Bench Marks!

Template for the Description of Cell-Based Toxicological Test Methods to Allow Evaluation and Regulatory Use of the Data

Alice Krebs[1,2], Tanja Waldmann[1], Martin F. Wilks[3], Barbara M. A. van Vugt-Lussenburg[4], Bart van der Burg[4], Andrea Terron[5], Thomas Steger-Hartmann[6], Joelle Ruegg[7], Costanza Rovida[8], Emma Pedersen[9], Giorgia Pallocca[1,8], Mirjam Luijten[10], Sofia B. Leite[11], Stefan Kustermann[12], Hennicke Kamp[14], Julia Hoeng[14], Philip Hewitt[15], Matthias Herzler[16], Jan G. Hengstler[17], Tuula Heinonen[18], Thomas Hartung[8,19], Barry Hardy[20], Florian Gantner[21], Ellen Fritsche[22], Kristina Fant[9], Janine Ezendam[10], Thomas Exner[20], Torsten Dunkern[23], Daniel R. Dietrich[24], Sandra Coecke[11], Francois Busquet[8,25], Albert Braeuning[26], Olesja Bondarenko[27], Susanne H. Bennekou[28], Mario Beilmann[29] and Marcel Leist[1,2,8]

[1]In vitro Toxicology and Biomedicine, Dept inaugurated by the Doerenkamp-Zbinden Foundation, University of Konstanz, Konstanz, Germany; [2]Konstanz Research School Chemical Biology (KoRS-CB), University of Konstanz, Konstanz, Germany; [3]Swiss Centre for Applied Human Toxicology, University of Basel, Basel, Switzerland; [4]BioDetection Systems BV, Amsterdam, The Netherlands; [5]European Food Safety Authority, Parma, Italy; [6]Investigational Toxicology, Drug Discovery, Pharmaceuticals, Bayer AG, Wuppertal, Germany; [7]Department of Organismal Biology, Uppsala University, Uppsala, Sweden; [8]CAAT-Europe, University of Konstanz, Konstanz, Germany; [9]RISE Research Institutes of Sweden, Göteborg, Sweden; [10]Centre for Health Protection, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands; [11]European Commission, Joint Research Centre (JRC), Ispra, Italy; [12]F. Hoffmann – La Roche, Pharma Research and Early Development, Pharmaceutical Sciences – Roche Innovation Center, Basel, Switzerland; [13]Experimental Toxicology and Ecology, BASF SE, Ludwigshafen, Germany; [14]Philip Morris International R&D, Neuchâtel, Switzerland; [15]Non Clinical Safety, Merck KGaA, Darmstadt, Germany; [16]German Federal Institute for Risk Assessment, Dept. Chemical Safety, Berlin, Germany; [17]Leibniz Research Centre for Working Environment and Human Factors (IfADo), Technical University of Dortmund, Dortmund, Germany; [18]FICAM, Faculty of Medicine and Life Sciences, Tampere University, Tampere, Finland; [19]Johns Hopkins University, Center for Alternatives to Animal Testing (CAAT), Baltimore, MD, USA; [20]Edelweiss Connect GmbH, Technology Park Basel, Basel, Switzerland; [21]Translational Medicine & Clinical Pharmacology, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany; [22]IUF – Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany; [23]Grünenthal GmbH, Aachen, Germany; [24]Human and Environmental Toxicology, University of Konstanz, Konstanz, Germany; [25]ALTERTOX SPRL, Ixelles, Bruxelles, Belgium; [26]German Federal Institute for Risk Assessment, Dept. Food Safety, Berlin, Germany; [27]Laboratory of Environmental Toxicology, National Institute of Chemical Physics and Biophysics, Tallinn, Estonia; [28]The National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark; [29]Boehringer Ingelheim Pharma GmbH & Co. KG, Nonclinical Drug Safety, Biberach, Germany

Abstract

Only few cell-based test methods are described by Organisation for Economic Co-operation and Development (OECD) test guidelines or other regulatory references (e.g., the European Pharmacopoeia). The majority of toxicity tests still falls into the category of non-guideline methods. Data from these tests may nevertheless be used to support regulatory decisions or to guide strategies to assess compounds (e.g., drugs, agrochemicals) during research and development if they fulfill basic requirements concerning their relevance, reproducibility and predictivity. Only a method description of sufficient clarity and detail allows interpretation and use of the data. To guide regulators faced with increasing amounts of data from non-guideline studies, the OECD formulated Guidance Document 211 (GD211) on method documentation

Disclaimer: The opinions expressed in this document strictly represent those of the authors and do not (necessarily) represent official views of the institutions they are affiliated with.

Received September 27, 2019;

ALTEX 36(4), 682-699. doi:10.14573/altex.1909271

Correspondence: Marcel Leist, PhD
In vitro Toxicology and Biomedicine, Dept inaugurated by the Doerenkamp-Zbinden foundation at the University of Konstanz, Universitaetsstr. 10, 78464 Konstanz, Germany
(marcel.leist@uni-konstanz.de)

for the purpose of safety assessment. As GD211 is targeted mainly at regulators, it leaves scientists less familiar with regulation uncertain as to what level of detail is required and how individual questions should be answered. Moreover, little attention was given to the description of the test system (i.e., cell culture) and the steps leading to it being established in the guidance. To address these issues, an annotated toxicity test method template (ToxTemp) was developed (i) to fulfill all requirements of GD211, (ii) to guide the user concerning the types of answers and detail of information required, (iii) to include acceptance criteria for test elements, and (iv) to define the cells sufficiently and transparently. The fully annotated ToxTemp is provided here, together with reference to a database containing exemplary descriptions of more than 20 cell-based tests.

1 Main test elements

Test, assay, test system, test method… All these terms are found in the literature and in discussions, but they need some definition to allow for specification of their background and requirements. "Test" is the shortest term and thus a good place to start: a toxicological "test" is a procedure to determine, in a quantifiable manner (with respect to damage and dose/concentration), whether a substance may harm/incapacitate an organism, a cell, or an essential component thereof. The terms "assay" and "test method" are used interchangeably with "test".

A test has various elements that are independent of one another to a large degree (Schmidt et al., 2017). The five main elements are: the test purpose (see Section 4), the test system, the exposure scheme, the endpoint, and the prediction model. Thus, the "test system" is one element of the "test method" and must not be confused with it. As this often causes confusion among non-specialists, it deserves some further explanation. The *in vivo* test method for acute toxicity assessment is a good example to explain the problem. The overall test method is defined by a test guideline, e.g., OECD TG 423 (acute oral toxicity) (OECD, 2002). The test system is, e.g., mouse, fasted for > 3 h prior to dosing; the exposure scheme involves single dosing by gavage and continued observation for 14 days; the endpoint is the number/percentage of dead animals; and the prediction model converts the test data into toxicity classes defined by the United Nation's Globally Harmonized System (GHS), e.g., category 2 (comprising compounds with an LD50 in the range of 5-50 mg/kg bodyweight). Clearly, the mouse is the test system and not the test method. Each element of the test method may be modified independently, e.g., changing from oral to dermal exposure; using time to death as endpoint; employing a binary prediction model (toxic/non-toxic) and using rats instead of mice. Any such change results in a new, different test method that may no longer be in accordance with the guideline method.

The same principles apply to cell-based assays, i.e., *in vitro* methods or new approach methods (NAM). For example, the test system may be neurons in 2D culture or neural organoids or liver cells; the exposure scheme may be 24 h exposure of cells cultured in standard medium plus test chemical or 72 h exposure in a special medium, with or without re-addition of test chemical every 24 h; the prediction model may be binary (toxic/non-toxic) or it may use a composite measure of several endpoints to define a continuous malformation index. Any combination of these independent elements results in a different overall test method.

2 Compliance issues with OECD Guidance Document 211 (GD211)

The key literature on how to describe non-guideline methods is OECD Guidance Document 211 (GD211) (OECD, 2017). In line with reports by many others (Freedman et al., 2015; Vogt et al., 2016; Hair et al., 2019), our own survey of the scientific literature has indicated that method descriptions still show an enormous heterogeneity in quality, detail and scope. Moreover, the clarity of the information, as well as the format that is used to provide the information, varies widely. This makes the extraction of necessary information as well as the use and interpretation of data produced using the method difficult. We further found that the items required to be addressed in GD211 are understood and interpreted in different ways by the users, or sometimes not understood at all.

Part of this might be attributable to the fact that GD211 looks at non-guideline methods from a regulatory perspective, i.e., it explicitly provides a format for reporting non-guideline methodology so that it can be used by regulatory toxicologists for the safety assessment of chemicals. It asks for information that is of specific value for regulators (e.g., information on validation, predictivity, standardization, applicability domain, etc.), but does not provide detailed background information on why regulators need such information and what they use it for. Experience has shown that the target audience, which includes method developers in academia or fundamental research, is not necessarily familiar with the underlying regulatory background.

Another reason for apparent non-compliance is that GD211 is a brief and highly condensed document. It often covers several distinct features of tests in a single question. Test developers may not consider all the different aspects of such a complex question without more specific guidance, and thus some issues may be missed entirely.

From the information recipient's point-of-view, non-compliance is not the only problem. Also, finding and retrieving the information within a report prepared in a not fully-standardized format can be a difficult task. For instance, it may be a time-consuming task to find out whether some information is absent or whether it is only placed or mentioned in another context than usual.

