International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

Statistical methods for the analysis of diagnostic test accuracy in clustered data settings - A protocol for a systematic review of methods

INPLASY202580006

doi: 10.37766/inplasy2025.8.0006

Received: 1 August 2025

Published: 2 August 2025

Dümmler, D; Weber, P; Vogel, F; Zapf, A; Rübsamen, N.

Corresponding author:
Daniel Dümmler

daniel.duemmler@uni-muenster.de

Author Affiliation:
Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany.

ADMINISTRATIVE INFORMATION

Support - This work is supported by the Deutsche Forschungsgemeinschaft (DFG), Grant No: 539658720.

Review Stage at time of this submission - Preliminary searches.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202580006

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 2 August 2025 and was last updated on 2 August 2025.

INTRODUCTION

Review question / Objective The primary aim of this systematic methodological review is to systematically identify, categorise, and evaluate available statistical methods for analysing diagnostic test accuracy (DTA) in clustered data settings. We will specifically examine methods addressing temporal and spatial clustering structures, across typical DTA measures, including sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve (AUC). The overarching goal is to support transparent, robust, and context-appropriate statistical analyses in diagnostic medicine.

Rationale DTA studies play a central role in evidence-based diagnostics and regulatory evaluation. While metrics such as sensitivity, specificity, and AUC are well established for evaluating test performance, certain diagnostic tests inherently generate complex clustered data structures, which require caution in estimation and interpretation (1–3). Clustered data arise when multiple, potentially correlated observations are obtained from a single individual: multiple readers interpreting the same diagnostic images, tests conducted at several anatomical sites or lesions, or measurements repeated over time (4). Examples include coronary magnetic resonance angiography for myocardial infarction, where data from multiple vessels within a single patient result in spatial clustering, or the use of smartwatches to detect atrial fibrillation, where repeated measurements over time introduce temporal clustering. These observations are typically positively correlated, and failing to account for these intra-individual dependencies may lead to biased estimates and underestimated variance (4).
Although various statistical methods, including generalized linear mixed models, generalized estimating equations, and Bayesian hierarchical models, have been proposed to address clustering in DTA studies, there remains no comprehensive synthesis or guideline to inform method selection in practice (4–6). Consequently, method choice often relies on ad hoc decisions, potentially leading to inconsistent evidence.

This systematic review aims to address this methodological gap by systematically identifying, evaluating, and comparing available statistical approaches, which will provide a structured evidence base and decision framework for robust, transparent, and context-appropriate analysis of clustered diagnostic accuracy data.

Condition being studied Diagnostic test accuracy in clinical or epidemiological applications involving clustered data, where clustering arises from spatial (e.g., multiple lesions), temporal (repeated measures in longitudinal settings), multi-reader, or multi-center structures.

METHODS

Search strategy The search strategy for this systematic review is being developed with guidance from the Cochrane Handbook for Diagnostic Test Accuracy Reviews, alongside recommendations from PRISMA-S, and PRESS (7–10). Search terms are organised into three main concept categories: core diagnostic terms, diagnostic accuracy metrics, and clustered data settings. Terms within each category are mapped and iteratively refined using terminology identified from marker studies, recent reviews, and expert feedback (11, 12). The strategy is tested using a reference set of known relevant studies, and all adaptations and changes are documented.

Search terms and controlled vocabulary are translated and adapted for each database, e.g. using MeSH terms in PubMed and EMTREE in Embase, and appropriate field tags and operators for each platform. Both free-text and thesaurus terms are used to ensure comprehensive retrieval, with all search syntax and translations fully documented. Where appropriate, proximity operators (such as “adj3”) are incorporated to capture studies using varied or complex terminology for clustered data structures.

Recognising the non-uniform terminology and inconsistent indexing in the literature on statistical methods for diagnostic accuracy research, supplementary identification techniques including citation tracking and hand searching are employed. This approach ensures comprehensive retrieval of relevant methodological studies, including those published in statistical or computational journals or described in applied research with non-standard nomenclature, as specified in the ClusterDiag project plan (12).

No date restrictions will be applied. Only peer-reviewed, full-text publications in English or German will be included. The final search strategy, including database-specific search strings and documentation of the iterative search development process, will be completed prior to screening and data extraction and will be reported in full as supplementary material in the final publication.

Participant or population This review targets studies that describe, apply, evaluate, or develop statistical methods to adjust for clustering in the analysis of diagnostic accuracy data. Eligible studies need to be of diagnostic relevance in clinical or epidemiological contexts, where clustering may occur spatially, temporally, or due to multi-level study designs (e.g. multi-reader, multi-center, or repeated measures).

