European Radiology (2019) 29:5386–5394
https://doi.org/10.1007/s00330-019-06045-7

EXPERIMENTAL

Comparative reviews of diagnostic test accuracy in imaging research: evaluation of current practices

Anahita Dehmoobad Sharifabadi[1] · Mariska Leeflang[2] · Lee Treanor[1] · Noemie Kraaijpoel[3] · Jean-Paul Salameh[4] · Mostafa Alabousi[5] · Nabil Asraoui[6] · Jade Choo-Foo[6] · Yemisi Takwoingi[7,8] · Jonathan J. Deeks[7,8] · Matthew D. F. McInnes[9]

Received: 21 November 2018 / Revised: 30 December 2018 / Accepted: 24 January 2019 / Published online: 21 March 2019

Abstract

Purpose The purpose of this methodological review was to determine the extent to which comparative imaging systematic reviews of diagnostic test accuracy (DTA) use primary studies with comparative or non-comparative designs.

Methods MEDLINE was used to identify DTA systematic reviews published in imaging journals between January 2000 and May 2018. Inclusion criteria: systematic reviews comparing at least two index tests (one of which was imaging-based); review characteristics were extracted. Study design and other characteristics of primary studies included in the systematic reviews were evaluated.

Results One hundred three comparative imaging reviews were included; 11 (11%) included only comparative studies, 12 (11%) included only non-comparative primary studies, and 80 (78%) included both comparative and non-comparative primary studies. For reviews containing both comparative and non-comparative primary studies, the median proportion of non-comparative primary studies was 81% (IQR 57–90%). Of 92 reviews that included non-comparative primary studies, 86% did not recognize this as a limitation. Furthermore, among 4182 primary studies, 3438 (82%) were non-comparative and 744 (18%) were comparative in design.

Conclusion Most primary studies included in comparative imaging reviews are non-comparative in design and awareness of the risk of bias associated with this is low. This may lead to incorrect conclusions about the relative accuracy of diagnostic tests and be counter-productive for informing guidelines and funding decisions about imaging tests.

Key Points

- *Few comparative accuracy imaging reviews include only primary studies with optimal comparative study designs. Among the rest, few recognize the risk of bias conferred from inclusion of primary studies with non-comparative designs.*
- *The demand for accurate comparative accuracy data combined with minimal awareness of valid comparative study designs may lead to counter-productive research and inadequately supported clinical decisions for diagnostic tests.*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-06045-7) contains supplementary material, which is available to authorized users.

✉ Matthew D. F. McInnes
mmcinnes@toh.on.ca

[1] Department of Radiology-Faculty of Medicine, University of Ottawa, Ottawa, Canada

[2] Amsterdam UMC, Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam Public Health, University of Amsterdam, Meibergdreef 9, Amsterdam, Netherlands

[3] Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

[4] Clinical Epidemiology Program, Ottawa Hospital Research Institute, School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada

[5] McMaster University, Hamilton, Canada

[6] Faculty of Medicine, University of Ottawa, Ottawa, Canada

[7] Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Birmingham B15 2TT, UK

[8] NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham B15 2TT, UK

[9] Department of Radiology, University of Ottawa, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Room c159 Ottawa Hospital Civic Campus, 1053 Carling Ave., Ottawa, ON K1Y 4E9, Canada

- *Using comparative accuracy imaging reviews with a high risk of bias to inform guidelines and funding decisions may have detrimental impacts on patient care.*

Keywords Diagnostic test, routine · Comparative effectiveness research · Sensitivity and specificity

Abbreviations

CT	Computed tomography
DTA	Diagnostic test accuracy
MRI	Magnetic resonance imaging
PRISMA – DTA	Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies
RCT	Randomized controlled trial
RSNA	Radiologic Society of North America
SPSS	Statistical Package for the Social Sciences
US	Ultrasound
VOICE	Value of Imaging through Comparative Effectiveness

Introduction

Research comparing the diagnostic accuracy of two or more imaging tests is essential to determine optimal diagnostic pathways and downstream treatment decisions. Comparative effectiveness research has recently been highly prioritized; there are a multitude of initiatives and increased funding to ensure high-quality comparative medical research [1–5]. In medical imaging, comparative accuracy studies are in high demand; the National Institute of Medicine has called for investigators to compare the effectiveness of imaging tests in diagnosis and monitoring of various target conditions [1]. Systematic reviews of comparative accuracy studies, where evidence on the diagnostic accuracy of two or more index tests are synthesized, are a major determinant of funding new imaging technologies and policies regarding their usage [6].

