Study designs for comparative diagnostic test accuracy

Yang, Bada ; Olsen, Maria ; Vali, Yasaman ; Langendam, Miranda W. ; Takwoingi, Yemisi; Hyde, Christopher; Bossuyt, Patrick M.; Leeflang, Mariska M G

DOI:
10.1016/j.jclinepi.2021.04.013

Citation for published version (Harvard):
Yang, B, Olsen, M, Vali, Y, Langendam, MW, Takwoingi, Y, Hyde, C, Bossuyt, PM & Leeflang, MMG 2021, 'Study designs for comparative diagnostic test accuracy: a methodological review and classification scheme', *Journal of Clinical Epidemiology*, vol. 138, pp. 128-138. https://doi.org/10.1016/j.jclinepi.2021.04.013

Journal of Clinical Epidemiology 138 (2021) 128–138

Journal of Clinical Epidemiology

Study designs for comparative diagnostic test accuracy: A methodological review and classification scheme

Bada Yang[a,*], Maria Olsen[a], Yasaman Vali[a], Miranda W. Langendam[a], Yemisi Takwoingi[b,c], Christopher J. Hyde[d], Patrick M.M. Bossuyt[a], Mariska M.G. Leeflang[a]

[a] *Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands*

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

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

[d] *Exeter Test Group and South West CLAHRC, University of Exeter Medical School, St Luke's Campus, Exeter, UK*

Accepted 19 April 2021; Available online 26 April 2021

Abstract

Objectives: (1) To identify and classify comparative diagnostic test accuracy (DTA) study designs; (2) to describe study design labels used by authors of comparative DTA studies.

Methods: We performed a methodological review of 100 comparative DTA studies published between 2015 and 2017, randomly sampled from studies included in 238 comparative DTA systematic reviews indexed in MEDLINE in 2017. From each study report, we extracted six design elements characterizing participant flow and the labels used by authors.

Results: We identified a total of 46 unique combinations of study design features in our sample, based on six design elements characterizing participant flow. We classified the studies into five study design categories based on how participants were allocated to receive each index test: 'fully paired' (n=79), 'partially paired, random subset' (n=0), 'partially paired, nonrandom subset' (n=2), 'unpaired randomized' (n=1) and 'unpaired nonrandomized' (n=3). The allocation method used in 15 studies was unclear. Sixty-one studies reported, in total, 29 unique study design labels but only four labels referred to specific design features of comparative studies.

Conclusion: Our classification scheme can help systematic review authors define study eligibility criteria, assess risk of bias, and communicate the strength of the evidence. A standardized labelling scheme could be developed to facilitate communication of specific design features. © 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Keywords: Diagnostic accuracy; Test comparison; Study design; Comparative accuracy studies; Bias

Conflicts of interest: The authors have no conflict of interests to declare.

Funding: Amsterdam UMC (The Netherlands) provided funding for this study. The funding organization had no role in the design, collection, analysis, and interpretation of the data or the decision to approve publication of the finished manuscript.

CRediT author statement: **B.Y.:** Conceptualization, Project administration, Methodology, Investigation, Formal analysis, Writing – Original Draft, Writing – Review & Editing. **M.O.:** Conceptualization, Methodology, Investigation, Formal analysis, Writing – Review & Editing. **Y.V.:** Investigation, Formal analysis, Writing – Review & Editing. **M.W.L.:** Conceptualization, Methodology, Writing – Review & Editing, Supervision. **Y.T.:** Methodology, Writing – Review & Editing. **C.H.:** Methodology, Writing – Review & Editing. **P.M.M.B.:** Methodology, Formal analysis, Writing – Review & Editing, Supervision. **M.M.G.L.:** Conceptualization, Methodology, Formal analysis, Writing – Review & Editing, Supervision.

* Corresponding author: Tel: +31(0)20 5666948

E-mail address: b.d.yang@amsterdamumc.nl (B. Yang).

https://doi.org/10.1016/j.jclinepi.2021.04.013

What is new?

Key findings

- Comparative diagnostic test accuracy (DTA) studies could be classified in five basic design categories based on how participants were allocated to index tests.
- Study design labels used by authors of comparative DTA studies were often nonspecific and seldom conveyed information regarding participant allocation.

What this adds to what is known

- First study to empirically examine variations in comparative DTA designs, with some designs susceptible to confounding.

What is the implication, what should change now

- Systematic review authors could use the proposed study design classification scheme when defining eligibility criteria and for tailoring risk of bias assessments.
- Investigators of primary DTA studies could use the classification scheme in selecting the best-fitting design for their study.
- Efforts may be undertaken to develop informative labels for comparative DTA studies.

1. Introduction

Of key interest to many clinicians and policymakers is an evidence-based answer to the question whether one diagnostic test performs better than others, for the same target condition. Selecting the best tests for an accurate classification of patients requires a comparative evaluation of the accuracy of the tests under consideration. Valid estimates of comparative accuracy can be obtained from comparative diagnostic test accuracy (DTA) studies: a comparison of accuracy of two or more tests (none of which is the reference standard) within a single study [1,2].

Several authors have described study designs for comparing the accuracy of two or more tests [1,3–6]. Probably best known are the design in which each participant undergoes all index tests and the design in which participants are randomly allocated to one of the index tests. Both designs aim to avoid confounding (Box 1) by comparing like-with-like [1]. Yet other designs exist, which may be more susceptible to bias.

Knowledge of the diversity of study designs is key when performing a systematic review. Review authors need to be able to identify and include designs relevant for their review question, assess studies' susceptibility to biases and analyze the results appropriately. An understanding of the difference between study designs is also important to primary study investigators, in order to choose the most appropriate one for their research question.