Intervention We will consider any statistical approach that accounts for such clustering when estimating standard diagnostic accuracy outcomes, including sensitivity, specificity, and the AUC. Several statistical methods have been proposed to address clustered diagnostic accuracy data, broadly categorised as parametric (e.g., generalized linear mixed models- GLMM), semiparametric (e.g., generalized estimating equation GEE), nonparametric approaches, and Bayesian hierarchical models. Each approach has distinct assumptions, strengths, limitations, and data requirements. Moreover, accuracy measures such as sensitivity and specificity or AUC may require different statistical strategies. For instance, while the ROC curve and AUC can summarise overall accuracy independent of a single threshold, sensitivity and specificity must often be considered as co-primary endpoints. Approaches can differ significantly depending on whether accuracy metrics are modelled individually ("two-model approaches") or jointly ("one-model approaches"), with hierarchical models offer a flexible structure between joint and separate modelling of accuracy metrics.

Comparator No specific comparison group. Studies that compare methods are considered eligible if they meet the stated inclusion criteria.

Study designs to be included We will include methodological articles, simulation studies, applied DTA studies, and systematic reviews that investigate statistical adjustment for clustering in DTA data. Any study phase or design is eligible. We include studies from all relevant medical, epidemiological, and clinical contexts without restriction, provided they explicitly address diagnostic accuracy in the presence of clustered data structures.

Eligibility criteria Inclusion: Studies that describe, develop, apply, or evaluate statistical methods specifically designed to handle clustering in DTA data; Reporting of diagnostic accuracy metrics adjusted for clustering, primarily, but not strictly limited to sensitivity, specificity, and AUC; Any methodological study type, including simulation studies, comparative method papers, applied DTA studies with explicit clustering adjustments, and systematic or scoping reviews.

Exclusion: Studies not addressing clustering or without any adjustment for intra-cluster correlation; Studies unrelated to diagnostic accuracy (e.g., pure prognostic or treatment effect analyses); Insufficient methodological description to ascertain assumptions, implementation, or interpretation; Insufficient detail on statistical method; irrelevant to DTA context; no clustering adjustment.

Information sources We will identify relevant studies through searches of major bibliographic databases in biomedical and diagnostic research, including MEDLINE (via PubMed), Web of Science Core Collection or Scopus, and Embase (via Ovid, subject to access). Searches will cover the full publication period from inception to the date of execution. To enhance completeness, we will apply supplementary methods including citation tracking (forward and backward), and reference list screening of included studies and relevant reviews. We will additionally explore relevant registries and curated collections of diagnostic studies, including those cited in systematic reviews or Cochrane resources, where applicable. Where relevant, we will also screen subject-specific repositories and preprint platforms with a focus on statistical methods (e.g., IEEE Xplore, arXiv, bioRxiv), particularly for methodological or simulation-based studies. Only peer-reviewed, full-text publications in English or German will be eligible, but non-English articles with English abstracts will be considered during screening. Grey literature, preprints, and unpublished methods will be excluded.

Main outcome(s) The primary outcome of this review is a structured comparative synthesis of statistical methods used for analysing clustered DTA data. Methods will be described and evaluated across the following domains: type of clustering addressed (e.g., spatial, temporal), statistical approach (e.g., GLMM, GEE, Bayesian), diagnostic metrics modelled (e.g., sensitivity, specificity, AUC), and practical features such as flexibility, robustness in small samples, covariate adjustment, and alignment with estimands. These outcomes will inform the development of a structured comparison table.

Additional outcome(s) We will also examine whether software, code, and model assumptions are made available and transparently reported and which author-stated strengths and limitations, and practical recommendations are stated.

Data management All records identified through database searches and supplementary sources will be imported into a reference management tool (Zotero or Covidence) for deduplication (13, 14). Title and abstract screening will be conducted independently by two reviewers using a standardised screening interface (e.g., Covidence or an equivalent systematic review platform). The screening will follow predefined inclusion and exclusion criteria, and a calibration phase will be conducted prior to screening to ensure consistency between reviewers.

Full texts of potentially eligible articles will be retrieved via institutional access or document delivery services; if necessary, authors will be contacted or interlibrary loan will be used. Two reviewers will independently assess full texts for final inclusion. Disagreements at any screening stage will be resolved through discussion or adjudication by a third reviewer. Reasons for exclusion at the full-text stage will be documented and presented in the PRISMA flow diagram. The reference lists of all included studies and relevant reviews will also be screened to identify additional eligible records.

Data extraction will be conducted independently by at least two reviewers using a standardised, pre-piloted extraction form. Discrepancies will be resolved through discussion or adjudicated by a third reviewer if needed. Extraction will be managed in Covidence or equivalent software. To ensure consistency, completeness, and comparability, extracted data will be organised into predefined domains that reflect both general study descriptors and method-specific features relevant to clustered diagnostic accuracy analysis.

These include study identifiers, study design characteristics, clustering structure, test and reference standard details, statistical method specifications (including assumptions, software, and adjustment strategies), reported diagnostic metrics (e.g., sensitivity, specificity, AUC), estimand alignment, flexibility, and transparency of reporting (e.g., availability of code and data).

All extracted data will be tabulated to support structured qualitative and comparative synthesis. Where appropriate, methods will be grouped across domains to enable cross-method comparisons. Authors of primary studies will be contacted for program codes.

