"In God we trust.
All others must bring data."
W. Edwards Deming
(1900-1993)
Professor and author

Food for Thought …

Toward Good *In Vitro* Reporting Standards

Thomas Hartung [1,2], Rob de Vries [3], Sebastian Hoffmann [4], Helena T. Hogberg [1], Lena Smirnova [1], Katya Tsaioun [1], Paul Whaley [5] and Marcel Leist [2]

[1]Johns Hopkins Bloomberg School of Public Health, Center for Alternatives to Animal Testing (CAAT), Baltimore, MD, USA; [2]University of Konstanz, CAAT-Europe, Konstanz, Germany; [3]SYRCLE (SYstematic Review Centre for Laboratory Animal Experimentation), Department for Health Evidence (section HTA), Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands; [4]seh consulting + services, Paderborn, Germany; [5]Lancaster Environment Centre, Lancaster University, Lancaster, UK

Abstract

A good experiment reported badly is worthless. Meaningful contributions to the body of science are made by sharing the full methodology and results so that they can be evaluated and reproduced by peers. Erroneous and incomplete reporting does not do justice to the resources spent on conducting the experiment and the time peers spend reading the article. In theory peer-review should ensure adequate reporting – in practice it does not. Many areas of biomedical science have developed reporting standards and checklists to support the adequate reporting of scientific efforts, but *in vitro* research still has no generally accepted criteria. It is characterized by a "Wild West" or "anything goes" attitude. Such a culture may undermine trust in the reproducibility of animal-free methods and thus parallel the "reproducibility crisis" discussed for other life science fields. The increasing data retrieval needs of computational approaches, especially "big data" and artificial intelligence, makes the reporting quality even more important to allow the scientific community to take full advantage of the results.

The first priority of reporting standards is to ensure the completeness and transparency of the provided information (data focus). The second tier is the quality of data display that makes information digestible and easy to grasp, compare, and further analyze (information focus). This article summarizes a series of initiatives geared towards improving the quality of *in vitro* work and its reporting. This shall ultimately lead to Good In Vitro Reporting Standards (GIVReSt).

1 Introduction

Science that is not reported is a hobby activity. The scientific community thrives on sharing, interacting, and building on the progress reports of research groups. However, the quality of reporting results is not consistent, often because we do not think carefully about all the elements that together make up the critical information and should be shared. Time pressure to publish and inadequate quality control add to sometimes sloppy reporting. Leon M. Cautillo rightly warned, "*The bitterness of poor quality remains long after the sweetness of meeting the schedule is forgotten*".

There are still some colleagues who believe that they can maintain an edge over their competition and hold off others from using the same approach by withholding critical details in a publication. This thinking is fundamentally flawed, as science is an exercise of sharing. The possible short-term gain of keeping others from using the results will likely turn against the authors as their results will be deemed irreproducible and will less likely be built upon by others, which would create citations and grow the recognition of the methodology described and the reputation of the authors.

Today, increasingly we use computational tools to retrieve and combine information. Wilson Mizner is quoted as saying "*If you steal from one author it's plagiarism; if you steal from many it's research*". Though there is some truth to this, many research results merit re-analysis and only come to full fruition when combined with others. Data-sharing and the increasing accessibility

Received December 19, 2018;

ALTEX 36(1), 3-17. doi:10.14573/altex.1812191

Correspondence: Thomas Hartung, MD PhD, Johns Hopkins Bloomberg School
of Public Health, Center for Alternatives to Animal Testing (CAAT),
615 N Wolfe St., Baltimore, MD, USA
(THartun1@jhu.edu)

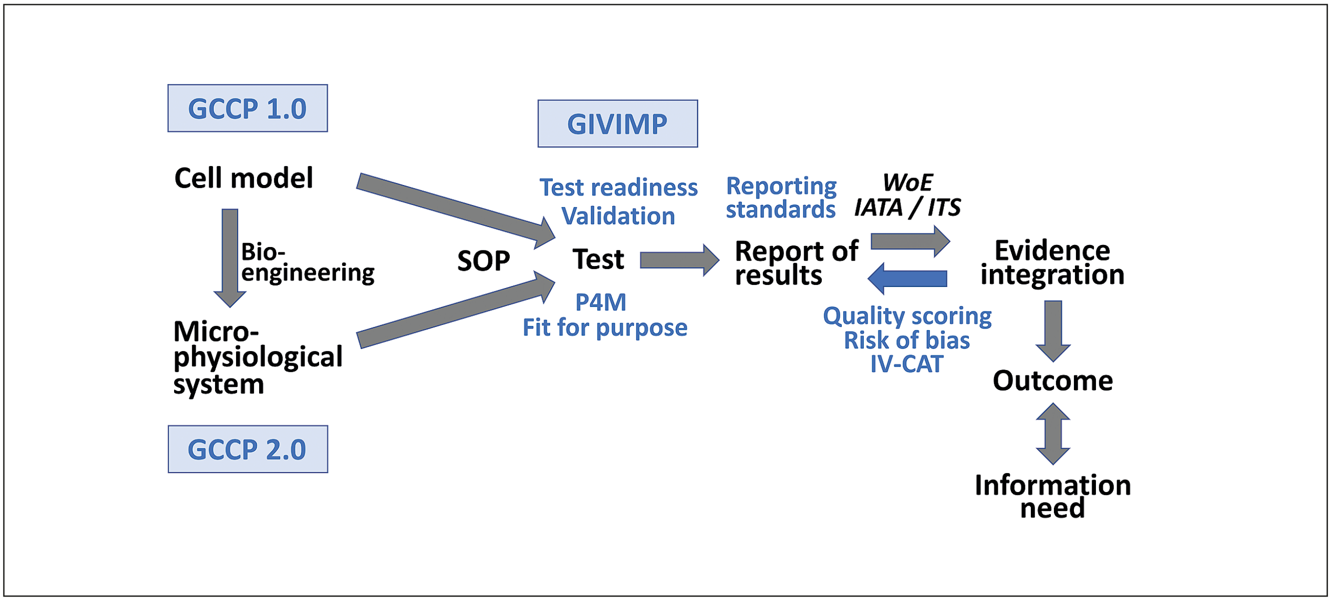

Fig. 1: The quality chain from cell model to satisfying an information need
A number of initiatives (in blue) detailed in the text aim to quality assure and control that *in vitro* approaches respond adequately to information needs: GCCP (Good Cell Culture Practice), GIVIMP (Good In Vitro Methods Practice), P4M (Public Private Partnership for Performance Standards of Microphysiological Systems) and IV-CAT (In-Vitro Critical Appraisal Tool). This is also critical for the integration of different types of test results / evidence streams by WoE (weight of evidence), IATA (integrated approaches to testing and assessment), and ITS (Integrated Testing Strategies).

of curated datasets and literature fuels this approach (Hartung, 2018). And the better the information is reported, the easier it can be retrieved and extracted. However, often evidence cannot be considered because of inadequate reporting. Identifying exclusion criteria for the respective pieces of evidence teaches us about the shortcomings of reporting.

A number of initiatives are converging toward the development of Good In Vitro Reporting Standards (GIVReSt): Good Cell Culture Practices (GCCP) (Coecke et al., 2005), which are currently being expanded to incorporate organotypic cultures, including microphysiological systems, as GCCP 2.0 (Pamies et al., 2017a, 2018; Pamies and Hartung, 2017) define minimum standards for cell culture and its documentation. The Guidelines for Transparency and Openness Promotion (TOP) of the Open Science Framework[1] promote more transparency when publishing work that may lead to better reproducibility of published data. The recent OECD guidance document for Good In Vitro Method Practices (GIVIMP)[2] (Eskes et al., 2017) sets standards for regulatory tests under Good Laboratory Practices (GLP). The development of evidence-based toxicology (Hoffmann and Hartung, 2006; Griesinger et al., 2009; Hartung, 2009; Stephens et al., 2013, 2016; Hoffmann et al., 2016, 2017), as promoted by the Evidence-based Toxicology Collaboration[3], includes the determination of quality, e.g., by quality appraisal, weight of evidence (WoE) assessment, and risk of bias analysis, as a component of systematic reviews and meta-analyses.

