DISSERTATION

Qualität der Berichterstattung diagnostischer Genauigkeitsstudien in der radiologischen Literatur

Quality of reporting of diagnostic accuracy studies published in radiological medical journals

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von
Ann-Christine Stahl

Erstbetreuung: Prof. Dr. med. Dr. h.c. Marc Dewey

Datum der Promotion: 29. November 2024

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

STARD	Standards for Reporting Diagnostic Accuracy
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses (bevorzugten Report Items für systematische Übersichten und Meta-Analysen)
IQA	Interquartilsabstand
95% KI	95% Konfidenzintervall

Zusammenfassung

Hintergrund: Diagnostische Genauigkeitsstudien weisen häufig Mängel in der Qualität ihrer Berichterstattung auf. *European Radiology* hat 2017 eine Empfehlung zur Verwendung der Standards for Reporting Diagnostic Accuracy (STARD) in ihren Einreichungsrichtlinien ergänzt, während *Radiology* die Verwendung der STARD-Checkliste 2016 verpflichtend gemacht hat. In dieser Arbeit haben wir die beiden Änderungen in den Einreichungsrichtlinien zum Anlass genommen, zu analysieren, ob eine Empfehlung oder Verpflichtung zur Verwendung der STARD-Checkliste zu einer Verbesserung der Qualität der Berichterstattung diagnostischer Genauigkeitsstudien geführt hat.

Methoden: Mit einer validierten Suchstrategie haben wir MEDLINE via PubMed nach diagnostischen Genauigkeitsstudien durchsucht, die in den Jahren 2015 und 2019 in den Journalen *European Radiology* und *Radiology* veröffentlicht wurden. Zwei Auswerterinnen haben mit Hilfe der aktuellen STARD-Checkliste unabhängig voneinander die Qualität der Berichterstattung der eingeschlossen Studien geprüft. Wurde ein Item der STARD-Checkliste suffizient wiedergegeben, konnte ein Punkt vergeben werden. Um einen STARD-Gesamtscore zu berechnen, wurden alle erzielten Punkte einer Studie addiert. Item 11 (Grund für die Wahl des Referenzstandards [wenn Alternativen existieren]) musste für diese Analyse ausgeschlossen werden. Denn nicht in jeder klinischen Situation existieren Alternativen für den Referenzstandard. Den beiden Auswerterinnen war es nicht möglich einzuschätzen, ob Item 11 in einer Studie nicht berücksichtigt wurde, weil es keine Alternativen zum Referenzstandard gab oder weil es nicht erfüllt wurde. T-Tests und Wilcoxon-Mann-Whitney-Tests wurden genutzt, um jeweils Unterschiede zwischen den beiden Jahren innerhalb eines Journals zu untersuchen sowie zwischen den beiden Journalen selbst.

Ergebnisse: 180 diagnostische Genauigkeitsstudien erfüllten die Einschlusskriterien. Der Mittelwert des STARD-Gesamtscores der eingeschlossenen Studien betrug 16,9 ± 2,9 von 29 Items (58,3%; Spannweite 9,5 - 24,5). In beiden Journalen verbesserte sich die Qualität der Berichterstattung mit den Änderungen der Einreichungsrichtlinien signifikant (*European Radiology:* 16,3 ± 2,7 versus 15,1 ± 2,3; $p < 0,02$; *Radiology:* 19,5 [IQA 18,5 - 21,5] versus 18,0 [IQA 15,5 - 19,5]; $p < 0,001$; Varghas und Delaneys A = 0,24).

Die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien war in *Radiology* (18,6 ± 2,4) signifikant besser als in *European Radiology* (15,9 ± 2,6; $p < 0,01$).

Fazit: Die Qualität der Berichterstattung ist in zwei führenden Journalen der radiologischen Literatur nach wie vor moderat. Die Empfehlung und Verpflichtung zur Verwendung der STARD-Checkliste hat in beiden Journalen zu einer Verbesserung der Qualität der Berichterstattung geführt. Herausgeber*innen können mit ergänzenden Verpflichtungen zur Verwendung bestimmter Reporting Guidelines einen Beitrag zur Verbesserung der Qualität der Berichterstattung leisten.

