Archives of Toxicology (2018) 92:1015–1020
https://doi.org/10.1007/s00204-017-2145-0

GUEST EDITORIAL

Highlight report: 'Big data in the 3R's: outlook and recommendations', a roundtable summary

C. Mahony[1] · R. Currie[2] · G. Daston[3] · N. Kleinstreuer[4] · B. van de Water[5]

Received: 13 December 2017 / Accepted: 14 December 2017 / Published online: 16 January 2018

Introduction

With the advances in animal alternative testing methods, computational models, and the emergence of integrated testing strategies, the status of big data in the 3R's (replacement, reduction and refinement) has come to the fore. A roundtable at the Tenth World Congress on Alternatives (WC-10) in Seattle set out to examine what big data means for the 3Rs and to review the progress being made. Big data indeed brings about the possibility of a deeper understanding of biology, which in turn could play a role in eventually redefining our decision-making process in chemical safety assessment, but the topic is not without its challenges. Opportunities exist to address these both now and in the future, as discussed herein.

Data sharing initiatives: the FAIR principles

In the era of big data, the format and methods by which the data are stored, accessed, and analysed have become increasingly important, and the concurrent rapid expansion of computational power and prowess is facilitating the synthesis and analysis of many types of data toward the goals of the 3Rs. These considerations apply not only to new high-dimensional data, e.g., from next generation sequencing technologies, but also to existing legacy data such as in vivo toxicology studies representing decades of work, vast numbers of animals, and millions of dollars. Varied groups have strong interests in creating common spaces where data can be shared, and in defining principles by which data sharing should occur. Scientific researchers want to publish their data and associated analyses, and to provide the opportunity for others to offer alternate interpretations. Journal editors and publishers are also under both internal and external pressure to ensure that publications are supported by transparent, easily accessible data sources, and funding agencies have recently enhanced their focus on proper data stewardship to ensure that grants are supporting valuable research. The Holdren memo, for example, mandates that work supported by the US federal taxpayer dollars should be delivered back to the public in an open manner. Similar ruling is provided by various European states and European Commission supported research. Finally, the data science community, including software and tool-builders, needs to have access to a broad, standardized swath of data to effectively process, analyse, and integrate multiple information sources to more efficiently and effectively advance scientific discovery. A number of recent initiatives and activities have brought these diverse stakeholders together to pool resources and experiences, discuss data sharing challenges and opportunities, and arrive upon a common set of ideals that should govern such processes.

One such initiative led to the publication of the FAIR principles (Table 1) for scientific data management and stewardship (Wilkinson et al. 2016). These four principles have become key objectives for data practises within the National Institutes of Health (NIH) and across the broader scientific community. With respect to the NIH, data objects that are federally funded, both intramurally and extramurally, must be findable, e.g., via a digital object identifier (DOI), and

✉ C. Mahony
mahony.c@pg.com

[1] Central Product Safety, Procter & Gamble Technical Centres Ltd., Whitehall Lane, Egham, Surrey, UK

[2] Toxicology and Health Science, Syngenta Jealotts Hill International Research Centre, Bracknell, Berkshire RG42 6EY, UK

[3] Central Product Safety, Procter & Gamble Company, Mason, OH, USA

[4] NIH/NIEHS/DNTP/The NTP Interagency Center for the Evaluation of Alternative Toxicological Methods, Research Triangle Park, NC 27713, USA

[5] Research Division of Drug Discovery and Safety, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands

Table 1 The FAIR guiding principles for scientific data management and stewardship [from (Wilkinson et al. 2016)]

Fair principles	
Findable	A data object should be uniquely and persistently identifiable
Accessible	Data is accessible by authorized users (human and machine) through a well-defined protocol
Interoperable	(Meta) data assigned to the data object is syntactically parseable and semantically machine accessible
Reusable	Data objects must comply with the above three principles and sufficiently documented to allow integration/linkage with other data sources