Quality assessment / Risk of bias analysis Given the methodological nature of this review, risk of bias and methodological quality assessment will be aligned with both the primary outcomes and the study type. We will apply established tools in a manner appropriate to the scope of this synthesis: QUADAS-2 will be used for primary diagnostic accuracy studies, AMSTAR 2 for systematic reviews, and a structured custom appraisal will be developed for simulation and methodological studies, drawing on reporting guidelines used by relevant journals and established checklists (e.g., for transparency, reproducibility, validation, and clarity of statistical assumptions) (15, 16). All assessments will be performed independently by two reviewers. Results will be presented narratively and, where appropriate, summarised in structured tables or visual formats (e.g., traffic light plots or summary matrices) (17).

While publication bias is not a primary concern in this review, potential selective reporting will be explored narratively if enough comparable studies are identified. GRADE will not be applied, as the review does not evaluate health interventions; however, we will qualitatively assess the consistency, scope, and methodological robustness of the included evidence to inform the development of the planned decision-support framework.

Strategy of data synthesis The review will employ a primarily narrative synthesis to summarise and compare the statistical methods identified. Extracted data will be grouped according to the type of clustering addressed (e.g., spatial, temporal, hierarchical), the statistical framework used (e.g., parametric, semi-parametric, nonparametric, Bayesian), and the diagnostic metrics modelled (e.g., sensitivity, specificity, AUC). Structured comparison tables will be developed to display key features, assumptions, and practical implications of each method. Where appropriate, descriptive statistics (e.g., frequency of method use or reporting of assumptions) will be used to complement the narrative. The findings can inform a decision-support framework, such as a method selection tree, and enables further simulation studies for formal comparisons, as outlined in the ClusterDiag project plan (12).

Subgroup analysis We will explore whether methodological characteristics vary systematically across relevant subgroups. These include differences by type of clustering structure (such as spatial vs. temporal data), statistical modelling approach (e.g., parametric versus nonparametric or Bayesian), type of diagnostic metric modelled (e.g., sensitivity and specificity vs. AUC), and by the clinical or application domain. These analyses will be conducted descriptively and narratively, in parallel with the structured synthesis tables.

Sensitivity analysis Not applicable.

Language restriction No restrictions during search. Inclusion limited to English or German full-text, or non-English articles with English abstracts enabling screening.

Country(ies) involved Germany.

Other relevant information This protocol follows the PRISMA-P checklist. Any protocol amendments will be documented in an updated INPLASY record and final publication.

Keywords Diagnostic accuracy; clustered data; Statistical methods; Sensitivity; Specificity; Area under the curve; Methods Comparison; Evidence Synthesis.

Dissemination plans Results will be disseminated via peer-reviewed publication and presentations at relevant methodological and diagnostic conferences.

Contributions of each author

Author 1 - Daniel Dümmler.
Email: daniel.duemmler@uni-muenster.de
Author 2 - Philipp Weber.
Email: p.weber@uke.de
Author 3 - Frederike Vogel.
Email: f.vogel@uke.de
Author 4 - Antonia Zapf.
Email: a.zapf@uke.de
Author 5 - Nicole Rübsamen.
Email: nicole.ruebsamen@uni-muenster.de

All authors contributed to the development and refinement of the review protocol, including conceptual design, methodological planning, and drafting. Roles in the review process will be assigned based on expertise and project needs.

References

1. Leeflang MMG. Syst revs & meta-analyses of DTA. Clin Microbiol Infect. 2014;20.
2. Gönen M, et al. Stat issues in diag imaging w/ multiple obs. Radiology. 2001;221.
3. Böhnke J, et al. DTA in longitudinal settings. J Clin Epidemiol. 2024;169:111314.
4. Genders TSS, et al. Sensitivity & specificity for clustered data. Radiology. 2012;265(3):910–6.
5. EMA. Guideline: clinical eval of diagnostic agents. 2010.
6. Zhou, X.-H., et al. Multi-reader/test analysis. In: Med Imaging. 2011;297–328.
7. Lefebvre C, et al. Ch.4: Search & select studies. Cochrane HB. 2024.
8. Spijker R, et al. Ch.6: Search & select studies. Cochrane COVID-19. 2023.

9. Stansfield C, et al. Ch.5: Identifying studies. Cochrane. 2023.
10. Rethlefsen ML, et al. PRISMA-S: search reporting. Syst Rev. 2021;10(1):1–19.
11. Vogel F, et al. Accuracy for clustered data. MEMTAB Conf, Birmingham. 2025.
12. DFG-GEPRIS. ClusterDiag proj. 2025. https://gepris.dfg.de
13. Corp Digital Scholarship. Zotero. http://www.zotero.org
14. Veritas Health Innov. Covidence. http://www.covidence.org
15. Shea BJ, et al. AMSTAR 2: SR appraisal tool. BMJ. 2017;358:j4008.
16. Whiting PF, et al. QUADAS-2. Ann Intern Med. 2011;155(8):529–36.
17. McGuinness LA, Higgins JPT. robvis tool. Res Synth Methods. 2021;12(1):55–61.