Quantifying the robustness of primary analysis results: A case study on missing outcome data in pairwise and network meta-analysis

Loukia M. Spineli1 | Chrysostomos Kalyvas2 | Katerina Papadimitropoulou3,4

1Midwifery Research and Education Unit, Hannover Medical School, Hannover, Germany
2Biostatistics and Research Decision Sciences, MSD Europe Inc., Brussels, Belgium
3Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
4Data Science and Biometrics, Danone Nutricia Research, Utrecht, The Netherlands

Correspondence
Loukia M. Spineli, Midwifery Research and Education Unit (OE 9210), Hannover Medical School, Carl-Neuberg-Straße 1, 30625 Hannover, Germany. Email: spineli.loukia@mh-hannover.de

Abstract
Conducting sensitivity analyses is an integral part of the systematic review process to explore the robustness of results derived from the primary analysis. When the primary analysis results can be sensitive to assumptions concerning a model’s parameters (e.g., missingness mechanism to be missing at random), sensitivity analyses become necessary. However, what can be concluded from sensitivity analyses is not always clear. For instance, in a pairwise meta-analysis (PMA) and network meta-analysis (NMA), conducting sensitivity analyses usually boils down to examining how ‘similar’ the estimated treatment effects are from different re-analyses to the primary analysis or placing undue emphasis on the statistical significance. To establish objective decision rules regarding the robustness of the primary analysis results, we propose an intuitive index, which uses the whole distribution of the estimated treatment effects under the primary and alternative re-analyses. This novel index is compared to an objective threshold to infer the presence or lack of robustness. In the case of missing outcome data, we additionally propose a graph that contrasts the primary analysis results to those of alternative scenarios about the missingness mechanism in the compared arms. When robustness is questioned according to the proposed index, the suggested graph can demystify the scenarios responsible for producing inconsistent results to the primary analysis. The proposed decision framework is immediately applicable to a broad set of sensitivity analyses in PMA and NMA. We illustrate our framework in the context of missing outcome data in both PMA and NMA using published systematic reviews.

Keywords
Kullback–Leibler divergence, missing outcome data, robustness, sensitivity analysis, systematic review
INTRODUCTION

Undertaking a systematic review involves making several decisions with respect to data collection and quantitative analysis. Sensitivity analysis is a critical element of the systematic review process to explore the impact of different data and/or modelling decisions on the results. If results remain consistent across the alternative re-analyses, they can be considered robust; otherwise, the primary analysis results should be interpreted with great caution. The most commonly used strategy for sensitivity analysis consists of excluding one trial at a time and examining the impact of the removal on the summary treatment effect.1,2 Often, we are more interested in exploring how sensitive the primary analysis results are to certain assumptions about the data, such as the missing at random (MAR) assumption regarding the outcomes of the missing participants.

Addressing missing outcome data (MOD) in the evidence synthesis framework faces many challenges due to the aggregate nature of the extracted data limiting the analysis strategies compared with the options at the patient-level.3 The literature advocates the importance of undertaking principled sensitivity analyses, especially when dealing with MOD, to assess the robustness of the primary analysis results. For instance, the report titled ‘The Prevention and Treatment of Missing Data in Clinical Trials’ by the National Research Council’s Committee on National Statistics explicitly recommends that ‘Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting’.4 Despite the guidance, authors of systematic reviews appear not to pursue sensitivity analysis regularly.

Recent empirical evidence has revealed that most systematic reviews addressing MOD in the primary analysis did not perform sensitivity analysis.5-7 Of those performing sensitivity analysis, the majority applied exclusion or imputation of MOD before the analysis. Imputation concerned binary outcomes and included clinically implausible scenarios for the outcome of missing participants in the compared interventions.5-7 For instance, all missing participants in both intervention arms were assumed to have experienced the outcome. Such sensitivity analyses are inadequate and inappropriate concerning (a) the scenarios for the missingness mechanisms (i.e., the reasons for MOD) and (b) the analysis methods to handle MOD. A proper sensitivity analysis must include a series of gradually stringent yet clinically plausible scenarios for the missingness mechanisms across the compared interventions. Then, an appropriate re-analysis of the data must include the simultaneous modelling of missing and observed outcomes for the different scenarios about the missingness mechanisms across the compared interventions while accounting for the uncertainty due to MOD.

To initiate a paradigm shift in the analysis of MOD in systematic reviews, we propose a novel framework for sensitivity analysis that promises to engage the researchers in a proper and thorough investigation of the impact of MOD on the pairwise meta-analysis (PMA) and network meta-analysis (NMA) results. To objectively infer the robustness of the primary analysis results, we developed an index that measures the overall ‘distance’ between the primary analysis results and the results obtained under a series of alternative re-analyses. In the present study, we illustrated this index within PMA and NMA for different scenarios about the missingness mechanism in the compared interventions. This index
considers the location and uncertainty of the estimated summary treatment effects under the different scenarios for the missingness mechanism. Hence, this index is potentially superior to alternative measures that merely use the magnitude and direction of the estimated treatment effects, or the statistical significance to infer the robustness of the primary analysis results. This index is particularly attractive in NMA, where the number of compared interventions renders the application of the sensitivity analysis in every pairwise comparison tedious.

Our proposed framework additionally includes a comprehensive and user-friendly graphical tool that illustrates the estimated parameter of interest (e.g., summary treatment effect), and uncertainty thereof, under progressively stringent yet clinically relevant scenarios for the missingness mechanisms using a single analysis. This tool aims to replace current suboptimal sensitivity analysis approaches, which (a) rely on clinically extreme scenarios for the missingness mechanisms in the compared interventions, and then, (b) apply these scenarios in separate analyses.

We have structured the article as follows. Section 2 introduces two motivating examples: a Cochrane review on a continuous primary outcome and a published systematic review with NMA on a binary primary outcome. In Section 3, we briefly describe the modelling framework for the appropriate handling of aggregate MOD (i.e., one-stage pattern-mixture model); then, we present the robustness index and the proposed graphical tool. In Section 4, we apply the proposed sensitivity analysis framework to the motivating examples. Discussion of our framework in the light of relevant published literature is given in Section 5. Brief recommendations for best practice in the analysis of MOD in systematic reviews are provided in Section 6.