Abbreviations
AC, acceptance criteria; GD, guidance document; OECD, Organisation for Economic Co-operation and Development; SOP, standard operating procedure; TG, test guideline; ToxTemp, toxicological test methods template

3 Annotations and guiding questions for a test method template

The problems that test developers face when trying to comply with the questions raised in GD211 may best be illustrated using an analogy from an entirely different field. Let us assume that a police recruit is asked in a questionnaire to "indicate body measures". Some would answer by giving only their height and weight. Others may include their shoe size and special measures for jackets and trousers. Few would give their head circumference (required for the uniform hat) and/or their glove size. Possibly even fewer would think to provide information on their sight, i.e., data essential for ordering sunglasses (e.g., whether they are myopic, distance between the eyes, etc.).

In the course of the European research project EU-ToxRisk (Daneshian et al., 2016), compliance of project case study results with GD211 became an important issue. While trying to implement high-quality assay documentation within the project, we realized that hardly any of the senior scientists and none of the junior faculty fully understood the requirements of GD211.

How are situations as exemplified above (police recruit or EU-ToxRisk examples) best avoided? How can it be ensured that all required information (i) is reported, (ii) is presented in a structured way so that the recipients can verify its completeness, and (iii) can be found easily? Two major strategies towards this goal are (a) to sub-divide complex questions into sets of more simple questions, each dealing with a single, defined issue, and (b) to explain the questions by adding additional notes, comments and guiding questions. These measures help to structure the answer and ensure that all relevant aspects are considered.

Such guidance is provided by the test method template (ToxTemp) shown in a compact version as Box 1 and provided as a printable version including additional notes and examples in the supplementary file[1]. Tab. 1 gives a synoptic overview of all items/questions of GD211 and of the respective counterparts in the ToxTemp (see also the supplementary file[1] for the detailed comparison). Moreover, an online methods database, based on the ToxTemp, is under construction[2]. Using Tab. 1 should enable anyone accustomed to the structure of GD211 to retrieve all relevant test method information from the database.

4 Understanding the test purpose

Any test (toxicological or not) is developed to probe a test hypothesis (e.g., whether a substance is toxic or not). The test design will always reflect that purpose and test parameters will ideally be optimized in order to achieve maximum certainty about whether the hypothesis should be accepted or rejected.

It is a basic scientific principle that test results should – within limits – only be used for the purpose they were designed for. This is not trivial for *in vitro* or new approach methods (NAM) developed for regulatory purposes. In this domain, the element "test purpose" plays a special role: Beyond the primary test purpose (e.g., determination of cytotoxicity), the results of a test may also be used for a secondary, regulatory purpose (regulation), and even a tertiary purpose (e.g., modelling a potential hazard in the population). Consider, e.g., *in vitro* tests used to predict Globally Harmonized System (GHS) classifications (e.g., moderate or strong skin sensitizer or eye irritant) (secondary purpose), which in turn are used to model the potential chemical hazard to which workers or the general population may be exposed (tertiary purpose).

Test developers might not be familiar with the limitations and requirements demanded of their test when it is to be used for secondary (or tertiary) purposes, and this may in the end lead to misinterpretation of results obtained from a test method used for purposes that it was not originally designed for. The problem is further complicated by the fact that an apparently simple regulatory statement (e.g., "The substance is a skin sensitizer.") in reality represents the outcome of a highly complex decision system based on a regulatory framework that has evolved over several decades, with paradigms and implicit assumptions that are often not obvious to the outsider.

To avoid such problems, communication between developers and regulatory recipients needs to be as transparent, comprehensive and precise as possible. Description of a test method, especially of the test purpose, in a way that allows such communication is essential to achieve this.

5 Distinctions between a test method description and the overall testing process

Reproducible toxicological research necessitates the comprehensive description of the testing process. Good information on this can be found in the series of guidance on Good Laboratory Practice (GLP) (e.g., OECD, 2005) or in the OECD Guidance Document on Good *In Vitro* Methods Practice (GIVIMP) (OECD, 2018). Several helpful tools are available, e.g., from the Science in Risk Assessment and Policy (SciRAP[3]) web resource, the DB-ALM methods summary (adapted from GD211), the EURL-ECVAM test submission template (used for structuring information for test validation), or the ALTEX BenchMarks series (Kisitu et al., 2019; Krebs et al., 2018).

Several earlier EU-funded projects devoted considerable resources to harmonize test method descriptions (Kinsner-Ovaskainen et al., 2009; Rovida et al., 2014), and similar activities are taking place in the USA (Flood et al., 2017). The overall reporting of *in vitro* experiments (data and methods) has been addressed by an NC3Rs initiative (RIVER) (Prior et al., 2019), large stakeholder workshops organized by CAAT-Europe (Hartung et al., 2019), and on the regulatory level (OECD harmonized templates for data reporting (OHT); OHT 201- Intermediate effects[4] is especially relevant).

[1] doi:10.14573/altex.1909271s

[2] https://eu-toxrisk.douglasconnect.com/public/

[3] http://scirap.org/Page/Index/aa44f63a-ce5d-4f26-bac3-346c27b34eb0/reporting-checklist

[4] http://www.oecd.org/ehs/templates/harmonised-templates-intermediate-effects.htm

Box. 1: Documentation of a test method and its readiness status, guided by a test method questionnaire
A print version of the complete test method questionnaire including notes and examples can be found in the supplementary file[1].

1. Overview

1.1 Descriptive full-text title
Provide a descriptive title using normal language without technical terms or acronyms.

1.2 Abstract
Please describe in no more than 200 words the following:
Which toxicological target (organ, tissue, physiological/biochemical function, etc.) is modelled? (8.1)
Which test system and readout(s) are used? (4.1; 5.2)
Which biological process(es) (e.g. neurite outgrowth, differentiation) and/or toxicological events (e.g. oxidative stress, cell death) are modelled/reflected by your test method? (8.1)
To which (human) adverse outcome(s) is your test method related or could be related? (8.1; 9.2; 9.3)
Which hazard(s) do(es) your test method (potentially) predict? (8.1; 8.6)
Does the test method capture an endpoint of current regulatory studies? (9.5)
If the method has undergone some form of validation/evaluation, give its status. (9.4)

2. General information

2.1 Name of test method
Provide the original/published name, as well as the potential tradename.

2.2 Version number and date of deposition
Provide the original deposition date of first version and date of current version.

2.3 Summary of introduced changes in comparison to previous version(s)
This only applies to updated versions. If this is the original version, state "original version".

2.4 Assigned data base name
Normal text names often do not uniquely define the method. Therefore, each method should be assigned a clearly and uniquely defined data base name.
These are some example data base names generated in the EU-ToxRisk project:
UKN1a_DART_NPC_Diff_6D_02
UKN1b_DART_NPC_Diff_4D_01
UKN2a_DART_NC_Migr_24h_04
The name is assembled (in more generic terms) from the following elements:
Axa_B_C_D_E
Axa: mandatory part of the identifier allowing unambiguous identification
A: Abbreviation/acronym of the partner depositing the assay
x: Consecutive number (referring to the partner's assay number)
a: Sub-specifier (for variants, i.e. very similar assays but e.g. different readout or medium); not mandatory, but 'Axa' must be specific (i.e. clearly identifying) for each assay variant.
B: Indication of the main intended use (max. 5 letters), e.g. DART, Neuro, Liver, Lung, Renal, Redox, Stress...
C: Specifier of test system, e.g. cell type such as NPC (neural precursor cells), NC (neural crest), Hep (liver cells), REN (kidney cells), PUL (lung cells) (max. 4 letters)
D: Identification of test endpoint, e.g. Diff_6D = Differentiation for 6 days; exp_24 h = exposure for 24 hours; RNA_6h = transcriptome after 6 hours (use max. 15 signs altogether; if desired in 2-3 blocks), name (and acronym) of the project partner home organisation.
E: version number.

2.5 Name and acronym of the test depositor
Include affiliation.

2.6 Name and email of contact person
Provide the details of the principal contact person.

2.7 Name of further persons involved
For example, the principal investigator (PI) of the lab, the person who conducted the experiments, etc.

2.8 Reference to additional files of relevance
Supply number of supporting files.
Describe supporting files (e.g. metadata files, instrument settings, calculation template, raw data file, etc.).

3. Description of general features of the test system source

3.1 Supply of source cells
Describe briefly whether the cells are from a commercial supplier, continuously generated by cell culture, or obtained by isolation from human/animal tissue (or other).

3.2 Overview of cell source component(s)
Give a brief overview of your biological source system, i.e. the source or starting cells that you use.
Which cell type(s) are used or obtained (e.g. monoculture/co-culture, differentiation state, 2D/3D, etc.)?
If relevant, give human donor specifications (e.g. sex, age, pool of 10 donors, from healthy tissue, etc.).

3.3 Characterization and definition of source cells
List quantitative and semi-quantitative features that define your cell source/starting cell population. For test methods that are based on differentiation, describe your initial cells, e.g. iPSC, proliferating SH-SY5Y; the differentiated cells are described in section 4.
Define cell identity, e.g. by STR signature (where available), karyotype information, sex (where available and relevant), ATCC number, passage number, source (supplier), sub-line (where relevant), source of primary material, purity of the cells, etc.
Describe defining biological features you have measured or that are FIRMLY established (use simple listing, limit to max. 0.5 pages), e.g. the cells express specific marker genes, have specific surface antigens, lack certain markers, have or lack a relevant metabolic or transporting capacity, have a doubling time of x hours, etc.
Transgenic cell lines have particular requirements concerning the characterization of the genetic manipulation (type of transgene, type of vector, integration/deletion site(s), stability, etc.).
Organoids and microphysiological systems (MPS) may need some special/additional considerations as detailed in Pamies et al. (2018) and Marx et al. (2016), e.g. ratio of cell types used, percent of normal cells in tumor spheroids created from resected tissue; derivation of cells for re-aggregating brain cultures.