Abstract

Background: Diagnostic accuracy studies are often judged to have a poor quality of reporting. *European Radiology* started recommending the use of the Standards for Reporting Diagnostic Accuracy (STARD) within its submission guidelines in 2017, while *Radiology* made the use of the STARD checklist mandatory for its submissions in 2016. We used these changes to the submission guidelines as an opportunity to analyze if recommending the use of the STARD checklist or making it mandatory has improved the quality of reporting of diagnostic accuracy studies.

Methods: MEDLINE (via PubMed) was searched for diagnostic accuracy studies published in *European Radiology* and *Radiology* with a validated search strategy. The search was limited to studies published in 2015 and 2019. Two independent reviewers assessed the quality of reporting of all included studies with the help of the updated STARD checklist. If an item was adequately reported, it scored a point. To calculate the total STARD score of each article the scored points were summed up. Item 11 (rationale for choosing the reference standard [if alternatives exist]) was excluded for this analysis because the reviewers were not able to determine whether the item was not reported because no alternatives exist or because the authors did not mention it. Student's t tests for independent samples as well as Wilcoxon–Mann–Whitney tests were used to analyze differences in the number of reported STARD items between studies published in 2015 and in 2019 within each journal. Differences between both journals were analyzed with student's t test as well.

Results: 180 diagnostic accuracy studies met the inclusion criteria. The mean total number of reported STARD items of all included studies was 16.9 ± 2.9 of 29 items (58.3%, range 9.5 - 24.5). The quality of reporting of diagnostic accuracy studies was significantly better after the changes in the submission guidelines in both journals (*European Radiology:* 16.3 ± 2.7 versus 15.1 ± 2.3, $p < 0.02$, *Radiology:* 19.5 [IQR 18.5 - 21.5] versus 18.0 [IQR 15.5 - 19.5], $p < 0.001$, Vargha and Delaney's A = 0.24). The quality of reporting of diagnostic accuracy studies was significantly better in *Radiology* (18.6 ± 2.4) than in *European Radiology* (15.9 ± 2.6, $p < 0.01$).

Conclusion: The quality of reporting of diagnostic accuracy studies published in two radiological medical journals is still moderate. Encouraging authors to follow the STARD checklist has improved the quality of reporting as well as making the use of the STARD checklist mandatory. Editors should consider adding mandatory reporting guidelines to the submission guidelines of their journals to make a contribution to improving the quality of reporting.

Einleitung

1.1 Diagnostische Genauigkeitsstudien

Diagnostische Genauigkeitsstudien vergleichen die Ergebnisse des Indextests mit denen des Referenzstandards (1, 2). Der Indextest stellt dabei die zu untersuchende Methode da, während der Referenzstandard die sensitivste und spezifischste bekannte sowie zur Verfügung stehende Methode zur Diagnosesicherung der untersuchten Erkrankung bildet (2). Dieser kann dabei ein einziger Test sein oder sich aus mehreren Testverfahren zusammensetzen (2). Die untersuchte Erkrankung kann eine bestimmte Krankheit darstellen oder einen veränderten klinischen Zustand beschreiben, der weitere Schritte, wie diagnostische Tests oder therapeutische Konsequenzen, notwendig macht (2). Der Begriff Genauigkeit beschreibt in diesen Studien den Grad an Übereinstimmung zwischen den Ergebnissen des Indextests und denen des Referenzstandards (2). Angegeben werden kann diese beispielsweise durch die Sensitivität, Spezifität, Likelihood-Ratio, den Vorhersagewert oder die Fläche unter der Grenzwertoptimierungskurve (2, 3).

