



OPEN ACCESS



Check for updates

# Recommendations for evaluating a single surrogate endpoint (REASSurE): consensus based, surrogate evaluation framework with explanation and elaboration

Wang Pok Lo,<sup>1,2</sup> Stuart G Baker,<sup>3</sup> Luca Bertolaccini,<sup>4,5</sup> Sylwia Bujkiewicz,<sup>6</sup> Catey Bunce,<sup>7</sup> Marc Buyse,<sup>8</sup> Oriana Ciani,<sup>9</sup> Robin Christensen,<sup>10,11</sup> Jane Daniels,<sup>12</sup> Sara Gandini,<sup>13</sup> Tamer A Gheita,<sup>14</sup> Peter B Gilbert,<sup>15</sup> Laura Jaeger,<sup>16</sup> Dorcas Kareithi,<sup>17,18</sup> Marissa Lassere,<sup>19</sup> Nicholas T Longford,<sup>20,21</sup> Anthony M Manyara,<sup>22</sup> Koji Oba,<sup>23</sup> Emily K Roberts,<sup>24</sup> Rod S Taylor,<sup>25</sup> Carlos J Toro-Huamanchumo,<sup>26,27</sup> Steff Lewis,<sup>28</sup> Hannah Ensor,<sup>28</sup> Christopher J Weir<sup>28</sup>

For numbered affiliations see end of the article

**Correspondence to:** C J Weir  
christopher.weir@ed.ac.uk  
on Bluesky;  
(ORCID 0000-0002-6494-4903)  
Additional material is published  
online only. To view please visit  
the journal online.

**Cite this as:** *BMJ* 2026;394:e100618  
<http://dx.doi.org/10.1136/bmj-2026-100618>

Accepted: 11 August 2026

The use of a surrogate endpoint to replace a primary outcome relevant to patients in a clinical trial may allow faster approval of a novel treatment, but at the cost of increased uncertainty about treatment benefit. This risk may be amplified if the surrogate endpoint was inadequately or inappropriately evaluated against the patient relevant outcome. Many components of this evaluation currently lack consensus, including the most important non-statistical factors, the choice of statistical method, and whether numerical thresholds should be used. This paper presents a consensus based, multi-stakeholder framework (REASSurE: Recommendations for Evaluating A Single Surrogate Endpoint) containing a list of mandatory and recommended items for surrogate endpoint evaluation, with guidance on how to assess each item.

Statisticians, trials methodologists, regulators, health technology assessment agencies, payers, and clinicians could use the REASSurE framework to achieve a holistic evaluation of a candidate surrogate endpoint, which is vital to ensure that the use of the surrogate endpoint is an appropriate decision.

Surrogate endpoints are often used in lieu of patient relevant outcomes to shorten the duration of clinical trials.<sup>1-3</sup> This decision results from the tension between efficiency and uncertainty in drug development. While the use of a surrogate endpoint may lead to faster approval of a treatment, there is an inherent risk that such a treatment may not provide adequate benefit to patients if a patient relevant outcome had been used to assess its efficacy.<sup>4</sup>

This risk may be alleviated if there is adequate evidence that the effect of the treatment on the surrogate endpoint reliably predicts the effect of the treatment on the patient relevant outcome.<sup>3 5</sup> Conversely, the risk may be increased if there is a lack of evidence, or evidence of poor predictive ability of the surrogate endpoint. Assessment of this risk is part of surrogate endpoint evaluation.<sup>6 7</sup>

This assessment ultimately matters to the patient; for instance, it is within the principles of informed consent that a patient planning to participate in a clinical trial fully understands the risk of participation, and the use of an inadequately assessed surrogate endpoint would increase this risk. As such, Manyara et al<sup>8</sup> further discussed the importance of patient and public involvement in the choice of endpoints during trial design. More generally, such involvement allows patients and the public to provide new perspectives on a specific research goal, which in turn may allow researchers to develop a deeper understanding of the opportunities, risks, and evidence base surrounding the goal (as concluded in an online workshop hosted by the Health Research Board Trials Methodology Research Network, Ireland, on 24 April 2025).

The statistical aspect of surrogate endpoint evaluation quantifies the predictive ability of a

## SUMMARY POINTS

Surrogate endpoint evaluation assesses the extent to which a candidate surrogate endpoint can replace a patient relevant outcome in a clinical trial such that the efficacy of a treatment can be adequately captured using the surrogate

There is currently no consensus on the most important non-statistical factors in surrogate endpoint evaluation, the most appropriate statistical method of surrogate evaluation, or whether numerical thresholds should be used to determine the validity of a candidate surrogate endpoint

The consensus based, multi-stakeholder framework (REASSurE) in this article presents a list of mandatory and recommended items for evaluation, both statistical and non-statistical, and provides guidance on how to assess whether a candidate surrogate endpoint satisfies each item

Statisticians, trials methodologists, regulators, health technology assessment agencies, payers, and clinicians could use the REASSurE framework to holistically evaluate a candidate surrogate endpoint

surrogate endpoint. Many statistical methods for this purpose have been proposed since the pioneering method developed by Prentice,<sup>9</sup> which have given rise to many statistical metrics including the proportion of treatment effect explained<sup>10 11</sup>; joint analyses of the individual level association between the surrogate endpoint and patient relevant outcome, and the trial level correlation between treatment effects on the surrogate endpoint and patient relevant outcome<sup>12</sup>; trial level association alone<sup>13</sup>; surrogate threshold effect<sup>14</sup>; causal effect predictiveness based on principal stratification<sup>15</sup>; individual causal association<sup>16</sup>; and sample size multiplier and true endpoint advantage.<sup>17</sup> No consensus currently exists on which statistical method is best or most appropriate, or whether the predictive ability is adequate if a statistical metric of the surrogate endpoint surpasses a certain numerical threshold.

A key concept in the statistical aspect of surrogate evaluation is the difference between:

- the prognostic association between the surrogate endpoint and the patient relevant outcome (at the level of the individual patient), termed individual level surrogacy (or synonymously, patient level surrogacy),<sup>18</sup> and
- the association between the treatment effect on the surrogate endpoint and the treatment effect on the patient relevant outcome, termed trial level surrogacy.<sup>18</sup>

Box 1 summarises the definitions of a surrogate endpoint, a clinical endpoint (termed patient relevant outcome prior to this point, and used henceforth), individual level analysis, and trial level analysis.

Many other non-statistical aspects also contribute to surrogate endpoint evaluation. The specific biological, clinical, or epidemiological context in which a surrogate endpoint is used may influence the quantity and/or quality of evidence that can be reasonably obtained, thereby affecting the statistical aspect of surrogate evaluation.<sup>1</sup> Unfortunately, consensus is lacking to date on the prioritisation of non-statistical factors for

consideration. These issues can be summarised by the following three questions:

- Issue 1: Which non-statistical factors are most important in surrogate evaluation?
- Issue 2: Should numerical thresholds be used to determine the validity of a surrogate?
- Issue 3: Which statistical method(s) is/are optimal to assess the validity of a surrogate?

Several surrogate evaluation frameworks have been published in the past 15 years. Three notable examples include:

- a three level hierarchy of evidence,<sup>19</sup> which provides a broad classification of levels of evidence from biological plausibility to trial level analysis;
- the Institute for Quality and Efficiency in Health Care (IQWiG) guidelines,<sup>20</sup> which use correlation and the surrogate threshold effect to statistically evaluate a surrogate endpoint (in addition to providing correlation thresholds), and additionally propose criteria to assess the reliability of evidence; and
- the Biomarker-Surrogacy Evaluation Schema,<sup>21</sup> which comprises four domains (study design, target outcome, statistical evaluation, generalisability) and also provides correlation thresholds.

However, none of these established frameworks deals with all three issues. The REASSurE (Recommendations for Evaluating A Single Surrogate Endpoint) project aimed to address all three issues in the form of a new consensus based surrogate evaluation framework that can guide stakeholders to evaluate a surrogate endpoint holistically, accounting for both non-statistical and statistical considerations.

### Development of the REASSurE framework

The REASSurE framework was developed in four stages: (stage 1) a two round online Delphi survey containing items addressing issues 1-3; (stage 2) an international online consensus meeting; (stage 3) piloting of the draft framework; and (stage 4) finalisation of the framework. Five stakeholder groups were targeted in this project: statisticians, trials methodologists, regulators, health technology assessment (HTA) experts, and clinicians. Stakeholders were recruited through emails to personal contacts of WPL, a LinkedIn dissemination post by WPL, and organisations representing each group, including: Statisticians in the Pharmaceutical Industry; International Drug Development Institute; American Statistical Association; Society for Clinical Trials; JISC AllStat mailing list; speakers in the 2022 Banff workshop (on statistical challenges in the identification, validation, and use of surrogate markers);<sup>22</sup> Trials Methodology Research Partnership Outcomes Working Group; Outcome Measures in Rheumatology Surrogate Outcomes Working Group; US Food and Drug Administration Centre for Drug Evaluation and Research; European Medicines Agency; National Cancer Institute in the National Institutes of Health; ISPOR Task Force (Surrogate Endpoint

#### Box 1: Definitions

##### Surrogate endpoint

An endpoint that is used in clinical trials as a substitute for a direct measure of how a patient feels; functions; or survives. A surrogate endpoint does not measure the clinical benefit of primary interest in and of itself; but rather is expected to predict that clinical benefit or harm based on epidemiological, therapeutic, pathophysiological, or other scientific evidence.<sup>28</sup>

##### Clinical endpoint (synonymously termed patient relevant outcome)

A characteristic or variable that reflects how a patient feels, functions, or survives.<sup>29</sup>

##### Individual level analysis

A statistical analysis estimating the prognostic association between the surrogate endpoint and the clinical endpoint.

##### Trial level analysis

A statistical analysis estimating the association between the treatment effect on the surrogate endpoint and the treatment effect on the clinical endpoint.

**Table 1 | Delphi survey and consensus meeting participation numbers by stakeholder group**

|                   | Total No of participants             | Statisticians | Trials methodologists | Regulators | Health technology assessment experts | Clinicians |
|-------------------|--------------------------------------|---------------|-----------------------|------------|--------------------------------------|------------|
| Delphi Round 1    | 62                                   | 37            | 26                    | 8          | 14                                   | 14         |
| Delphi Round 2    | 45                                   | 30            | 21                    | 3          | 9                                    | 9          |
| Consensus meeting | 18 (excluding chair and facilitator) | 13            | 8                     | 3          | 3                                    | 3          |

Counts could overlap because participants may belong to more than one stakeholder group.

Statistical Assessment for HTA Decision Making); Association of Physicians of Great Britain and Northern Ireland; UK Clinical Research Collaboration. The protocol for this project, including the development of the first round of the Delphi survey, is available in the supplementary materials.

The initial set of items for the Delphi survey were sourced from the three surrogate evaluation frameworks listed above,<sup>19-21</sup> in addition to items or themes from three multi-stakeholder, consensus based studies on approaches to evaluate surrogate endpoints.<sup>23-25</sup> Full details are contained in the protocol (see supplementary materials).

Table 1 provides a breakdown of participant numbers by stakeholder group in (stage 1) each round of the Delphi survey, and (stage 2) the consensus meeting. In stage 1, consensus on a survey item was reached if at least 75% of participants selected the same option. Survey sections that required consensus were:

- Section 1 (Definitions): options “yes,” “no,” and “don’t know”;
- Section 2 (Context of use): options “important” (subdivided into “very important” and “important”), “somewhat important,” “not important” (subdivided into “not very important” and “not important at all”), and “don’t know”;
- Section 3 (Statistical thresholds): options “yes, both  $R^2$  and STEP,” “yes, only for  $R^2$ ,” “yes, only for STEP,” “no,” and “don’t know” (in this case, the 75% threshold was assessed by combining the first and second options for  $R^2$ , and combining the first and third options for STEP);
- Section 5 (Interpretation and implementation of statistical methods): either with each option being a selected statistical method, or with options “yes,” “no,” and “don’t know.”

The threshold of 75% is consistent with common practice in consensus studies, where Diamond et al<sup>26</sup> found that the “most common definition for consensus was percent agreement (25 studies), with 75% being the median threshold to define consensus.”<sup>26</sup> All further details regarding rounds 1 and 2, including a full breakdown of votes of each survey item, and free text responses, are available in the supplementary materials.