When comparing healthcare interventions, randomized controlled trials (RCTs) are common and the highest standard for comparison [7]. Systematic reviews of RCTs are used in clinical decision making for selecting optimal interventions. Systematic reviews that compare the accuracy of diagnostic tests often use primary studies that have investigated only one of the index tests and make uncontrolled between-study comparisons [8, 9]. These primary study comparisons are termed non-comparative as they evaluate index tests in different and/or non-randomized study populations with varying characteristics and different reference standards or methodologies. In contrast to indirect comparisons and network meta-analyses of interventional studies, there is typically no common control to adjust for the differences in accuracy of common tests between studies, and their usage at the same point in the diagnostic pathway is often incorrectly assumed [10–12]. As a result, the high degree of heterogeneity observed in DTA meta-analyses [13] raises concern that results derived from comparing non-comparative primary studies may be at high risk of bias [9].

Unlike studies comparing therapeutic interventions, well-designed comparative studies in diagnostic test accuracy (DTA) research are relatively uncommon [9, 14]. High-quality comparative methods would include primary studies that apply the index tests to every study participant or randomly allocate participants to receive one of the index tests [8, 9]. These studies are expected to be at lower risk of bias since they enable comparisons of like-with-like, either within the same participants or between randomized groups [15].

DTA systematic reviews using non-comparative primary studies are likely to report different results and are at higher risk of bias, compared with systematic reviews using comparative primary studies [9]. Methods to adjust for bias introduced due to the inclusion of non-comparative primary studies, such as adjusting for study features or using individual patient data, have not demonstrated success [16, 17]. If comparative imaging reviews largely rely on non-comparative primary studies, the conclusions regarding the comparative accuracy of index tests may be biased. As such, an evaluation of the current practices in conducting imaging comparative accuracy research is an important first step in understanding the potential extent of such bias.

The purpose of this methodological review was to determine the extent to which comparative imaging systematic reviews of DTA use primary studies with comparative or non-comparative designs.

Methods

Research ethics board approval is not required for this study type at our institution. The study protocol was registered in the Open Science Framework (DOI https://doi.org/10.17605/OSF.IO/P7X3C).

Terminology

A “comparative imaging review” is defined as a systematic review of DTA that includes at least one comparison of test accuracy of imaging tests.

For classification of primary studies that were included in the comparative imaging reviews, we used the term “non-comparative primary study” to refer to studies that either (a) evaluated a single test or (b) compared the accuracy of two or more tests between non-randomly allocated groups. We used the term "comparative primary study" to refer to a study that aimed to compare either (a) the accuracy of at least two index tests in the same population of participants (for which participants had both tests and the reference standard) or (b) one that randomly allocated patients to receive each index test and reference standard. Furthermore, we classified a primary study as comparative only if at least two of the index tests compared are the same as the competing index tests in the comparative imaging review. Figure 1 provides a schematic of the classification scheme for primary study comparative methods.

