American Journal of Therapeutics 30, e56–e71 (2023)

OPEN

Residual Disease After Primary Surgical Treatment for Advanced Epithelial Ovarian Cancer, Part 2: Network Meta-analysis Incorporating Expert Elicitation to Adjust for Publication Bias

Andrew Bryant,[1]* Michael Grayling,[1] Ahmed Elattar,[2] Ketankumar Gajjar,[3] Dawn Craig,[1] Luke Vale,[1] and Raj Naik[4]

Background: Previous work has identified a strong association between the achievements of macroscopic cytoreduction and improved overall survival (OS) after primary surgical treatment of advanced epithelial ovarian cancer. Despite the use of contemporary methodology, resulting in the most comprehensive currently available evidence to date in this area, opponents remain skeptical.

Areas of Uncertainty: We aimed to conduct sensitivity analyses to adjust for potential publication bias, to confirm or refute existing conclusions and recommendations, leveraging elicitation to incorporate expert opinion. We recommend our approach as an exemplar that should be adopted in other areas of research.

Data Sources: We conducted random-effects network meta-analyses in frequentist and Bayesian (using Markov Chain Montel Carlo simulation) frameworks comparing OS across residual disease thresholds in women with advanced epithelial ovarian cancer after primary cytoreductive surgery. Elicitation methods among experts in gynecology were used to derive priors for an extension to a previously reported Copas selection model and a novel approach using effect estimates calculated from the elicitation exercise, to attempt to adjust for publication bias and increase confidence in the certainty of the evidence.

Therapeutic Advances: Analyses using data from 25 studies (n = 20,927 women) all showed the prognostic importance of complete cytoreduction (0 cm) in both frameworks. Experts accepted publication bias was likely, but after adjustment for their opinions, published results overpowered the informative priors incorporated into the Bayesian sensitivity analyses. Effect estimates were attenuated but conclusions were robust in all analyses.

Conclusions: There remains a strong association between the achievement of complete cytoreduction and improved OS even after adjustment for publication bias using strong informative priors formed from an expert elicitation exercise. The concepts of the elicitation survey should be strongly considered for utilization in other meta-analyses.

Keywords: advanced epithelial ovarian cancer, residual disease, expert elicitation, publication bias, Bayesian network meta-analysis

[1]*Population Health Sciences Institute, Newcastle University, Newcastle Upon Tyne, United Kingdom;* [2]*Pan-Birmingham Gynaecological Oncology Cancer Centre, Birmingham, United Kingdom;* [3]*Nottingham City Hospital, Obstetrics and Gynaecology, Nottingham, United Kingdom; and* [4]*Northern Gynaecological Oncology Centre, Gateshead, United Kingdom.*

The authors have no conflicts of interest to declare.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.americantherapeutics.com).

A. Elattar, K. Gajjar, and R. Naik provided clinical expertise and contributed to the discussion sections of the paper; A. Bryant conceptualized the elicitation exercise and adjustment applied in Part B of the exercise and drafted the methodological, results and discussion sections of the paper. M. Grayling provided statistical expertise and applied and critically reviewed all aspects of methodology in the paper. D. Craig and L. Vale provided expert guidance and critical review of the paper. All authors agreed the final version.

**Address for correspondence: Population Health Sciences Institute, Newcastle University, 4th Floor, Idley Building 1, Queen Victoria Road, Newcastle upon Tyne NE1 7RU. E-mail: andy.bryant@ncl.ac.uk*

INTRODUCTION

Ovarian cancers remain a major concern to women worldwide.[1,2] In advanced disease, surgery and platinum-based chemotherapy are the standard treatment options. Traditionally, this included upfront primary debulking surgery (PDS) which is performed to remove as much visible disease as possible. This is because the amount of residual tumor is one of the most important prognostic factors for survival of epithelial ovarian cancer (EOC).[1] Chemotherapy followed by interval debulking surgery is an alternative primary treatment option for women diagnosed with advanced ovarian cancer, and evidence in this area is emerging. We focus our research on PDS because there is an established evidence base in which to apply our suggested methodology. A more extensive description of the aims of primary surgery in achieving "optimal cytoreduction" has been described in previous publications.[1,2] Evidence suggests that where there is "complete cytoreduction" (surgery that completely removes all visible tumour), survival is significantly improved compared with less-than-complete cytoreduction.[3–5] However, publication bias[6] leaves room for uncertainty as to the true value of complete cytoreduction, and opponents to the approach have raised concerns regarding the strength of the evidence base.

The Gynecological Cancer InterGroup defined "optimal" cytoreduction as having no macroscopic residual disease which is often reported in the literature as RD0 (residual disease (RD) = 0 cm), near-optimal RD (<1 cm), and suboptimal RD (>1 cm).[7] Although there is now less controversy about the prognostic importance of maximum cytoreduction, there remains divided opinion about the effects of any remaining RD after PDS and about what attempts should be made for maximal efforts at debulking. Different philosophies are evident within the surgical community, but there are also other important considerations, such as surgical skills, training, the woman's fitness for more radical treatment, morbidity, mortality, and quality of life. These are all considerations when assessing publication bias and the reliability of the effect estimates in published studies. There is also the issue about unreported studies that show "negative" results, which in this context may be a study showing no benefit of complete cytoreduction.

Indeed, publication bias is a well-known threat to the validity of meta-analyses.[6,8] Negative or statistically insignificant findings typically have less chance of being published; therefore, available studies tend to be a biased sample. This leads to an inflation of effect size estimates of unknown degree.[9] Consequently, it has been argued that attempting to correct for bias is typically better than incorporating no correction at all because publication bias is inevitable in most meta-analyses. This includes when no publication biases are detected, as available tests to ascertain the presence of publication bias typically have low power.[10] Ultimately, using adequate methods of bias correction can add confidence to the certainty of effect estimates in a meta-analysis.

Accordingly, this research had two main aims. First, to compare the results of a Bayesian network meta-analysis (NMA) using a noninformative prior[11] with ones attempting to adjust for selective reporting of outcomes and publication bias[6,12] by using expert elicitation methodology.[13–15] Elicitation was conducted using expert members of the British Gynaecological Cancer Society (BGCS). The adjustment for publication bias is a key component to this research because many skeptics refute conclusions in this area despite sound methodology being applied previously.[1,2,16–19] The use of novel NMA methodology in this area has been previously deployed,[20,21] but it is important to disseminate findings to the wider surgical community and not just proponents of aggressive surgery. This is only achievable by reporting effect estimates that are more likely to be closer to the true effects by removing a degree of bias. Secondly, and of paramount importance, is to encourage the use of this methodology in other areas of research, particularly where the magnitude of effects are disputed, affecting the certainty of the evidence. We promote the use of our methodology throughout the article and encourage others to attempt to implement the methods in their own research.

Specifically, to assess the potential effects of publication bias, we implement a modified version of the selection model described by Mavridis et al[22] (see also Chootrakool et al[23] and Mavridis et al).[24] This approach extends the popular Copas selection model for a conventional two-group meta-analysis[25–27] to the general NMA setting. It is, particularly, dependent on the specification of probabilities for the chance that "small" and "large" studies would be published, which we nominate using the results of an expert elicitation exercise. Although a small number of studies have previously performed this type of analysis in a NMA, they have specified these parameters somewhat arbitrarily (e.g., to reflect perceived levels of "low" and "high" publication bias). We are unaware of any previous work that has elicited these key parameters from experts.

We also use an alternative approach to adjusting for publication bias in a NMA, which to the best of our knowledge has not been considered previously, which leverages informative priors in an otherwise

conventional Bayesian NMA. In our case, the informative priors are formed based on the opinion of expert members of the BGCS.[28,29] We believe this approach would be easy to mimic for all oncology settings that use survival outcomes where an estimate of the control arm event rate can be reliably estimated.

METHODS

Search strategy and selection criteria

The NMAs reported in this article synthesized studies according to good research principles following the methods outlined by Bryant et al.[2] Bibliographic databases were searched from 1950 up to September 2021 (results of search are shown in Figure 1). We applied the same search and inclusion criteria as outlined by Bryant et al.[2] The population of interest was women who had received primary cytoreductive surgery followed by adjuvant platinum-based chemotherapy.[1] Included studies reported overall survival (OS) for comparisons of RD thresholds after surgery and used the same statistical adjustment constraints by Bryant et al[2] to minimize selection bias.[20,30] We sifted references identified from the search, extracted data on pertinent items, and assessed risk of bias in accordance with the Cochrane guidelines,[20] following on from the systematic review that underpins this analysis and the subsequent frequentist NMA.[1,2]

Expert elicitation exercise and statistical considerations

An expert elicitation exercise[31] was sent to members of the BGCS by the organizing committee. The elicitation exercise was conducted before the completion of the systematic review,[1] and the findings from this exercise used to adjust the meta-analyses for perceived publication bias. In the elicitation exercise, we asked participants to account for the sort of studies that have been conducted but not published, the plausible magnitude and direction of any publication bias and possible explanations for why and how the publication bias occurs. The survey consisted of two main parts, part A and part B, and is given in Supplementary Material. The results were used to perform the sensitivity analyses adjusting for publication bias, as described further below.