Previously, Tajik et al. used systematic review methodology to identify and classify the designs of studies for evaluating treatment selection markers [7]. This was done by analyzing the flow of patients in each study and grouping studies with similar design features. By grouping comparative DTA studies in a similar way, based on key features, we may be able to classify such studies, with implications for risk of bias assessment and statistical analysis.

In this methodological literature review, we examined a sample of published comparative DTA studies to document the range of study designs for answering comparative accuracy questions. Based on our findings, we proposed a study design classification scheme. As a secondary objective, we documented the labels that study authors have used to describe their study, to examine the availability of widely used informative labels, which could be adopted by current or future design classification schemes.

Box 1 Confounding in the context of comparative accuracy.

A pertinent question in comparative accuracy research is whether any observed differences in accuracy between two or more tests are caused by the characteristics of the tests themselves or by confounding factors. Confounding is a bias in estimating causal effects [8]. Confounding in the context of comparative accuracy occurs when a variable that influences test accuracy also affects the choice of the index test used.

For example, suppose we are comparing the accuracy of computed tomography (CT) and ultrasound for the diagnosis of appendicitis without any randomization, and patients with a severe clinical presentation are more likely to receive CT rather than ultrasound. If appendicitis is more often detected on imaging in patients with a more severe clinical presentation, the relative accuracy of CT compared to ultrasound will be overestimated. Here, severity of presentation is a confounder.

2. Methods

2.1. Study design

A literature survey of comparative DTA studies. The study protocol was registered at the Open Science Framework (https://osf.io/6xkr3).

2.2. Data sources

We sampled comparative DTA studies from all studies included in an existing set of 238 comparative DTA systematic reviews indexed in MEDLINE and published between January 1st and December 31st 2017 [9].

2.3. Eligibility criteria

Any comparative DTA study in humans was eligible. We defined a comparative DTA study as a study that evaluated two or more index tests and for which the study report contained at least one statement, anywhere in the article, in which the accuracy of the index tests was compared. The comparison could be qualitative (for example: higher/lower, superior/inferior, better/best/worst/worse, and versus) or quantitative. We excluded non-English language studies and studies for which the full-text report could not be retrieved.

2.4. Study selection

We retrieved the references of all primary studies included in the 238 comparative DTA systematic reviews. We focused on studies published between 2015 and 2017, the three most recent years in these reviews. We then evaluated the eligibility of primary studies, randomly selected from the list. We assigned a random number to each study, using a random number generator on Google Sheets software (Google, Mountain View, California, U.S.) and evaluated studies for eligibility starting from the lowest study numbers. We did so first by assessing the title and abstract and then the full-text report, until we could include 100 primary comparative DTA studies. We did not stratify the selection by review or test type. Title and abstract screening and full-text assessment were done in duplicate. Disagreements were resolved by consensus, or by consulting a senior author.

2.5. Data extraction

We extracted the following data items in duplicate; disagreements were resolved by consensus. If a comparative DTA study contained more than one test comparison, we extracted data only on the first comparison reported in the study.

- Study information: study ID, number of index tests, type of index test, target condition
- Study design labels: any terms that authors use to identify the study design
- Study design features: the following six design elements (i.e. grouping of design features) that characterize the flow of participants through the study
  1. Direction of data collection (prospective or retrospective)
  2. Number of ‘gates’, i.e. sets of eligibility criteria for recruiting diseased and nondiseased participants (single or multiple gates)
  3. Participant sampling method (consecutive, random, or neither)
  4. Method of allocating participants to index tests (any method reported in the study; including, but not limited to: each participant receiving all tests, random allocation, nonrandom allocation, other)
  5. Number of reference standards (single or multiple)
  6. Limited verification design (verification of some but not all participants by design: yes or no)

The selection of these design elements was based on the authors’ opinions, without a formal development procedure. The authors brainstormed about an initial list of design features that would characterize participant flow. This list was updated during the data extraction process to include unanticipated design features. In Table 1 we describe the definitions and relevance of these items. Similar items are used to assess risk of bias in diagnostic accuracy studies [10,11].

Additionally, we recorded *post hoc* whether studies used a method to reduce confounding in their design or analysis. Methods to adjust for confounding included but were not limited to restriction, stratification, inverse probability weighting, multivariable regression, and other techniques.

For identifying design features, we aimed to make our own assessment independent of the shorthand terms reported by study authors. For example, if a study reported that data collection was prospective, while it was clear from the description that data had been collected prior to study initiation, we classified the study as retrospective. If no information was reported, other than the shorthand terms reported by the authors, we relied on the shorthand terms. If no descriptions or terms were reported, we marked the feature as ‘unclear’.

2.6. Data synthesis

We calculated the frequency of each design feature and combinations thereof. Based on our findings and discussions within the author team, we aimed to classify study designs according to features that would be (1) specific for comparative DTA studies (rather than single test evaluations) and (2) could have consequences for risk of bias assessment and for the preferred approach to statistical analysis (e.g. whether or not to use methods that take correlated data into account). For each design category, we produced a flow diagram with a brief discussion of its strengths and weaknesses, and included an example from our sample of studies. We examined and categorized study design labels based on whether the label conferred information on (1) identification of a DTA study, (2) a comparison, and (3) any design features specific for comparisons, such as the method for allocating participants to tests.

Table 1. Six comparative DTA study design elements characterizing participant flow.

