# Appendices

## Appendix 1: Agreement on domains and questions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Current version** | | **Original version as per protocol** | **Consensus in first round** | **Consensus in second round** | **Consensus in third round** |
| **Domainsa,b,c** | |  |  |  |  |
| 1. Clinical translatability of results to human disease or condition (construct validity) | | 1. Generalisability of results to human disease or condition (construct validity) | - | 73.3% | 86.7% |
| 2. Reproducibility of results in a range of clinically relevant conditions (external validity) | | 2. External validity | - | 66.7% | 86.7% |
| 3. Bias (internal validity) | | 3. Internal validity | - | 80.0% | 93.3% |
| 4. Experimental design and data analysis | | 7. Appropriate analysis and interpretation | - | 93.3% | 100.0% |
| 5. Reproducibility and replicability of methods and results in the same model | | 5. Reproducibility of results | - | 80.0% | 85.7% |
| 6. Implications of the study findings (study conclusions)d | | - | - | 91.7% | 92.3% |
| 7. Research integrity | | 6. Research integrity | - | 85.7% | 91.7% |
| 8. Research transparencyd | | - | - | 100.0% | 100.0% |
| - | | 4. Measurement errorse | - | 100.0% | 100.0% |
| **Signalling questions** | |  |  |  |  |
| 1.1 Did the authors use a model that adequately represents the human disease? | | 1.1 Are the results from the model transferable to humans? | 68.4% | 76.5% | 76.5% |
| 1.2 Did the authors sufficiently identify and characterise the model? | | 2.1 Is the model used in the preclinical research really the model that the researcher wants to study? | 55.6% | 64.7% | 70.6% |
| 1.3 Were the method and timing of the intervention in the specific model relevant to humans? | | 1.2 Are the dose, route of administration, and timing of the intervention transferable to humans? | 84.2% | 100.0% | 100.0% |
| 1.4 If the study used an 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? | | 1.3 Is the correlation between the surrogate outcome measured at the appropriate time (chosen in the preclinical research) and clinical outcome strong, consistent and independent? | 89.5% | 82.4% | 82.4% |
| 1.5 If the study used an 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? | | 1.4 Have randomised trials involving the same or different class of drugs consistently demonstrated that improvement in surrogate outcome(s) led to improvement in clinical outcomes? | 83.3% | 87.5% | 87.5% |
| 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? | | 1.5 Did a systematic review demonstrate that the effect of an intervention on a preclinical model is similar to that in humans? | 61.1% | 66.7% | 73.3% |
| 2.1 Did the authors describe sample size calculations? | | 7.1 Was an appropriate level of random error chosen for sample size calculations? | 84.2% | 88.2% | 88.2% |
| 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? | | 7.3 Were appropriate statistical tests performed? This depends upon the type of data, the distribution of data, and the type of comparison. | 95.0% | 100.0% | 100.0% |
| - | | 7.5 In the case of multivariate analysis, were the assumptions of analysis true?f | 89.5% | 93.8% | 81.3% |
| 2.3 Did the authors make adjustment for multiple hypothesis testing? | | 7.4 Was adjustment made for multiple testing? | 84.2% | 93.8% | 94.1% |
| 2.4 If a dose-response analysis was conducted, did the authors describe the results? | | 7.6 If a dose-response relation is feasible, was the dose-response relationship reported which strengthens the association and conclusions? | 77.8% | 86.7% | 88.2% |
| 2.5 Did the authors assess and report accuracy? 2.6 Did the authors assess and report precision? 2.7 Did the authors assess and report sampling error? | | 4.1 Has the measurement error been assessed appropriately?g | 85.0% | 82.4% | 82.4% |
| 2.8 Was the measurement error low or was the measurement error adjusted in statistical analysis? | | 4.2 Has the measurement error been adjusted appropriately? | 65.0% | 70.6% | 70.6% |
| 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? | | 3.1 SYRCLE's risk of bias tool for animal research (in vivo research) should be used to assess the risk of bias in animal research | 68.4% | 88.2% | 88.2% |
| 4.1 Were the results reproduced with alternative preclinical models of the disease/condition being investigated? | | 2.3 Can the results be reproduced with a different pre-clinical model of the disease condition investigated? | 78.9% | 87.5% | 93.8% |
| 4.2 Were the results consistent across a range of clinically relevant variations in the model? | | 2.2 Are the results generalisable across different types of individuals of the model? | 88.9% | 93.8% | 93.8% |
| 4.3 Did the authors report take existing evidence into account when choosing the comparators? | | 2.4 Has an appropriate comparator been chosen? | 73.7% | 93.8% | 87.5% |
| 5.1 Did the authors describe the experimental protocols/methods sufficiently to allow their replication? | | 5.1 Are the experimental procedures described sufficiently to allow reproduction of the results? | 94.7% | 100.0% | 100.0% |
| 5.2 Did an independent group of researchers replicate the experimentald protocols/methods? | | - | - | 100.0% | 100.0% |
| 5.3 Did the authors or an independent group of researchers reproduce the results in similar and different laboratory conditions? | | 5.2 Were the results reproduced under the same laboratory conditions?h | 90.0% | 94.1% | 100.0% |
| - | | 5.3 Were the results reproduced under different laboratory conditions?h | 75.0% | 82.4% | 88.2% |
| - | | 5.5 Were the results reproduced by a different group of researchers or was this a study that showed reproducibility of results of experiments by a different group of researchers?h | 80.0% | 82.4% | 82.4% |
| 6.1 Did the authors’ conclusions represent the study findings, taking its limitations into account? | | 7.7 Were appropriate conclusions reached? | 80.0% | 88.2% | 88.2% |
| 6.2 Did the authors provide details on additional research required to conduct first-in-human studies?d | | - | - | 83.3% | 76.9% |
| 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?d | | - | - | 84.6% | 84.6% |
| 7.2 Did the authors take steps to prevent unintentional changes to data? | | 6.3 Were appropriate steps taken to prevent inadvertent errors in data collection? | 60.0% | 76.5% | 70.6% |
| 8.1 Did the authors describe the experimental procedures sufficiently in a protocol that was registered prior to the start of the research? 8.2 Did the authors describe any deviations from the registered protocol? | | 6.1 Were the experimental procedures described sufficiently in a registered protocol and any deviations from the protocol described appropriately?i | 85.0% | 94.1% | 94.1% |
| 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? | | 5.4 Is the raw data available for other groups of researchers to analyse? | 85.0% | 88.2% | 88.2% |
| - If computer codes were used for analysis, did the authors make them available to allow reanalysis of the data?d,j | | - | - | 66.7% | 71.4% |
| **Signalling questions for which consensus was not reached** | |  |  |  |  |
|  | | | | | |
| - | | 3.2 A new risk of bias tool for in-vitro research should be developed and used to assess the risk of bias in in-vitro research. | 57.9% | 53.3% | 60.0%k |
| - | | 6.2 Was the research overseen by an independent steering committee? | 55.0% | 58.8% | 56.3% |
| - | | 7.2 Were the observed differences and variation in the outcome similar to the difference used for sample size calculations? | 73.7% | 76.5% | 25.0% |

