Clinical relevance assessment of animal preclinical research (RAA) tool: Development and explanation

Kurinchi S Gurusamy [Corresp., 1, 2], **David Moher** [3, 4], **Marilena Loizidou** [1], **Irfan Ahmed** [5], **Marc Avey** [3, 4], **Carly Barron** [3, 4, 6], **Brian Davidson** [1], **Miriam Dwek** [7], **Christian Gluud** [8], **Gavin Jell** [1], **Kiran Katakam** [8], **Joshua Montroy** [9], **Timothy D McHugh** [10], **Nicola Osborne** [11], **Merel Ritskes-Hoitinga** [12], **Kees van Laarhoven** [13], **Jan Vollert** [14, 15], **Manoj Lalu** [9]

[1] Research Department of Surgical Biotechnology, University College London, London, England, United Kingdom

[2] Surgery and Interventional Trials Unit, University College London, London, England, United Kingdom

[3] Centre for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, The Ottawa Hospital, Ottawa, Canada

[4] School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canada

[5] Department of Surgery, NHS Grampian, Aberdeen, Scotland, United Kingdom

[6] Department of Medicine, McMaster University, Hamilton, Canada

[7] School of Life Sciences, University of Westminster, London, England, United Kingdom

[8] Copenhagen Trial Unit, Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copehagen, Denmark

[9] Department of Anesthesiology and Pain Medicine, Blueprint Translational Research Group, Clinical Epidemiology and Regenerative Medicine Programs, Ottawa Hospital Research Institute, Ottawa Hospital, Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Canada

[10] UCL Centre for Clinical Microbiology, Division of Infection & Immunity, University College London, London, England, United Kingdom

[11] Responsible Research in Practice, London, England, United Kingdom

[12] SYRCLE, Department for Health Evidence, Radboud University Medical Center, Nijmegen, Netherlands

[13] Department of Surgery, Radboud Institute for Health Sciences, Nijmegen, Netherlands

[14] Pain Research, Department of Surgery & Cancer, Imperial College, London, England, United Kingdom

[15] Center of Biomedicine and Medical Technology Mannheim CBTM, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

Corresponding Author: Kurinchi S Gurusamy
Email address: k.gurusamy@ucl.ac.uk

Background. Only a small proportion of preclinical research (research performed in animal models prior to clinical trials in humans) translates into clinical benefit in humans. Possible reasons for the lack of translation of the results observed in preclinical research into human clinical benefit include the design, conduct, and reporting of preclinical studies. There is currently no formal domain-based assessment of the clinical relevance of preclinical research. To address this issue, we have developed a tool for the assessment of the clinical relevance of preclinical studies, with the intention of assessing the likelihood that therapeutic preclinical findings can be translated into improvement in the management of human diseases.

Methods. We searched the EQUATOR network for guidelines that describe the design, conduct, and reporting of preclinical research. We searched the references of these guidelines to identify further relevant publications and developed a set of domains and signalling questions. We then conducted a modified Delphi-consensus to refine and develop the tool. The Delphi panel members included specialists in evidence-based (preclinical) medicine specialists, methodologists, preclinical animal researchers, a veterinarian, and clinical researchers. A total of 20 Delphi-panel members completed the first round and 17 members from 5 countries completed all three rounds.

Results. This tool has eight domains (construct validity, external validity, risk of bias, experimental

design and data analysis plan, reproducibility and replicability of methods and results in the same model, research integrity, and research transparency) and a total of 28 signalling questions and provides a framework for researchers, journal editors, grant funders, and regulatory authorities to assess the potential clinical relevance of preclinical animal research.

Conclusion. We have developed a tool to assess the clinical relevance of preclinical studies. This tool is currently being piloted.

Clinical Relevance Assessment of Animal preclinical research (RAA) tool: Development and Explanation

Kurinchi S Gurusamy[1,2], David Moher[3,4], Marilena Loizidou[1], Irfan Ahmed[5], Marc Avey[3,4], Carly Barron[3,4,6], Brian Davidson[1], Miriam Dwek[7], Christian Gluud[8], Gavin Jell[1], Kiran Katakam[8], Joshua Montroy[9], Timothy D McHugh[10], Nicola Osborne[11], Merel Ritskes-Hoitinga[12], Kees van Laarhoven[13], Jan Vollert[14,15], Manoj Lalu[9]

[1] Research Department of Surgical Biotechnology, University College London, London, England, United Kingdom

[2] Surgery and Interventional Trials Unit, University College London, London, England, United Kingdom

[3] Centre for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, The Ottawa Hospital, Ottawa, Canada

[4] School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canada

[5] Department of Surgery, NHS Grampian, Aberdeen, Scotland

[6] Department of Medicine, McMaster University, Hamilton, Canada

[7] School of Life Sciences, University of Westminster, London, England, United Kingdom

[8] Copenhagen Trial Unit, Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark

[9] Department of Anesthesiology and Pain Medicine, Blueprint Translational Research Group, Clinical Epidemiology and Regenerative Medicine Programs, Ottawa Hospital Research Institute, Ottawa Hospital, Department of Cellular and Molecular Medicine, Ottawa, University of Ottawa, Canada

[10] UCL Centre for Clinical Microbiology, Division of Infection & Immunity, University College London, London, England, United Kingdom

[11] Responsible Research in Practice, London, England, United Kingdom

[12] SYRCLE, Department for Health Evidence, Radboud University Medical Center, Nijmegen, Netherlands

[13] Department of Surgery, Radboud Institute for Health Sciences, Nijmegen, Netherlands

[14] Pain Research, Department of Surgery & Cancer, Imperial College, London, England, United Kingdom

[15] Center of Biomedicine and Medical Technology Mannheim CBTM, Medical Faculty Mannheim, Heidelberg University, Germany

Corresponding author: Professor Kurinchi Gurusamy, 9th floor, Royal Free Hospital, Royal Free Campus, University College London, Rowland Hill Street, London. NW3 2PF. Email: k.gurusamy@ucl.ac.uk

Emails

Kurinchi S Gurusamy: k.gurusamy@ucl.ac.uk

David Moher: dmoher@ohri.ca (ORCID 0000-0003-2434-4206)

Marilena Loizidou: m.loizidou@ucl.ac.uk

Irfan Ahmed: irfanahmed2@nhs.net

Marc Avey: marc.avey@ICF.com

Carly Barron: carly.c.barron@gmail.com

Brian Davidson: b.davidson@ucl.ac.uk

Miriam Dwek: M.V.Dwek@westminster.ac.uk

Christian Gluud: cgluud@ctu.dk

Gavin Jell: g.jell@ucl.ac.uk

Kiran Katakam: kirandenmark@gmail.com

Joshua Montroy: jmontroy@ohri.ca

Timothy D McHugh: t.mchugh@ucl.ac.uk

Nicola Osborne: nikki@responsibleresearchinpractice.co.uk

Merel Ritskes-Hoitinga: Merel.Ritskes-Hoitinga@radboudumc.nl

53 Kees van Laarhoven: Kees.vanLaarhoven@radboudumc.nl

54 Jan Vollert: j.vollert@imperial.ac.uk

55 Manoj Lalu: mlalu@toh.ca (ORCID 0000-0002- 0322-382X)

56

Abstract

Background. Only a small proportion of preclinical research (research performed in animal models prior to clinical trials in humans) translates into clinical benefit in humans. Possible reasons for the lack of translation of the results observed in preclinical research into human clinical benefit include the design, conduct, and reporting of preclinical studies. There is currently no formal domain-based assessment of the clinical relevance of preclinical research. To address this issue, we have developed a tool for the assessment of the clinical relevance of preclinical studies, with the intention of assessing the likelihood that therapeutic preclinical findings can be translated into improvement in the management of human diseases.

Methods. We searched the EQUATOR network for guidelines that describe the design, conduct, and reporting of preclinical research. We searched the references of these guidelines to identify further relevant publications and developed a set of domains and signalling questions. We then conducted a modified Delphi-consensus to refine and develop the tool. The Delphi panel members included specialists in evidence-based (preclinical) medicine specialists, methodologists, preclinical animal researchers, a veterinarian, and clinical researchers. A total of 20 Delphi-panel members completed the first round and 17 members from 5 countries completed all three rounds.

Results. This tool has eight domains (construct validity, external validity, risk of bias, experimental design and data analysis plan, reproducibility and replicability of methods and results in the same model, research integrity, and research transparency) and a total of 28 signalling questions and provides a framework for researchers, journal editors, grant funders, and regulatory authorities to assess the potential clinical relevance of preclinical animal research.

Conclusion. We have developed a tool to assess the clinical relevance of preclinical studies. This tool is currently being piloted.

Introduction

Only a small proportion of preclinical research (research performed on animals prior to clinical trials) translates into clinical benefit in humans. In a study evaluating the translation of preclinical research into clinical benefit, a total of 101 technologies (including drugs, devices, and gene therapy) were assessed in preclinical models and considered to be promising. Of these, 27 (27%) were subsequently tested in human randomised clinical trials within 20 years of their preclinical publication. Of these 27 human translational attempts, only one technology resulted in clinical benefit (1%; 95% confidence interval 0.2% to 5.4%) (Contopoulos-Ioannidis, Ntzani et al. 2003). In a 2014 report, of 100 potential drugs giving objective improvement when evaluated in a commonly used mouse model for the treatment of amyotrophic lateral sclerosis, none were found to be clinically beneficial (Perrin 2014). Finally, a systematic review found that there were significant differences in the estimates of treatment effectiveness in animal experiments compared to that observed in human randomised controlled trials with some interventions being beneficial in animals but harmful in humans (Perel, Roberts et al. 2007). Some of the reasons for the lack of translation of the beneficial results observed in preclinical research into human clinical benefit could relate to the design, conduct, and reporting of preclinical studies (Collins and Tabak 2014, Begley and Ioannidis 2015, Ioannidis 2017). Further information of the reasons and explanations for the lack of translation of the beneficial results observed in preclinical research into human clinical benefit is provided under the explanations for the relevant domains and signalling questions.

Why is this project needed?

A domain-based tool is a tool that assesses different aspects that impact the outcome of interest (in this case, clinical relevance of preclinical research). Such domain-based tools are preferred by methodologists to assess clinical studies (Higgins, Green et al. 2011, Whiting, Rutjes et al. 2011, Sterne, Hernan et al. 2016, Whiting, Savovic et al. 2016); however, as indicated below, no such tool exists to assess the potential clinical relevance of preclinical research.

Aim of this project

The aim of this project was to design a domain-based tool to assess the clinical relevance of a preclinical research study in terms of the likelihood that therapeutic preclinical findings can be translated into improvement in the management of human diseases. As part of the process, the scope and applicability of this tool was defined to include only in vivo animal interventional studies.

Who is this intended for?

This tool is intended for all preclinical researchers and clinical researchers considering translation of preclinical findings to first-in-human clinical trials, the funders of such studies, and regulatory agencies that approve first-in-human studies.

Materials & Methods

We followed the Guidance for Developers of Health Research Reporting Guidelines (Moher, Schulz et al. 2010) as there is no specific guidance for developers of tools to assess clinical relevance of preclinical tools. The registered protocol is available at http://doi.org/10.5281/zenodo.1117636 (Zenodo registration: 1117636). The study did not start until the protocol for the current study was registered. The overall process is summarised in Figure 1.

Search methods

First, we established whether there is any domain-based assessment tool for preclinical research. We searched the EQUATOR Network's library of reporting guidelines using the terms 'animal' or 'preclinical' or 'pre-clinical'. We included any guidelines or tools that described the design, conduct, and reporting of preclinical research. We searched the references of these guidelines to identify further relevant publications. We searched only the EQUATOR Network's library as it contains a comprehensive search of the existing reporting guidelines. A scoping search of Pubmed using the terms 'animal[tiab] AND (design[tiab] OR conduct[tiab] OR report[tiab])' returned nearly 50,000 records and initial searching of the first 1000 of them did not indicate any relevant publications. Therefore, the more efficient strategy of searching the EQUATOR Network's library was used to find any publications of a domain-based tool related to design, conduct, or reporting guidelines of preclinical research.

Development of domains and signalling questions

We recorded the topics covered in the previous guidance on preclinical research to develop a list of domains and signalling questions to be included in the formal domain-based assessment of preclinical research. The first author identified and included all the topics covered in each of the publications and combined similar concepts. The initial signalling questions were developed after preliminary discussions with and comments from the all the Delphi panel members (please see below) prior to finalising the initial list of signalling questions. The full list of the topics covered in the publications and the initial signalling questions are available in the supplementary information Appendix 1 (second column). The signalling questions are questions that help in the assessment of a domain. Additional details about how domains and signalling questions can be used are listed in Box 1.

Box 1

1. Signalling questions are questions that help in the assessment about a domain. As such, the overall domain assessment is more important that the answers for individual signalling questions.
2. Depending upon the nature and purpose for the research, certain domains may be more important than the other. For example, if the purpose is to find out whether there is enough information to perform a first-in-human study, the clinical translatability and reproducibility domain is of greater importance than if the report was about the first interventional study on a newly developed experimental model.

Selection of experts and consensus

Next, we approached experts in the field of preclinical and clinical research to participate in the development process. The group of experts were purposively sampled using snowballing principles (used to identify people with a rich knowledge base on a topic) (Heckathorn 2011): people who perform only preclinical research, people who perform only clinical research, people who perform both preclinical and clinical research, and methodologists, all of whom had interest in improving the clinical relevance of preclinical research were approached and asked to suggest other experts who could contribute to the process. We conducted a modified Delphi-consensus method to refine and develop the tool. The Delphi-consensus method was based on that described by Jones et al. (Jones and Hunter 1995). The steps in the Delphi process is shown in box 2. All were completed electronically using an excel file.

