Protocol for a scoping review to identify all available NMA-DTA models

Areti Angeliki Veroniki[1,2,3,*], Sofia Tsokani[1], Gerta Rücker[4], Yemisi Takwoingi[5,6], Dimitris Mavridis[1,7]

Author Details:

[1] Department of Primary Education, School of Education, University of Ioannina, Ioannina, Greece

[2] Knowledge Translation Program, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, Ontario, Canada

[3] Institute of Reproductive and Developmental Biology, Department of Surgery & Cancer, Faculty of Medicine, Imperial College, London, United Kingdom

[4] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Stefan-Meier-Strasse 26, 79104 Freiburg, Germany

[5] Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, UK

[6] NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK

[7] Paris Descartes University, Sorbonne Paris Cité, Faculté de Médecine, Paris, France

***Corresponding author:**

Areti Angeliki Veroniki, MSc, PhD

Research Fellow at the Department of Primary Education, School of Education, University of Ioannina, Ioannina, Greece

Phone: +30 26510 04324; fax: +30 26510 05854; e-mail: averoniki@uoi.gr

Background

The diagnosis of the clinical condition of a patient is usually the first and more crucial step before initiating treatment. Diagnostic tests are routinely used in healthcare settings for confirming or excluding a target condition. However, diagnostic tests are rarely 100% accurate. Evaluation of diagnostic tests to identify the most accurate test(s) for a particular condition contributes to the prevention of unjustified treatment, as well as unnecessary healthcare costs and risks to patients. For some conditions, such as cancer, diagnostic test accuracy (DTA) studies aim to identify a new diagnostic test that is as or more accurate than the current standard of care, yet less costly and/or less invasive.

Rarely are clinical and policy decisions made on the basis of the results of a single DTA study but rather on evidence from multiple DTA studies addressing the same research question. Sensitivity (i.e., the probability of the test being positive when the target condition is present) and specificity (i.e. the probability of the test being negative when the target condition is absent) are measures commonly used to express the accuracy of a test.

Several statistical models have been proposed for the synthesis of the sensitivity and specificity of diagnostic tests across studies. One approach is the Moses-Littenberg model[1], which fits a straight regression line to variables derived from the logits of sensitivity and the false positive rate (1−specificity) of each study. However, this method has many statistical flaws and is not recommended[2,3]. Two hierarchical models have been proposed as alternative statistical methods: the hierarchical summary receiver operating characteristic (HSROC) model and the bivariate random-effects model[2,3]. Both models take into account the potential correlation of sensitivity and specificity across studies, and they are mathematically equivalent when there are no covariates in the models[4].

To summarise and compare the accuracy of two or more index tests, one of the following two approaches is usually considered: 1) separate meta-analysis for each index test and formal comparison of the pooled results using statistical tests (e.g. z-test) or informally by comparing confidence intervals; or 2) meta-regression using test type as a categorical variable in a hierarchical model[5-9]. The former is generally considered inappropriate because the tests have not been compared in the same model and the latter is the commonly recommended approach. Although, these hierarchical meta-regression models can accommodate studies of a single test as well as comparative accuracy studies, Takwoingi et al. showed, empirically, that meta-analytic estimates from comparative studies, where two diagnostic tests are directly compared, may differ from those estimated from non-comparative studies[10]. Comparative accuracy studies may randomize participants to an index test or participants may receive all index tests (often referred to as paired studies). The hierarchical meta-regression approach typically assumes independence between the index tests evaluated in a comparative study regardless of the study design. However, when the same individuals receive all index tests, as in most comparative DTA studies, sensitivities (and similarly specificities) among diagnostic tests are likely to be correlated and should be taken into account to avoid misleading findings.

To date, most DTA studies have focused on the accuracy of a single index test, but in many cases, as in the diagnosis of cervical cancer, it is necessary to compare the accuracy of at least 3 tests. For the cervical cancer diagnosis, the tests of HPV DNA, HPV mRNA, and co-testing (Pap test + HPV DNA or mRNA test) can be used. The key question is: Which test is the best? Although direct DTA comparisons (head-to-head comparison) offer the most valid design, they are not always available. The accuracy of different tests can be compared indirectly through a common comparator test. When evaluating at least three healthcare interventions, the network meta-analysis (NMA), combining both direct and indirect evidence[11], has been introduced as an extension of pairwise meta-analysis.