3.4 Acceptance criteria for source cell population
Describe the acceptance criteria (AC) for your initial cells (i.e. the quality criteria for your proliferating cell line, tissue for isolation, organism, etc.). Which specifications do you consider to describe the material, which quality control criteria have to be fulfilled (e.g. pathogen-free)? Which functional parameters (e.g. certain biological responses to reference substances) are important?
For iPSC maintenance: How do you control pluripotency? How stable are your cells over several passages? Which passage(s) are valid?
For primary cells: Show stability and identity of supply; demonstrate stability of function (e.g. xenobiotic metabolism).
Quantitative definitions for AC should be given based on this defining information. Exclusion criteria (features to be absent) are also important.
As in 3.3., special/additional requirements apply to genetically-modified cells and microphysiological systems.

3.5 Variability and troubleshooting of source cells
Name known causes of variability of the initial cells/source cells.
Indicate critical consumables or batch effects (e.g. relevance of the plate format and supplier, batch effects of fetal calf serum (FCS) or serum replacement, critical additives like type of trypsin, apo-transferrin vs. holo-transferrin, etc.).
Indicate critical handling steps and influencing factors (e.g. special care needed in pipetting, steps that need to be performed quickly, cell density, washing procedures, etc.).
As in 3.3., special/additional requirements apply to genetically-modified cells and microphysiological systems, e.g. dependence on matrix chemistry and geometry, dependence on microfluidics system, consideration of surface cells vs core cells, etc.
Give recommendations to increase/ensure reproducibility and performance.

3.6 Differentiation towards the final test system
Describe the principles of the selected differentiation protocol, including a scheme and graphical overview, indicating all phases, media, substrates, manipulation steps (medium change/re-plating, medium additives, etc.). Special/additional requirements apply to microphysiological systems and organoids: e.g. cell printing, self-aggregation/self-organisation process, interaction with the matrix, geometrical characterization (size/shape), etc.

3.7 Reference/link to maintenance culture protocol
Provide the SOP of the general maintenance procedure as a database link. This should also include the following information:
How are the cells maintained outside the experiment (basic cell propagation)?
How pure is the cell population (average, e.g. 95% of iPSC cells Oct4-positive)?
What are the quality control measures and acceptance criteria for each cell batch?
Which number(s) passage(s) can be used in the test?
Is Good Cell Culture Practice (GCCP) and/or Good In Vitro *Method Practice (GIVIMP) followed?*
How long can same cell batches be used?
How are frozen stocks and cell banks prepared?
For primary cells: how are they obtained in general and what are they characterized for (and what are inclusion and exclusion criteria).

4. Definition of the test system as used in the method

4.1 Principles of the culture protocol
Describe the test system as it is used in the test.
If the generation of the test system involves differentiation steps or complex technical manipulation (e.g. formation of microtissues), this is described in 3.6.
Give details on the general features/principles of the culture protocol (collagen embedding, 3D structuring, addition of mitotic inhibitors, addition of particular hormones/growth factors, etc.) of the cells that are used for the test.
What is the percentage of contaminating cells; in co-cultures what is the percentage of each subpopulation?

Are there subpopulations that are generally more sensitive to cytotoxicity than others, and could this influence viability measures?
Is it known whether specific chemicals/chemical classes show differential cytotoxicity for the cell sub-populations used?

4.2 Acceptance criteria for assessing the test system at its start

What are the endpoint(s) that you use to control that your culture(s) is/are as expected at the start of toxicity testing (e.g. gene expression, staining, morphology, responses to reference chemicals, etc.)?
Describe the acceptance criteria for your test system, i.e. the quality criteria for your cells/tissues/organoids: Which endpoints do you consider to describe the cells or other source material, which parameters are important?
Describe the (analytical) methods that you use to evaluate your culture (PCR, ATP measurement) and to measure the acceptance criteria (AC).
Which values (e.g. degree of differentiation or cell density) need to be reached/should not be reached?
Historical controls: How does your test system perform with regard to the acceptance criteria, e.g. when differentiation is performed 10 times, what is the average and variation of the values for the acceptance criteria parameters)? Indicate actions if the AC are not met.
Examples: cell are > 90% viable, or > 98% of cells express marker x (e.g. AP-2), or > 80% of the cells attach, etc.

4.3 Acceptance criteria for the test system at the end of compound exposure

Describe the acceptance criteria for your test system, i.e. the quality criteria for your cells/tissues/organoids: Which endpoints do you consider to describe the cells or other source material, which parameters are important?
Which values (e.g. degree of differentiation or cell density) need to be reached/should not be reached?
Historical controls: How does your test system perform with regard to the acceptance criteria, e.g. when differentiation is performed 10 times, what is the average and variation of the values for the acceptance criteria parameters)? Indicate actions if the AC are not met.
Examples: Usual neurite length is 50 ±15 µm; experiments with average neurite length below 25 µm in the negative controls (NC) are discarded. Usual nestin induction is 200 ±40 fold, experiments with inductions below 80-fold for NC are discarded.

4.4 Variability of the test system and troubleshooting

Give known causes of variability for final test system state.
Indicate critical consumables or batch effects (e.g. plate format and supplier, batch effects of FCS or serum replacement, additives).
Indicate critical handling steps, and/or influencing factors identified (e.g. special care needed in pipetting, steps that need to be performed quickly, cell density).
Indicate positive and negative controls and their expected values, and accepted deviation within and between the test repeats.
Give recommendations to increase/ensure reproducibility and performance.

4.5 Metabolic capacity of the test system

What is known about endogenous metabolic capacity (CYP system (phase I); relevant conjugation reactions (phase II))?
What is known about other pathways relevant to xenobiotic metabolism?
What specific information is there on transporter activity?

4.6 Omics characterization of the test system

Are there transcriptomics data or other omics data available that describe the test system (characterization of cells without compounds)?
Briefly list and describe such data.
Indicate the type of data available (e.g. RNASeq or proteomics data).
Refer to data file, data base or publication.

4.7 Features of the test system that reflect the *in vivo* tissue

Give information on where the test system differs from the mimicked human tissue and which gaps of analogy need to be considered.

4.8 Commercial and intellectual property rights aspects of cells

Are there elements of the test system that are protected by patents or any other means?

4.9 Reference/link to the culture protocol

Fill only if section 3 has not been answered.
Provide the SOP for the general maintenance procedure as a database link. This should also include the following information:
How are the cells maintained outside the experiment (basic cell propagation)?
How pure is the cell population (average, e.g. 95% of iPSC cells Oct4-positive)?
What are the quality control measures and acceptance criteria for each cell batch?
Which number(s) passage(s) can be used in the test?
Is Good Cell Culture Practice (GCCP) and/or Good In Vitro *Method Practice (GIVIMP) followed?*
How long can same cell batches be used?
How are freezing stocks and cell banks prepared?
For primary cells: How are they obtained in general and what are they characterized for (and what are inclusion and exclusion criteria).

5. Test method exposure scheme and endpoints

5.1 Exposure scheme for toxicity testing

Provide an exposure scheme (graphically, show timelines, addition of medium supplements and compounds, sampling, etc.), within the context of the overall cell culture scheme (e.g. freshly re-plated cells or confluent cells at start, certain coatings, etc.).

Include medium changes, cell re-plating, whether compounds are re-added in cases of medium change, critical medium supplements, etc.

5.2 Endpoint(s) of the test method

Define the specific endpoint(s) of the test system that you use for toxicity testing (e.g. cytotoxicity, cell migration, etc.).
Indicate whether cytotoxicity is the primary endpoint.
What are secondary/further endpoints?
Also describe here potential reference/normalization endpoints (e.g. cytotoxicity, protein content, housekeeping gene expression) that are used for normalization of the primary endpoint.

5.3 Overview of analytical method(s) to assess test endpoint(s)

Define and describe the principle(s) of the analytical methods used. Provide here a general overview of the method's key steps (e.g. cells are fixed or not, homogenized sample or not, etc.), sufficient for reviewers/regulators to understand what was done, but not in all detail for direct repetition.
If you have two or more endpoints (e.g. viability and neurite outgrowth), do you measure both in the same well, under same conditions in parallel, or independently of each other?
For imaging endpoints: Explain in general how quantification algorithm or how semi-quantitative estimates are obtained and how many cells are imaged (roughly).

5.4 Technical details (of e.g. endpoint measurements)

Provide information on machine settings, analytical standards, data processing and normalization procedures.
For imaging endpoints: provide detailed algorithm.
This information should also be covered in an SOP, preferably in DB-ALM format (see link in 6.6).

5.5 Endpoint-specific controls/mechanistic control compounds (MCC)

MCC are chemicals/manipulations that show biologically plausible changes of the endpoint. List such controls (up to 10), indicate why you consider them as MCC, and describe expected data on such controls. Highlight the compounds to be used for testing day-to-day test performance, i.e. for setting acceptance criteria (AC).
If available, indicate MCC that each increase or decrease the activity of the relevant pathway. Do pathway inhibitions or activations correlate with the test method response?

5.6 Positive controls

What chemicals/manipulations are used as positive controls? Describe the expected data on such controls (signal and its uncertainty)? How good are in vivo *reference data on the positive controls? Are* in vivo *relevant threshold concentrations known?*

5.7 Negative and unspecific controls

What chemicals/manipulations are used as negative controls? Describe the expected data on such controls (signal and its uncertainty)? (Such data define the background noise of the test method)
What is the rationale for the concentration setting of negative controls?
Do you use unspecific controls? If yes, indicate the compounds and the respective rationale for their use and the concentration selection.

5.8 Features relevant for cytotoxicity testing

Does the test system have a particular apoptosis sensitivity or resistance?
Is cytotoxicity hard to capture for minor cellular subpopulations?
In multicellular systems, which cell population is the most sensitive? Are specific markers known for each cell population?
Are there issues with distinguishing slowed proliferation from cell death?
For repeated/prolonged dosing: Is early death and compensatory growth considered?
For very short-term endpoints (e.g. electrophysiology measured 30 min after toxicant exposure): Is a delayed measure of cytotoxicity provided?