Box 2

1. The first round included questions regarding scope and necessity (i.e. should the tool include all types of preclinical research or only preclinical in vivo animal research and whether a domain or signalling question should be included in the final tool) in addition to the signalling questions available in the second column of Appendix 1.
2. The signalling questions were already classified into domains and were supported by explanations and examples in the first Delphi round. The original classification of signalling questions is available in Appendix 1.
3. The Delphi panel ranked the questions by importance on a scale of 1 to 9 with 1 being of lowest importance and 9 being highest importance.
4. The ranking scores were then grouped into three categories: 1 to 3 being strong disagreement about the importance of the question, 4 to 6 being weak to moderate agreement about the question, and 7 to 9 being strong agreement about the question. The questions were phrased in such a way that higher scores supported inclusion into the tool and lower scores indicated exclusion (of the scope, domain, or signalling question). Consensus was considered to have been reached when 70% or more participants scored 7

or more. There is variability in the definition of consensus and 70% or more participants scoring 7 or more is within the previously reported range for consensus agreement (Sumsion 1998, Hasson, Keeney et al. 2000, Diamond, Grant et al. 2014). This is a commonly used percentage for defining consensus in the Delphi process (Kleynen, Braun et al. 2014, Kirkham, Davis et al. 2017).

5. A total of three rounds were conducted. The panel members were allowed to add new domains or signalling questions in the first round. The panel members could also suggest revisions to the existing domains or questions (for example, revision of the explanation, examples, or by combining some domains or questions and splitting others) in all the rounds.
6. After the first round, the Delphi panel were shown their previous rank for the question and the median rank (and interquartile range) of questions of the Delphi panel. In addition, the Delphi panel were also asked to choose the best version of any revisions to the questions and provide ranks for any additional questions identified in the first round.
7. The panel members were able to retain or change the rank in each of the rounds after the first round.
8. For calculation of median and interquartile range of ranks and consensus, non-responses were ignored.
9. At the end of the third round, the aspects which have been ranked with a score of 7 or above for necessity by at least 70% of the panel were included in the final tool.
10. There was no restriction on the Delphi panel to consult others while ranking the questions. However, only one final response on the set of questions was accepted from each Delphi panel member.

Then, we refined the signalling questions and explanation by iterative electronic communications. Finally, we piloted the tool in biomedical researchers who perform animal preclinical research and those who perform first-in-human studies to clarify the signalling questions and explanations.

Results

Deviations from protocol

There were two deviations from our protocol. Firstly, we did not exclude questions even when consensus was reached on the necessity of the questions: this was because the phrasing of the domain/signalling question, the explanation, the domain under which the signalling question is located, and combining or splitting the domains were still being debated. Secondly, we did not conduct an online meeting of the panel members between the second and third rounds of the Delphi process because listing and summarising the comments from different contributors achieved the aim of providing information to justify or revise the ranking.

Search results

Twenty-one publications were identified (Idris, Becker et al. 1996, Sena, van der Worp et al. 2007, Bath, Macleod et al. 2009, Fisher, Feuerstein et al. 2009, Macleod, Fisher et al. 2009, Bouxsein, Boyd et al. 2010, Hooijmans, Leenaars et al. 2010, Kilkenny, Browne et al. 2010, van der Worp, Howells et al. 2010, Begley and Ellis 2012, Landis, Amara et al. 2012, Hooijmans, Rovers et al. 2014, NIH 2014, Perrin 2014, Bramhall, Florez-Vargas et al. 2015, Czigany, Iwasaki et al. 2015, Andrews, Latremoliere et al. 2016, Biophysical Journal 2017, Open Science Framework 2017, Osborne, Avey et al. 2018, Smith, Clutton et al. 2018). The main topics covered in these publications were bias, random errors, reproducibility, reporting, or a mixture of these elements which result in lack of translation of preclinical research into clinical benefit. One publication was based on a consensus meeting (Andrews, Latremoliere et al. 2016) and five were based on expert working groups (Hooijmans, Leenaars et al. 2010, Kilkenny, Browne et al. 2010, Landis, Amara et al. 2012, NIH 2014, Osborne, Avey et al. 2018); and the remaining were opinions of the authors (Idris, Becker et al. 1996, Sena, van der Worp et al. 2007, Bath, Macleod et al. 2009, Fisher, Feuerstein et al. 2009, Macleod, Fisher et al. 2009, Bouxsein, Boyd et al. 2010, van der Worp, Howells et al. 2010, Begley and Ellis 2012, Hooijmans, Rovers et al. 2014, Perrin 2014, Bramhall, Florez-Vargas et al. 2015, Czigany, Iwasaki et al. 2015, Biophysical Journal 2017, Open Science Framework 2017, Smith, Clutton et al. 2018). All five publications based on consensus meeting or expert working groups were reporting guidelines (Hooijmans, Leenaars et al. 2010, Kilkenny, Browne et al. 2010, Landis, Amara et al. 2012, NIH 2014, Osborne, Avey et al. 2018).

Survey respondents

A total of 20 Delphi-panel members completed the first round and 17 members from 5 countries completed all three rounds. The panel members included specialists representing a broad scope of stakeholders, including target users that would evaluate interventions for potential 'bench-to-bedside' translation: evidence-based (preclinical) medicine specialists, methodologists, preclinical researchers, veterinarian, and clinical researchers from UK, Canada, Denmark, and Netherlands. The mean and standard deviation age of people who completed was 48.4 and 10.9 at the time of registration of protocol. Of the 17 respondents completing all the three rounds, 12 were males and 5 were females; eleven of these 17 respondents were Professors or had equivalent senior academic grade at the time of registration. There were no conflicts of interest for the survey respondents other than those listed in the 'Conflicts of interest' section of this document.

The reasons for drop-out included illness (one member) and concerns about the scope and applicability of the tool (two members). These were aspects that were developed as part of the process of the registered protocol. Therefore, clarity on the scope and applicability was available only at the end of the Delphi process and not in the first round of the Delphi-process.

Domains and signalling questions

The Delphi panel agreed on eight domains, which constitutes the tool. Table 1 lists the domains and signalling questions for which consensus agreement was reached. The first four domains relate to the study design and analysis that are within the control of the research team (clinical translatability of results to human disease or condition (construct validity), experimental design and data analysis, bias (internal validity), and reproducibility of results in a different disease-specific model (external validity). The fifth domain relates to replicability of results for which the research team may have to rely on other research teams (reproducibility and replicability of methods and results in the same model); however, these aspects can be integrated as part of the same study. The sixth domain relates to study conclusions which considers the study design, analysis, and reproducibility and replicability of results. The last two domains relate to factors that increase or decrease the confidence in the study findings (research integrity and research transparency).

These eight domains cover a total of 28 signalling questions. The number of questions in each domain range from 1 to 8, with a median of 3 questions in each domain. All the signalling questions have been phrased in such a way that a classification of 'yes' or 'probably yes' will result in low concerns about the clinical relevance of the study for the domain.

Scope and applicability of the tool

The scope of the tool is only for assessment of the clinical relevance of a preclinical research study in terms of the likelihood that therapeutic preclinical findings can be translated into improvement in the management of human diseases and not for assessment of the quality of the study, i.e. how well the study was conducted, although we refer to tools that assess how well the study was conducted. It is important to make this distinction as even a very well-designed and conducted preclinical study may not translate to improvement in the management of human diseases, as is the case of clinical research.

As part of the Delphi process, the scope was narrowed to include only *in vivo* laboratory based preclinical animal research evaluating interventions. Therefore, our tool is not intended for use on other forms of preclinical research such as *in vitro* work (e.g. cell cultures), *in silico* research, or veterinary research. This tool is not applicable in the initial exploratory phase of development of new animal models of disease, although the tool is applicable in interventional studies using such newly developed models.

The domains and signalling questions in each round of the Delphi process and post-Delphi process are summarised in Figure 2.

Classification of signalling questions and domains

Consistent with existing domain based tools , responses to each signalling question can be classified as 'yes', 'probably yes', 'probably no', 'no', or 'no information' (Sterne, Hernan et al. 2016, Whiting, Savovic et al. 2016), depending upon the information described in the report or after obtaining the relevant information from the report's corresponding author, although the study authors may provide answers that the assessor asks because of cognitive bias. A few questions can also be classified as 'not applicable'. These questions start with the phrase 'if'. For classification of the concerns in the domain, such questions are excluded from the analysis.

A domain can be classified as 'low concern' if ***all*** the signalling questions under the domain were classified as 'yes' or 'probably yes', 'high concern' if ***any*** of the signalling questions under the domain were classified as 'no' or 'probably no', and as 'moderate concern' for all other combinations.

Overall classification of the clinical relevance of the study

A study with 'low concerns' for all domains will be considered as a study with high clinical relevance in terms of translation of preclinical results with similar magnitude and direction of effect to improve management of human diseases. A study with unclear or high concerns for one or more domains will be considered as a study with uncertain clinical relevance in terms of translation of preclinical results with similar magnitude and direction of effect to improve management of human diseases.

However, depending upon the nature and purpose for use of the research, certain domains may be more important than the other and the users can decide in advance whether a particular domain is important (or not). For example, if the purpose is to find out whether there is enough information to perform a first-in-human study, the clinical translatability and reproducibility domain is of greater importance than if the report was about the first interventional study on the model.

At the design and conduct stage, researchers, funders, and other stakeholders can specifically look at the domains that are assessed as unclear or high concern and improve the design and conduct to increase the clinical relevance. At the reporting stage, researchers, funders, and other stakeholders can use this tool to design, fund, or give approval for further research.

Practical use of the tool

The tool should be used with a clinical question in mind. This should include the following aspects of the planned clinical study as a minimum: population in whom the intervention or diagnostic test is used, intervention and control, and the outcomes (PICO).

We recommend that the tool is used after successfully completing the training material, which includes examples of how the signalling questions can be answered and assessment of understanding the use of the tool (the training material is available at: https://doi.org/10.5281/zenodo.4159278) and at least two assessors using the tool independently.

A schema for the practical use of the tool is described in Figure 3.

Scoring

The tool has not been developed to obtain an overall score for clinical relevance assessment. Therefore, modifying the tool by assigning scores to individual signalling questions or domains is likely to be misleading.

Panel agreement

Appendix 1 summarises the Delphi panel agreement on the different domains and signalling questions. As shown in the Appendix 1, the domains, the signalling questions, and the terminologies used have improved significantly from the starting version of the tool. Appendix 1 also demonstrates that there was a change in the agreement in the questions indicating that the panel members were receptive to others' views while ranking the questions.

Rationale and explanation of domains and signalling questions

Domain 1: Clinical translatability of results to human disease or condition (construct validity)

The purpose of this domain is to assess whether statistically positive results in the reports of the preclinical animal research studies could result in clinical benefit. This evaluation focuses on both primary outcomes and secondary outcomes, or the 'main findings' if the reports do not explicitly declare primary and secondary outcomes.

1.1 Did the authors use a model that adequately represents the human disease?

This question assesses biological plausibility. We have used the term 'model' to refer to the animal model used as a substitute for human disease, for example, a mouse model of multiple myeloma. We have also used this term to refer to induction methods in animals in a non-diseased state that progress to a diseased state, for example, a rat model of behavioural alterations (forced swim test) mimicking depression (Yankelevitch-Yahav, Franko et al. 2015), or animals which have been exposed to a treatment even if they did not have any induced human disease, for example, a rabbit model of liver resection, a canine model of kidney transplantation. Studies have shown that animal researchers frequently use disease models that do not adequately

represent or relate to the human disease (de Vries, Buma et al. 2012, Sloff, de Vries et al. 2014, Sloff, Simaioforidis et al. 2014, Zeeff, Kunne et al. 2016).

Specific characteristics to consider include species and/or strain used, age, immune competence, and genetic composition as relevant. Other considerations include different methods of disease induction in the same or different species.

This signalling question considers whether the researchers reporting the results ('authors') have described on information such as characteristics of model, different methods of disease induction (if appropriate), and biological plausibility while choosing the model, and have the researchers provided evidence for the choice of animal model. The assessment of these questions may require subject content expertise.

1.2 Did the authors identify and characterise the model?

This question assesses whether after choosing the appropriate model (species, sex, genetic composition, age), the authors have performed studies to characterise the model. For example, sepsis is often induced through caecal ligation and puncture; however, the effects of this procedure can produce variable sepsis severity. Another example is when genes that induce disease may not be inherited reliably: the resulting disease manifestation could be variable and interventions may appear to be less effective or more effective than they actually are (Perrin 2014). Therefore, it is important to ensure that the genes that induce the disease are correctly identified and that such genes are inherited. Another example is when the authors want to use a knockout model to understand the mechanism of how an intervention works based on the assumption that the only difference between the knockout mice and the non-knockout mice is the knockout gene. However, the animals used may still contain the gene that was intended to be removed or the animals may have other genes introduced during the process of creating the knockout mice (Eisener-Dorman, Lawrence et al. 2009). Therefore, it is important to understand and characterise the baseline model prior to testing an experimental intervention.

1.3 Were the method and timing of the intervention in the specific model relevant to humans?

For pharmacological or biological interventions, this question refers to the dose and route of administration. For other types of interventions, such as surgery or device implementation, the question refers to whether the method used in the animal model is similar to that in humans.

For pharmacological interventions, there may be a therapeutic dose and route which is likely to be safe and effective in humans. It is unlikely the exact dose used in animals is studied in humans, at least in the initial human safety studies. Therefore, dose conversion is used in first-in-human studies. Simple practice guides and general guidance for dose conversion between animals and humans are available (FDA 2005, Nair and Jacob 2016, EMA 2017). However,

some researchers may use doses in animals at levels that would be toxic when extrapolated to humans and therefore unlikely to be used. Dose conversion guides (Nair and Jacob 2016) can help with the assessment of whether the dose used is likely to be toxic. The effectiveness of an intervention at such toxic doses is not relevant to humans. It is preferable to use the same route of administration for animal studies as planned in humans, since different routes may lead to different metabolic fate and toxicity of the drug.