aThe first number indicates the domain number. The second number (after the decimal point) indicates the number of the signalling question under the domain.

bWe have ordered the tool according to the finally agreed version.

cIn the first round, the agreement on overall structure of the domain was sought. So, there are no figures available for individual domains. In subsequent rounds, agreement on each domain was sought.

dThis is a new domain or signalling question

eThe most frequently preferred option of the Delphi panel was to include the questions under this domain under ‘Experimental design and analysis’. Therefore, this domain does not feature in the final tool despite consensus agreement that the questions under domain were important.

fThe most frequently preferred option of the panel members was to include this question in the explanation for question 6.3 current version, 7.3 initial version. Therefore, this question was removed despite achieving consensus.

gThe most frequently preferred option of the panel members was to split this signalling question into three different signalling questions.

hThree questions (original protocol version 5.2, 5.3, and 5.5) were combined into a single signalling question as this was the most frequently preferred option by the panel

iThe most frequently preferred option of the panel members was to split this signalling question into two different signalling questions.

jAlthough, this was added as a new signalling question and a consensus reached on inclusion, it was decided by the panel members that this can be combined with 8.3.

kThis reflects the lack of consensus about including in vitro research for this tool rather than the lack of consensus about the necessity for a tool to assess the reliability of in vitro research.

## Appendix 2 Determining association between surrogate outcome and clinical outcome

There are many ways of determining the association between two variables (Lee Rodgers and Nicewander 1988). Two common ways of determining the association between a surrogate outcome and a clinical outcome (or its animal equivalent) are correlation coefficient and relative risk ratio (or relative odds ratio) of the treatment effect of the surrogate outcome compared with that of the clinical outcome (Institute for Quality and Efficiency in Health Care 2011, Ciani, Buyse et al. 2013). There are no universally accepted threshold values for interpretation of high association for neither of these measures.