Data set and notation

The impact on OS of optimal and suboptimal cytoreduction for primary advanced disease was assessed using several RD thresholds that have been reported in the literature. Accordingly, our data set consists of the results of n studies, comparing a total of T RD thresholds (or arms; labeled 1,2,..). We use the terms, arms and RD thresholds interchangeably for the benefit of those mimicking our methods because it is likely that they will be applying the methodology to study arms in an RCT setting. We use the term design to refer to the set of RD thresholds compared in a given study, that is, a design is some subset of at least 2 RD thresholds in the network. Let $d = 1, \ldots, D$ index the designs used in our network, and n_d be the number of studies included in the network that used the dth design. Set also T_d as the number of RD thresholds in design d. Then, we have designs with T_d = 2,3,4. The designs in our data set are presented in Figure 2; we have n = 28, $T = 9$, and $D = 8$.

From a study of design d, the information used is: (a) $T_d - 1$ estimated effects (log hazard ratios, in our case) and their standard errors and; (b) $(T_d - 1)(T_d - 2)/2$ correlations between the $T_d - 1$ effects. We use subscript indices to identify this study and its design and superscript indices to denote the contrast being evaluated such that $y_{i,d}^{(a,b)}$ refers to the effect size for the ab comparison (where a and b are 2 RD thresholds) in the ith study that has the dth design. Similarly, we let $s_{i,d}^{(a,b)}$ denote the corresponding observed standard error (SE). Our data set, in this notation, is available on request.

Part A: Copas model approach

Part A of the elicitation exercise asked clinicians about their perceived probability of publication of individual studies relating to the standard error of their effect sizes. Part A was conducted to facilitate the conduct of a previously proposed method of adjusting for publication bias in a NMA.[24] We now describe this methodology.

Measurement model

Each observed effect $y_{i,d}^{(a,b)}$ in a two-threshold study (i.e., any study with $T_d = 2$) is modeled as a normal distribution:

$$y_{i,d}^{(a,b)} \sim N\left(\theta_{i,d}^{(a,b)}, \left(s_{i,d}^{(a,b)}\right)^2\right).$$

A random-effects model is assumed because publication bias is confounded with heterogeneity. Thus, it is assumed that the mean relative treatment effect is modeled as $\theta_{i,d}^{(a,b)} = \lambda^{(a,b)} + \delta_{i,d}^{(a,b)}$, where the random effects $\delta_{i,d}^{(a,b)}$ are normally distributed as $\delta_{i,d}^{(a,b)} \sim N(0, \tau^2)$.

In multithreshold study i of design d, the vector of $T_d - 1$ contrasts is modeled as a multivariate normal distribution. For example, if arms a, b, and c are included, then:

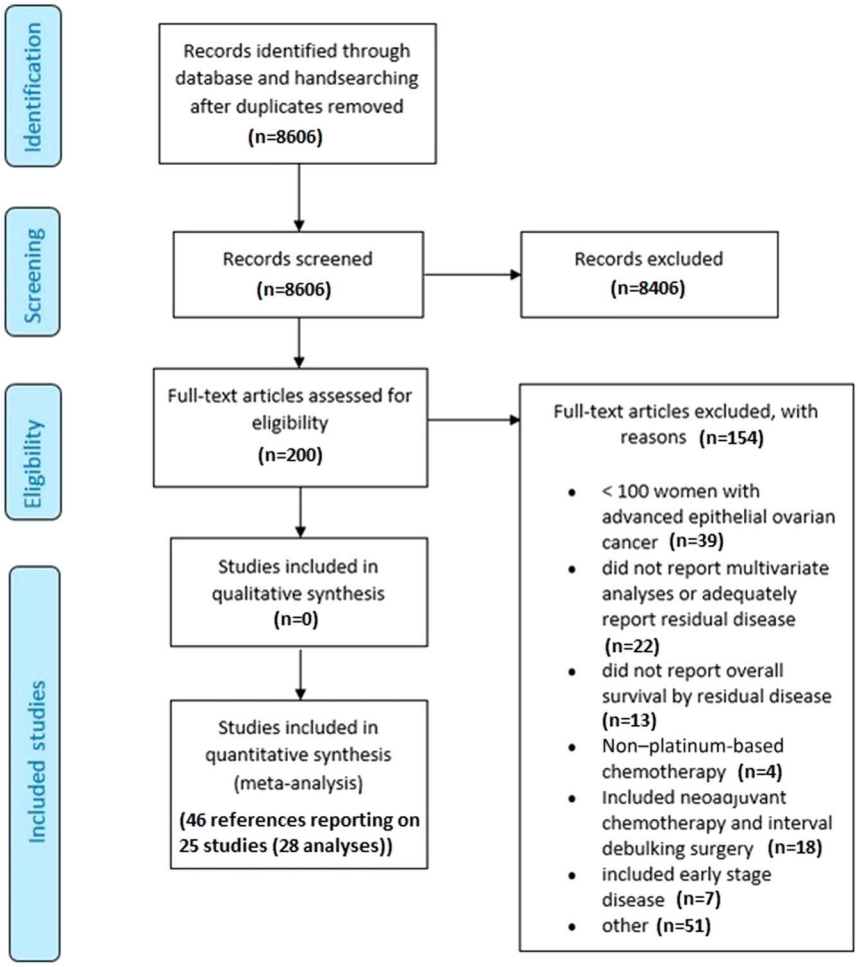
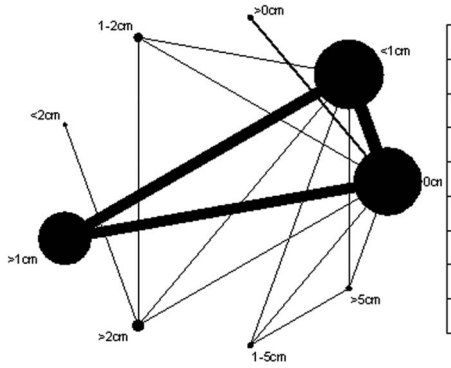


FIGURE 1. PRISMA flowchart.



Design (d)	Arms	T_d	n_d
1	1, 2, 4, 7	4	1
2	1, 2, 8, 9	4	1
3	1, 2, 6	3	15
4	2, 4, 7	3	1
5	1, 2	2	2
6	1, 3	2	4
7	2, 6	2	2
8	5, 7	2	2

FIGURE 2. Network diagram and summary of designs showing RD comparisons after primary cytoreductive surgery for advanced EOC. Arms 1–9 correspond to the following categories: 1 (0 cm), 2 (<1 cm), 3 (>0 cm), 4 (1–2cm), 5 (<2 cm), 6 (>1 cm), 7 (>2 cm), 8 (1–5cm), and 9 (>5 cm).

$$\begin{pmatrix} ny_{i,d}^{(a,b)} \\ ny_{i,d}^{(a,c)} \end{pmatrix} \mid \begin{matrix} n\theta_{i,d}^{(a,b)} \\ n\theta_{i,d}^{(a,b)} \end{matrix} \sim N\left(\begin{pmatrix} n\lambda^{(a,b)} \\ n\lambda^{(a,c)} \end{pmatrix} + \begin{pmatrix} n\delta_{i,d}^{(a,b)} \\ n\delta_{i,d}^{(a,c)} \end{pmatrix}, \begin{pmatrix} \left(s_{i,d}^{(a,b)}\right)^2 & cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,c)}\right) \\ cov\left(y_{i,d}^{(ab)}, y_{i,d}^{(a,c)}\right) & \left(s_{i,d}^{(a,c)}\right)^2 \end{pmatrix} \right).$$

Assuming a common heterogeneity parameter across treatment comparisons, the random effects are:

$$\begin{pmatrix} n\delta_{i,d}^{(a,b)} \\ n\delta_{i,d}^{(a,c)} \end{pmatrix} \sim N\left(\begin{pmatrix} n0 \\ n0 \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right).$$

Similarly, if arms *a*, *b*, *c*, and *e* are included, then:

$$\begin{pmatrix} ny_{i,d}^{(a,b)} & n\theta_{i,d}^{(a,b)} \\ ny_{i,d}^{(a,c)} & n\theta_{i,d}^{(a,c)} \\ ny_{i,d}^{(a,e)} & n\theta_{i,d}^{(a,e)} \end{pmatrix} \sim N\left(\begin{pmatrix} n\lambda^{(a,b)} \\ n\lambda^{(a,c)} \\ n\lambda^{(a,e)} \end{pmatrix} + \begin{pmatrix} n\delta_{i,d}^{(a,b)} \\ n\delta_{i,d}^{(a,c)} \\ n\delta_{i,d}^{(a,e)} \end{pmatrix}, \begin{pmatrix} \left(s_{i,d}^{(a,b)}\right)^2 & cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,c)}\right) & cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,e)}\right) \\ cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,c)}\right) & \left(s_{i,d}^{(a,c)}\right)^2 & cov\left(y_{i,d}^{(a,c)}, y_{i,d}^{(a,e)}\right) \\ cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,e)}\right) & cov\left(y_{i,d}^{(a,c)}, y_{i,d}^{(a,e)}\right) & \left(s_{i,d}^{(a,e)}\right)^2 \end{pmatrix} \right),$$

$$\begin{pmatrix} n\delta_{i,d}^{(a,b)} \\ n\delta_{i,d}^{(a,c)} \\ n\delta_{i,d}^{(a,e)} \end{pmatrix} \sim N\left(\begin{pmatrix} n0 \\ n0 \\ n0 \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix} \right)$$

We assume a common between-study variance (heterogeneity) τ^2 across treatment comparisons; although arguably not realistic, this is common and often necessary in practice (given there are few studies per comparison).