For non-pharmacological interventions for which similar interventions have not been tested in humans, feasibility of use in humans should be considered. For example, thermal ablation is one of the treatment options for brain tumours. Ablation can, for example, also be achieved by irreversible electroporation, which involves passing high voltage electricity and has been attempted in human liver and pancreas (Ansari, Kristoffersson et al. 2017, Lyu, Wang et al. 2017). However, the zone affected by irreversible electroporation has not been characterised fully: treatment of human brain tumours using this technique can only be attempted when human studies confirm that there are no residual effects of high voltage electricity in the surrounding tissue (not requiring ablation). Until then, the testing of irreversible electroporation in animal models of brain tumours is unlikely to progress to human trials and will not be relevant to humans regardless of how effective it may be.

The intervention may also be effective only at a certain time point in the disease (i.e. 'therapeutic window'). It may not be possible to recognise and initiate treatment during the therapeutic window because of the delays in appearance of symptoms and diagnosis. Therefore, there is no rationale in performing preclinical animal studies in which the intervention cannot be initiated during the likely therapeutic window. Finally, the treatment may be initiated prior to induction of disease in animal models: this may not reflect the actual use of the drug in the human clinical situation.

1.4 If the study used a surrogate outcome, was there a clear and reproducible relationship between an intervention effect on the surrogate outcome (measured at the time chosen in the preclinical research) and that on the clinical outcome?

A 'surrogate outcome' is an outcome that is used as a substitute for another (more direct) outcome along the disease pathway. For example, in the clinical scenario, an improvement in CD4 count (surrogate outcome) leads to a decrease in mortality (clinical outcome) in people with human immune deficiency (HIV) (Bucher, Guyatt et al. 1999). The relationship between the effect of the intervention (a drug that improves the CD4 count) on the surrogate outcome (CD4 count) and a clinical outcome (mortality after HIV infection) should be high, should be shown in multiple studies, and should be independent of the type of intervention for a surrogate outcome to be valid (Bucher, Guyatt et al. 1999). This probably applies to preclinical research as well. For example, the relationship between the effect of an intervention (a cancer drug) on the surrogate outcome (apoptosis) and a clinical outcome or its animal equivalent (for example, mortality in

the animal model) should be high, shown in multiple studies and independent of the type of intervention for a surrogate outcome to be valid in the preclinical model.

If the surrogate outcome is the only pathway or the main pathway between the disease, intervention, and the clinical outcome (or its animal equivalent) (Figure 4), the surrogate outcome is likely to be a valid indirect surrogate outcome (Fleming and DeMets 1996). This, however, should be verified in clinical studies. For example, preclinical animal research studies may use gene or protein levels to determine whether an intervention is effective. If the gene (or protein) lies in the only pathway between the disease and animal equivalent of the clinical outcome, a change in expression, levels, or activity of the gene (or protein) is likely to result in an equivalent change in the animal equivalent of the clinical outcomes. To simplify this even further this signalling question can be simplified to the context in which it is used for example, "Is apoptosis at 24 hours (surrogate outcome) in the preclinical animal model correlated with improved survival in animals (animal equivalent of a clinical outcome)"? Another example of this signalling question simplified to the context of the research can be "Are aberrant crypt foci (surrogate outcome) in animal models correlated to colon cancer in these models (animal equivalent of a clinical outcome)"?

This signalling question assesses whether the authors have provided evidence for the relationship between surrogate outcome and the clinical outcome (or its animal equivalent). There is currently no guidance as to what a high level of association is in terms of determining the relationship between surrogate outcomes and the clinical outcomes (or its animal equivalent). Some suggestions are mentioned in Appendix 2.

1.5 If the study used a surrogate outcome, did previous experimental studies consistently demonstrate that change in surrogate outcome(s) by a treatment led to a comparable change in clinical outcomes?

This question aims to go further than the evaluation of association between surrogate outcome and the clinical outcome (or its animal equivalent). A simple association between a surrogate outcome and clinical outcome may be because the surrogate outcome may merely be a good predictor. For example, sodium fluoride caused more fractures despite increasing bone mineral density, even though, low bone mineral density is associated with increased fractures (Bucher, Guyatt et al. 1999). If a change in the surrogate outcome by a treatment results in a comparable change in the clinical outcome (or its animal equivalent), the surrogate outcome is likely to be a valid surrogate outcome (Figure 4). This change has to be consistent, i.e. most studies showing that a treatment results in a comparable improvement in the clinical outcome (or its animal equivalent). Note that it is possible that there may not a fully comparable change, for example, a 50% improvement in the surrogate outcome may result only in a 25% improvement in the animal equivalent of the clinical outcome. In such situations, it is possible to use the 'proportion explained' approach proposed by Freedman et al. (Freedman, Graubard et al. 1992), a concept

which was extended to randomised controlled trials and systematic reviews by Buyse et al. (Buyse, Molenberghs et al. 2000). This involves calculating the association between the effect estimates of the surrogate outcome and clinical outcome (or its animal equivalent) from the different trials or centres within a trial (Buyse, Molenberghs et al. 2000) (although, one can obtain a more reliable estimate of this association using individual participant data) (Tierney, Pignon et al. 2015).

Generally, few surrogate outcomes are validated substitutes for clinical outcomes: an example of a valid surrogate outcome is CD4 count in people with human immune deficiency (HIV) (Bucher, Guyatt et al. 1999). Even if an association exists between the surrogate outcome and the clinical outcome, failure to demonstrate that changes in surrogate outcome by a treatment led to changes in clinical outcome can have disastrous effects (Bucher, Guyatt et al. 1999, Yudkin, Lipska et al. 2011, Kim and Prasad 2015, Rupp and Zuckerman 2017) (Appendix 3).

1.6 Did a systematic review with or without meta-analysis demonstrate that the effect of an intervention or a similar intervention in animal model was similar to that in humans?

The best way to find consistent evidence to support or refute the validity of surrogate outcomes (covered in the previous signalling questions) and the comparability of the animal equivalent of the clinical outcomes to that in humans is by systematic reviews. For example, if an intervention results in better functional recovery in a mouse model of stroke, then does it also result in better functional recovery in humans with stroke? If so, other interventions can be tested in this model. Systematic reviews help in calculating the association between the effect estimates of the surrogate outcome and clinical outcome (or its animal equivalent) from the different trials or centres within a trial, as mentioned previously (Buyse, Molenberghs et al. 2000).

Failure to conduct a systematic review of preclinical studies prior to the start of the clinical research and presenting selective results to grant funders or patients is scientifically questionable, likely to be unethical, and can lead to delays in finding suitable treatments for diseases by investing resources in treatments that could have been predicted to fail (Cohen 2018, Ritskes-Hoitinga and Wever 2018). Therefore, this signalling question assesses whether the authors provide evidence from systematic reviews of preclinical animal research studies and clinical studies that the intervention or a similar intervention showed treatment effects that were similar in preclinical research studies and clinical studies in humans.

Domain 2: Experimental design and data analysis plan

The purpose of this domain is to assess the experimental study design and assess the analysis performed by the authors with respect to random errors and measurement errors. There are very good online resources that can help with the experimental design and statistical analysis in

preclinical studies (Bate and Clark 2014, Festing 2016, Nature Collection 2018). These resources can help in the assessment of this domain.

2.1 Did the authors describe sample size calculations?

Sample size calculations are performed to control for random errors (i.e. ensure that a difference of interest can be observed) and should be used in preclinical studies that involve hypothesis testing (for example, a study conducted to find out whether a treatment is likely to result in benefit). This signalling question assesses whether the authors have described the sample size calculations to justify the number of animals used to reliably answer the research question.

2.2 Did the authors plan and perform statistical tests taking the type of data, the distribution of data, and the number of groups into account?

The statistical tests that are performed depend upon the type of data (for example, categorical nominal data, ordinal data, continuous quantitative data, continuous discrete data), distribution of data (for example, normal distribution, binomial distribution, Poisson distribution, etc.), and the number of groups compared. The authors should justify the use of statistical tests based on the above factors. The hypothesis testing should be pre-planned. This signalling question assesses whether the authors planned and performed statistical tests taking type of data, distribution of data, and the number of groups compared into account.

The authors may use multivariable analysis (analysis involving more than one predictor variable) or multivariate analysis (analysis involving more than one outcome variable), although these terms are often used interchangeably (Hidalgo and Goodman 2013). Some assumptions about the data are made when multivariable analysis and multivariate analysis are performed (Casson and Farmer 2014, Nørskov, Lange et al. 2020) and the results are reliable only when these assumptions are met. Therefore, assessment of whether the authors have reported about the assumptions should be considered as a part of this signalling question.

The authors may have also performed unplanned hypothesis testing after the data becomes available, which is a form of ‘data dredging’ and can be assessed in the next signalling question. The authors may also have made other changes to the statistical plan. This aspect can be assessed as part of signalling question 8.2.

2.3 Did the authors make adjustment for multiple hypothesis testing?

This signalling question assesses whether study authors have made statistical plans to account for multiple testing.

When multiple hypotheses are tested in the same research, statistical adjustments are necessary to achieve the planned alpha and beta errors. Testing for more than two groups is a form of

multiple testing: the statistical output usually adjusts for more than two groups. However, testing many outcomes is not usually adjusted in the statistical software output and has to be adjusted manually (or electronically) using some form of correction. This is not necessary when the study authors have a single primary outcome and base their conclusions on the observations on the single primary outcome. However, when multiple primary outcomes are used, adjustments for multiple hypothesis testing should be considered (Streiner 2015). For example, if the effectiveness of a drug against cancer is tested by apoptosis, cell proliferation, and metastatic potential, authors should consider statistical adjustments for multiple testing.

Multiple analyses of the data with the aim of stopping the study once statistical significance is reached and data dredging (multiple unplanned subgroup analyses to identify an analysis that is statistically significant; other names include 'P value fiddling' or 'P-hacking') are other forms of multiple testing and should be avoided (Streiner 2015). Methods for interim analysis to guide stopping of clinical trials such as sequential and group sequential boundaries have been developed (Grant, Altman et al. 2005). Implementation of group sequential designs may improve the efficiency of animal research (Neumann, Grittner et al. 2017).

2.4 If a dose-response analysis was conducted, did the authors describe the results?

In pharmacological testing in animals, it is usually possible to test multiple doses of a drug. This may also apply to some non-pharmacological interventions, where one can test the intervention at multiple frequencies or duration (for example, exercise for 20 minutes versus exercise for 10 minutes versus no exercise). A dose-response relationship indicates that the effect observed is greater with an increase in the dose. Animal studies incorporating dose-response gradients were more likely to be replicable to humans (Hackam and Redelmeier 2006). This signalling question assesses whether the authors have reported the dose-response analysis if it was conducted.

2.5 Did the authors assess and report accuracy?

Accuracy is the nearness of the observed value (using the method described) to the true value. Depending upon the type of outcome, these can be assessed by Kappa statistics, Bland-Altman method, correlation coefficient, concordance correlation coefficient, standard deviation, or relative standard deviation (Bland and Altman 1986, Bland and Altman 1996, Bland and Altman 1996, Bland and Altman 1996, van Stralen, Jager et al. 2008, Watson and Petrie 2010, Zaki, Bulgiba et al. 2012). This signalling question assesses whether the authors have provided a measure of accuracy by using an equipment for which accuracy information is available, or used a reference material (material with known values measured by an accurate equipment) to assess accuracy.

2.6 Did the authors assess and report precision?

Precision, in the context of measurement error, is the nearness of values when repeated measurements are made in the same sample (technical replicates). The same methods used for assessing accuracy can be used for assessing precision, except that instead of using a reference material, the comparison is between the measurements made in the same sample for assessing precision. The width of confidence intervals can also provide a measure of the precision. This signalling question assesses whether the authors have measured and reported precision.

2.7 Did the authors assess and report sampling error?

In some situations, errors arise because of the non-homogenous nature of the tissues or change of values over time, for example, diurnal variation. The same methods used to assess accuracy can be used for assessing sampling error, except that instead of using a reference material, the comparison is between the measurements made in samples from different parts of cancer/diseased tissue (biological replicates) or samples from different times. This signalling question assesses whether the authors have measured and reported sampling error.

2.8 Was the measurement error low or was the measurement error adjusted in statistical analysis?

This signalling question assesses whether the measurement errors (errors in one or more of accuracy, precision, sampling error) were low or were reported as adjusted in statistical analysis. There are currently no universally agreed values at which measurement errors can be considered low. This will depend upon the context and the measure used to assess measurement error. For example, if the differences between the groups is in cm and the measurement error is non-differential (i.e. the error does not depend upon the intervention) and is a fraction of a mm, then the measurement error is unlikely to cause a major difference in the conclusions. On the other hand, if the measurement error is differential (i.e. the measurement error depends upon the intervention) or large relative to the effect estimates, then this has to be estimated and adjusted during the analysis. Measurement error can be adjusted using special methods such as ANOVA repeated measurements, general linear model repeated measurements, regression calibration, moment reconstruction, or simulation extrapolation (Vasey and Thayer 1987, Carroll 1989, Lin and Carroll 1999, Littell, Pendergast et al. 2000, Freedman, Fainberg et al. 2004, Freedman, Midthune et al. 2008).

Domain 3: Bias (internal validity)

Even if an animal model with good construct validity is chosen, biases such as selection bias, confounding bias, performance bias, detection bias, and attrition bias can decrease the value of

the study (Higgins, Green et al. 2011). The purpose of this domain is to assess the risks of bias in the study.

3.1 Did the authors minimise the risks of bias such as selection bias, confounding bias, performance bias, detection bias, attrition bias, and selective outcome reporting bias?