With the increasing use of *in vitro* data, these approaches have to adapt also to *in vitro* information. The different quality initiatives outlined below follow a "one-quality" concept for all elements of the *in vitro* information value chain: Only good (relevant) cell models can be used to build meaningful tests, which have to be adequately described together with their results to allow evaluation of their value for evidence integration so as to achieve an outcome responding to the information need (Fig. 1). No element of the quality chain can be weak, otherwise the information need is not adequately satisfied. Obviously, many elements overlap, first of all because some principles of quality assurance and quality control are universal and also apply to cell culture. Others differ mainly with respect to whether they are applied prospectively or retrospectively, e.g., whether advising how to design or report a test, or how to assess a test result. This is a two-way-street as learning experiences gained by analyzing scientific reports teach us how to improve reporting. Other initiatives differ more with respect to addressing regulatory needs, like GIVIMP[2] in the context of Good Laboratory Practice. The regulatory use of *in vitro* methods has brought about the concepts that feed into validation (Leist et al., 2012). Again, there is a lot of cross-talk with reporting standards as the method definition in

[1] https://osf.io/ud578/
[2] https://bit.ly/2KwBtOh
[3] http://www.ebtox.org

Tab. 1: Quality assurance *versus* quality control for *in vitro* reporting

Quality Assurance (QA)	Quality Control (QC)
QA is a set of activities for ensuring quality in the processes by which, in our case, reports are prepared.	QC is a set of activities for ensuring quality in the report. The activities focus on identifying defects in the actual reports produced.
QA is a managerial tool, typically implemented by the laboratory head. The implementation of GIVReSt would be such a tool.	QC is a corrective tool, which can be delegated to someone checking report quality. The implementation of IV-CAT would be such a tool.
QA aims to prevent defects with a focus on the process used to make the report. It is a proactive quality process.	QC aims to identify (and correct) defects in the finished report. Quality control, therefore, is a reactive process.
The goal of QA is to improve development and processes so that defects do not arise when the report is being developed.	The goal of QC is to identify defects after a report is developed and before it is released. This can be on the side of the laboratory, the peer reviewer, the publisher, or the data extractor for further analyses.
Prevention of quality problems in reporting through planned and systematic activities including adequate documentation of the work to be reported.	The activities used to achieve and maintain the report quality.
Establish a good quality management system by training and supervision of all involved as well as the assessment of its adequacy. Periodic conformance audits of the operations of the system.	Finding and eliminating sources of quality problems so that requirements are continually met. These include guidance documents and checklists used by authors and journals.

form of a Standard Operating Protocol (SOP) is the fundamental basis for any method evaluation. The concepts associated with validation are necessarily in constant evolution as they must be adapted to the individual new methods (Hartung, 2007), use scenarios and rigor; different names have been attributed to these variants (e.g., qualification, fit-for-purpose validation, performance standards, quality management, etc.) but they all include critical information on method definition, reporting, and validation status, which cross-fertilize with GIVReSt.

2 The principles of quality assurance (QA) and quality control (QC)

According to the World Health Organization (WHO), quality assurance is a "*wide ranging concept covering all matters that individually or collectively influence the quality of a product*"[4]. The International Organization for Standardization (ISO) developed the ISO 9000 series standards, i.e., a set of international standards on quality management and quality assurance developed to help organizations effectively document the elements that need to be implemented to maintain an efficient quality system. The standards, initially published in 1987, are not specific to any particular industry, product or service. They underwent major revision in 2000 and now include ISO 9000:2005 (definitions), ISO 9001:2008 (requirements), ISO 9004:2009 (continuous improvement), and ISO 9001: 2015 (risk management). ISO 9000[5] defines quality assurance (QA) as "*part of quality management focused on providing confidence that quality requirements will be fulfilled*". ISO 9000, clause 3.2.10, defines quality control (QC) as "*part of quality management focused on fulfilling quality requirements*". It is the part of Good Manufacturing Practice (GMP) concerned with sampling, specification, and testing, documentation, and release procedures, which ensures that the necessary and relevant tests are performed and the product is released for use only after ascertaining its quality. Simply put, QA focuses on the process of quality, while QC focuses on the quality of output. Table 1 further differentiates QA and QC (see also[6]) for reports on *in vitro* work.

QA thus is a verification activity that verifies that you are doing the right thing in the right manner, which prevents the bugs. QC is a validation activity that validates the product against the requirements, i.e., it identifies the bugs and gets them fixed. QA and QC share tools, including defining processes, quality audit, selection of tools, and training. A quality culture should be established in which everyone feels responsible for maintaining the quality of the product.

An important QC tool is the quality evaluation of articles, which can in turn instruct which information needs to be reported to meet the quality criteria. A scoping review of quality scoring tools in the field of toxicology was carried out by EBTC (Samuel et al., 2016). Relevant here, seven documents addressed the methodological quality of *in vivo* and/or *in vitro* studies, three were aimed primarily at the toxicological or environmental health communities, while four were aimed more broadly. Fifteen criteria for addressing the methodological quality of *in vivo* and *in vitro* studies were proposed

[4] https://www.who.int/medicines/areas/quality_safety/quality_assurance/en/

[5] https://asq.org/quality-resources/iso-9000

[6] https://reqtest.com/testing-blog/quality-assurance-vs-quality-control-differences-2/

in at least four of the relevant documents; eight of these 15 criteria are readily aligned with standard risk of bias categories: Selection bias (baseline characteristics, similarity/appropriate control group selection, allocation concealment, randomization); performance bias (blinding of researchers); detection bias (blinding of outcome assessors); attrition bias (complete outcome data); reporting bias (selective outcome reporting); confounding bias (account for confounding variables); appropriate statistical methods (sample size determination, statistical analysis); appropriate/controlled exposure (incl. characterization); optimal time window used; statement of conflict of interest / funding source; test substance / treatment details; test organism / system. These criteria can now be translated into reporting recommendations.

3 Good Cell Culture Practice (GCCP) and its relation to reporting standards

GCCP was initiated by one of the authors (T.H.) in 1996 by organizing a symposium in Berlin, Germany. In a well-documented process, it was developed toward a guidance document in 2005 (Gstraunthaler and Hartung, 1999; Hartung et al., 2002; Coecke et al., 2005). Two workshops in 2015 (Pamies et al., 2017a, 2018) reopened the guidance development in order to adapt it to technical progress, especially the increasing importance of stem cells and bioengineering technologies that enable the establishment of organotypic cultures (microphysiological systems) (Marx et al., 2016).

Fundamentally, GCCP should form the basis of *in vitro* reporting standards. As a key principle, GCCP includes needs for documentation "*Principle 3: Documentation of the information necessary to track the materials and methods used, to permit the repetition of the work, and to enable the target audience to understand and evaluate the work*" (Coecke et al., 2005, referred to as GCCP 1.0). Though this is intended primarily for internal communication, it also forms the basis for external communication, although usually even more information is needed for transfer of a method to another laboratory as the infrastructure elsewhere is different.

GCCP 1.0. includes "*Table 2: Examples of requirements for documentation concerning the origins of cells and tissues*". Features include: ethical and safety, species/strain, source, sex, age, number of donors, health status, any special pre-treatment, organ/tissue of origin, cell type(s) isolated, isolation technique, date of isolation, operator, supplier, informed consent, material transfer agreement, medical history, pathogen testing, shipping conditions, state of material on arrival, cell line identification and authentication, and mycoplasma testing. GCCP 1.0 also includes "*Table 3: Examples of requirements for documentation concerning the handling, maintenance and storage of cells and tissues*". Features suggested for documentation are: ethical and safety, morphology, histopathology, purity of isolation, phenotype, type of culture, culture medium, feeding cycles, growth and survival, characteristics, initial passage number, confluency at subculture, subculturing details, induction of differentiation, identification and authentication, ageing, mycoplasma testing. GCCP 1.0 further includes "*Table 4: Details to be included in papers for publication in journals, using the example of mouse 3T3 cells*". Suggested reporting features include: type of culture, cell/tissue type, species, origin, description, catalogue/product number, basic culture medium, serum, antibiotics, other additives, frequency of medium change, culture flasks for stock cells, culture plates for test, culture well inserts, surface coating, subculture frequency, subculture split ratio, detachment solution, usable passage range, passage number at receipt, passage number at use, maintenance conditions, storage conditions, relevant standard operating procedures/guidelines, references, and further comments.

