Risk scores to guide referral decisions for people with suspected ovarian cancer in secondary care: a systematic review and cost-effectiveness analysis

Marie Westwood,[1*] Bram Ramaekers,[2] Shona Lang,[1] Sabine Grimm,[2] Sohan Deshpande,[1] Shelley de Kock,[1] Nigel Armstrong,[1] Manuela Joore[2] and Jos Kleijnen[3]

[1]Kleijnen Systematic Reviews Ltd, York, UK
[2]Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre and CAPHRI, School for Public Health and Primary Care, Department of Health Services Research, Maastricht University, the Netherlands
[3]School for Public Health and Primary Care, Care and Public Health Research Institute (CAPHRI), Maastricht University, the Netherlands

*Corresponding author marie@systematic-reviews.com

Declared competing interests of authors: none

Published August 2018
DOI: 10.3310/hta22440

Scientific summary

Referral decisions for people with suspected ovarian cancer in secondary care

Health Technology Assessment 2018; Vol. 22: No. 44
DOI: 10.3310/hta22440

NIHR Journals Library www.journalslibrary.nihr.ac.uk

Scientific summary

Background

Current guidance [National Collaborating Centre for Cancer. *Ovarian Cancer: The Recognition and Initial Management of Ovarian Cancer*. Clinical guideline (CG122). Manchester: National Institute for Health and Care Excellence; 2011] recommends that serum cancer antigen 125 (CA125) levels should be measured in secondary care, in all people with suspected ovarian cancer for whom serum CA125 levels have not already been measured in primary care. CG122 specifically recommends the calculation of a Risk Malignancy Index 1 (RMI 1) score, which includes CA125 levels, morphological features seen on ultrasound and menopausal status, with referral to a specialist multidisciplinary team (SMDT) for people with a RMI 1 score of ≥ 250. An evaluation of current evidence is needed to assess the clinical utility and cost-effectiveness of alternative methods of risk-scoring.

Objectives

The overall objective of this assessment was to summarise the evidence on the clinical effectiveness and cost-effectiveness of using alternative risk scores that include CA125 levels, human epididymis protein 4 (HE4) levels or morphological features seen on ultrasound {Risk of Ovarian Malignancy Algorithm [ROMA], International Ovarian Tumour Analysis [IOTA] group's simple ultrasound rules the IOTA group's Assessment of Different NEoplasias in the adneXa [ADNEX] model, Overa [multivariate index assay, second generation (MIA2G)], and RMI 1 at thresholds other than 250} to guide referral decisions for women with suspected ovarian cancer in secondary care. The following research questions were defined:

- What is the accuracy of alternative risk scores (including alternative RMI 1 score thresholds), which include HE4 and CA125 levels and morphological features seen on ultrasound, compared with the RMI 1 score with a referral threshold of ≥ 250 (current practice), in which the target condition is histologically confirmed ovarian cancer?
- What are the effects of using alternative risk scores (including alternative RMI 1 score thresholds), which include HE4 and CA125 levels and morphological features seen on ultrasound, compared with the RMI 1 score with a referral threshold of ≥ 250 (current practice), on clinical management decisions and clinical outcomes?
- What is the cost-effectiveness of alternative risk scores (including alternative RMI 1 score thresholds), which include HE4 and CA125 levels and morphological features seen on ultrasound, compared with the RMI 1 score with a referral threshold of ≥ 250 (current practice), when routinely used in secondary care to guide decisions about referral to a SMDT for women with suspected ovarian cancer?

Methods

Assessment of clinical effectiveness

Twenty-one databases, including MEDLINE and EMBASE, research registers and conference proceedings, were searched from inception to November 2016. Search results were screened for relevance independently by two reviewers. A full-text inclusion assessment, data extraction and a quality assessment were conducted by one reviewer and checked by a second. Study quality was assessed using the quality assessment of diagnostic accuracy studies 2 (QUADAS-2) tool and PROBAST (Prediction model study Risk Of Bias Assessment Tool). A meta-analysis using weighted averages and random-effects modelling was used to estimate summary sensitivity and specificity with 95% confidence intervals (CIs). Analyses were conducted separately for each assay, threshold and target condition (all malignancy, ovarian cancer and borderline cancer) for which data were available.