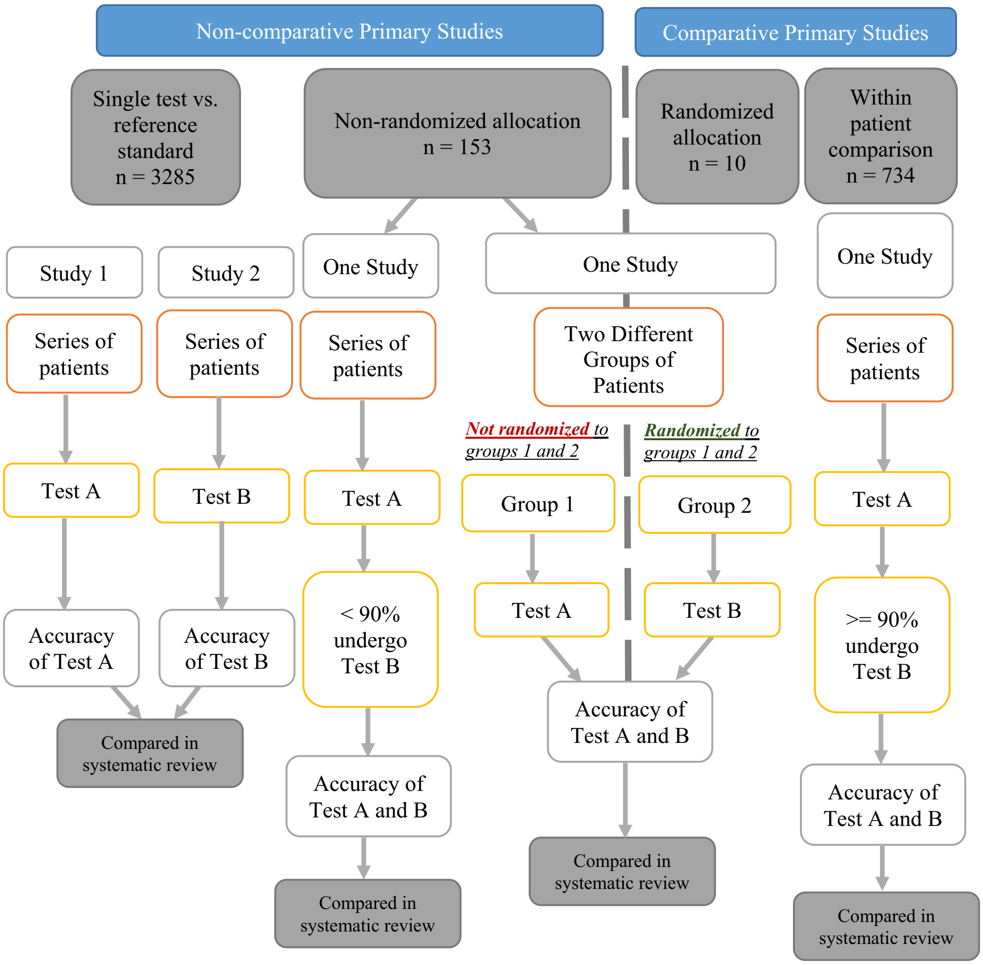
Data sources, searches, and inclusion

MEDLINE was searched to identify DTA systematic reviews published between January 1st, 2000, and May 6th, 2018 (last date of search). The Montori method was used as a search strategy to identify systematic reviews (search details in Appendix 1) [18]. The search was restricted to imaging-specific journals based on the 2015 Thomas Institute of Science Information list of imaging journals [19], corresponding to 72 journals (Appendix 2). We did not use specific keywords filtering comparative imaging reviews. There were no restrictions on language and type of index test.

Comparative imaging reviews were included if the following criteria were met: (1) the diagnostic accuracy of at least two index tests (at least one of which must be an imaging test: CT, MRI, US, x-ray, nuclear medicine, mammography) were evaluated in human subjects against a reference standard; (2) if the objective of the review was comparative; and (3) if the review contained at least one sentence in which the review authors made a comparison between index tests regarding their relative diagnostic accuracy. Reviews were excluded if the full-text article was irretrievable, if the primary studies were not identifiable through the references, or if comparisons were made between different thresholds of the same index test.

A single investigator (AD) performed the title and abstract screen. Two investigators (AD and MA) independently assessed full texts of relevant studies for inclusion. Disagreements were resolved by discussion, or by consulting a third reviewer (MDFM).

Fig. 1 Comparative and non-comparative primary study designs


Non-comparative Primary Studies
Comparative Primary Studies
Single test vs. reference standard n = 3285
Non-randomized allocation n = 153
Randomized allocation n = 10
Within patient comparison n = 734
Study 1
Study 2
One Study
One Study
One Study
Series of patients
Series of patients
Series of patients
Two Different Groups of Patients
Series of patients
Not randomized to groups 1 and 2
Randomized to groups 1 and 2
Test A
Test B
Test A
Group 1
Group 2
Test A
Test A
Test B
Accuracy of Test A
Accuracy of Test B
< 90% undergo Test B
>= 90% undergo Test B
Accuracy of Test A and B
Compared in systematic review
Accuracy of Test A and B
Accuracy of Test A and B
Compared in systematic review
Compared in systematic review
Compared in systematic review

Data extraction from comparative imaging reviews

Two reviewers (any two of LT, ADS, JC, NA medical students; JPS, an epidemiology graduate student; and NK, MA medical residents) independently extracted the following data from the full text of included comparative imaging reviews: first author, journal, journal impact factor, publication year, target condition, index test evaluated, imaging modality/subspecialty, number of included primary studies, whether comparative primary study design was an inclusion criterion, and whether the inclusion of non-comparative primary studies was recognized as a limitation. The meta-analytic method for comparing the accuracy of the index tests was extracted and categorized as follows: (a) between meta-analysis (e.g., separate models fitted for each test and comparisons made between meta-analytic summaries using statistical tests); (b) within meta-analysis (e.g., test comparisons made within a single meta-analytic model, e.g., meta-regression); (c) none (e.g., informal comparisons or narrative reviews); (d) unclear (not reported). If either a separate analysis limited to comparative primary studies or a regression analysis with comparative primary study design was a covariate was performed, its impact on the results was recorded.

Data extraction from primary studies

For each included comparative imaging review, the abstracts (and if necessary full texts) of the primary studies were retrieved to extract the following: first author, publication year, and study design (described in Fig. 1). If < 90% of patients underwent both index tests, the study was classified as non-randomized allocation of patients to index tests. For reviews evaluating > 2 index tests, each pairwise test comparison was considered separately. For reviews with multiple comparisons, the largest meta-analysis was used for data extraction; for example, if a comparative imaging review included both per-lesion and per-patient meta-analyses, the analysis with the largest dataset was included. Primary studies were classified under randomized allocation if this was explicitly stated in the design of the study.