Selection model

To model the probability each study is selected for publication, we assume that there is a latent variable underlying each study. This latent variable takes positive values if the specific study is published and negative values otherwise. Thus, there are as many latent variables as study designs, and each latent variable represents the propensity for publication given the design of that specific study. The propensity for publication for each design and study is denoted by $z_{i,d}$. It is modeled as a function of two parameters, α_d and β_d, and a function, $f(i,d)$, of the particular study and its design:

$$z_{i,d} = \alpha_d + \frac{\beta_d}{f(i,d)} = u_{i,d} + \xi_{i,d}.$$

Here, $\xi_{i,d} \sim N(0,1)$, and we constrain $\beta_d \geq 0$ because this will reflect the belief that larger studies are more likely to be published. In a two-threshold study i of design d that compares a and b, we set $f(i,d) = s_{i,d}^{(a,b)}$. Following Chootrakool et al (2011)[23] for a multithreshold trial, we use the average of the standard errors in this study. For example, in study i of design d that compares a, b, c, and e, we set:

$$f(i,d) = \frac{s_{i,d}^{(a,b)} + s_{i,d}^{(a,c)} + s_{i,d}^{(a,e)}}{3}.$$

With the above, the probability that study i with design d is published is equal to:

$$\mathbb{P}(z_{i,d} > 0) = \Phi\left(\alpha_d + \frac{\beta_d}{f(i,d)}\right) = \Phi(u_{i,d}).$$

This provides us with an interpretation of the parameters α_d and β_d. Informally, parameter α_d is the marginal probability that a study with design d is published, assuming it has infinite variance (not accounting for the approach taken to multithreshold studies). Parameter β_d is a discrimination parameter, discriminating the probabilities of publication between studies with difference variances.

Combined measurement and selection model

The measurement and selection models do not share common parameters but are connected through their residual terms. Specifically, we set $corr(y_{i,d}^{(ab)}, z_{i,d}) = \rho_d$ such that ρ_d controls how the effect size affects the probability of the study being published. Then, for two-threshold (thresholds *a* and *b*), three-threshold (thresholds *a*, *b*, and *c*), and four-threshold (thresholds *a*, *b*, *c*, and *e*) studies, the joint distribution for its effect sizes and propensity for publication are as follows:

$$\begin{pmatrix} ny_{i,d}^{(a,b)} \\ nz_{i,d} \end{pmatrix} \sim N\left(\begin{pmatrix} n\theta_{i,d}^{(a,b)} \\ nu_{i,d} \end{pmatrix}, \begin{pmatrix} \left(s_{i,d}^{(a,b)}\right)^2 & \rho_d s_{i,d}^{(a,b)} \\ \rho_d s_{i,d}^{(a,b)} & 1 \end{pmatrix} \right) I_{z_{i,d}>0},$$

$$\begin{pmatrix} ny_{i,d}^{(a,b)} \\ ny_{i,d}^{(a,c)} \\ nz_{i,d} \end{pmatrix} \sim N\left(\begin{pmatrix} n\theta_{i,d}^{(a,b)} \\ n\theta_{i,d}^{(a,c)} \\ nu_{i,d} \end{pmatrix}, \begin{pmatrix} \left(s_{i,d}^{(a,b)}\right)^2 & cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,c)}\right) & \rho_d s_{i,d}^{(a,b)} \\ cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,c)}\right) & \left(s_{i,d}^{(a,c)}\right)^2 & \rho_d s_{i,d}^{(a,c)} \\ \rho_d s_{i,d}^{(a,b)} & \rho_d s_{i,d}^{(a,c)} & 1 \end{pmatrix} \right) I_{z_{i,d}>0},$$

where I_X is the indicator variable for event X.

$$\begin{pmatrix} ny_{i,d}^{(a,b)} \\ ny_{i,d}^{(a,c)} \\ ny_{i,d}^{(a,e)} \\ nz_{i,d} \end{pmatrix} \sim N\left(\begin{pmatrix} n\theta_{i,d}^{(a,b)} \\ n\theta_{i,d}^{(a,c)} \\ n\theta_{i,d}^{(a,e)} \\ nu_{i,d} \end{pmatrix}, \begin{pmatrix} \left(s_{i,d}^{(a,b)}\right)^2 & cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,c)}\right) & cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,e)}\right) & \rho_d s_{i,d}^{(a,b)} \\ cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,c)}\right) & \left(s_{i,d}^{(a,c)}\right)^2 & cov\left(y_{i,d}^{(a,c)}, y_{i,d}^{(a,e)}\right) & \rho_d s_{i,d}^{(a,c)} \\ cov\left(y_{i,d}^{(a,b)}, y_{i,d}^{(a,e)}\right) & cov\left(y_{i,d}^{(a,c)}, y_{i,d}^{(a,e)}\right) & \left(s_{i,d}^{(a,e)}\right)^2 & \rho_d s_{i,d}^{(a,e)} \\ \rho_d s_{i,d}^{(a,b)} & \rho_d s_{i,d}^{(a,c)} & \rho_d s_{i,d}^{(a,e)} & 1 \end{pmatrix} \right) I_{z_{i,d}>0},$$

Prior distributions for model selection parameters

To fit the above model, prior distributions for the selection model parameters α_d and β_d are required. To form the priors, we need to specify lower and upper bounds, P_d^{low} and P_d^{high}, for the probability that a study of design d is published, where these extremes relate to small and large possible values of $f(i,d)$. P_d^{low} and P_d^{high} are modeled as random variables to reflect the uncertainty around them. Then, α_d and β_d are calculated using the inequalities:

$$P_d^{low} \leq \mathbb{P}(z_{id}>0|f(i,d)) \leq P_d^{high} \forall\, d.$$

Specifically, this gives:

$$\alpha_d + \frac{\beta_d}{\max\{f(i,d)\}} = \Phi^{-1}\left(P_d^{low}\right),$$

$$\alpha_d + \frac{\beta_d}{\min\{f(i,d)\}} = \Phi^{-1}\left(P_d^{high}\right).$$

Unlike Mavridis et al,[22] rather than setting $\min\{f(i,d)\}$ and $\max\{f(i,d)\}$ as the observed minimal and maximal values in the data set, we use the results of an elicitation exercise in which we asked experts about the probability studies of certain sizes would be published. For the population under investigation, we describe below why $\mathrm{SE}(\log HR) = \sqrt{6.25/n}$ is a reasonable assumption. Using this, on plugging in the minimal and maximal sample sizes from the elicitation exercise, the formulae for $f(i,d)$ gives:

$$\min\{f(i,d)\} = \begin{cases} 0.1 & : T_d = 2, \\ 0.122 & : T_d = 3, \\ 0.141 & : T_d = 4, \end{cases}$$

$$\max\{f(i,d)\} = \begin{cases} 0.25 & : T_d = 2, \\ 0.306 & : T_d = 3, \\ 0.354 & : T_d = 4. \end{cases}$$

All that then remains is to specify prior distributions $P_d^{low} \sim U(L_{1,d}, L_{2,d})$ and $P_d^{high} \sim U(U_{1,d}, U_{2,d})$. For those two-threshold designs that include the 0 cm arm, we are able to directly use the results from Part A of the survey. For those studies that did not contain the 0 cm arm, we calculate, similarly, swapping in their reference category for 0 cm (e.g., the probability of publication of <1 cm versus >2 cm would be taken as the elicited values for 0 cm versus >2 cm); the results are unlikely to be sensitive to this assumption because the number of studies that do not contain the reference category is small. For multi-threshold studies, we conservatively use the minimum probabilities across the various pairwise comparisons. We then consider three combinations of values for $L_{1,d}$, $L_{2,d}$, $U_{1,d}$, and $U_{2,d}$. We take them as the 0th and 50th (median) percentiles, the 25th (lower quartile) and 75th (upper quartile) percentiles, and the 50th and 100th percentiles of the elicited probabilities (with $L_{1,d}$ and $L_{2,d}$ set using the results for the smallest trial size we asked experts about and $U_{1,d}$ and $U_{2,d}$ set using the results for the largest trial size we asked experts about). We denote the elicited q^{th} percentile for the small study size by $P_{s,d,q}$ and similarly, $P_{l,d,q}$ for the large. The percentiles are then presented in Table 2.