These lists already give us a fairly comprehensive overview of reporting features for traditional cell cultures. GCCP 2.0 is trying to complement these for stem cell-derived models and microphysiological systems (MPS). Stem cells require information about the donors similar to primary cell work, but also on reprogramming and maintenance of stemness, differentiation protocols, etc. The work with MPS requires similar documentation efforts as they often work with stem cells and their derivatives. MPS have the peculiarity that they are often cultured for prolonged periods of time, which might require additional documentation. Since MPS mimic organ functions, it is important to document which functions these are and how they are assessed. Organ-on-a-chip MPS require additional reporting, including information about the utilized device, such as the material it is made of and technical specifications, e.g., the medium flow rate in microfluidic devices, frequency of medium addition, nutritional support, etc.

The Pamies et al. (2017a) workshop recommends: "*A high-quality scientific report should cover the objective of the work, the protocols and SOPs used, planning and experimental design, execution of the study, definition of the test conditions, test procedure, test acceptance criteria, data collection and analysis as well as a discussion of the outcome. The extent to which the study adheres to relevant standards, regulations, guidelines or guidance documents should be stated, along with adherence to safety and quality assurance procedures. This could also include a statement of compliance with the GCCP principles. Reports on cell and tissue culture work should address a minimum set of information that covers the origins of the cells, characterization, maintenance, handling, and traceability of the cells, and the procedures used*". The respective Table 7 lists examples of information needs: safety information, ethical issues, species, strain, source, sex, age, race, number of donors, health status, tissue of origin, cell type(s) isolated, date of isolation, isolation technique, operator, supplier, informed consent, material transfer agreement, medical history of donor, pathogen testing, shipping conditions, condition of material on arrival, identification and authentication, and mycoplasma testing.

4 The reporting of tests

The difference between a model and a test is of critical importance. To give an example, primary rat liver cells are a model; they can be cultured under various conditions. A test serves an information need and follows a defined protocol, often referred to as a Standard Operating Protocol (SOP), which details and defines how the model is employed to obtain test results.

This includes ideally:

- A definition of the scientific purpose of the method
- A description of its mechanistic basis
- The case for its relevance
- The protocol, including: standard operation procedures, including a defined exposure scheme; specification of endpoints and endpoint measurements; derivation, expression, and interpretation of results, i.e., the (preliminary) prediction model; and inclusion of adequate controls.

In general, SOPs, which mainly contain detailed descriptions of each analytical method, are essential for maintaining the same analytical quality over a long period of time. The procedures are a prerequisite for correct transfer of methods from one laboratory to another.

The contents of an SOP according to the Guidelines on Standard Operating Procedures for Clinical Chemistry[7] are typically as follows: (1) introduction, (2) principle of method, (3) specimen type, collection and storage, (4) reagents, standards and control – preparation and storage, (5) equipment, glassware and other accessories, (6) detailed procedure, (7) calculations, calibration curve, (8) analytical reliabilities – (QC and statistical assessment), (9) hazardous reagents, (10) reference range and clinical interpretation, (11) limitations of method (e.g., interfering substances and troubleshooting), (12) references, (13) date and signature of authorization, (14) (effective date + schedule for review).

A very comprehensive description of test methods was developed by OECD guidance document 211 (Box 1) (OECD, 2014). While some of this is geared toward GLP and the context of regulatory use, it provides a very valuable list of aspects to be reported.

Box 1 Summary of features to report according to OECD Guidance Document No. 211 "Guidance Document for Describing Non-Guideline *In Vitro* Test Methods"

1. General information
1.1 Assay name (title)
1.2 Summary
1.3 Date of Method Description (MD)
1.4 MD author(s) and contact details
1.5 Date of MD update(s) and contacts
1.6 Assay developer(s)/Laboratory and contact details
1.7 Date of assay development and/or publication
1.8 Reference(s) to main scientific papers
1.9 Availability of information about the assay in relation to proprietary elements
1.10 Information about the throughput of the assay
1.11 Status of method development and uses:
i) Development status
ii) Known uses
iii) Evaluation study
iv) Validation study
v) Regulatory use
1.12 Abbreviation and Definitions

2. Test Method Definition
2.1 Purpose of the test method
2.2 Scientific principle of the method.
2.3 Tissue, cells or extracts utilised in the assay and the species source
2.4 Metabolic competence of the test system
2.5 Description of the experimental system exposure regime
2.6 Response and Response Measurement
2.7 Quality / Acceptance criteria:

- Experimental data (storage/archiving), indicate unit of measurement of the raw data, not only transformed data
- Experimental system(s) used
- Equipment used, calibration program
- Availability of internal standards (e.g., positive and negative controls, reference chemicals,
- Performance benchmarks
- Standards followed such as good cell culture practice, if relevant
- Criteria to accept or reject experimental data
- Limit of detection and limit of quantification, detection range.

2.8. Known technical limitations and strengths
2.9 Other related assays that characterise the same event as in 2.1

3. Data interpretation and prediction model
3.1 Assay response(s) captured in the prediction model
3.2 Data analysis
3.3 Explicit prediction model
3.4 Software name and version for algorithm/prediction model generation

4. Test Method Performance
4.1 Robustness of the method
4.2 Reference chemicals/chemical libraries, rationale for their selection and other available information
4.3 Performance measures/predictive capacity (if known)
4.4 Scope and limitations of the assay, if known

5. Potential Regulatory applications
5.1 Context of use

6. Bibliography

7. Supporting information

[7] http://apps.searo.who.int/PDS_DOCS/B0218.pdf

Recently, we tried to define test-readiness criteria (Bal-Price et al., 2018) to support an assessment of suggested tests for consideration in a regulatory context. Such an approach is meant to render the evaluation/validation of a test method more flexible. The idea is to score all aspects of test readiness, under the assumption that various applications of the test may have more or less stringent requirements. For instance, a given test method may be used as a pre-screen to identify compounds for further evaluation. The requirements for specificity and robustness would be moderate in this case, but if a large number of compounds is to be tested, the assay needs to score highly for throughput. The same assay may also be used to predict adverse effects of a drug before it enters human trials. In such a case, the criteria for robustness, documentation, and accuracy would be much more stringent, while throughput is not an important criterion. Altogether, more than 60 scoring items for test readiness were defined and bundled into 13 groups (e.g., description of the test system/model; the test prediction model; or the suitability for high-throughput applications). Guidance was also provided on which readiness groups would need to score highly for various types of assay use, and an exemplary scoring was performed for new approach methods (NAM) in the area of developmental neurotoxicity. The readiness criteria are broadly defined and may be applied to any toxicological field. For instance, the test methods of the EU-ToxRisk project (Daneshian et al., 2016) are now being evaluated according to this scoring scheme.

4.1 Test quality

Test quality in this context means the status of assessing the quality of an assay, i.e., its fitness for purpose to satisfy a certain information need. This assessment will change over time with increasing accumulated experience. As this assessment cannot be done every time a test is to be used, qualification procedures have been put into place. Arguably, the most comprehensive (and rigid) of these are formal validations as applied to alternative methods in a regulatory context. This often has been a decade-long process costing several hundred thousand $ per test. For this reason, there is a perceived need to communicate also less stringent but more practical indications of test quality such as test readiness, compliance with performance standards, and other steps toward validity assessment.

Readiness qualification for a test may differ depending on its application. Official validated OECD Test Guidelines used for hazard assessment follow the OECD Guidance Document 34 (GD 34) (OECD, 2005). However, this is not always suitable for qualifying NAMs as these can be non-standardized tests and may not be transferable to other laboratories due to very specific equipment, e.g., chip platforms, robotic systems, and fluidics. Complementary to GD 34 is OECD GD 211 for non-guideline *in vitro* test methods (OECD, 2014), which proposes how to harmonize the reporting of these new methods to assess their relevance, quality, and readiness. The guidance document is divided into five main parts: (1) general information; (2) test method definition, including the purpose, rationale, intended use of the test, and detailed information and protocol on both model and assays (the test). In addition, the test method definition should identify quality and acceptance criteria based on reference compounds and detail how to accept or discard experimental data; (3) data interpretation and prediction model, if applicable as stand-alone test or in a testing battery; (4) test method performance; including robustness of the method (reproducibility and transferability, within and between laboratories), reference chemicals, performance measures and predictive capacity. The last part is (5) potential regulatory applications. Depending on the intended use of the test, these criteria have different importance. If the test is intended to be used in academia with the aim to study plausible mechanisms, the test might have high readiness despite not being ready for a regulatory context (Bal-Price et al., 2018), e.g., without sufficient data on (4) and (5). To better judge the relevance and quality of data produced by NAMs, performance standards with defined acceptance criteria should be identified.