2 | MOTIVATING EXAMPLES

To investigate whether different scenarios for the missingness mechanism across the compared interventions may compromise the robustness of the primary analysis results with respect to treatment effectiveness, we considered the following motivating examples. Both examples refer to a harmful outcome; therefore, a negative value of the effect measure (see Section 3.6) favours the first intervention of the comparison, whereas a positive value of the effect measure favours the second intervention of the comparison.

2.1 | A pairwise meta-analysis on a continuous outcome

Taylor et al. compared inositol with glucose (placebo) in terms of the resolution of a depressive episode using the Hamilton Depression Rating Scale. We selected this Cochrane review from previous work for having at least a moderate number of MOD that were unbalanced in the compared arms in all included trials. Namely, both the percentage of total MOD and the difference in the percentage of MOD in the compared arms were more than 5% in every trial (Table 1 and Tables S1 and S2 in supporting information).

2.2 | A network meta-analysis on a binary outcome

Baker et al. compared seven pharmacologic treatments with each other and placebo in chronic obstructive pulmonary disease (COPD) patients (Figure S1, supporting information). We focused on the exacerbation of COPD as the primary binary outcome. The authors ‘lumped’ the interventions in four classes, but we considered the interventions separately as reported in the original trials. For illustrative purposes, this network was selected from a previous work for having at least a moderate number of MOD in most trials and observed comparisons (Table 2 and Table S3 in supporting information).

3 | METHODS

3.1 | One-stage pattern-mixture model for binary and continuous outcomes

Consider a collection of N trials with different sets of T interventions for a pre-specified population and outcome (binary or continuous). In arm k of trial i, we extract information on the number of participants completing the trial (hereinafter named completers), cik, the number randomised, nik, and the measured outcome (e.g., the number of participants responding to the randomised intervention, or the mean outcome and standard deviation as measured among the completers). Using a pattern-mixture model, we model missing and observed outcomes simultaneously, and hence, we maintain the randomised sample. In arm k of trial i, we define the pattern-mixture model as follows:

\[ \theta_{ik} = \theta_{ik}^c (1 - q_{ik}) + \theta_{ik}^m q_{ik}, \]

where \( \theta_{ik} \) is the underlying outcome in arm k of trial i given the randomised participants, \( \theta_{ik}^c \) and \( \theta_{ik}^m \) are the underlying outcomes given the completers and missing participants, respectively, and \( q_{ik} \) is the probability of MOD. In the case of a binary outcome,
TABLE 1  Inositol versus placebo for depression (continuous outcome)\(^8\)

| Trial          | Inositol | Placebo | Total MOD (%) | Diff. in %MOD |
|----------------|----------|---------|---------------|---------------|
|                | y\(^o\)  | s\(^o\) | m c           |               |
| Levine 1995    | 21.6     | 10      | 8 13          |               |
| Chengappa 2000 | -7.4     | 6.4     | 0 12          |               |
| Levine 1999    | 20.4     | 13      | 5 13          |               |
| Nemets 1999    | 21.7     | 11      | 5 18          |               |

Note: Values in bold indicate a substantial risk of attrition bias (>20%), and values in italics indicate a moderate risk of attrition bias.
Abbreviations: c, number of completers; Diff., absolute difference; m, number of missing outcome data (MOD); s\(^o\), standard deviation given the completers; y\(^o\), mean given the completers.

TABLE 2  Distribution of %MOD in each comparison and intervention of the network\(^9\)

| Characteristic | Trial | Minimum | First quartile | Median | Third quartile | Maximum |
|----------------|-------|---------|----------------|--------|----------------|---------|
| Comparisons with placebo |       |         |                |        |                |         |
| BUD vs. PBO    | 1     | 38      | 38             | 38     | 38             | 38      |
| BUD+ vs. PBO   | 1     | 36      | 36             | 36     | 36             | 36      |
| FOR vs. PBO    | 3     | 16      | 20             | 23     | 31             | 38      |
| FLU vs. PBO    | 4     | 6       | 8              | 20     | 32             | 39      |
| FLU+ vs. PBO   | 4     | 6       | 8              | 20     | 32             | 35      |
| SAL vs. PBO    | 9     | 5       | 12             | 15     | 20             | 33      |
| TIO vs. PBO    | 5     | 6       | 10             | 11     | 13             | 15      |
| Comparisons with an active intervention |       |         |                |        |                |         |
| FLU+ vs. FLU   | 2     | 28      | 30             | 32     | 34             | 36      |
| SAL vs. FLU    | 2     | 29      | 31             | 32     | 33             | 34      |
| FOR vs. BUD    | 1     | 32      | 32             | 32     | 32             | 32      |
| BUD+ vs. BUD   | 1     | 30      | 30             | 30     | 30             | 30      |
| BUD+ vs. FOR   | 1     | 30      | 30             | 30     | 30             | 30      |
| FLU+ vs. SAL   | 3     | 3       | 17             | 30     | 30             | 30      |
| SAL vs. TIO    | 2     | 7       | 8              | 9      | 10             | 11      |
| FOR vs. TIO    | 1     | 1       | 1              | 1      | 1              | 1       |
| Interventions  |       |         |                |        |                |         |
| BUD            | 1     | 31      | 31             | 31     | 31             | 31      |
| BUD+           | 1     | 28      | 28             | 28     | 28             | 28      |
| FLU+           | 4     | 5       | 10             | 21     | 30             | 31      |
| FLU            | 4     | 0       | 10             | 20     | 30             | 40      |
| PBO            | 19    | 5       | 13             | 18     | 22             | 44      |
| FOR            | 4     | 1       | 11             | 18     | 24             | 32      |
| SAL            | 10    | 2       | 9              | 14     | 18             | 32      |
| TIO            | 7     | 1       | 4              | 8      | 9              | 12      |

Abbreviations: BUD, budesonide; BUD+, budesonide plus formoterol; FLU, fluticasone; FLU+, fluticasone plus salmeterol; FOR, formoterol; MOD, missing outcome data; PBO, placebo; SAL, salmeterol; TIO, tiotropium.
\[ \theta_{ik} = P(I_{ikj} = 1|M_{ikj} = 1 \cup M_{ikj} = 0), \]

\[ \theta_{ik}' = P(I_{ikj} = 1|M_{ikj} = 1), \]

\[ \theta_{ik}'' = P(I_{ikj} = 1|M_{ikj} = 0), \]

and in the case of a continuous outcome,

\[ \theta_{ik} = E(Y_{ikj}|M_{ikj} = 1 \cup M_{ikj} = 0), \]

\[ \theta_{ik}' = E(Y_{ikj}|M_{ikj} = 1), \]

\[ \theta_{ik}'' = E(Y_{ikj}|M_{ikj} = 0), \]

where \( I_{ikj} \) and \( M_{ikj} \) are dummy variables that indicate whether participant \( j \) experienced the event or completed arm \( k \) in trial \( i \), respectively, and \( Y_{ikj} \) indicates the continuous outcome measured in participant \( j \).