All data extraction was duplicated independently, with disagreements resolved through discussion and if necessary with a third reviewer (MDFM).

Data synthesis and analysis

Descriptive statistics were used to calculate the proportion of comparative imaging reviews relying on (a) only comparative primary studies, (b) only non-comparative primary studies, and (c) both comparative and non-comparative primary studies. Additionally, other review characteristics extracted were classified by review type (only comparative, only non-comparative, both comparative, and non-comparative). The proportion of each type of primary study design was examined by review type and publication year. Analyses were performed using SPSS software, version 24 [20].

Results

Selection of comparative imaging reviews

One-hundred three comparative imaging reviews were included, comprising 2280 unique primary studies. Screening and inclusion details are outlined in Fig. 2.

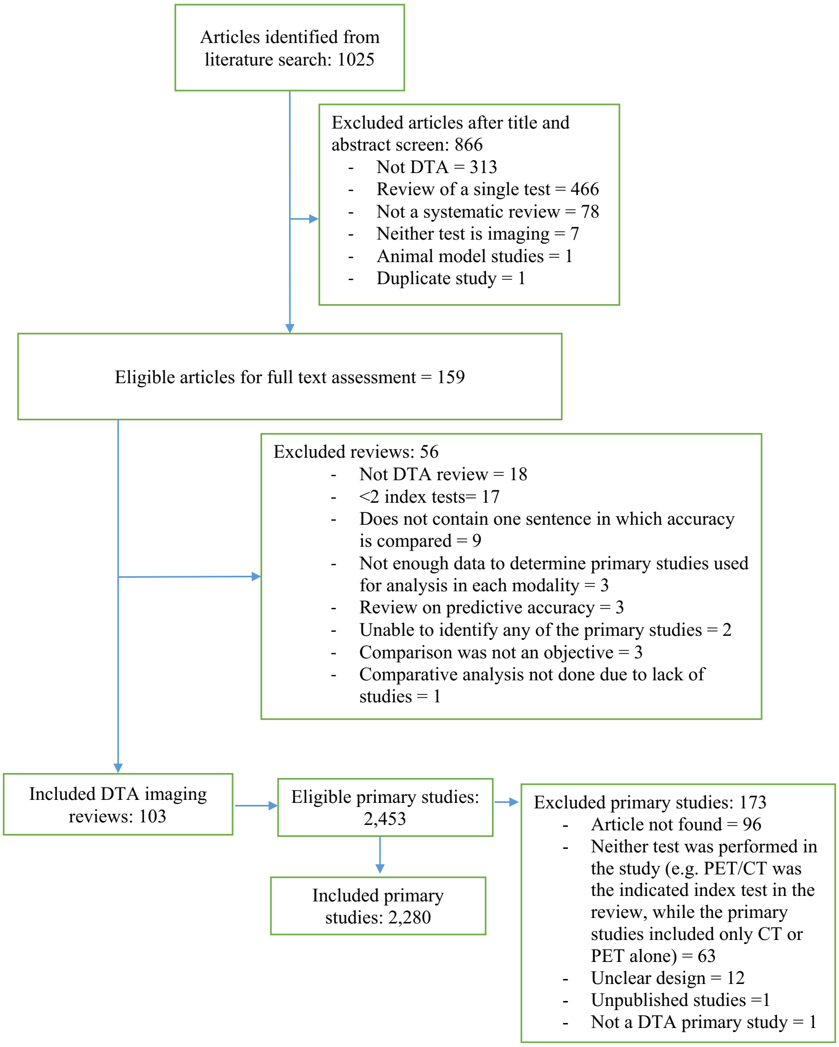
Comparative imaging review characteristics

Comparative imaging review characteristics are outlined in Table 1. Overall, 11/103 (11%) included only comparative primary studies, 12/103 (11%) included only non-comparative primary studies, and 80/103 (78%) included both comparative and non-comparative primary studies (Table 2). The median percentage of non-comparative primary studies for all reviews was 81 (IQR 38–93). For reviews containing both comparative and non-comparative primary studies, the median percentage of non-comparative primary studies was 81 (IQR 57–90%). Of the comparative imaging reviews that included non-comparative primary studies, 79/92 (86%) did not recognize the limitation of including non-comparative primary studies. Among comparative imaging reviews that included both comparative and non-comparative primary studies, 6/80 (8%) included a subgroup analysis for comparative primary studies. The impact of the subgroup on the final results of the review was not reported for 2 [21, 22], did not impact the final results for 3 [23–25], and changed the direction of results for 1 [26]. Comparative imaging review characteristics stratified by overall review type (i.e., based on the design of their included primary studies) are outlined in Table 2.