In stage 2, after the results of round 2 (levels of consensus of each survey item and themes of free text responses) were collated, the lead researcher proposed a first version of the REASSurE framework for discussion in an online consensus meeting involving selected round 2 participants. A document was sent

to consensus meeting delegates (in advance of the meeting), of which a redacted form is available in the supplementary materials. The meeting was held on 26 April 2024 on Microsoft Teams and lasted 3 hours. The meeting involved one chair, one facilitator, and 18 delegates across six continents (four from North America, one from South America, 10 from Europe, one from Africa, one from Asia, and one from Oceania). Consensus on a framework item was assessed using three options: “include in framework” (item is mandatory), “append to framework” (item is recommended), and “exclude from framework” (item is removed). The consensus threshold was set at 75%, where an item would be included if at least 75% of the votes were “include”, and an item would be appended if at least 75% of the votes were either “include” or “append.” Voting was conducted using version 3 of the Online Surveys platform (<https://app.onlinesurveys.jisc.ac.uk/login>).

A revised version of the REASSurE framework was developed after analysing the results of the consensus meeting (key revisions presented in the supplementary materials). This version was then piloted (led by CJT-H), implementing the framework within an HTA to assess the validity of metastasis-free survival as a surrogate for overall survival in non-metastatic, castration resistant, prostate cancer. The pilot testing session involved one moderator, two co-leads, and five session stakeholders, with the stakeholders participating in the voting and feedback components of the meeting. The final evaluations using the framework aligned with the conclusion by the HTA experts of the lack of support for validity of metastasis-free survival as a surrogate. The feedback obtained from the session (key revisions presented in supplementary materials) was used to inform further revision of the framework.

The framework was then finalised, following feedback from consensus meeting delegates on the revised version. The final version is reported in this article: the framework, presented in the form of a table of items; and an explanation and elaboration section that describes the requirements of each item in greater detail.

The REASSurE tabulated framework (table 2) provides guidelines to arbitrating whether a surrogate endpoint is acceptable for use in the development of drugs and medical devices, and/or the health technology assessment of drugs and medical devices. Justifications of decisions should be based, at minimum, on the items presented and their associated grades where applicable. As the weights assigned

Table 2 | REASSurE tabulated framework: recommendations for evaluating a single surrogate endpoint

| Domain or item                                    | Item No | Subitem                                                                                                                                                                            | Grade                                                                                                                                                                                                 |
|---------------------------------------------------|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Domain 0: Context of use</b>                   |         |                                                                                                                                                                                    |                                                                                                                                                                                                       |
| Treatment                                         | 0.1a    | State the indication and study population of the treatment                                                                                                                         | Not applicable                                                                                                                                                                                        |
|                                                   | 0.1b    | State the class of the treatment                                                                                                                                                   | Not applicable                                                                                                                                                                                        |
|                                                   | 0.1c    | State the line of the treatment                                                                                                                                                    | Not applicable                                                                                                                                                                                        |
| Endpoints                                         | 0.2a    | State the surrogate endpoint                                                                                                                                                       | Not applicable‡                                                                                                                                                                                       |
|                                                   | 0.2b    | State the clinical endpoint                                                                                                                                                        | Not applicable‡                                                                                                                                                                                       |
| Phase                                             | 0.3     | State the clinical trial phase in which the surrogate endpoint may be used                                                                                                         | Not applicable                                                                                                                                                                                        |
| <b>Domain 1: Practical motivation</b>             |         |                                                                                                                                                                                    |                                                                                                                                                                                                       |
| Disease prevalence*                               | 1.1     | Disease prevalence*                                                                                                                                                                | Rare; or common                                                                                                                                                                                       |
| Clinical need*                                    | 1.2     | Clinical need*                                                                                                                                                                     | Urgent or unmet; or neither                                                                                                                                                                           |
| Time saved                                        | 1.3     | Time saved using the surrogate endpoint rather than the clinical endpoint relative to the increase in uncertainty in the predicted treatment effect on the clinical endpoint       | Worthwhile; not worthwhile; or undetermined                                                                                                                                                           |
| Other resources saved*                            | 1.4     | Resources saved using the surrogate endpoint rather than the clinical endpoint relative to the increase in uncertainty in the predicted treatment effect on the clinical endpoint* | Worthwhile; not worthwhile; or undetermined                                                                                                                                                           |
| <b>Domain 2: Clinical and biological evidence</b> |         |                                                                                                                                                                                    |                                                                                                                                                                                                       |
| Meta-analyses†                                    | 2.1a    | Quantity                                                                                                                                                                           | Multiple; single; or none                                                                                                                                                                             |
|                                                   | 2.1b    | Quality (risk of bias) of each meta-analysis                                                                                                                                       | Guidance on risk of bias assessment‡                                                                                                                                                                  |
|                                                   | 2.1c    | Sample size (number of trials in each meta-analysis)                                                                                                                               | Not applicable‡                                                                                                                                                                                       |
|                                                   | 2.1d    | Individual level association between the surrogate and clinical endpoints*                                                                                                         | Consistency of association across all meta-analyses: high; moderate; low; or not applicable. State all numerical values and metrics.‡ See also domain 3 for guidance.                                 |
|                                                   | 2.1e    | Trial level association between the treatment effect on the surrogate endpoint and the treatment effect on the clinical endpoint                                                   | Consistency of association across all meta-analyses: high; moderate; low; or not applicable. State all numerical values and metrics.‡ See also domain 3 for guidance.                                 |
| Meta-analyses: Overall conclusion                 |         |                                                                                                                                                                                    | Favours the surrogate endpoint; against the surrogate endpoint; uncertain; or no meta-analysis                                                                                                        |
| Randomised controlled trials†                     | 2.2a    | Quantity                                                                                                                                                                           | Multiple; single; or none                                                                                                                                                                             |
|                                                   | 2.2b    | Quality (risk of bias) of each trial                                                                                                                                               | Guidance on risk of bias assessment‡                                                                                                                                                                  |
|                                                   | 2.2c    | Sample size (number of participants in each trial)                                                                                                                                 | Not applicable‡                                                                                                                                                                                       |
|                                                   | 2.2d    | Individual level association between the surrogate and clinical endpoints*                                                                                                         | Consistency of association across all meta-analyses: high; moderate; low; or not applicable. State all numerical values and metrics.‡ See also domain 3 for guidance.                                 |
|                                                   | 2.2e    | Trial level association between the treatment effect on the surrogate endpoint and the treatment effect on the clinical endpoint                                                   | Consistency of association across all meta-analyses: high; moderate; low; or not applicable. State all numerical values and metrics.‡ See also domain 3 for guidance.                                 |
| Randomised controlled trials: Overall conclusion  |         |                                                                                                                                                                                    | Favours the surrogate endpoint; against the surrogate endpoint; uncertain; or no randomised controlled trials                                                                                         |
| Observational studies*                            | 2.3a    | Quantity                                                                                                                                                                           | Multiple; single; none                                                                                                                                                                                |
|                                                   | 2.3b    | Quality (risk of bias) of each study                                                                                                                                               | Guidance on risk of bias assessment‡                                                                                                                                                                  |
|                                                   | 2.3c    | Sample size (number of participants in each study)                                                                                                                                 | Not applicable‡                                                                                                                                                                                       |
|                                                   | 2.3d    | Individual level association between the surrogate and clinical endpoints*                                                                                                         | Consistency of association across all meta-analyses: high; moderate; low; or not applicable. State all numerical values and metrics.‡ See also domain 3 for guidance.                                 |
|                                                   | 2.3e    | Study level association between the treatment effect on the surrogate endpoint and the treatment effect on the clinical endpoint                                                   | Consistency of association across all meta-analyses: high; moderate; low; or not applicable. State all numerical values and metrics.‡ See also domain 3 for guidance.                                 |
| Observational studies: Overall conclusion         |         |                                                                                                                                                                                    | Favours the surrogate endpoint; against the surrogate endpoint; uncertain; or no observational studies                                                                                                |
| Biological plausibility                           | 2.4     | Suggested biological plausibility factors‡                                                                                                                                         | Good; fair; or limited understanding                                                                                                                                                                  |
| Clinical plausibility                             | 2.5a    | Causality: Surrogate endpoint in relation to the disease process towards the clinical endpoint                                                                                     | Lies directly on the causal pathway; lies on at least one, but not all, of many causal pathways; well correlated with, but does not lie on, a causal pathway; or causal pathways unknown or uncertain |
|                                                   | 2.5b    | Suggested clinical plausibility factors* ‡                                                                                                                                         | Not applicable                                                                                                                                                                                        |
| Generalisability                                  | 2.6a    | Treatment classes                                                                                                                                                                  | Multiple other treatment classes; one other treatment class; or no generalisability                                                                                                                   |
|                                                   | 2.6b    | Risk populations                                                                                                                                                                   | Multiple other risk populations; one other risk population; or no generalisability                                                                                                                    |
| <b>Domain 3: Statistical methods</b>              |         |                                                                                                                                                                                    |                                                                                                                                                                                                       |
| Individual level analysis*                        | 3.1     | Choice of methods                                                                                                                                                                  | Not applicable‡                                                                                                                                                                                       |
| Trial level analysis                              | 3.2a    | R <sup>2</sup>                                                                                                                                                                     | Not applicable‡                                                                                                                                                                                       |
|                                                   | 3.2b    | Surrogate threshold effect*                                                                                                                                                        | Not applicable‡                                                                                                                                                                                       |
|                                                   | 3.2c    | Choice of methods                                                                                                                                                                  | Not applicable‡                                                                                                                                                                                       |
| Thresholding*                                     | 3.3     | R <sup>2</sup> *                                                                                                                                                                   | Not applicable‡                                                                                                                                                                                       |

Overall grades to assign: Is this surrogate endpoint acceptable for use in the development of the drug or medical device, and/or health technology assessment of the drug or medical device? No set of overall grades is proposed due to the importance and flexibility of context. However, any answers to the questions above must be justified with a comprehensive synthesis of evidence from each of domains 0, 1, 2, and 3.

All items without a footnote symbol are mandatory. REASSurE=Recommendations for Evaluating A Single Surrogate Endpoint.

\*Non-mandatory items or subitems.

†At a minimum, either item 2.1 or 2.2 is mandatory.

‡See explanation and elaboration section below.

to the domains and synthesised items are context dependent, no holistic scoring procedure is proposed. Where relevant, qualitative grades are assigned to subitems. As each domain assesses different facets of evidence, using this framework in its entirety is a multi-stakeholder task.

### Explanation and elaboration of the REASSurE framework

The detailed descriptions of framework subitems, their requirements, and limitations must be read and considered alongside the tabulated framework (table 2), which cross references this material. REASSurE is intended to be applied in the context of drugs and medical device development only. The tabulated framework applies only when the final endpoint is a clinical endpoint.

#### Domain 0: Context of use

The items in this domain provide the context of use: the characteristics of the treatment being evaluated, the clinical trial endpoints used, and the phase of the clinical trial that the surrogate endpoint is proposed to be used in. These items are based on the definition of context of use stated within the FDA (US Food and Drug Administration) Biomarker Qualification Programme. In REASSurE, the term “biomarker” in the definition below is replaced by “surrogate endpoint.” The FDA programme states: “The [context of use] is a concise description of the biomarker’s specified use in drug development . . . Note that a drug development use may also include descriptive information such as the patient population, disease or disease stage or model system; stage of drug development; and/or mechanism of action of the therapeutic intervention.”<sup>27</sup>

In REASSurE, we adopt the BEST (Biomarkers, Endpoints, and other Tools) definition of a surrogate endpoint and the NIH (US National Institutes of Health) definition of a clinical endpoint. These definitions are stated below. BEST states: “An endpoint that is used in clinical trials as a substitute for a direct measure of how a patient feels; functions; or survives. A surrogate endpoint does not measure the clinical benefit of primary interest in and of itself; but rather is expected to predict that clinical benefit or harm based on epidemiologic; therapeutic; pathophysiologic; or other scientific evidence.”<sup>28</sup> NIH states: “A characteristic or variable that reflects how a patient feels, functions, or survives.”<sup>29</sup> The time and/or frequency of measurement of both the surrogate and clinical endpoints must be clearly stated in 0.2 endpoints. Only the clinical endpoint that the surrogate endpoint is being evaluated against should be stated in subitem 0.2b.