### 3.1.1 Informative missingness parameters

White et al.\(^{11}\) and Mavridis et al.\(^{12}\) proposed replacing the missingness parameter, \( \theta_{ik}' \), with the informative missingness odds ratio (IMOR) parameter in the logarithmic scale,

\[ q_{ik} = \logit(\theta_{ik}'') - \logit(\theta_{ik}'), \]

or the informative missingness difference of means (IMDoM) parameter,

\[ \psi_{ik} = \theta_{ik}'' - \theta_{ik}', \]

for a binary or continuous outcome, respectively.

These parameters measure the departure from the MAR assumption: non-zero departure implies informative (non-ignorable) missingness mechanisms, whereas zero departure implies MAR. Parameters \( q_{ik} \) and \( \psi_{ik} \) are unknown, and their values can be either suggested or estimated – the latter is possible when Bayesian methods are used. Under the Bayesian framework, it is straightforward to estimate these parameters by assigning a plausible prior distribution on them, which enables learning about the missingness mechanism. A normal prior distribution is a natural choice for \( q_{ik} \) and \( \psi_{ik} \). We may structure the prior distribution to be specific to the intervention, the trial, or the trial-arm, as well as to be fixed, exchangeable or independent across interventions, trials, or trial-arms.\(^{10,13}\) In the present article, we applied intervention-specific log IMORs and IMDoMs that are exchangeable across the trials,

\[ q_{ik}, \psi_{ik} \sim N(\Delta_{ik}, \sigma_{ik}^2), \]

\[ \Delta_{ik} \sim N(\xi, 1) \text{ and } \sigma_{ik} \sim \text{Unif}(0, 1), \]

for \( i = 1, 2, \ldots, N \) and \( k = 1, 2, \ldots, a_i \), where \( a_i \) is the number of arms in trial \( i \).

### 3.1.2 Clinically plausible scenarios for sensitivity analysis

In line with published recommendations, we considered the MAR assumption (i.e., \( \xi = 0 \)) to be the primary analysis for being a reasonably plausible assumption.\(^{14,15}\) We selected the following values of \( \xi \) for log IMOR to reflect a different degree of informative missingness:

\[-\log(3), -\log(2), \log(1), \log(2), \text{ and } \log(3),\]

where \( e^{-\log(\alpha)} \) indicates that the odds of an event is \( \alpha \) times more likely in completers than in missing participants, whereas \( e^{\log(\alpha)} \) indicates the opposite. For IMDoM, we considered the following values of \( \xi \):

\[-2, -1, 0, 1, \text{ and } 2,\]

where \(-X\) indicates that the outcome increases by \( X \) on average in completers as compared to missing participants, whereas \( X \) indicates the opposite. Therefore, we get a \( 5^2 \times 2 \) matrix with 25 unique pairs of values (hereinafter named ‘scenario matrix’) that indicate 25 different scenarios for the pairwise comparison (first column for the active intervention, second column for the control intervention):

| Scenario | log IMOR Active | log IMOR Control | IMDoM Active | IMDoM Control |
|----------|--------------|-----------------|-------------|--------------|
| 1        | -log(3)      | -log(3)         | -2          | -2           |
| 2        | -log(3)      | -log(2)         | -2          | -1           |
| ...      | ...          | ...             | ...         | ...          |
| 13 (MAR) | log(1)       | log(1)          | 0           | 0            |
| ...      | ...          | ...             | ...         | ...          |
| 24       | log(3)       | log(2)          | 2           | 1            |
| 25       | log(3)       | log(3)          | 2           | 2            |

While the scenarios mentioned above are directly applicable in a star-shaped network, as all active interventions...
are compared with a single control intervention, they are not immediately applicable in a non-star-shaped network, where an intervention may serve as the control in one trial and as the active treatment in another trial. Failure to ensure transitive scenarios for the missingness mechanisms across the observed comparisons of the non-star-shaped network may compromise the validity of the NMA results.\textsuperscript{16} A viable solution is to consider the same scenarios for all active interventions of the network.\textsuperscript{16} This solution requires that we select a reference intervention for the non-star-shaped network that receives different IMDoM/IMOR values from the rest of the interventions for scenarios that differ in the compared arms.\textsuperscript{16} In the supporting information, we illustrate the concept of transitivity in the missingness mechanisms in a triangle network. For instance, for a non-star-shaped network with four interventions (A, B, C, and D, where D is the reference intervention), the scenario matrix for log IMOR is the following:

| Scenario | A   | B   | C   | D (reference) |
|----------|-----|-----|-----|---------------|
| 1        | −log(3) | −log(3) | −log(3) | −log(3) |
| 2        | −log(3) | −log(3) | −log(3) | −log(2) |
| ...      | ...   | ...   | ...   | ...         |
| 13 (MAR) | log(1) | log(1) | log(1) | log(1) |
| ...      | ...   | ...   | ...   | ...         |
| 24       | log(3) | log(3) | log(3) | log(2) |
| 25       | log(3) | log(3) | log(3) | log(3) |