Primary study characteristics

Among the primary studies assessed, 3438/4182 (82%) were non-comparative and 744/4182 (18%) were comparative; these numbers are higher than the number of included primary studies (2280), because in multiple comparison reviews, primary studies were considered separately for each pairwise comparison. Of non-comparative primary studies, 3285/3438 (96%) were single test vs. reference standard studies and 153/3438 (4%) were non-randomized studies. Within the non-randomized primary studies, 94/153 (61%) conducted both index test in < 90% of all patients and 59/153 (39%) compared non-randomized groups of patients. Of comparative primary studies, 734/744 (99%) were within-patient comparisons and 10/744 (1%) were randomized. Detailed information regarding individual primary studies stratified by overall review types and their publication year are provided in Table 3.

Fig. 2 Selection of comparative imaging reviews


Articles identified from literature search: 1025
Excluded articles after title and abstract screen: 866
- Not DTA = 313
- Review of a single test = 466
- Not a systematic review = 78
- Neither test is imaging = 7
- Animal model studies = 1
- Duplicate study = 1
Eligible articles for full text assessment = 159
Excluded reviews: 56
- Not DTA review = 18
- <2 index tests= 17
- Does not contain one sentence in which accuracy is compared = 9
- Not enough data to determine primary studies used for analysis in each modality = 3
- Review on predictive accuracy = 3
- Unable to identify any of the primary studies = 2
- Comparison was not an objective = 3
- Comparative analysis not done due to lack of studies = 1
Included DTA imaging reviews: 103
Eligible primary studies: 2,453
Included primary studies: 2,280
Excluded primary studies: 173
- Article not found = 96
- Neither test was performed in the study (e.g. PET/CT was the indicated index test in the review, while the primary studies included only CT or PET alone) = 63
- Unclear design = 12
- Unpublished studies =1
- Not a DTA primary study = 1

Discussion

This methodologic review of 103 comparative imaging systematic reviews identified that a minority (11%) include only primary studies with comparative design. Among the remaining comparative imaging systematic reviews, primary studies with non-comparative design comprised the majority of included studies (> 80%); few of them recognized the risk of bias conferred by the inclusion of primary studies with non-comparative design. The bias introduced by including primary studies with non-comparative designs raises major concerns regarding the validity of conclusions that can be drawn from these comparative imaging reviews.

Our findings are consistent with a previously published study on general diagnostic accuracy comparative reviews (including non-imaging topics) in which 11% of comparative reviews included only primary studies with comparative design [9]. However, our finding that 82% of primary studies included in comparative accuracy imaging systematic reviews were non-comparative in design is higher than reported by Takwoingi et al, who identified that 69% of included primary studies were non-comparative [9]. This discrepancy may be related to the relative difficulty in conducting index tests; for example, conducting multiple imaging tests on study participants is likely more logistically challenging than multiple laboratory tests.

In our cohort, there was a higher proportion of multiple index test comparisons in reviews that included primary studies with non-comparative design; all reviews that included only comparative design primary studies evaluated only two index tests. The increased likelihood of including non-comparative primary studies in multiple comparison reviews illustrates the difficult balance between limiting inclusion to comparative design primary studies (to minimize the risk of bias) vs. allowing all primary study-design types to include more studies (higher precision) in the comparative imaging review. Authors evaluating more than two index tests should consider techniques such as network meta-analysis which may overcome some of the bias from non-comparative primary studies [27].

Our findings raise concern regarding meta-analytic approaches used in comparative imaging reviews. Most reviews

Table 1 Characteristics of included comparative imaging reviews

Characteristics	Number
Year of publication, median (IQR)	2013 (2010–2017)
Total journals	25
Radiology	16 (15)
European Radiology	15 (15)
European Journal of Radiology	12 (12)
Other	60 (58)
Journal impact factor, median (IQR)	2.57 (2.13–3.97)
Imaging modalities, *n* (%)[a]	
MRI	56 (54)
CT	52 (50)
Nuclear medicine	41 (40)
US	36 (35)
Plain films	7 (7)
Mammography	3 (3)
Fluoroscopy	2 (2)
Interventional radiology	1 (1)
Other	9 (9)
Target conditions	60
Hepatic lesions	7 (7)
Coronary artery disease	5 (5)
Breast cancer	5 (5)
Bone metastases	5 (5)
Appendicitis	4 (4)
Pulmonary nodule	4 (4)
Prostate cancer	3 (3)
Ovarian carcinoma	3 (3)
Thyroid lesion	3 (3)
Pulmonary embolism	2 (2)
Cholecystitis	2 (2)
Pancreatic cancer	2 (2)
Acetabular tears	2 (2)
Other	56 (54)
Imaging specialties	13
Oncologic imaging	35 (34)
Gastrointestinal	17 (17)
Musculoskeletal	11 (11)
Chest	7 (7)
Cardiac	6 (6)
Breast	5 (5)
Other	22 (21)

[a] Multiple modalities are reported per review

(80%) reported "between" meta-analysis comparisons (separate meta-analyses for each index test). The comparison of accuracy estimates from these separate meta-analyses is problematic since it does not account for both within- and between-study variability in one statistical model [13]. Authors should consider methods such as hierarchical meta-regression (including test type as a covariate) which would better account for these sources of variability and allow more valid comparisons of accuracy [28].