1. Direction of data collection	
Definition	Whether data was collected prior to or after study initiation. We considered 'data collection' to include the following study processes: (1) enrolling study participants, (2) performing the tests, and (3) interpreting test results. *Prospective* (defined as data collection after study initiation) *Retrospective* (defined as not prospective)
Relevance	Knowledge of direction of data collection allows (partial) reconstruction of participant flow when the study is poorly reported. Moreover, purposefully collected data from a prospective study may be of higher quality (e.g. consistent coding of variables, less missing data) than routinely collected data.
2. Number of gates	
Definition	Whether a single set or multiple sets of eligibility criteria were used for recruiting participants with and without the target condition. *Single gate* (defined as a single set of eligibility criteria for all participants, most commonly those in whom the target condition is suspected) *Multiple gate* (defined as separate sets of eligibility criteria, e.g. one set for patients with the target condition and a second set for healthy controls)
Relevance	Sampling participants with and without the target condition separately, rather than sampling a single group of participants suspected of the target condition, may inflate estimates of diagnostic test accuracy [12]
3. Participant sampling method	
Definition	*Consecutive* (enrolment of all eligible participants in sequence) *Random (*random sampling of participants) *Neither consecutive nor random* (if the sampling was not consecutive or random, for example sampling based on convenience or availability)
Relevance	Consecutive or random sampling strategies may help reduce bias.
4. Method of allocating participants to index tests	
Definition	Method used by investigators to decide which index test a participant would receive. We extracted any method described in the study report.
Relevance	The index test groups being compared should be comparable with respect to factors that may affect test accuracy. Some allocation methods, such as randomization of a large group, are expected to produce comparable groups.
5. Number of reference standards	
Definition	Whether a single or multiple reference standards were used. *Single reference standard* (one reference standard for all participants) *Two or more reference standards* (if so, we extracted information on the process for selecting the reference standard for a given participant)
Relevance	Ideally, index test results should be verified by a single reference standard. If more than one reference standard is used, the risk of bias also depends on how a reference standard is chosen for each participant.
6. Limited verification	
Definition	Verification of the presence/absence of the target condition in a subset of participants, e.g. based on index test results to increase efficiency *Yes* (if so, we extracted which participants were verified) *No* (all participants were verified, or the study started with participants who were already verified)
Relevance	Full verification of all index test results is not always necessary to obtain valid estimates of comparative accuracy. 'Limited verification' designs can increase study efficiency (for example, by verifying only discordant index test results) but also limits what measures of comparative accuracy can be used to express study results [3,13–15]

3. Results

3.1. Results of the search

The 238 comparative DTA systematic reviews contained a total of 5,789 references to primary studies. Of the 5,789 studies, 946 studies had been published between 2015 and 2017. We assigned a random study number to these 946 studies and ranked these from first to last. As we expected to exclude studies, we selected the first 320 studies for title and abstract screening. We excluded 175 studies during this phase (five were non-English study reports) and ranked the remaining 145 from first to last random study number. The first 113 full text reports were assessed until the inclusion of 100 comparative DTA studies, as we had to exclude 13 studies (Appendix 1 contains the flow diagram and Appendix 2 contains a list of all included and the 13 excluded studies, with reasons).

Table 2. Frequency of study design features.

Study design features	**N**
Total	100
Direction of data collection	
Prospective	43
Retrospective	42
Unclear	15
Number of gates (sets of eligibility criteria)	
Single	73
Multiple	26
Unclear	1
Participant sampling method	
Consecutive	36
Random	2
Neither consecutive nor random	0
Unclear	62
Method of allocating participants to index tests	
Each participant received all index tests	79
Some participants received all index tests, others only one of the index tests (randomly)	0
Some participants received all index tests, others only one of the index tests (nonrandomly)	2
Random allocation	1
Nonrandom allocation	3
Unclear*	15
Number of reference standards	
Single	72
Multiple; choice depended on index test results	2
Multiple; choice depended on a third test not in the comparison	3
Multiple; choice was based on clinical indication	1
Multiple; choice was unexplained	12
Unclear	10
Limited verification	
No	93
Yes, verification of only participants with a positive index test result	1
Yes, verification of only participants with a positive index test result, and a random sample of participants with a negative index test result	2
Unclear	4

* In 9 of 15 studies, some participants received all index tests but it was unclear whether this was the case for all participants.

3.2. Characteristics of included studies

Most comparative DTA studies evaluated biochemical ($n = 50$) and imaging tests ($n = 47$). The most frequent target conditions were neoplasms ($n = 54$), disorders of the digestive system (n = 14) and infectious diseases ($n = 10$). Approximately half of the studies compared two index tests ($n = 49$), but comparisons of three ($n = 17$), four ($n = 15$) and five or more ($n = 15$) index tests were also seen. See Appendix 3 for a more detailed overview.

3.3. Study design features

Table 2 summarizes the frequency of the study design features captured by the six design elements relating to participant flow. Based on the study report, we inferred that the direction of data collection was prospective in 43 and retrospective in 42, while it was unclear in 15 studies. Most studies used a single gate ($n = 73$); 26 studies used multiple gates and eligibility criteria were unclear in one study. Participant sampling was consecutive in 36 studies and random in two; it could not be identified in 62 studies.

We identified four different strategies for allocating participants to index tests. The most common method was for each study participant to receive all index tests, which we refer to as ‘pairing’ ($n = 79$). In two other studies, some participants received all index tests, while a selected subset received only one of the index tests and the selection was not random. In one study, participants were randomly allocated to one of the index tests; in three other studies,