Since 2014 several models, including NMA models, have been proposed for the comparative meta-analysis of multiple tests[12-17], but further research is needed to establish the benefits and properties of these models. It should be noted that the NMA methods developed for the comparison of multiple interventions

cannot be directly applied to the meta-analysis of multiple diagnostic tests, due to differences (e.g., in the design) between DTA studies and intervention studies. A major difference is that NMA-DTA models jointly synthesise sensitivity and specificity, whereas the NMA of multiple interventions considers a single effect measure (e.g., relative risk). Within a study a pair of sensitivity and specificity are independent (they refer to different populations, diseased and non-diseased), whereas across studies they are potentially correlated due to differences in implicit or explicit thresholds, and thus specialized (i.e. bivariate) methodology is needed. Another difference is that intervention studies usually compare different patient groups (patients are usually randomly allocated to the intervention groups), whereas diagnostic tests are often evaluated in the same individuals within a study. Hence, the appropriate use of NMA-DTA methodology will directly impact the validity of research, advance knowledge translation activities and facilitate complex policy decisions, with an aim to prevent unjustified treatment, unnecessary healthcare costs, and risks to patients.

The validity of NMA-DTA models has not yet been examined. We will review existing meta-analytic methods for comparing the accuracy of multiple diagnostic tests. To date, no study has evaluated the accuracy of multiple tests for the diagnosis of cervical cancer in a single model, which would permit the identification of the best strategy to diagnose cervical cancer. The hierarchy of these diagnostic tests according to their accuracy in terms of sensitivity and specificity, and identification of the best available test to detect cervical cancer will help avoid unnecessary screening, colposcopy, and treatment (e.g., surgery) associated with undesirable effects, such as preterm births and miscarriages at 2nd trimester[18]. Therefore, following our evaluation of NMA-DTA methods, we will apply a valid approach to synthesise and compare the tests.

Objectives

1) Identify gaps and methodological deficiencies in the existing literature with regard to the statistical methods developed for conducting an NMA-DTA. We aim to:
    a. conduct a scoping review of the statistical methods developed for meta-analysis comparing the accuracy of at least 3 diagnostic tests.
    b. collect all available NMA-DTA methods and identify potential methodological deficiencies and strengths
    c. present a theoretical framework and specific steps of NMA-DTA methods
2) Apply identified methods to determine the test with a high specificity and sensible sensitivity (to be used as a triage of the primary screening test and identify which women will need referral colposcopy), or determine the tests with a trade-off between sensitivity and specificity that could be used as a single primary screening test without need for triage for the diagnosis of invasive cervical cancer (CIN2+). The purpose is:
    a. synthesise all available evidence on cervical cancer tests in an NMA-DTA
    b. rank tests from 'most-accurate' to 'least accurate' and according to their sensitivity and specificity, and identify the best test for the diagnosis of cervical cancer
    c. empirically assess the properties of DTA-NMA models identified in objective 1 using data from a collection of comparative DTA reviews identified from a previous empirical study.

Methods

Scoping review process (stage 1): To facilitate the identification of gaps and methodological deficiencies in the existing literature, we will conduct a scoping review of statistical methods for comparative meta-analysis of at least 3 index tests.

To address *objective 1* we will use the following process:

- **Study inclusion criteria:** We will include all studies that apply, describe, or evaluate an NMA of multiple diagnostic tests. Any study design will be eligible. Only English language studies will be eligible. There will be no restriction on time of publication.

- **Search strategy and literature searches:** We will search PubMed, JSTOR, and Web of Science for relevant publications. A search of the 'grey' literature, including unpublished material and conference abstracts through Google will be also conducted. The search strategy will include the terms: diagnostic tests, SROC method, sensitivity, specificity, diagnostic test accuracy, diagnostic studies, which will be combined using the AND operator with the terms: network meta-analysis, multiple treatment meta-analysis, mixed treatment comparison, indirect comparison, meta-analysis, multivariate analysis.

- **Study selection:** We will use the *abstrackr* tool (http://abstrackr.cebm.brown.edu/account/login) to import the search results and screen citations. We will conduct a training exercise prior to commencing screening. All reviewers will screen a random sample of 10% of citations from the search. Two reviewers will screen titles and abstracts for inclusion (Level 1) and full-texts (Level 2), independently. Conflicts will be resolved by discussion or involvement of a 3rd reviewer. We will scan references from included studies. Team members will also use their extensive networks with experts in the field to identify further articles, dissertations and ongoing research.