accessible, meaning that they can be read and interpreted by both humans and machines. Datasets should be well-described by metadata, using standardized ontologies, in an interoperable way that allows for proper cataloguing and storage to ensure that they can be integrated with other data sources and are therefore reusable. Ongoing efforts within the National Institute of Environmental Health Sciences (NIEHS), and the associated National Toxicology Program (NTP), provide a snapshot of the larger research community as data scientists work to put these principles into practise. NIEHS, and other parts of the NIH, deal with a heterogeneous mix of data systems and technologies, data management practises, metadata capture and standards, funding mechanisms and resources for building & sustaining systems, and policies around usage of data. NIEHS is building a Data Commons, which will serve as a common platform for management of research data, and is also developing a metadata catalogue for terminologies/ontologies that can be used to curate new and existing datasets. Significant efforts are underway to improve the capturing of data provenance, search functionality, and visual analytics, amenable to both user interfaces and web application programming interfaces (APIs) that will provide a bridge between The Chemical Effects in Biological Systems (CEBS: which houses all of the NTP's data in a web-accessible format https://www.niehs.nih.gov/research/resources/databases/cebs) and other resources such as the EPA's Chemistry Dashboard (https://comptox.epa.gov/dashboard), PubChem (https://pubchem.ncbi.nlm.nih.gov/), and NLM ToxNet (https://toxnet.nlm.nih.gov/) databases.

Specific to data sharing with respect to the 3Rs, the NTP Interagency Center for Evaluation of Alternative Toxicological Methods (NICEATM) has built the Integrated Chemical Environment (ICE: https://ice.ntp.niehs.nih.gov/) to apply FAIR principles to both non-animal in vitro and in silico data as well as legacy in vivo animal data (Bell et al. 2017). The data integrator portion of ICE is a portal through which users can compare alternative approaches and build predictive models using the existing animal data as anchoring endpoints, to help establish scientific confidence in new approaches. In coordination with the Interagency Coordinating Committee for the Validation of Alternative Methods (ICCVAM), and the 16 federal agencies who sit on the committee, NICEATM is also helping to develop a U.S. Strategic Roadmap for Modernizing Safety Testing of Chemicals and Medical Products (https://ntp.niehs.nih.gov/go/natl-strategy). One of the strategic goals of the roadmap is to foster the use of efficient, flexible, and robust practises to establish confidence in new methods (Casey 2016). Some specific objectives related to this goal include identifying and collating sources of high-quality human toxicological and exposure data, creating centralized data access points that are publicly available and easily accessible, actively soliciting the submission and collation of parallel data from existing animal studies and new alternative methods, and leveraging partnerships and complimentary initiatives, all of which necessitate FAIR data sharing practises.

Certain aspects of the FAIR principles are especially challenging, for both NIH and the broader scientific community. Interoperability and reusability of data depend largely upon the questions at hand, and require agreement and coordination across many parties. There are other, purely practical, issues that arise due to the size of datasets and policies around sensitive information that result in situations where the data may not be movable, and computation must be moved to the data. Further, under the reality of finite resources, the scientific community must guide prioritization of data storage, access, analysis, and maintenance.

Data use: big data versus informative data

The classical view of the central dogma of biology is that the coded genetic information in DNA is transcribed into messenger RNA (mRNA) which contains the program for synthesis of protein. Modern biology is extending this and viewing all scales as a complex system of interacting networks. For instance, although cellular components are ultimately encoded by the genome via regulatory RNAs and sequences for protein, cellular phenotypes emerge from the interactions of signal transduction and gene regulatory networks controlling the protein–protein interaction and metabolic networks. Organ and organism phenotypes emerge from developmental pathways, the structural connections between cells (the connectome) and the extracellular matrix deposited by those cells as well environmental conditions.

And finally, ecosystem networks emerge from the evolution, organism trophic relationships and the environmental conditions. The curation and understanding of such data requires bioinformatics and advanced analytical techniques. However, the deep understanding of these data has not to date been the domain of toxicology, which has focussed on understanding the adverse effects of chemical substances on living organisms. Toxicologists have historically applied the latter data to address safety assessment questions. Regardless of whether we consider old versus new approaches, the process starts with a problem formulation which must identify a question to solve. The next step is to build a testing strategy of how that problem may be solved, and this defines what types of data may be necessary to solve it. Increasingly the scientific and regulatory community is working with new approach methods which bring challenges in how to structure, store and analyse. Experts are being brought in to structure data on biology, exposure and toxicology, and large compendia such as these are suggested to be Big Data. The application of these data to toxicology-related questions needs to bring communities together to both define the need and make the requisite data available.