5.9 Acceptance criteria for the test method

Which rule do you apply to test whether a test run is within the normal performance frame?
How do you document this decision?
Indicate actions if the AC are not met.

5.10 Throughput estimate

Indicate "real data points per month" (not per week/per quarter, etc.): count three working weeks per month. Each concentration is a data point. Necessary controls that are required for calibration and for acceptability criteria are NOT counted as data points. All technical replicates of one condition are counted as one single data point (see notes for explanation)
Indicate possibility/extent of repeated measures (over time) from same dish.
Explain your estimate.

6. Handling details of the test method

6.1 Preparation/addition of test compounds

Give an overview of the range of volumes, particular lab ware/instruments for dispensing, temperature/lighting considerations, particular media/buffers for dilution, decision rules for the solvent, tests of solubility as stocks and in culture medium, etc.
How are compound stocks prepared (fold concentration, verification, storage, etc.)?

How are dilutions prepared? What solvent is used? Is filtering used to obtain sterility?
How does the final addition to the test system take place?
Give details of addition of test compounds to test systems (e.g. in which compartment of compartmentalized cultures, in which volume, before after or during medium change, etc.).

6.2 Day-to-day documentation of test execution
How are day-to-day procedures documented (type of 'lab book' organisation, templates)?
Define lab-specific procedures used for each practical experiment on how to calculate test compound concentrations (and to document this).
How are plate maps defined and reported?
Detailed information should also be covered in an SOP, preferably in DB-ALM format (see link in 6.6).

6.3 Practical phase of test compound exposure
How is the time plan of pipetting established, followed, and documented?
How is adherence to plate maps during pipetting documented?
What are the routine procedures to document intermediate steps with potential errors, mistakes and uncertainties?
How are errors documented (e.g. pipetting twice in one well)?
How are the plate wells used sequentially – following which pattern?
Detailed information should also be included in an SOP, preferably in DB-ALM format (see link in 6.6).

6.4 Concentration settings
How is the concentration range of test compounds defined (e.g. only single concentrations, always 1:10 serial dilutions or variable dilution factors, ten different concentrations, etc.)? Is there a rule for defining starting dilutions?
For functional endpoints that may not provide full concentration-response, how is the test concentration defined? E.g. EC_{10} of viability data are usually tested for gene expression endpoints.
Detailed information should also be included in an SOP, preferably in DB-ALM format (see link in 6.6).

6.5 Uncertainties and troubleshooting
What types of compounds are problematic, e.g. interference with analytical endpoint, low solubility, precipitation of medium components, etc.?
What experimental variables are hard to control (e.g. because they are fluorescent)?
What are critical handling steps during the execution of the assay?
Robustness issues, e.g. known variations of test performance due to operator training, season, use of certain consumable or unknown causes, etc.
Describe known pitfalls (or potential operator mistakes).

6.6 Detailed protocol (SOP)
Ideally the SOP follows the DB-ALM or a comparable format:
https://ecvam-dbalm.jrc.ec.europa.eu/home/contribute
Refer to additional file(s) (containing information covered in sections 3 and 4), containing all details and explanations.
Has the SOP been deposited in an accessible data base?
Has the SOP been reviewed externally and if yes, how?

6.7 Special instrumentation
Does the method require specialized instrumentation that is not found in standard laboratories?
Is there a need for custom-made instrumentation or material?
Is there a need for equipment that is not commercially available (anymore)?

6.8 Possible variations
Describe possible variations, modifications and extensions of the test method:
a) other endpoints,
b) other analytical methods for same endpoint,
c) other exposure schemes (e.g. repeated exposure, prolonged exposure, etc.),
d) experimental variations (e.g. use of a specific medium, presence of an inhibitor or substrate that affects test outcome, etc.)

6.9 Cross-reference to related test methods
Indicate the names (and database names) of related tests and give a short description (including a brief comment on differences to the present method).
If the test method has been used for high throughput transcriptomics or deep sequencing as alternative endpoint, this should be indicated.

7. Data management

7.1 Raw data format
What is the data format?
Raw data: give general explanation. Upload an exemplary file of raw data (e.g. Excel file as exported out of plate reader).
Provide an example of processed data at a level suitable for general display and comparison of conditions and across experiments and methods.

If the file format is not proprietary or binary, include a template. This will help other users to provide their data in a similar way to the general data infrastructure.
Example as used in EU-ToxRisk: Excel sheet with columns specifying line number, assay name, date of experiment, identifier for reference to partner lab book, compound, concentration (in: -log[M]), line number of corresponding control, number of replicates, endpoints, data of endpoint(s), etc.

7.2 Outliers
How are outliers defined and handled?
How are they documented?
Provide the general frequency of outliers.

7.3 Raw data processing to summary data
How are raw data processed to obtain summary data (e.g. EC_{50}, BMC15, ratios, PoD, etc.) in your lab?
Describe all processing steps from background correction (e.g. measurement of medium control) to normalization steps (e.g. if you relate treated samples to untreated controls).

7.4 Curve fitting
How are data normally handled to obtain the overall test result (e.g. concentration response fitting using model X, determination of EC_{50} by method Y, use of EC_{50} as final data)?
How do you model your concentration response curve (e.g. LL.4 parameter fit) and which software do you use (e.g. GraphPad Prism, R, etc.)?
Do you usually calculate an uncertainty measure of your summary data (e.g. a 95% confidence interval for the BMC or a BMCL), and with which software?
Can you give uncertainty for non-cytotoxicity or no-effect?
How do you handle non-monotonic curve shapes or other curve features that are hard to describe with the usual mathematical fit model?

7.5 Internal data storage
How and how long are raw and other related data stored?
What backup procedures are used (how frequently)?
How are data versions identified?

7.6 Metadata
How are metadata documented and stored (lab book, Excel files, left in machine, etc.)?
How are they linked to raw data?
What metadata are stored/should be stored?

7.7 Metadata file format
Give example of the metadata file (if available).
If metadata or data format (see 7.1) are pre-defined in the project, state here "as pre-defined in project xxx" (e.g. EU-ToxRisk).

8. Prediction model and toxicological application

8.1 Scientific principle, test purpose and relevance
What is the scientific rationale to link test method data to a relevant in vivo *adverse outcome?*
Which toxicological target (organ, tissue, physiological/biochemical function, etc.) is modelled?
Which biological process(es) (e.g. neurite outgrowth, differentiation) are modelled/reflected by your test method?
Which toxicological events (e.g. oxidative stress, cell death) are modelled/reflected by your test method?
To which (human) adverse outcome(s) is your test method related?
Which hazard(s) do(es) your test method (potentially) predict?

8.2 Prediction model
Provide the statistics of your benchmark response (threshold and variance):
(i) For dichotomized data, provide your prediction model. When do you consider the result as toxic or not toxic?
(ii) For pseudo-dichotomized outcomes (two classes with borderline class in between): define borderline range.
(iii) For multi-class or continuous outcomes: provide definitions and rationale.
What is the rationale for your threshold? This can be on a mathematical (e.g. 3-fold standard deviation) or a biological basis (e.g. below 80% viability).
Is there a toxicological rationale for the threshold settings and definitions of your prediction model?
What are the limitations of your prediction model?
What is a 'hit' if the test is used in screening mode (= hit definition, if different from above)?

8.3 Prediction model setup
How was the prediction model set up (using which test set of chemicals to train the model; using probing with what kind of classifiers/ statistical approaches)?
Has the prediction model been tested (what was the test set of chemicals)? List chemicals or give n, if n > 50.
Is the process documented (publication)?
Does the prediction model (PM) apply to changes to both sides of controls (up/down)? If the PM is one-sided (e.g. toxicants leading

to a decrease vs. control), how are data in the opposite direction handled and interpreted? If the PM is two-sided, do different rules, characteristics and interpretations apply to the two sides (e.g. is a decrease in viability or an increase in viability both interpreted as an effect/toxicity; are thresholds and performance characteristics to both sides the same?).

8.4 Test performance

Indicate here basic performance parameters or, if possible, preliminary estimates (label as such): Baseline variation (noise) within assays AND between assays.
What is the signal/noise ratio (signal = standard positive control)?
Is the z-factor determined?
Give the specificity of the test method. How is it determined?
Give the sensitivity of the test method. How is it determined?
Give measures of the uncertainty of your test method. How are they determined?
What is the detection limit (required change of endpoint to become measurable)?
If available, give limit of detection (LOD) and limit of quantification (LOQ).
What are inter-operator variations?
Are there data of 'historical controls' over a longer time period?

8.5 *In vitro – in vivo* extrapolation (IVIVE)

Describe parameters important for the determination of free compound concentrations in the medium.
Indicate the lipid and protein content of the medium and the cells.
Indicate the volume of the cells.
Indicate volume (medium volume) and surface area of culture dish.
Is there information/literature on IVIVE strategies/data in the test?
Has the test been used earlier for IVIVE?
Are there special considerations that are relevant for IVIVE (e.g. potential for compound accumulation due to frequent medium changes and compound re-addition, glycoprotein (MDR1) expression, capacity for xenobiotic metabolism of test system)?

8.6 Applicability of test method

Which compounds is the test likely to pick up correctly, where is it likely to fail?
How does the test method react to mixtures and UVCBs?
Are there areas (according to industry sector, compound chemistry, physical-chemical properties) that need to be excluded from testing, or that are particularly suitable?
Which compound class cannot be detected (e.g. neurotransmitters for which the receptors are not expressed, endocrine disruptors in absence of respective pathway)?
Are any compounds known to interfere with the test system (e.g. fluorescent or colored chemicals)?