#### Case study 1

In the phase 3 randomised trial PROTECT, Heerspink et al<sup>30</sup> investigated the efficacy of sparsentan—“a novel, non-immunosuppressive, single-molecule, dual endothelin and angiotensin receptor antagonist” in “adults with IgA nephropathy who continue to have

persistent proteinuria despite receiving maximised treatment with angiotensin-converting enzyme (ACE) inhibitors or angiotensin receptor blockers (ARBs).” As such, sparsentan was evaluated as a first-in-class,<sup>31</sup> second line therapy.

In PROTECT, the surrogate endpoint was the primary endpoint of “change from baseline in urine protein-creatinine ratio based on a 24-h urine sample at week 36.”<sup>30</sup> The clinical endpoint was “kidney failure (confirmed 40% eGFR reduction, end-stage kidney disease, or all-cause mortality) . . . up to 110 weeks from randomisation.”<sup>32</sup> The justification was that the “primary and confirmatory endpoints of the trial, proteinuria reduction and rate of decline in estimated glomerular filtration rate (eGFR), respectively, were selected on the basis of emerging evidence supporting these endpoints as surrogates for kidney failure.”<sup>33–35”</sup><sup>30</sup>

Papapetropoulos et al<sup>36</sup> reported that sparsentan was approved by the FDA under accelerated approval, based on evidence from PROTECT as the pivotal trial. Sparsentan was approved for the following therapeutic use: “proteinuria in adults with primary immunoglobulin A nephropathy at risk of rapid disease progression.”<sup>36</sup>

#### Domain 1: Practical motivation

This domain provides guidance on assessing the practical motivation to use a surrogate endpoint.

##### Item 1.1: Disease prevalence

This item may be considered in surrogate evaluation, because disease rarity may be a rationale for using a surrogate endpoint. Rare diseases often have unique or distinct characteristics compared to common diseases. If this item is considered, these characteristics must be accounted for when holistically evaluating biological evidence. In this framework, rarity is determined solely by prevalence. There is no universal prevalence threshold for a disease to be considered rare; for instance, in the UK, a disease is rare if its prevalence is less than 1 in 2000,<sup>37</sup> whereas in the USA, a disease is rare if fewer than 200 000 people are affected.<sup>38</sup> As such, the geographical region to which the decision to use a surrogate endpoint based on this framework would apply (eg, regional, national, or continental agencies) should be used to determine the prevalence of a disease, and to select the appropriate prevalence threshold. There are also conditions for which regional, national, or continental consensus papers recommend the use of disease incidence to assess rarity, rather than disease prevalence. For example, in the EU, the Joint Action on Rare Cancers by consensus recommended that “the best criterion for a definition of rare cancers is incidence, rather than prevalence,”<sup>39</sup> with the incidence threshold being fewer than six in 100 000 per year.<sup>39</sup> For these conditions, incidence should be used corresponding to the relevant geographical region.

#### Case study 2

Amyotrophic lateral sclerosis is a “rare, serious, and life-threatening disease that causes muscle weakness

and death due to respiratory failure” with an estimated prevalence of “around 16,000 cases overall” in the United States.<sup>40</sup> This estimated prevalence is consistent with that reported by Mehta et al<sup>41</sup> In 2023, the FDA approved tofersen “for the treatment of amyotrophic lateral sclerosis in adults who have a mutation in the superoxide dismutase 1 (SOD1) gene” under the accelerated pathway.<sup>42</sup> The surrogate endpoint of change in plasma neurofilament light chain was used in the pivotal trial<sup>43</sup> that was in turn accepted as a reasonably likely surrogate endpoint by the FDA.<sup>42 44</sup> In this case, the grade is “rare.”

#### Item 1.2: Clinical need

This item may be considered in surrogate evaluation, because clinical need may be a rationale to use a surrogate endpoint. If this item is considered, the risk-benefit trade-off must be evaluated. The cost of not considering a surrogate endpoint may outweigh the risks if clinical need is urgent or unmet, even if the biological, clinical, or statistical performance of the surrogate may not be as strong as desired. However, urgent or unmet clinical need does not necessarily imply that a surrogate endpoint should be accepted regardless of other biological, clinical or statistical considerations. This item is not mutually exclusive to item 1.1 (disease prevalence). For instance, the decision to use a surrogate endpoint in the context of a rare disease with unmet need, despite limited evidence, may be more favourable than that without unmet need (items 1.1 and 1.2). Consulting patients and the public as part of assessing whether clinical need is urgent or unmet in a given context may also be valuable.

#### Continuation of case study 2

“Before FDA [US Food and Drug Administration] approval of tofersen, there were no approved therapies specifically for the treatment of SOD1-ALS [amyotrophic lateral sclerosis (ALS) due to the *SOD1* gene mutation]. Most available treatments for ALS are intended to relieve symptoms such as cramps and spasticity and improve the patient’s quality of life. The three FDA-approved treatments for ALS: (1) riluzole, (2) edaravone, and (3) sodium phenylbutyrate with taurursodiol, do not effectively treat SOD1-ALS, with most patients experiencing continued progression of their disease resulting in respiratory failure and death. Therefore, there was a significant unmet medical need for effective treatments for this familial form of ALS.”<sup>40</sup> In this case, the grade is considered to be “unmet.”

#### Case study 3

In the context of kidney transplantation therapies, Naesens et al<sup>45</sup> supported the use of the conditional marketing authorisation (CMA) pathway by the European Medicines Agency: “Late graft failure (loss of graft function >1 year post transplant) is a condition with unmet medical need. Therefore, CMA should be considered for interventions that demonstrate potential benefits. Late graft failure is a seriously debilitating,

life-threatening disease for which no specific preventive or treatment options are available. CMA of therapies aimed at preventing late graft failure could be based on trials that show benefit on a validated surrogate endpoint for graft failure.” Furthermore, “graft failure is a highly relevant hard endpoint in clinical studies, but it is a late endpoint. This hampers the feasibility of using graft failure as an endpoint in clinical trials that aim at improving late graft failure rates. Surrogates for late graft failure are therefore needed but require robust definitions.”<sup>45</sup> Naesens et al<sup>46</sup> discussed the appropriateness of several such candidate surrogates. In this case, the grade is considered to be “unmet.”

#### Item 1.3: Time saved

This item must be considered in surrogate evaluation, as a trade-off for the increased uncertainty by using a surrogate endpoint. This item refers to the time saved by using the surrogate endpoint rather than the clinical endpoint in a clinical trial. Even though the increase in uncertainty must first be determined statistically using domain 3, this item is placed here owing to its contribution in the practical assessment of a surrogate endpoint.

#### Case study 4

In a meta-analysis, Saad et al<sup>47</sup> evaluated “disease-free survival as a surrogate for overall survival in patients with HER2-positive, early breast cancer in trials of adjuvant trastuzumab for up to 1 year.” The Kaplan-Meier curves (fig 1S in the supplementary appendix of Saad et al<sup>47</sup>) demonstrated that reaching 80% probability of survival required a follow-up time of around 72 months as measured by disease-free survival, compared to around 162 months for overall survival. Therefore, the follow-up time could be shortened by around 7.5 years using disease-free survival as a surrogate in this context.

In terms of increase in uncertainty, Saad et al<sup>47</sup> reported that “[p]atient-level associations between disease-free and overall survival were strong ( $r=0.90$  [95% CI 0.89 [to] 0.90]). Trial-level associations gave rise to values of  $R^2$  of 0.75 (95% CI 0.50 [to] 1.00) for the full set.” Therefore, with substantial time saved and reasonable associations at the individual level and trial level, the grade is considered to be “worthwhile.”

#### Item 1.4: Other resources saved

This item may be considered in surrogate evaluation. This item refers to resources saved (other than time) by using the surrogate endpoint rather than the clinical endpoint to obtain initial approval of the drug or medical device. Resources saved may include functionality saved (from a patient’s perspective) and costs saved (such as trial costs, or costs a patient may incur for clinical care). Evidence of a favourable trade-off may include quantitative modelling procedures such as economic evaluation of cost effectiveness or cost utility. Details on the conduct of cost effectiveness analyses are beyond the scope of this framework, and thus no case study is included here.

**Domain 2: Clinical and biological evidence**

Clinical evidence is based on human epidemiological or clinical studies. Both the variety and quantity of evidence should be considered. Evidence from at least one meta-analysis of several randomised controlled trials is important. One large, high quality observational study is preferable to several observational studies with small sizes and poor quality.

*Item: 2.1 Meta-analyses*

At minimum, either this item or 2.2 (Randomised controlled trials) must be considered in surrogate evaluation.

**Item 2.1a: Quantity**

Classify the number of meta-analyses of randomised controlled trials as multiple (more than one), single, or none. Meta-analyses of observational studies are not considered in this framework. If the number is none, move to item 2.2 (Randomised controlled trials).

**Item 2.1b: Quality**

The ROBIS<sup>48</sup> (Risk of Bias in Systematic Reviews) and ROB-ME<sup>49</sup> (Risk of Bias due to Missing Evidence) assessment tools may be used where a more formal risk of bias assessment is desired.

**Item 2.1c: Sample size**

For each meta-analysis, state the number of randomised controlled trials and the total sample size. Characteristics described in items 0.1 (Treatment) or 2.4 (Biological plausibility) may justify a subgroup analysis by trial.

**Item 2.1d: Individual level association**

If individual level association is available, state all numerical values and metrics of individual level association (together with confidence or credible intervals). The metric, in addition to its value, is important as different metrics give rise to different statistical interpretations. If individual level association(s) are available in more than one meta-analysis, evaluate the consistency of the numerical values of association as high, moderate, or low. For further details, see item 3.1 (Individual level analysis).

**Item 2.1e: Trial level association**

State all numerical values and metrics of trial level association (together with confidence or credible intervals). The metric, in addition to its value, is important as different metrics give rise to different statistical interpretations. If trial level association(s) are available in more than one meta-analysis, evaluate the consistency of the numerical values of association as high, moderate, or low. For further details, see item 3.2 (Trial level analysis).

**Case study 5**

Cabrit et al<sup>50</sup> conducted a meta-analysis of individual patient data that evaluated disease-free survival as a

surrogate endpoint for overall survival in oesophageal cancer:

- 2.1a: The quantity is single.
- 2.1b: Details of the systematic review on which the meta-analysis was based can be found in Faron et al.<sup>51</sup>
- 2.1c: "In total, individual data of 4518 patients from 22 trials was analysed (46 patients in the smallest trial and 802 in the largest)."<sup>50</sup>
- 2.1d: Kendall's  $\tau$  was used as the individual level metric. Across all patients, the values of Kendall's  $\tau$  were 0.75 (0.74 to 0.76), 0.77 (0.75 to 0.84), and 0.70 (0.69 to 0.72) using an unadjusted Clayton copula, an unadjusted Plackett copula, and a Poisson model respectively. Parentheses contain 95% confidence intervals. Cabrit et al<sup>50</sup> further computed Kendall's  $\tau$  in three subgroup analyses. Because only one meta-analysis was available, the consistency grade is considered to be "not applicable."
- 2.1e: Trial level  $R^2$  and the surrogate threshold effect were used as trial level metrics. Across all patients, the values of  $R^2$  were 0.95 (0.89 to 0.99), 0.95 (0.89 to 0.99), and 0.98 (0.90 to 1.00) using an unadjusted Clayton copula, an unadjusted Plackett copula, and a Poisson model, respectively. Parentheses contain 95% confidence intervals. Across all patients, the surrogate threshold effect was 0.86 in all three models. Only one meta-analysis was available, so the consistency grade is "not applicable."
- Overall conclusion: Based on this meta-analysis alone, the grade is "favours the surrogate endpoint."

**Case study 6**

Baechle et al<sup>52</sup> conducted an aggregate data meta-analysis evaluating glycated haemoglobin (HbA1c) as a surrogate endpoint for all cause mortality in type 2 diabetes mellitus.