In the new approach, an initial challenge in the assessment of a new chemical is to determine toxicity hypotheses to test. Toxicology already has conceptual frameworks that define ways to approach this problem. Read-across and chemical categorisation may suggest toxicity endpoints that are known to be associated with a particular type of chemistry. Big data approaches may help here, but in practise the number of examples is quite modest. Frequently for predictive toxicology we need to answer the questions: (1) what are the likely targets of a chemical, and (2) what are the known, or inferred, toxic effects of interfering with a target? The current Ensembl human genome database (https://www.ensembl.org/index.html) has 20338 coding genes and ChEMBL (https://www.ebi.ac.uk/chembl/) has approximately 1.7 million distinct compounds with 14 million activities measured on 11,538 targets from many species. We might consider to use these data to build QSAR models of chemical binding to protein targets to help answer our first question. In practise, although this seems like a big data set, the reality is that data is deep over a narrow range of targets, shallow over many more targets and nonexistent over most. The challenge for big data analysis using data like this is then to fill in the gaps by learning from existing data and applying it to other targets. Once we have identified the potential targets a chemical can interact with; the prediction of which toxicities may occur is challenging given the complexity that modern molecular biology has revealed. Conversely the target deconvolution of a toxicity is also challenging and consequently the mechanisms of most toxicities have not been determined. Building compendia of toxicity and underling mechanisms may help us deal with this complexity, and advanced machine leaning approaches will be important in gaining these insights.

Several big data trends may enable future disruption to toxicology with potential gains for the 3Rs. Increasingly individuals will generate data to build a “quantified self” with applications in personalised medicine. Sensors for real time detection are increasingly cheap and can be linked to mobile phones, creating a digital wake of GPS location data. Similarly, there is a growing internet of ‘things’ vis-a-vis a network of physical devices, and more use of precision approaches in, for example, agriculture can probe information on the tracking of items in commerce. If we can learn to integrate these data, might we enable a better understanding of exposure and links to health for improved understanding of epidemiology, and at the same time be able to measure the actual human toxicity pathway perturbations?

Integration of large data streams for toxicity prediction: reorganizing toxicology

Big data are already being used for predictive toxicology in several ways. A tremendous amount of data exists on the toxicity of chemicals, exposures, and other relevant information such as physical chemistry properties. These data have been compiled into databases that are searchable by a variety of means, most notably by chemical structure. This has facilitated risk assessment by analogy (read-across) and created groupings of chemicals that are toxicologically similar. EPA’s ACToR (https://actor.epa.gov/actor/home.xhtml), for example, aggregates data from hundreds of sources, on hundreds of thousands of chemicals (Judson et al. 2012). Such large aggregations of data make it possible to do read-across in a consistent, systematic way in which all the relevant analogues can be considered (Shah et al. 2016). Data from high-throughput test systems like ToxCast, or high-content data from toxicogenomics, are being used to generate biological effect data on far more chemicals than could be tested by conventional means, which is making it possible to consider biological potency in setting priorities for which chemicals require more attention. Previously, it was only possible to use rudimentary exposure surrogates (such as production volume or potential for widespread use) or the simplest toxicity heuristics to prioritize chemicals. Big data and computational approaches to estimate exposure have greatly improved our ability to get through the backlog of chemicals needing assessment (Wambaugh et al. 2013). That said, as highlighted previously, there are still significant challenges to overcome in compiling data, assuring its quality, and optimizing interoperability of data sets, as well as in determining how to interpret these data for risk assessment.

One of the key challenges is that the large data streams characterizing biological effects cover the molecular or cellular level, whereas risk assessments are carried out based on adverse effects at the organ or organismal level. Solving this challenge will require considerable research to identify and quantitate the key steps between initial effect and ultimate outcome. A first, foundational step in doing so will be to re-organize toxicology from a field based on endpoints to one based on modes of action. Projects to create mode of action ontologies (ECETOC 2016) will be useful as organizing structures, and in delimiting the universe of possible modes of action. Having such ontology will identify where there are gaps in high-throughput batteries, and will permit the tailoring of testing based on the plausible modes of action for a particular chemical structure. The challenge of connecting mode of action with adverse outcome will be at least partially met by constructing ontologies at multiple levels (i.e., on chemistry, gene expression, cellular response, adverse outcome) for chemicals that have been well studied. However, identifying and quantifying key events in the pathway from mechanism to outcome will require sophisticated in vitro models and computational simulations working in an iterative fashion. Multiple data streams, organized according to mode of action, should also facilitate the identification of patterns in the data through machine learning.