### 3.2 The Kullback–Leibler divergence measure

Before presenting the robustness index, we introduce the Kullback–Leibler divergence (KLD) measure, a commonly used measure to compare two probability distributions.\textsuperscript{17} In the present study, the KLD measure is a function of the posterior mean and posterior standard deviation of the summary treatment effect under the MAR assumption (i.e., $\xi = 0$) and each informative scenario (i.e., $\xi \neq 0$). The true missingness mechanism is not known, and we consider the MAR assumption to be a reasonably plausible assumption for the primary analysis following the recommendations of the relevant literature.\textsuperscript{14,15} Note that the prior and posterior distributions of a summary treatment effect are conjugate; in this case, both distributions follow the normal distribution.\textsuperscript{18} Therefore, the KLD measure from the informative scenario $i$ ($i \in S$, where $S = \{1, 2, ..., 25\} \setminus \{13\}$ with \{13\} being the MAR assumption) to the MAR assumption is calculated as follows:

$$D_{0,i} = D(N_0|N_i) = \frac{1}{2} \left\{ \left( \frac{s_0}{s_i} \right)^2 + \frac{(\mu_i - \mu_0)^2}{s_i^2} - 1 + 2 \ln \frac{s_i}{s_0} \right\},$$

where $N_i$ and $N_0$ refer to the normal posterior distribution under informative scenario $i$ and the MAR assumption, respectively, $\mu_i$ and $\mu_0$ are the corresponding posterior means, and $s_i$ and $s_0$ are the corresponding posterior standard deviations. The KLD values range from zero to infinity, with zero indicating a perfect match of the distributions. Intuitively, it follows that the lower the KLD value, the ‘closer’ the two distributions are, and thus, approximating the MAR assumption with an informative MOD scenario does not essentially change the information conveyed from the former. The KLD measure is a divergence and not a distance measure, which means that it does not share the symmetrical properties of a distance measure (see supporting information).

### 3.3 The robustness index

The ultimate goal is to combine the KLD values properly across all informative scenarios to infer the robustness of the primary analysis results. Therefore, we propose the robustness index (RI), which quantifies the overall ‘distance’ between the primary analysis results and the results from the alternative re-analyses, and is a function of the KLD measure. In the context of MOD, RI measures the overall distance of the KLD measures from zero for all $|S|$ informative scenarios (here, $|S| = 24$)

$$RI = \sqrt{\sum_{i \in S} D_{0,i}^2}. \quad (1)$$

Equation (1) is similar to the Euclidean distance; however, we are only interested in the overall distance from zero – not in the distance among all $|S|$ KLD values. We calculate RI only for the summary treatment effects due to the conjugate ‘nature’ of their prior and posterior distributions. Note that we can calculate as many RI(s) as the number of possible comparisons in the network. For a network of $T = 4$ interventions, we have six possible comparisons (i.e., $T(T - 1)/2$), and hence, we can calculate a total of six RI. For a PMA, we obtain only one RI.

A ‘typical’ sensitivity analysis uses only a fraction of the information concerning the estimated summary treatment. This usually involves the evaluation of the point estimate and the credible interval (CrI) of the summary treatment effect to infer whether its magnitude, direction, and conclusion differ in different re-analyses of the same outcome. Contrariwise, the RI, being a function of the
KLD measure, retains the richness of information as it considers the whole posterior distribution of the estimated summary treatment effect under the alternative re-analyses.

### 3.4 Setting the threshold of robustness

To our knowledge, there are no universally accepted directions on what constitutes similar results when several re-analyses are undertaken and compared to the primary analysis for the summary treatment effects. In this work, we introduce an intuitive rule, quantified by the RI, to classify the overall ‘distance’ between the primary and alternative re-analysis results as ‘low’ and ‘substantial’ translating into presence and lack of robustness, respectively. Specifically, a ‘low’ overall distance is defined as a distance from zero that is less than one between-trial standard deviation ($\tau$) of low statistical heterogeneity: $RI < \tau_{\text{low}}$. As low statistical heterogeneity, we considered the median of the empirically-based distribution for $\tau^2$ in the case of a general health-care setting. In the SMD and log OR scale, this median is equal to 0.03 in the standardised mean difference (SMD) scale and 0.08 in the log odds ratio (OR) scale. Then, the corresponding threshold of robustness is equal to $\sqrt{0.03} = 0.17$ in the SMD scale and $\sqrt{0.08} = 0.28$ in the log OR scale (i.e., 1.32 after exponentiation). Therefore, $RI < 0.17$ or $RI < 0.28$ indicates ‘robustness’, and $RI \geq 0.17$ or $RI \geq 0.28$ implies ‘lack of robustness’ in the SMD and log OR scale, respectively.

Different prior distributions for $\tau^2$ will result in different thresholds for robustness to some extent: the larger the median of the distribution (as a threshold of ‘low’ statistical heterogeneity), the larger the threshold of robustness. For instance, in the case of the half-normal prior distribution for $\tau$ with variance 0.5 or 1, the median is equal to 0.34 and 0.67, respectively, and thus, the corresponding $D_T$ values are 0.34 and 0.67. Assuming a binary outcome, $D_T$ of 0.34 or 0.67 on the log OR scale is equivalent to 1.40 or 1.95 on the OR scale. Particularly, the latter value indicates considerable information loss and should not be used to infer the robustness of the primary analysis. In other words, the selection of the prior distribution for $\tau^2$ to perform Bayesian NMA should not dictate the threshold of robustness. We provide some discussion on the selection of robustness thresholds later on.

In the NMA setting, it is more challenging to infer the overall robustness of the primary analysis results since RI may differ across comparisons: for instance, robustness may be inferred for some comparisons, but not for others. To determine the presence or lack of robustness in the whole network, we recommend the following straightforward decision framework:

1. presence of robustness, when RI is less than the selected threshold of robustness for all possible comparisons in the network;
2. lack of robustness, when RI is equal to or exceeds the selected threshold of robustness for at least one comparison in the network.

Using the heatmap, we can illustrate RI for all possible comparisons in the network and use green and red colours to indicate the presence and lack of robustness, respectively.

### 3.5 Graphical tool: Enhanced balloon plot

To illustrate the summary treatment effects of the compared interventions for the scenarios presented in Section 3.1.2, we propose the ‘enhanced balloon plot’. The following description refers to a PMA, and hence, the terms ‘active arm’ and ‘control arm’. For NMA, the corresponding terms are ‘non-reference intervention’ and ‘reference intervention’. The proposed plot has the following characteristics:

1. For the pairwise comparison, the plot’s x-axis and y-axis refer to the same values of IMDoM/IMOR for the active and control arm, respectively. Therefore, each pair of coordinates (i, j) indicates a specific scenario from the scenario matrix. Figure 1 illustrates the main structure of the balloon plot with scenarios for the IMDoM parameter (Figure 1a) and IMOR parameter (Figure 1b). The four corners reflect the most extreme MOD scenarios. Scenarios above the main diagonal (grey dashed line) indicate a larger IMDoM or IMOR value for the control arm than the active arm, and scenarios below the main diagonal indicate the opposite. Scenarios on the main diagonal refer to the same IMDoM or IMOR value in both arms. The points (0, 0) and (1, 1) correspond to the MAR assumption for IMDoM and IMOR, respectively. Scenarios closer to MAR (the white area) are considered optimistic, and scenarios further from MAR (the light grey area) are sceptical. Scenarios outside the range (the grey area) are considered implausible.
2. For each (i, j), the bubble (point) indicates the estimated effect size. The colour of each bubble reflects the uncertainty around the estimate. The size of each bubble is proportional to the magnitude of the effect: the larger the treatment effect, the larger the bubble.
3. We use ‘crossed bubble’ to indicate scenarios associated with conclusive evidence (i.e., zero is not included in the 95% credible interval of SMD or log OR) and a ‘filled bubble’ for scenarios with inconclusive evidence.
The enhanced balloon plot can be constructed for any parameter of the PMA or NMA model similar to the summary treatment effects. While the focus is on the summary treatment effects, we additionally present the enhanced balloon plot for the between-trial heterogeneity parameter, $\tau^2$. In the case of $\tau^2$, we use ‘crossed bubble’ to indicate scenarios that are associated with considerable statistical heterogeneity (i.e., posterior median of $\tau^2 \geq$ median of the selected empirically-based prior distribution for $\tau^2$) and a ‘filled bubble’ for scenarios with low statistical heterogeneity.

In Section 4, we illustrate how the proposed RI and the enhanced balloon plot can be used jointly in the proposed sensitivity framework for a PMA and a network of interventions. Since RI signals whether a comparison is associated (or not) with robust primary results, the enhanced balloon plot can demystify those scenarios with substantially deviant results from MAR that may contribute to the lack of robustness in that comparison.

### 3.6 Model implementation

For each scenario, we performed one-stage Bayesian random-effects PMA for the Cochrane review, and one-stage Bayesian random-effects NMA with consistency equations for the published systematic review while incorporating the pattern-mixture model. We considered the log OR and the SMD as the effect measures which are intuitively related to the log IMOR and IMDoM parameters. For the NMA, we considered the scenario matrix in Section 3.1.2 for non-star-shaped networks with placebo as the reference intervention. For the location parameters, we assigned normal prior distribution with mean 0 and variance 10,000. We assigned empirically-based prior distributions on $\tau^2$: for the binary outcome, log-normal prior distribution with mean $-2.06$ and variance $1.51^2$ (median: 0.13, 95% range: 0.007–2.10), and for the continuous outcome, log-$t$ prior distribution with mean $-2.99$, variance $2.16^2$, and 5 degrees of freedom (median: 0.049, 95% range: 0.001–4.70). These prior distributions refer to the ‘symptoms reflecting the continuation of condition’ and the ‘mental health indicators’ outcome-types,
respectively, for the ‘pharmacological versus placebo’ comparison-type.19,20 We ran three chains of different initial values with 100,000 iterations; 10,000 burn-in; and thinning equal to 5 and 10 for the PMA and NMA, respectively. We assessed convergence using the Gelman–Rubin convergence diagnostic and visual inspection of trace-plots and autocorrelation plots.23 We used JAGS24 via the R-package R2jags25 (statistical software R, version 3.6.126) to implement the models, and the R-package ggplot227 to draw the enhanced balloon plot and the heatmap. All functions related to this manuscript are publicly available at https://github.com/LoukiaSpin/Quantifying-Robustness-in-Meta-analysis.git.

4 | RESULTS

4.1 | Pairwise meta-analysis: Continuous outcome

The summary treatment effects (SMDs) from the primary and sensitivity analyses are shown in the enhanced balloon plot (Figure 2a). Under the MAR assumption, the posterior mean of SMD was equal to −0.10 (95% CrI: −0.60 to 0.39) in favour of inositol, but the CrI included zero (inconclusive evidence). Under the best-case scenario for placebo (bottom right), SMD increased to zero, concluding both interventions equally effective. Contrariwise, in all other scenarios, inositol prevailed placebo (SMD range: −0.19 to −0.02).

Scenarios that assumed a larger IMDoM in the placebo than inositol (scenarios above the main diagonal that favour inositol) led to a larger posterior mean of SMD (SMD range: −0.19 to −0.11) compared to the primary analysis. On the contrary, scenarios that assumed a larger IMDoM in inositol than placebo (scenarios below the main diagonal that favour placebo) led to the same or a smaller posterior mean of SMD (SMD range: −0.08 to 0.00) as in the primary analysis. As expected, since the outcome was harmful, a larger IMDoM in the experimental than control arm corresponded to a relatively worse outcome in the former as compared to the latter. Therefore, SMD was pulled towards zero in these scenarios.

FIGURE 2 The enhanced balloon plot on the summary standardised mean difference (SMD) under different scenarios of the informative missingness difference of means (IMDoM) parameter for placebo (y-axis) and inositol (x-axis) (plot A). Bubbles refer to the posterior mean of SMD with their size being proportional to the magnitude of SMD after rescaling into [0, 1]. The framed number refers to the primary analysis. Negative SMDs favour inositol, and positive SMDs favour placebo. Darker tones of blue refer to smaller values of the posterior standard deviation of SMD, and darker tones of red refer to larger values of the posterior standard deviation of SMD. Crossed bubbles indicate scenarios associated with conclusive evidence (i.e., zero is not included in the 95% credible interval of SMD) and filled bubbles indicate scenarios with inconclusive evidence. Barplot on the Kullback–Leibler divergence measure as calculated for extreme, sceptical and optimistic scenarios about IMDoM (plot B). Red bars refer to scenarios of low proximity in the compared arms, orange bars refer to scenarios of large proximity in the compared arms, and green bars refer to the same scenarios in the compared arms. IMDoM, informative missingness difference of means [Colour figure can be viewed at wileyonlinelibrary.com]
Contrariwise, a smaller IMDoM in the experimental than control arm corresponded to a better outcome in the former than in the latter, and hence, SMD was pulled towards larger negative values.

Scenarios that assigned the same IMDoM in both interventions (main diagonal) yielded SMDs closer to the primary analysis, as compared to scenarios with IMDoM of opposite sign in the compared interventions or informative scenario in one arm but MAR in the other arm. However, the evidence was inconclusive in all scenarios (i.e., zero was included in the 95% CrI). Overall the uncertainty around SMD was very similar across the different scenarios (posterior standard deviation range: 0.25–0.26).

