



Submitted by:

Rosemary Zaleski, ExxonMobil Biomedical Sciences, rosemary.t.zaleski@exxonmobil.com
Hal Zenick, retired US EPA
Timothy Buckely, US EPA

Proposal title:

Big Data: Exploring Possibilities to Improve Predictive Exposure Capabilities

Describe the problem to be addressed. Why is the issue important? To whom is this issue important?

There is growing appreciation for the enormous anthropogenic chemical landscape associated with modern society. It is estimated that there are thousands of different chemicals in commerce, many of which find their way into our homes through consumer products (Egeghy et al., 2012). Given the wide range of both natural and manmade chemical substances as well as potential exposure pathways for each one, it becomes apparent that measurement of all possible exposures is not feasible. Traditional approaches to evaluating chemical exposures have tended to take an observational “under the lamp post” approach where tens of chemicals are considered at a time and the exposure is evaluated after the fact. Although important and valuable in the evolution of exposure science, there is recognition that a new approach is needed in order to be predictive and to meaningfully address the enormous chemical landscape associated with modern society. Expanding the predictive capability of exposure models is therefore a priority.

Advances in measurement technologies, informatics, and computational capabilities paired with broad adoption of “smart” personal devices have created unparalleled opportunity to generate large data sets. While often generated for other reasons, these novel data sources may include information relevant to understanding exposure potential. For example, large databases exist that can shed light on consumer behaviors and exposure potential, such as product purchase, purchase location, activity patterns, and product composition information. However, no systematic assessment has been made of these new data sources in the context of either improving exposure tools or interpreting model predictions.

The goal of this proposal is for HESI to strategically identify, enable and facilitate the integration and application of emerging advances in exposure science to transform and inform human health assessment. As risk assessment requires consideration of both hazard and exposure, it must be understood that to move risk science forward advances must be made in both hazard and exposure aspects.

This proposal aims to identify new and expanded data sources and evaluate potential applications to improve understanding of exposure aspects of risk. It also aims to address important considerations for using big data in general, including aspects such as confidentiality and appropriate development of data linkages, relevant to multiple scientific areas.

It is envisioned that this integration can provide a step change advancement for evaluation of predictive exposure models and expansion of their capabilities.

To manage scope and to make most use of ongoing data mining in this area (i.e., ExpoCast); the workshop will be focused on data related to consumer product purchase and use, consumer behavior, and measurement data relevant to residential exposures.



WHAT CAN HESI DO TO CONTRIBUTE TO THESE ISSUES?

We propose a HESI project in which a strategic subset of these new technologies, i.e. big data, is identified, integrated and evaluated in an initial pilot program. We propose establishment of a working group to identify data types most useful to exposure assessment as well as possible data sources, followed by a workshop to address considerations important for appropriate use and linkage of big data sets as well as identify one or more data sets that could then be tested in a pilot program. The results of this project are expected to:

- a) improve current understanding of exposure,
- b) shape future integrated data collection programs that can be used to validate and expand predictive capabilities for exposure models, and
- c) provide guidance on addressing issues that may arise in any science area where efforts are being made to obtain and integrate big data from multiple sources.

Reference:

Egeghy PP, Judson R, Gangwal S, Mosher S, Smith D, Vail J, Cohen Hubal EA. (2012). The exposure data landscape for manufactured chemicals. *Sci Total Environ* 414:159–166.

Describe the basic project steps or stages to the best of your ability, including an expected timeline, milestones, and deliverables for the first two years.

Part I: Establish a working group to identify exposure estimation needs and data sources of potential interest

Establish a working group of government, industry and academic scientists to:

1. Identify types of data most informative for improving exposure models and associated modeling needs. This activity will summarize and potentially build upon existing experience from current /recent strategic initiatives, such as EPA's ExpoCast program, HESI's RISK21 project, ACC's ExpoDat activity, and advances in regulatory exposure approaches (REACH).
2. Identify and characterize big data sets and other novel data sources potentially relevant for exposure application. Some include: Nielsen, Walmart, SAS, Google, Eurostat, UN, Multinational Time Use Database, International Association of Time Use Research, Apple Research Kit, Euromonitor International, Amazon, etc. Relevant data source information such as description of the data types, relevant population, how the information was gathered, and access/availability considerations will be summarized (where available).
3. Develop a prioritized list of considerations to be addressed in order to apply this type of information in an exposure context. Include consideration of capturing full distributional characterization of data to support probabilistic analyses.