Correlation coefficients range from -1.00 to +1.00. Values close to -1.00 or +1.00 indicate high negative or positive correlation, while values close to 0 indicate lack of correlation between two variables. Institute for Quality and Efficiency in Health Care suggests that the correlation can be considered high when lower limit of the 95% confidence interval (CI) of the correlation coefficient is ≥ 0.85, low when the upper limit of the 95% CI is ≤ 0.70, and moderate otherwise (Institute for Quality and Efficiency in Health Care 2011).

There is also no guidance on the interpretation of relative odds ratio or relative risk ratio values. Values close to 1.00 indicate that the treatment effects of surrogate outcomes and clinical outcomes are similar. Values below 1.00 indicate overestimation of treatment effects and values above 1.00 indicate underestimation of treatment effects by the surrogate outcome for a ‘bad outcome’ (an outcome where lower proportions or values are preferable, for example proportion dead or symptom score). The reverse is true for a ‘good outcome’ (an outcome where higher proportions or values are preferable, for example, cure or health-related quality of life). One can find whether the differences in the effect estimates between the surrogate outcome and clinical outcome are statistically significant by using methods described by Altman et al. to compare the effect estimates (Altman and Bland 2003), although some adjustments have to be made to the calculations of the pooled standard error because the estimates have been obtained from the same samples (Borenstein, Hedges et al. 2009).

## Appendix 3 Examples of surrogate outcomes that failed to be a good substitute for clinical outcomes

There are several examples of treatments that cause an improvement in the surrogate outcome (which is correlated with the clinical outcome) but actually result in harm to patients. Ventricular ectopic beats are associated with adverse prognosis in patients with myocardial infarction and class 1 antiarrhythmic agents effectively suppress ventricular arrhythmias in animals and humans; however, these drugs increased human mortality when evaluated in randomised controlled trials (Bucher, Guyatt et al. 1999). This probably resulted in tens of thousands of deaths in people with non-lethal arrhythmias because of the reliance on the surrogate outcomes (Bucher, Guyatt et al. 1999). In people with heart failure, angiotensin-converting enzyme inhibitors improved exercise capacity and decreased mortality as demonstrated by randomised controlled trials, but other classes of drugs such as milrinone and epoprostenol caused increased cardiovascular mortality despite improving the exercise tolerance (Bucher, Guyatt et al. 1999). In diabetes, glycosylated haemoglobin (HbA1c) is used as a surrogate marker for clinical outcomes. However, some oral hypoglycaemic drugs that reduce HbA1c increase the risk of cardiovascular events (Yudkin, Lipska et al. 2011). In cancer, Kim et al. found that only about 15% of the 36 cancer drugs approved by the US FDA between 2008 and 2012 on the basis of reduction in tumour size or volume, progression-free survival, or disease-free survival improved overall survival (Kim and Prasad 2015). In 50% of the cancer drugs, it was clearly demonstrated that there was no improvement in overall survival with uncertainty remaining in the remaining 35% of the drugs (Kim and Prasad 2015). Rupp et al. evaluated whether the 50% of the cancer drugs with no improvement in overall survival improved the health-related quality of life (HRQoL). They found information on health-related quality of life for only 7 of the 18 drugs: only one of the drugs improved HRQoL; two drugs worsened the HRQoL; there was no difference in HRQoL or inconsistent results about HRQoL for the remaining drugs (Rupp and Zuckerman 2017).

# References

Altman, D. G. and J. M. Bland (2003). "Interaction revisited: the difference between two estimates." BMJ **326**(7382): 219.

Borenstein, M., L. Hedges, J. Higgins and H. Rothstein (2009). "Chapter 4: Effect sizes based on means." Introduction to meta-analysis.

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.

Ciani, O., M. Buyse, R. Garside, T. Pavey, K. Stein, J. A. Sterne and R. S. Taylor (2013). "Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study." BMJ **346**: f457.

Institute for Quality and Efficiency in Health Care (2011). "Validity of surrogate endpoints in oncology: Executive summary of rapid report A10-05, Version 1.1." <https://www.ncbi.nlm.nih.gov/books/NBK198799/> (accessed 24 August 2018).

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

Lee Rodgers, J. and W. A. Nicewander (1988). "Thirteen Ways to Look at the Correlation Coefficient." The American Statistician **42**(1): 59-66.

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

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