In order to increase the quality and validity of comparative imaging reviews, more primary studies with comparative design are needed [14]. Within-patient comparison studies (all patients receive both index tests) may be challenging to conduct. For example, randomization may be unethical when one of the diagnostic pathways may lead to earlier adequate therapy [29]. Furthermore, conducting more than one index test in one population may be challenging due to the high cost of imaging and potential for increased time to treatment. Certain procedural techniques may also not be technically feasible to perform twice on one patient; for example, it may not be feasible to obtain image-guided biopsies of one lesion twice [29]. Despite this, randomization of patients to index tests in comparative primary studies should be considered as an alternate high-quality design. Randomization may decrease the logistic challenges of having all patients undergo two tests and may be appropriate and ethical in scenarios where the relative accuracy of the two tests is considered comparable. Interestingly, we identified only ten primary studies that used randomized allocation indicating that this design may be underutilized.

The median year of publication for primary studies with comparative design was higher (2007) than that for primary studies with non-comparative design (2003). This may represent progress related to efforts to promote high-quality comparative research [30]. Recent initiatives such as courses and workshops on comparative research by national and international agencies, such as the RSNA Value of Imaging through Comparative Effectiveness (VOICE) program, should further promote optimal methods for comparative accuracy primary studies and systematic reviews [5, 30].

The goal of comparative accuracy reviews should be to allow valid comparisons between imaging tests. Valid comparisons are not feasible unless review authors are aware of the risk of bias from inclusion of primary studies with non-comparative design. Using non-comparative designs to compare two index tests may be futile in terms of cost and time as decisions made in clinical practice from these reviews may be confounded. Non-comparative primary studies may have used the index tests in various clinical contexts and diagnostic pathways, making comparisons between the two tests inaccurate [9]. To address this source of confounding, review authors can consider only including primary studies with comparative design; however, given the relative lack of these studies in the imaging literature, this may limit sample size. Alternatively, if primary studies with non-comparative design are included, statistical methods to explore the impact primary studies design (e.g., meta-regression using study design as a covariate) should be considered [9].

Various guidelines have been developed to improve the conduct and reporting of systematic reviews that compare

Table 2 Characteristics of included reviews stratified by type of included studies

	Only comparative studies[b]	Only non-comparative studies[c]	Both comparative and non-comparative studies[d]	Total
Reviews, *n* (%)	11 (11)	12 (11)	80 (78)	103 (100)
Number of primary studies per review				
Median (range)	7 (4–12)	17 (11–42)	24 (3–90)	19 (3–90)
IQR	6–11	13–26	13–39	12–33
Number of index tests, *n* (%)				
2	11 (100)	11 (92)	41 (51)	63 (61)
3	0	1 (8)	22 (28)	23 (22)
4	0	0	11 (14)	11 (11)
≥ 5	0	0	6 (7)	6 (6)
Year of publication, *n* (%)[a]				
Median (range)	2014 (2007–2018)	2014 (2000–2016)	2013 (2002–2018)	2013 (2000–2018)
IQR	2008–2018	2012–2016	2010–2017	2010–2017
Comparative primary studies as an inclusion criterion, *n* (%)	10 (91)	1 (9)	6 (7)	17 (17)
Reference standard, *n* (%)				
Composite (pathology and/or follow-up)	2 (18)	3 (25)	32 (40)	37 (36)
Imaging and/or clinical follow-up	1 (9)	3 (25)	14 (17)	18 (17)
Pathology	4 (36)	4 (33)	17 (21)	25 (24)
Surgery/intra-operative	1 (9)	0	4 (5)	5 (5)
Not specified	1 (9)	1 (8)	4 (5)	6 (6)
Other	2 (18)	1 (8)	9 (11)	12 (12)
Method, *n* (%)				
Between meta-analysis	9 (82)	10 (83)	63 (79)	82 (80)
Within meta-analysis	0	1 (8)	3 (4)	4 (4)
None	2 (18)	1 (8)	14 (17)	17 (16)

[a] By total median split

[b] Only comparative refers to comparative imaging reviews containing only comparative primary studies

[c] Only non-comparative refers to comparative imaging reviews containing only non-comparative primary studies

[d] Both non-comparative and comparative refers to comparative imaging reviews containing a combination of comparative and non-comparative primary studies

Table 3 Distribution of primary studies across overall review categories based on included primary studies (only comparative, only non-comparative, and combination) and primary study publication year