The performance standard concept was first introduced as ECVAM's Modular Approach (Hartung et al., 2004) and adopted into OECD GD34 a year later. As per OECD, "*Performance standards are based on an adequately validated test method and provide a basis for evaluating the comparability of a proposed test method that is mechanistically and functionally similar… Three main elements of performance standards (1) essential method components; (2) list of reference substances; (3) accuracy and reliability performance values*". When new methods are developed, which are similar in performance to the validated methods, a larger set of chemicals, ideally avoiding the reference compounds, should be used for this development. Reference compounds then should be used to demonstrate that the test performance is comparable to the validated test method. The examples of OECD-defined performance standards for some *in vitro* tests are available[8].

The originally defined concept of performance standards (Hartung et al., 2004; OECD, 2005) is currently undergoing a revival, owing to rapid advances in biotechnology, such as microphysiological systems, bioprinting, and organ-on-a-chip. These promising but biologically complex technologies require reconsideration of several aspects of formal validation, and the definition of performance standards is part of it. The new technologies are much more complex and are not only designed to replace a given animal test. Instead, they are multipurpose models that allow many different tests to be set up, which complicates the validation process. For example, our human BrainSpheres model ("mini-brains") (Pamies et al., 2017b) can be used to test neurotoxicity, developmental neurotoxicity, de- and remyelination, and to study many different disease aspects such as infection, cancer, trauma, stroke, neurodegeneration, etc. We discussed the concept of performance standard-based validation in this series earlier this year (Smirnova et al., 2018, and Figure 3 therein).

Performance standards for complex and multipurpose *in vitro* systems such as organ-on-a-chip are going beyond "minimum

[8] http://www.oecd.org/chemicalsafety/testing/performance-standards.htm

performance standards" for a new method, similar to an existing validated one. The purpose of the system and the engineering goals should be defined, and quality assurance should be provided (which will include both correlative and mechanistic validation, Hartung et al., 2013). Basically, a performance standard, defined by expert consensus, should meet the test purpose ("fit-for-purpose" concept) and anchor different elements of correlative and/or mechanistic validation. This can be compared to the point of reference concept, introduced at an ECVAM workshop (Hoffmann et al., 2008). The test should comply with a point of reference, defined by experts, rather than being compared to a traditional animal test.

5 The reporting of test results

5.1 Laboratory proficiency

This is primarily relevant for regulatory applications, though it would actually benefit all reported scientific data. Proficiency refers to the training and quality assurance / control a laboratory and its operators have undergone to demonstrate that they are capable of running the given test. This might include results of ring trials, intra-laboratory reproducibility assessments, historical controls, assessment of reference materials, etc. While this is certainly not (yet) a common scientific reporting standard, elements of this could be encouraged.

5.2 Test execution *versus* reporting

In theory, execution and reporting should be the same, in practice this is not always the case. For example, while reference to method papers or public protocols is a preferred way to ensure comprehensive reporting, any deviations from such protocols need to be meticulously documented. In longer studies, exchange of materials and machines can occur, which also may require reporting.

5.3 Data analysis and representation

Appropriate data analysis and representation is not an experimental aspect that only should be considered once the data have been obtained; it is determined to a large extent when an experiment is planned. Once the experimental purpose, e.g., a scientific hypothesis to be tested, is clearly defined, an appropriate experimental approach to inform the purpose or to test the hypothesis can be designed. The characteristics of the data, such as the experimental unit, the data type, and the (assumed) distribution, will guide various experimental features, including different levels of replication, number of test concentrations, and experimental repetitions (Lovell and Omori, 2008).

Once the data have been obtained, any deviation from the planned analyses needs to be considered carefully and should be justified. Most importantly, no data should be omitted. Exclusion of aberrant data – often called outliers – can usually not be justified with statistical methods, but should be clearly linked to observations made, e.g., a handling error. Unjustified exclusion of data disqualifies all experimental data as it can be abused to bias results.

The data thus should be summarized and analyzed as planned when designing the experiment. Any statistical analysis should be specified in sufficient detail allowing reproduction. The choice of the selected statistical approaches as such, but also choices made in applying them, should be justified. Graphical representation should provide as much information as possible, e.g., including a measure of dispersion for summary data such as the mean, and be closely linked to the statistics (if any). Interestingly, the widely used bar chart is often not the best option (Krzywinski and Altman, 2014). However, graphs should not be overburdened with information, possibly distracting from the most important features.

5.4 Publication

The TOP guidelines[9], which outline a framework that over 1,000 journals and publishers have elected to follow, has the following underlying principles: "*While specific needs and expectations vary across fields, the effective use of research findings relies on the availability of core information about research materials, data, and analysis*". A workshop hosted by the Center for Open Science in September 2017 and subsequent discussions led to a position paper suggesting a standardized approach to reporting and a working group of journal editors and experts joining to develop a minimal set of reporting standards for research in the life sciences. This working group drew from the collective experience of journals implementing a range of different approaches designed to enhance reporting and reproducibility (e.g., STAR Methods), existing life science checklists (e.g., the *Nature Research* reporting summary), and results of recent meta-research studying the efficacy of such interventions (e.g., Macleod et al., 2017; Han et al., 2017) to devise a set of minimal expectations that journals could agree to ask their authors to meet.

The working group aims for three key deliverables:

- *A "minimal standards" framework setting out minimal expectations across four core areas of materials (including data and code), design, analysis, and reporting (MDAR)*
- *A "minimal standards" checklist intended to operationalize the framework by serving as an implementation tool to aid authors in complying with journal policies, and editors and reviewers in assessing reporting and compliance with policies*
- *An "elaboration" document or user guide providing context for the "minimal standards" framework and checklist*

They state: "*While all three outputs are intended to provide tools to help journals, researchers, and other stakeholders with adoption of the minimal standards framework, we do not intend to be prescriptive about the precise mechanism of implementation, and we anticipate that in many cases they will be used as a yardstick within the context of an existing reporting system. Nevertheless, we hope these tools will provide a consolidated view to help raise reporting standards across the life sciences*". According to their website, draft versions of these tools are anticipated by spring 2019. They want to work with a wider group of journals as well

[9] https://cos.io/our-services/top-guidelines/

as funders, institutions, and researchers to gather feedback and seek consensus towards defining and applying these minimal standards. As part of this feedback stage, a "community pilot" involving interested journals to test application of the tools shall be conducted. The "minimal standards" working group includes Karen Chambers (Wiley), Andy Collings (eLife), Chris Graf (Wiley), Veronique Kiermer (Public Library of Science), David Mellor (Center for Open Science), Malcolm Macleod (University of Edinburgh), Sowmya Swaminathan (Nature Research/Springer Nature), Deborah Sweet (Cell Press/Elsevier), and Valda Vinson (Science/AAAS). Veronique Kiermer (vkiermer@plos.org) and Sowmya Swaminathan (s.swaminathan@us.nature.com) are given as contacts for interested parties.

5.5 Peer-review *versus* reporting

In vitro studies are becoming an increasingly important source of data in chemical risk assessment. There are concerns, however, about the methodological quality of these studies, and multiple initiatives are being undertaken to improve this. These initiatives to improve study quality focus on three target groups: (1) researchers designing, conducting and reporting primary *in vitro* exposure studies, (2) peer reviewers of journals advising on whether to publish a submitted manuscript, and (3) authors of systematic reviews aiming to assess the risk of bias/study quality of the primary *in vitro* studies included in their review. The Evidence-based Toxicology Collaboration is currently conducting a project, led by Paul Whaley and Rob de Vries, aimed at developing a tool specifically targeted at the second group, peer reviewers. The objective of this tool, called IV-CAT (In-Vitro Critical Appraisal Tool; de Vries and Whaley, 2018), is to help ensure comprehensive and exacting peer-review of *in vitro* toxicology studies to increase the quality (understood as "fitness-for-purpose") of published *in vitro* research. Completeness of reporting will be an important criterion within IV-CAT, but publishability also depends on, e.g., the scientific importance of the study and its internal validity.

The development of IV-CAT will consist of five steps, which are extensively described in a prespecified protocol[10]. The first step, a systematic review of existing critical appraisal tools and reporting standards for *in vitro* research in order to collect criteria that should be included in IV-CAT, is currently being conducted.