Working group representatives from a range of disciplines within HESI will be invited to participate to help facilitate cross-functional, innovative discussion. The working group will develop summaries of each activity for presentation at the Workshop.

Part II: Host a workshop to consider both general considerations for integrating and applying big data sets and then more specifically exposure related applications for identified data sources

As a second step, a workshop would be convened of the HESI working group, representatives familiar with select key big data sets (where available), and representatives from other disciplines that may bring unique perspectives, additional data sources, or information relevant for successful linkage and application of big data in an exposure assessment context. The workshop would be aimed at developing a detailed understanding of information present in the identified data sets and potential applications, and its strengths and limitations in a general context and in the context of exposure prediction. Also, the workshop would address more general questions related to linkage of multiple data sets in a meaningful way, including assessing quality for use in novel non-intended applications, data confidentiality, and other general aspects of applying big data to outside-of-the-box questions.



Approaches for building partnerships to enable data access and use will also be discussed. Establishing the appropriate domain around application of data sets will be an important aspect – i.e., strengths and limitations of data as they impact data use and interpretation. This information should be relevant not only for potential exposure application, but for other efforts that may seek to maximize the information value of available data sets. Anticipated workshop work products include:

1. A list of potential opportunities to apply novel data sources to improve exposure characterization and model interpretation capabilities
2. A prioritized list of considerations to be addressed in order to apply this type of information in both a general context as well as an exposure specific context
3. Identification of the most promising application opportunities

Part III: Develop Report(s)

Following the workshop, the HESI working group would lead development of a white paper describing the workshop output, including a path forward to pilot application of at least one of the big data sets for use in improving exposure estimation capabilities. See Attachment 2 for a model report outline. A second paper that described results of the data set testing and its interpretation/translation is also envisioned.

Timeline:

HESI has a successful record in influencing and advancing exposure science through its scientific committees, including the Integration of Biomonitoring Exposure Data into the Risk Assessment Process Technical Committee (sunset in 2012) and the Risk Assessment in the 21st Century (RISK21) Technical Committee (concluded in 2014). The timing is ideal for this new proposal to advance exposure science.



Attachment 1: Possible Workshop Agenda Ideas

Key Activities:

Activity 1: Review of Exposure Characterization and Modeling Needs
Presentation by Workgroup Members

Activity 2: Review of Data Sets and Sources
Presentation by data set representatives and/or Workgroup members
Access to data sets
Novel ways to generate data sets

Activity 3: Discussion of data source limitations and application considerations (both general considerations and data source/model specific considerations)
Discussion Leader(s) TBD

Activity 4: Discussion of potential application opportunities for identified data sources to meet identified exposure assessment needs
Discussion Leader(s) TBD

Activity 5: Identification and prioritization of promising application opportunities; development of recommended path forward

Activity 6: Interpretation/translation of output for target audiences.

Key Work Products:

1. A list of potential opportunities to apply novel data sources to improve exposure characterization and model interpretation capabilities
2. A prioritized list of considerations to be addressed in order to assess, integrate, and apply big data in general, and also specifically in an exposure context
3. Identification of the most promising application opportunities for improving consumer exposure prediction



Attachment 2: Outline of possible topics to be covered in Workshop White Paper

1. Overview/Executive Summary
2. Introduction
 - 2.1. Background & Rationale
 - 2.2. Project Description
3. Summary of exposure estimation opportunities and model needs
4. Summary of data sets
5. Overview of potential applications, with further detail for top ideas
6. Discussion of potential limitations and application considerations (General and Specific)
7. Interpretation/translation of output for relevant audiences.
8. Proposed path forward to pilot test most promising application(s)