MIAME standard (FGDS, 2010), and the Checklist for Exchange and Interpretation of Data from a Toxicology Study (Fostel et al, 2007) were resources for question development. After a spreadsheet of questions was generated, multiple rounds of testing were performed by scientists to refine and determine the appropriate weight for questions, and ultimately validate user agreement across a test set of published studies.

Methods

Source of questions

The initial questions used to develop the SOAR tool were derived from three main peer-reviewed sources: 1) MIAME (FGDS, 2010), 2) ToxRTool (Schneider et al, 2009), and 3) the Checklist for Exchange and Interpretation of Data from a Toxicology Study (Fostel et al, 2007). The questions that pertained directly to microarray data came from MIAME, while general questions on information needed for repeating a toxicological study are drawn from the ToxRTool and the Checklist. A few questions were

also written based on expert guidance because they were not included elsewhere. The ToxRTool in particular was also used as a general guide for how to design and structure this type of tool.

Development of the tool

Questions from the source materials were organized in a Google Drive Spreadsheet (see supplemental material for links to the spreadsheet of questions). A "Preliminary Questions" section was developed to screen out manuscripts that do not have three or more biological replicates or do not pertain to a chemical exposure and are thus not relevant to chemical risk assessment. This section also asked questions that determined the type of study (*in vivo, in vitro,* etc) in order to tailor the questions asked in the subsequent sections (these questions did not affect the score). The remaining questions were organized into five sections: 1) Test System (including separate sets of questions for *in vivo* and *in vitro* studies), 2) Test Substance, 3) Experimental Design, 4) Microarray Data, and 5) Suitability for Benchmark Dose (BMD) modeling, as seen in Table 1. Each question had a "yes" or "no" answer, with a few questions also containing a "Not Applicable" option. Initially, weights were set to one for every question, with an "NA" answer causing the weight to drop to zero. After testing, weights were adjusted to range from 0 to 1 depending on the importance of the information, as determined by participating microarray experts.

The spreadsheet format allowed for the use of drop-down response menus, automatic calculation of weighted scores for each section of questions, sections for rater comments, and automated scripts that adjust the questions that users were presented based on the type of data in the study, as well as automatic bibliographic data entry. Additionally, *mouse-over* comments were added to the spreadsheet to provide more information and examples of how to find the answer to the question within a published manuscript. Questions were edited first internally using a training set of four manuscripts, shown in Table 2, for which the pass/fail designation was determined *a priori* (Fertuck et al, 2003; Fracchiolla et

al, 2011; Frericks et al, 2011; Permenter et al, 2011). During the course of testing, some questions were re-worded for clarity, other questions were removed because the evaluation team did not find them informative, and the weights of the questions were adjusted to better reflect their importance in determining suitability for use in an assessment.

Testing

Seven scientists with diverse backgrounds and experience with toxicogenomic data were recruited to assist in assessing and validating the SOAR tool (see Table 3 for details on participants) over the course of four separate rounds conducted over nine weeks. During the first two rounds, the scientists were asked to focus on editing, clarifying, reformatting, or suggesting questions for addition or removal. In the third round, statistics on user agreement were calculated to focus on improving the wording and the weights of the questions, specifically where users disagreed. In the fourth and final round, six of the experts rated the same four manuscripts (n = 6; one scientist dropped out of the study before this round) to validate the tool. Because of the small sample size throughout the study, percent agreement between users on the final pass/fail outcome for a manuscript was the only statistic used.

Papers used for testing and evaluating SOAR were identified by performing a PubMed (http://www.ncbi.nlm.nih.gov/pubmed) literature search using the search string: microarray AND exposure. Nineteen papers were chosen that were pertinent to risk assessment of chemicals and not coauthored by the participating scientists. Papers were assigned to participants so that each paper was rated at least twice in one round and no participant rated the same paper more than once (n = 2-5 per paper per round, n = 6-7 per paper total; see Table 4 for exact sample sizes per paper per round).

At the beginning of each round, the scientists were given PDF copies of their assigned manuscripts for that round with author, affiliation, date, and journal information removed. The participants were also given PDF copies of information pertaining to the raw data (e.g. a print out of the manuscript's entry in

the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) database), also with author and date information removed. Table 2 gives reference information for the papers used for testing. Participants were given approximately 10 days to answer all of the questions in the tool for the four papers in a round. After all participants had completed a round, feedback was collected on questions to edit, remove, or add and changes were made accordingly. The weighting was also modified and a pass/fail threshold was developed based on participant feedback.