Different scenarios about IMDoM in the compared interventions had a negligible impact on the posterior median of $\tau^2$ (range: 0.029–0.035) (Figure S2, supporting information). Most scenarios yielded a relatively similar posterior standard deviation of $\tau^2$ to the primary analysis (MAR: 0.35), apart from the ‘best-case for inositol’ and the ‘better outcome among MOD in both arms’ that yielded a higher posterior standard deviation equal to 0.53 and 0.51, respectively. The posterior median of $\tau^2$ was lower than the median of the empirically-based prior distributions, and hence, reflected low statistical heterogeneity across all scenarios. However, the posterior standard deviation of $\tau^2$ indicated some uncertainty in estimating the parameter as the 95% CrI ranged from very low (around 0.0001) to considerable values of $\tau^2$ (around 0.60) across all scenarios. The small number of trials may have contributed to the wide overall 95% CrI of $\tau^2$.

Figure 2b illustrates the distribution of the KLD measures across extreme, sceptical and optimistic scenarios for IMDoM. Despite the moderate and substantial risk of attrition in the included trials, the KLD measure ranged from 0.00 to 0.07, indicating low ‘distance’ overall from every informative scenario to the primary analysis. The RI confirmed the robustness of the primary analysis results when compared to the different scenarios of IMDoM for inositol versus placebo ($RI = 0.02$). Robustness may be an implication of adjusting the within-trial results by modelling MOD under different scenarios for the missingness mechanisms.

### 4.2 Network meta-analysis: Binary outcome

We have obtained the estimated treatment effects of 28 possible pairwise comparisons in total (i.e., $\binom{7}{2}$ with $T = 8$) for every scenario of the scenario matrix. In this case, it is not advisable to create 28 balloon plots and inspect them separately as it would be challenging to decide on the robustness of the primary analysis results in every comparison. Thus, we propose to start with the assessment of the heatmap on the RI of all possible comparisons and then create the balloon plots only for those comparisons whose results were contentious compared with the primary analysis results.

#### 4.2.1 Heatmap on the robustness index for all possible comparisons in the network

According to the heatmap (Figure 3), the results of the primary analysis were not robust to different IMOR scenarios in all comparisons with placebo (range of RI: 0.39–0.57). For illustration, we focused on the four comparisons with placebo that yielded the largest RI: fluticasone, fluticasone plus salmeterol, formoterol, and tiotropium versus placebo. The KLD measure approached or exceeded the threshold of robustness (i.e., $RI = 0.28$) for the worst-case scenarios (under extreme scenarios) for these comparisons (Figure 4). Overall, more distant scenarios yielded larger values of the KLD measure compared with less or no distant scenarios (Figure 4). For each comparison, among the more distant scenarios, the extreme ones resulted in larger values of the KLD measure, followed by the sceptical and optimistic ones.

The comparisons without placebo demonstrated the robustness of the primary analysis results (range of RI: 0.03–0.20, Figure 3). A plausible explanation may be that, after adjusting for MOD, any residual bias in the log OR of comparisons with placebo (the reference intervention) may be cancelled out to some extent in the log OR of the comparisons without the reference intervention due to the consistency equation.²⁸ Using the decision framework of robustness in the whole network (Section 3.4), we concluded that the credibility of the primary analysis should be called into question for the analysed network.

#### 4.2.2 Enhanced balloon plots of comparisons with lack of robustness

Figure 5 presents a series of enhanced balloon plots for the four aforementioned comparisons with placebo that failed to demonstrate the robustness of the primary analysis results. In line with the results from the PMA (Section 4.1), assigning the same IMOR in the compared interventions (main diagonal) yielded results closer to the primary analysis, followed by scenarios with the same direction but different IMOR value, as compared to scenarios with opposite direction. In line with Figure 4, the ORs obtained under the two extreme scenarios (the corners in the top
left and in the bottom right of the balloon plot) deviated from the primary analysis the most in all comparisons.

In fluticasone versus placebo (Figure 5a), all scenarios indicated fluticasone as more efficacious than placebo, while different informative scenarios affected the magnitude of OR substantially. Assuming a larger IMOR in the placebo than fluticasone led to larger OR favouring fluticasone (range: 0.63–0.75) compared to the other way around (range: 0.78–0.94), yet this evidence was inconclusive in all scenarios. Overall, there was some variability in the uncertainty around the log OR across the scenarios: according to the colour scale, assigning IMOR <1 in both arms yielded slightly more precise results than assigning IMOR >1 in both arms. The same conclusions were drawn for tiotropium versus placebo; however, the evidence was conclusive favouring tiotropium across all scenarios (Figure 5b).

In Figure 5c, scenarios with a larger IMOR in the placebo than formoterol yielded larger OR favouring the latter (range: 0.75–0.88). On the contrary, scenarios with a larger IMOR in formoterol than in placebo yielded OR closer to or slightly larger than one (range: 0.93–1.12). The evidence was inconclusive in all scenarios. Overall the uncertainty around log OR was similar across the scenarios (posterior standard deviation range: 0.26–0.29).

For the combination intervention versus placebo (Figure 5d), conclusions were similar to those for fluticasone and tiotropium versus placebo. The four most extreme of the scenarios that assigned larger IMOR in the placebo than the combination intervention yielded conclusive evidence in favour of the latter.
Contrary to the results from the PMA (Section 4.1), different scenarios about IMOR in the compared interventions had a negligible impact both on the posterior median and posterior standard deviation of $\tau^2$ (Figure S3, supporting information). There was low statistical heterogeneity in all scenarios according to the posterior median of $\tau^2$ (range: 0.02–0.03) and low uncertainty in estimating the parameter according to the posterior standard deviation (range: 0.027–0.036).

Under a harmful outcome, a larger IMOR in the placebo than the active interventions yielded a smaller OR favouring the latter. Furthermore, it changed the conclusions for the combination intervention versus placebo. A possible explanation may be that these scenarios led to a relative increase in the odds of exacerbations in the placebo compared to the active interventions. On the contrary, a larger IMOR in the active interventions than placebo yielded a larger OR favouring the latter and changed its direction in formoterol versus placebo.
Furthermore, IMOR < 1 in both arms corresponds to logit(θ_m) < logit(θ_0), and hence, to a smaller variance in logit(θ) as compared to logit(θ_m). In combination with q < 0.5 observed in all trials of the network, θ_m had more weight in the estimation of the arm-specific risk θ and variance thereof as compared to θ_m (via the pattern-mixture model). Consequently, these scenarios yielded more precise log OR compared to scenarios with IMOR > 1 in both arms. The precision in the estimation of log OR in the remaining scenarios was somewhere in-between.