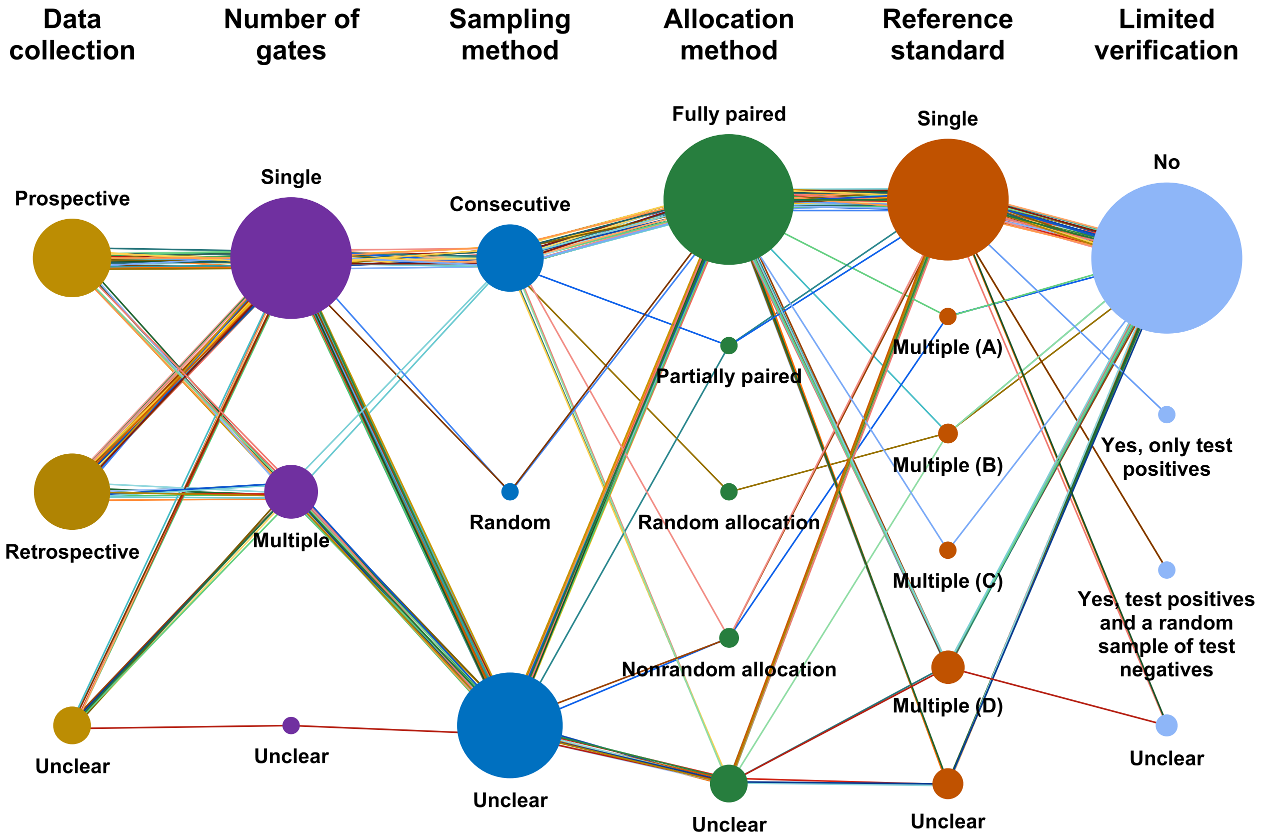


Figure 1. Network diagram of 100 comparative DTA studies and their design features.
This diagram illustrates the data in Table 2 as well as the relation between the design features of 100 studies. Six columns with different colored nodes are the six design elements relating to participant flow. The nodes indicate the design features, and the size of each node is proportional to the number of studies with that specific design feature. Each colored line represents one study, which connects the design features from left to right. Colour version of the figure is available online.
Abbreviations: Multiple (A): ≥2 reference standards, choice depended on index test results; Multiple (B): ≥2 reference standards, choice depended on a third test not in the comparison; Multiple (C): ≥2 reference standards, choice was based on clinical indication; Multiple (D): ≥2 reference standards, choice was unexplained

allocation was not random. In 15 studies the method of allocation was unclear.

Most studies used a single reference standard ($n = 72$). In 18 studies that used two or more reference standards, the choice of the reference standard depended on either the index test results ($n = 2$), a third test not in the comparison ($n = 3$), clinical indication ($n = 1$), or the choice was unexplained ($n = 12$). The number of reference standards was unclear in 10 studies. Limited verification was used in three studies. One study verified only participants with a positive index test result. Two studies verified participants with a positive index test result and a random sample of participants with a negative result. None of the studies (21 of which were not fully paired) relied on methods to account for confounding.

Figure 1 displays the relation between the design features of all 100 studies. Including ‘unclear’ design features, we observed a total of 46 unique combinations of design features in our sample (a full list is available in Appendix 4). Of these, the most frequently occurring design feature combination (n=13) was a prospective study using a single gate, with consecutive sampling, each participant receiving all index tests, a single reference standard, and verification of all participants. Excluding studies with at least one ‘unclear’ design feature produced a total of 11 unique combinations (Appendix 4).

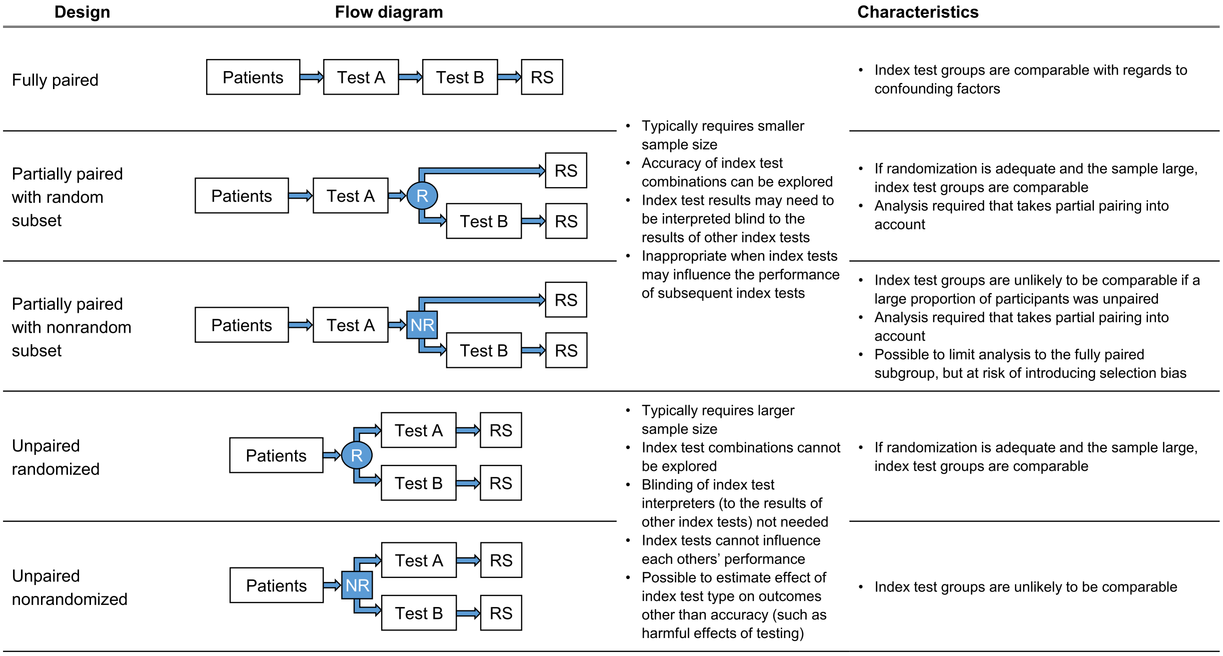
3.4. Labels used to describe comparative DTA studies