Part B: alternative novel approach

Part B involved an alternative approach that asked clinicians to estimate the number of studies for key comparisons that they believed would be conducted but unpublished, and thus unidentified in the NMA. They were then subsequently asked to specify sample and effect sizes for each such missing study. The approach in Part B is a particularly novel aspect of this research because it can be used as prior information to inform adjustment of meta-analyses for publication bias in a way we believe to be previously unexplored. Here, we outline how this could be achieved.

We note that this is only one potential way to form a prior based on the elicited data and that a sensitivity analysis should certainly be conducted. For example, in our elicitation exercise, the choice of the number of miss-

ing studies was left open ended as to not lead experts to a choice and bias the results. Consequently, a sensitivity analysis could be conducted removing high estimates of unpublished studies if it was judged that unrealistic entries were unduly inflating an average.

Given an assumed 5-year survival rate of 36%[32–34] and a minimum sample size of n = 100 to meet the criteria for inclusion in the NMA, small studies might be underpowered and, furthermore, null findings might be due to deficiencies in the study design and conduct. Hence, including these studies might not lead to an appropriate adjustment of meta-analysis estimates. This is why we included studies with a minimum sample size of 100 patients in the systematic review,[1] and a minimum 64 events (deaths, d) are expected with 36 participants being alive and censored at the end of this study:

$$d = 100(1 - 0.36) = 64.$$

Generalizing this result, we assume that d can be related to n in general through the following formula:

$$n = \frac{d}{1 - (5 \text{ year suvival rate})} = \frac{d}{0.64}.$$

The standard error of the log hazard ratio $(\text{SE}(\log HR))$ can then be related to n by rearranging the following formula:

$$d = \frac{4}{\text{SE}(\log HR)^2},$$

$$\Rightarrow \text{SE}(\log HR) = \sqrt{\frac{4}{d}} = \sqrt{\frac{4}{0.64n}} = \sqrt{\frac{6.25}{n}}.$$

Next, we denote by m_{cij} the number of missing studies according to expert responder $c = 1, \ldots, C$, with a HR of HR_j and a sample size of n_i, where:

$$n_1 = 100, n_2 = 200, n_3 = 300, n_4 = 400, n_5 = 500, n_6 = 625,$$

$$HR_1 = 1, HR_2 = 0.9, HR_3 = 0.8, HR_4 = 0.7, HR_5 = 0.6, HR_6 = 0.5.$$

We compute the average number of missing studies of type ij, across the responders, as:

$$m_{ij} = \frac{1}{C} \sum_{c=1}^{C} m_{cij}.$$

We use this to form an average sample size of missing studies with a HR of HR_j through:

$$m_j = \frac{\sum_i n_i m_{ij}}{\sum_i m_{ij}}.$$

With this, we assume that information from missing studies with a HR of HR_j can be categorized through the following distribution:

$$P_j \sim N\left(\log HR_j, \frac{6.25}{m_j}\right).$$

The P_j can then be combined in a weighted manner, giving more weight to those values of j with a larger value of m_j, through conflation. This gives an elicited prior of:

$$P \sim N\left(\frac{\sum_j \frac{m_j \log HR_j}{6.25}}{\sum_j \frac{m_j}{6.25}}, \frac{1}{\sum_j \frac{m_j}{6.25}}\right) = N\left(\frac{\sum_j m_j \log HR_j}{\sum_j m_j}, \frac{6.25}{\sum_j m_j}\right).$$

This elicited estimate can then be used as prior information and be applied in a Bayesian analysis[35–37] that reflects the results of the expert opinion in the elicitation exercise.[22,38]

Data analysis

We compare the results of a frequentist approach[2] with a NMA conducted within a Bayesian framework in WinBUGS 1.4.3 (MRC Biostatistics Unit, Cambridge, UK),[39,40] using two chains each with 100,000 simulations and a burn-in period of 30,000 simulations. The base case Bayesian analysis (analogous to the frequentist analysis) used vague noninformative priors and adjusted for multiarm trials using conditional distributions. Figure 2 shows a network diagram[41] of the thresholds (nodes) and comparisons (lines) available and a summary of designs in our network. Convergence of the model in the two chains was assessed using Brooks–Gelman–Rubin, trace and autocorrelation plots.[40]

Transitivity and design inconsistency were not deemed an issue because of restrictive inclusion criteria.[2] Consistency, which is measured in agreement of direct and indirect evidence, was assessed by comparing the individual data point's posterior mean deviance contributions for the consistency and inconsistency model.[42–44] Owing to the volume of sensitivity analyses, we did not conduct any further node splitting[42–44] because this was previously performed in the base case analysis.[2]

We present the results of the Bayesian NMA of optimal RD thresholds using effect sizes reported as posterior median HRs and 95% credible intervals (CrIs). All the thresholds are relative to the 0 cm macroscopic

RD reference threshold. We also present rankograms, which ranked RD thresholds from having highest probability of survival (ranked 1) to the lowest (ranked 9). In addition, we report the probability of being the best RD threshold and the surface under the cumulative ranking curves (SUCRAs).[45]

Sensitivity analyses (SA) form the crucial basis of this research; we perform a number of analyses that attempt to adjust the base case estimates for publication bias. We a priori focus on macroscopic RD to 0 cm, RD <1 cm, and suboptimal RD >1 cm. Other thresholds will add strength to the network but are not of direct interest. We use this approach in a complex situation that includes multiple RD thresholds (arms) and studies that included multiple thresholds (up to four in a study). In practice, it should be more straightforward following and applying the methodology to other analyses in different areas that have simpler networks and in a conventional intervention setting.

We repeated the base case Bayesian analysis above and used the elicitation exercise to use the Copas selection model (part A) and incorporate informative priors (part B) in place of the vague (noninformative) ones. For those wanting to restrict to a frequentist setting, in Part B, an analogous analysis in the frequentist framework is possible by including the elicited missing studies from the average experts' responses (artificially) in the observed studies in the NMA. However, we recommend applying the proposed Bayesian methods and formulating priors.

RESULTS

Summary of studies

The flow of the literature is presented in the PRISMA flowchart (Figure 1). The search strategy identified 8606 unique references, of which 200 progressed to full-text screening. Forty-six references, reporting on 25 primary studies which included 28 analyses (n = 20,927), met our inclusion criteria. Full details of searches along with a PRISMA flowchart, characteristics of included studies, and risk of bias assessments are provided by Bryant et al.[2]

The network diagram[41] and summary of designs in our network show the range of RD threshold comparisons after optimal cytoreductive surgery for advanced EOC (Figure 2). The most common RD thresholds were complete (0 cm) and near-optimal (<1 cm), while this was also the most widely reported comparison.

Base case analysis

The results of the base case Bayesian NMA were consistent with the frequentist analysis, and there was also no evidence of inconsistency in the network.[2] There were no issues with model convergence in WinBUGS,[39] as indicated by Brooks–Gelman–Rubin, trace and autocorrelation plots, with the number of simulations used adequate (see Appendix Figures 3–8, http://links.lww.com/AJT/A124, http://links.lww.com/AJT/A125, http://links.lww.com/AJT/A126, http://links.lww.com/AJT/A127, http://links.lww.com/AJT/A128, http://links.lww.com/AJT/A129, and http://links.lww.com/AJT/A130, respectively). The results of the base case analyses demonstrate prolonged OS if primary cytoreductive surgery achieved macroscopic RD to 0 cm compared with any other RD threshold (Table 1). Macroscopic RD to 0 cm was overwhelmingly the best ranked threshold because it was consistently ranked first (Table 1 and rankogram in Figure 9 in Supplementary Material, http://links.lww.com/AJT/A131), with a very high probability of being the best RD threshold (SUCRA and P best ranged from 98.4% to 99.9%). Sensitivity analyses using different random number seeds resulted in all Bayesian models being correct to one decimal place (data not shown). Low values in MC error terms in the model indicated reliability in estimates to good precision.[40]

Expert elicitation exercise

Eighteen expert members of the BGCS participated in the expert elicitation exercise. They were given the sample sizes (based on observed data for each RD threshold) and were asked in Part A of the elicitation exercise to state probabilities of publication for a study comparing different RD thresholds with complete cytoreduction (macroscopic RD to 0 cm). Table 2 presents the distribution of the elicited probabilities for each RD threshold, for the smallest and largest considered study sizes (full details of the expert clinician elicitation exercise are provided by Bryant et al[31]). In summary, responses suggest that publication bias may be quite likely in studies where the sample size was small. For example, the average response suggested that experts believed there was a 55% chance that a comparison of RD < 1 cm versus RD 0 cm would be reported for a study with a sample size of 100 participants. Responders seemed to indicate that the probability of publication was lowest for comparisons involving greater macroscopic disease volume [largest elicited median probability 20% (interquartile range 10–75) in macroscopic disease involving RD > 2 cm versus RD 0 cm and as low as 3.5% (interquartile range: 0–50) for RD > 5 cm vs. RD 0 cm]. However, respondents seemed to dismiss the threat of publication bias for comparisons of RD < 1 cm versus RD 0 cm and RD > 1 cm versus RD 0 cm in larger studies. Comparisons involving suboptimal

Table 1. Results of base case frequentist and Bayesian NMA of optimal RD threshold after primary cytoreductive surgery for advanced EOC.