Some sources, examples, and rationale for the risk of bias in animal studies are available in the SYRCLE's risk of bias assessment tool for animal research, National Research Council's guidance of description of animal research in scientific publications, US National Institute of Neurological Disorders and Stroke's call for transparent reporting, and National Institute of Health's principles and guidelines for Reporting Preclinical Research (National Research Council 2011, Landis, Amara et al. 2012, Hooijmans, Rovers et al. 2014, NIH 2014). These risks of bias should have been minimised in the study. While many researchers are familiar with most of these types of bias, selective outcome reporting warrants further discussion. Selective outcome reporting is a form of bias where study authors selectively report the results that favour the intervention. The selective outcome reporting bias should, as a minimum, cover whether the choice of results to be reported (in tables, text, or figures) were predetermined. Changing the outcomes is prevalent in human clinical trials (Jones, Keil et al. 2015, Altman, Moher et al. 2017, Howard, Scott et al. 2017). There are no studies that investigate the prevalence of changing the outcomes in preclinical animal research; however, one can expect that it is at least as prevalent in preclinical animal research as in clinical research. It is now also possible to register preclinical animal studies at www.preclinicaltrials.eu and www.osf.io before they start, which can help with the assessment of selective outcome reporting bias.

In some situations, International Organization for Standardization (ISO) standards (Chen and Wang 2018) and National Toxicology Program recommendations (2018) may also be applicable.

Domain 4: Reproducibility of results in a range of clinically relevant conditions (external validity)

The purpose of this domain is to assess whether the results were reproduced in a range of clinically relevant conditions (different methods of disease induction, different genetic composition, different ages, sex, etc).

4.1 Were the results reproduced with alternative preclinical models of the disease/condition being investigated?

The underlying rationale behind preclinical animal research is the genetic, anatomical, physiological, and biochemical similarities (one or more of the above) between animals and humans. Different animals have different levels of genetic similarities with humans and between

each other (Gibbs, Weinstock et al. 2004, Church, Goodstadt et al. 2009, Howe, Clark et al. 2013), which leads to anatomical, physiological, and biochemical differences between the different species. This can lead to differences in the treatment effects between different animal species or different models of induction of disease. The differences may be in the direction (for example, the intervention is beneficial is some species and harmful in others) or in the magnitude (for example, the intervention is beneficial in all the species, but the treatment effects differ in differ species). Even if the inconsistency is only in the magnitude of effect, this indicates that the treatment effects in humans may also be different from those observed in different species. Therefore, consistent treatment effects observed across different animal species or different models of induction of disease may increase the likelihood of similar treatment effects being observed in humans. This signalling question assesses the consistency across different preclinical models.

4.2 Were the results consistent across a range of clinically relevant variations in the model?

In the clinical setting, a treatment is used in people of different ages, sex, genetic composition, and with associated comorbidities. These differences within species and existing comorbidities can lead to different treatment effects even if the same species and the model of induction is used. Therefore, this signalling question assesses whether animals of multiple ages, sex, genetic compositions, and existing comorbidities were used and whether the treatment effect was consistent across these clinically relevant variations.

4.3 Did the authors report take existing evidence into account when choosing the comparators?

Researchers may choose an inactive control rather than an established active treatment as the control to show that a drug is effective. They may also choose a weak control such as a dose lower than the effective dose or an inappropriate route for control to demonstrate a benefit of the intervention. Therefore, in these last examples, experimental results that the intervention is better than control are applicable only for the comparison of the intervention with a weak control, which may not be clinically relevant. This signalling question assesses whether the authors chose an established active treatment at the correct dose or route (in the case of pharmacological interventions) as control.

Domain 5: Reproducibility and replicability of methods and results in the same model

In a survey of more than 1500 scientists conducted by Nature, more than 70% of researchers tried and failed to reproduce another scientist's experiments, and more than half failed to reproduce their own experiments (Baker 2016). About 90% of scientists surveyed thought that there was a slight or significant ‘reproducibility crisis’ (Baker 2016). This domain assesses the reproducibility (the ability to achieve similar or nearly identical results using comparable materials and methodologies) and replicability (the ability to repeat a prior result using the same

source materials and methodologies) (FASEB journal 2016) of the methods and results in the same animal model and differs from external validity, which focusses on whether the results were reproduced in a different clinically relevant model.

5.1 Did the authors describe the experimental protocols sufficiently to allow their replication?

One of the methods of improving replication is to describe the experimental protocols sufficiently. This signalling question assesses whether the authors have described the experimental protocols sufficiently to allow their replication.

5.2 Did an independent group of researchers replicate the experimental protocols?

This signalling question is different from the 5.1, above. The previous question assesses whether the protocols were described sufficiently, while this signalling question assesses whether these protocols were actually replicated by an independent group of researchers. The independent group of researchers could be part of the author team and could be from the same or different institutions, as long as they repeated the experiments independently. The results of replication of experimental protocols can be part of the same report, but could also be another report.

5.3 Did the authors or an independent group of researchers reproduce the results in similar and different laboratory conditions?

This signalling question is different from the 5.1, above. The previous question assesses whether the protocols were protocols could be replicated. This signalling questions assesses whether the results could be reproduced in similar and different laboratory conditions. Even when the protocols/methods are replicated by an independent group of researchers, the results may not be replicated or reproduced in similar and/or different laboratory conditions (Baker 2016). This signalling question assesses whether the results were replicated or reproduced. Attempts to replicate or reproduce the results can be a part of the same report, but could also be another report, particularly if the attempt to replicate or reproduce the results is made by an independent group of researchers.

Domain 6: Implications of the study findings (study conclusions)

The purpose of the domain is to assess whether the authors have made conclusions that reflect the study design and results.

6.1 Did the authors' conclusions represent the study findings, taking its limitations into account?

This signalling question assesses whether the study authors considered all the findings and limitations of the study while arriving at conclusions. The study authors may have made conclusions based on extrapolations of their results and not on their data, which is poor research

practice. This should also be considered while assessing this signalling question. Studies designed to look at pathophysiology of disease or mechanism of action of treatment should demonstrate evidence of similarity between disease process in animal and human disease before arriving at conclusions regarding these aspects.

6.2 Did the authors provide details on additional research required to conduct first-in-human studies?

Researchers should consider the limitations of their study before recommending first-in-human studies. For example, this may be the first experimental study on this research question; therefore, the research question may not have been conducted in multiple centres. The authors should highlight the need for studies that reproduce the results by a different group of researchers. If the current study was a study to attempt reproduction of the results of a previous study, then the authors should clarify whether further preclinical studies are required or whether the intervention should be evaluated in humans with justifications: repeating the study in preclinical models can be justified if the intervention needs to evaluated after a modification; recommending evaluation in humans can be justified if efficacy and safety has been demonstrated consistently in multiple preclinical models and centres.

This signalling question assesses whether the study authors have made future research recommendations based on the study design and results from this study in the context of other studies on this issue. A study on investigator brochures in Germany demonstrated that animal study results were not evaluated well, for example, by systematic reviews of animal studies before clinical trials (Wieschowski, Chin et al. 2018), highlighting that the further research recommendations should be made taking other studies on the topic into account.

Domain 7: Research integrity

The purpose of this domain is to ensure that the authors adhered to the principles of research integrity during the design, conduct and reporting of their research. If the authors did not adhere to the principles of research integrity, the results can be unreliable even if the study experimental design and analysis were reliable. Lack of research integrity can decrease the confidence in the study findings.

7.1 Did the research team obtain ethical approvals and any other regulatory approvals required to perform the research prior to the start of the study?

Animal research should be performed ethically in a humane way. While university ethics boards can confirm the existence of ethical approval, additional licensing requirements (for example, Home Office License in UK) may be necessary before the research can be conducted. This is to ensure that the principles of replacement (methods which avoid or replace the use of animals),

reduction (methods which minimise the number of animals used per experiment), and refinement (methods which minimise animal suffering and improve welfare) are followed during scientific research (NC3Rs , UK Government 1986). In some countries like the UK, preclinical studies conducted to justify human clinical trials are required to follow Good Laboratory Practice Regulations (UK Government 1999). This signalling question assesses whether the study authors have provided the details of ethics approval or any other regulatory approvals and standards that they used in their research.

7.2 Did the authors take steps to prevent unintentional changes to data?

Unintentional human errors when handling data ('data corruption') has the potential to affect the quality of study results and a possible reason for lack of reproducibility as they can cause misclassification of exposure or outcomes (Van den Broeck, Cunningham et al. 2005, Ward, Self et al. 2015). 'Data cleaning' is the process of identifying and correcting these errors, or at least attempting to minimise the impact on study results (Van den Broeck, Cunningham et al. 2005). Methods used for data cleaning can have a significant impact on the results (Dasu and Loh 2012, Randall, Ferrante et al. 2013). The best way to minimise data errors is to avoid them in the first place. While there are many 'data handling' guidelines about the protection of personal data, there is currently no guidance on the best method to avoid 'data corruption'. The UK Digital Curation Centre (www.dcc.ac.uk) provides expert advice and practical help to research organisations wanting to store, manage, protect, and share digital research data. Maintenance of laboratory logs, accuracy measures between laboratory logs and data used, and use of password-protected data files can all decrease the risks of unintentional changes to data. This signalling question assesses whether the authors took steps to prevent unintentional changes to data.

Domain 8: Research transparency

The purpose of this domain is to assess whether the animal study authors were transparent in their reporting. Transparent reporting increases the confidence in the study findings and promotes replicability of the research findings. Reporting guidelines such as ARRIVE guidelines 2.0, Gold Standard Publication Checklist to improve the quality of animal studies, and National Research Council's guidance on description of animal research in scientific publications can help with transparent reporting (Hooijmans, Leenaars et al. 2010, National Research Council 2011, Percie du Sert, Ahluwalia et al. 2020).

8.1 Did the authors describe the experimental procedures sufficiently in a protocol that was registered prior to the start of the research?

While selective outcome reporting is covered under the bias (internal validity) domain, the authors may have changed the protocol of the study in various other ways, for example, the disease-specific model, intervention, control, or the methods of administration of the intervention

and control. The experimental protocols should be registered prior to the start of the study in a preclinical trial registry, such as, https://www.preclinicaltrials.eu/, which allows registration of animal studies and is searchable. Studies can also be registered in Open Science Framework (https://osf.io/). Alternatively, posting the protocol in open access preprint servers such as https://www.biorxiv.org/ or https://arxiv.org/, print or online journals, in an institutional or public data repository such as https://zenodo.org/ is another option. The study authors should provide a link to this registered protocol in their study report.

The focus of this signalling question is about availability of a registered protocol prior to research commencement, which had enough details to allow replication, while the signalling question 5.1 refers to the description of the final protocol used (after all the modifications to the registered protocol) in sufficient detail to allow replication.

8.2 Did the authors describe any deviations from the registered protocol?

There may be justifiable reasons for alteration from a registered protocol. The authors should be explicit and describe any deviations from their plans and the reasons for them. In addition to registries, repositories, and journals for registering preclinical trials, some journals also offer 'registered reports' publishing format, which involves peer review of the study design and methodology, and if successful, results in a conditional acceptance for publication prior to the research being undertaken (Hardwicke and Ioannidis 2018). This will also allow evaluation of the deviations from the registered protocol.

8.3 Did the authors provide the individual subject data along with explanation for any numerical codes/substitutions or abbreviations used in the data to allow other groups of researchers to analyse?

In addition to making the protocol available, the key aspects of reproducibility and replicability in research involving data are the availability of the raw data from which results were generated, the computer code that generated the findings, and any additional information needed such as workflows and input parameters (Stodden, Seiler et al. 2018). Despite the journal policies about data sharing, only a third of computational and data analysis could be reproduced in a straightforward way or with minor difficulty (Stodden, Seiler et al. 2018). The remaining required substantial revisions for reproduction or could not be reproduced (Stodden, Seiler et al. 2018).

During the analysis, the authors may have processed the data to allow analysis. This may be in the form of transformation of data (for example, log-transformation or transformation from continuous or ordinal data into binary data), substitutions of texts with numbers (for example, intervention may be coded as 1 and control may be coded as 0; similarly, the characteristics and/or outcomes may have been coded), or may have used abbreviations for variable names to

allow easy management and meet the requirements for the statistical software package used. Some authors may use complex computer codes to perform the analysis. This is different from the transformation or substitution codes and refers to a set of computer commands that are executed sequentially by the computer. While the authors may provide the individual subject data as part of data sharing plan or as a journal requirement, this data is unlikely to be useful for analysis if the transformation codes, substitution codes, abbreviations, or computer codes are not available. Therefore, the individual participant data should be provided along with any transformation codes, substitution codes, and abbreviations to allow other researchers to perform analysis. The individual participant data can be provided either as a supplementary appendix in the journal publication or can be provided in open access repositories such as https://zenodo.org/ or university open access repositories. This signalling question assesses whether individual subject data with sufficient details to reanalyse were available.

Discussion

Using a modified Delphi consensus process, we have developed a tool to assess the clinical relevance of a preclinical research study in terms of the likelihood that therapeutic preclinical research methods and findings can be translated into improvement in the management of human diseases. We searched for existing guidelines about the design, conduct, and reporting of preclinical research and developed domains and signalling questions by involving experts. A modified Delphi consensus process was used to develop new domains and signalling questions and refine the existing domains and signalling questions to improve the understanding of the people who assess the clinical relevance of animal research. We have included only questions for which consensus was achieved (i.e. at least 70% of the Delphi panel members considered the question important to evaluate the clinical relevance of animal research). This tool provides a framework for researchers, journal editors, grant funders, and regulatory authorities to assess the clinical relevance of preclinical animal research with the aim to achieve better design, conduct, and reporting of preclinical animal research.