- **Data abstraction**: Abstracted data will include study characteristics (e.g. publication year, study design) and characteristics of the corresponding methods (e.g. description of methods, advantages/disadvantages as described by the authors). The data abstraction form will be piloted on a random sample of 10% of included articles and modified as required. To ensure accuracy, 2 reviewers will independently abstract all data using a predefined Excel form; discrepancies will be resolved by discussion or a 3rd reviewer.

- **Data charting and collation:** Data analysis will involve quantitative (e.g. frequency analysis) and qualitative (i.e. content analysis) methods. We will collect information on the NMA-DTA methods and will extract the specific steps for each method. The statistical properties of the identified models will be discussed in the team.

Application to empirical data (stage 2):

Multiple DTA studies have been conducted to identify the best screening strategy for cervical cancer[18,19]. Cervical cancer is the 4th most frequently occurring cancer in women around the world and can affect them during their reproductive years[20]. Since the development of the Papanicolaou (Pap) test, screening has been essential in identifying cervical cancer at a treatable stage. With the identification of the human papillomavirus (HPV) as the causative agent of essentially all cervical cancer cases, HPV molecular screening tests and HPV vaccines for primary prevention against the virus have been developed. In the majority of women, HPV infection will be cleared by the immune system. However, when the immune system does not eliminate the virus, persistent HPV infection can cause abnormal cervical cells. These lesions are known as cervical 'precancer' because over time they can progress to cervical cancer if left untreated[19]. Cervical cancer is a serious public health problem worldwide, including Greece. In 2018, approximately 311 000 women died from cervical cancer; >85% of these deaths occurring in low- and middle-income countries[20]. In Greece, 300 new cases of cervical cancer are estimated to occur per year.

We will use results from stage 1 to complete an NMA-DTA for the diagnosis of cervical cancer. The NMA methodology allows for ranking of all available diagnostic tests according to their sensitivity and specificity, even when some tests have not been directly compared in a head-to-head accuracy study.

To address *objective 2* we will use the following process:

- **Obtain cervical cancer data, prepare and format data for analysis**: We will collect all available data from DTA studies comparing the accuracy of different tests for the diagnosis of cervical

cancer. We will search for systematic review and meta-analyses on this topic in order to identify relevant studies; if we do not find any recent systematic review, we will perform a new search. Once the data are collected, we will obtain the 2x4 tables which cross classifies pairs of index tests in the diseased and non-diseased groups. Where these are not available then a 2x2 table for each index test (number of true positives, true negatives, false positives, and false negatives) will be obtained. We will attempt to obtain missing data by contacting the study authors.

- **Build network geometry and conduct data analysis:** In this task we will model the cervical cancer data using the models identified in objective 1. This task consists of translating the data into the relevant software (e.g., Stata, R, WinBUGS), programming the model including all relevant assumptions regarding the statistical distribution of the model parameters and assessing model fit. In particular, we will:
  - Determine the tests that will be used in the network
  - Program the model in the relevant statistical software (e.g., Stata, R, WinBUGS)
  - Synthesise direct and indirect evidence using NMA-DTA models
- **Model assessment and presentation of cervical cancer results:** The collected data will be used to identify the most accurate tests for clinical practice and health-care decision making. The NMA results for all pairwise contrasts will be visually compared using forest plots and the degree of the overlap of the confidence intervals (CI) will be assessed. We will also present a test hierarchy according to their accuracy, between-study heterogeneity, and inconsistency across test comparisons in the included models. In particular, we will:
  - Present results from all NMA-DTA models, the relative accuracy between all pairs of tests and a hierarchy of the tests using ranking statistics and a rank-heat plot[21-23]
  - Assess prerequisite NMA assumptions, including network homogeneity, consistency, and transitivity
  - Explore heterogeneity through study and patient characteristics
  - Explore distribution of potential effect modifiers across test comparisons. We will liaise with clinicians and experts in cervical cancer to assess transitivity and will apply statistical methods to compare direct and indirect evidence (i.e. consistency)
  - Assess properties of models empirically through the cervical cancer data
- **Interpretation of results:**
  - Interpret results according to model assumptions
  - Repeat analyses for different pre-planned subgroup/meta-regression analyses, where data are available
- **Manuscript preparation for publication**
  - Describe models and empirical results for cervical cancer in a manuscript
  - Submit manuscript for publication in an open-access peer-reviewed journal