6 Reporting enabling evidence retrieval

In the United States, the National Libraries of Medicine (NLM) provide access to many online toxicology and environmental health resources, for example, LiverTox®[11] and the TOXNET®[12] integrated system of many databases. The overall coverage of NLM's resources includes chemicals and drugs, diseases and the environment, environmental health, exposure science, occupational safety and health, poisoning, emergency response, risk assessment and regulations, and toxicology. Users of NLM's toxicology and environmental health resources include first responders, academic researchers, students, government agencies (e.g., CPSC, EPA, FDA, OSHA, and NIOSH), industry, healthcare providers, and planners.

Currently, consistency in terminology to report study methods and findings can be low, as is being found in, e.g., the epigenetic literature and especially for mechanistic research with its variety of assays and outcomes that it investigates and attempts to relate into disease progression pathways (Whaley et al., *in preparation*). This presents an obstacle to identifying and organizing relevant literature, as indexing systems will not always keep up with or be able to accommodate researcher-generated synonyms for research concepts, resulting in potentially relevant evidence for a systematic review being missed due to the use of incorrect or unconventional terminology in study reports. While interventions to encourage the use of consistent terminology can be made, they are unlikely to be completely effective considering the volume of papers being published, the number of authors engaged in research, and the natural variation in language use between different research communities.

In this context, what is more achievable and of higher practical use in evidence retrieval is the definition of universal ontologies (Whetzel et al., 2011; Hardy et al., 2012a,b) to cover the concepts being used by researchers as they describe their research practices and study results, and the relationships between them. The National Center for Biomedical Ontology (NCBO)[13] defines an ontology as "*a kind of controlled vocabulary of well-defined terms with specified relationships between those terms, capable of interpretation by both humans and computers*". Rather than focusing exclusively on consistent reporting by researchers, ontologies allow research data to be labeled and mapped onto a set of ontological concepts as a universal meta-language describing day-to-day research practices. This provides a structure for organizing, cataloguing, and retrieving data from the literature, which can be created by groups such as GCCP and applied on top of the literature rather than relying on individual researchers to address issues regarding, e.g., common use of terminology, by themselves on a case-by-case basis. Readier identification of when different researchers are investigating the same (or related) things, even though they are using different terminology in describing what they are doing and may be unaware of those similarities or relations themselves, improves rates of retrieval of relevant evidence from literature databases and provides additional opportunities for synthesizing comparable studies in understanding health impacts of exposure to chemical substances.

A classic, long-running example of addressing a number of aspects of the ontology challenge is PubMed. The taxonomy in PubMed is in the process of adapting MeSH terms for toxicol-

[10] https://zenodo.org/record/1493498

[11] https://livertox.nih.gov/

[12] https://toxnet.nlm.nih.gov/

[13] https://www.bioontology.org/community/learning-about-ontologies

ogy. Currently, the toxicology medical search heading (MeSH) term (Toxicology [H01.158.891]) has only three sub-categories:

- Ecotoxicology [H01.158.891.211]
- Forensic Toxicology [H01.158.891.424]
- Toxicogenetics [H01.158.891.850]

In order to facilitate literature searching and retrieval of relevant articles, in particular for systematic reviews, the toxicology taxonomy needs to be more granular and needs to include the ability to search for toxicology endpoints in different organisms (human, vertebrate animals, invertebrate animals, cell models, plants, etc.), organ systems (heart, liver, skin, neuro-, etc.), and should include mechanisms of toxicity. Toxicity endpoints need to be easily associated with the papers that mention them (e.g., following ECHAs classification system). Finally, more detailed indexing of the chemicals being studied would be helpful in retrieving appropriate articles (i.e., matching assays to chemicals the way it is done by PubChem[14]).

To start addressing these deficiencies, NLM convened a workshop on April 6, 2018 on "Re-envisioning NLM's Toxicology and Environmental Health Resources" led by Drs. Pertti (Bert) Hakkinen and Florence Chang. EBTC participated in the workshop as one of the stakeholders along with EFSA, US EPA, US FDA, and other academic and government stakeholders. The goal of the workshop was to optimize NLM's electronic resources in the areas of toxicology and environmental health by assessing and re-envisioning them to ensure that they continue to meet the changing needs of a diverse set of users. The purpose of the workshop was to gain critical insight from participants to help NLM's Toxicology and Environmental Health Information Program (TEHIP) ensure its continued relevance and to identify opportunities to enhance the types of content offered for the many and varied users of NLM's toxicology and environmental health resources. The workshop built upon the insights gained from a web survey of workshop participants and other users of NLM toxicology and environmental health resources. The workshop provided insights that will be used to thoughtfully inform the development of the new generation of NLM resources for toxicology, environmental health, and other relevant information.

Text mining and artificial intelligence-enabled literature retrieval are examples of needs and opportunities as new technologies emerge and mature. There is growing interest in using machine learning approaches to priority rank studies and reduce human burden in screening literature (Cohen et al., 2006) when conducting systematic reviews. In addition, identifying addressable questions during the problem formulation phase of systematic review can be challenging, especially for topics that have a large literature base. SWIFT-Review (Howard et al., 2016), SysRev[15], DistillerSR (Matwin et al., 2010), and other software applications are in the process of developing various priority ranking algorithms to identify studies relevant to a given research question and may be used at various stages of the systematic review process: for question formulation, categorization of literature, de-duplication, or visualization of research results. These approaches hold great promise for facilitating accurate literature retrieval and facilitating systematic review, but in order for the software to be more accurate and to be able to read and classify full-text articles, large literature databases such as NLM's PubMed, as well as BIOSIS[16] and Web of Science[17] can make this work easier by improving the taxonomy and the ontologies. Publishers and journal editors can contribute by putting together requirements and enforcing measures for structured abstracts[18] (i.e., an abstract with distinct, labeled sections, e.g., Introduction, Methods, Results, Discussion, for rapid comprehension) and complete data reporting including raw data. In the long term, structured abstracts and requirements for complete transparent data reporting will contribute not only to the literature retrieval accuracy, but also to the fidelity of the results, reproducibility, and credibility of science.

7 Reporting enabling data integration

A major role in promoting the quality of studies can be played by systematic reviews. These comprehensive, structured, and transparent summaries of existing evidence are already standard practice within clinical medicine, but until recently were relatively rare in other fields such as preclinical research and toxicology. Over the last years, however, the number of systematic reviews in these latter fields has been steadily increasing (Hoffmann et al., 2017; Whaley et al., 2016); where we have a rudimentary count of them, they are showing exponential growth (Whaley et al., 2016).

Although the basic principles of systematic reviews are very similar for all fields of application, applying the principles to these new fields can generate substantial challenges. One clear example is the poor reporting of primary studies, which hampers assessment of the actual risk of bias of those studies. A standard step within systematic reviews is the assessment of the internal validity of the included studies. An experiment can be considered internally valid if the differences in results observed between the experimental groups can, apart from random error, be attributed to the intervention under investigation. This validity is threatened by certain types of bias, where systematic differences between experimental groups other than the intervention of interest are introduced, either intentionally or unintentionally (de Vries et al., 2014). The risk of bias assessment within a systematic review tries to assess the actual risks of bias in the primary studies, for example the risk of selection bias due to a lack of proper randomization.

[14] https://pubchem.ncbi.nlm.nih.gov

[15] www.sysrev.us

[16] https://www.ebsco.com/products/research-databases/biosis-previews

[17] http://wokinfo.com

[18] https://www.nlm.nih.gov/bsd/policy/structured_abstracts.html

SYRCLE[19], the Systematic Review Centre for Laboratory animal Experimentation, has been conducting and supporting systematic reviews in the field of preclinical research (both animal and *in vitro* studies, e.g., Golbach et al., 2016) for nearly 10 years. In their experience, in 80-90% of the cases it is not possible to assess the actual risk of bias, because crucial details of the design and conduct of the studies are not reported. In these cases, the risks of bias have to be scored as "unclear"; after all, the researchers might have applied measures to avoid bias, such as randomization and blinding, but just not reported them (Hooijmans et al., 2014). Moreover, if in so many cases it is not possible to assess the actual risk of bias, it is also not possible to assess the impact of potential biases, for example by conducting a subgroup analysis comparing studies that did and did not randomize allocation.

At this point in time, it is therefore difficult to determine to what extent the results reported in the primary studies are over- or underestimations of the true effects of the experimental interventions applied. Improving reporting standards is therefore not only important for reproducibility but also for assessing reliability.

8 Outcome reporting *versus* information needs

While this might sound trivial, any scientific report should have a purpose, an information need, to which the study and in consequence the report or article is responding. This information need should be stated explicitly and also an assessment should be made to which extent the need has been met and what is needed to further approach this goal.