8.7 Incorporation in test battery

Does the test fit into a test battery? If yes, into which test battery and are there any restrictions?
Indicate potential strengths and weaknesses of the system in a test battery (e.g. method is a good confirmation assay, good for creating alerts, mechanistic follow-up, screening, etc.).
Compare performance to similar tests.
Which gaps in a known or potential battery does the test method fill?
Should the test preferentially be used in the first tier or later tiers, are complementary assays required or is it a stand-alone method?

9. Publication/validation status

9.1 Availability of key publications

Refer to published literature on the test AND indicate in detail deviations from published descriptions (e.g. plastic plate supplier, cell number, endpoint measurement, timing, etc.).
Provide the most relevant publications that describe/give a comprehensive overview of (a) your test system and/or (b) your test method. Describe what aspects are covered therein.
Give a prioritized (according to importance) list of further publications on the test method or its application.
Give short comments on which type(s) of information can be obtained from these publications (e.g. contains test chemical lists, contains more positive/negative controls, contains validation against other tests, contains incorporation in test battery, demonstrates use by other lab, etc.).

9.2 (Potential) linkage to AOPs

Indicate whether the test method has been or could be linked to an AOP (or several AOPs) and in which form (e.g. test of KE activation).
Can the test method cover an AOP MIE/KE?
Reference relevant AOP and if in AOP-wiki, refer to status.

9.3 Steps towards mechanistic validation

Indicate/summarize information on mechanistic validation, e.g. by omics approaches or by use of endpoint specific controls (MCC; section 5.5).
Has it been explored in how far the system reflects human biology, signaling, tissue organization relevant to the form of toxicity to be assessed (e.g. nigrostriatal neurons should contain dopamine, liver tests relevant to cholestasis may need to contain bile canalicular structures, etc.)?

Are there interventions (knock-out, knockdown, chemical inhibitors, specific pathway triggering) that support the use of the test for certain toxicological questions and that corroborate expectations to the test system?
Is there a form of mechanistic validation?
Do toxicant-altered genes (or other biomarkers) correspond to changes in mimicked human tissue (after poisoning or in relevant pathologies)?

9.4 Pre-validation or validation
Indicate/summarize activities for test qualification, pre-validation or validation.
Indicate e.g. ring trials, full (pre-)validations.
Give an overview of compounds or libraries that have been tested.

9.5 Linkage to (e.g. OECD) guidelines/regulatory use
Indicate whether the test method is linked to an OECD Test Guideline (how, and which) or other regulatory guidance (e.g. EMA).

10. Test method transferability

10.1 Operator training
What experience is required?
How are new operators trained in your laboratory?
How much training/experience is required for smooth assay performance?

10.2 Transfer
Has the test system been transferred to other labs?
Has the test method been used by various operators (over a long time period)?
Has the test method been transferred to other labs?
Is there data on inter-laboratory variability?
What are procedures and how was the performance (experience) of the transfer?

11. Safety, ethics and specific requirements

11.1 Specific hazards; issues of waste disposal
Are there special legal requirements for running the test in your lab; are there special hazards associated with the test that may affect operators, bystanders, others (e.g. through waste).

11.2 Safety data sheet (SDS)
Are the SDSs for all hazardous reagents used in the test method available?
Are the SDSs for all hazardous test compounds stored?
Describe where and how the SDSs are stored internally. How is safe handling ensured?
Is the exposure scenario for the hazardous reagents used in the test method available?

11.3 Specific facilities/licenses
Are special permits (e.g. genetic work, stem cells, radioactivity, etc.) required?
Are special facilities required?
Is special ethical approval necessary (indicate approval document).

11.4 Commercial aspects/intellectual property of material/procedures
List elements of the test method (e.g. consumables, chemicals, analytical methods, equipment) that are protected by patents or any other means. Indicate the type of protection and where the element (or license for it) may be obtained.

The many types of essential information can be grouped as belonging to four partially overlapping packages: (i) the overall test method description, (ii) the technical test procedure (as outlined in a standard operating procedure (SOP) involving, e.g., defined labware, consumables and pipetting steps), (iii) the characterization of test and reference materials/chemicals, and (iv) all issues relating to data processing and archiving. In a wider sense, an additional package (v) addresses the test purpose, the test limitations (i.e., information on its applicability) and the criteria to be used for interpreting test results. Here, the focus is on the overall test method description (packages i + v). Notably, this description also involves some information from the other packages (ii-iv) – not in full detail but as far as required to provide an overall understanding of the test method.

6 Why data are meaningless without a test method description

There is an intricate relationship between a test method and the data it generates. It is evidently clear that a test method itself has no value in toxicology if it does not generate data. The reverse condition is often overlooked: Is there any value in data if the test method is not sufficiently documented or disclosed? This question is not easily answered. One reason for this is that "naked data", i.e., data without reference to a test method, do not exist for *in vivo* tests. Most data inherently contain some information on test conditions. For instance, if one talks about data on the LD50 (lethal dose, 50%) test for oral toxicity in rat (example data: 4.5 ±1 mg/kg bodyweight), then the data contain information on

the test system (rat) and the exposure scheme (oral dosing, timing). Also, information on the endpoint and a normalization of the dose to a test unit can be easily derived from the data. In the above example, this would be death (as endpoint) and dose per body weight unit (as data normalization procedure). Without such inherent test method information, a data point (e.g., 5 mg or 117 g) would be meaningless.

The situation is even more extreme for cell-based tests or new approach methods (NAM) in general. Data may be something like "10%". Depending on the test, this may refer to a 10% reduction in viability (i.e., hardly any effect), 10% remaining viability (a drastic effect), a 10% increase in neurite growth (a moderate effect of unclear relevance) or 10% remaining level of acetylcholine activity (extreme adversity). Even if this was specified, it would not be possible to interpret the data, as they may have been obtained after exposure for 5 min to 10 nM compound or after exposure for one year to 1 mM compound. This example, trivial as it may appear, is intended to explain an important point: Data have no meaning if the respective test method with all its elements is not disclosed fully and transparently (Leist and Hengstler, 2018).

Large studies (mainly focusing on animal experimentation) have reported that test method descriptions are often poor (Freedman et al., 2015; Hair et al., 2019; Vogt et al., 2016; Ingre-Khans et al., 2019). This lack of test method transparency may contribute significantly to reproducibility issues often reported for biomedical research (Ioannidis, 2012; Begley and Ellis, 2012; Prinz et al., 2011; Hartung et al., 2019). In the toxicological context, data obtained using test method descriptions that are not sufficiently clear cannot be considered "valid", i.e., "suitable for a regulatory context or in a context where major scientific or commercial decisions depend on the data interpretation" (Hartung and Leist, 2008; Leist et al., 2014). Given the experience that many essential details on methods are often not disclosed in scientific publications, it is of prime importance to define meticulously what is required of a method description and also to provide guidance on structuring this information (OECD, 2018; Leist et al., 2010, 2012; Hartung et al., 2019).

7 How to define a cell

The test system of many *in vitro* test methods is a cell culture (in two- or three-dimensional format). Therefore, the definition of the cells is an important element of a test method description. This is also clearly stated in GD211. However, the problem of defining a cell appears to have been underestimated – especially with respect to many modern, highly dynamic and complex cell cultures (Pamies et al., 2017, 2018; Balmer et al., 2012; Bal-Price et al., 2015). To appreciate the problem, it is helpful to initially look at some attempts to define a cell and to examine the problems associated with them:

1. *Quoting the catalogue number or another apparent cell identifier*. Often the ATCC (American Type Culture Collection) number, the colleague who provided the cells, or the tissue from which the (primary) cells were obtained is given. This may work to some extent if all principles of Good Cell Culture Practice (GCCP) are considered and adhered to (Coecke et al., 2005). Yet, many problems have been described in this regard. For instance, many cells have been misidentified or cross-contaminated with other cells (Drexler et al., 2003; Gignac et al., 1993; Horbach and Halffman, 2017; Masters, 2002; Nardone, 2007, 2008; Stacey, 2000; Stacey et al., 1992), or microbiologically contaminated (e.g., with mycoplasms). Even two batches of cells from a single source can differ significantly (Ben-David et al., 2018; Frattini et al., 2015; Kleensang et al., 2016; Liu et al., 2019). Even more importantly, this form of identification only applies to cell cultures that are based on well-defined cell lines. Even in apparently simple cases (e.g., using the HepG2 cell line or primary rodent hepatocytes or neuronal cells), cultures of the same cells have different properties depending on their culture medium, cell density, cell-matrix, etc. (Brigelius-Flohe et al., 1995; Latta et al., 2000; Leist et al., 1999; Ramaiahgari et al., 2014; Delp et al., 2019; Falsig et al., 2006; Gantner et al., 1996; Gerhardt et al., 2001; Zimmer et al., 2012).
2. *Determination of the cells' genotype*. Sometimes, genotyping, e.g., by short tandem repeat (STR) profiling, by extensive single nucleotide polymorphism (SNP) profiling or by array comparative genome hybridization (aCGH) (Zhang et al., 2017; Matsuda, 2017; Cao et al., 2015) is used for defining a test system. However, the genotype gives no information on the differentiation state or the epigenetic state. This problem is also not circumvented by complete genome sequencing (Gutbier et al., 2018).
3. *Reference to the original source*. For primary cells, this has in the past been considered as sufficiently defining (e.g., for hepatocytes, neurons, broncho-alveolar cells or blood cell populations) (Kruglikov et al., 1976; Schildknecht et al., 2011, 2013; Gerhardt et al., 2001). However, it is now ackowldeged that this approach does not account for the large variability in cell purity, viability, activation/inflammatory state, de-differentiation, history within the organism (e.g., drug treatment), etc. Even the culture setup can make a dramatic difference. This is exemplified by broncho-alveolar cells grown submerged in medium (standard type of culture) or at the air-liquid interphase, which promotes baso-lateral polarization. Even more variation is introduced if one considers that the cells may be confluent (with tight junctions) or non-confluent, at an early or late passage number, etc.
4. *Reference to a differentiation protocol*. Many modern test systems are derived from stem cells that have been differentiated. If they are obtained from commercial sources, the protocol used to generate the cells is mostly confidential, and variability between lots is unknown. Even if all protocol details are known, the issue still remains that one given cell (with a defined genome) can give rise to dozens of differentiated progeny cell types and that those may have multiple health and proliferation states (Stiegler et al., 2011; Zimmer et al., 2011). The differentiation protocol is not sufficiently defining for the final population.