Sixty-one study reports used 29 unique study labels (Table 3), while the remaining 39 studies did not use labels to describe their design. Most labels were not DTA-specific, of which the labels ‘retrospective study’ and ‘prospective study’ were most frequently used (19 and 18 times, respectively). Three labels indicated a comparison but communicated no specific information about the design (‘comparative study’; ‘head-to-head comparison’; ‘prospective comparison’). Two DTA-specific labels were used (‘prospective diagnostic study’; ‘comparative diagnostic accuracy’), of which the latter indicated that the study had compared the accuracy of tests without further information about the design. Four labels contained information

Table 3. Labels for describing the study design used in 61 comparative DTA studies.

Type (number of labels) *	Labels (number of studies using a label)**
Non-DTA-specific labels that refer to various study types (18)	Retrospective study (19); Prospective study (18); Cross-sectional study (3); Case series (2); Case-control study; Case-control validation study; Cross-sectional analysis of a prospective cohort; Inception cohort study; Monocenter, retrospective, exploratory, institutional review board-approved cohort study; Multicenter retrospective cohort study; Multicenter retrospective study; Prospective cohort design; Prospective, multicenter, blinded study; Prospective, multicenter, observational study; Prospective, observational study; Prospective ongoing study; Prospective, single-center study; Single-institution prospective trial;
DTA-specific labels (1)	Prospective diagnostic study
Labels indicating a comparison (3)	Head-to-head comparison (3); Comparative study (2); Prospective comparison
DTA-specific labels indicating a comparison (1)	Comparative diagnostic accuracy
Labels indicating a specific design feature for comparisons (4)	Prospective randomized clinical trial; Prospective randomized study; Prospective, non-randomized international multicenter study with within-patient comparison; Within-patient study
Unclear labels (2)	Single-blind research; Verification study

* Some studies reported more than one label.
** If a label is reported by only a single study, the number of studies is omitted.



Fully paired:
- Index test groups are comparable with regards to confounding factors

Paired designs (fully and partially paired):
- Typically requires smaller sample size
- Accuracy of index test combinations can be explored
- Index test results may need to be interpreted blind to the results of other index tests
- Inappropriate when index tests may influence the performance of subsequent index tests

Partially paired with random subset:
- If randomization is adequate and the sample large, index test groups are comparable
- Analysis required that takes partial pairing into account

Partially paired with nonrandom subset:
- Index test groups are unlikely to be comparable if a large proportion of participants was unpaired
- Analysis required that takes partial pairing into account
- Possible to limit analysis to the fully paired subgroup, but at risk of introducing selection bias

Unpaired designs:
- Typically requires larger sample size
- Index test combinations cannot be explored
- Blinding of index test interpreters (to the results of other index tests) not needed
- Index tests cannot influence each others' performance
- Possible to estimate effect of index test type on outcomes other than accuracy (such as harmful effects of testing)

Unpaired randomized:
- If randomization is adequate and the sample large, index test groups are comparable

Unpaired nonrandomized:
- Index test groups are unlikely to be comparable

Figure 2. Classification of comparative DTA study designs based on participant allocation method.
Flow diagrams are drawn assuming a single gate for participant enrolment and two index tests in the comparison. The flow diagrams for partially paired designs are an example and they can be illustrated in multiple ways, with the criterium being that a subset of participants received multiple index tests. Abbreviations: NR: nonrandom allocation; R: random allocation; RS: reference standard.

on design features specific for comparisons, namely the method of allocating participants to index tests (for example, 'prospective, non-randomized international multicenter study with within-patient comparison').

3.5. Classification of study designs

Relying on available study design features, we propose a classification of comparative DTA designs based on the methods for allocating participants to index tests. This classification centers on two characteristics: whether participants receive one or more index tests and whether participants are randomly allocated or not. As the labels used by study authors were heterogeneous, we used provisional, descriptive names for the studies in our dataset: fully paired design ($n = 79$), partially paired design with random subset ($n = 0$), partially paired design with nonrandom subset (n=2), unpaired randomized design ($n = 1$) and unpaired nonrandomized design ($n = 3$). The allocation method used in the remaining 15 studies was unclear. See Figure 2 for flow diagrams of each design. Below, we briefly discuss each design, with its advantages and limitations.

4. Study design categories

4.1. Fully paired design

These are comparative DTA studies in which each participant receives all index tests. We call this design 'fully paired' (even though more than two index tests can be compared) because of the pairing of test results within the same study participant.