- 2.1a: The quantity is single.
- 2.1b: Details of the systematic review on which the meta-analysis was based, including risk of bias assessment, can be found in Palmer et al.<sup>53</sup>
- 2.1c: "For the final analysis, we were able to use 346 pairs (i.e., with information on the treatment effect for HbA1c as well as for all-cause mortality) of effect estimates from 205 single trials. Follow-up information could be included for 20 pairs from 13 trials. The 346 pairs of effect estimates were informed by 361 deaths from 122,245 observations".<sup>52</sup>
- 2.1d: No individual patient data were analysed, so no individual level associations were available to report. The consistency grade is therefore "not applicable."
- 2.1e: The slope and intercept (derived from weighted linear regression), and Pearson's correlation coefficient were used as trial level metrics, following the Daniels and Hughes<sup>13</sup> criteria. When risk differences were used to

measure treatment effect, the slope, intercept and correlation coefficient were -0.031% (95% confidence interval -0.179% to 0.117%), 0.036% (-0.076% to 0.148%), and -0.089 (-0.232 to 0.060), respectively. When logarithmic relative risks were used to measure treatment effect, the slope, intercept, and correlation coefficient were 0.129 (-0.043 to 0.302), 0.022 (-0.125 to 0.170), and -0.010 (-0.145 to 0.134), respectively. Only one meta-analysis was available, so the consistency grade is “not applicable.”

- Overall conclusion: Based on this meta-analysis alone, the grade is “against the surrogate endpoint.”

#### *Item 2.2: Randomised controlled trials*

At minimum, either this item or item 2.1 (Meta-analyses) must be considered in surrogate evaluation.

##### **Item 2.2a: Quantity**

Classify the number of randomised controlled trials as multiple (more than one), single, or none. If none, move to item 2.3 (Observational studies). Use item 2.1 (Meta-analyses) for any randomised controlled trials already analysed in a meta-analysis, to avoid duplication of evidence. The quantity should be the number of randomised controlled trials that have not been analysed in a meta-analysis. For example, if six randomised controlled trials were obtained, of which five were analysed in a meta-analysis and detailed in item 2.1 (Meta-analyses), the quantity of randomised controlled trials listed in item 2.2 (Randomised controlled trials) would be “single.”

##### **Item 2.2b: Quality**

For each trial, quality factors to consider may include study design (including assessment of adequate power) and the reporting of study results. The Cochrane RoB 2 (Risk of Bias 2)<sup>54</sup> assessment tool may be used where a more formal risk of bias assessment is desired.

##### **Item: 2.2c: Sample size**

For each trial, state the number of participants. Additionally, state the number of units of analysis if it is different to the number of participants.

##### **Item 2.2d: Individual level association**

If individual level association is available, state all numerical values and metrics of individual level association (together with confidence or credible intervals). The metric, in addition to its value, is important because different metrics give rise to different statistical interpretations. If individual level association(s) are available in more than one randomised controlled trial, evaluate the consistency of the numerical values of association as high, moderate, or low. For further details, see item 3.1 (Individual level analysis).

##### **Item 2.2e: Trial level association**

State all numerical values and metrics of trial level association (together with confidence or credible

intervals). The metric, in addition to its value, is important because different metrics give rise to different statistical interpretations. The grades for consistency apply only (1) when a trial level association is a single trial metric and is available in more than one trial; or (2) if only one trial is available, a trial level association is a multi-trial metric where centres within the trial are used to represent the component trials. In these instances, evaluate the consistency of the values of this metric as high, moderate or low. Otherwise, grading is unnecessary. For further details, see item 3.2 (Trial level analysis).

#### **Case study 7**

D'Amico et al<sup>55</sup> evaluated PSA end (level of prostate specific antigen immediately after treatment) >0.5 ng/mL and PSA nadir (lowest level of PSA reached after treatment) >0.5 ng/mL as candidate surrogate endpoints for prostate cancer specific mortality in localised or locally advanced prostate cancer. In this case study, only the evaluation of PSA end >0.5 ng/mL will be illustrated.

- 2.2a: The quantity is multiple.
- 2.2b: Two randomised controlled trials were analysed: DFCI<sup>56</sup> and TROG.<sup>57</sup> Both trials had at least 80% power. Participants were randomised using a permuted blocks randomisation algorithm in DFCI, whereas participants were randomised by minimisation in TROG. Even though the results of DFCI showed equal allocation, this was not stated in the methodology.<sup>56</sup> In TROG, “an initial randomisation ratio of 2:1:1 was adjusted once it was clear that targets would be reached, to ensure equal final numbers of patients in each group.”<sup>57</sup> Both trials were open label. A more complete risk of bias assessment may be carried out using the Cochrane RoB 2 (Risk of Bias 2)<sup>54</sup> assessment tool.
- 2.2c: The DFCI and TROG trials randomised 206 and 537 participants, respectively.
- 2.2d: Hazard ratios were used as the individual level metrics. In the DFCI trial, the hazard ratio was 4.35 (95% confidence interval 1.46 to 12.94) in the univariable analysis, and the adjusted hazard ratio was 2.76 (1.20 to 6.31) in the multivariable analysis. In the TROG trial, the hazard ratio was 3.22 (1.92 to 5.41) in the univariable analysis, and the adjusted hazard ratio was 2.60 (1.46 to 4.63) in the multivariable analysis. The consistency grade is therefore “high” across the two trials.
- 2.2e: The proportion of treatment effect explained (PTE) and partial proportion of variation explained (pPVE) were used as trial level metrics. In the DFCI trial, the PTE was 0.75 (95% confidence interval 0.48 to 1.19) and the pPVE was 0.0016 (0 to 0.0100). In the TROG trial, the PTE was 0.81 (0.49 to 1.33) and the pPVE was 0.0016 (0 to 0.0140). For both the PTE and pPVE, the consistency grade is therefore “high” across the two trials.
- Overall conclusion: Based on these two trials, the grade is “favours the surrogate endpoint.”

*Item 2.3: Observational studies*

- Features of these studies may be considered in surrogate evaluation:
- Item 2.3a: Quantity. Classify the number of observational studies as multiple (more than one), single, or none.
- Item 2.3b: Quality. For each study, quality factors to consider may include study design and the reporting of study results. The OSQE<sup>58</sup> (Observational Study Quality Evaluation) assessment tool may be used where a more formal risk of bias assessment is desired.
- Item 2.3c: Sample size. For each study, state the number of participants. Additionally, state the number of units of analysis if it is different to the number of participants.
- Item 2.3d: Individual level association. State all numerical values and metrics of individual level association (together with confidence or credible intervals). The metric, in addition to its value, is important because different metrics give rise to different statistical interpretations. If individual level association(s) are available in more than one observational study, evaluate the consistency of the numerical values of association as high, moderate, or low. For further details, see item 3.1 (Individual level analysis).
- Item 2.3e: Study level association. Note that study level is the terminological equivalent to trial level for observational studies. If study level association is available, state all numerical values and metrics of study level association (together with confidence or credible intervals). The metric, in addition to its value, is important because different metrics give rise to different statistical interpretations. The grades for consistency apply only when a study level association is a single study metric that is available from more than one study. In this instance, evaluate the consistency of the values of this metric as high, moderate, or low, otherwise the grades may be ignored. A study level association computed from observational non-randomised study data is insufficient for surrogate evaluation, because evidence from non-randomised studies is insufficient to fully assess the treatment effect on the surrogate endpoint or the clinical endpoint. For further details, see item 3.2 (Trial level analysis).

**Case study 8**

Bonati et al<sup>59</sup> conducted a prospective observational study evaluating mean fat fraction (MFF) of thigh muscle, measured by quantitative muscle MRI (qMRI) as a surrogate endpoint for motor function measure (MFM) in Duchenne muscular dystrophy (DMD).

- 2.3a: The quantity is single.
- 2.3b: The OSQE<sup>58</sup> was used for quality assessment. Because the study protocol of Bonati et al<sup>59</sup> was based on Fischmann et al,<sup>60</sup> the earlier article was also used to assess quality. The results are presented in the supplementary materials, where

an aggregate score of 8 was determined from 16 questions (1 point per question). Areas where points were lost included insufficient information on inclusion/exclusion criteria and participant background characteristics, a short observational period (which Bonati et al<sup>59</sup> noted as a limitation), and missing data. As a result, this study was not considered to be of “high quality.”

- 2.3c: “Twenty patients with genetically confirmed DMD who participated in a transversal study on muscle MRI were re-enrolled for this additional longitudinal evaluation.”<sup>59</sup>
- 2.3d: Pearson’s product moment correlation coefficient ( $r$ ) was used as the individual level metric to assess the association between MFF and MFM. Across all thigh muscles, the values of  $r$  at baseline and at 1 year follow-up were  $-0.917$  ( $P \leq 0.01$ ) and  $-0.907$  ( $P \leq 0.01$ ), respectively. Across all thigh muscles, the value of  $r$  between the annual change of MFF and the annual change of MFM was  $-0.713$  ( $P \leq 0.01$ ). Bonati et al<sup>59</sup> computed further values of  $r$  by subdividing the MFF of thigh muscles into that of quadriceps, hamstrings, and adductors. Because only one observational study was available, the consistency grade is “not applicable.”
- 2.3e: Study level metrics were absent as no treatment was investigated, and thus the consistency grade is “not applicable.”
- Overall conclusion: Based on this observational study alone, the grade is “uncertain.”

Biological evidence refers to the biological or epidemiological considerations that do not require conduct of, or results of, human epidemiological or clinical studies. The following item provides guidance on assessing the biological background behind the use of a surrogate endpoint.

*2.4 Biological plausibility*

This item must be considered in surrogate evaluation. Well established biological plausibility forms an important part of the justification for use of a surrogate endpoint. Biological factors to consider may include the treatment mechanism of action; disease pathogenesis; the temporal nature of the surrogate endpoint on the biological path; treatment responses from credible animal models; and the biological relationship between the surrogate and clinical endpoints.<sup>61</sup> This list is not intended to be exhaustive, because plausibility always depends on the context of use. This item is not mutually exclusive to items 1.1 (Disease prevalence) and 1.2 (Clinical need). For instance, biological plausibility is irrelevant without clinical need (items 2.4 and 1.2), and it may not always be possible to obtain the same depth of understanding of the pathogenesis of a rare disease compared to that of a more common disease (items 2.4 and 1.1).

**Case study 9**

HIV replicates in infected CD4+ cells, which subsequently die because of immune responses such

as apoptosis and pyroptosis.<sup>62</sup> This increase in plasma HIV RNA leads to immunosuppression through the gradual depletion of CD4+ cells, which eventually leads to death due to opportunistic infections.<sup>63</sup> Plasma HIV RNA is therefore a natural choice of surrogate endpoint for HIV disease progression. Nucleoside reverse transcriptase inhibitors such as zidovudine inhibit HIV replication,<sup>64</sup> and the action of zidovudine on suppressing HIV infection has been demonstrated in a mouse model.<sup>65</sup> The grade is therefore “good understanding.”

### Case study 10

The Cardiac Arrhythmia Suppression Trial (CAST)<sup>66</sup> evaluated the effect of class Ic anti-arrhythmics encainide and flecainide in patients with left ventricular dysfunction following a myocardial infarction. The surrogate endpoint was suppression of ventricular ectopy, and the clinical endpoint was “death or cardiac arrest with resuscitation, either of which was due to arrhythmia.”<sup>66</sup> Even though the anti-arrhythmics successfully suppressed ventricular ectopy, more deaths were reported in the treatment group compared to placebo—an example of the surrogate paradox.<sup>67</sup> The CAST investigators concluded that the “mechanisms underlying the excess mortality during treatment with encainide or flecainide remain unknown.”<sup>66</sup> This example illustrates the risks of using a surrogate endpoint as the primary endpoint when the treatment mechanism of action is poorly understood, so the grade for biological plausibility was therefore “limited understanding.”

### Item 2.5 Clinical plausibility

Items 2.5 (Clinical plausibility) and 2.6 (Generalisability) should be assessed based on clinical evidence; that is, human epidemiological or clinical studies.

- Item 2.5a: Causality. This item must be considered in surrogate evaluation. This item assesses the position of the surrogate endpoint in relation to the disease progression towards the clinical endpoint.
- Item 2.5b: Other. Clinical plausibility depends on context, and the above item 2.5a (Causality) may not be exhaustive. Other items of clinical plausibility to consider may include safety (whether there is a low or high risk of adverse events).