Results

Round 1 & 2 (General Question and Format Editing)

Significant changes were made to the tool after the first two rounds of question adjustment. The number of questions dropped from a maximum of 61 questions to a maximum of 34 questions, as shown in Table 1. Several subjective questions were removed from the tool, along with questions that did not come from a peer-reviewed source. Originally there were 11 questions from MIAME, 23 questions from ToxRTool, 5 questions from Fostel et al (2007), 12 questions from the Benchmark Dose Technical Guidance document (USEPA, 2012), and 17 questions from domain experts. Section 5 pertaining to BMD modeling was removed because it required highly specific questions about the data and a level of understanding and time commitment beyond what should be expected from a first pass screening tool. Additionally, after the participating scientists rated a paper that involved human subjects (papers 4, 7, and 8), it became apparent that a separate set of questions was needed specifically for human studies under the "Test System" section. Originally the "Test System" questions were broken up into *in vivo* and *in vitro* sections but did not consider the human subject. With guidance from the participating scientists, a section was added for "*In vivo*, human" test subjects.

Finally, the "Microarray Data" section was split into two different sets of questions depending on whether or not raw data was available for the study. Less information is needed about how the normalized data were processed if interested scientists can access the data in raw form. After making these revisions, the final version of the tool involved five main sections with the first section setting up the tool and the remaining four sections used to score the paper. The final version contained 11 questions from MIAME, 19 questions from the ToxRTool, 4 questions from Fostel et al, and 6 questions from domain experts (if a question was repeated in two of the guidance sources it was only cited as being from one of the two).

Throughout the editing process the weights of the questions were also set. It was determined that a paper would be recommended for further consideration in a human health risk assessment if it received a score of at least 80% for each section.

Round 3 (Targeted Question Editing)

Results from the third round of testing are shown in Figure 1. Of the seven papers tested in this round, there were only two where the experts disagreed on the pass/fail outcome (i.e., there was not a unanimous pass/fail determination). For paper 13 there was no agreement between the three experts rating this paper, though further inquiry showed that this was caused by rater misunderstanding of the presented data. One scientist had incorrectly interpreted the study as being *in vitro*, while the other two answered as *in vivo*. Of the two scientists who determined it was an *in vivo* study, one failed it by answering "no" to the question "Is frequency and duration of exposure to the test substance explained?" while the other scientist answered "yes."

For paper 17, two of the three experts were in agreement that the paper should fail. The third expert did not agree, making the percent agreement 66%. The main disagreement was on the answer to the

question: “Are the study endpoint(s) and their method(s) of determination clearly described?” which may be considered subjective to some users.

Round 4 (Validation)

Round 4, where all scientists rated the same papers as validation, produced 100% agreement on the final outcome (pass/fail) of all 4 papers (n=6), as shown in Table 4. Concordance was achieved only after discussing the responses of one participant. The results were reviewed when there was disagreement on the pass/fail status of a paper. Each response given by the scientist who disagreed was examined and it was discovered that the scientist had incorrectly answered a single question that caused Papers 9 and 10 to fail (“Is frequency and duration of exposure to the test substance explained?”). The frequency and duration information was pointed out in the manuscript to the scientist who had answered “no.” This scientist realized that they missed this information while rating the manuscript and chose to revise their response, bringing their results into concordance with the rest of the group. Though there was some other disagreement between answers to specific questions for all of the papers, none of the differences were significant enough to change the pass/fail outcome of the tool.

For access to the full electronic version of the tool and all of the questions included in it, see the supplemental material.

Discussion

The SOAR tool was designed to provide a transparent method for risk assessors to determine the suitability of specific, published microarray data for consideration in assessments. The goals are similar to those of the ToxRTool but with a focus on issues of data analysis and study design specific to transcriptomic microarrays. The tool was developed through four rounds of testing with experts who have microarray experimental design and analysis experience. This repetitive testing allowed for a

thorough evaluation of the wording, the appropriateness, and the weights applied to each question, as well as the general ease of use of the spreadsheet format. By the final validation round, all six experts agreed on whether the four papers would pass or fail.