	Frequentist				Bayesian			
RD	HR (95% CI)	Mean rank	P (best), %	SUCRA, %	HR (95% CrI)	Median rank	P (best), %	SUCRA, %
0 cm	Reference	1	99	99.9	Reference	1 (1–1)	98.42	99.8
<1 cm	1.98 (1.76–2.24)	3.4	0	70.2	1.99 (1.76–2.27)	3(2–5)	0	69.88
>0 cm	1.95 (1.48–2.58)	3.4	0	70.6	1.95 (1.46–2.63)	3(2–6)	0.005	70.43
1–2 cm	3.34 (2.04–5.47)	7.3	0	21.8	3.57 (2.14–5.99)	8(5–9)	0	18.58
<2 cm	2.82 (1.58–5.04)	6.0	0	36.9	2.89 (1.57–5.34)	7(2–8)	0.044	36.75
>1 cm	2.57 (2.26–2.93)	5.8	0	40.0	2.58 (2.26–2.97)	6(4–8)	0	40.91
>2 cm	4.36 (2.69–7.04)	8.7	0	3.4	4.47 (2.72–7.43)	9(7–9)	0	4.17
1–5cm	1.85 (1.11–3.08)	3.2	1	72.0	1.85 (1.06–3.22)	3(2–7)	1.498	71.93
>5 cm	2.75 (1.62–4.67)	6.2	0	35.3	2.75 (1.55–4.89)	6(2–9)	0.033	37.54

RD, residual disease; CI, confidence interval; P (best), probability that RD threshold is the best; CrI, credible interval; EOC, epithelial ovarian cancer; HR, hazard ratio; SUCRA, surface under the cumulative ranking curves.

RD (greater macroscopic disease volume) were considered to have a low probability of not being published for both the small and larger studies (but lower in smaller studies).

In part B of the elicitation exercise, the mean number of missing studies estimated by experts for comparison of RD < 1 cm versus RD 0 cm was 17.8. The average number of estimated missing studies was lower for the comparisons involving suboptimal macroscopic disease volume (RD thresholds that are > 1 cm).[31] In the comparison of RD < 1 cm versus RD 0 cm, on average, 9.4 of the 17.8 studies would be associated with a HR of 1. As the HR increased, fewer studies were believed to be missing such that, when the detected HR was 0.5, the average number of studies believed to be missing was less than 1 (Table 3). The weighted average HR of the effect size from the missing studies was 0.83 (95% CI 0.77–0.90) for the comparison of RD < 1 cm compared with RD 0 cm. This HR was calculated based on a total of 3906 participants in the estimated missing studies and 2500 deaths given a 5-year survival rate of 36% (Table 3). This corresponded to a log HR of −0.19 and SE log HR of 0.04; thus, we used ~N(−0.19, 0.04) as the distribution for our elicited prior for the <1 cm versus 0 cm comparison. Similarly, the mean number of missing

Table 2. The distribution of elicited probabilities for each RD threshold for the smallest and largest considered study sizes.

	Small study publication probabilities					Large study publication probabilities				
Design (*d*)	$P_{s,d,0}$	$P_{s,d,25}$	$P_{s,d,50}$	$P_{s,d,75}$	$P_{s,d,100}$	$P_{l,d,0}$	$P_{l,d,25}$	$P_{l,d,50}$	$P_{l,d,75}$	$P_{l,d,100}$
1	0	10	20	70	100	0	15	30	80	100
2	0	0	3.5	50	95	0	0	10	80	100
3	0	20	45	80	100	40	75	95	99	100
4	0	10	20	70	100	0	15	30	80	100
5	0	30	55	80	100	80	90	99.5	100	100
6	0	20	50	80	95	0	70	80	99	100
7	0	20	45	90	100	40	75	95	99	100
8	0	10	20	75	100	0	15	30	80	100

These are computed using the results of the elicitation exercise.
P,(*s,l*),*d*,(percentiles 0, 25, 50, 75, 100), probability that a small/large study is published with a specified design in a number of percentiles. Design (*d*) 1, arms 1,2,4,7; *d*(2), arms 1,2,8,9; *d*(3), arms 1,2,6; *d*(4), 2,4,7; *d*(5), arms 1,2; *d*(6), arms 1,3; *d*(7), arms 2.6; *d*(8), arms 5,7 where arms 1–9 correspond to the following categories: 1 (0 cm), 2 (<1 cm), 3 (>0 cm), 4 (1–2cm), 5 (<2 cm), 6 (>1 cm), 7 (>2 cm), 8 (1–5cm), and 9 (>5 cm).

Table 3. Breakdown of distribution of size and magnitude of elicited unpublished studies of near-optimal RD < 1 cm versus complete cytoreduction (0 cm).

N=321 (n=17.8)	Estimated effect size					
	HR = 1	HR = 0.9	HR = 0.8	HR = 0.7	HR = 0.6	HR ≤ 0.5
Assumed 5-year survival: 36%	RD <1 cm and 0 cm are the same	10% less chance of mortality favoring RD <1 cm	20% less chance of mortality favoring RD <1 cm	30% less chance of mortality favoring RD <1 cm	40% less chance of mortality favoring RD <1 cm	>=50% less chance of mortality favoring RD <1 cm
Sample size[h]						
n < 100			STUDY EXCLUDED			
n = 100	122.08[g]	19.12	22.7	1.34	2.14	1.14
n = 200	25.08	11.12	12.62	4.38	2.18	2.18
n = 300	6.04	4.04	1.04	2.04	0	0
n = 400	10.37	9.37	9.37	9.37	9.37	9.37
n = 500	1.04	1.04	3.04	1.04	0	0
n > 500	5.08	4.04	4.04	3.04	1.04	1.04
Total studies[a] (mean)	169.7 (9.4)	48.7 (2.7)	52.8 (2.9)	21.2 (1.2)	14.7 (0.8)	13.7 (0.8)
Effective *n*[b] (mean)	26,879 (1493.3)	12,141 (674.5)	12,899 (716.6)	7790 (432.8)	5048 (280.4)	4948 (274.9)
Effective *d*[c] (mean)	17,203 (956)	7770 (432)	8255 (459)	4986 (277)	3231 (179)	3167 (176)
SElogHR ($\sqrt{(4/d)}$)[d]	0.065	0.096	0.093	0.120	0.149	0.151
95% CI for HR[e]	0.88–1.14	0.75–1.09	0.67–0.96	0.55–0.89	0.45–0.80	0.37–0.67
Elicited estimate[f]	HR 0.83 (95% CI 0.77–0.90), logHR −0.19 SElogHR 0.04 (n = 3906, *d* = 2500)					
Elicited prior	~N(−0.19, 0.04)					

[a]Absolute number of estimated missing studies elicited from responders with mean (simply absolute number divided by 18 (number of responders)) given in parentheses ().
[b]Absolute number of estimated missing participants elicited based on total studies with mean given in parentheses.
[c]Absolute number of deaths estimated from the number of participants assuming a 5-year survival rate of 36% with mean in parentheses ().
[d]Approximation of the standard error (SE) of the log HR using formula derived by Parmar, namely the square root of 4 divided by the mean number of deaths.
[e]95% confidence interval for HR calculated using logHR ±1.96 multiplied by standard error of log HR then transforming back by taking the exponential.
[f]Elicited HR with 95% confidence interval using mean responses for all aggregated effect sizes.
[g]Number of studies given in the breakdown were rescaled in 3 respondents to correspond to the total number estimated. Therefore, any noninteger numbers in the table are due to this rescaling.
[h]Size of studies missed that could have been included in the analysis.

studies estimated in the elicitation exercise for comparison of RD > 1 cm versus RD 0 cm was 8.6.[31] The weighted average HR of the missing studies led to formulating ~N(−0.26, 0.05) as a prior. The mean number of missing studies estimated by responders for comparison of RD > 2 cm versus RD 0 cm was 6.2.[31] The weighted average HR of the missing studies led to formulating ~N(−0.24, 0.06) as a prior. However, there seemed to be widespread feeling among experts that publication bias was of much less concern

in suboptimal RD thresholds, and this is reflected in some of the sensitivity analyses (Table 5). A worked example surrounding derivation of priors based on these estimates is presented in Table 3 for the comparison of macroscopic RD with 0 cm and near-optimal cytoreduction to <1 cm.