This tool is different from the ARRIVE guidelines 2.0 (Percie du Sert, Ahluwalia et al. 2020) and the NIH effort on improving preclinical research (NIH 2014) as our tool is a domain-based assessment tool rather than a reporting guideline. Furthermore, as opposed to a reporting guideline where the questions relate to clarity of reporting, the questions in this tool assess the likelihood of the results being clinically relevant. This tool is also different from the SYRCLE risk of bias of tool, as this tool goes beyond the risk of bias in the research (Hooijmans, Rovers et al. 2014). While many of the issues have been covered by other reporting guidance on preclinical research, the issue of measurement errors (errors in accuracy, precision, or sampling error) have not been addressed in existing guidance on preclinical research. Measurement error in exposure or outcome is often neglected in medical research despite the potential to cause biased estimation of the effect of an exposure or intervention (Hernan and Cole 2009, Brakenhoff, Mitroiu et al. 2018, Brakenhoff, van Smeden et al. 2018). Even though preclinical animal research often

involves repeated measurements, the measurement error is generally not reported or not taken into account during the analysis. This Delphi panel arrived at a consensus that measurement errors should be taken into account during the analysis if necessary and should be reported to enable an assessment of whether the preclinical research is translatable to humans.

We are now piloting this tool to improve it. This is in the form of providing learning material to people willing to pilot this tool and requesting them to assess the clinical relevance of preclinical animal studies. Financial incentives are being offered for piloting the tool. We intend to pilot the tool with 50 individuals including researchers performing or planning to perform preclinical or clinical studies. If the percentage agreement for classification of a domain is less than 70%, we will consider refining the question, explanation, or training by an iterative process to improve the agreement. The link for the learning material is available at: https://doi.org/10.5281/zenodo.4159278. The tool can be completed using an Excel file, which is available in the same link.

Conclusions

We have developed a tool to assess the clinical relevance of preclinical studies. This tool is currently being piloted.

Acknowledgements

References

(2018). "National Toxicology Program." https://ntp.niehs.nih.gov/ (accessed on 24 August 2018).

Altman, D. G., D. Moher and K. F. Schulz (2017). "Harms of outcome switching in reports of randomised trials: CONSORT perspective." BMJ **356**: j396.

Andrews, N. A., A. Latremoliere, A. I. Basbaum, J. S. Mogil, F. Porreca, A. S. Rice, C. J. Woolf, G. L. Currie, R. H. Dworkin, J. C. Eisenach, S. Evans, J. S. Gewandter, T. D. Gover, H. Handwerker, W. Huang, S. Iyengar, M. P. Jensen, J. D. Kennedy, N. Lee, J. Levine, K. Lidster, I. Machin, M. P. McDermott, S. B. McMahon, T. J. Price, S. E. Ross, G. Scherrer, R. P. Seal, E. S. Sena, E. Silva, L. Stone, C. I. Svensson, D. C. Turk and G. Whiteside (2016). "Ensuring transparency and minimization of methodologic bias in preclinical pain research: PPRECISE considerations." Pain **157**(4): 901-909.

Ansari, D., S. Kristoffersson, R. Andersson and M. Bergenfeldt (2017). "The role of irreversible electroporation (IRE) for locally advanced pancreatic cancer: a systematic review of safety and efficacy." Scand J Gastroenterol **52**(11): 1165-1171.

Baker, M. (2016). "1,500 scientists lift the lid on reproducibility." Nature **533**(7604): 452-454.

Bate, S. T. and R. A. Clark (2014). "The Design and Statistical Analysis of Animal Experiments." http://www.cambridge.org/gb/academic/subjects/life-sciences/quantitative-biology-biostatistics-and-mathematical-modellin/design-and-statistical-analysis-animal-experiments?format=HB (acccessed 25 August 2018).

Bath, P. M., M. R. Macleod and A. R. Green (2009). "Emulating multicentre clinical stroke trials: a new paradigm for studying novel interventions in experimental models of stroke." Int J Stroke **4**(6): 471-479.

Begley, C. G. and L. M. Ellis (2012). "Drug development: Raise standards for preclinical cancer research." Nature **483**(7391): 531-533.

Begley, C. G. and J. P. Ioannidis (2015). "Reproducibility in science: improving the standard for basic and preclinical research." Circ Res **116**(1): 116-126.

Biophysical Journal (2017). "Guidelines for the reproducibility of Biophysics Research." http://www.cell.com/pb/assets/raw/journals/society/biophysj/PDFs/reproducibility-guidelines.pdf (accessed on 11 June 2017).

Bland, J. M. and D. G. Altman (1996). "Measurement error." Bmj **313**(7059): 744.

Bland, J. M. and D. G. Altman (1996). "Measurement error and correlation coefficients." Bmj **313**(7048): 41-42.

Bland, J. M. and D. G. Altman (1996). "Measurement error proportional to the mean." Bmj **313**(7049): 106.

Bland, M. J. and D. G. Altman (1986). "Statistical methods for assessing agreement between two methods of clinical measurement." The Lancet **327**(8476): 307-310.

Bouxsein, M. L., S. K. Boyd, B. A. Christiansen, R. E. Guldberg, K. J. Jepsen and R. Muller (2010). "Guidelines for assessment of bone microstructure in rodents using micro-computed tomography." J Bone Miner Res **25**(7): 1468-1486.

Brakenhoff, T. B., M. Mitroiu, R. H. Keogh, K. G. M. Moons, R. H. H. Groenwold and M. van Smeden (2018). "Measurement error is often neglected in medical literature: a systematic review." J Clin Epidemiol **98**: 89-97.

Brakenhoff, T. B., M. van Smeden, F. L. J. Visseren and R. H. H. Groenwold (2018). "Random measurement error: Why worry? An example of cardiovascular risk factors." PLOS ONE **13**(2): e0192298.

Bramhall, M., O. Florez-Vargas, R. Stevens, A. Brass and S. Cruickshank (2015). "Quality of methods reporting in animal models of colitis." Inflamm Bowel Dis **21**(6): 1248-1259.

Bucher, H. C., G. H. Guyatt, D. J. Cook, A. Holbrook, F. A. McAlister and G. for the Evidence-Based Medicine Working (1999). "Users' guides to the medical literature: XIX. applying clinical trial results a. how to use an article measuring the effect of an intervention on surrogate end points." JAMA **282**(8): 771-778.

Buyse, M., G. Molenberghs, T. Burzykowski, D. Renard and H. Geys (2000). "The validation of surrogate endpoints in meta-analyses of randomized experiments." Biostatistics **1**(1): 49-67.

Carroll, R. J. (1989). "Covariance analysis in generalized linear measurement error models." Stat Med **8**(9): 1075-1093; discussion 1107-1078.

Casson, R. J. and L. D. Farmer (2014). "Understanding and checking the assumptions of linear regression: a primer for medical researchers." Clin Experiment Ophthalmol **42**(6): 590-596.

Chen, P. W. and H. M. Wang (2018). "Randomized controlled trial of scleroligation versus band ligation for eradication of gastroesophageal varices." Gastrointestinal Endoscopy **87**(3): 904-904.

Church, D. M., L. Goodstadt, L. W. Hillier, M. C. Zody, S. Goldstein, X. She, C. J. Bult, R. Agarwala, J. L. Cherry, M. DiCuccio, W. Hlavina, Y. Kapustin, P. Meric, D. Maglott, Z. Birtle, A. C. Marques, T. Graves, S. Zhou, B. Teague, K. Potamousis, C. Churas, M. Place, J. Herschleb, R. Runnheim, D. Forrest, J. Amos-Landgraf, D. C. Schwartz, Z. Cheng, K. Lindblad-Toh, E. E. Eichler, C. P. Ponting and C. The Mouse Genome Sequencing (2009). "Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse." PLOS Biology **7**(5): e1000112.

Cohen, D. (2018). "Oxford vaccine study highlights pick and mix approach to preclinical research." BMJ **360**: j5845.

Collins, F. S. and L. A. Tabak (2014). "Policy: NIH plans to enhance reproducibility." Nature **505**(7485): 612-613.

Contopoulos-Ioannidis, D. G., E. Ntzani and J. P. Ioannidis (2003). "Translation of highly promising basic science research into clinical applications." Am J Med **114**(6): 477-484.

Czigany, Z., J. Iwasaki, S. Yagi, K. Nagai, A. Szijarto, S. Uemoto and R. H. Tolba (2015). "Improving Research Practice in Rat Orthotopic and Partial Orthotopic Liver Transplantation: A Review, Recommendation, and Publication Guide." Eur Surg Res **55**(1-2): 119-138.

Dasu, T. and J. M. Loh (2012). "Statistical distortion: consequences of data cleaning." J Proc. VLDB Endow **5**(11): 1674-1683.

de Vries, R. B., P. Buma, M. Leenaars, M. Ritskes-Hoitinga and B. Gordijn (2012). "Reducing the number of laboratory animals used in tissue engineering research by restricting the variety of animal models. Articular cartilage tissue engineering as a case study." Tissue Eng Part B Rev **18**(6): 427-435.

Diamond, I. R., R. C. Grant, B. M. Feldman, P. B. Pencharz, S. C. Ling, A. M. Moore and P. W. Wales (2014). "Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies." J Clin Epidemiol **67**(4): 401-409.

Eisener-Dorman, A. F., D. A. Lawrence and V. J. Bolivar (2009). "Cautionary insights on knockout mouse studies: the gene or not the gene?" Brain Behav Immun **23**(3): 318-324.

EMA (2017). "Guideline on strategies to identify and mitigate risks for first-in-human and early clinical trials with investigational medicinal products." https://www.ema.europa.eu/documents/scientific-guideline/guideline-strategies-identify-mitigate-risks-first-human-early-clinical-trials-investigational_en.pdf (accessed on 11 November 2018).

FASEB journal (2016). "Enhancing research reproducibility: Recommendations from the Federation of American Societies for Experimental Biology " https://www.faseb.org/Portals/2/PDFs/opa/2016/FASEB_Enhancing%20Research%20Reproducibility.pdf (accessed on 11 November 2018).

FDA (2005). "Guidance for industry: Estimating the maximum safe starting dose in initial clinical trials for therapeutics in adult healthy volunteers." https://www.fda.gov/downloads/Drugs/Guidances/UCM078932.pdf%23search=%27guidekines+for+industry+sfe+starting%27 (accessed 11 November 2018).

Festing, M. (2016). "The Design of Animal Experiments " https://uk.sagepub.com/en-gb/eur/the-design-of-animal-experiments/book252408#contents (accessed 25 August 2018).

Fisher, M., G. Feuerstein, D. W. Howells, P. D. Hurn, T. A. Kent, S. I. Savitz, E. H. Lo and S. Group (2009). "Update of the stroke therapy academic industry roundtable preclinical recommendations." Stroke **40**(6): 2244-2250.

Fleming, T. R. and D. L. DeMets (1996). "Surrogate end points in clinical trials: are we being misled?" Ann Intern Med **125**(7): 605-613.

Freedman, L. S., V. Fainberg, V. Kipnis, D. Midthune and R. J. Carroll (2004). "A new method for dealing with measurement error in explanatory variables of regression models." Biometrics **60**(1): 172-181.

Freedman, L. S., B. I. Graubard and A. Schatzkin (1992). "Statistical validation of intermediate endpoints for chronic diseases." Stat Med **11**(2): 167-178.

Freedman, L. S., D. Midthune, R. J. Carroll and V. Kipnis (2008). "A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression." Stat Med **27**(25): 5195-5216.