Big data and dynamics of cell signalling: the future of toxicolog(y)(ists)

Despite the fact that there is an enormous amount of data available in the public domain, including EPA's Chemistry Dashboard (https://comptox.epa.gov/dashboard), LINCS (http://lincsportal.ccs.miami.edu/dcic-portal/), and TG-GATEs (http://toxico.nibiohn.go.jp/english/), we still have a limited understanding on how each of the individual signalling components are lined up to contribute to the adverse outcomes. Future big data should further contribute to this understanding, with a focus on the quantitative relationships between particular cell signalling activation states and the contribution to adversity. Past research has used diverse assays to capture mode of action for a multitude of compounds, but clear direct relationships to the relevant adversity were not necessarily considered in the same test system. For example, single endpoint reporter assays for measuring ERα activation in different cell types are not associated with measurement of increased cell proliferation due to ERα signalling, prohibiting conclusions between mode of action and cell biological response (Huang et al. 2014). To ensure that we gain insights into these relationships, the toxicology community will require quantitative temporal concentration response datasets, preferably at the single cell level. Such activities have recently been initiated in the Innovative Medicine Initiative TransQST project (http://transqst.org/), that aims to capture quantitative information on pathway activation, relate this to adverse outcome, and link in vitro data to in vivo situations, including human data. Deriving detailed insight in these relationships will ensure an explosion of data points, since it will require both different doses and time points for a wide variety of compounds and model systems. Several technological advances will contribute to this next level understanding of chemical-biological interactions. Novel targeted high-throughput transcriptomics technologies will ensure a cost effective quantitative evaluation of stress response pathway activation upon chemical exposure (Niepel et al. 2017). This will allow clear definition of both benchmark dose/concentrations for pathway activation by individual chemicals, as well as definition of tipping points, where certain stress pathway activation levels would switch from an adaptive to an adverse state (Sand et al. 2017). While this technology will allow fast assessment at the whole cell population level, we will not be able to discern the individual cell state level. This may be tackled via different technological approaches. Firstly, through application of single cell RNA sequencing technologies we can discern the contribution of individual cells to the entire population (Macaulay and Voet 2014); it would also allow for the identification of gene expression events that are activated in all cells of the population, and mark the most suitable quantitative markers for pathway activation. This technology is still in its infancy and is not yet used in understanding population responses at the single cell level in toxicology settings. When RNA sequencing further decreases in price, this technology will certainly also impact toxicology and boost the level of big data for us to deal with. Secondly, live cell imaging allows the assessment of the dynamics of stress pathway activation at the single cell level (Wink et al. 2014). Through fluorescent protein genetic engineering approaches such as BAC-GFP transgenomics and CRISPR/Cas9 technology, the behaviour of relevant intracellular molecular markers derived from gene expression data can be followed in individual cells over time. With the advantage of high-throughput microscopy technology we can now monitor the dynamics of pathway activation at single cell level in hundred parallel treatment conditions with hundreds of cell per conditions (Wink et al. 2017). Parallel perturbation conditions allow the fast screening of temporal concentration responses of thousands of compounds or the application of RNA interference screening to identify relevant signalling molecules that determine the pathway activation in connection to the adverse outcome. Understandably, these high-content imaging approaches generate a massive amount of images and single cell information within each image: hence big data. Storage of these data in the public domain is still limited although possible.

All above novel technologies that will likely be part of the toolbox for the future modern molecular and cellular toxicologist clearly defines a new era for toxicology. Are current toxicologists fit for this big data analysis? There will be a high demand for the toxicologist that can both understand the scientific problems, design and perform the large scale experiments, but also be able to work with the Big Data. Current teaching programs typically lack behind here. We should be aware that the future toxicologists should possess a mix of molecular/cellular toxicology and data scientist skills. This will require investment in toxicology training programs so that our future toxicologist is able to extract the information from big data that allows them to answer the crucial scientific questions that will move the safety sciences forward. The time is now to review our toxicology training programs and make them fit for what is already at our doorsteps.