9 Earlier *in vitro* reporting standard initiatives

There is no shortage of guidance on how to report science – the Equator network (Enhancing the quality and transparency of health research) identified 408 reporting guidelines[20]. They suggest templates and checklists with the message "*your article is not complete until you have done all of these things*". For example, a group of journal editors and experts in reproducibility and transparent reporting have been developing a framework for minimal reporting standards in the life sciences[21]. Table 2 lists some of the reporting guidance relevant to *in vitro* work (extracted and modified from the NIH website[22] and Wikipedia[23]).

[19] https://bit.ly/2Gw1yjU

[20] https://www.equator-network.org/reporting-guidelines/

[22] http://crosstalk.cell.com/blog/towards-minimal-reporting-standards-for-life-scientists

[23] https://www.nlm.nih.gov/services/research_report_guide.html

[12] https://en.wikipedia.org/wiki/IMRAD

Tab. 2: Existing reporting guidance relevant to *in vitro* work

Organization / Guideline	Description	References
American Medical Association Manual of Style[a]	A manuscript style guide for medical science. Current 10th edition to be updated soon.	Christiansen, 2008
CAAT GIVReST – Good In Vitro Reporting Standards initiative	Ongoing project involving about 90 experts aiming to develop *in vitro* reporting standards.	Leist et al., 2010; this article
Common Data Elements[b]	Common data elements are standardized terms for the collection and exchange of data. This portal provides access to NIH-supported CDE initiatives and other resources for investigators developing data collection protocols.	
CoBRA – Citation of bioresources in journal articles[c]	Developed by members of the journal editors' subgroup of the Bioresource Research Impact Factor (BRIF) for citing bioresources, such as biological samples, data, and databases.	Bravo et al., 2015
EQUATOR – Enhancing the quality and transparency of health research[d]	Seeks to improve the quality of research (includes additional resources and links to other reporting guidelines).	
EASE – Guidelines for authors and translators of scientific articles to be published in English[e]	Seeks quality reporting of all scientific literature.	
ENTREQ – Enhancing transparency in reporting the synthesis of qualitative research[f]	Provides a framework for reporting the synthesis of qualitative health research fostered by various universities.	
FAIRsharing[g] (formerly Biosharing)	A curated, informative, and educational resource on data and metadata standards, inter-related to databases and data policies.	
ICMJE – International Committee of Medical Journal Editors[h] (formerly known as the URM)	Uniform requirements for manuscripts submitted to biomedical journals (also called the Vancouver style)	

Organization / Guideline	Description	References
International Congress on Peer Review and Biomedical Publication[i]	Aims to improve the quality and credibility of scientific peer review and publication and to help advance the efficiency, effectiveness, and equitability of the dissemination of biomedical information.	Rennie and Flanagin, 2018
ISA-Tab – Mayfield Handbook Investigation/Study/Assay (ISA) tab-delimited (TAB) format[j]	A general-purpose framework with which to collect and communicate complex metadata (i.e., sample characteristics, technologies used, type of measurements made) from omics-based experiments employing a combination of technologies.	
MIAME – Minimum information about a microarray experiment[k]	Describes the basic data needed to enable the unambiguous interpretation of the results and to possibly replicate the experiment.	Brazma et al., 2001
MIBBI – Minimum information for biological and biomedical investigations[l]	Portal of almost 40 checklists, which can be used when reporting biological and biomedical science research.	
OECD harmonised templates (OHTs) to report test results[m]	Stakeholder-endorsed OHT 201 for reporting on "intermediate effects" being observed via *in vitro* assays and possibly other non-animal test methods (computational predictions, etc.).	
OECD GD 211 for describing non-guideline *in vitro* test methods[n]	Guidance document initiated by the Advisory Group on Molecular Screening and Toxicogenomics (EAGMST), operating under the Working Group of the National Coordinators of the Test Guidelines Programme (WNT) at the OECD.	
OECD GD on Good In Vitro Method Practices (GIVIMP)[o]	Guidance document aiming to reduce the uncertainties in cell and tissue-based *in vitro* method derived predictions by applying all necessary good scientific, technical and quality practices from *in vitro* method development to *in vitro* method implementation for regulatory use.	Eskes et al., 2017
PLOS editorial and publishing policies: Reporting guidelines for specific study designs[p]	PLOS requires that authors comply with field-specific standards for preparation and recording of data and select repositories appropriate to their field.	
Principles and guidelines for reporting preclinical research[q], National Institutes of Health (NIH)	Set of principles to enhance rigor and further support research that is reproducible, robust, and transparent, which a number of journals have agreed to endorse.	Landis et al., 2012; Collins and Tabak, 2014; Clayton and Collins, 2014; Lorsch et al., 2014
SHERPA[r] (Securing a hybrid environment for research preservation and access)	Develops open-access institutional repositories in universities and provides compliance-checking tools against publication bias and against excessive corporate influence on scientific integrity.	
SRQR[s] – Standards for reporting qualitative research: a synthesis of recommendations	Guidance how to report qualitative research.	O'Brien et al., 2014
Structured Abstracts[t], National Library of Medicine (NLM)	Description of structured abstracts and how they are formatted for MEDLINE.	Nakayama et al., 2005
TOP – Transparency and openness promotion (TOP), Open Science Framework[u]	The guidelines cover eight transparency standards: citation, data transparency, analytic methods transparency, research materials transparency, design and analysis transparency, preregistration of studies, preregistration of analysis plans, replication	Foster and Deardorff, 2017
TRANSPOSE – Transparency in scholarly publishing for open scholarship evolution[v]	Database of journal policies focusing on open peer review, co-reviewing, and detailed preprinting policies.	
WAME[w]: World Association of Medical Editors	Association of editors of medical journals who seek to foster cooperation among and education of medical journal editors.	

[a]https://bit.ly/1F0K2O6; [b]https://www.nlm.nih.gov/cde/index.html; [c]http://www.equator-network.org/wp-content/uploads/2015/03/Cobra-check-list.pdf; [d]http://www.equator-network.org; [e]www.ease.org.uk/publications/author-guidelines-authors-and-translators/; [f]www.equator-network.org/reporting-guidelines/entreq/; [g]https://fairsharing.org; [h]http://www.icmje.org; [i]https://peerreviewcongress.org/index.html; [j]http://www.dcc.ac.uk/resources/metadata-standards/isa-tab; [k]http://fged.org/projects/miame/; [l]https://bit.ly/1njFpnb; [m]http://www.oecd.org/ehs/templates/harmonised-templates-intermediate-effects.htm; [n]https://bit.ly/2k8PxTf; [o]https://bit.ly/2KwBtOh; [p]https://journals.plos.org/plosone/s/best-practices-in-research-reporting; [q]https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research; [r]http://www.sherpa.ac.uk; [s]http://www.equator-network.org/reporting-guidelines/srqr/; [t]https://www.nlm.nih.gov/bsd/policy/structured_abstracts.html; [u]https://cos.io/our-services/top-guidelines/; [v]https://transpose-publishing.github.io; [w]http://www.wame.org/index.php

10 The CAAT In Vitro Reporting Standards (GIVReSt) initiative

Leist et al. (2010) published considerations on test descriptions in toxicology earlier in this series of articles (Hartung, 2017). These were later updated for some aspects (Leist et al., 2012; Schmidt et al., 2017). This work focused on general considerations of (i) which elements are required to fully define a test method, (ii) what are the common mistakes, and (iii) how can they be avoided. In parallel, a broader activity was started to comprehensively assemble reporting standards. Experts, coordinated by CAAT, assembled suggestions for such standards (Tab. 3), which were then presented in a satellite workshop of a SOT meeting to 60 scientists. They gave further input to the standards, so that a well-balanced collection was established and can soon be published. It is important to note here that the focus of this initiative is different from many others as compiled in Table 2. The first priority of reporting standards, as defined by various organizations, is to ensure the completeness and transparency of information given (data focus). This involves many technical aspects. For instance, for microarray or sequencing data such standards assure that all information can be found, is interoperable, and can be linked to the relevant metadata. This applies not only to experimental data from NAM, but also to *in silico* data and the description of models used to derive them. Also, e.g., for models, the standards mainly define parameters that need to be described and all assumptions used for building the model to be listed. All these initiatives neglect to some extent a second level of required information: The "data display standards" of this tier are intended to ensure reporting in a form that makes information digestible and quick to grasp, compare, and further analyze. Thus, this level focuses less on the data itself, but more on the information that can be extracted from the data.