### Case study 11

This case study illustrates item 2.5a (Causality) only. Blette et al<sup>68</sup> conducted a meta-analysis and mediation analysis of individual patient data that evaluated low risk status as a surrogate endpoint for clinical worsening in pulmonary arterial hypertension. In mediation analysis, a larger indirect effect and smaller direct effect between the treatment and clinical endpoint are preferable because they imply that most of the treatment effect has been mediated by the surrogate endpoint. Blette et al<sup>68</sup> found from the mediation analysis that “attainment of low-risk status

was not a strong mediator for the effect of treatment on time to clinical worsening for any of the risk scores, with proportionally small and only marginally significant indirect treatment effects. The proportions of the effects mediated through the risk scores were at best modest, ranging from 0.07 to 0.13.” As a result, it is reasonable to suggest the causality grade be “causal pathways unknown or uncertain.”

### 2.6 Generalisability

This item must be considered in surrogate evaluation, because evidence supporting the use of the surrogate endpoint (from items 2.1 (Meta-analyses), 2.2 (Randomised controlled trials), and 2.3 (Observational studies)) that is generalisable to a) different treatment classes and/or b) different risk populations is both important and desirable. This item may be considered in tandem with items 1.1 (Disease prevalence), 1.2 (Clinical need), and 2.4 (Biological plausibility). For instance, such evidence is of particular importance in the context of urgent clinical need while lacking direct evidence within a given target population as part of the treatment indication. To quantitatively assess levels of generalisability, bayesian methods are emerging regarding sharing information about surrogate relationships across treatment classes,<sup>69</sup> treatments,<sup>70</sup> and indications.<sup>71</sup>

### Case study 12

Lassere et al<sup>21</sup> evaluated blood pressure as a surrogate endpoint for stroke prevention through a systematic review of randomised trials. Table 1 of Lassere et al's article<sup>21</sup> presented evidence from 18 randomised controlled trials, with treatments spanning diuretics, beta blockers, angiotensin converting enzyme inhibitors, calcium channel blockers and angiotensin receptor blockers. In addition, Lassere et al<sup>21</sup> commented that “under assumptions of minimal uncertainty (reliability coefficient 0.9), systolic blood pressure reduction is a B+ surrogate endpoint for stroke prevention as there is good statistical evidence from multiple (n=17) randomised controlled trials, of irreversible morbidity (stroke), across different drug class mechanisms (at least 5) and clinical risk populations (gender, age, ethnicity and primary/secondary prevention).”

- 2.6a: The grade is therefore “multiple other treatment classes,” however, a subgroup analysis would further confirm this.
- 2.6b: The grade is therefore “multiple other risk populations,” however, a subgroup analysis would further confirm this.

### Domain 3: Statistical methods

This domain provides guidance on the types of statistical analysis to conduct and report for surrogate evaluation.

### Item 3.1: Individual level analysis

This item may be considered in surrogate evaluation. An individual level method of analysis investigates

the association between the surrogate and clinical endpoints within each treatment arm, and may also consider estimation of causal parameters that measure how the surrogate links to treatment efficacy against the clinical endpoint. Individual patient data from randomised controlled trials may not always be available, owing to data sharing issues. Real world data can be used equally well to assess the prognostic ability of the surrogate endpoint, but they are generally inadequate to estimate causal individual level treatment effects, because of the lack of randomisation to eliminate confounding in treatment choices. A strong individual level association between the surrogate and clinical endpoints does not necessarily imply that a positive treatment effect on the surrogate endpoint will yield a positive treatment effect on the clinical endpoint.<sup>18–72</sup> Examples of individual level methods of analysis are provided by Ensor et al<sup>73</sup> and Weir and Taylor.<sup>74</sup>

#### *Item 3.2: Trial level analysis*

This item must be considered in surrogate evaluation. In particular, trial level analysis from randomised trials or their meta-analysis is mandatory in surrogate evaluation. A trial level method of analysis investigates the association between the treatment effects on the surrogate endpoint and the treatment effects on the clinical endpoint.

#### **Item 3.2a: $R^2$**

This item must be considered in surrogate evaluation. A trial level  $R^2$  must be presented; however, the methods for deriving this metric are not prescribed. An  $R^2$  value might not be the only statistical metric used to conclude the validity of a surrogate endpoint.

#### **Item 3.2b: Surrogate threshold effect**

The surrogate threshold effect<sup>14</sup> is used to apply surrogate endpoints in a new trial. The existence of surrogate threshold effects may be considered in surrogate evaluation. Mathematically, the surrogate threshold effect exists only when the trial level  $R^2$  value is sufficiently high. When  $R^2$  is too low, its confidence intervals are too wide to define any surrogate threshold effect. In future, there may be alternative metrics developed that measure a similar concept, in which case this item would also apply to these metrics.

#### **Item 3.2c: Further guidance**

It is recommended to use methods that account for uncertainty in treatment effects on both the surrogate endpoint and the clinical endpoint. Further, it is recommended to conduct a sensitivity analysis to test the flexibility of any inherent model based assumptions such as normality. Statistical methodological review articles such as those by Ensor et al,<sup>73</sup> Weir and Taylor,<sup>74</sup> Elliott<sup>75</sup> and National Institute for Health and Care Excellence<sup>76</sup> provide a wide range of trial level methods of analysis. Regarding the mention of credible intervals in items 2.1e and 2.2e (Trial level association) and item 2.3e (Study level association), examples of

bayesian methods of surrogate endpoint evaluation can be found in Daniels and Hughes,<sup>13</sup> Bujkiewicz et al,<sup>77</sup> Bujkiewicz et al,<sup>78</sup> and Wheaton et al.<sup>79</sup>

#### *Item 3.3: $R^2$ thresholding*

This item may be considered in surrogate evaluation. If thresholds are considered as a criterion for statistical validity, consider the two following points: lack of generalisability, where thresholds do not generalise across therapeutic areas; and limitations of use, where thresholds must not be applied in practice without further consideration of the wider package of statistical evidence that may be available. Declaring that  $R^2$  is sufficiently large to recommend use of the surrogate endpoint depends on the context. Declaring that  $R^2$  is sufficiently small to recommend not using the surrogate endpoint may be acceptable with less context. While thresholds may provide guidance, they are, importantly, not informed by the consequences of inappropriately declaring either a sufficiently large or sufficiently small  $R^2$  value.

#### **Conclusion**

In conclusion, REASSurE fully resolves issues 1 and 2, and partly resolves issue 3.

#### **Issue 1: Which non-statistical factors are most important in surrogate evaluation?**

Mandatory factors are a favourable trade-off between time saved versus increased uncertainty, biological plausibility, either evidence from randomised controlled trial(s) or evidence from a meta-analysis, clinical plausibility, and generalisability of a candidate surrogate endpoint. The time-uncertainty trade-off is a qualitative comparison between the reduction in trial duration by using the surrogate endpoint as the primary endpoint, and the statistical association between the surrogate and clinical endpoints. The larger the reduction in the trial duration from using a surrogate endpoint, the sooner the results of the trial can reach new patients; and the weaker the association between the surrogate and clinical endpoints, manifested in wider 95% confidence or credible intervals for predicting the effect of the surrogate endpoint on the clinical endpoint. What makes a substantial reduction in duration can be guided by both the magnitude of the reduction rate and its clinical significance that may in turn be informed by patients' opinions. What makes a substantial association is also subjective, especially as no numerical thresholds are offered (discussed in issue 2 below). Weighing the favourability of this trade-off between earlier results and less precise predictions depends on the application.

Recommended, non-mandatory factors to consider are disease prevalence, clinical need, a favourable trade-off between resources saved (other than time) versus increased uncertainty, and evidence from observational studies. If disease prevalence is considered in the context of a rare disease, the relaxation of criteria for accepting a candidate surrogate endpoint may depend on clinical urgency

(extent of morbidity and/or mortality), and the practical extent to which biological plausibility and/or clinical evidence can be established, in light of the fact that realistic sample sizes may be too small to conduct high quality, randomised controlled trials.

If clinical need is considered, concluding the urgency or unmet nature of this need may depend on the speed of disease progression to mortality or high morbidity in patient groups, and the lack of treatments or insufficiently efficacious treatments approved to counteract this progression. In the context of high unmet clinical need, the relaxation of criteria for accepting a candidate surrogate endpoint may depend on the factors described above for rare diseases, together with time or resources saved and any evidence of generalisability across populations or indications that may support or oppose the use of the surrogate in those patient subgroups with high unmet need that lack direct evidence.

The framework provides no general formula for setting the minimum level of plausibility, evidence, or generalisability (as a function of disease prevalence or clinical need) at which a surrogate can be accepted. The decision to accept or reject a candidate surrogate endpoint in light of mitigating factors such as disease rarity or urgent need must be considered on a case-by-case basis, and a decision must be reached by consensus of all the relevant stakeholder groups.

### **Issue 2: Should numerical thresholds be used to determine the validity of a surrogate?**

Thresholding was only considered (but not made mandatory) in the context of  $R^2$  metrics for surrogacy; that is, the squared correlation between the treatment effect on the surrogate endpoint and the treatment effect on the patient relevant (clinical) outcome. This resolution departs from the IQWiG guidelines<sup>20</sup> and the Biomarker-Surrogacy Evaluation Schema,<sup>21</sup> where in both frameworks, thresholding was a mandatory component of surrogate evaluation. The reassurance that a candidate surrogate endpoint surpassing a certain threshold can bring is acknowledged, but the dangers of thresholding should not be understated. The acceptability of thresholding depends on the context. If thresholds are used, they should not be used alone for decision making.

Therefore, REASSurE does not prescribe a fixed minimum value of any surrogacy metric that is required for surrogate acceptance. Recommending such a value would likely be at odds with the non-statistical components that the framework attempts to harmonise, particularly because the quantitative performance of a surrogate endpoint would likely depend on many factors within its contextual environment described in domains 0, 1, and 2. However, in future, it may be useful to explore the extent to which evidence of statistical generalisability across patient subgroups, indications, or treatments can inform acceptable values of surrogacy metrics in the context of a different patient subgroup, indication, or treatment.

The items of individual level and trial level associations in domain 2 involve stating both the point estimate of each association value and its confidence or credible interval, because the confidence or credible interval may enhance or weaken the evidence conveyed by the point estimate. For instance, an extreme point estimate may markedly hint at weak or strong surrogacy, but too wide a confidence interval would lead to an inconclusive result. As with thresholding, REASSurE does not prescribe fixed numerical values of dispersion that is required for conclusive evidence, because this again depends on contextual factors that may in turn influence pragmatic sample sizes.

It remains unexplored whether the statistical evaluation of a surrogate endpoint should primarily be based on the point estimate, or the lower bound of its corresponding confidence interval. Among the existing surrogacy evaluation frameworks, only the IQWiG guidelines<sup>20</sup> mandate assessment of the lower bound to demonstrate statistical validity, and in specific scenarios, further check whether the entire confidence interval lies above the surrogate threshold effect. These proposals were not discussed during the consensus meeting, so this is an area of potential refinement of the framework in future studies.

### **Issue 3: Which statistical method(s) is/are optimal to assess the validity of a surrogate?**

No consensus was reached on the optimality of a statistical method. However, trial level analysis was agreed to be a mandatory item in surrogate evaluation, with individual level analysis being a recommended (non-mandatory) item. The latter result (that individual level analysis is not a mandatory item) is a consequence of two factors: barriers to sharing individual patient data, such as data becoming unavailable, lack of intellectual property rights (from the researcher's perspective<sup>80</sup>), or lack of public trust (from the patient's perspective<sup>81</sup>); and the fact that a strong individual level association need not imply a strong trial level association. For a trial level analysis, trial level correlation was agreed to be a mandatory item in surrogate evaluation, with the surrogate threshold effect being a recommended (non-mandatory) item.