Summary

The perspectives and discussion at the WC-10 roundtable have illustrated the growing size and complexity of data in the 3Rs quest, the culmination of which is four main recommendations;

- To harness big data and create meaning from it, the infrastructure and resources for storage and analysis need to be up to par. People therefore need to care for, curate and describe their data (and metadata) in a consistent way, meaning we need standardized naming ontologies that are universally adopted to facilitate transitions across data sets.
- More, shared data and models will allow for better problem solving. If everyone were to share their data, it would increase our resources for better chemical assessment and improved product design.
- As with any research effort having a clear hypothesis, demonstrating transparent methodology and establishing scientific confidence in results is paramount. Whilst big data and associated methods can be used to help solve problems, people are still needed to define what the problems are. To this end, future 21st century toxicologists will need to be trained to include a focus on cell biology and informatics, and recognition must be given to the complex infrastructure and expertise that different disciplines and departments bring to the interpretation of big data. Realistically, we cannot expect the future to be a 'click and play' technology.
- Finally, toxicology data will need to be re-organised by mode of action to take advantage of big data and machine learning. Understanding big data and computational model outputs will still require human expertise, in particular in the areas of biological relevance and mechanistic interpretation.

Thus, we proffer these points, as highlighted in the WC-10 roundtable *'Big data in the 3R's—Outlook and Recommendations'*, to contribute to progress in systems biology/toxicology. Namely, the need for; (1) standard ontologies that are universally adopted to facilitate transition across data sets, (2) more shared data adherent to FAIR principles, (3) people to define what the problems are, and (4) toxicology data re-organised by mode of action.

Acknowledgements The authors thank Dr. Charles Schmitt (NIEHS) for his contributions and thoughtful review of this article. The views expressed in this manuscript do not necessarily represent the official positions of any federal agency. BvdW is supported through funding from the European Union's Horizo2020 research and innovation programme under Grant Agreement No 681002.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Bell SM, Phillips J, Sedykh A et al (2017) An integrated chemical environment to support 21st-century toxicology. Environ Health Perspect 125(5):054501. https://doi.org/10.1289/EHP1759

Casey WM (2016) Advances in the development and validation of test methods in the United States. Toxicol Res 32(1):9–14. https://doi.org/10.5487/TR.2016.32.1.009

ECETOC (2016) Building a prenatal developmental toxicity ontology. ECETOC, Brussels

Huang R, Sakamuru S, Martin MT et al (2014) Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. Sci Rep 4:5664. https://doi.org/10.1038/srep05664

Judson RS, Martin MT, Egeghy P et al (2012) Aggregating data for computational toxicology applications: the U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) system. Int J Mol Sci 13(2):1805–1831. https://doi.org/10.3390/ijms13021805

Macaulay IC, Voet T (2014) Single cell genomics: advances and future perspectives. PLoS Genet 10(1):e1004126. https://doi.org/10.1371/journal.pgen.1004126

Niepel M, Hafner M, Duan Q et al (2017) Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. Nat Commun 8(1):1186. https://doi.org/10.1038/s41467-017-01383-w

Sand S, Parham F, Portier CJ, Tice RR, Krewski D (2017) Comparison of points of departure for health risk assessment based on high-throughput screening data. Environ Health Perspect 125(4):623–633. https://doi.org/10.1289/EHP408

Shah I, Liu J, Judson RS, Thomas RS, Patlewicz G (2016) Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information. Regul Toxicol Pharmacol 79:12–24. https://doi.org/10.1016/j.yrtph.2016.05.008

Wambaugh JF, Setzer RW, Reif DM et al (2013) High-throughput models for exposure-based chemical prioritization in the ExpoCast project. Environ Sci Technol 47(15):8479–8488. https://doi.org/10.1021/es400482g

Wilkinson MD, Dumontier M, Aalbersberg IJ et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

Wink S, Hiemstra S, Huppelschoten S et al (2014) Quantitative high content imaging of cellular adaptive stress response pathways in toxicity for chemical safety assessment. Chem Res Toxicol 27(3):338–355. https://doi.org/10.1021/tx4004038

Wink S, Hiemstra S, Herpers B, van de Water B (2017) High-content imaging-based BAC-GFP toxicity pathway reporters to assess chemical adversity liabilities. Arch Toxicol 91(3):1367–1383. https://doi.org/10.1007/s00204-016-1781-0