It is sometimes argued that this has mainly esthetic aspects, and that in theory any data display can be re-constructed, if only the primary data and their meta data connection is known. In practice, the situation is different. This becomes most obvious in the field of big data. The information contained therein is often only made accessible by the type of display chosen – and depending on this choice, the message is transferred well or not at all. In these fields, the type of data display has become an own research topic, and boundaries between data analysis and display are vanishing, e.g., in the computer sciences sub-field of immersive analytics (human user-guided data exploration in virtual visual environments). When dealing with the exploration of big data sets, the aim is often general knowledge or generally applicable rules. However, it is frequently neglected that such information can only be understood by many end-users in the form of examples. Thus, such exemplifications are an important feature of good presentation standards. This also applies to *in vitro* technologies and computational models. Example cases should be presented in parallel with the more general information outcome. They provide the most immediate tool for the recipient to judge the performance of a method or the implications of a data set.

An important example for presentation issues of classical low-throughput experimental NAM is the visualization of the experimental conditions and the exposure scheme used to generate the data. If this is not done, important information is often lost, and the interpretation and understanding of the data is drastically slowed or misguided.

These few examples illustrate that there is an urgent need for reporting standards, focusing on the transfer of information, even when we reach a state where at least the data as such are made available in a findable, accessible, interoperable, and re-usable (FAIR) way (Corpas et al., 2018).

Tab. 3: Chapter structure and responsibilities of CAAT GIVReST

Chapter no.	Chapter structure	Responsibilities
1	General considerations concerning the "message" of a publication	James McKim, Francesca Caloni, Bas Blaauboer, James Yager
2	Description of materials	Rodger Curren, Tuula Heinonen, Erwin Roggen
3	Cells and tissues	Joshua Harrill, Robert Chapin, Andras Dinnyes, Gerhard Gstraunthaler
4	Description of methods for assessing endpoints of interest	Tim Shafer, Nina Hasiwa, Pamela J. Lein
5	*In silico* methods / data processing	Kevin Crofton, Bob Burrier, Andre Kleensang, Alan Smith
6	Statistical considerations	Sebastian Hoffmann, Daniel Dietrich, Christoph van Thriel
7	General data presentation	Marcel Leist, Mardas Daneshian, Jan Hengstler
8	Presentation of test features and characteristics	Matteo Goldoni, Rosella Alinovi, Marta Barenys, Ellen Fritsche, Silvana Pinelli, Sara Tagliaferri, Olavi Pelkonen
9	Special considerations concerning response dynamics beyond general presentations	

11 Discussion

This article makes the case for the need for *in vitro* reporting standards, tentatively named Good In Vitro Reporting Standards (GIVReSt). It summarizes, probably not even comprehensively, the activities known to and/or involving the authors. These activities are closely linked to Good Cell Culture Practice (GCCP), and thus already span 22 years. Arguably, already that guidance includes the substance of what is needed to document *in vitro* work. The lack of application by authors, peer-reviewers, and publishers, however, shows that such guidance has to be delivered on a silver platter and that former dissemination activities were simply not sufficient.

Most progress with respect to quality advances has been in the regulatory field as Good Laboratory Practice is increasingly applied to *in vitro* methods. Sure, the severity of regulatory decisions and the need to mutually accept test results between countries are strong driving forces. However, the reproducibility crisis (Baker, 2016; Jarvis and Williams, 2016) in science alone should prompt and cement such quality assurance. Along with our mantra "*the most important omics is economics*" (Meigs et al., 2018), this has also to be seen as an economic problem. Freedman et al. (2015) have estimated how this translates to wasted money: "*the cumulative (total) prevalence of irreproducible preclinical research exceeds 50%, resulting in approximately US$28,000,000,000 (US $28B)/year spent on preclinical research that is not reproducible – in the United States alone*". It is incredible how much research is apparently done and reported in a way that is not reproducible. This can only be characterized as students and scientists toying around and is not only a waste of resources: Wrong results put forward mislead other researchers and the society as a whole. Martin Van Buren is quoted as saying "*It is easier to do a job right than to explain why you didn't*". In this sense, we should embrace a culture of quality in how science is done and how it is reported. Henry Ford's "*Quality means doing it right when no one is looking*" describes such a culture of quality. However, at this stage, we have to look and speak out loudly about the quality deficits in the life sciences and the way they are being reported.

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature 533*, 452-454. doi:10.1038/533452a

Bal-Price, A., Hogberg, H., Crofton, K. et al. (2018). Recommendation on test readiness criteria for new approach methods in toxicology: Exemplified for developmental neurotoxicity. *ALTEX 35*, 306-352. doi:10.14573/altex.1712081

Bravo, E., Calzolari, A., De Castro, P. et al. (2015). Developing a guideline to standardize the citation of bioresources in journal articles (CoBRA). *BMC Med 13*, 33. doi:10.1186/s12916-015-0266-y

Brazma, A., Hingamp, P., Quackenbush, J. et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet 29*, 365-371. doi:10.1038/ng1201-365

Christiansen, S. L. (2008). Ethical and legal guidance in biomedical publishing: The AMA manual of style, tenth edition. *Chest 134*, 1344-1346. doi:10.1378/chest.08-1165

Clayton, J. A. and Collins, F. S. (2014). Policy: NIH to balance sex in cell and animal studies. *Nature 509*, 282-283. doi:10.1038/509282a

Coecke, S., Balls, M., Bowe, G. et al. (2005). Guidance on good cell culture practice. *Altern Lab Anim 33*, 261-287.

Cohen, A. M., Hersh, W. R., Peterson, K. and Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc 13*, 206-219. doi:10.1197/jamia.M1929

Collins, F. S. and Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature 505*, 612-613. doi:10.1038/505612a

Corpas, M., Kovalevskaya, N. V., McMurray, A. and Nielsen, F. G. (2018). FAIR guide for data providers to maximise sharing of human genomic data. *PLoS Comput Biol 14*, e1005873. doi:10.1371/journal.pcbi.1005873

Daneshian, M., Kamp, H., Hengstler, J. et al. (2016). Highlight report: Launch of a large integrated European in vitro toxicology project: EU-ToxRisk. *Arch Toxicol 90*, 1021-1024. doi:10.1007/s00204-016-1698-7

de Vries, R. B.., Wever, K. E., Avey, M. T. et al. (2014). The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *ILAR J 55*, 427-437. doi:10.1093/ilar/ilu043

de Vries, R. and P. Whaley. (2018). In vitro critical appraisal tool (IV-CAT): Tool development protocol. *Zenodo*. doi:10.5281/zenodo.1493498#.XAQ4GpTA4A8.mendeley

Eskes, C., Boström, A.-C., Bowe, G. et al. (2017). Good cell culture practices & in vitro toxicology. *Toxicol In Vitro 45*, 272-277. doi:10.1016/j.tiv.2017.04.022

Foster, E. D. and Deardorff, A. (2017). Open science framework (OSF). *J Med Libr Assoc 105*, 203-206. doi:10.5195/JMLA.2017.88

Freedman, L. P., Cockburn, I. M. and Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS Biol 13*, e1002165. doi:10.1371/journal.pbio.1002165

Golbach, L. A., Portelli, L. A., Savelkoul, H. F. et al. (2016). Calcium homeostasis and low-frequency magnetic and electric field exposure: A systematic review and meta-analysis of in vitro studies. *Environ Int 92-93*, 695-706. doi:10.1016/j.envint.2016.01.014

Griesinger, C., Hoffmann, S., Kinsner-Ovaskainen, A. et al. (2009). Proceedings of the First International Forum Towards Evidence-Based Toxicology. Conference Centre Spazio Villa Erba, Como, Italy. 15-18 October 2007. *Human Exp Toxicol, Spec Issue: Evidence-Based Toxicology (EBT) 28*, 83-163.

Gstraunthaler, G. and Hartung, T. (1999). Bologna declaration toward good cell culture practice. *Altern Lab Anim 27*, 206.

Han, S., Olonisakin, T. F., Pribis, J. P. et al. (2017). A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review. *PLoS One 12*, e0183591. doi:10.1371/journal.pone.0183591

Hardy, B., Apic, G., Carthew, P. et al. (2012a). A toxicology ontology roadmap. *ALTEX 29*, 129-137. doi:10.14573/altex.2012.2.129

Hardy, B., Apic, G., Carthew, P. et al. (2012b). Toxicology ontology perspectives. *ALTEX 29*, 139-156. doi:10.14573/altex.2012.2.139

Hartung, T., Balls, M., Bardouille, C. et al. (2002). Report of ECVAM task force on good cell culture practice (GCCP). *Altern Lab Anim 30*, 407-414.