Box 2 Example of a fully paired design.

Agorastos and colleagues compared the accuracy of human papillomavirus (HPV) testing and liquid-based cytology (LBC) for the diagnosis of cervical intraepithelial neoplasia grade 2 or worse [16]. The study enrolled 4009 women taking part in cervical cancer screening. Each woman underwent both tests: a sample was collected for initial LBC evaluation and an aliquot of the remaining sample was used for HPV testing. The cytologists and the molecular biologists performing the HPV test were blinded to each other's' test results. Women testing positive for either HPV or LBC received the reference standard colposcopy (with or without subsequent biopsy) and a random sample of 106 women negative for HPV and LBC received colposcopy as well. The study reported 'two-by-four' table data (Appendix 5) and used McNemar's test for paired data for comparing the sensitivity and specificity of HPV and LBC.

There are several advantages to a fully paired design. First of all, participants receiving both index test A and index test B are identical in terms of factors affecting test accuracy (i.e. no confounding). Second, this design typically has greater statistical power compared to unpaired designs, as between-subject variability is minimized. Lastly, a paired design allows explorations of the accuracy of index test combinations [17]. For instance, one may conduct a paired study comparing computed tomography (CT) versus magnetic resonance imaging (MRI), and be able to assess the accuracy of CT, or MRI, and of CT and MRI combined. Subsequently, the accuracy of a range of testing strategies can be compared in the same study: CT only, MRI only, a strategy of CT followed by MRI if CT negative, a strategy of MRI followed by CT if MRI negative, and others (see Laméris et al. [18] for an example of a comparison of multiple testing strategies).

While the paired design has many appeals, it can also have disadvantages. It may not be feasible or ethical to expose each participant to multiple index tests (for example, if the index tests are invasive). Biases may be introduced if one index test influences the performance of subsequently performed tests (for example, biopsy needle A may disrupt the histological architecture before biopsy needle B is used). Furthermore, it may be necessary to blind each of the index test interpreters to other index test results, if the test interpretation has a subjective component.

It should be noted that, when a paired design is used, test results may be correlated and appropriate statistical methods should be used to take this correlation into account [13,19]. A paired study enables the construction of a contingency table (sometimes called 'two-by-four' table or 'joint classification' table), which cross-classifies results of two index tests against each other separately for participants with and those without the target condition. Two-by-four tables should ideally be reported to allow for meta-analysis of comparative accuracy data [20].

4.2. Partially paired design with nonrandom subset

We refer to a study as being 'partially paired' if some participants receive multiple index tests but others receive only one of the index tests.

Box 3 Example of a partially paired design with nonrandom subset.

A study compared the accuracy of MRI, ultrasound and mammography for detecting residual breast cancer after neoadjuvant chemotherapy [21] in 150 breast cancer cases from 143 women. All cancer cases ($n = 150$) underwent MRI and ultrasound after chemotherapy, but only 131 of them also underwent mammography for reasons unknown. All participants subsequently underwent surgery and histopathology to determine the presence of residual cancer.

Whether the index test groups are comparable in such a design depends on the proportion of participants that are paired (as a larger proportion of paired participants implies that the groups are more comparable) and the process by which some participants received only one of the tests. If this process is nonrandom, there is a risk of confounding. For instance, suppose that we conducted a study comparing CT versus MRI and 60 of 100 study participants received CT and MRI, while 40 received CT only. Participants who were severely ill (requiring immediate surgical intervention) received CT only as they could not undergo further MRI testing, while participants with less severe disease were able to receive both CT and MRI. The comparison between the tests is then confounded by disease severity. We call this a partially paired design with a 'nonrandom subset'.

Since some participants receive multiple index tests, partially paired studies are also susceptible to the biases that may affect fully paired studies: one index test may influence the performance of subsequent tests, and the interpretation of one index test may be biased by the knowledge of the results of another index test.

Partially paired studies with a nonrandom subset can be analyzed by using data from all participants, or by analyzing only the subgroup of participants who received all index tests. The former approach requires statistical methods that take partial pairing into account [22–24] and is at risk of confounding. The latter approach reduces the risk of confounding, but also reduces the study sample to participants who received all index tests, which may not be representative of the study population.

4.3. Partially paired design with random subset

Participants in a partially paired design could also be randomly allocated to receive one or multiple index tests. If the random allocation was adequate (that is, the allocation sequence was randomly generated and allocation concealed) we can assume that the group with a single index test and the group with multiple index tests are comparable in terms of confounding variables. This design could be an attractive option if not all participants can undergo the second index test, for example because the second test is expensive, invasive, or scarce. In our sample, we did not identify any studies with this design.

4.4. Unpaired randomized design

If a paired design is unfeasible, unethical, or inappropriate for other reasons, investigators can allocate each participant to one of the index tests, ideally by randomization.

Box 4 Example of an unpaired randomized design.

Carrara and colleagues conducted a randomized study to compare the accuracy of two needles (25 and 22 gauge) for endoscopic ultrasound-guided fine-needle aspiration of solid gastrointestinal masses [25]. The authors did not explicitly state a target condition, although it was most likely any pancreatic or peripancreatic malignancy. Participants were allocated in a 1:1 ratio to either needle using a computer-generated random sequence. The reference standard for the final diagnosis was surgery, or follow-up if the tumor was unresectable.

If the sample size is sufficiently large, randomization is expected to produce index test groups that are comparable in terms of confounding variables [26]. Analogous to randomized trials of interventions, the allocation sequence should be randomly generated and allocation concealed. Unpaired randomized designs also allow the estimation of the effect of index tests on outcomes other than accuracy, such as harms and other direct health effects of tests [3]. Since the data are unpaired, statistical power is typically lower than in paired designs, and combinations of index tests cannot be explored. Unpaired studies produce two-by-two contingency tables separately for each index test.