Gibbs, R. A., G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, S. Scherer, G. Scott, D. Steffen, K. C. Worley, P. E. Burch, G. Okwuonu, S. Hines, L. Lewis, C. DeRamo, O. Delgado, S. Dugan-Rocha, G. Miner, M. Morgan, A. Hawes, R. Gill, C. R. A. Holt, M. D. Adams, P. G. Amanatides, H. Baden-Tillson, M. Barnstead, S. Chin, C. A. Evans, S. Ferriera, C. Fosler, A. Glodek, Z. Gu, D. Jennings, C. L. Kraft, T. Nguyen, C. M. Pfannkoch, C. Sitter, G. G. Sutton, J. C. Venter, T. Woodage, D. Smith, H.-M. Lee, E. Gustafson, P. Cahill, A. Kana, L. Doucette-Stamm, K. Weinstock, K. Fechtel, R. B. Weiss, D. M. Dunn, E. D. Green, R. W. Blakesley, G. G. Bouffard, P. J. de Jong, K. Osoegawa, B. Zhu, M. Marra, J. Schein, I. Bosdet, C. Fjell, S. Jones, M. Krzywinski, C. Mathewson, A. Siddiqui, N. Wye, J. McPherson, S. Zhao, C. M. Fraser, J. Shetty, S. Shatsman, K. Geer, Y. Chen, S. Abramzon, W. C. Nierman, R. A. Gibbs, G. M. Weinstock, P. H. Havlak, R. Chen, K. James Durbin, A. Egan, Y. Ren, X.-Z. Song, B. Li, Y. Liu, X. Qin, S. Cawley, G. M. Weinstock, K. C. Worley, A. J. Cooney, R. A. Gibbs, L. M. D'Souza, K. Martin, J. Qian Wu, M. L. Gonzalez-Garay, A. R. Jackson, K. J. Kalafus, M. P. McLeod, A. Milosavljevic, D. Virk, A. Volkov, D. A. Wheeler, Z. Zhang, J. A. Bailey, E. E. Eichler, E. Tuzun, E. Birney, E. Mongin, A. Ureta-Vidal, C. Woodwark, E. Zdobnov, P. Bork, M. Suyama, D. Torrents, M. Alexandersson, B. J. Trask, J. M. Young, D. Smith, H. Huang, K. Fechtel, H. Wang, H. Xing, K. Weinstock, S. Daniels, D. Gietzen, J. Schmidt, K. Stevens, U. Vitt, J. Wingrove, F. Camara, M. Mar Albà, J. F. Abril, R. Guigo, A. Smit, I. Dubchak, E. M. Rubin, O. Couronne, A. Poliakov, N. Hübner, D. Ganten, C. Goesele, O. Hummel, T. Kreitler, Y.-A. Lee, J. Monti, H. Schulz, H. Zimdahl, H. Himmelbauer, H. Lehrach, H. J. Jacob, S. Bromberg, J. Gullings-Handley, M. I. Jensen-Seaman, A. E. Kwitek, J. Lazar, D. Pasko, P. J. Tonellato, S. Twigger, C. P. Ponting, J. M. Duarte, S. Rice, L. Goodstadt, S. A. Beatson, R. D. Emes, E. E. Winter, C. Webber, P. Brandt, G. Nyakatura, M. Adetobi, F. Chiaromonte, L. Elnitski, P. Eswara, R. C. Hardison, M. Hou, D. Kolbe, K. Makova, W. Miller, A. Nekrutenko, C. Riemer, S. Schwartz, J. Taylor, S. Yang, Y. Zhang, K. Lindpaintner, T. D. Andrews, M. Caccamo, M. Clamp, L. Clarke, V. Curwen, R. Durbin, E. Eyras, S. M. Searle, G. M. Cooper, S. Batzoglou, M. Brudno, A. Sidow, E. A. Stone, J. Craig Venter, B. A. Payseur, G. Bourque, C. López-Otín, X. S. Puente, K. Chakrabarti, S. Chatterji, C. Dewey, L. Pachter, N. Bray, V. B. Yap, A. Caspi, G. Tesler, P. A. Pevzner, D. Haussler, K. M. Roskin, R. Baertsch, H. Clawson, T. S. Furey, A. S. Hinrichs, D. Karolchik, W. J. Kent, K. R. Rosenbloom, H. Trumbower, M. Weirauch, D. N. Cooper, P. D. Stenson, B. Ma, M. Brent, M. Arumugam, D. Shteynberg, R. R. Copley, M. S. Taylor, H. Riethman, U. Mudunuri, J. Peterson, M. Guyer, A. Felsenfeld, S. Old, S. Mockrin and F. Collins (2004). "Genome sequence of the Brown Norway rat yields insights into mammalian evolution." Nature **428**: 493.

Grant, A. M., D. G. Altman, A. B. Babiker, M. K. Campbell, F. J. Clemens, J. H. Darbyshire, D. R. Elbourne, S. K. McLeer, M. K. Parmar, S. J. Pocock, D. J. Spiegelhalter, M. R. Sydes, A. E. Walker, S. A. Wallace and D. s. group (2005). "Issues in data monitoring and interim analysis of trials." Health Technol Assess **9**(7): 1-238, iii-iv.

Hackam, D. G. and D. A. Redelmeier (2006). "Translation of research evidence from animals to humans." JAMA **296**(14): 1731-1732.

1000 Hardwicke, T. E. and J. P. A. Ioannidis (2018). "Mapping the universe of registered reports." Nature
1001 Human Behaviour **2**(11): 793-796.

1002 Hasson, F., S. Keeney and H. McKenna (2000). "Research guidelines for the Delphi survey technique." J
1003 Adv Nurs **32**(4): 1008-1015.

1004 Heckathorn, D. D. (2011). "Snowball Versus Respondent-Driven Sampling." Sociol Methodol **41**(1): 355-
1005 366.

1006 Hernan, M. A. and S. R. Cole (2009). "Invited Commentary: Causal diagrams and measurement bias." Am
1007 J Epidemiol **170**(8): 959-962; discussion 963-954.

1008 Hidalgo, B. and M. Goodman (2013). "Multivariate or multivariable regression?" Am J Public Health
1009 **103**(1): 39-40.

1010 Higgins, J., S. Green and (editors) (2011). "Cochrane Handbook for Systematic Reviews of Interventions
1011 Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-
1012 handbook.org."

1013 Hooijmans, C. R., M. Leenaars and M. Ritskes-Hoitinga (2010). "A gold standard publication checklist to
1014 improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews
1015 more feasible." Altern Lab Anim **38**(2): 167-182.

1016 Hooijmans, C. R., M. M. Rovers, R. B. de Vries, M. Leenaars, M. Ritskes-Hoitinga and M. W. Langendam
1017 (2014). "SYRCLE's risk of bias tool for animal studies." BMC Medical Research Methodology **14**(1): 43.

1018 Howard, B., J. T. Scott, M. Blubaugh, B. Roepke, C. Scheckel and M. Vassar (2017). "Systematic review:
1019 Outcome reporting bias is a problem in high impact factor neurology journals." PLOS ONE **12**(7):
1020 e0180986.

1021 Howe, K., M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot, M. Muffato, J. E. Collins, S. Humphray, K.
1022 McLaren, L. Matthews, S. McLaren, I. Sealy, M. Caccamo, C. Churcher, C. Scott, J. C. Barrett, R. Koch, G.-J.
1023 Rauch, S. White, W. Chow, B. Kilian, L. T. Quintais, J. A. Guerra-Assunção, Y. Zhou, Y. Gu, J. Yen, J.-H.
1024 Vogel, T. Eyre, S. Redmond, R. Banerjee, J. Chi, B. Fu, E. Langley, S. F. Maguire, G. K. Laird, D. Lloyd, E.
1025 Kenyon, S. Donaldson, H. Sehra, J. Almeida-King, J. Loveland, S. Trevanion, M. Jones, M. Quail, D. Willey,
1026 A. Hunt, J. Burton, S. Sims, K. McLay, B. Plumb, J. Davis, C. Clee, K. Oliver, R. Clark, C. Riddle, D. Elliott, G.
1027 Threadgold, G. Harden, D. Ware, S. Begum, B. Mortimore, G. Kerry, P. Heath, B. Phillimore, A. Tracey, N.
1028 Corby, M. Dunn, C. Johnson, J. Wood, S. Clark, S. Pelan, G. Griffiths, M. Smith, R. Glithero, P. Howden, N.
1029 Barker, C. Lloyd, C. Stevens, J. Harley, K. Holt, G. Panagiotidis, J. Lovell, H. Beasley, C. Henderson, D.
1030 Gordon, K. Auger, D. Wright, J. Collins, C. Raisen, L. Dyer, K. Leung, L. Robertson, K. Ambridge, D.
1031 Leongamornlert, S. McGuire, R. Gilderthorp, C. Griffiths, D. Manthravadi, S. Nichol, G. Barker, S.
1032 Whitehead, M. Kay, J. Brown, C. Murnane, E. Gray, M. Humphries, N. Sycamore, D. Barker, D. Saunders,
1033 J. Wallis, A. Babbage, S. Hammond, M. Mashreghi-Mohammadi, L. Barr, S. Martin, P. Wray, A. Ellington,
1034 N. Matthews, M. Ellwood, R. Woodmansey, G. Clark, J. D. Cooper, A. Tromans, D. Grafham, C. Skuce, R.
1035 Pandian, R. Andrews, E. Harrison, A. Kimberley, J. Garnett, N. Fosker, R. Hall, P. Garner, D. Kelly, C. Bird,
1036 S. Palmer, I. Gehring, A. Berger, C. M. Dooley, Z. Ersan-Ürün, C. Eser, H. Geiger, M. Geisler, L. Karotki, A.
1037 Kirn, J. Konantz, M. Konantz, M. Oberländer, S. Rudolph-Geiger, M. Teucke, C. Lanz, G. Raddatz, K.

1038 Osoegawa, B. Zhu, A. Rapp, S. Widaa, C. Langford, F. Yang, S. C. Schuster, N. P. Carter, J. Harrow, Z. Ning,
1039 J. Herrero, S. M. J. Searle, A. Enright, R. Geisler, R. H. A. Plasterk, C. Lee, M. Westerfield, P. J. de Jong, L. I.
1040 Zon, J. H. Postlethwait, C. Nüsslein-Volhard, T. J. P. Hubbard, H. R. Crollius, J. Rogers and D. L. Stemple
1041 (2013). "The zebrafish reference genome sequence and its relationship to the human genome." Nature
1042 **496**: 498.

1043 Idris, A. H., L. B. Becker, J. P. Ornato, J. R. Hedges, N. G. Bircher, N. C. Chandra, R. O. Cummins, W. Dick,
1044 U. Ebmeyer, H. R. Halperin, M. F. Hazinski, R. E. Kerber, K. B. Kern, P. Safar, P. A. Steen, M. M. Swindle, J.
1045 E. Tsitlik, I. von Planta, M. von Planta, R. L. Wears and M. H. Weil (1996). "Utstein-style guidelines for
1046 uniform reporting of laboratory CPR research. A statement for healthcare professionals from a task
1047 force of the American Heart Association, the American College of Emergency Physicians, the American
1048 College of Cardiology, the European Resuscitation Council, the Heart and Stroke Foundation of Canada,
1049 the Institute of Critical Care Medicine, the Safar Center for Resuscitation Research, and the Society for
1050 Academic Emergency Medicine. Writing Group." Circulation **94**(9): 2324-2336.

1051 Ioannidis, J. P. (2017). "Acknowledging and Overcoming Nonreproducibility in Basic and Preclinical
1052 Research." JAMA **317**(10): 1019-1020.

1053 Jones, C. W., L. G. Keil, W. C. Holland, M. C. Caughey and T. F. J. B. M. Platts-Mills (2015). "Comparison of
1054 registered and published outcomes in randomized controlled trials: a systematic review." BMC Medicine
1055 **13**(1): 282.

1056 Jones, J. and D. Hunter (1995). "Consensus methods for medical and health services research." BMJ
1057 **311**(7001): 376-380.

1058 Kilkenny, C., W. J. Browne, I. C. Cuthill, M. Emerson and D. G. Altman (2010). "Improving bioscience
1059 research reporting: the ARRIVE guidelines for reporting animal research." PLoS Biol **8**(6): e1000412.

1060 Kim, C. and V. Prasad (2015). "Cancer drugs approved on the basis of a surrogate end point and
1061 subsequent overall survival: An analysis of 5 years of us food and drug administration approvals." JAMA
1062 Internal Medicine **175**(12): 1992-1994.

1063 Kirkham, J. J., K. Davis, D. G. Altman, J. M. Blazeby, M. Clarke, S. Tunis and P. R. Williamson (2017). "Core
1064 Outcome Set-STAndards for Development: The COS-STAD recommendations." PLOS Medicine **14**(11):
1065 e1002447.

1066 Kleynen, M., S. M. Braun, M. H. Bleijlevens, M. A. Lexis, S. M. Rasquin, J. Halfens, M. R. Wilson, A. J.
1067 Beurskens and R. S. Masters (2014). "Using a Delphi technique to seek consensus regarding definitions,
1068 descriptions and classification of terms related to implicit and explicit forms of motor learning." PLoS
1069 One **9**(6): e100227.

1070 Landis, S. C., S. G. Amara, K. Asadullah, C. P. Austin, R. Blumenstein, E. W. Bradley, R. G. Crystal, R. B.
1071 Darnell, R. J. Ferrante, H. Fillit, R. Finkelstein, M. Fisher, H. E. Gendelman, R. M. Golub, J. L. Goudreau, R.
1072 A. Gross, A. K. Gubitz, S. E. Hesterlee, D. W. Howells, J. Huguenard, K. Kelner, W. Koroshetz, D. Krainc, S.
1073 E. Lazic, M. S. Levine, M. R. Macleod, J. M. McCall, R. T. Moxley, 3rd, K. Narasimhan, L. J. Noble, S. Perrin,
1074 J. D. Porter, O. Steward, E. Unger, U. Utz and S. D. Silberberg (2012). "A call for transparent reporting to
1075 optimize the predictive value of preclinical research." Nature **490**(7419): 187-191.

1076 Lin, X. and R. J. Carroll (1999). "SIMEX variance component tests in generalized linear mixed
1077 measurement error models." Biometrics **55**(2): 613-619.

1078 Littell, R. C., J. Pendergast and R. Natarajan (2000). "Modelling covariance structure in the analysis of
1079 repeated measures data." Stat Med **19**(13): 1793-1819.

1080 Lyu, T., X. Wang, Z. Su, J. Shangguan, C. Sun, M. Figini, J. Wang, V. Yaghmai, A. C. Larson and Z. Zhang
1081 (2017). "Irreversible electroporation in primary and metastatic hepatic malignancies: A review."
1082 Medicine (Baltimore) **96**(17): e6386.

1083 Macleod, M. R., M. Fisher, V. O'Collins, E. S. Sena, U. Dirnagl, P. M. Bath, A. Buchan, H. B. van der Worp,
1084 R. Traystman, K. Minematsu, G. A. Donnan and D. W. Howells (2009). "Good laboratory practice:
1085 preventing introduction of bias at the bench." Stroke **40**(3): e50-52.

1086 Moher, D., K. F. Schulz, I. Simera and D. G. Altman (2010). "Guidance for developers of health research
1087 reporting guidelines." PLoS Med **7**(2): e1000217.

1088 Nair, A. B. and S. Jacob (2016). "A simple practice guide for dose conversion between animals and
1089 human." J Basic Clin Pharm **7**(2): 27-31.

1090 National Research Council (2011). "Guidance for the description of animal research in scientific
1091 publications " https://www.nap.edu/catalog/13241/guidance-for-the-description-of-animal-research-in-
1092 scientific-publications (accessed on 27 Nov 2018).

1093 Nature Collection (2018). "Statistics for biologists." https://www.nature.com/collections/qghhqm
1094 (accessed on 25 August 2018).