Is it at all possible to unambiguously define a cell culture? At the present state of the art, this may indeed not be possible (Gutbier et al., 2018; Liu et al., 2019) unless the use of the culture is specified. With a specific use scenario in mind (use of the cells in a

Tab. 1: Synopsis of information found in OECD GD211 and the extensive test method description questionnaire (ToxTemp) compiled here

The left column lists all chapters/items of GD211 (original numbering) (OECD, 2017), while the right column indicates where the corresponding information can be found in ToxTemp. Also, GD211 sometimes lists several important issues without assignment to sub-items (see, e.g., item 3 "Data interpretation and prediction model" of GD211). Such compilations of several aspects are often covered by several distinct chapters of ToxTemp. Two examples for the diverging levels of detail are highlighted in blue: (i) chapter 2.3 of GD211 addresses the test system (cells). This information is covered in 15 sub-chapters of ToxTemp; (ii) in chapter 2.6 of GD211, there are several sub-items (bullets) that are highly divers. These are therefore covered in ToxTemp in very different chapters. A more detailed comparison is found in Tab. S2[1].

GD211 chapter and chapter name		Information to be found in chapter(s) of ToxTemp
1	**General information**	
	1.1 Assay Name (title) / 1.2 Summary / 1.3 Date of MD / 1.4 MD author(s) and contact details / 1.5 Date of MD update(s) and contacts / 1.6 Developer(s) laboratory and contact details / 1.7 Date of assay development and/or publication / 1.8 Reference(s) to scientific papers / 1.9 Information about the assay in relation to proprietary elements / 1.10 Information about the throughput / 1.11 Status of method development and uses / 1.12 Abbreviations and Definitions	1.1; 1.2; 2.2; 2.3; 2.5; 2.6; 2.7; 4.8; 5.10; 8.1; 9.1; 9.4; 9.5; 10.2; 11.4
2.	**Test method definition**	
2.1	Purpose of the test method	4.7; 8.1; 8.7; 9.2; 9.3
2.2	Scientific principle of the method	4.7; 5.5; 8.1; 9.2; 9.3
2.3	Tissue, cells or extracts utilised in the assay and the species source	3.1 - 3.7; 4.2 - 4.8
2.4	Metabolic competence of the test system	4.5
2.5	Description of the experimental system exposure regime	5.1; 5.6; 5.7; 5.10; 6.4; 6.7; 8.6; 11.3
2.6	Response and response measurement • Response here makes reference to any biological effect, process or activity that can be measured • Specify precisely and describe the response and its measurement • Specify the precise response or activity investigated as applicable e.g. "IC50" • And how it is calculated	1.2; 5.2; 5.3; 7.3; 7.4; 8.1; 9.3 5.2 Endpoint(s) of the test method 1.2 Abstract 8.1 Scientific principle, test purpose and relevance 9.3 Steps towards mechanistic validation 5.2 Endpoint(s) of the test method 5.3 Overview on analytical method(s) to assess test endpoint(s) 7.3 Raw data processing to summary data 7.4 Curve fitting
2.7	Quality / Acceptance criteria	3.2; 3.4; 4.2; 4.3; 5.4; 5.5; 5.6; 5.7; 6.7; 7.1; 7.5; 8.4; 9.5
2.8	Known technical limitations and strengths	5.8; 8.6; 8.7
2.9	Other related assays that characterise the same event as in 2.1	6.9
3.	**Data interpretation and prediction model**	8.2; 8.3; 9.2; 9.3
3.1	Assay response(s) captured in the prediction model	5.2; 8.2
3.2	Data analysis	7.3; 7.4; 8.4
3.3	Explicit prediction model	8.2
3.4	Software name and version for algorithm/prediction model generation	7.4
4.	**Test method performance**	
4.1	Robustness of the method	8.4; 10.2
4.2	Reference chemicals/chemical libraries, rationale for their selection and other available information	5.5; 5.6; 5.7; 8.3; 8.6
4.3	Performance measures/predictive capacity	7.4; 8.3; 9.4
4.4	Scope and limitations of the assay	8.2; 8.6; 8.7
5.	**Potential regulatory applications**	8.1
5.1	Context of use	8.2; 8.4; 8.6; 8.7
6.	**Bibliography**	9.1
7.	**Supporting information**	2.8

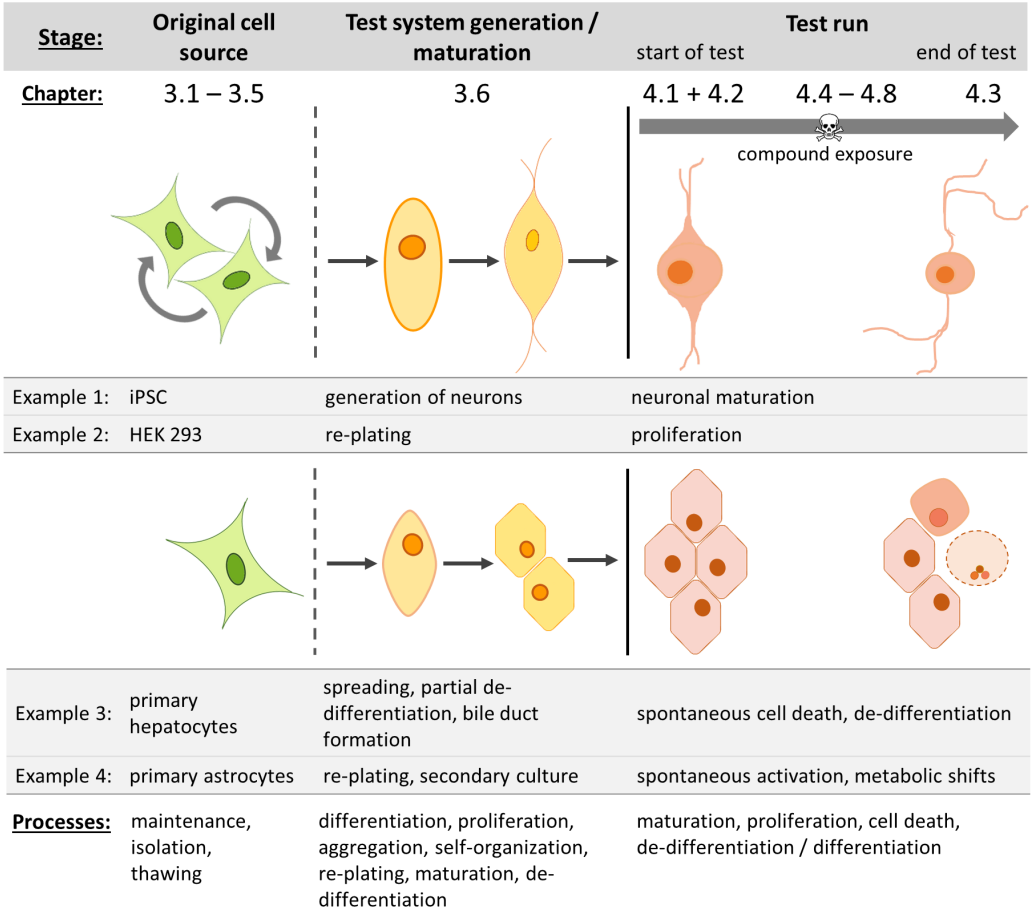

Fig. 1: Documentation of cell culture stages relevant for their use in an *in vitro* test method

The ToxTemp presented here assesses test system features and the respective acceptance criteria (AC) at different stages. In chapter 3, the original source of cells and their characteristics (green cells) as well as their differentiation/maturation (yellow cells) are covered. Chapter 4 focuses on the test system stages at the beginning of the test, i.e., at the start of chemical exposure, and at the final stage (end of test; red cells). Four examples of cellular test systems and their stages during test preparation and testing are given. In addition, exemplary processes (e.g., proliferation) that can change a test system, which should therefore be documented and taken into account for a comprehensive test method documentation, are indicated.

given test method), a fit-for-purpose definition is feasible and realistic (Lorge et al., 2016).

The ToxTemp includes two sections that help to better define cells. First, a cell culture is not seen as one static and given system. Therefore, it contains a whole suite of questions, notes and guiding questions that address different stages of the cells and are designed to define the entire process leading from a maintenance culture (or an original tissue) to the final test system (Fig. 1). Second, the definition of acceptance criteria (AC) for the test system is explained and requested. This ensures that the state of the cell culture can be defined in a pre-determined range that is fit-for-purpose concerning the use of the test system within the given test method (Blaauboer et al., 2012; Hansson et al., 2000; Hirt et al., 2000).

8 Acceptance criteria (AC)

The ToxTemp places a strong emphasis on acceptance criteria, not only for the overall test method but also for the test system at different stages. Going one step further, the questions address specification of the test system at the start and at the end of testing. This is necessary in many cases when the cell culture changes significantly during the test (chemical exposure phase). Examples for this are the use of differentiating stem cells (shift of subpopulations) or of primary cells (shift of composition, differentiation state, viability, activation state and other features).