4.5. Unpaired nonrandomized design

In unpaired designs, participants can be nonrandomly allocated to index test groups, for example based on clinicians' preference [27].

Box 5 Example of an unpaired nonrandomized design.

Sheridan and colleagues compared MRI and MR arthrography (MRA) for the diagnosis of superior labrum anterior posterior (SLAP) lesions of the shoulder [28]. They retrospectively reviewed routinely collected data of patients who underwent arthroscopy (the reference standard) and included patients ($n = 444$) when they had previously received MRI (n=234) or MRA (n=210). The authors did not report reasons why some patients received a particular test. The analysis did not account for confounding; overall accuracy, sensitivity, specificity, and predictive values of MRI and MRA were compared by examining whether the 95% confidence intervals of the tests overlapped.

Since the index test groups resulting from nonrandom allocation are unlikely to be comparable, comparative accuracy estimates from such studies will only be meaningful when confounding is addressed in the design or analysis. Analogous to observational studies of interventions and exposures, assumptions regarding the underlying causal structure could be made explicit using a directed acyclic graph [29] and confounding reduced in the analysis using methods such as matching,[30] regression analysis,[19] inverse probability weighting, or other techniques. However, none of the three unpaired nonrandomized studies included in our review used methods to reduce confounding.

5. Discussion

In this methodological review, we aimed to document the range of study designs in published comparative DTA studies and to produce a classification scheme. From each study we extracted data on six design elements pertaining to participant flow. We found 46 unique combinations of design features. Our findings show that, while comparative DTA studies can be designed in various ways, certain features are more common than others. These are: a single gate for participant enrolment, pairing of index tests, the use of a single reference standard, and verification of all participants. As we assumed that participant allocation to index tests is probably the characteristic most strongly associated with risk of bias (although empirical confirmation of bias is admittedly lacking), we used this feature as the basis for our classification scheme.

The existence of comparative DTA studies and potential sources of bias in such studies have received relatively little attention in the literature. A recent overview of 238 comparative DTA systematic reviews found that only 24% of reviews mentioned comparative DTA studies as part of their inclusion criteria; only 0.8% had planned or performed risk of bias assessment for these [9]. It is possible that authors lack the knowledge and tools necessary to set appropriate eligibility criteria and identify sources of bias across various comparative designs. We hope that review authors find our scheme helpful in searching for and including appropriate designs, and in tailoring risk of bias assessments to different types of designs. Our scheme may also benefit primary study investigators, who might find it difficult to select the most valid study design for their comparative question and consequently risk using sub-optimal or inefficient designs. We hope that, based on the strengths and limitations of each design, investigators will be able to choose the appropriate design for their research question.

Study design labels facilitate efficient communication of different design features but only when they are unambiguous and their use is standardized among investigators. We found a range of labels used by authors of comparative DTA studies, with 61 studies in our sample reporting 29 different study design labels. An examination of these revealed that most labels do not distinguish comparative DTA studies from other study types, nor do they provide basic information other than that different tests were compared. Only four labels communicated essential methodological information regarding participant allocation. Overall, a clear standardized labeling framework for comparative DTA studies appears to be lacking. There is room for the development of informative, consensus-based labels for efficient communication of various design features of comparative DTA studies.

Our review represents, to our knowledge, a first attempt at classifying the designs of comparative DTA studies based a systematic survey of the literature. Our basic scheme represents a starting point that can be further expanded and characterized (with other features) to allow for a more detailed communication of the way comparative DTA studies are organized.

5.1. Limitations of this review

There are a number of limitations in this review. First, there is currently no agreed definition of comparative DTA studies and others may disagree with our definition. Some may prefer to restrict comparative DTA studies to studies with an explicit comparative objective or hypothesis. (Sixty-four of 100 studies in our sample would satisfy that redefinition.) Second, our decision to sample 100 comparative DTA studies was based on feasibility rather than on achieving data saturation. Due to our limited sample size, we may have failed to capture designs that are relatively uncommon. Third, we sampled from studies included in systematic reviews. Therefore, the frequency of design features (Table 1) reflects the review topics and the in/exclusion criteria of such reviews. The 100 comparative DTA studies in our sample originated from 61 systematic reviews, of which 11 (18%) had one or more in/exclusion criteria restricting to specific study design features (Appendix 6). As a consequence, the frequency of some design features in our sample (for example prospective data collection and a single gate for participant enrolment) may be higher compared to studies included in reviews without such restrictions. We also admit that our classification scheme has not yet been externally validated in a different cohort of studies. Future methodological studies could examine the merits of the classification scheme for describing variation in study designs in a different body of evidence.

6. Recommendations

Our classification scheme for comparative DTA study designs is intended to help systematic reviewers in defining study eligibility criteria, in assessing risk of bias, and in communicating the strength of the evidence. In addition, researchers could use the scheme to select optimal designs for future primary comparative DTA studies. Since existing labels describing comparative DTA studies were generally heterogeneous and nonspecific, efforts could be undertaken to develop an agreed set of informative labels for comparative DTA studies.

Acknowledgements

We thank Pieter Zwanenburg, MSc (University of Amsterdam), for his comments and suggestions on improving a previous draft. Yemisi Takwoingi is funded by a UK National Institute for Health Research (NIHR) Postdoctoral Fellowship, and is supported by the NIHR Birmingham Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or the Department of Health and Social Care.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.04.013.

References

[1] Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical Evidence of the importance of comparative studies of diagnos-

tic test accuracy. Ann Intern Med 2013;158(7):544. doi:10.7326/0003-4819-158-7-201304020-00006.

[2] Leeflang MMG, Reitsma JB. Systematic reviews and meta-analyses addressing comparative test accuracy questions. Diagnostic Progn Res 2018;2(1):17. doi:10.1186/s41512-018-0039-0.