1095 NC3Rs "The 3Rs." https://www.nc3rs.org.uk/the-3rs (accessed on 12 November 2018).

1096 Neumann, K., U. Grittner, S. K. Piper, A. Rex, O. Florez-Vargas, G. Karystianis, A. Schneider, I. Wellwood,
1097 B. Siegerink, J. P. Ioannidis, J. Kimmelman and U. Dirnagl (2017). "Increasing efficiency of preclinical
1098 research by group sequential designs." PLoS Biol **15**(3): e2001307.

1099 NIH (2014). "Principles and Guidelines for Reporting Preclinical Research."
1100 https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-
1101 research (accessed on 11 June 2017).

1102 Nørskov, A. K., T. Lange, E. E. Nielsen, C. Gluud, P. Winkel, J. Beyersmann, J. de Uña-Álvarez, V. Torri, L.
1103 Billot, H. Putter, J. Wetterslev, L. Thabane and J. C. Jakobsen (2020). "Assessment of assumptions of
1104 statistical analysis methods in randomised clinical trials: the what and how." BMJ Evid Based Med.

1105 Open Science Framework (2017).
1106 "Guidelines for Transparency and Openness Promotion (TOP) in Journal Policies and Practices “The TOP
1107 Guidelines” " https://osf.io/ud578/?show=revision (accessed on 11 June 2017).

1108 Osborne, N., M. T. Avey, L. Anestidou, M. Ritskes-Hoitinga and G. Griffin (2018). "Improving animal
1109 research reporting standards." HARRP, the first step of a unified approach by ICLAS to improve animal
1110 research reporting standards worldwide **19**(5).

1111 Percie du Sert, N., A. Ahluwalia, S. Alam, M. T. Avey, M. Baker, W. J. Browne, A. Clark, I. C. Cuthill, U.
1112 Dirnagl, M. Emerson, P. Garner, S. T. Holgate, D. W. Howells, V. Hurst, N. A. Karp, S. E. Lazic, K. Lidster, C.
1113 J. MacCallum, M. Macleod, E. J. Pearl, O. H. Petersen, F. Rawle, P. Reynolds, K. Rooney, E. S. Sena, S. D.
1114 Silberberg, T. Steckler and H. Würbel (2020). "Reporting animal research: Explanation and elaboration
1115 for the ARRIVE guidelines 2.0." PLOS Biology **18**(7): e3000411.

1116 Perel, P., I. Roberts, E. Sena, P. Wheble, C. Briscoe, P. Sandercock, M. Macleod, L. E. Mignini, P. Jayaram
1117 and K. S. Khan (2007). "Comparison of treatment effects between animal experiments and clinical trials:
1118 systematic review." BMJ **334**(7586): 197.

1119 Perrin, S. (2014). "Preclinical research: Make mouse studies work." Nature **507**(7493): 423-425.

1120 Randall, S. M., A. M. Ferrante, J. H. Boyd, J. B. J. B. M. I. Semmens and D. Making (2013). "The effect of
1121 data cleaning on record linkage quality." **13**(1): 64.

1122 Ritskes-Hoitinga, M. and K. Wever (2018). "Improving the conduct, reporting, and appraisal of animal
1123 research." BMJ **360**: j4935.

1124 Rupp, T. and D. Zuckerman (2017). "Quality of life, overall survival, and costs of cancer drugs approved
1125 based on surrogate endpoints." JAMA Internal Medicine **177**(2): 276-277.

1126 Sena, E., H. B. van der Worp, D. Howells and M. Macleod (2007). "How can we improve the pre-clinical
1127 development of drugs for stroke?" Trends Neurosci **30**(9): 433-439.

1128 Sloff, M., R. de Vries, P. Geutjes, J. IntHout, M. Ritskes-Hoitinga, E. Oosterwijk and W. Feitz (2014).
1129 "Tissue engineering in animal models for urinary diversion: a systematic review." PLoS One **9**(6): e98734.

1130 Sloff, M., V. Simaioforidis, R. de Vries, E. Oosterwijk and W. Feitz (2014). "Tissue engineering of the
1131 bladder--reality or myth? A systematic review." J Urol **192**(4): 1035-1042.

1132 Smith, A. J., R. E. Clutton, E. Lilley, K. E. A. Hansen and T. Brattelid (2018). "PREPARE: guidelines for
1133 planning animal research and testing." Laboratory Animals **52**(2): 135-141.

1134 Sterne, J. A., M. A. Hernan, B. C. Reeves, J. Savovic, N. D. Berkman, M. Viswanathan, D. Henry, D. G.
1135 Altman, M. T. Ansari, I. Boutron, J. R. Carpenter, A. W. Chan, R. Churchill, J. J. Deeks, A. Hrobjartsson, J.
1136 Kirkham, P. Juni, Y. K. Loke, T. D. Pigott, C. R. Ramsay, D. Regidor, H. R. Rothstein, L. Sandhu, P. L.
1137 Santaguida, H. J. Schunemann, B. Shea, I. Shrier, P. Tugwell, L. Turner, J. C. Valentine, H. Waddington, E.
1138 Waters, G. A. Wells, P. F. Whiting and J. P. Higgins (2016). "ROBINS-I: a tool for assessing risk of bias in
1139 non-randomised studies of interventions." BMJ **355**: i4919.

1140 Stodden, V., J. Seiler and Z. Ma (2018). "An empirical analysis of journal policy effectiveness for
1141 computational reproducibility." Proc Natl Acad Sci U S A **115**(11): 2584-2589.

1142 Streiner, D. L. (2015). "Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether
1143 and how to correct for many statistical tests." The American Journal of Clinical Nutrition **102**(4): 721-
1144 728.

1145 Sumsion, T. (1998). "The Delphi Technique: An Adaptive Research Tool." British Journal of Occupational
1146 Therapy **61**(4): 153-156.

1147 Tierney, J. F., J. P. Pignon, F. Gueffyier, M. Clarke, L. Askie, C. L. Vale, S. Burdett and I. P. D. M.-a. M. G.
1148 Cochrane (2015). "How individual participant data meta-analyses have influenced trial design, conduct,
1149 and analysis." J Clin Epidemiol **68**(11): 1325-1335.

1150 UK Government (1986). "Animals (Scientific Procedures) Act 1986."
1151 https://www.legislation.gov.uk/ukpga/1986/14/contents (accessed on 12 November 2018).

1152 UK Government (1999). "The Good Laboratory Practice Regulations 1999."
1153 http://www.legislation.gov.uk/uksi/1999/3106/contents/made (accessed on 24 August 2018).

1154 Van den Broeck, J., S. A. Cunningham, R. Eeckels and K. Herbst (2005). "Data cleaning: detecting,
1155 diagnosing, and editing data abnormalities." PLoS Med **2**(10): e267.

1156 van der Worp, H. B., D. W. Howells, E. S. Sena, M. J. Porritt, S. Rewell, V. O'Collins and M. R. Macleod
1157 (2010). "Can animal models of disease reliably inform human studies?" PLoS Med **7**(3): e1000245.

1158 van Stralen, K. J., K. J. Jager, C. Zoccali and F. W. Dekker (2008). "Agreement between methods." Kidney
1159 International **74**(9): 1116-1120.

1160 Vasey, M. W. and J. F. Thayer (1987). "The Continuing Problem of False Positives in Repeated Measures
1161 ANOVA in Psychophysiology: A Multivariate Solution." **24**(4): 479-486.

1162 Ward, M. J., W. H. Self and C. M. Froehle (2015). "Effects of Common Data Errors in Electronic Health
1163 Records on Emergency Department Operational Performance Metrics: A Monte Carlo Simulation." Acad
1164 Emerg Med **22**(9): 1085-1092.

1165 Watson, P. F. and A. Petrie (2010). "Method agreement analysis: A review of correct methodology."
1166 Theriogenology **73**(9): 1167-1179.

1167 Whiting, P., J. Savovic, J. P. Higgins, D. M. Caldwell, B. C. Reeves, B. Shea, P. Davies, J. Kleijnen, R.
1168 Churchill and R. group (2016). "ROBIS: A new tool to assess risk of bias in systematic reviews was
1169 developed." J Clin Epidemiol **69**: 225-234.

1170 Whiting, P. F., A. W. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. Leeflang, J. A.
1171 Sterne, P. M. Bossuyt and Q.-. Group (2011). "QUADAS-2: a revised tool for the quality assessment of
1172 diagnostic accuracy studies." Ann Intern Med **155**(8): 529-536.

1173 Wieschowski, S., W. W. L. Chin, C. Federico, S. Sievers, J. Kimmelman and D. Strech (2018). "Preclinical
1174 efficacy studies in investigator brochures: Do they enable risk-benefit assessment?" PLoS Biol **16**(4):
1175 e2004879.

1176 Yankelevitch-Yahav, R., M. Franko, A. Huly and R. Doron (2015). "The forced swim test as a model of
1177 depressive-like behavior." J Vis Exp(97).

1178 Yudkin, J. S., K. J. Lipska and V. M. Montori (2011). "The idolatry of the surrogate." BMJ **343**.

1179 Zaki, R., A. Bulgiba, R. Ismail and N. A. Ismail (2012). "Statistical Methods Used to Test for Agreement of
1180 Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic
1181 Review." PLOS ONE **7**(5): e37908.

1182 Zeeff, S. B., C. Kunne, G. Bouma, R. B. de Vries and A. A. Te Velde (2016). "Actual Usage and Quality of
1183 Experimental Colitis Models in Preclinical Efficacy Testing: A Scoping Review." Inflamm Bowel Dis **22**(6):
1184 1296-1305.

1185

Figure 1

Overall process

The outline of the process is shown in this figure. A total of three rounds were conducted. Consensus agreement was reached when at least 70% of panel members strongly agreed (scores of 7 or more) to include the domain or signalling question.

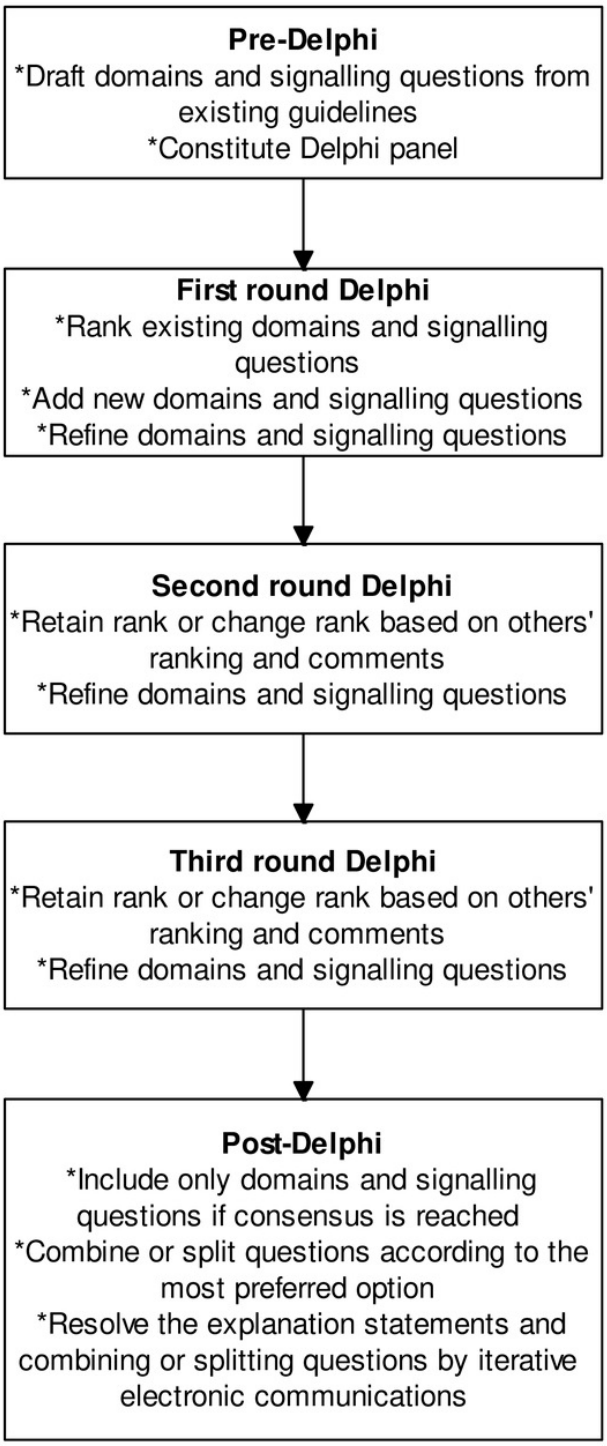

Pre-Delphi
*Draft domains and signalling questions from existing guidelines
*Constitute Delphi panel
First round Delphi
*Rank existing domains and signalling questions
*Add new domains and signalling questions
*Refine domains and signalling questions
Second round Delphi
*Retain rank or change rank based on others' ranking and comments
*Refine domains and signalling questions
Third round Delphi
*Retain rank or change rank based on others' ranking and comments
*Refine domains and signalling questions
Post-Delphi
*Include only domains and signalling questions if consensus is reached
*Combine or split questions according to the most preferred option
*Resolve the explanation statements and combining or splitting questions by iterative electronic communications