Adjustment for publication bias

Tables 4 and 5 present the estimated effect sizes for RD thresholds for the sensitivity analyses incorporating an adjustment for publication bias. Models were constructed using responses from parts A and B of the expert elicitation exercise.

All analyses were based on 100,000 Markov Chain Monte Carlo simulations with a burn-in period of 30,000 draws, from two chains (as in Mavridis et al[22]). We present the median OS estimate for each RD group relative to the reference category (0 cm), along with its 95% CrI, SUCRA values, the median (and 95% CrI) rank for each group, and the estimated probability each group provides the best OS are also given.

Bayesian NMAs were fitted in a series of sensitivity analyses that used informative priors based on estimates obtained from the expert elicitation exercise (see above). We set out to explore a range of sensitivity analyses, from ones that best reflected the experts' views to more extreme scenarios that fully tested the robustness of the base case analysis presented in Table 1. Specifically, the main focus of our work was to examine the conclusions in the unadjusted analysis that identified three clear and distinct categories of RD groups after primary cytoreductive surgery, namely complete (0 cm), near-optimal (<1 cm), and suboptimal (>1 cm). Other reported RD thresholds contributed toward the network and added strength to the NMA, but clearly some comparisons such as when RD 1–2 cm is compared with macroscopic RD to 0 cm were not of paramount importance and would not necessarily be a widely reported and expected comparison. Therefore, it would not be appropriate to focus on publication bias in this example. Accordingly, sensitivity analyses focused on the main RD categories of complete RD to 0 cm, RD<1 cm, and suboptimal RD >1 cm.

Part A: Copas selection model

Table 4 presents the results of the selection model analyses. As would be expected, the introduction of increasing levels of publication bias adjustment typically results in greater reductions in the estimated OS benefit for the reference category. However, in almost all instances the results change little compared with the base case frequentist and Bayesian analyses (Table 1). The 0 cm category retains at least an 87.48% estimated chance of providing the best OS.

Part B: alternative novel approach

Sensitivity analysis (SA) 1 incorporated prior information using the estimates derived above (~N(−0.24, 0.06)) for RD <1 cm and RD > 0 cm, ~N(−0.26, 0.05) for RD >1 cm, and ~N(−0.24, 0.06) for RD>2 cm). In SA 2, informative priors were used for RD <1 cm, >0 cm, 1–2cm, <2 cm, and >1 cm and only RD < 1 cm and >0 cm in SA 3. SA 4 used informative priors for RD <1 cm, >1 cm, and >2 cm, and SA 5 incorporated priors for all RD thresholds. SA 5 was thus the most extreme sensitivity analysis.

SA 6 and 7 grouped RD < 2 cm into the RD < 1 cm threshold to reduce the number of RD groups to eight. This was due to the fact that RD < 2 cm was sparsely reported in the observed NMA comparisons because suboptimal RD is now clearly defined as >1 cm in the guidelines and the RD < 2 cm group was obtaining undue influence in the ranking statistics, which was wholly implausible. SA 6 incorporated prior information for RD < 1 cm, RD > 0 cm, >1 cm, and >2 cm. SA 7 incorporated prior information for RD < 1 cm, >0 cm, and >1 cm.

All sensitivity analyses were in line with the base case analysis and demonstrated prolonged OS if primary cytoreductive surgery achieved macroscopic RD to 0 cm compared with any other RD threshold (Table 5). However, the effect estimates were attenuated in comparisons involving macroscopic RD to 0 cm, although not to any suggestion of changing the existing conclusions. This was even the case in the most extreme sensitivity analysis (SA 5) that used all RD thresholds, including ones that would not have been expected to have been widely reported in reality. There remained three clear and distinct categories of RD thresholds after primary cytoreductive surgery. Complete macroscopic RD to 0 cm is still by far the best surgical option, with near-optimal (<1 cm) debulking a consolation if this is not possible. Suboptimal can therefore be defined as RD > 1 cm.

There were no issues with model convergence or other diagnostics in WinBUGS[39] in any of the sensitivity analyses, as previously indicated in the base case analysis.

DISCUSSION

There is a high level of uncertainty facing women undergoing treatment for advanced EOC, especially given differences in practice between surgeons in the United Kingdom and internationally. There are many

Table 4. Results of the selection model analyses are given, for the 3 considered sets of priors for the publication probabilities.

	$L_{1,d} = P_{s,d,0}, L_{2,d} = P_{s,d,50}, U_{1,d} = P_{l,d,0}, U_{2,d} = P_{l,d,50}$				$L_{1,d} = P_{s,d,25}, L_{2,d} = P_{s,d,75}, U_{1,d} = P_{l,d,25}, U_{2,d} = P_{l,d,75}$				$L_{1,d} = P_{s,d,50}, L_{2,d} = P_{s,d,100}, U_{1,d} = P_{l,d,50}, U_{2,d} = P_{l,d,100}$			
RD	HR (95% CrI)	Median rank (95% CrI)	P (best), (%)	SUCRA (%)	HR (95% CrI)	Median rank (95% CrI)	P (best), (%)	SUCRA (%)	HR (95% CrI)	Median rank (95% CrI)	P (best), (%)	SUCRA (%)
0 cm	—	1 (1–2)	96.47	99.55	—	1 (1–2)	87.48	98.38	—	1 (1–2)	96.53	99.56
<1 cm	1.93 (1.72–2.19)	3 (2–5)	0	68.58	1.93 (1.72–2.19)	4 (2–6)	0	63.11	1.95 (1.73–2.22)	3 (2–5)	0	68.95
>0 cm	1.89 (1.41–2.53)	3 (2–6)	0	69.45	1.92 (1.44–2.60)	4 (2–7)	0	62.61	1.94 (1.44–2.60)	3 (2–6)	0	68.40
1–2cm	3.28 (1.97–5.48)	8 (4–9)	0	19.91	3.36 (1.98–5.67)	8 (5–9)	0	17.49	3.42 (2.04–5.78)	8 (4–9)	0	19.06
<2 cm	2.72 (1.49–5.00)	6 (2–8)	0.07	36.89	2.80 (1.52–5.21)	7 (2–8)	0.06	32.26	2.80 (1.50–5.18)	7 (2–8)	0.06	36.80
>1 cm	2.49 (2.19–2.85)	6 (4–8)	0	39.43	2.51 (2.20–2.89)	6 (5–8)	0	34.92	2.53 (2.22–2.91)	6 (4–8)	0	39.65
>2 cm	4.10 (2.51–6.81)	9 (7–9)	0	5.13	4.28 (2.57–7.13)	9 (7–9)	0	3.52	4.30 (2.59–7.24)	9 (7–9)	0	4.79
1–5cm	1.75 (0.96–3.17)	3 (1–7)	3.28	71.99	1.45 (0.79–2.61)	2 (1–6)	10.99	80.27	1.77 (0.96–3.22)	3 (1–7)	3.26	72.22
>5 cm	2.58 (1.39–4.79)	6 (2–9)	0.18	39.07	1.98 (1.05–3.70)	4 (2–8)	1.47	57.45	2.59 (1.37–4.83)	6 (2–9)	0.15	40.56

$P_{(s,l),d}$(percentiles 50, 75, 100)=probability that a small/large study is published with a specified design in 50th, 75th, and 100th percentiles; RD: residual disease; P (best): probability that RD threshold is the best.

reviews and guidelines assessing the effect of remaining RD on OS after primary surgery. Unfortunately, many include low quality studies prone to selection and other biases due to poor design or inadequate conduct of statistical analyses. A NMA with stringent inclusion criteria that minimized selection bias was required to synthesize the evidence in this important clinical area. Because it is an area of great equipoise, to convince opponents and proponents of maximal efforts of surgical debulking alike, estimated effects need to report "fair" effect estimates. Making an adjustment for publication bias using elicited views of gynecological experts, we argue is the best approach to achieving this.

There is limited current guidance on methods for adjusting meta-analyses for publication bias, including strategies for choosing an informative prior.[46] Many systematic reviewers neglect to examine or discuss publication bias.[30,47] We are unaware of any literature on how often authors adjust for publication bias in their primary analyses or the methods they apply when adjustment is performed. However, the Copas model for NMAs has been infrequently cited, and inspection of the citations seems to indicate that nobody has elicited the parameters for its employment previously. More generally, elicitation within the context of meta-analyses is rare, likely because of its associated burden. Part B of our elicitation exercise, which elicits the average magnitude of the effect in missing studies is, to the best of our knowledge, novel. In conventional use of Bayesian methods, when prior information is scarce, it is advantageous to collaborate closely with experts. Prior information can be obtained systematically, and the information gathered can easily be formalized into prior distributions,[48] as we showed in our elicitation exercise. Often prior specifications should use available information because it can be the key to answering questions about populations that otherwise remain unanswered. The search for prior information may be intensive and time consuming, but the rewards are obvious because meta-analyses are almost all exclusively subject to some degree of reporting bias; therefore, we can improve the reliability of effect estimates by adjusting for publication bias. In our case, we used expert elicitation, but this could use some other systematic approach, with the key message that applying some kind of sensible adjustment for publication bias is better than doing nothing in most cases. However, incorporating prior information that disagrees with the information contained in data can lead to spurious conclusions, particularly if the prior is too informative. Obviously, there is no way of knowing this when estimating publication bias a priori, but substantial gains can be

Table 5. Results for the series of sensitivity analyses using elicited priors from part B of elicitation exercise.