The tool should be used by at least two different risk assessors familiar with microarray data for each manuscript being scored. If the two raters cannot agree on whether the manuscript passes or fails the tool, a third risk assessor should be consulted to make the final determination on the manuscript. The final round of validation was performed with this method in mind. Specific answers were examined only when an expert did not agree with the pass/fail designation of the rest of the group, as we would expect to occur in actual use. The situation discussed in the results of the validation, where one user made an honest mistake in their response that caused the papers to incorrectly fail, is a prime example of how multiple users will ensure the accuracy of the scores. Choosing to only have such comprehensive discussions when there was disagreement on the ultimate pass/fail result of the paper removed the need to discuss every question in the tool when the overall outcome was the same and benefited the users by reducing the overall length of time spent considering the literature.

Notably, there are disadvantages to taking such a broad look at the results. The main concern is that all users could make mistakes on a single paper that would result in an incorrect pass/fail designation. This could occur if the mistakes were made on the same question or on different questions. Additionally, these “mistakes” could occur in two different ways: 1) as the result of typing the incorrect response (choosing “no” for a question when the user meant to choose “yes”), or 2) as the result of differing interpretations of the questions or of the information in the manuscript being rated. If only the overall pass/fail result is examined in a case where multiple users make “mistakes,” both users may end up having incorrectly passed or failed a study. The remedy for this, which was also performed in the present study, was to have one person quickly compare the individual results from multiple users. Then,

if answers differed on questions with high weights or on a significant number of questions, regardless of the final pass/fail designation, these can be brought to the attention of users.

Using repetitive testing with the same group of experts can result in the experts being trained in the meaning of the questions. By the final round their agreement in scoring may have been based on their collective understanding of the meaning of the questions and not on the innate clarity of the wording. This could mean that the tool would not produce such concordant results with new users who have less experience with the questions. In order to combat this issue, the majority of the questions were given comments in the spreadsheet with an alternate wording or clarifying details. Training would need to be provided for risk assessors to familiarize themselves with microarrays and their data, so specific training on the SOAR tool could be provided at that time.

The ultimate goal for the SOAR tool is to use natural language processing to enable computers to perform the first pass screen of all papers resulting from a literature search. If the computer gives a manuscript a “pass” then it will be sent to a human for further consideration and potential analysis. Many questions that could be considered subjective were removed by the final round of testing in an attempt to make the transition to natural language processing easier. There are questions on data quality that computers will not be capable of answering in the foreseeable future and these were set aside for human consideration after the first pass screening has taken place. As a result, the tool does not examine some of the more important aspects of data quality, such as overall reproducibility of the results. However, the goal is that after using the tool, risk assessors will be much better informed on the details of the paper and the study, as well as possible weaknesses and strengths so that they can make a final decision on whether or not it is appropriate to include in their assessment.

It is important to note that the results from the tool are not meant to be used as a strict cut-off; the opinion of an experienced expert should always take precedence over the result of the tool, which is

intended only to make the process of identifying suitable studies more systematic and transparent. However, if agencies and risk assessors employ the SOAR tool, the information and the record created by collecting that information will be a critical step in fulfilling the need for transparent and thorough decisions on the quality of the omics studies.

Acknowledgements

The author would like to acknowledge the scientists who participated in developing this tool: Lyle Burgoon, Shannon Bell, Edward Perkins, Natalia Garcia-Reyero, Ping Gong, and Rong-Lin Wang. Additional thanks to Drs. Channa Keshava, Jennifer Nichols, Reeder Sams, and John Vandenberg for their helpful comments on the tool and the manuscript.

References

Andreasen, E.A., Mathew, L.K., Tanguay, R.L. (2006). Regenerative Growth Is Impacted by TCDD: Gene Expression Analysis Reveals Extracellular Matrix Modulation. *Toxicol. Sci*. **92**(1), 254-269.

Boyle, J.O., Gümüş, Z.H., Kacker, A., Choksi, V.L., Bocker, J.M., Zhou, X.K., Yantiss, R.K., Hughes, D.B., Du, B., Judson, B.L., Subbaramaiah, K., Dannenberg, A.J. (2010). Effects of Cigarette Smoke on the Human Oral Mucosal Transcriptome. *Cancer. Prev. Res*. **3**(3), 266-278.

Carolan, B.J., Heguy, A., Harvey, B.G., Leopold, P.L., Ferris, B., Crystal, R.G. (2006) Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase L1 gene in human airway epithelium of cigarette smokers. *Cancer Res*. **66**(22), 10729-40.