Assessment of cost-effectiveness

The base-case analysis included seven risk scores:

1. Risk of Malignancy Index RMI 1 score (at a threshold of 250)
2. Risk of Ovarian Malignancy Algorithm score using Abbott Diagnostics' ARCHITECT CA125 and HE4 assays (Abbott Diagnostics, Abbott Park, IL, USA)
3. Risk of Ovarian Malignancy Algorithm using Roche Diagnostics' Elecsys CA125 and HE4 assays (Roche Diagnostics, Rotkreuz, Switzerland)
4. Overa (MIA2G; Vermillion, Inc., Austin, TX, USA; at a threshold of 5 units)
5. International Ovarian Tumour Analysis Simple Rules (inconclusive, assumed to be malignant)
6. International Ovarian Tumour Analysis ADNEX model (at a threshold of 10%)
7. Risk of Malignancy Index (threshold of 200).

This assessment used the economic model from CG122 as a starting point to develop a de novo model adapted to better fit the scope of the current assessment; consistent with the CG122 model, the population age was assumed to be 40 years.

In the de novo health economic model, the mean expected costs and quality-adjusted life-years (QALYs) were calculated for each alternative risk assessment strategy. These long-term consequences were estimated based on the accuracy of the different strategies to detect ovarian cancer, followed by referral to a SMDT and treatment in tertiary care, or no tertiary referral. It was also taken into account that a small proportion of patients with pelvic masses are diagnosed with colorectal cancer (consistent with CG122).

A decision tree and a Markov model were developed. The decision tree was used to model the short-term outcomes. It was assumed that patients who are found to have a high risk of malignancy [i.e. who receive a high-risk test result (either true or false positive)] are referred to a SMDT, and patients who receive a low-risk test result (either true or false negative) are not referred to a SMDT.

Results

Assessment of clinical effectiveness

Fifty-one diagnostic cohort studies (65 publications and one unpublished interim report) were included in the systematic review. Sixteen studies were identified for the ROMA score, 18 for the IOTA group's simple ultrasound rules, seven for the IOTA group's ADNEX model, three for Overa (MIA2G) and 10 for different thresholds of the RMI 1; some studies assessed more than one risk score. The main potential sources of bias in the included studies related to patient flow (not all patients were included in the analysis) and the applicability of the index text (test performed before referral, retrospective application of variables, use of experienced ultrasound practitioners and risk score-specific pre-study training).

The ROMA score, using the Abbott Diagnostics' ARCHITECT or Roche Diagnostics' Elecsys tumour marker assays, did not offer any clear performance advantage over the RMI 1. The only ROMA score study ($n = 213$ participants) using the Abbott Diagnostics ARCHITECT assay, which included all participants in the analysis, reported similar sensitivity and specificity estimates for the ROMA score and the RMI 1 at a decision threshold of 200, 75% (95% CI 60.4% to 86.4%) versus 77.1% (95% CI 62.7% to 88.0%), and 87.9% (95% CI 81.9% to 92.4%) versus 81.8% (95% CI 75.1% to 87.4%), respectively. By contrast, when participants with borderline tumours and/or those with malignancies other than epithelial ovarian cancer were excluded from the analyses (two studies, $n = 1172$ participants), the summary specificity estimate for the ROMA score (53.3%, 95% CI 50.0% to 56.7%) was significantly lower than that for the RMI 1 score at a decision threshold of 200 (80.3%, 95% CI 77.5% to 82.9%), and the summary sensitivity estimates were similar and higher, at 96.4% (95% CI 93.6% to 98.2%) and 93.4% (95% CI 90.0% to 95.9%). The only study to report a direct comparison of the ROMA score, using Roche Diagnostics' Elecsys tumour marker assays and the RMI 1 score at a decision threshold of 200, included all study participants in the analysis, irrespective of final histological diagnosis, but classified participants with borderline tumours as disease negative. In this study, the sensitivity estimate for the ROMA score appeared to be slightly higher than that for the RMI 1 score, at 83.8% (95% CI 73.4% to 91.3%) versus 78.4% (95% CI 67.3% to 87.1%), respectively, and the specificity estimate for the ROMA score appeared to be slightly lower than that for the RMI 1 score, at 68.8% (95% CI 61.6% to 75.4%) versus 79.6% (95% CI 73.1% to 85.1%), respectively, but neither difference was statistically significant. The summary estimates of sensitivity and specificity for the ROMA score, using Roche Diagnostics' Elecsys tumour marker assays at the manufacturer's recommended thresholds, were derived from non-comparative accuracy studies in which all participants were included in the analysis (with the target condition being all malignancy) were 79.1% (95% CI 74.2% to 83.5%) and 79.1% (95% CI 76.3% to 81.6%), respectively (two studies, $n = 1252$ participants). In studies in which the manufacturer's recommended cut-off points were used, the performance of the ROMA score did not differ significantly between premenopausal women and postmenopausal women. Limited data indicated that patients with borderline tumours and those with non-ovarian primaries accounted for disproportionately high numbers of those with false-negative, low-risk ROMA scores. There were no studies evaluating the ROMA score using CA125 and HE4 assays on the Fujirebio Diagnostics' LUMIPULSE® G automated chemiluminescent enzyme immunoassay system (Fujirebio Diagnostics, Gothenburg, Sweden).