	SA 1		SA 2		SA 3		SA 4		SA 5		SA 6		SA 7		
RD	HR (95% CrI)	SUCRA	HR (95% CrI)	SUCRA	HR (95% CrI)	SUCRA	HR (95% CrI)	SUCRA	HR (95% CrI)	SUCRA	HR (95% CrI)	SUCRA	HR (95% CrI)	SUCRA	Rank
0 cm	Ref	88.53	Ref	85.32	Ref	99.6	Ref	91.43	Ref	60.06	Ref	97.02	Ref	99.3	1
<1 cm	1.57 (1.23–1.86)	45.68	1.54 (1.22–1.83)	42.75	1.84 (1.60–2.08)	65.69	1.62 (1.32–1.88)	55.61	1.38 (1.07–1.69)	18.76	1.54 (1.20–1.82	60.01	1.75 (1.48–2.00)	63.18	2
>1 cm	2.00 (1.53–2.40)	19.3	1.97 (1.52–2.38)	18.5	2.45 (2.13–2.82)	40.38	2.07 (1.66–2.43)	23.89	1.78 (1.33–2.23)	0.8332	1.99 (1.51–2.38)	26.09	2.19 (1.83–2.52)	39.67	3

RD: Residual disease; SA: sensitivity analysis; SUCRA: Surface under cumulative ranking curve (using whole network of RD thresholds).
cm: Centimeters; Ref: Reference RD threshold is 0 cm; SD: standard deviation; ~N(mean, SD): Normal distribution with specified mean and SD; Rank: Based on highest median ranking between considered RD thresholds.
SA 1: ~N(−0.19,0.04) used to incorporate prior information for RD <1 cm and RD>0 cm, ~N(−0.26,0.05) for RD >1 cm, and ~N(−0.24,0.06) for RD>2 cm.
SA 2: ~N(−0.19,0.04) used to incorporate prior information for RD <1 cm and RD>0 cm, ~N(−0.26,0.05) for RD 1–2cm, <2 cm, and >1 cm.
SA 3: ~N(−0.19,0.04) used to incorporate prior information for RD <1 cm and RD>0 cm.
SA 4: ~N(−0.19,0.04) used to incorporate prior information for RD <1 cm, ~N(−0.26,0.05) for RD>1 cm, and ~N(−0.24,0.06) for RD>2 cm.
SA 5: ~N(−0.19,0.04) used to incorporate prior information for RD <1 cm and RD>0 cm, ~N(-0.26,0.05) for RD 1–2 cm, <2 cm and >1 cm, and ~N(−0.24,0.06) for RD>2 cm, 1–5 cm, and >5 cm.
SA 6: ~N(−0.19,0.04) used to incorporate prior information for RD <1 cm and RD>0 cm, ~N(−0.26,0.05) for RD >1 cm, and ~N(−0.24,0.06) for RD>2 cm where RD<2 cm grouped in RD<1 cm threshold.
SA 7: ~N(−0.19,0.04) used to incorporate prior information for RD <1 cm and RD>0 cm and ~N(−0.26,0.05) for RD >1 cm where RD<2 cm grouped in RD<1 cm threshold.

achieved when the inclusion of this information is appropriate. The use of well-specified informative priors can result in improved parameter estimation, where effect estimates will be more reliable.[49]

A good systematic review will eliminate some forms of reporting biases by following good research practice, at least by conforming to PRISMA guidelines.[50] However, many forms of assessment for publication bias, such as funnel plots and formal tests of asymmetry, as well as methods for addressing it, such as the trim and fill method, multiple imputation, and extensive searches of gray literature, are not adequate.[10,51] We also showed that the sophisticated selection models used in our analyses using results from part A of the elicitation exercise may also not have made the kind of adjustment for publication bias that reflected the opinions of the experts who participated. This was because the adjustment in part A was minimal. The novel methodology used in part B of the exercise where a prior was formulated from the average number of missing studies with their effect sizes may offer a simple and highly desirable approach. However, the adjustments in part A and B do not give different results leading to different conclusions, but Part B seemed to adjust effect estimates that seemed to more reflect the opinions of the experts. Consequently, there could be more scope for the results in Parts A and B to differ if this exercise was repeated in the future. In either approach, it is important to specify methods a priori as to not abuse the results by making post hoc adjustments. If the results from the two methods do differ in any future exercise in another research discipline, researchers can use our recommended sensitivity analyses to assess the impact on their conclusions. If there are widespread differences in results, then the confidence in any adjustment for publication bias will be low and it may be most appropriate to report the unadjusted effect estimates as the primary result.

In most areas where a meta-analysis is feasible, there will be experts in the area, so it is advisable to approach these or organizations like we did with the BGCS. Gynecological cancers are common in women, but other diseases may be more rare and not have the same kind of membership in such a society. Therefore, attempts to invite individual experts to participate in any elicitation exercise may be the only option. Thought should be given as to how experts could contribute to such exercises. When conducting a Bayesian analysis, it is important to always provide the origin of and reason behind the priors. We achieved this through our detailed elicitation exercise and a critique of the answers that each responder gave. We also provided the exact specifications of the priors.[48] We also strongly advise those replicating our methods to

conduct a series of sensitivity analyses as we did in Parts A and B of the elicitation exercise when we applied the results to priors. We clearly demonstrated the impact of various priors on the posterior estimates, ranging from noninformative to skeptical priors, to test the robustness of the conclusions.[48,52] A different research area may differ in the impact of such a wide range of priors, if for example, it was known evidence would be very limited in, say, a rare disease. Setting overly skeptical priors in this such a setting may not be sensible, but it is important to set out these justifications a priori if adjustments for publication bias are to have an impact. It is important to understand and interpret any differences between analyses with different priors.

Previous analyses[2] had shown a clear survival benefit of complete cytoreduction to no visible disease after primary cytoreductive surgery in women with advanced EOC. In a Bayesian framework, extreme value sensitivity analyses examined the plausibility of overturning conclusions in the base case analyses. There seemed to be little likelihood that the existing conclusions were not reliable. The selection model indicates that our findings are robust to large levels of publication bias. The elicited estimate used in Part B of the elicitation exercise as an adjustment for publication bias was also robust to the base case results, but seemed to be more representative of the strength of feeling in the experts' opinions. For example, the mean number of missing studies estimated by experts for comparison of RD $<$ 1 cm versus RD 0 cm was 17.8, corresponding to the derivation of an informative prior (N($-$0.24, 0.06)). Clearly, this had some impact on diluting the magnitude of effect to reflect the omission of unpublished studies in the base case meta-analysis and may compute more unbiased and representative estimates. Further research is now extremely unlikely to change our confidence in the existing estimates of effect. The estimates from the set of sensitivity analyses from the Copas selection model did not have the same desired impact. However, we believe the framework applied as an extension to NMA by incorporating multiarm studies was unique and will be of use in other settings.

Evidence from the literature is not the sole determining factor for clinical decision making. Clinicians also have a preference for 'consensus-based decision making.' This is often through relatively informal sources, such as conversations with clinical colleagues and fellow academic experts. Discussing and trading perspectives can be invaluable in gathering information to form judgments.[53,54] Empirical evidence that incorporates expert elicitation in areas of uncertainty may be of paramount importance to the development of clinical guidelines, enabling the disadvantages of contemporary statistical methodologies to be combined with previously implicit expert consensus. This NMA represents a major update and extension to previous analyses. The NMA adjusted for publication bias using elicited information on the three main comparisons of RD (namely, macroscopic RD to 0 cm, RD $<$ 1 cm, and RD $>$ 1 cm). The sensitivity analyses that use prior information and incorporate this into the effect estimates may help to remove any potential skepticism in the previous findings. The overall certainty of the evidence remains moderate despite a dilution of effect estimates in comparisons involving RD 0 cm. We believe our analyses, which have used advanced statistical methodology and expert opinion, offer the best and most comprehensive currently available evidence base to emphasize the reward for making every effort to perform aggressive surgery in women with advanced EOC, if at all feasible. Our findings should inform clinical guidelines and assist the shared decision-making process between patients, carers, and clinicians in routine practice on selecting the most appropriate choice of primary surgical approach for women with advanced EOC. if at all feasible. This work represents the best available evidence at this time.