[3] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332(7549):1089–92. doi:10.1136/bmj.332.7549.1089.

[4] Wade R, Corbett M, Eastwood A. Quality assessment of comparative diagnostic accuracy studies: our experience using a modified version of the QUADAS-2 tool. Res Synth Methods 2013;4(3):280–6. doi:10.1002/jrsm.1080.

[5] Colli A, Fraquelli M, Casazza G, et al. The architecture of diagnostic research: From bench to bedside-research guidelines using liver stiffness as an example. Hepatology 2014;60(1):408–18. doi:10.1002/hep.26948.

[6] Bossuyt P., Leeflang M. Chapter 6: Developing Criteria for Including Studies. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4 [Updated September 2008]*. The Cochrane Collaboration; 2008. Available at: https://methods.cochrane.org/sdt/handbook-dta-reviews.

[7] Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: a systematic review and classification. Clin Cancer Res 2013;19(17):4578–88. doi:10.1158/1078-0432.CCR-12-3722.

[8] Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health 2001;22(1):189–212. doi:10.1146/annurev.publhealth.22.1.189.

[9] Yang B, Vali Y, Dehmoobad Sharifabadi A, et al. Risk of bias assessment of test comparisons was uncommon in comparative accuracy systematic reviews: an overview of reviews. J Clin Epidemiol 2020;127:167–74. doi:10.1016/j.jclinepi.2020.08.007.

[10] Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529. doi:10.7326/0003-4819-155-8-201110180-00009.

[11] Yang B., Mallett S., Takwoingi Y., et al. Development of QUADAS-C, a risk of bias tool for comparative diagnostic accuracy studies. doi:10.17605/OSF.IO/HQ8MF.

[12] Rutjes A, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PMM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51(8):1335–41. doi:10.1373/clinchem.2005.048595.

[13] Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. J Clin Epidemiol 2010;63(8):883–91. doi:10.1016/j.jclinepi.2009.08.024.

[14] Irwig L, Macaskill P, Farnsworth A, et al. A randomized crossover trial of PAPNET for primary cervical screening. J Clin Epidemiol 2004;57(1):75–81. doi:10.1016/S0895-4356(03)00259-2.

[15] Chock C, Irwig L, Berry G, Glasziou P. Comparing dichotomous screening tests when individuals negative on both tests are not verified. J Clin Epidemiol 1997;50(11):1211–17. doi:10.1016/S0895-4356(97)00122-4.

[16] Agorastos T, Chatzistamatiou K, Katsamagkas T, et al. Primary screening for cervical cancer based on high-risk human papillomavirus (HPV) detection and HPV 16 and HPV 18 genotyping, in comparison to cytology. PLoS One 2015;10(3):e0119755. doi:10.1371/journal.pone.0119755.

[17] Macaskill P, Walter SD, Irwig L, Franco EL. Assessing the gain in diagnostic performance when combining two diagnostic tests. Stat Med 2002;21(17):2527–46. doi:10.1002/sim.1227.

[18] Laméris W, Van Randen A, Wouter Van Es H, et al. Imaging strategies for detection of urgent conditions in patients with acute abdominal pain: Diagnostic accuracy study. BMJ 2009;339(7711):29–33. doi:10.1136/bmj.b2431.

[19] Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press; 2003.

[20] Takwoingi Y. Meta-analytic approaches for summarising and comparing the accuracy of medical tests. Univ Birmingham Res Arch 2016(March).

[21] Schaefgen B, Mati M, Sinn HP, et al. Can routine imaging after neoadjuvant chemotherapy in breast cancer predict pathologic complete response? Ann Surg Oncol 2016;23(3):789–95. doi:10.1245/s10434-015-4918-0.

[22] Thomson PC. A hybrid paired and unpaired analysis for the comparison of proportions. Stat Med 1995;14(13):1463–70. doi:10.1002/sim.4780141306.

[23] Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. Med Decis Mak 1998;18(1):110–21. doi:10.1177/0272989X9801800118.

[24] Gallas BD, Pesce LL. Comparison of ROC methods for partially paired data. Med Imaging 2009 Image Perception, Obs Performance, Technol Assess. 2009;7263(301):72630V. doi:10.1117/12.813688.

[25] Carrara S, Anderloni A, Jovani M, et al. A prospective randomized study comparing 25-G and 22-G needles of a new platform for endoscopic ultrasound-guided fine needle aspiration of solid masses. Dig Liver Dis 2016;48(1):49–54. doi:10.1016/j.dld.2015.09.017.

[26] Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010:340. doi:10.1136/bmj.c869.

[27] Tuite MJ, Rutkowski A, Enright T, Kaplan L, Fine JP, Orwin J. Width of high signal and extension posterior to biceps tendon as signs of superior labrum anterior to posterior tears on MRI and MR arthrography. Am J Roentgenol 2005;185(6):1422–8. doi:10.2214/AJR.04.1684.

[28] Sheridan K, Kreulen C, Kim S, Mak W, Lewis K, Marder R. Accuracy of magnetic resonance imaging to diagnose superior labrum anterior–posterior tears. Knee Surgery, Sport Traumatol Arthrosc 2015;23(9):2645–50. doi:10.1007/s00167-014-3109-z.

[29] Lederer DJ, Bell SC, Branson RD, et al. Control of confounding and reporting of results in causal inference studies: guidance for authors from editors of respiratory, sleep, and critical care journals. Ann Am Thorac Soc 2018;16(1) AnnalsATS.201808-564PS. doi:10.1513/AnnalsATS.201808-564PS.

[30] Del Turco MR, Mantellini P, Ciatto S, et al. Full-field digital versus Screen-film mammography: comparative accuracy in concurrent screening cohorts. Am J Roentgenol 2007;189(4):860–6. doi:10.2214/ajr.07.2303.