The summary estimates of sensitivity, derived from direct comparison studies that included all study participants in their analyses [two studies, $n =$ (confidential information has been removed)], were significantly higher for both the ADNEX model, at 96% (95% CI 94.5% to 97.1%), and IOTA group's simple ultrasound rules, at 92.8% (95% CI 90.9% to 94.3%), than for the RMI 1 score at a decision threshold of 200: 66% (95% CI 62.9% to 69%) (confidential information has been removed). Conversely, the summary estimates of specificity, for both the ADNEX model, at 67% (95% CI 64.2% to 69.6%), and the IOTA group's simple ultrasound rules, at 71.6% (95% CI 68.9% to 74.1%), were significantly lower than those for the RMI 1 score at a decision threshold of 200: 89% (95% CI 87% to 90.7%) (confidential information has been removed). In order to achieve similar levels of sensitivity to those provided by the ADNEX model and the IOTA group's simple ultrasound rules, a very low RMI 1 score decision threshold (25) would be needed; the summary sensitivity and specificity estimates for the RMI 1 score at this threshold were 94.9% (95% CI 91.5% to 97.2%) and 51.1 (95% CI 47.0% and 55.2%), respectively.

No studies were identified that directly compared Overa (MIA2G) to the RMI 1.

Studies evaluating the RMI 1 score at different thresholds indicated no significant difference in performance between thresholds of 200 and 250.

Assessment of cost-effectiveness

In the base-case analysis, the RMI 1 with a threshold of 250 was the least effective [16.926 life-years (LYs), 13.820 QALYs] and the second cheapest (£5669). The IOTA group's simple ultrasound rules (inconclusive, assumed to be malignant), was the cheapest (£5667) and the second most effective (16.954 LYs, 13.841 QALYs), and thereby dominated the RMI 1 (at both the 200 and 250 thresholds). The IOTA group's ADNEX model (threshold of 10%), with a cost of £5699, was the most effective (16.957 LYs, 13.843 QALYs), and compared with the IOTA group's simple ultrasound rules, resulted in an incremental cost-effectiveness ratio of £15,304 per QALY gained. The remaining risk scores [ROMA using Abbott Diagnostics' ARCHITECT, ROMA using Roche Diagnostics' Elecsys and Overa (MIA2G) by Vermillion] were dominated. As a result, the incremental analysis indicated that, up to thresholds of £15,304 per QALY gained, the IOTA group's simple ultrasound rules are cost-effective, whereas the IOTA group's ADNEX model (threshold of 10%) is cost-effective for higher thresholds. Consequently, at willingness-to-pay thresholds of both £20,000 and £30,000 per QALY, the RMI 1, at a threshold of 250, had a probability of being cost-effective of 1%. For the IOTA group's simple ultrasound rules and the IOTA group's ADNEX model (threshold of 10%), this was 39% and 60%, respectively (at the £20,000 threshold), and 23% and 75%, respectively (at the £30,000 threshold). The probabilities for the other risk scores were $< 1\%$ for these thresholds.