	Review categories			Publication year
Primary study	Only comparative, *n* (%)	Only non-comparative, *n* (%)	Both comparative and non-comparative, *n* (%)	Median (range) IQR
Non-comparative primary study				
Single test vs. reference standard	0	263 (99)	3022 (79)	2003 (1972–2017) 1997–2008
Non-randomized allocation to index tests	0	3 (1)	150 (4)	2007 (1987–2016) 2001–2010
Total non-comparative	0	266 (100)	3172 (83)	2003 (1972–2017) 1997–2008
Comparative primary study				
Within-patient comparison of index tests[a]	80 (100)	0	654 (17)	2007 (1979–2018) 2002–2010
Randomized allocation to index tests	0	0	10 (0.003)	2009 (2000–2012) 2002–2010
Total comparative	80 (100)	0	664 (17)	2007 (1979–2018) 2002–2010

[a] Within-patient comparison of index tests refers to a primary study where both index tests were performed in one patient as well as the reference standard

index tests. The Cochrane Collaboration's Handbook for Systematic Reviews of Diagnostic Test Accuracy provides further guidance for review authors on appropriate analysis and interpretation of results for DTA comparative systematic reviews [31]. Furthermore, the PRISMA-DTA checklist facilitates more comprehensive and transparent reporting of reviews to contextualize conclusions regarding index test superiority in comparisons [32, 33]. However, some tools designed for DTA systematic reviews are not optimized for comparative accuracy (e.g., QUADAS-2 for assessing risk of bias in included studies); further work to tailor these for comparative reviews would be helpful [34].

This study has several strengths. The search included reviews published since 2000 and a broad set of imaging journals. Pilot extraction forms were used in addition to duplicate independent extraction to reduce variability. The data extraction included several thousand primary studies and used contemporary guidance regarding study design classification.

This work has limitations. The search was limited to imaging journals only; imaging reviews published in non-imaging journals were not included. The analysis did not explore other primary study characteristics which could further bias comparative reviews; for example, factors known to confer a high risk of bias such as blinding of the index tests to each other and to the reference standard were not considered [32, 34]. Evaluation of further sources of bias in comparative imaging reviews is warranted in future research. Furthermore, a 90% threshold for studies that did not implement both index tests in the study sample was arbitrary. However, given the small number of studies in this category, it is unlikely to significantly impact our findings. In addition, it was important to identify these studies to illustrate the discrepancies between our definition of comparative studies and those used in the included reviews.

In conclusion, our findings illustrate challenges in comparative imaging reviews: (1) a minority of the reviews rely exclusively on primary studies with comparative design; (2) the majority of primary studies included in comparative accuracy reviews are non-comparative in design; and (3) awareness of the risk of bias conferred from inclusion of primary studies with non-comparative design is low. The recent prioritization of comparative accuracy data and simultaneous minimal awareness of valid comparative designs may lead to counterproductive research and inadequately supported clinical decisions for diagnostic tests and subsequent treatment. Given the rapid growth of systematic reviews in imaging journals and the fundamental role of imaging in clinical practice, it is vital that high-quality comparative accuracy imaging research is encouraged in primary studies, and that the risk of bias from including non-comparative primary studies in reviews is better understood and acknowledged [35, 36].

Funding This study has received funding from the University of Ottawa, Department of Radiology Research Stipend Program.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Matthew McInnes.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry Several authors have significant statistical expertise (Drs McInnes, Leeflang, Deeks).

Informed consent Written informed consent was not required for this study because evaluation of published literature is N/A.

Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was not required because this is an evaluation of published literature.

Methodology
- retrospective
- cross-sectional study
- multicenter study

References

1. Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care (2009) Learning what works: infrastructure required for comparative effectiveness research: workshop summary. Appendix C, Comparative Effectiveness Research Priorities: IOM Recommendations Washington, DC: National Academies Press (US). Available via https://www.nap.edu/read/12214/chapter/2. Accessed 11 Oct 2018
2. Godlee F (2010) More research is needed - but what type? BMJ 341:c4662
3. Comparative Effectiveness Research Prioritization: National Academies of Sciences, Engineering, Medicine. Available via http://www.nationalacademies.org/hmd/Activities/Research/CERPriorities.aspx. Accessed 13 Aug 2018
4. America RSoN. RSNA/ASNR comparative effectiveness research training (CERT) program. Available via https://www.rsna.org/education/workshops/comparative-effectiveness-research-training. Accessed 11 Oct 2018
5. A collaborative training program in Biomedical Big Data and Comparative Effectiveness Research (2018) Value of Imaging through Comparative Effectiveness (VOICE)
6. National Institute for Health and Care Excellence (NICE) (2013) Guide to the methods of technology appraisal. NICE process and methods guides
7. Concato J, Shah N, Horwitz RI (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342(25):1887–1892
8. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM (2008) Systematic reviews of diagnostic test accuracy. Ann Intern Med 149(12):889–897
9. Takwoingi Y, Leeflang MM, Deeks JJ (2013) Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med 158(7):544–554
10. Sutton A, Ades AE, Cooper N, Abrams K (2008) Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics 26(9):753–767