Chen, J., Carney, S.A., Peterson, R.E., Heideman, W. (2008). Comparative genomics identifies genes mediating cardiotoxicity in the embryonic zebrafish heart. *Physiol. Genomics*. **33**, 148-158.

Dreij, K., Rhrissorrakrai, K., Gunsalus, K.C., Geacintov, N.E., Scicchitano, D.A. (2010). Benzo[a]pyrene diol epoxide stimulates an inflammatory response in normal human lung fibroblasts through a p53 and JNK mediated pathway. *Carcinogenesis*. **31**(6), 1149-1157.

Fertuck, K.C., Eckel, J.E., Gennings, C., Zacharewski, T.R. (2003). Identification of temporal patterns of gene expression in the uteri of immature, ovariectomized mice following exposure to ethynylestradiol. *Physiol. Genomics*. **15**, 127-141.

Fostel, J.M., Burgoon, L.D., Zwickl, C., Lord, P., Corton, J.C., Bushel, P.R., Cunningham, M., Fan, L., Edwards, S.W., Hester, S., Stevens, J., Tong, W., Waters, M., Yang, C., Tennant, R.. (2007). Toward a checklist for exchange and interpretation of data from a toxicology study. *Toxicol. Sci.* **99**(1), 26-34.

Fracchiolla, N.S., Todoerti, K., Bertazzi, P.A., Servida, F., Corradini, P., Carniti, C., Colombi, A., Cecilia Pesatori, A., Neri, A., Deliliers, G.L. (2011). Dioxin exposure of human CD34+ hemopoietic cells induces gene expression modulation that recapitulates its in vivo clinical and biological effects. *Toxicology.* **283**(1), 18-23.

Frericks, M., Burgoon, L.D., Zacharewski, T.R., Esser, C. (2011). Promoter analysis of TCDD-inducible genes in a thymic epithelial cell line indicates the potential for cell-specific transcription factor crosstalk in the AhR response. *Toxicol. Appl. Pharmacol.* **232**(2), 268-279.

[FGDS] Functional Genomics Data Society (2010). Minimum Information About a Microarray Experiment - MIAME 2.0. http://www.mged.org/Workgroups/MIAME/miame_2.0.html.

Gebel, S., Diehl, S., Pype, J., Friedrichs, B., Weiler, H., Schüller, J., Xu, H., Taguchi, K., Yamamoto, M., Müller, T. (2010). The Transcriptome of Nrf2_/_ Mice Provides Evidence for Impaired Cell Cycle Progression in the Development of Cigarette Smoke–Induced Emphysematous Changes. *Toxicol. Sci.* **115**(1), 238-252.

Gottipolu, R.R., Wallenborn, J.G., Karoly, E.D., Schladweiler, M.C., Ledbetter, A.D., Krantz, T., Linak, W.P., Nyska, A., Johnson, J.A., Thomas, R., Richards, J.E., Jaskot, R.H., Kodavanti, U.P. (2008). One-Month Diesel Exhaust Inhalation Produces Hypertensive Gene Expression Pattern in Healthy Rats. *Environ. Health Perspect*. **117**, 38-46.

Hirano, M., Tanaka, S., Asami, O. (2011). Classification of polycyclic aromatic hydrocarbons based on mutagenicity in lung tissue through DNA microarray. *Environ. Toxicol*. DOI 10.1002/tox.20761

Heiden, T.C., Struble, C.A., Rise, M.L., Hessner, M.J., Hutz, R.J., Carvan, M.J. 3rd. (2008). Molecular targets of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) within the zebrafish ovary: Insights into TCDD-induced endocrine disruption and reproductive toxicity. *Reprod. Toxicol*. **25**(1), 47-57.

Institute of Medicine. (2011). Finding What Works in Health Care: Standards for Systematic Reviews. Washington, DC: The National Academies Press.

Kong, E.C., Allouche, L., Chapot, P.A., Vranizan, K., Moore, M.S., Heberlein, U., Wolf, F.W. (2010). Ethanol-Regulated Genes That Contribute to Ethanol Sensitivity and Rapid Tolerance in Drosophila. *Alcohol Clin. Exp. Res.* **34**(2), 302-316.