Ideally, trial level analyses should also be conducted using individual patient data, not only for data accuracy reasons such as data integrity verification, data completeness or measurement error adjustments, but also for the following data analysis reasons:

- some statistical models that provide trial level metrics cannot be based solely on aggregate data,<sup>12</sup> so their use would be unnecessarily ruled out;
- the same surrogate endpoint used across multiple randomised trials may be defined and measured differently in each trial, which cannot be statistically accounted for by standardisation in aggregate analyses; and
- aggregate data are insufficient to perform (sensitivity) analyses within patient subsets or treatment subsets.

As a result, these limitations and potential biases must be borne in mind when evaluating surrogate evidence based on trial level aggregate analyses; for instance, in the context of a therapeutic area in which there is a lack of standardisation of endpoint definitions. Nonetheless, REASSurE does not mandate the use of individual patient data in recognition of current barriers to access described above, nor conversely delineate the sufficiency of an aggregate data analysis.

Discrepancies between the strengths of individual level surrogacy and trial level surrogacy are not unusual. Reasonable/high individual level surrogacy coupled with low trial level surrogacy is a common combination, and is often why candidate surrogate endpoints fail to progress to validity. Explanations and examples are described in Baker and Kramer,<sup>72</sup> and Buyse et al.<sup>18</sup> Because the framework research question (Is this surrogate endpoint acceptable for (1) use in the development of the drug or medical device, and/or (2) health technology assessment of the drug or medical device?) is focused on predicting treatment efficacy, trial level surrogacy (association between treatment effects) is therefore more important than individual level surrogacy (individual prognosis). The case of low individual level surrogacy coupled with reasonable or high trial level surrogacy may be due to existence of alternative causal pathways to the clinical endpoint (so prognosis based on surrogate evidence is unreliable), while the treatment operates primarily on the path of the surrogate endpoint. If the treatment is approved based on this evidence, it is vital to communicate to patients that the effect of the treatment can be predicted well using their surrogate measurements collectively, but individually, having a better surrogate value does not necessarily mean a better outcome, nor conversely.

The REASSurE framework provides a comprehensive and structured approach to surrogate endpoint

evaluation, accounting for statistical and non-statistical aspects that are of importance to multiple stakeholder groups. It is linked to the SPIRIT-Surrogate<sup>8</sup> and CONSORT-Surrogate<sup>82</sup> extension checklists for the reporting of surrogate endpoints in randomised trials. As such, surrogate evaluation using REASSurE would support the completion of SPIRIT/CONSORT-Surrogate for new trials. Table 3 outlines the areas of overlap between the SPIRIT/CONSORT-Surrogate checklists and REASSurE.

The Delphi survey formed the backbone of each domain of REASSurE. From table 1, statisticians formed the majority of participants (59.7%) in round 1, with a relative under-representation of clinicians, health technology assessment experts, and regulators. In the spirit of the concluding message of Piedbois and Miller Crowell,<sup>83</sup> this project has helped bridge the gap in communication between statisticians and clinicians on the topic of surrogate evaluation, since its results accounted for the opinions of both stakeholder groups. However, even though clinicians were included, the imbalance in stakeholder representation suggests that the results may be less generalisable to the overall views of the clinical community, which might have been mitigated through an improved recruitment strategy targeting more clinician groups. In round 2, the overall retention rate was 73% (45/62), although it was variable by stakeholder group: 81% of statisticians and trials methodologists (30/37 and 21/26, respectively), and 64% of health technology assessment experts and clinicians (9/14 in each group) were retained, compared to 38% (3/8) of regulators. During the project, measures were taken to mitigate the dropout rate by ensuring that the survey was not too long or tedious; and sending a maximum of three reminder emails across both rounds. Nonetheless, this imbalance suggests that the items in REASSurE may be skewed in favour of the views of statisticians and trials methodologists, with more uncertainty about the true

**Table 3 | Overlap between SPIRIT/CONSORT-Surrogate and REASSurE**

| SPIRIT-Surrogate item No | CONSORT-Surrogate item No | Item statement                                                                                                                                                                                                                             | Links to REASSurE                                                                                                                                                                                                           |
|--------------------------|---------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 8.1                      | 1b.1, 2.1                 | State (a) that the primary outcome is a surrogate endpoint, and (b) the target outcome(s) whose intervention effect is being substituted for.                                                                                              | Covered by item 0.2 (Endpoints)                                                                                                                                                                                             |
| 12.1                     | 6a.1                      | State the practical or scientific reason(s) for using a surrogate endpoint as a primary outcome.                                                                                                                                           | Covered by domain 1 and item 2.4 (Biological plausibility)                                                                                                                                                                  |
| 12.3                     | 6a.2                      | Justification for selected surrogate endpoint: (a) evidence (or lack thereof) of surrogate endpoint validation; and (b) evidence (or lack thereof) of validity being specific to the context used (eg, intervention, disease, population). | Covered by domain 2 and domain 3                                                                                                                                                                                            |
| 14.1                     | 7.1                       | Clarify if the sample size will be estimated to demonstrate that a minimum effect on the surrogate endpoint would be predictive of a benefit on the target outcome(s).                                                                     | Not covered, but linked to item 3.2b (Surrogate threshold effect)                                                                                                                                                           |
| 22.1                     | 22.2                      | Comment on whether the trial design (including sample size and follow-up period), given the use of a surrogate endpoint, adequately captures the potential harms of the intervention being tested.                                         | Partly covered by item 2.5 (Clinical plausibility); in particular, "Other items of clinical plausibility to consider may include safety (whether there is a low or high risk of adverse events)."                           |
| 31c.1                    | 26.2                      | If surrogate and target outcome data will be collected in the trial, state the open access arrangements for the data for future secondary research.                                                                                        | Not covered, but linked to item 3.1 (Individual level analysis); in particular, "It should be noted that individual patient data from randomised controlled trials may not always be available due to data sharing issues." |

SPIRIT=Standard Protocol Items: Recommendations for Interventional Trials; CONSORT=Consolidated Standards of Reporting Trials; REASSurE=Recommendations for Evaluating A Single Surrogate Endpoint.

range of opinions from regulatory representatives. It will therefore be a useful task to match REASSurE with surrogate evaluation procedures of regulatory agencies across different countries, from which any similarities and differences could be collaboratively discussed.

The consensus meeting gave rise to the finalised set of items of REASSurE, agreed on by 18 stakeholders. While the participation had strong geographical coverage, the stakeholder groups were limited to those targeted in the Delphi survey. A broader approach to recruiting survey participants, such as through established organisations for patients, patient advocates, manufacturers, or pharmaceutical industry representatives, would likely increase the robustness of the framework items and their explanation and elaboration sections. In particular, patients and/or their advocates may have unique insights into what truly constitutes unmet clinical need based on their lived experiences, together with what time or resources saved could mean for them in practice. Manufacturers and pharmaceutical representatives could offer methodologies to assess costs saved, which was the only section of the framework that was out of scope for a case study. The initial design of this project also included a component on patient and public involvement, in which focus groups would be organised online involving patients and clinicians to share their views on the use and safety of surrogate endpoints, by way of the nominal group technique. This component was designed to be separate from this consensus project, as a majority of the survey questions concerned niche statistical issues, and thus the survey may not have been as readily accessible to patients. This component ultimately did not materialise due to shortage of time, so launching this patient-centric project in the near future would aptly complement the findings in REASSurE, especially those related to issue 1.

The REASSurE framework itself is limited by its ability to only evaluate a single surrogate endpoint. In recent years, substantial progress has been made on developing statistical methodologies to evaluate multiple surrogate endpoints against a single clinical endpoint,<sup>84-87</sup> with the idea that a collection of surrogate endpoints may together statistically perform better than a single surrogate endpoint. The evaluation of multiple surrogate endpoints may follow practices different to the recommendations in REASSurE; for instance, multiple surrogacy may affect how the items 2.4 (Biological plausibility) and 2.5a (Causality) in table 2 are defined or graded. No consensus based frameworks exist on multiple surrogacy evaluation to our knowledge. Developing extensions to REASSurE, such as REASSurE-Multiple for multiple surrogate evaluation, may be a natural next step. Furthermore, as the framework was developed by experts across six continents, producing translated versions of REASSurE would further facilitate global efforts to rigorously evaluate candidate surrogate endpoints.

Assessment of the items of REASSurE requires several fields of expertise. We therefore recommend

that researchers use the REASSurE framework in collaboration with all relevant stakeholder groups (recently alternatively termed as interest holder groups<sup>88</sup>) to ensure a well informed, holistic evaluation of a candidate surrogate endpoint before deciding whether to apply it as the primary endpoint of a clinical trial.

#### AUTHOR AFFILIATIONS

<sup>1</sup>Centre for Population Health Sciences, Usher Institute, School of Population Health Sciences, University of Edinburgh, Edinburgh, UK

<sup>2</sup>School of Medicine and Biomedical Sciences, Green Templeton College, University of Oxford, Oxford, UK

<sup>3</sup>Independent researcher, Columbia, MD, USA

<sup>4</sup>Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy

<sup>5</sup>Department of Thoracic Surgery, IEO, European Institute of Oncology IRCCS, Milan, Italy

<sup>6</sup>Biostatistics Research Group, Division of Public Health and Epidemiology, School of Medical Sciences, University of Leicester, Leicester, UK

<sup>7</sup>International Centre for Eye Health (ICEH), London School of Hygiene and Tropical Medicine, London, UK

<sup>8</sup>International Drug Development Institute, Louvain-La-Neuve, Belgium

<sup>9</sup>Centre for Research on Health and Social Care Management, SDA Bocconi, Milan, Italy

<sup>10</sup>Section for Biostatistics and Evidence-Based Research, the Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark

<sup>11</sup>Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, Odense, Denmark

<sup>12</sup>Nottingham Clinical Trials Unit, University of Nottingham, Nottingham, UK

<sup>13</sup>Department of Experimental Oncology, IRCCS European Institute of Oncology, Milan, Italy

<sup>14</sup>Rheumatology Department, Faculty of Medicine, Cairo University, Cairo, Egypt

<sup>15</sup>Fred Hutchinson Cancer Center, Seattle, WA, USA

<sup>16</sup>Office of New Drugs, Center for Drug Evaluation and Research, United States Food and Drug Administration, Silver Spring, MD, USA

<sup>17</sup>Biostatistics Research Group, Newcastle University, Newcastle, UK

<sup>18</sup>Jasiri Cancer Research Foundation, Embu, Kenya

<sup>19</sup>St George Hospital and School of Population Health, University of New South Wales, Sydney, NSW, Australia

<sup>20</sup>SNTL Statistics Research and Consulting, London, UK

<sup>21</sup>Institute of Econometrics, Warsaw School of Economics (SGH), Warsaw, Poland

<sup>22</sup>Global Health and Ageing Research Unit, Bristol Medical School, University of Bristol, Bristol, UK

<sup>23</sup>Interfaculty Initiative in Information Studies, University of Tokyo, Tokyo, Japan

<sup>24</sup>Department of Biostatistics, University of Iowa, Iowa City, IA, USA

<sup>25</sup>Robertson Centre for Biostatistics & Primary Care and General Practice, School of Health and Wellbeing, University of Glasgow, Glasgow, UK

<sup>26</sup>Research Unit for Generation and Synthesis Evidence in Health, Saint Ignatius of Loyola University, Lima, Peru

<sup>27</sup>OBEMET Center for Obesity and Metabolic Health, Lima, Peru

<sup>28</sup>Edinburgh Clinical Trials Unit, Usher Institute, School of Population Health Sciences, University of Edinburgh, Edinburgh EH16 4UX, UK

We thank the organisations Outcome Measures in Rheumatology (OMERACT), and the UKCRC Registered CTU Network for disseminating the first round of the Delphi survey to their respective members. We also thank Jane Blazeby, Sandro Gsteiger, Spencer Hey, Takako Kaneyasu, Lindsay Renfro, Marzia Scelsi, Maria Pia Sormani, Anthony Staines, Florian Stijnen, Christopher Sweeney, Dario Trapani, and Adrian Vickers for their responses to the Delphi survey. We are grateful to Sandra Chavez-Malpartida and Sebastian Medina-Ramirez for their

roles as co-leads and presenters of a pilot testing session testing the framework in the context of a health technology assessment. We thank Derek Stewart for providing his patient perspective on the use of surrogate endpoints. REASSurE project team: WPL (lead researcher); CJW, SL, and HE (PhD supervisory team). REASSurE consensus group: WPL (chair), AMM (facilitator), SGB, LB, SB, CB, MB, OC, JD, SG, TAG, PBG, LJ, DK, ML, NTL, KO, EKR, RST, and CJT-H. This article reflects the views of the authors and should not be construed to represent FDA's views or policies. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