The definition and application of acceptance criteria (AC) ensures that the specifications of test method elements are within a

fit-for-purpose state that allows the performance of the test under robust/reproducible conditions. The criteria and methods to control acceptability, including information on historic control data, are an indispensable information requirement for a test that may be used in a regulatory context and/or may need to be transferred from one laboratory to another.

9 Validation status

"Validation" generally describes the process of establishing that a test method is fit for the test purpose. In chemical risk assessment this often translates into demonstrating the adequacy, relevance and reliability of the test method for the purpose in question (e.g., OECD Manual for the Investigation of High Production Volume Chemicals, chapter 3[5]; Hartung et al., 2004; Balls et al., 2006; Hartung, 2007; Hoffmann and Hartung, 2006; Coecke et al., 2014, 2016).

As described above, an *in vitro* test method often has a primary purpose, i.e., to investigate the biological activity of a chemical with respect to a cell-based endpoint, but may also be used for secondary or further regulatory purposes. It is well-established that a test method for which principal fitness for the primary purpose has not been actively established should not be used for any regulatory purpose. However, a test that is fit for a primary purpose is not automatically fit for a secondary or other regulatory purpose. For instance, even if an assay has been found "fit" to reliably predict stress gene activation, and there is a hypothesis that such activation is responsible for a certain type of liver toxicity, and therefore the test method may be useful as an indirect way (a model) to predict this type of toxicity in humans, additional steps have to be taken to demonstrate the fitness of the test method for this secondary regulatory purpose, e.g., by comparison with the respective human data or with other, already established models.

While this general notion may be acceptable for most test method developers, the need for "formal validation" (by dedicated institutions, according to standardized and harmonized workflows) is discussed more controversially. As background for such discussions, it may be useful to rationalize that chemical risk assessment relies on a highly defined methodological framework that is harmonized internationally to the extent possible. In this system, mutual acceptance of a test method is achieved most readily if an internationally trusted, independent body (such as the OECD) has confirmed that it is fit-for-purpose. Therefore, the formal validation status is often considered an important element of test method descriptions, and such information is also requested by GD211.

Notably, non-guideline methods, the main subject of GD211, have usually not been formally validated. Many of them are also not pre-validated or in the progress of being validated. Thus, answers to the question on the validation status may be relatively information-poor, i.e., taking the form of "non-applicable", "unknown" or "no specific information / no validated state". More information on the status of a test method may be obtained if test readiness is defined more broadly, allowing for different readiness stages, depending on the purpose/application of the test method (e.g., screening versus regulatory use) (Bal-Price et al., 2018; Fritsche et al., 2017). Within such a framework, different elements of a test method may be evaluated separately, e.g., in a modular validation process (Hartung et al., 2004), or according to specified readiness criteria and scoring lists (Bal-Price et al., 2018). The ToxTemp gives guidance on how to provide such detailed information.

10 Conclusions

Interactions with many researchers from academia and industry have shown that the most common roadblock to practical implementation of GD211 is a lack of understanding of the requirements, and that additional notes and guiding questions are necessary to overcome this. Moreover, it was recognized that a comprehensive test method description needs more focus on practical aspects of the testing process (such as details on data handling or of the test system) and on the documentation of method limitations and troubleshooting. We found that the project partners of the EU-ToxRisk project were all able, compliant and motivated to deliver comprehensive method descriptions when guided by the ToxTemp. This shows that there is a positive attitude and a good motivation amongst test developers to describe their assay. A tool/template such as the one presented here is likely to act as support and catalyst towards a culture of fully transparent and complete test method descriptions.

In conclusion, the ToxTemp presented here will make it easier for the many test developers not deeply familiar with regulatory environments to describe their assays in sufficient detail. It will also allow to better understand and compare different testing methods (e.g., for the same endpoint). If the ToxTemp were broadly implemented, it would not only give guidance to test developers but could change the overall culture of method documentation. Scientists fostered in an environment that values sound descriptions of methods and data will not only become better (more reliable) researchers but will also be better trained for entering responsible positions in industry and regulatory agencies.

References

Bal-Price, A., Crofton, K. M., Leist, M. et al. (2015). International stakeholder network (ISTNET): Creating a developmental neurotoxicity (DNT) testing road map for regulatory purposes. *Arch Toxicol 89*, 269-287. doi:10.14573/altex.1402121

Bal-Price, A., Hogberg, H. T., Crofton, K. M. 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

Balls, M., Amcoff, P., Bremer, S. et al. (2006). The principles of weight of evidence validation of test methods and

[5] http://www.oecd.org/chemicalsafety/risk-assessment/49191960.pdf

testing strategies. The report and recommendations of ECVAM workshop 58. *Altern Lab Anim 34*, 603-620. doi:10.1177/026119290603400604

Balmer, N. V., Weng, M. K., Zimmer, B. et al. (2012). Epigenetic changes and disturbed neural development in a human embryonic stem cell-based model relating to the fetal valproate syndrome. *Hum Mol Genet 21*, 4104-4114. doi:10.1093/hmg/dds239

Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature 483*, 531-533. doi:10.1038/483531a

Ben-David, U., Siranosian, B., Ha, G. et al. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. *Nature 560*, 325-330. doi:10.1038/s41586-018-0409-3

Blaauboer, B. J., Boekelheide, K., Clewell, H. J. et al. (2012). The use of biomarkers of toxicity for integrating in vitro hazard estimates into risk assessment for humans. *ALTEX 29*, 411-425. doi:10.14573/altex.2012.4.411

Brigelius-Flohe, R., Lotzer, K., Maurer, S. et al. (1995). Utilization of selenium from different chemical entities for selenoprotein biosynthesis by mammalian cell lines. *Biofactors 5*, 125-131.

Cao, M. D., Balasubramanian, S. and Boden, M. (2015). Sequencing technologies and tools for short tandem repeat variation detection. *Brief Bioinform 16*, 193-204. doi:10.1093/bib/bbu001

Coecke, S., Balls, M., Bowe, G. et al. (2005). Guidance on good cell culture practice. A report of the second ECVAM task force on good cell culture practice. *Altern Lab Anim 33*, 261-287. doi:10.1177/026119290503300313

Coecke, S., Bowe, G., Milcamps, A. et al. (2014). Considerations in the development of in vitro toxicity testing methods intended for regulatory use. In A. Bal-Price and P. Jennings (eds.), *In Vitro Toxicology Systems*. New York, NY: Springer New York. doi:10.1007/978-1-4939-0521-8_25

Coecke, S., Bernasconi, C., Bowe, G. et al. (2016). Practical aspects of designing and conducting validation studies involving multi-study trials. *Adv Exp Med Biol 856*, 133-163. doi:10.1007/978-3-319-33826-2_5

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

Delp, J., Funke, M., Rudolf, F. et al. (2019). Development of a neurotoxicity assay that is tuned to detect mitochondrial toxicants. *Arch Toxicol 93*, 1585-1608. doi:10.1007/s00204-019-02473-y

Drexler, H. G., Dirks, W. G., Matsuo, Y. et al. (2003). False leukemia-lymphoma cell lines: An update on over 500 cell lines. *Leukemia 17*, 416-426. doi:10.1038/sj.leu.2402799

Falsig, J., Porzgen, P., Lund, S. et al. (2006). The inflammatory transcriptome of reactive murine astrocytes and implications for their innate immune function. *J Neurochem 96*, 893-907. doi:10.1111/j.1471-4159.2005.03622.x

Flood, S., Houck, K. and Grulke, C. (2017). Development of a Context-Rich Database of ToxCast Assay Annotations. doi:10.23645/epacomptox.5178610.v1

Frattini, A., Fabbri, M., Valli, R. et al. (2015). High variability of genomic instability and gene expression profiling in different HeLa clones. *Sci Rep 5*, 15377. doi:10.1038/srep15377

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

Fritsche, E., Crofton, K. M., Hernandez, A. F. et al. (2017). OECD/EFSA workshop on developmental neurotoxicity (DNT): The use of non-animal test methods for regulatory purposes. *ALTEX 34*, 311-315. doi:10.14573/altex.1701171

Gantner, F., Leist, M., Kusters, S. et al. (1996). T cell stimulus-induced crosstalk between lymphocytes and liver macrophages results in augmented cytokine release. *Exp Cell Res 229*, 137-146. doi:10.1006/excr.1996.0351

Gerhardt, E., Kugler, S., Leist, M. et al. (2001). Cascade of caspase activation in potassium-deprived cerebellar granule neurons: Targets for treatment with peptide and protein inhibitors of apoptosis. *Mol Cell Neurosci 17*, 717-731. doi:10.1006/mcne.2001.0962

Gignac, S. M., Steube, K., Schleithoff, L. et al. (1993). Multiparameter approach in the identification of cross-contaminated leukemia cell lines. *Leuk Lymphoma 10*, 359-368. doi:10.3109/10428199309148561

Gutbier, S., May, P., Berthelot, S. et al. (2018). Major changes of cell function and toxicant sensitivity in cultured cells undergoing mild, quasi-natural genetic drift. *Arch Toxicol 92*, 3487-3503. doi:10.1007/s00204-018-2326-5

Hair, K., Macleod, M. R., Sena, E. S. et al. (2019). A randomised controlled trial of an intervention to improve compliance with the ARRIVE guidelines (IICARus). *Res Integr Peer Rev 4*, 12. doi:10.1186/s41073-019-0069-3