Discussion

Diagnostic tests have been around for a long time, and as our understanding of biology and disease increases, along with advances in technology, new tests emerge. The plethora of DTA studies has led to the use of meta-analysis, where health professionals seek to determine the accuracy of available diagnostic tools. Decision making involves selecting among multiple testing strategies; therefore, studies that compare several test strategies and estimate differences in sensitivity and specificity are more informative than those that evaluate the accuracy of a single index test. Several organizations commission NMAs, such as the National Institute for Health and Clinical Excellence (NICE) in UK, the Institute for Quality and Efficiency in Healthcare (IQWiG) in Germany, and the Canadian Agency for Drugs and Technologies in Health (CADTH), because they allow decision-makers to identify the most effective and safe intervention across various alternatives[33-38]. The comparison of multiple diagnostic test strategies using NMA-DTA can impact clinical decision-making and patient health. Additional research studies are required to establish and

disseminate NMA-DTA methods, so that national and international organizations (e.g., NICE, IQWiG, CADTH, WHO [World Health Organization]) can base health care decision making on reliable results.

References

1. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in medicine.* Jul 30 1993;12(14):1293-1316.
2. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in medicine.* Oct 15 2001;20(19):2865-2884.
3. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology.* Oct 2005;58(10):982-990.
4. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics.* Apr 2007;8(2):239-251.
5. Roberts E, Ludman AJ, Dworzynski K, et al. The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting. *Bmj.* Mar 4 2015;350:h910.
6. Patrick DL, Cheadle A, Thompson DC, Diehr P, Koepsell T, Kinne S. The validity of self-reported smoking: a review and meta-analysis. *American journal of public health.* Jul 1994;84(7):1086-1093.
7. Doebler P, Holling H, Bohning D. A mixed model approach to meta-analysis of diagnostic studies with binary test outcome. *Psychological methods.* Sep 2012;17(3):418-436.
8. Arbyn M, Sankaranarayanan R, Muwonge R, et al. Pooled analysis of the accuracy of five cervical cancer screening tests assessed in eleven studies in Africa and India. *International journal of cancer.* Jul 1 2008;123(1):153-160.
9. Rücker G. Network Meta-Analysis of Diagnostic Test Accuracy Studies. In: G. B-Z, ed. *Diagnostic Meta-Analysis*. Cham: Springer; 2018.
10. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Annals of internal medicine.* Apr 2 2013;158(7):544-554.
11. Li T, Puhan MA, Vedula SS, Singh S, Dickersin K, Ad Hoc Network Meta-analysis Methods Meeting Working G. Network meta-analysis-highly attractive but more methodological research is needed. *BMC medicine.* Jun 27 2011;9:79.
12. Dimou NL, Adam M, Bagos PG. A multivariate method for meta-analysis and comparison of diagnostic tests. *Statistics in medicine.* Sep 10 2016;35(20):3509-3523.
13. Nyaga VN, Aerts M, Arbyn M. ANOVA model for network meta-analysis of diagnostic test accuracy data. *Statistical methods in medical research.* Jun 2018;27(6):1766-1784.
14. Hoyer A, Kuss O. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach. *Statistical methods in medical research.* May 2018;27(5):1410-1421.
15. Menten J, Lesaffre E. A general framework for comparative Bayesian meta-analysis of diagnostic studies. *BMC medical research methodology.* Aug 28 2015;15:70.
16. Nyaga VN, Arbyn M, Aerts M. Beta-binomial analysis of variance model for network meta-analysis of diagnostic test accuracy data. *Statistical methods in medical research.* Aug 2018;27(8):2554-2566.
17. Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. *Research synthesis methods.* Dec 2014;5(4):294-312.

18. Kyrgiou M, Athanasiou A, Paraskevaidi M, et al. Adverse obstetric outcomes after local treatment for cervical preinvasive and early invasive disease according to cone depth: systematic review and meta-analysis. *Bmj.* Jul 28 2016;354:i3633.
19. Kyrgiou M, Mitra A, Arbyn M, et al. Fertility and early pregnancy outcomes after conservative treatment for cervical intraepithelial neoplasia. *The Cochrane database of systematic reviews.* Sep 29 2015(9):CD008478.
20. World Health Organization WHO. Human papillomavirus (HPV) and cervical cancer. 2019.
21. Veroniki AA, Straus SE, Fyraridis A, Tricco AC. The rank-heat plot is a novel way to present the results from a network meta-analysis including multiple outcomes. *Journal of clinical epidemiology.* Aug 2016;76:193-199.
22. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of clinical epidemiology.* Feb 2011;64(2):163-171.
23. Veroniki AA, Straus SE, Rucker G, Tricco AC. Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *Journal of clinical epidemiology.* Aug 2018;100:122-129.