**Contributors:** WPL, CJW, SL, and HE contributed to the project conception and methodology. WPL conducted the investigation, formal analysis and data curation of this project, and wrote the original draft of the manuscript. CJW, SL, and HE formed the supervisory team of this project. All authors contributed to the reviewing and editing of the manuscript draft. All authors approved the final version. WPL is the guarantor. The corresponding author CJW attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

**Funding:** WPL was supported in this work by an MRC-NIHR Trials Methodology Research Partnership PhD studentship funded by the University of Edinburgh, and by a Foulkes Foundation fellowship. SB was funded by the Medical Research Council, methodology research grant (MR/T025166/1); SB received support from the Leicester NIHR Biomedical Research Centre. RC reports that the Section for Biostatistics and Evidence-Based Research, the Parker Institute, Bispebjerg and Frederiksberg Hospital is supported by a core grant from the Oak Foundation (OFIL-24-074). The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

**Competing interests:** All authors have completed the ICMJE uniform disclosure form at [www.icmje.org/disclosure-of-interest/](http://www.icmje.org/disclosure-of-interest/) and declare: support from the Medical Research Council, University of Edinburgh, and Foulkes Foundation for the submitted work. WPL was supported in this work by an MRC-NIHR Trials Methodology Research Partnership PhD studentship funded by the University of Edinburgh, and by a Foulkes Foundation fellowship. CJW was supported in this work by funding from the UK Medical Research Council (project references MR/S014357/1 and MR/V038400/1). SB is a member of the NICE (National Institute for Health and Care Excellence) Decision Support Unit and the NICE Guidelines Technical Support Unit; has served as a paid consultant to NICE, CROs (contract research organisations), and the pharmaceutical industry providing methodological advice; and has received payments for educational events from Roche and the University of Bristol, funding to attend conferences from CROs and Roche, research funding from the European Federation of Pharmaceutical Industries and Associations, and Johnson & Johnson, and research support in kind from AstraZeneca and Roche. MB owns stocks in IDDI, CluePoints, and One2Treat. RC reports personal consulting fees from ZPD (ScanDroitin), Roche Pharmaceuticals, Elos Medtech, Image Analysis (Image Analysis Group, UK), Compass Communications, and Ascendis Pharma, for advice related to biostatistics, trial design, statistical inference, evidence synthesis, or clinical epidemiology; receives personal remuneration for editorial roles as statistical editor for *Osteoarthritis and Cartilage* and *Acta Orthopaedica*; and serves in unpaid scientific and methodological roles within the OMERACT technical advisory group and strategic advisory group, the GRADE Working Group, Cochrane, and on editorial or advisory boards of several journals, including the *Journal of Clinical Epidemiology*, *BMJ Open*, *Arthritis Care & Research*, and *Arthritis Research & Therapy*. KO reports an institutional grant from the Fiscal Year 2025 Research Programme for the Practical Application of Innovative Cancer Medicine; honorariums for lectures and presentations from DOT World, WDB coco, Boehringer Ingelheim Japan, IQVIA, and Chugai Pharmaceutical; and participation on advisory boards for Asahi-Kasei and EA Pharma, all outside the submitted work. RST declares as lead investigator that his institution has received past research funding from the UK Medical Research Council (grant MR/V038400/1) for the development of SPIRIT and CONSORT extensions for surrogate primary outcomes in randomised controlled trials. All other authors declare no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, and no other relationships or activities that could appear to have influenced the submitted work.

**Ethical approval:** This project was approved by the Edinburgh Medical School's research ethics committee on 10 December 2023 (approval reference 22-EMREC-024\_NSA\_01).

**Data sharing:** Data available on reasonable request to WPL ([pok.lo@gtc.ox.ac.uk](mailto:pok.lo@gtc.ox.ac.uk)).

**Transparency:** WPL affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

**Patient and public involvement:** Derek Stewart commented on the manuscript and emphasised the wider importance of involving patients and the public in discussions on surrogate endpoints, due to the risks outlined in the introduction.

**Dissemination to participants and related patient and public communities:** The REASSurE framework will be disseminated to the public through conference presentations, seminars, and social media.

**Provenance and peer review:** Not commissioned; externally peer reviewed.

This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: <http://creativecommons.org/licenses/by/4.0/>.

- Boissel JP, Collet JP, Moleur P, et al. Surrogate endpoints: a basis for a rational approach. *Eur J Clin Pharmacol* 1992;43:235-44. doi:10.1007/BF02333016
- Manyara AM, Davies P, Stewart D, et al. Definitions, acceptability, limitations, and guidance in the use and reporting of surrogate end points in trials: a scoping review. *J Clin Epidemiol* 2023;160:83-99. doi:10.1016/j.jclinepi.2023.06.013
- Ciani O, Manyara AM, Davies P, et al. A framework for the definition and interpretation of the use of surrogate endpoints in interventional trials. *EclinicalMedicine* 2023;65:102283. doi:10.1016/j.eclinm.2023.102283
- Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996;125:605-13. doi:10.7326/0003-4819-125-7-199610010-00011
- Christensen R, Ciani O, Manyara AM, et al. Surrogate endpoints: a key concept in clinical epidemiology. *J Clin Epidemiol* 2024;167:111242. doi:10.1016/j.jclinepi.2023.111242
- Ellenberg S, Hamilton JM. Surrogate endpoints in clinical trials: cancer. *Stat Med* 1989;8:405-13. doi:10.1002/sim.4780080404
- Katz R. Biomarkers and surrogate markers: an FDA perspective. *NeuroRx* 2004;1:189-95. doi:10.1602/neuroRx.1.2.189
- Manyara AM, Davies P, Stewart D, et al. Reporting of surrogate endpoints in randomised controlled trial protocols (SPIRIT-Surrogate): extension checklist with explanation and elaboration. *BMJ* 2024;386:e078525. doi:10.1136/bmj-2023-078525
- Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989;8:431-40. doi:10.1002/sim.4780080407
- Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* 1992;11:167-78. doi:10.1002/sim.4780110204
- Parast L, McDermott MM, Tian L. Robust estimation of the proportion of treatment effect explained by surrogate marker information. *Stat Med* 2016;35:1637-53. doi:10.1002/sim.6820
- Buyse M, Molenberghs G, Burzykowski T, et al. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000;1:49-67. doi:10.1093/biostatistics/1.1.49
- Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Stat Med* 1997;16:1965-82.
- Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharm Stat* 2006;5:173-86. doi:10.1002/pst.207
- Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. *Biometrics* 2008;64:1146-54. doi:10.1111/j.1541-0420.2008.01014.x
- Alonso A, Van der Elst W, Molenberghs G, et al. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics* 2015;71:15-24. doi:10.1111/biom.12245
- Baker SG, Lassere MND, Lo WP. Surrogate endpoint metaregression: useful statistics for regulators and trialists. *J Clin Epidemiol* 2024;175:111508. doi:10.1016/j.jclinepi.2024.111508
- Buyse M, Saad ED, Burzykowski T, et al. Surrogacy Beyond Prognosis: The Importance of Trial-Level Surrogacy. *Oncologist* 2022;27:266-71. doi:10.1093/oncolo/oyac006
- Ciani O, Buyse M, Drummond M, et al. Time to Review the Role of Surrogate End Points in Health Policy: State of the Art and the Way Forward. *Value Health* 2017;20:487-95. doi:10.1016/j.jval.2016.10.011