Hartung, T., Bremer, S., Casati, S. et al. (2004). A modular approach to the ECVAM principles on test validity. *Altern Lab Anim 32*, 467-472.

Hartung, T. (2007). Food for thought … on validation. *ALTEX 24*, 67-72. doi:10.14573/altex.2007.2.67

Hartung, T. (2009). Food for thought … on evidence-based toxicology. *ALTEX 26*, 75-82. doi:10.14573/altex.2009.2.75

Hartung, T., Stephens, M. and Hoffmann, S. (2013). Mechanistic validation. *ALTEX 30*, 119-130. doi:10.14573/altex.2013.2.119

Hartung, T. (2017). Food for Thought ... The first ten years. ALTEX 34, 187-192. doi:10.14573/altex.1703311

Hartung, T. (2018). Making big sense from big data. *Frontiers In Big Data 1*, 5. doi:10.3389/fdata.2018.00005

Hoffmann, S. and Hartung, T. (2006). Towards an evidence-based toxicology. *Human Exp Toxicol 25*, 497-513. doi:10.1191/0960327106het648oa

Hoffmann, S., Edler, L., Gardner, I. et al. (2008). Points of reference in validation – The report and recommendations of ECVAM Workshop. *Altern Lab Anim 36*, 343-352.

Hoffmann, S., Hartung, T. and Stephens, M. (2016). Evidence-based toxicology. *Adv Exp Med Biol 856*, 231-241. doi:10.1007/978-3-319-33826-2_9

Hoffmann, S., de Vries, R. B., Stephens, M. L. et al. (2017). A primer on systematic reviews in toxicology. *Arch Toxicol 91*, 2551-2575. doi:10.1007/s00204-017-1980-3

Hooijmans, C. R., Rovers, M. M., de Vries, R. B. et al. (2014). SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol 414*, 43. doi:10.1186/1471-2288-14-43

Howard, B. E., Phillips, J., Miller, K. et al. (2016). SWIFT-Review: A text-mining workbench for systematic review. *Syst Rev 5*, 87. doi:10.1186/s13643-016-0263-z

Jarvis, M. F. and Williams, M. (2016). Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. *Trends Pharmacol Sci 37*, 290-302. doi:10.1016/j.tips.2015.12.001

Krzywinski, M. and Altman, N. (2014). Visualizing samples with box plots. *Nat Methods 11*, 119-120. doi:10.1038/nmeth.2813

Landis, S. C., Amara, S. G., Asadullah, K. et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature 490*, 187-191. doi:10.1038/nature11556

Leist, M., Efremova, L. and Karreman, C. (2010). Food for thought … Considerations and guidelines for basic test method descriptions in toxicology. *ALTEX 27*, 309-317. doi:10.14573/altex.2010.4.309

Leist, M., Hasiwa, M., Daneshian, M. and Hartung, T. (2012). Validation and quality control of replacement alternatives – Current status and future challenges. *Toxicol Res 1*, 8. doi:10.1039/c2tx20011b

Lorsch, J. R., Collins, F. S. and Lippincott-Schwartz, J. (2014). Fixing problems with cell lines. *Science 346*, 1452-1453. doi:10.1126/science.1259110

Lovell, D. P. and Omori, T. (2008). Statistical issues in the use of the comet assay. *Mutagenesis 23*, 171-178. doi:10.1093/mutage/gen015

Macleod, M. R. and the NPQIP Collaborative group (2017). Findings of a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution. *bioRxiv*, 187245. doi:10.1101/187245

Marx, U., Andersson, T. B., Bahinski, A. et al. (2016). Biology-inspired microphysiological system approaches to solve the prediction dilemma of substance testing using animals. *ALTEX 33*, 272-321. doi:10.14573/altex.1603161

Matwin, S., Kouznetsov, A., Inkpen, D. et al. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *J Am Med Inform Assoc 17*, 446-453. doi:10.1136/jamia.2010.004325

Meigs, L., Smirnova, L., Rovida, C. et al. (2018). Animal testing and its alternatives – The most important omics is economics. *ALTEX 35*, 275-305. doi:10.14573/altex.1807041

Nakayama, T., Hirai, N., Yamazaki, S. and Naito, M. (2005). Adoption of structured abstracts by general medical journals and format for a structured abstract. *J Med Libr Assoc 93*, 237-242.

O'Brien, B. C., Harris, I. B., Beckman, T. J. et al. (2014). Standards for reporting qualitative research: A synthesis of recommendations. *Acad Med. 89*, 1245-1251. doi:10.1097/ACM.0000000000000388

OECD (2005). Guidance Document on the Validation and, International Acceptance of New or Updated Test Methods for Hazard Assessment. *OECD Series on Testing and Assessment Number 34*. OECD Publishing, Paris.

OECD (2014). Guidance Document for Describing Non-Guideline In Vitro Test Methods. *OECD Series on Testing and Assessment No. 211*, OECD Publishing, Paris. doi:10.1787/9789264274730-en

Pamies, D., Bal-Price, A., Simeonov, A. et al. (2017a). Good cell culture practice for stem cells and stem-cell-derived models. *ALTEX 34*, 95-132. doi:10.14573/altex.1607121

Pamies, D., Barreras, P., Block, K. et al. (2017b). A human brain microphysiological system derived from iPSC to study central nervous system toxicity and disease. *ALTEX 34*, 362-376. doi:10.14573/altex.1609122

Pamies, D. and Hartung, T. (2017). 21st century cell culture for 21st century toxicology. *Chem Res Toxicol 30*, 43-52. doi:10.1021/acs.chemrestox.6b00269

Pamies, D., Bal-Price, A., Chesne, C. et al. (2018). Advanced good cell culture practice for human primary, stem cell-derived and organoid models as well as microphysiological systems. *ALTEX 35*, 353-378. doi:10.14573/altex.1710081

Rennie, D. and Flanagin, A. (2018). Three decades of peer review congresses. *JAMA 319*, 350-353. doi:10.1001/jama.2017.20606

Samuel, G. O., Hoffmann, S., Wright, R. et al. (2016). Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: A scoping review. *Environ Int 92-93*, 630-646. doi:10.1016/j.envint.2016.03.010

Schmidt, B. Z., Lehmann, M., Gutbier, S. et al. (2017). In vitro acute and developmental neurotoxicity screening: An overview of cellular platforms and high-throughput technical possibilities. *Arch Toxicol 34*, 1-33. doi:10.1007/s00204-016-1805-9

Smirnova, L., Kleinstreuer, N., Corvi, R. et al. (2018). 3S – Systematic, systemic, and systems biology and toxicology. *ALTEX 35*, 139-162. doi:10.14573/altex.1804051

Stephens, M. L., Andersen, M., Becker R. A. et al. (2013). Evidence-based toxicology for the 21st century: Opportunities and challenges. *ALTEX 30*, 74-104. doi:10.14573/altex.2013.1.074

Stephens, M. L., Betts, K., Beck, N. B. et al. (2016). The emergence of systematic review in toxicology. *Toxicol Sci 152*, 10-16. doi:10.1093/toxsci/kfw059

Whetzel, P. L., Noy, N. F., Shah, N. H. et al. (2011). BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res 39*, W541-545. doi:10.1093/nar/gkr469

Whaley, P., Letcher, R. J., Covaci, A. and Alcock, R. (2016). Raising the standard of systematic reviews published in Environment International. *Environ Int 97*, 274-276. doi:10.1016/j.envint.2016.08.007

Whaley, P., Watford S., Allard P. et al. (in preparation). Enhancing systematic review and evidence mapping with shared ontologies and semantic matching: An example from epigenetic research.

Conflict of interest

Thomas Hartung is the founder of Organome LLC, which aims to commercialize the mini-brains mentioned.

Acknowledgements

The enormous effort by Dr Mardas Daneshian in organizing the earlier work by CAAT toward reporting standards is gratefully appreciated. The colleagues who wrote the draft document and who participated in the first consensus workshop on the quality of publications dealing with *in vitro* systems, San Francisco, 2012 as well as those listed in Table 3 are thanked for their respective input into these discussions.