Die Bedeutung von diagnostischen Genauigkeitsstudien hat in den vergangenen Jahren stark zugenommen (4). Durch die stetige Entwicklung von neuen diagnostischen Testverfahren ist es immer wieder notwendig, die Genauigkeit dieser Verfahren valide zu prüfen (4-6). Problematisch ist dabei jedoch, dass diagnostische Genauigkeitsstudien oftmals anfällig für Bias und Variation sind (7). Besonders davon betroffen sind unter anderem die Bereiche demographische Charakteristika, Prävalenz und Schwere der untersuchten Krankheit beziehungsweise des klinischen Zustandes, Clinical Review Bias sowie Variation der Beobachter*innen (7). Studienergebnisse, die durch beispielsweise Bias verzerrt wurden, münden möglicherweise in verfälschten Empfehlungen bezüglich des untersuchten Testverfahrens (Indextest) und schränken die allgemeine Gültigkeit der Aussagen der Studie ein (2, 8). Darüber hinaus schränkt eine schlechte Qualität der Berichterstattung die Beurteilung von Bias und Variation in diesen Studien massiv ein (8).

1.2 Standards for Reporting Diagnostic Accuracy

Um diesen Mängeln zu begegnen, wurde im Jahr 2003 das Standards for Reporting Diagnostic Accuracy (Standards für die Berichterstattung diagnostischer Genauigkeitsstudien) Statement (STARD-Statement) veröffentlicht (4). Das STARD-Statement umfasst eine Checkliste, die 25 Items beinhaltet und alle Bereiche einer Studie, wie den Abstract, die Einleitung, die Methodik, die Ergebnisse sowie die Diskussion, abdeckt (4). Zusätzlich wurde eine Vorlage für ein Flussdiagramm entwickelt, das Autor*innen von diagnostischen Genauigkeitsstudien helfen soll, ihre Selektion von in die Studie einzuschließenden Patient*innen mit Gründen für den Ausschluss darzustellen (4). Dabei sollen auch die Ergebnisse des Indextests berücksichtigt werden (4). 2015 wurde eine überarbeitete Version des STARD-Statements veröffentlicht (9, 10). Es wurde versucht, die einzelnen Items verständlicher zu beschreiben (9). Einige Items wurden zusammengefügt oder getrennt und andere Items wiederum wurden neu hinzugefügt (9). Die aktuelle Fassung des STARD-Statements umfasst nun eine Checkliste von 30 Items, die Autor*innen dabei unterstützen sollen, die wichtigsten Punkte ihrer Studie in ihrer Publikation wiederzugeben (9, 10). So sollen Bias und Variation besser erkannt, aber auch von vornherein vermieden werden, wenn die Checkliste bereits bei der Entwicklung einer Studie zur Rate gezogen wird (10).

Seit der Veröffentlichung des STARD-Statements haben diverse Studien die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien untersucht (11-16). Das überwiegende Ergebnis war, dass weiterhin Mängel bei der Berichterstattung bestehen und auch die Besserung über die Zeit nicht durchgreifend ist (12, 14). Wichtige Bereiche, die besonders davon betroffen sind, sind beispielsweise unerwünschte Ereignisse bei der Durchführung des Indextests oder Referenzstandards, Kalkulationen der beabsichtigten Stichprobengröße und die Studienregistrierung (16-18). Die bereits durchgeführten Studien haben aber auch herausgefunden, dass diagnostische Genauigkeitsstudien, die in Journalen veröffentlicht werden, die die Verwendung der STARD-Checkliste empfehlen oder voraussetzen, von einer vollständigeren Berichterstattung zeugen als solche, die in Journalen publiziert wurden, die von einer Empfehlung noch absehen (15, 17, 19).

1.3 Fragestellung

Im Jahr 2017 hat *European Radiology* die Empfehlung, die STARD-Richtlinien bei der Einreichung von diagnostischen Genauigkeitsstudien zu berücksichtigen, in den Einreichungsrichtlinien auf der *European Radiology* Webseite ergänzt. Diese Information haben wir auf Nachfrage via E-Mail vom Scientific Publications Department von *European Radiology* erhalten. *Radiology* hingegen hat die Verwendung der STARD-Checkliste im Januar 2016 für alle eingereichten diagnostischen Genauigkeitsstudien verpflichtend gemacht (20).