- 20 Institute for Quality and Efficiency in Health Care. Validity of surrogate endpoints in oncology: Executive summary of rapid report a10-05, version 1.1 [internet], 2011. Institute for Quality and Efficiency in Health Care (IQWiG). Cologne, Germany. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK198799>. 17 August 2025.
- 21 Lassere MN, Johnson KR, Schiff M, et al. Is blood pressure reduction a valid surrogate endpoint for stroke prevention? An analysis incorporating a systematic review of randomised controlled trials, a by-trial weighted errors-in-variables regression, the surrogate threshold effect (STE) and the Biomarker-Surrogacy (BioSurrogate) Evaluation Schema (BSES). *BMC Med Res Methodol* 2012;12:27. doi:10.1186/1471-2288-12-27
- 22 Parast L, Gilbert P, Wu L. *Statistical Challenges in the Identification, Validation, and Use of Surrogate Markers*. Banff International Research Station for Mathematical Innovation and Discovery. 2022.
- 23 Lassere MN, Johnson KR, Boers M, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. *J Rheumatol* 2007;34:607-15.
- 24 Schievink B, Lambers Heerspink H, Leufkens H, et al. The use of surrogate endpoints in regulating medicines for cardio-renal disease: opinions of stakeholders. *PLoS One* 2014;9:e108722. doi:10.1371/journal.pone.0108722
- 25 Ciani O, Grigore B, Taylor RS. Development of a framework and decision tool for the evaluation of health technologies based on surrogate endpoint evidence. *Health Econ* 2022;31:44-72. doi:10.1002/hec.4524
- 26 Diamond IR, Grant RC, Feldman BM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. *J Clin Epidemiol* 2014;67:401-9. doi:10.1016/j.jclinepi.2013.12.002
- 27 FDA. Context of use, 2021. Available from: <https://www.fda.gov/drugs/biomarker-qualification-program/context-use>. 17 August 2025.
- 28 FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource 2016-. Silver Spring (MD): Food and Drug Administration (US). Co-published by National Institutes of Health (US), Bethesda (MD). Available from: <https://www.ncbi.nlm.nih.gov/books/NBK326791>. 17 August 2025.
- 29 Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001;69:89-95. doi:10.1067/mcp.2001.113989
- 30 Heerspink HJL, Radhakrishnan J, Alpers CE, et al. Sparsentan in patients with IgA nephropathy: a prespecified interim analysis from a randomised, double-blind, active-controlled clinical trial. *Lancet* 2023;401:1584-94. doi:10.1016/S0140-6736(23)00569-X
- 31 Chiu AW, Bredenkamp N. Sparsentan: A First-in-Class Dual Endothelin and Angiotensin II Receptor Antagonist. *Ann Pharmacother* 2024;58:645-56. doi:10.1177/10600280231198925
- 32 Rovin BH, Barratt J, Heerspink HJL, et al. Efficacy and safety of sparsentan versus irbesartan in patients with IgA nephropathy (PROTECT): 2-year results from a randomised, active-controlled, phase 3 trial. *Lancet* 2023;402:2077-90. doi:10.1016/S0140-6736(23)02302-4
- 33 Levey AS, Gansevoort RT, Coresh J, et al. Change in Albuminuria and GFR as End Points for Clinical Trials in Early Stages of CKD: A Scientific Workshop Sponsored by the National Kidney Foundation in Collaboration With the US Food and Drug Administration and European Medicines Agency. *Am J Kidney Dis* 2020;75:84-104. doi:10.1053/j.ajkd.2019.06.009
- 34 Thompson A, Carroll K, A Inker L, et al. Proteinuria Reduction as a Surrogate End Point in Trials of IgA Nephropathy. *Clin J Am Soc Nephrol* 2019;14:469-81. doi:10.2215/CJN.08600718
- 35 Inker LA, Heerspink HJL, Tighiouart H, et al. Association of Treatment Effects on Early Change in Urine Protein and Treatment Effects on GFR Slope in IgA Nephropathy: An Individual Participant Meta-analysis. *Am J Kidney Dis* 2021;78:340-349.e1. doi:10.1053/j.ajkd.2021.03.007
- 36 Papapetropoulos A, Topouzis S, Alexander SPH, et al. Novel drugs approved by the EMA, the FDA, and the MHRA in 2023: A year in review. *Br J Pharmacol* 2024;181:1553-75. doi:10.1111/bph.16337
- 37 Department of Health and Social Care. The UK Rare Diseases Framework, 2021. Available from: <https://www.gov.uk/government/publications/uk-rare-diseases-framework/the-uk-rare-diseases-framework>. 17 August 2025.
- 38 FDA. Rare diseases at FDA, 2022. Available from: <https://www.fda.gov/patients/rare-diseases-fda>. 17 August 2025.
- 39 Casali PG, Trama A. Rationale of the rare cancer list: a consensus paper from the Joint Action on Rare Cancers (JARC) of the European Union (EU). *ESMO Open* 2020;5:e000666. doi:10.1136/esmoopen-2019-000666
- 40 FDA. Case Study: Tofersen (Qalsody), 2025. Available from: <https://www.fda.gov/about-fda/accelerating-rare-disease-cures-arc-program/learning-and-education-advance-and-empower-rare-disease-drug-developers-leader-3d#case-studies>. 17 August 2025.
- 41 Mehta P, Raymond J, Nair T, et al. Amyotrophic lateral sclerosis estimated prevalence cases from 2022 to 2030, data from the national ALS Registry. *Amyotroph Lateral Scler Frontotemporal Degener* 2025;26:290-5. doi:10.1080/21678421.2024.2447919
- 42 FDA. Application Number 215887Orig1s000 Integrated Review, 2023. Available from: [https://www.accessdata.fda.gov/drugsatfda\\_docs/nda/2023/215887Orig1s000IntegratedR.pdf](https://www.accessdata.fda.gov/drugsatfda_docs/nda/2023/215887Orig1s000IntegratedR.pdf). 11 December 2025.
- 43 ClinicalTrials.gov. An Efficacy, Safety, Tolerability, Pharmacokinetics and Pharmacodynamics Study of BILB067 (Tofersen) in Adults With Inherited Amyotrophic Lateral Sclerosis (ALS) (VALOR (Part C)). NCT02623699. 2023. Available from: <https://clinicaltrials.gov/study/NCT02623699>. 17 August 2025.
- 44 FDA. Table of Surrogate Endpoints That Were the Basis of Drug Approval or Licensure, 2024. Available from: <https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure>. 17 August 2025.
- 45 Naesens M, Loupy A, Hilbrands L, et al. Rationale for Surrogate Endpoints and Conditional Marketing Authorization of New Therapies for Kidney Transplantation. *Transpl Int* 2022;35:10137. doi:10.3389/ti.2022.10137
- 46 Naesens M, Budde K, Hilbrands L, et al. Surrogate Endpoints for Late Kidney Transplantation Failure. *Transpl Int* 2022;35:10136. doi:10.3389/ti.2022.10136
- 47 Saad ED, Squifflet P, Burzykowski T, et al. Disease-free survival as a surrogate for overall survival in patients with HER2-positive, early breast cancer in trials of adjuvant trastuzumab for up to 1 year: a systematic review and meta-analysis. *Lancet Oncol* 2019;20:361-70. doi:10.1016/S1470-2045(18)30750-2
- 48 Whiting P, Savović J, Higgins JPT, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol* 2016;69:225-34. doi:10.1016/j.jclinepi.2015.06.005
- 49 Page MJ, Sterne JAC, Boutron I, et al. ROB-ME: a tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. *BMJ* 2023;383:e076754. doi:10.1136/bmj-2023-076754
- 50 Cabrit N, Cheugoua-Zanetsie M, Tierney J, et al. Disease-free survival as surrogate for overall survival in esophageal cancer: An individual patient data meta-analysis of neoadjuvant chemotherapy and chemoradiotherapy. *Eur J Cancer* 2025;218:115292. doi:10.1016/j.ejca.2025.115292
- 51 Faron M, Cheugoua-Zanetsie M, Tierney J, et al. Individual Participant Data Network Meta-Analysis of Neoadjuvant Chemotherapy or Chemoradiotherapy in Esophageal or Gastroesophageal Junction Carcinoma. *J Clin Oncol* 2023;41:4535-47. doi:10.1200/JCO.22.02279
- 52 Baechle C, Scherler W, Lang A, et al. Is HbA1c a valid surrogate for mortality in type 2 diabetes? Evidence from a meta-analysis of randomized trials. *Acta Diabetol* 2022;59:1257-63. doi:10.1007/s00592-022-01887-y
- 53 Palmer SC, Mavridis D, Nicolucci A, et al. Comparison of Clinical Outcomes and Adverse Events Associated With Glucose-Lowering Drugs in Patients With Type 2 Diabetes: A Meta-analysis. *JAMA* 2016;316:313-24. doi:10.1001/jama.2016.9400
- 54 Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898. doi:10.1136/bmj.l4898
- 55 D'Amico AV, Chen MH, de Castro M, et al. Surrogate endpoints for prostate cancer-specific mortality after radiotherapy and androgen suppression therapy in men with localised or locally advanced prostate cancer: an analysis of two randomised trials. *Lancet Oncol* 2012;13:189-95. doi:10.1016/S1470-2045(11)70295-9
- 56 D'Amico AV, Chen MH, Renshaw AA, et al. Androgen suppression and radiation vs radiation alone for prostate cancer: a randomized trial. *JAMA* 2008;299:289-95. doi:10.1001/jama.299.3.289
- 57 Denham JW, Steigler A, Lamb DS, et al. Short-term neoadjuvant androgen deprivation and radiotherapy for locally advanced prostate cancer: 10-year data from the TROG 96.01 randomised trial. *Lancet Oncol* 2011;12:451-9. doi:10.1016/S1470-2045(11)70063-8
- 58 Drukker M, Weltens I, van Hooijdonk CFM, et al. Development of a Methodological Quality Criteria List for Observational Studies: The Observational Study Quality Evaluation. *Front Res Metr Anal* 2021;6:675071. doi:10.3389/frma.2021.675071
- 59 Bonati U, Hafner P, Schädelin S, et al. Quantitative muscle MRI: A powerful surrogate outcome measure in Duchenne muscular dystrophy. *Neuromuscul Disord* 2015;25:679-85. doi:10.1016/j.nmd.2015.05.006
- 60 Fischmann A, Hafner P, Gloor M, et al. Quantitative MRI and loss of free ambulation in Duchenne muscular dystrophy. *J Neurol* 2013;260:969-74. doi:10.1007/s00415-012-6733-x

- 61 Temple R. Are surrogate markers adequate to assess cardiovascular disease drugs? *JAMA* 1999;282:790-5. doi:10.1001/jama.282.8.790
- 62 Doitsh G, Greene WC. Dissecting How CD4 T Cells Are Lost During HIV Infection. *Cell Host Microbe* 2016;19:280-91. doi:10.1016/j.chom.2016.02.012
- 63 Swinkels H, Nguyen A, Gulick P. *HIV and AIDS*. StatPearls. 2024 Jul 27.
- 64 Edwards Z, Ingold CJ, Azmat CE. Zidovudine. In: StatPearls [Internet]. 2023 Jun 5. StatPearls Publishing.
- 65 McCune JM, Namikawa R, Shih CC, et al. Suppression of HIV infection in AZT-treated SCID-hu mice. *Science* 1990;247:564-6. doi:10.1126/science.2300816
- 66 Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med* 1991;324:781-8. doi:10.1056/NEJM199103213241201
- 67 Tereshchenko LG, Manyara AM, Ciani O, et al. A call for transparency, improved reporting, and interpretation of trials using surrogate end points in cardiac electrophysiology. *Heart Rhythm* 2025;22:e492-9. doi:10.1016/j.hrthm.2025.02.038
- 68 Blette BS, Moutchia J, Al-Naamani N, et al. Is low-risk status a surrogate outcome in pulmonary arterial hypertension? An analysis of three randomised trials. *Lancet Respir Med* 2023;11:873-82. doi:10.1016/S2213-2600(23)00155-8
- 69 Papanikos T, Thompson JR, Abrams KR, et al. Bayesian hierarchical meta-analytic methods for modeling surrogate relationships that vary across treatment classes using aggregate data. *Stat Med* 2020;39:1103-24. doi:10.1002/sim.8465
- 70 Ren S, Singh J, Gsteiger S, et al. Evaluating amyloid-beta as a surrogate endpoint in trials of anti-amyloid drugs in Alzheimer's disease: a Bayesian meta-analysis. *arXiv preprint*. arXiv:2504.06807. 2025.
- 71 Singh J, Anwer S, Palmer S, et al. Multi-indication Evidence Synthesis in Oncology Health Technology Assessment: Meta-analysis Methods and Their Application to a Case Study of Bevacizumab. *Med Decis Making* 2025;45:17-33. doi:10.1177/0272989X241295665
- 72 Baker SG, Kramer BS. A perfect correlate does not a surrogate make. *BMC Med Res Methodol* 2003;3:16. doi:10.1186/1471-2288-3-16
- 73 Ensor H, Lee RJ, Sudlow C, et al. Statistical approaches for evaluating surrogate outcomes in clinical trials: A systematic review. *J Biopharm Stat* 2016;26:859-79. doi:10.1080/10543406.2015.1094811
- 74 Weir CJ, Taylor RS. Informed decision-making: Statistical methodology for surrogacy evaluation and its role in licensing and reimbursement assessments. *Pharm Stat* 2022;21:740-56. doi:10.1002/pst.2219
- 75 Elliott MR. Surrogate endpoints in clinical trials. *Annu Rev Stat Appl* 2023;10:75-96.
- 76 National Institute for Health and Care Excellence. Surrogate endpoints in cost-effectiveness analysis for use in health technology assessment, 2024. Available from: <https://www.nice.org.uk/what-nice-does/our-research-work/our-projects-and-partners/international-collaboration>. 18 August 2025.
- 77 Bujkiewicz S, Thompson JR, Spata E, et al. Uncertainty in the Bayesian meta-analysis of normally distributed surrogate endpoints. *Stat Methods Med Res* 2017;26:2287-318. doi:10.1177/0962280215597260
- 78 Bujkiewicz S, Achana F, Papanikos T, Riley RD, Abrams KR. NICE DSU Technical Support Document 20: multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints, 2019. Available from: <https://sheffield.ac.uk/sites/default/files/2022-02/TSD20-mvmeta-final.pdf>. 18 August 2025.
- 79 Wheaton L, Papanikos A, Thomas A, et al. Using Bayesian Evidence Synthesis Methods to Incorporate Real-World Evidence in Surrogate Endpoint Evaluation. *Med Decis Making* 2023;43:539-52. doi:10.1177/0272989X231162852
- 80 van Middelkoop M, Lohmander S, Bierma-Zeinstra SMA. Sharing data-taming the beast: barriers to meta-analyses of individual patient data (IPD) and solutions. *Br J Sports Med* 2020;54:822-4. doi:10.1136/bjsports-2019-101892
- 81 Kalkman S, van Delden J, Banerjee A, et al. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. *J Med Ethics* 2022;48:3-13. doi:10.1136/medethics-2019-105651
- 82 Manyara AM, Davies P, Stewart D, et al. Reporting of surrogate endpoints in randomised controlled trial reports (CONSORT-Surrogate): extension checklist with explanation and elaboration. *BMJ* 2024;386:e078524. doi:10.1136/bmj-2023-078524
- 83 Piedbois P, Miller Crosswell J. Surrogate endpoints for overall survival in advanced colorectal cancer: a clinician's perspective. *Stat Methods Med Res* 2008;17:519-27. doi:10.1177/0962280207081865
- 84 Bujkiewicz S, Thompson JR, Riley RD, et al. Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Stat Med* 2016;35:1063-89. doi:10.1002/sim.6776
- 85 Parast L, Cai T, Tian L. Evaluating multiple surrogate markers with censored data. *Biometrics* 2021;77:1315-27. doi:10.1111/biom.13370
- 86 Flórez AJ, Molenberghs G, Van der Elst W, et al. An efficient algorithm to assess multivariate surrogate endpoints in a causal inference framework. *Comput Stat Data Anal* 2022;172:107494. doi:10.1016/j.csda.2022.107494
- 87 Van der Elst W, Ong F, Stijnen F, et al. Multiple Surrogates in the Meta-Analytic Setting for Normally Distributed Endpoints. *Statistics in Biopharmaceutical Research* 2024;17:536-47. doi:10.1080/19466315.2024.2429387
- 88 Akl EA, Khabisa J, Petkovic J, et al. Interest-holders: A new term to replace stakeholders in the context of health research and policy. *Cochrane Evid Synth Methods* 2024;2:e70007. doi:10.1002/cesm.70007

## Supplementary materials