Figure 2

Flow of domains and signalling questions

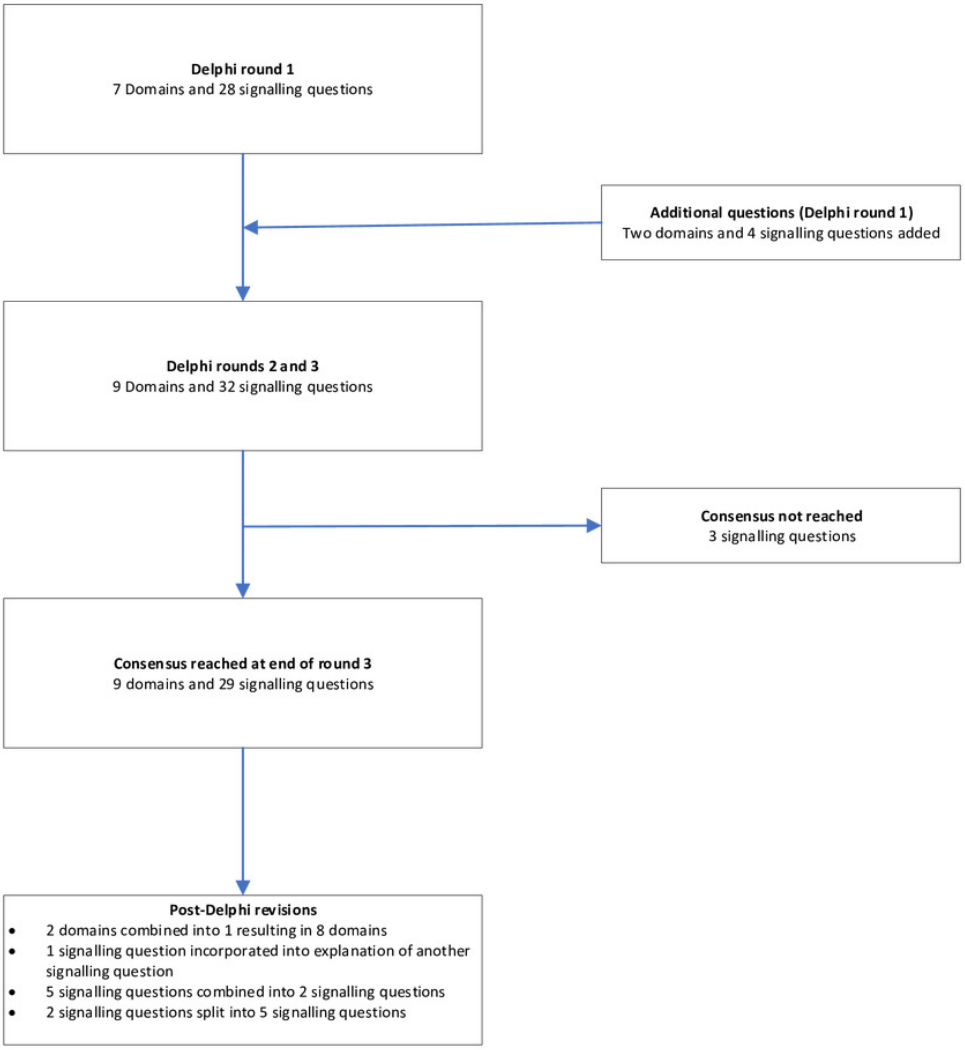

Delphi round 1
7 Domains and 28 signalling questions
Additional questions (Delphi round 1)
Two domains and 4 signalling questions added
Delphi rounds 2 and 3
9 Domains and 32 signalling questions
Consensus not reached
3 signalling questions
Consensus reached at end of round 3
9 domains and 29 signalling questions
Post-Delphi revisions
2 domains combined into 1 resulting in 8 domains
1 signalling question incorporated into explanation of another signalling question
5 signalling questions combined into 2 signalling questions
2 signalling questions split into 5 signalling questions

Figure 3

Schema for use of the tool

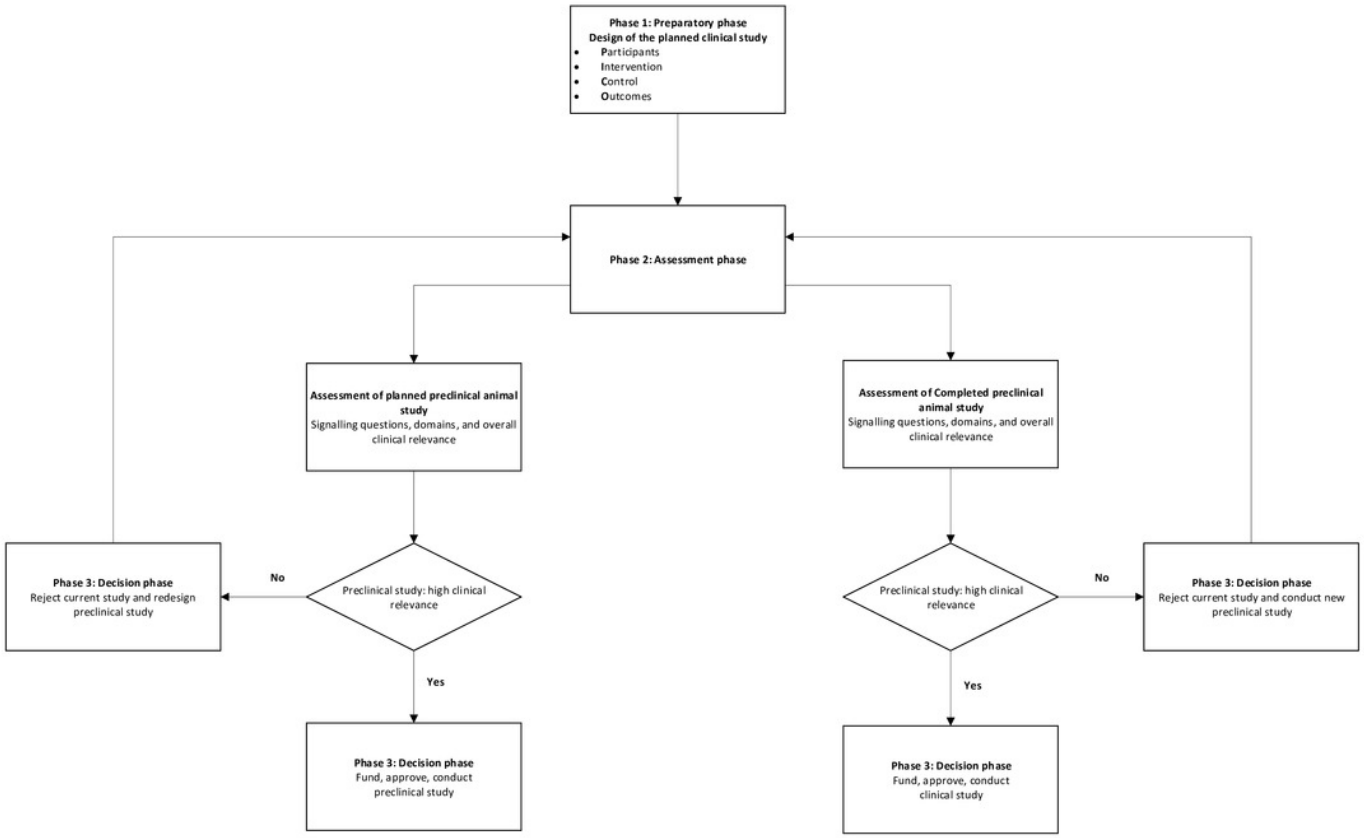


Figure 4

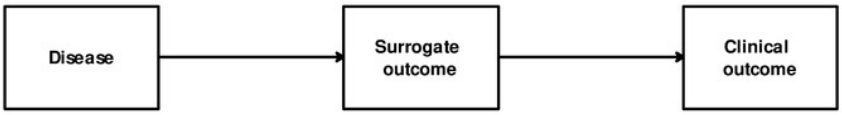
Situation when a surrogate outcome is likely to be valid

(A) The surrogate outcome is the only pathway that the disease can cause the clinical outcome. (B) The intervention acts in this pathway and causes a change in surrogate outcome, leading to a change in the clinical outcome. (C) If there are other pathways (which are not affected by the intervention) through which the disease can cause the clinical outcome, then the validity of the surrogate outcome will be decreased. If the intervention affects the clinical outcome through pathways unrelated to the surrogate outcome, then the validity of the surrogate outcome will be decreased.

Figure 2 Situation when a surrogate outcome is likely to be valid

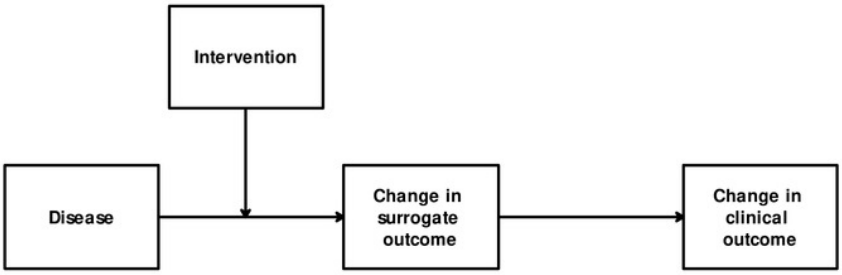
decreased. If the intervention affects the clinical outcome through pathways unrelated to the surrogate outcome, then the validity of the surrogate outcome will be decreased.

Figure 2a



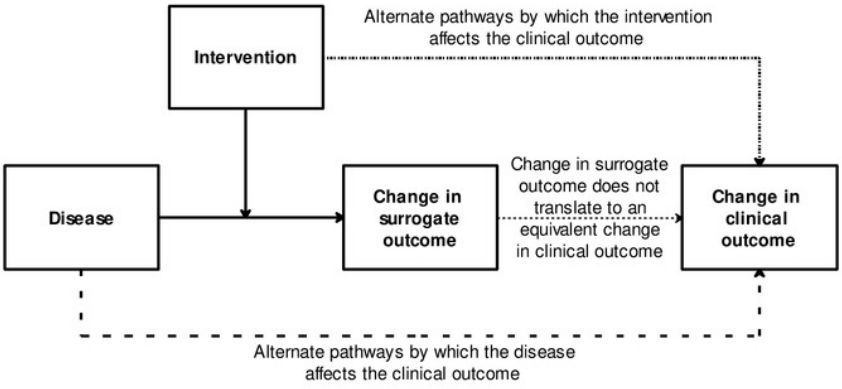
The surrogate outcome is the only pathway that the disease can cause the clinical outcome.

Figure 2b



The intervention acts in this pathway and causes a change in surrogate outcome, leading to a change in the clinical outcome.

Figure 2c



If there are other pathways (which are not affected by the intervention) through which the disease can cause the clinical outcome, then the validity of the surrogate outcome will be

Table 1(on next page)

Domains and signalling questions

1 **Table 1 Domains and signalling questions**

Domain or signalling question	Classification
Domain 1: Clinical translatability of results to human disease or condition (construct validity)	**Low concern/Moderate concern/High concern**
1.1 Did the authors use a model that adequately represents the human disease?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
1.2 Did the authors sufficiently identify and characterise the model?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
1.3 Were the method and timing of the intervention in the specific model relevant to humans?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
1.4 If the study used a surrogate outcome, was there a clear and reproducible correlation between surrogate outcome measured at the appropriate time (chosen in the preclinical research) and clinical outcome?	'Not applicable' /'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
1.5 If the study used a surrogate outcome, did previous experimental studies consistently demonstrate that change in surrogate outcome(s) by a treatment led to change in clinical outcomes?	'Not applicable' /'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
1.6 Did a systematic review with or without meta-analysis demonstrate that the effect of an intervention or a similar intervention on a preclinical model was similar to that in humans?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
Domain 2: Experimental design and analysis	**Low concern/Moderate concern/High concern**
2.1 Did the authors describe sample size calculations?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
2.2 Did the authors plan and perform statistical tests taking the type of data, the distribution of data, and the number of groups into account?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
2.3 Did the authors make adjustment for multiple hypothesis testing?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
2.4 If a dose-response analysis was conducted, did the authors describe the results?	'Not applicable' /'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
2.5 Did the authors assess and report accuracy?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
2.6 Did the authors assess and report precision?	'Yes'/ 'Probably yes'/ 'Probably

	no'/ 'No'/ 'No information'
2.7 Did the authors assess and report sampling error?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
2.8 Was the measurement error low or was the measurement error adjusted in statistical analysis?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
Domain 3: Bias (internal validity)	**Low concern/Moderate concern/High concern**
3.1 Did the authors minimise the risks of bias such as selection bias, confounding bias, performance bias, detection bias, attrition bias, and selective outcome reporting bias?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
Domain 4: Reproducibility of results in a range of clinically relevant conditions (external validity)	**Low concern/Moderate concern/High concern**
4.1 Were the results reproduced with alternative preclinical models of the disease/condition being investigated?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
4.2 Were the results consistent across a range of clinically relevant variations in the model?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
4.3 Did the authors report take existing evidence into account when choosing the comparators?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
Domain 5: Replicability of methods and results in the same model	**Low concern/Moderate concern/High concern**
5.1 Did the authors describe the experimental protocols/methods sufficiently to allow their reproduction?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
5.2 Did an independent group of researchers reproduce the experimental protocols/methods?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
5.3 Did the authors or an independent group of researchers reproduce the results in similar and different laboratory conditions?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
Domain 6: Implications of the study findings (study conclusions)	**Low concern/Moderate concern/High concern**
6.1 Did the authors' conclusions represent the study findings, taking its limitations into account?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
6.2 Did the authors provide details on additional research required to conduct first-in-human studies?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
Domain 7: Research integrity	**Low concern/Moderate concern/High concern**
7.1 Did the authors or the research team obtain ethical approvals and any other regulatory approvals required to	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'

perform the research prior to the start of the study?	
7.2 Did the authors take steps to prevent unintentional changes to data?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
Domain 8: Research transparency	**Low concern/Moderate concern/High concern**
8.1 Did the authors describe the experimental procedures sufficiently in a protocol that was registered prior to the start of the research?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
8.2 Did the authors describe any deviations from the registered protocol?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'
8.3 Did the authors provide the individual subject data along with explanation for any numerical codes/substitutions or abbreviations used in the data to allow other groups of researchers to analyse?	'Yes'/ 'Probably yes'/ 'Probably no'/ 'No'/ 'No information'

2