Landi, M.T., Dracheva, T., Rotunno, M., Figueroa, J.D., Liu, H., Dasgupta, A., Mann, F.E., Fukuoka, J., Hames, M., Bergen, A.W., Murphy, S.E., Yang, P., Pesatori, A.C., Consonni, D., Bertazzi, P.A., Wacholder, S., Shih, J.H., Caporaso, N.E., Jen, J. (2008). Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One.* **3**(2).

McHale, C.M., Zhang, L., Lan, Q., Li, G., Hubbard, A.E., Forrest, M.S., Vermeulen, R., Chen, J., Shen, M., Rappaport, S.M., Yin, S., Smith, M.T., Rothman, N. (2009). Changes in the Peripheral Blood Transcriptome Associated with Occupational Benzene Exposure Identified by Cross-Comparison on Two Microarray Platforms. *Genomics.* **93**(4), 343-349.

[NRC] National Research Council. (2011). Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: The National Academies Press.

Nilsson, E., Larsen, G., Manikkam, M., Guerrero-Bosagna, C., Savenkova, M.I., Skinner, M.K. (2012). Environmentally Induced Epigenetic Transgenerational Inheritance of Ovarian Disease. *PLoS One.* **7**(5).

Pedersen, M.B., Skov, L., Menné, T., Johansen, J.D., Olsen, J. (2007). Gene Expression Time Course in the Human Skin during Elicitation of Allergic Contact Dermatitis. *J. Invest. Dermatol.* **127**, 2585-2595.

Permenter, M.G., Lewis, J.A., Jackson, D.A. (2011). Exposure to Nickel, Chromium, or Cadmium Causes Distinct Changes in the Gene Expression Patterns of a Rat Liver Derived Cell Line. *PLoS One.* **6**(11).

Schneider, K., Schwarz, M., Burkholder, I., Kopp-Schneider, A., Edler, L., Kinsner-Ovaskainen, A., Hartung, T., Hoffmann, S. (2009). 'ToxRTool', a new tool to assess the reliability of toxicological data. *Toxicol. Lett.* **189**(2), 138-144.

Song, M.O., Li, J., Freedman, J. (2009). Physiological and toxicological transcriptome changes in HepG2 cells exposed to copper. *Physiol. Genomics.* **38**, 386-401.

Song, R., Duarte, T.L., Almeida, G.M., Farmer, P.B., Cooke, M.S., Zhang, W., Sheng, G., Fu, J., Jones, G.D. (2009). Cytotoxicity and gene expression profiling of two hydroxylated polybrominated diphenyl ethers in human H295R adrenocortical carcinoma cells. *Toxicol. Lett.* **185**(1), 23-31.

Stevens, T., Krantz, Q.T., Linak, W.P., Hester, S., Gilmour, M.I. (2008). Increased transcription of immune and metabolic pathways in naive and allergic mice exposed to diesel exhaust. *Toxicol Sci.* **102**(2), 359-70.

Suvorov, A. and Takser, L. (2010). Global Gene Expression Analysis in the Livers of Rat Offspring Perinatally Exposed to Low Doses of 2,2',4,4'-Tetrabromodiphenyl Ether. *Environ. Health Perspect.* **118**, 97-102.

Woods, C.G., Fu, J., Xue, P., Hou, Y., Pluta, L.J., Yang, L., Zhang, Q., Thomas, R.S., Andersen, M.E., Pi, J. (2009). Dose-dependent transitions in Nrf2-mediated adaptive response and related stress responses to hypochlorous acid in mouse macrophages. *Toxicol. Appl. Pharmacol.* **238**(1), 27-36.

[USEPA] United States Environmental Protection Agency (2012). Benchmark Dose Technical Guidance. Risk Assessment Forum, Washington, DC; EPA/100/R-12/001.

Table 1: Question sections included in the original version of the SOAR tool compared to the final version. The section "Test System" has different questions based on the type of study. The maximum number of questions a paper can require is 34, though only 29 of them would be scored. The first five basic questions are used to exclude inappropriate papers and to set up the questions required, and are therefore not given a score.

Question sections	Original # of Questions	Final # of Questions
Preliminary Questions	4	5
Test System (*in vivo* human, *in vivo* non-human, or *in vitro*)	7-10	3-10
Test Substance	6	6
Experimental Design	11	5
Microarray Data (either including raw data or not)	18	5-8
Suitability for Benchmark Dose Modeling	12	-

Table 2: The papers used to develop and test the SOAR tool. The first four were used only during internal development of the questions. Papers 1-8 were used by seven experts (internal and external) for 2 rounds of revising the questions. The last 11 were used by the same group to validate the tool and determine inter-rater reliability. Papers were chosen by performing a broad literature search and removing any that were affiliated with the expert in this study.