Our analyses are also easily replicated in different oncology areas and other diseases. Although various different options for priors are available with various statistical approaches, we strongly recommend applying our methodology. Part B particularly is a very simple but highly effective and desirable approach and the use of incorporating the representative views of experts in their field, results will also be highly relevant, and most importantly, effect estimates should be reliable. Although it is important to test the robustness of the conclusions across all the specified sensitivity analyses, we intentionally used a more complicated exemplar as in practice most reviewers following the methods will be able to apply to more routine meta-analyses. In many cases, this will probably involve just two arm studies and potentially in just two comparison interventions. Previous experience with very straightforward exemplars then encountering difficulties when applying to real life data can be very demotivating, so we were conscious to ensure our methods can be repeated in almost all settings. To adopt the methods in Part B of the elicitation exercise, it merely requires a reliable survival estimate, for example, 5-year survival then follow the steps outlined in the paper. The number of experts required for the elicitation exercise and how many experts would constitute a representative number and the resources available to the research team.

REFERENCES

1. Bryant A, Hiu S, Kunonga P, et al. Impact of residual disease as a prognostic factor for survival in women with advanced epithelial ovarian cancer after primary surgery. *Cochrane Database Syst Rev.* 2021:9. doi: 10.1002/14651858.CD015048.
2. Bryant A, Johnson E, Grayling M, et al. Residual disease threshold after primary surgical treatment for advanced epithelial ovarian cancer (EOC): a network meta-analysis. *BMC Syst Rev.* 2022. [submitted].
3. Colombo N, Van Gorp T, Parma G, et al. Ovarian cancer. *Crit Rev Oncol/Hematol*. 2006;60:159–179.
4. Vergote I, De Wever I, Tjalma W, et al. Neoadjuvant chemotherapy or primary debulking surgery in advanced ovarian carcinoma: a retrospective analysis of 285 patients. *Gynecol Oncol.* 1998;71:431–436.
5. Vergote I, Trimbos BJ. Treatment of patients with early epithelial ovarian cancer. *Curr Opin Oncol.* 2003;15:452–455.
6. Marks-Anglin A, Chen Y. A historical review of publication bias. *Res Syn Meth*. 2020;11:725–742.
7. Stuart GC, Kitchener H, Bacon M, et al. Gynecologic cancer InterGroup (GCIG) consensus statement on clinical trials in ovarian cancer: report from the fourth ovarian cancer consensus conference. *Int J Gynecol Cancer*. 2011;21:750–755.
8. Deeks J, Higgins JPT, Altman DG. Chapter 10: analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 6.3 (Updated February 2022)*. Cochrane; 2022.
9. Ropovik I, Adamkovic M, Greger D. Neglect of publication bias compromises meta-analyses of educational research. *PLoS ONE*. 2021;16:e0252415.
10. Page MJ, Higgins JPT, Sterne JAC. Chapter 13: assessing risk of bias due to missing results in a synthesis. In: Higgins JPT, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 6.3 (Updated February 2022)*. Cochrane; 2022.
11. Spiegelhalter DJ. Incorporating bayesian ideas into health-care evaluation. *Stat Sci.* 2004;19:156–174.
12. Dwan K, Gamble C, Williamson PR, et al. Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS One*. 2013;8:e66844.
13. Iglesias CP, Thompson A, Rogowski WH, et al. Reporting guidelines for the use of expert judgement in model-based economic evaluations. *Pharmacoeconomics*. 2016;34:1161–1172.
14. Bojke L, et al. *Developing a Reference Protocol for Expert Elicitation in Healthcare Decision Making*. Health Technology Assessment Reports; 2019. [In Press].
15. Morgan MG. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proc Natl Acad Sci USA*. 2014;111:7176–84.
16. Use Committee for Medicinal Products for Human Use (CHMP). In: Agency EM, ed. *Guideline on Adjustment for Baseline Covariates in Clinical Trials*; 2015.
17. Altman DG. *Covariate Imbalance, Adjustment for 2005*; 2005. https://doi.org/10.1002/0470011815.b2a01015.
18. Reeves BC, Deeks JJ, Higgins JPT, et al. Chapter 24: including non-randomized studies on intervention effects. In: Higgins JPT, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 62 (Updated February 2021)*; 2021.
19. McKenzie JE, Brennan SE, Ryan RE, et al. Chapter 3: defining the criteria for including studies and how they will be grouped for the synthesis. In: Higgins JPT, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 62 (Updated February 2021)*; 2021.
20. Higgins JPT, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions Version 6.0 Cochrane*; 2019. (updated July 2019).
21. Higgins JPT, Welton NJ. Network meta-analysis: a norm for comparative effectiveness? *Lancet*. 2015;386:628–630.
22. Mavridis D, Welton NJ, Sutton A, et al. A selection model for accounting for publication bias in a full network meta-analysis. *Stat Med.* 2014;33:5399–5412.
23. Chootrakool H, Shi JQ, Yue R. Meta-analysis and sensitivity analysis for multi-arm trials with selection bias. *Stat Med.* 2011;30:1183–1198.
24. Mavridis D, Sutton A, Cipriani A, et al. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Stat Med.* 2013;32:51–66.
25. Copas J, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*. 2000;1:247–262.
26. Copas JB. What works? Selectivity models and meta-analysis. *J Roy Stat Soc (Series A)*. 1999;162:95–109.
27. Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Stat Meth Med Res.* 2001;10:251–265.
28. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (with discussion). *J Roy Statist Soc Ser A.* 1994;157:357–416.
29. Spiegelhalter DJ, Myles J, Jones D, et al. Bayesian methods in health technology assessment: a review. *Health Technol Assess Rep.* 2000;4:1–130.
30. Ahmed I, Riley RD. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ*. 2012;344:d7762.
31. Bryant A, Grayling M, Hiu S, et al. Residual disease after primary surgery for advanced epithelial ovarian cancer: expert elicitation exercise to explore opinions about potential impact of publication bias in a systematic review and meta-analysis. *BMJ Open*. 2022;0:e060183. doi: 10.1136/bmjopen-2021-060183.
32. (OCRA) Ocra. What is the survival rate for Stage 3 ovarian cancer? Available at: https://ocrahope.org/patients/about-ovarian-cancer/staging/#:~:text=Most%20women%20diagnosed%20with%20Stage%20III%20ovarian%20cancer%20have%20a,survival%20rate%20of%20approximately%2039%25. Accessed June 14, 2022.
33. American Cancer Society. *Survival Rates for Ovarian Cancer*; 2022.

34. Siegel RL, Miller KD, Jemal A. Cancer statistics. *CA A Cancer J Clin.* 2020;70:7–30.
35. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, UK: John Wiley & Sons; 2004.
36. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res.* 2001;10:277–303.
37. Sutton AJ, Abrams KR, Jones DR, et al. *Methods for Meta-Analysis in Medical Research*. Chichester, UK: John Wiley & Sons; 2000.
38. Wilson ECF, Usher-Smith JA, Emery J, et al. Expert elicitation of multinomial probabilities for decision-analytic modeling: an application to rates of disease progression in undiagnosed and untreated melanoma. *Value in Health*. 2018;21:669–676.
39. Lunn DJ, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput.* 2000;10:325–337.
40. Dias S, Welton NJ, Sutton AJ, et al. *NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials*. NICE Decision Support Unit; 2016.
41. Chaimani A, Higgins JPT, Mavridis D, et al. Graphical tools for network meta-analysis in STATA. *PLoS ONE*. 2013;8:e76654.
42. Dias S, Sutton AJ, Caldwell DM, et al. *NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based on Randomised Controlled Trials*. National Institute for Health and Care Excellence (NICE); 2014.
43. Higgins JP, Barrett JK, Lu G, et al. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods.* 2012;3:98–110.
44. Dias S, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med.* 2010;29:932–944.
45. Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev.* 2017;6:79.
46. Mueller KF, Meerpohl JJ, Briel M, et al. Methods for detecting, quantifying, and adjusting for dissemination bias in meta-analysis are described. *Rev Environ Econ Pol.* 2016;80:25–33.
47. Ross A, Cooper C, Gray H, et al. Assessment of publication bias and systematic review findings in top-ranked otolaryngology journals. *JAMA Otolaryngol Head Neck Surg.* 2019;145:187–188.
48. Zondervan-Zwijnenburg M, Peeters M, Depaoli S, et al. Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Res Hum Develop.* 2017;20:305–320.
49. Quick H, Huynh T, Ramachandran G. A method for constructing informative priors for bayesian modeling of occupational hygiene data. *Ann Work Exposures Health.* 2017;61:67–75.
50. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
51. Page MJ, Sterne JAC, Higgins JPT, et al. Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: a review. *Res Synth Methods*. 2021;12:248–259.
52. McShane BB, Bockenholt U, Hansen KT. Adjusting for publication bias in meta-analysis: an evaluation of selection methods and some cautionary notes. *Perspect Psychol Sci.* 2016;11:730–749.
53. Gupta DM, Boland RJ, Aron DC. The physician's experience of changing clinical practice: a struggle to unlearn. *Implement Sci.* 2017;12:28.
54. Kristensen N, Nymann C, Konradsen H. Implementing research results in clinical practice- the experiences of healthcare professionals. *BMC Health Serv Res.* 2016;16:48.