BioSpectrum
the business of Bio & Health Sciences
ASIA EDITION

Transparent Toxicology: Towards improved reproducibility and data reusability

08 February 2017 | Opinion | By BioSpectrum Bureau

Transparent Toxicology: Towards improved reproducibility and data reusability

The concept of reproducibility is one of the foundations of scientific practice and the bedrock by which scientific validity can be established. However, the extent to which reproducibility is being achieved in the sciences is currently under question. Several studies have shown that much peer-reviewed scientific literature is not reproducible. One crucial contributor to the obstruction of reproducibility is the lack of transparency of original data and methods. Reproducibility, the ability of scientific results and conclusions to be independently replicated by independent parties, potentially using different tools and approaches, can only be achieved if data and methods are fully disclosed.

In the biomedical sciences, the issue is further complicated by the fact that we now typically deal with very large, multi-faceted and highly complex datasets (largely a result of technological advances which have led to a rapid growth in data generation). To truly facilitate reproducibility, not only does this data need to be fully disclosed, it also needs to be shared in a way that is practical for other scientists to search and scrutinise: it needs to be reusable.

Several initiatives have been established to facilitate data reusability and drive improved reproducibility. The outcome of one such recent initiative, designed and endorsed by a multi-disciplinary group of stakeholders, are known as the 'FAIR Principles'. They advise that data be findable, accessible, interoperable and reusable:

â€¢ Findable
(Meta)data are assigned a globally unique and persistent identifier
Data are described with rich metadata (defined in the 'reusability' bullet below)
Metadata clearly and explicitly include the identifier of the data it describes
(Meta)data are registered or indexed in a searchable resource

â€¢ Accessible
(Meta)data are retrievable by their identifier using a standardised communications protocol
The protocol is open, free, and universally implementable
The protocol allows for an authentication and authorisation procedure, where necessary

Metadata are accessible, even when the data are no longer available

â€¢ Interoperable
(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
(Meta)data use vocabularies that follow FAIR principles
(Meta)data include qualified references to other (meta)data

â€¢ Reusable
Meta(data) are richly described with a plurality of accurate and relevant attributes
(Meta)data are released with a clear and accessible data usage license
(Meta)data are associated with detailed provenance
(Meta)data meet domain-relevant community standards

At Philip Morris International (PMI), the concepts of scientific reproducibility and data reusability are particularly important. The U.S. Family Smoking Prevention and Tobacco Control Act defines a modified risk tobacco product (MRTP) as any that is 'sold or distributed to reduce the harm or risk of tobacco-related disease associated with commercially marketed tobacco products'. Through technological innovation and rigorous scientific assessment, we are developing a range of MRTPs, also called Reduced-Risk Products (RRPs). RRPs is the term we use to refer to products that present, are likely to present, or have the potential to present less risk of harm to smokers who switch to these products versus continued smoking. We have a range of RRPs in various stages of development, scientific assessment and commercialisation. Because our RRPs do not burn tobacco, they produce far lower quantities of harmful and potentially harmful compounds than found in cigarette smoke.

We are leading a full-scale effort to ensure that RRPs ultimately replace cigarettes. Recognising the need for RRP-related science to be reviewed, scrutinised and verified by the external scientific community, as well as by regulatory bodies such as the U.S. Food and Drug Administration, it is crucial that our methods and data are transparently shared in a way that allows easy review and understanding.

While there are several open data repositories that aim to integrate toxicological datasets, there are few that provide comprehensive, integrated and harmonised toxicological evidence for respiratory analysis and assessment (the area in which we are specifically concerned at PMI). As such, we have launched a proof-of-concept database and website (INTERVALS) that, in line with the FAIR Principles, shares results from studies conducted by PMI to assess RRPs. Our goal is to grow this initiative and establish a public repository for all preclinical RRP assessment methods and data.

INTERVALS

INTERVALS is an inhalation toxicology repository for MRTPs. It entails a database and a searchable web portal designed to allow the scientific community to easily retrieve preclinical information relevant to RRPs, from a single place and in a reusable format. Its purpose is to share this information across the scientific and regulatory communities, so as to facilitate reproducibility in predictive toxicology and risk assessment.

Concepts of Infrastructure and Data Sharing

API: Application Programming Interfaces

For proof-of-concept, a number of data-sets from PMI's RRP assessment studies have been integrated into the platform. These range from large in vivo inhalation studies to novel in vitro studies using three-dimensional human buccal and nasal tissue cultures. The studies investigated traditional physiological endpoints as well as genomics, proteomics, transcriptomics, and lipidomics profiles through a variety of advanced systems toxicology analytical techniques.

Alongside raw data, rich metadata are also provided to describe experiments, data production, data processing, and additional relevant information. Metadata are provided as ISA-Tab files, a standardised format for the collection and dissemination of complex metadata, categorised under three distinct headings: Investigation, Study, and Assay (I-, S-, and A-tabs, respectively). The I-tab summarises general information on the complete investigation, including people involved in the investigation, related publications, and protocol descriptions. The S-tab contains information on the study subjects and/or samples, their characteristics, and any treatments applied. The A-tab describes the smallest complete unit of experimentation that produces data associated with a subject. The S- and A-tabs are also linked through additional metadata provided in the I-tab.

In order to interpret these multi-faceted datasets, INTERVALS also utilises Garuda technology which connects different

databases, applications and services using a language-independent interface. It allows users to build customisable research workflows and bespoke tools by which to visualise and analyse data. An open, community-driven platform, Garuda is similarly aligned with the principles of transparency that underlie the INTERVALS platform. Ultimately, it enables researchers to integrate and extract meaningful information from large scale, multi-omics datasets.

IMPLICATIONS AND APPLICATIONS

The INTERVALS platform is one example of how PMI is sharing its data with the scientific community and encouraging external review and scrutiny. It has been designed to facilitate the practical reuse of our methods and data, and thus to demonstrate the reproducibility of our scientific findings. We also hope that the platform sparks wider interest and that other industries and institutions producing data relevant to RRPs submit that data to INTERVALS for the benefit of the entire community.

In addition, the INTERVALS concept has the potential to be useful in many other areas of societal concern: assessment of pharmaceutical products, air-pollution risk management, nanotechnology innovation, to name only a few. Essentially, any field which deals with large and complex data could benefit from this approach as the contemporary issues concerning transparency, data reusability and reproducibility are general across the sciences.

The INTERVALS proof-of-concept platform sits under the sbv IMPROVER umbrella, a collaborative initiative led and funded by PMI that aims to develop a robust methodology for verifying scientific methods and results. Based on the principles of crowd-sourcing, sbv IMPROVER is facilitating enhanced dialogue within the scientific community, transparency of research processes and open innovation in scientific discovery. Further information about sbv IMPROVER is available at: http://www.sbvimprover.com. The INTERVALS platform can be accessed at: https://systox.sbvimprover.com/.