11. Lumley T (2002) Network meta-analysis for indirect treatment comparisons. Stat Med 21(16):2313–2324
12. Bossuyt PM, Irwig L, Craig J, Glasziou P (2006) Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 332(7549):1089–1092
13. Dinnes J, Deeks J, Kirby J, Roderick P (2005) A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess 9(12):1–113 iii
14. Leeflang MMG, Reitsma JB (2018) Systematic reviews and meta-analyses addressing comparative test accuracy questions. Diagn Progn Re 2(17)
15. Zhou X-H, Obuchowski NA, McClish DK (2011) Statistical methods in diagnostic medicine. John Wiley & Sons, Hoboken. https://doi.org/10.1002/9780470906514
16. Leeflang M, Nisio M, Rutjes A, Zwinderman AH, Bossuyt P (2011) Adjusting for indirectness in comparative test accuracy meta-analyses. Cochrane Database Syst Rev Supplement
17. Wang J, Bossuyt P, Geskus R et al (2015) Using individual patient data to adjust for indirectness did not successfully remove the bias in this case of comparative test accuracy. J Clin Epidemiol 68(3):290–298
18. Shojania KG, Bero LA (2001) Taking advantage of the explosion of systematic reviews: an efficient MEDLINE search strategy. Eff Clin Pract 4(4):157–162
19. Web of Science: Clarivate Analytics. Available via https://login.webofknowledge.com/. Accessed 11 Oct 2018
20. IBM Statistics for Mac (2016). 24 ed: Corp IBM
21. Issa Y, Kempeneers MA, van Santvoort HC, Bollen TL, Bipat S, Boermeester MA (2017) Diagnostic performance of imaging modalities in chronic pancreatitis: a systematic review and meta-analysis. Eur Radiol 27(9):3820–3844
22. Kiewiet JJ, Leeuwenburgh MM, Bipat S, Bossuyt PM, Stoker J, Boermeester MA (2012) A systematic review and meta-analysis of diagnostic performance of imaging in acute cholecystitis. Radiology 264(3):708–720
23. Laméris W, van Randen A, Bipat S, Bossuyt PM, Boermeester MA, Stoker J (2008) Graded compression ultrasonography and computed tomography in acute colonic diverticulitis: meta-analysis of test accuracy. Eur Radiol 18(11):2498–2511
24. Vilgrain V, Esvan M, Ronot M, Caumont-Prim A, Aubé C, Chatellier G (2016) A meta-analysis of diffusion-weighted and gadoxetic acid-enhanced MR imaging for the detection of liver metastases. Eur Radiol 26(12):4595–4615
25. Wang Z, Wang Y, Sui X et al (2015) Performance of FLT-PET for pulmonary lesion diagnosis compared with traditional FDG-PET: a meta-analysis. Eur J Radiol 84(7):1371–1377
26. Berger N, Luparia A, Di Leo G et al (2017) Diagnostic performance of MRI versus galactography in women with pathologic nipple discharge: a systematic review and meta-analysis. AJR Am J Roentgenol 209(2):465–471
27. McGrath TA, Bossuyt PM, Cronin P et al (2018) Best practices for MRI systematic reviews and meta-analyses. J Magn Reson Imaging. https://doi.org/10.1002/jmri.26198
28. Rutter CM, Gatsonis CA (2001) A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 20(19):2865–2884
29. Bossuyt PM, Lijmer JG, Mol BW (2000) Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 356(9244):1844–1847
30. Kang SK, Rawson JV, Recht MP (2018) Supporting imagers' VOICE: a national training program in comparative effectiveness research and big data analytics. J Am Coll Radiol 15(10):1451–1454
31. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y (2010) Chapter 10: Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C (eds) Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy The Cochrane Collaboration
32. McInnes MDF, Moher D, Thombs BD et al (2018) Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 319(4):388–396
33. Frank RA, Bossuyt PM, McInnes MDF (2018) Systematic reviews and meta-analyses of diagnostic test accuracy: the PRISMA-DTA statement. Radiology. https://doi.org/10.1148/radiol.2018180850
34. Whiting PF, Rutjes AW, Westwood ME et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155(8):529–536
35. Alabousi M, Alabousi A, McGrath TA et al (2018) Epidemiology of systematic reviews in imaging journals: evaluation of publication trends and sustainability? Eur Radiol. https://doi.org/10.1007/s00330-018-5567-z
36. Pandharipande PV, Gazelle GS (2009) Comparative effectiveness research: what it means for radiology. Radiology 253(3):600–605

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.