The sensitivity and scenario analyses indicated that the hazard ratio for SMDT referral versus no SMDT referral (for patients with ovarian cancer) was the most influential parameter in the model, and that the results were reasonably robust. Most scenario analyses indicated that at thresholds of £20,000 and £30,000 per QALY gained, the IOTA group's ADNEX model (threshold of 10%) remained the cost-effective strategy. In two scenario analyses, the IOTA group's simple ultrasound rules (inconclusive, assumed to be malignant) was considered to be cost-effective at a threshold of £20,000 and/or £30,000 per QALY gained. For the scenario comparing the optimal sensitivity RMI 1 threshold, which was found to be 25 (at all thresholds of £2890 per QALY gained or higher), the RMI 1 was still dominated.

For the premenopausal and postmenopausal subgroups, the IOTA group's ADNEX model (threshold of 10%) remained cost-effective at thresholds of £20,000 and £30,000 per QALY gained.

Conclusions

Implications for service provision

There is evidence to suggest that using either the ADNEX model or the IOTA group's simple ultrasound rules to assess the risk of malignancy in women with an adnexal mass may offer increased sensitivity relative to current practice (the RMI 1 at a decision threshold of 250 or 200); that is, a higher proportion of those women who have a malignant tumour would be referred to a SMDT. A similar sensitivity could be achieved with the RMI 1 by using a very low decision threshold (25); however, this is associated with a lower specificity and a greater number of unnecessary referrals than those achievable using either the ADNEX model or the IOTA group's simple ultrasound rules. The limited available evidence suggests that the ROMA score does not offer any clear performance advantage over the RMI 1. Although Overa (MIA2G) appears to have higher sensitivity than the ROMA score, there are no data to support a direct comparison between Overa (MIA2G) and the RMI 1.

Overall, the cost-effectiveness model provides evidence to strongly prioritise sensitivity over specificity. As a result, the IOTA group's ADNEX model (threshold of 10%), which had the highest sensitivity (96.3%), was considered to be cost-effective.

Suggested research priorities

Further studies or analyses of the IOTA data set are needed to understand the role of menopausal status and other potentially relevant factors, such as family history of ovarian cancer, in the performance of both the IOTA and the ADNEX tests. Large diagnostic cohort studies are needed to fully evaluate the performance of the ROMA score (using different manufacturers' tumour marker assays) and of Overa (MIA2G), compared with the RMI 1, at a decision threshold of 250 or 200. These studies should be conducted in a population that includes the full spectrum of differential diagnoses likely to be present in those referred to secondary care for the investigation of an adnexal mass. Further studies are required to explore the distribution of histological diagnoses among patients with false-negative, low-risk classifications. A more complete exploration of the types of patients who are likely to be misclassified as being at a low risk of having ovarian cancer using the various risk-scoring options available, as well as an investigation of the downstream clinical consequences for these patients, is required.

Study registration

This study is registered as PROSPERO CRD42016053326.

Funding

Funding for this study was provided by the Health Technology Assessment programme of the National Institute for Health Research.

Health Technology Assessment

HTA/HTA TAR

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Impact factor: 4.513

Health Technology Assessment is indexed in MEDLINE, CINAHL, EMBASE, The Cochrane Library and the Clarivate Analytics Science Citation Index.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Technology Assessment* journal

Reports are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

The HTA programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

For more information about the HTA programme please visit the website: http://www.nets.nihr.ac.uk/programmes/hta

This report

The research reported in this issue of the journal was funded by the HTA programme as project number 16/30/02. The contractual start date was in December 2016. The draft report began editorial review in May 2017 and was accepted for publication in October 2017. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health and Social Care.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).