Hansson, O., Castilho, R. F., Kaminski Schierle, G. S. et al. (2000). Additive effects of caspase inhibitor and lazaroid on the survival of transplanted rat and human embryonic dopamine neurons. *Exp Neurol 164*, 102-111. doi:10.1006/exnr.2000.7406

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. doi:10.1177/026119290403200503

Hartung, T. (2007). Food for thought ... On validation. *ALTEX 24*, 67-80. doi:10.14573/altex.2007.2.67

Hartung, T. and Leist, M. (2008). Food for thought ... On the evolution of toxicology and the phasing out of animal testing. *ALTEX 25*, 91-102. doi:10.14573/altex.2008.2.91

Hartung, T., De Vries, R., Hoffmann, S. et al. (2019). Toward good in vitro reporting standards. *ALTEX 36*, 3-17. doi:10.14573/altex.1812191

Hirt, U. A., Gantner, F. and Leist, M. (2000). Phagocytosis of nonapoptotic cells dying by caspase-independent mechanisms. *J Immunol 164*, 6520-6529. doi:10.4049/jimmunol.164.12.6520

Hoffmann, S. and Hartung, T. (2006). Designing validation studies more efficiently according to the modular approach: Retrospective analysis of the Episkin test for skin corrosion. *Altern Lab Anim 34*, 177-191. doi:10.1177/026119290603400209

Horbach, S. and Halffman, W. (2017). The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. *PLoS One 12*, e0186281. doi:10.1371/journal.pone.0186281

Ingre-Khans, E., Agerstrand, M., Beronius, A. et al. (2019). Toxicity studies used in registration, evaluation, authorisation and restriction of chemicals (REACH): How accurately are they reported? *Integr Environ Assess Manag 15*, 458-469. doi:10.1002/ieam.4123

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspect Psychol Sci 7*, 645-654. doi:10.1177/1745691612464056

Kinsner-Ovaskainen, A., Rzepka, R., Rudowski, R. et al. (2009). Acutoxbase, an innovative database for in vitro acute toxicity studies. *Toxicol In Vitro 23*, 476-485. doi:10.1016/j.tiv.2008.12.019

Kisitu, J., Hougaard Bennekou, S. and Leist, M. (2019). Chemical concentrations in cell culture compartments (C5) – Concentration definitions. *ALTEX 36*, 154-160. doi:10.14573/altex.1901031

Kleensang, A., Vantangoli, M. M., Odwin-DaCosta, S. et al. (2016). Genetic variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function. *Sci Rep 6*, 28994. doi:10.1038/srep28994

Krebs, A., Nyffeler, J., Rahnenfuhrer, J. et al. (2018). Normalization of data for viability and relative cell function curves. *ALTEX 35*, 268-271. doi:10.14573/1803231

Kruglikov, R. I., Getsova, V. M. and Uniial, M. (1976). (Effect of an excess of serotonin in the brain on consolidation of temporary connections). *Zh Vyssh Nerv Deiat Im I P Pavlova 26*, 1208-1213.

Latta, M., Kunstle, G., Leist, M. et al. (2000). Metabolic depletion of ATP by fructose inversely controls CD95- and tumor necrosis factor receptor 1-mediated hepatic apoptosis. *J Exp Med 191*, 1975-1985. doi:10.1084/jem.191.11.1975

Leist, M., Maurer, S., Schultz, M. et al. (1999). Cytoprotection against lipid hydroperoxides correlates with increased glutathione peroxidase activities, but not selenium uptake from different selenocompounds. *Biol Trace Elem Res 68*, 159-174. doi:10.1007/bf02784404

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, N., Daneshian, M. et al. (2012). Validation and quality control of replacement alternatives – Current status and future challenges. *Toxicol Res 1*, 8-22. doi:10.1039/c2tx20011b

Leist, M., Hasiwa, N., Rovida, C. et al. (2014). Consensus report on the future of animal-free systemic toxicity testing. *ALTEX 31*, 341-356. doi:10.14573/altex.1406091

Leist, M. and Hengstler, J. G. (2018). Essential components of methods papers. *ALTEX 35*, 429-432. doi:10.14573/altex.1807031

Liu, Y., Mi, Y., Mueller, T. et al. (2019). Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat Biotechnol 37*, 314-322. doi:10.1038/s41587-019-0037-y

Lorge, E., Moore, M. M., Clements, J. et al. (2016). Standardized cell sources and recommendations for good cell culture practices in genotoxicity testing. *Mutat Res 809*, 1-15. doi:10.1016/j.mrgentox.2016.08.001

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

Masters, J. (2002). Re: False cell lines. *Exp Cell Res 272*, 216. doi:10.1006/excr.2001.5439

Matsuda, K. (2017). PCR-based detection methods for single-nucleotide polymorphism or mutation: Real-time PCR and its substantial contribution toward technological refinement. *Adv Clin Chem 80*, 45-72. doi:10.1016/bs.acc.2016.11.002

Nardone, R. M. (2007). Eradication of cross-contaminated cell lines: A call for action. *Cell Biol Toxicol 23*, 367-372. doi:10.1007/s10565-007-9019-9

Nardone, R. M. (2008). Curbing rampant cross-contamination and misidentification of cell lines. *Biotechniques 45*, 221-227. doi:10.2144/000112925

OECD (2002). Test No. 423: Acute Oral Toxicity – Acute Toxic Class Method. *OECD Guidelines for the Testing of Chemicals, Section 4*. OECD Publishing, Paris. doi:10.1787/9789264071001-en

OECD (2005). *Good Laboratory Practice (GLP): OECD Principles and Guidance for Compliance Monitoring*. OECD Publishing, Paris. doi:10.1787/9789264012837-en

OECD (2017). Guidance Document for Describing Non-Guideline In Vitro Test Methods. *Series on Testing and Assessment No. 211*. OECD Publishing, Paris. doi:10.1787/9789264274730-en

OECD (2018). *Guidance Document on Good In Vitro Method Practices (GIVIMP)*. OECD Publishing, Paris. doi:10.1787/9789264304796-1-en

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

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

Prinz, F., Schlange, T. and Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nat Rev Drug Discov 10*, 712. doi:10.1038/nrd3439-c1

Prior, H., Casey, W., Kimber, I. et al. (2019). Reflections on the progress towards non-animal methods for acute toxicity testing of chemicals. *Regul Toxicol Pharmacol 102*, 30-33. doi:10.1016/j.yrtph.2018.12.008

Ramaiahgari, S. C., den Braver, M. W., Herpers, B. et al. (2014). A 3D in vitro model of differentiated HepG2 cell spheroids with improved liver-like properties for repeated dose high-throughput toxicity studies. *Arch Toxicol 88*, 1083-1095. doi:10.1007/s00204-014-1215-9

Rovida, C., Vivier, M., Garthoff, B. et al. (2014). ESNATS conference – The use of human embryonic stem cells for novel toxicity testing approaches. *Altern Lab Anim 42*, 97-113. doi:10.1177/026119291404200203

Schildknecht, S., Pape, R., Muller, N. et al. (2011). Neuroprotection by minocycline caused by direct and specific scavenging of peroxynitrite. *J Biol Chem 286*, 4991-5002. doi:10.1074/jbc.m110.169565

Schildknecht, S., Karreman, C., Poltl, D. et al. (2013). Generation of genetically-modified human differentiated cells for toxicological tests and the study of neurodegenerative diseases. *ALTEX 30*, 427-444. doi:10.14573/altex.2013.4.427

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 91*, 1-33. doi:10.1007/s00204-016-1805-9

Stacey, G. N., Bolton, B. J. and Doyle, A. (1992). DNA fingerprinting transforms the art of cell authentication. *Nature 357*, 261-262. doi:10.1038/357261a0

Stacey, G. N. (2000). Cell contamination leads to inaccurate data: We must take action now. *Nature 403*, 356. doi:10.1038/35000394

Stiegler, N. V., Krug, A. K., Matt, F. et al. (2011). Assessment of chemical-induced impairment of human neurite outgrowth by multiparametric live cell imaging in high-density cultures. *Toxicol Sci 121*, 73-87. doi:10.1093/toxsci/kfr034

Vogt, L., Reichlin, T. S., Nathues, C. et al. (2016). Authorization of animal experiments is based on confidence rather than evidence of scientific rigor. *PLoS Biol 14*, e2000598. doi:10.1371/journal.pbio.2000598

Zhang, C., Cerveira, E., Romanovitch, M. et al. (2017). Array-based comparative genomic hybridization (aCGH). *Methods Mol Biol 1541*, 167-179. doi:10.1007/978-1-4939-6703-2_15

Zimmer, B., Kuegler, P. B., Baudis, B. et al. (2011). Coordinated waves of gene expression during neuronal differentiation of embryonic stem cells as basis for novel approaches to developmental neurotoxicity testing. *Cell Death Differ 18*, 383-395. doi:10.1038/cdd.2010.109

Zimmer, B., Lee, G., Balmer, N. V. et al. (2012). Evaluation of developmental toxicants and signaling pathways in a functional test based on the migration of human neural crest cells. *Environ Health Perspect 120*, 1116-1122. doi:10.1289/ehp.1104489

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by the Doerenkamp-Zbinden foundation, the Land-BW (INVITE), the BMBF (e:ToP program, SysBioTop), grants by EFSA and DK-EPA, Estonian Research Council grant PUT1015 as well as the projects from the European Union's Horizon 2020 research and innovation programme EU-ToxRisk (grant agreement No 681002), ENDpoiNTs (grant agreement No 825759), and the project CERST (Center for Alternatives to Animal Testing) of the Ministry for Culture and Science of the State of North-Rhine Westphalia, Germany [233-1.08.03.03-121972]. We are grateful to many contributors not listed as authors. Especially a whole team of experts from EURL-ECVAM gave important advice.