	Reference	PMID	Study Type	Study Compound	Rounds used
Papers Used During Internal Development					
	Fertuck et al. (2003)	12915738	In vivo, mouse	ethynylestradiol	Development
	Permenter et al. (2011)	22110744	In vitro, rat	nickel, chromium, cadmium	Development
	Frericks et al. (2011)	18691609	In vitro, mouse	TCDD	Development
	Fracchiolla et al. (2011)	21296121	In vitro, human	TCDD	Development
Papers Used for General Question Editing and Formatting					
1	Woods et al. (2009)	19376150	In vitro, mouse	Hypochlorous acid	Round 1 (n=3) Round 2 (n=4)
2	Chen et al. (2008)	18230668	In vivo, zebrafish	Retinoic acid, TCDD	Round 1 (n=3) Round 2 (n=4)
3	Kong et al. (2010)	19951294	In vivo, Drosophila	Ethanol	Round 1 (n=4) Round 2 (n=2)
4	Pedersen et al. (2007)	17597826	In vivo, human	Nickel	Round 1 (n=5) Round 2 (n=2)
5	Nilsson et al. (2012)	22570695	In vivo, rat	Multiple pesticides, plastics, TCDD, and jet fuel	Round 1 (n=4) Round 2 (n=3)
6	Song MO, et al. (2009)	19549813	In vitro, human	Copper	Round 1 (n=4) Round 2 (n=3)
7	Boyle et al. (2010)	20179299	In vivo, human	Cigarette Smoke	Round 1 (n=2) Round 2 (n=5)
8	Carolan et al. (2006)	17108109	In vivo, human	Cigarette Smoke	Round 1 (n=2) Round 2 (n=5)
Papers Used for Targeted Question Editing					
13	Andreasen et al. (2006)	16443690	In vivo, zebrafish	TCDD	Round 3 (n=3)
14	Song R, et al. (2009)	19095052	In vivo, human	PBDEs	Round 3 (n=4)
15	Gottipolu et al. (2008)	19165385	In vivo, rat	Diesel exhaust	Round 3 (n=3)
16	King Heiden et al. (2008)	17884332	In vivo, zebrafish	TCDD	Round 3 (n=5)
17	Dreij et al. (2010)	20382639	In vitro, human	Benzo[a]pyrene diol epoxide	Round 3 (n=3)
18	Suvorov et al. (2010)	20056577	In vivo, rat	BDE-47	Round 3 (n=5)
19	McHale et al. (2009)	19162166	Epidemiological	Benzene	Round 3 (n=4)
Papers Used for Validation					
9	Stevens et al. (2008)	18192680	In vivo, mice	Diesel exhaust	Round 4 (n=6)
10	Gebel et al. (2010)	20133372	In vivo, mice	Cigarette Smoke	Round 4 (n=6)
11	Landi et al. (2008)	18297132	Epidemiological	Cigarette Smoke	Round 4 (n=6)
12	Hirano et al. (2011)	21887816	In vitro, human	PAHs	Round 4 (n=6)

Table 3: Experts who participated in editing and validating the SOAR tool, their affiliations, and expertise.

Expert Name	**Affiliation**	**Expertise**
Shannon Bell	ORISE Fellow at NHEERL, USEPA, Research Triangle Park, NC	Systems biology, large data analysis
Lyle Burgoon	NCEA, USEPA, Research Triangle Park, NC	Systems biology, bioinformatics, data mining
Natalia Garcia-Reyero	Mississippi State University, Starkville, MS	Ecotoxicogenomics
Ping Gong	Badger Technical Services, Vicksburg, MS	Ecotoxicogenomics
Emma McConnell	ORISE Fellow at NCEA, USEPA, Research Triangle Park, NC	Ecotoxicology and environmental health
Edward Perkins	USACE, Vicksburg, MS	Toxicogenomics
Rong-Lin Wang	NERL, USEPA, Cincinnati, OH	Genomics, bioinformatics, data mining

Table 4: Results from Round 4, validation. Though some authors disagreed on specific answers to certain questions, the disagreement was not significant enough to change the final outcome for the papers. Paper 9 and 10 passed; paper 11 and 12 failed. For paper 12, EM, SB, LB, and NGR failed the paper in the "Basic Questions" section based on a lack of sufficiently biological replicates (tool requires n≥3), and therefore the following question sections were not answered. RW and PG did complete all the question sections, however, the paper still failed.