5 | DISCUSSION

We provide a decision framework to infer robustness of the primary analysis results in PMA and NMA by
conducting appropriate sensitivity analyses. This work addresses two key limitations in the current application of sensitivity analysis in systematic reviews. First, we promote a comprehensive sensitivity analysis process founded on plausible scenarios with respect to the investigated characteristic (here, MOD) and proper statistical analysis for all scenarios in one stage. Second, we offer a straightforward index of robustness and an intuitive decision framework to conclude on the robustness of the primary analysis results objectively. The RI is particularly useful in NMA for an efficient evaluation of the robustness of the primary analysis results in the whole network.

The decision to illustrate the proposed framework in the context of MOD was driven by the inherent uncertainty about the missingness mechanisms, which renders the assessment of the primary analysis results in a sensitivity analysis imperative. However, the RI is relevant to a broader set of sensitivity analyses, for instance, when we investigate the robustness of the summary treatment effects to different prior distributions for \( \tau^2 \), or to different assumptions for the probability of publication bias in NMA.\(^{30}\) The researchers have the liberty to adjust the threshold of robustness to the aims of the systematic review. For instance, if the aim is to develop health policy guidelines (e.g., through HTAs), a more stringent robustness threshold may be preferred. If the systematic review is fundamentally exploratory (e.g., to appraise the quality of the available relevant evidence and elicit knowledge gaps), the researchers may adopt our thresholds of robustness. Moreover, we have proposed a threshold of robustness that is specific to the effect measure. Since different effect measures are associated with a different extent of statistical heterogeneity,\(^{31}\) we do not recommend using a ‘common’ threshold of robustness for all effect measures. The importance of sensitivity analysis and the relevance of our proposed framework is not restricted to the health field, as systematic reviews are also popular in social sciences (e.g., Campbell reviews) and they are subject to the same considerations for transparent and rigorous analysis.

Furthermore, the size and number of trials, as well as the frequency of the outcome, and the structure of the network can affect the posterior distribution of the summary treatment effects, and by extent, the KLD measure and the index RI. For instance, in a sparse network (i.e., a network with few connections between interventions informed by few trials) or a meta-analysis with few small trials on a rare outcome, the posterior distribution of the summary treatment effects is sensible to the prior distribution for \( \tau^2 \) due to the limited available evidence.\(^{32}\) In this case, we advise the researchers to prefer stringent thresholds of robustness to avoid spurious conclusions on the robustness of the primary analysis results. In more dense networks, the researchers can use our proposed thresholds.

We have not incorporated any information on the clinical relevance of the accepted deviation that defined our proposed thresholds of robustness and that partly comprises a limitation of our study. However, by using an empirically-based distribution for \( \tau^2 \), we have attempted to reduce the subjectivity in determining the threshold of robustness. Ideally, a clinically accepted difference in the treatment effects between the primary and sensitivity analyses should be explicitly defined in the systematic review protocol, and it should be driven by the aims of the systematic review and the expert opinion on the interventions and condition under investigation. To mitigate the subjectivity in the expert opinion, we advise the researchers to consult several experts to elicit the clinically relevant accepted deviation in treatment effects and calculate the weighted average of the suggested thresholds using the years of experience as weights.\(^{33}\) The researchers are encouraged to use our proposed threshold to investigate the sensitivity of the clinically relevant threshold to a more or less conservative threshold (if the elicited threshold is higher or lower than the proposed one) to preclude possible spurious robustness.

Our proposed sensitivity analysis framework is highly essential for rating the certainty in the evidence concerning the risk of attrition bias via the GRADE approach. The current GRADE approach lacks the objectivity in demonstrating robustness: ‘to the extent that pooled estimates remain similar when making progressively more stringent assumptions (and in particular, results remain statistically significant), one would conclude that the results are robust to the missing data [...]’\(^{34}\) To infer the similarity between two or more summary treatment effects, we need to quantify similarity, and then employ an objective decision framework to support our conclusions. With our proposed RI, we have achieved to offer a framework to infer robustness objectively. Rather than relying on the location and statistical significance alone, as the GRADE approach currently encourages, we consider the whole posterior distribution of the parameter in the calculation of the index.

The present study lacks a clinician’s advice on plausible missingness scenarios tailored to the motivating examples. In addition, we have considered five scenarios for the IMOR and IMDOM values with a relatively narrow range, but more options could be explored. Researchers are encouraged to consider more scenarios or scenarios in a broader range, if plausible. However, the convergence time will be inevitably longer, especially in large networks of interventions. We do not recommend defining less than five scenarios for the missingness parameter as the sensitivity analysis may be inadequate.
However, not a limitation per se, we have used Bayesian methods for offering flexibility in the analysis of aggregate MOD and for being popular in NMA. Our proposed RI can be applied straightforwardly in the frequentist framework since the estimated summary treatment effect and uncertainty thereof are the necessary components to calculate the KLD measure and obtain the RI. Lastly, the proposed RI is not immediately applicable to the assessment of possible inconsistency, as the threshold of robustness has been exclusively defined on summary treatment effects. We need to define what constitutes considerable inconsistency (measured as the difference between the direct and indirect effect) to infer whether different missingness scenarios compromise the consistency assumption.

6 | CONCLUSIONS

Consistent with previous recommendations for proper handling of aggregate MOD, we propose and strongly advise the following sensitivity analysis plan. As a first step, researchers should seek a clinician’s opinion for: (a) the MOD scenarios that are plausible for the interventions and conditions under investigation, and (b) the threshold of robustness that is proper for the aims of the systematic review. Ideally, this consultation should occur during the protocol development of the systematic review. The second step entails appropriate synthesis of trials with MOD to obtain the posterior distribution of the summary treatment effects for all possible scenarios in one stage and then, calculate the RI. The final steps of the proposed process are to compare the RI with the selected threshold to conclude the robustness of the primary analysis results and present the enhanced balloon plots for the summary treatment effect of comparisons where lack of robustness was found (if any). Then, the enhanced balloon plots can help detect the missingness scenario that led to deviant results from the MAR assumption regarding the magnitude, uncertainty, and conclusiveness of the estimated parameter (summary treatment effect or between-trial variance).