		Scores by Author					
		EM	**SB**	**RW**	**PG**	**LB**	**NGR**
Paper 9	I. Test Organism (In vivo)	97	97	83	100	97	97
	II. Test Substance	100	100	82	100	100	100
	III. Experimental Design	100	100	100	100	100	100
	IV. Microarray Data	85	85	85	85	85	85
	Final Result:	**PASS**	**PASS**	**PASS**	**PASS**	**PASS**	**PASS**
Paper 10	I. Test Organism (In vivo)	80	97	97	90	93	97
	II. Test Substance	100	100	100	100	100	100
	III. Experimental Design	87	100	100	100	100	100
	IV. Microarray Data	96	92	96	92	92	85
	Final Result:	**PASS**	**PASS**	**PASS**	**PASS**	**PASS**	**PASS**
Paper 11	I. Human Subjects (In vivo)	69	81	69	100	100	81
	II. Test Substance	100	100	69	100	100	69
	III. Experimental Design	67	77	77	77	77	77
	IV. Microarray Data	81	75	38	69	62	38
	Final Result:	**FAIL**	**FAIL**	**FAIL**	**FAIL**	**FAIL**	**FAIL**
Paper 12	I. Test System (In vitro)	-	-	100	100	-	-
	II. Test Substance	-	-	94	94	-	-
	III. Experimental Design	-	-	87	87	-	-
	IV. Microarray Data	-	-	39	25	-	-
	Final Result:	**FAIL**	**FAIL**	**FAIL**	**FAIL**	**FAIL**	**FAIL**

Figure Legends

Figure 1: Percent agreement between experts on final pass/fail result of papers tested in Round 3. Each paper was tested through the SOAR tool by 3-5 expert experts. Paper 13 had no agreement between the three experts due to misunderstanding of the data presented in the paper. Paper 17 had 1 of 3 experts disagree.

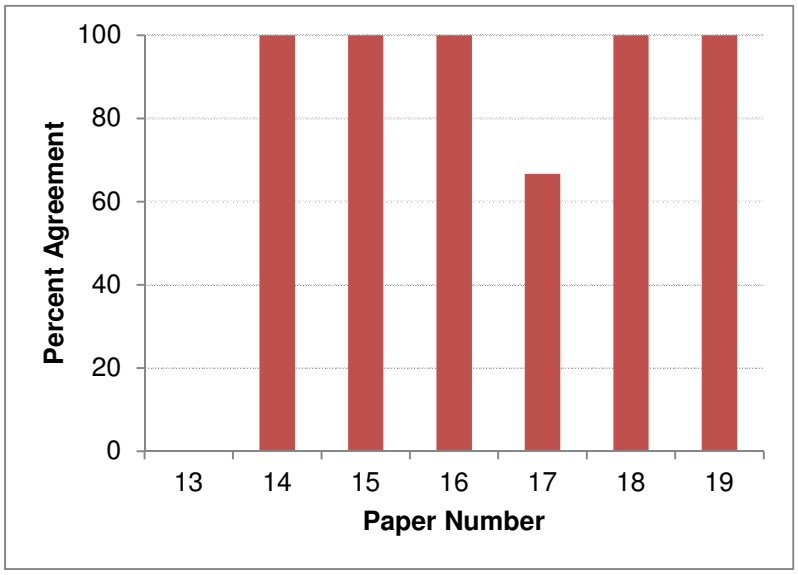
100
80
60
40
20
0
Percent Agreement
13
14
15
16
17
18
19
Paper Number

Figure 1

Supplemental Materials:

For a copy of the original version of the SOAR tool used in the first round of question editing, see the following link:

https://docs.google.com/spreadsheet/ccc?key=0AmmkQbxxSwwKdDNqYjBxaGhYTHFPX3NhaTMyT1A2WXc

For a copy of the final version of the SOAR tool created from this study, see the following link:

https://docs.google.com/a/ordweb.epa.gov/spreadsheet/ccc?key=0AmmkQbxxSwwKdFkwWkVHWW9TV3ptSFhSeTM0Z0dRQ3c#gid=0