ACKNOWLEDGEMENTS

Loukia M. Spineli is supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) (grant SP 1664/1-3). Chrysostomos Kalyvas is employed by Merck Sharp & Dohme. Katerina Papadimitropoulou is a Ph.D. candidate at the Department of Clinical Epidemiology of Leiden University Medical Center. Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

LMS conceived the idea of the study. All authors designed the study. LMS performed the analysis. CK and KP checked the code for correctness. LMS drafted the article. All authors revised the article critically for important intellectual content and approved the final version of the article.

ORCID

Loukia M. Spineli https://orcid.org/0000-0001-9515-582X
Chrysostomos Kalyvas https://orcid.org/0000-0003-0606-4518
Katerina Papadimitropoulou https://orcid.org/0000-0002-5732-4044

REFERENCES

1. Patsopoulos NA, Evangelou E, Ioannidis JP. Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. Int J Epidemiol. 2008;37(5):1148-1157.
2. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for Meta-Analysis in Medical Research. Chichester, England: Wiley; 2000.
3. Little RJA, Rubin D. Statistical Analysis with Missing Data. 2nd ed. John Wiley & Sons: Hoboken, NJ; 2002.
4. National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press; 2010. Retrieved from www.nap.edu
5. Spineli LM, Pandis N, Salanti G. Reporting and handling missing outcome data in mental health: a systematic review of Cochrane systematic reviews and meta-analyses. Res Synth Methods. 2015;6(2):175-187.
6. Spineli LM, Yepes-Nuñez JJ, Schünemann HJ. A systematic survey shows that reporting and handling of missing outcome data in networks of interventions is poor. BMC Med Res Methodol. 2018;18(1):115.
7. Kahale LA, Diab B, Brignardello-Petersen R, et al. Systematic reviews do not adequately report or address missing outcome data in their analyses: a methodological survey. J Clin Epidemiol. 2018;99:14-23.
8. Taylor MJ, Wilder H, Bhagwagar Z, Geddes J. Inositol for depressive disorders. Cochrane Database Syst Rev. 2004;2:CD004049. https://doi.org/10.1002/14651858.CD004049.pub2.
9. Baker WL, Baker EL, Coleman CI. Pharmacologic treatments for chronic obstructive pulmonary disease: a mixed-treatment comparison meta-analysis. Pharmacotherapy. 2009;29(8):891-905.
10. Spineli LM. An empirical comparison of Bayesian modelling strategies for missing binary outcome data in network meta-analysis. BMC Med Res Methodol. 2019;19(1):86.
11. White IR, Higgins JP, Wood AM. Allowing for uncertainty due to missing data in meta-analysis—part 1: two-stage methods. Stat Med. 2008;27(5):711-727.
12. Mavridis D, White IR, Higgins JP, Cipriani A, Salanti G. Allowing for uncertainty due to missing continuous outcome data in pairwise and network meta-analysis. Stat Med. 2015;34(5):721-741.
13. Turner NL, Dias S, Ades AE, Welton NJ. A Bayesian framework to account for uncertainty due to missing binary outcome data in pairwise meta-analysis. *Stat Med*. 2015;34(12):2062-2080.

14. Higgins JP, White IR, Wood AM. Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clin Trials*. 2008;5(3):225-239.

15. White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clin Trials*. 2012;9(4):396-407.

16. Spineli LM. Modeling missing binary outcome data while preserving transitivity assumption yielded more credible network meta-analysis results. *J Clin Epidemiol*. 2019;105:19-26.

17. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat*. 1951;22(1):79-86.

18. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making*. 2013;33(5):607-617.

19. Rhodes KM, Turner RM, Higgins JP. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol*. 2015;68(1):52-60.

20. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015;34(6):984-998.

21. Spiegelhalter DJ, Abrams KR, Myles J. Hierarchical priors. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, England: John Wiley & Sons; 2004:167-174.

22. Spineli LM, Kalyvas C, Papadimitropoulou K. Continuous(ly) missing outcome data in network meta-analysis: a one-stage pattern-mixture model approach. *Stat Methods Med Res*. 2021;962280220983544. https://doi.org/10.1177/0962280220983544.

23. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457-472.

24. Plummer M. JAGS: Just Another Gibbs Sampler, version 4.3.0 user manual. 2017. pp. 1–74.

25. Su YS, Yajima M. R2jags: Using R to Run ‘JAGS’. R package version 0.6-1. 2020. Retrieved from https://CRAN.R-project.org/package=R2jags

26. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020. Retrieved from https://www.r-project.org

27. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag; 2016.

28. Spineli LM, Papadimitropoulou K, Kalyvas C. Pattern-mixture model in network meta-analysis of binary missing outcome data: one-stage or two-stage approach? *BMC Med Res Methodol*. 2021;21(1):12.

29. Turner RM, Domínguez-Islas CP, Jackson D, Rhodes KM, White IR. Incorporating external evidence on between-trial heterogeneity in network meta-analysis. *Stat Med*. 2019;38(8):1321-1335.

30. Mavridis D, Welton NJ, Sutton A, Salanti G. A selection model for accounting for publication bias in a full network meta-analysis. *Stat Med*. 2014;33(30):5399-5412.

31. Rhodes KM, Turner RM, Higgins JP. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. *Res Synth Methods*. 2016;7(4):346-370.

32. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. *Res Synth Methods*. 2017;8(1):79-91.

33. Chaimani A, Porcher R, Sbidian É, Mavridis D. A Markov chain approach for ranking treatments in network meta-analysis. *Stat Med*. 2021;40(2):451-464.

34. Guyatt GH, Ebrahim S, Alonso-Coello P, et al. GRADE guidelines 17: assessing the risk of bias associated with missing participant outcome data in a body of evidence. *J Clin Epidemiol*. 2017;87:14-22.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**Appendix S1: Supporting Information**

**How to cite this article:** Spineli LM, Kalyvas C, Papadimitropoulou K. Quantifying the robustness of primary analysis results: A case study on missing outcome data in pairwise and network meta-analysis. *Res Syn Meth*. 2021;12:475–490. https://doi.org/10.1002/jrsm.1478