Wir haben diese Änderungen in den Einreichungsrichtlinien der beiden Journale zum Anlass genommen, folgende Fragestellungen zu untersuchen:

1. Sorgt die Empfehlung auf der *European Radiology* Webseite für eine bessere Qualität der Berichterstattung diagnostischer Genauigkeitsstudien?
2. Führt die Verpflichtung zur Verwendung der STARD-Checkliste in Studien publiziert in *Radiology* zu einer besseren Qualität der Berichterstattung?
3. In welchem Journal ist die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien besser?
4. Welche Bereiche sind besonders von einer mangelnden Qualität der Berichterstattung diagnostischer Genauigkeitsstudien betroffen beziehungsweise welche Bereiche werden bereits ausreichend erfüllt?

Dabei sind zwei Publikationen entstanden, deren Studienergebnisse die Grundlage des folgenden Manteltextes bilden. Die erste Publikation wurde unter dem Titel *„Has the STARD statement improved the quality of reporting of diagnostic accuracy studies published in European Radiology?“* im Januar 2023 in *European Radiology* veröffentlicht (21). Die zweite Publikation erschien im Mai 2023 mit dem Titel *„Has the Quality of Reporting Improved since it Became Mandatory to Use the Standards for Reporting Diagnostic Accuracy?“* in *Insights into Imaging* (22).

Methodik

Beide Studien dieser Arbeit konnten nicht im internationalen, prospektiven Register für systematische Übersichtsarbeiten, PROSPERO, registriert werden, weil sie nicht alle notwendigen Kriterien erfüllten (23). Denn PROSPERO akzeptiert ausschließlich systematische Übersichtsarbeiten, die ein gesundheitsbezogenes Ergebnis (Outcome) untersuchen (23), was unsere Arbeiten nicht erfüllten. Dies wurde uns auf Nachfrage via E-Mail an das Team von PROSPERO bestätigt, weswegen wir keinen Antrag auf Registrierung gestellt haben. Dennoch wurden in beiden Studien die Punkte der Checkliste der „bevorzugten Report Items für systematische Übersichten und Meta-Analysen“ (PRISMA-Statement) berücksichtigt (24).

2.1 Studiensuche

Um diagnostische Genauigkeitsstudien zu selektieren, die in den Jahren 2015 und 2019 in den Journalen *European Radiology* und *Radiology* veröffentlicht wurden, haben wir sowohl auf MEDLINE als auch auf den jeweiligen Webseiten der Journale nach potenziell einzuschließenden diagnostischen Genauigkeitsstudien gesucht. Die Datenbank MEDLINE haben wir über PubMed mit folgender validierter Suchstrategie durchsucht (11, 25): „sensitivity AND specificity.sh” OR „specificity.tw” OR „false negative.tw” OR „accuracy.tw”. Dabei steht „.sh” für „MEDLINE subheading“ und „.tw” für „text word“. Außerdem galt für unsere Suche die folgende Zeitraumlimitation: 2015/1/1 bis 2015/12/31 AND 2019/1/1 bis 2019/12/31. Ergänzend haben wir die Webseiten der Journale jeweils manuell mit dem Suchbegriff „Diagnostic accuracy studies“ durchsucht. PubMed wurde zuletzt am 8. April 2020 von uns durchsucht und die Webseiten der beiden Journale zuletzt am 23. Juni 2020.

2.2 Studienselektion

Zunächst haben zwei Auswerterinnen (A.S., eine fortgeschrittene Medizinstudentin mit drei Jahren Erfahrung in der Recherche und Analyse von diagnostischen Genauigkeitsstudien sowie A.T., eine Zahnärztin mit einem Jahr Erfahrung in der Recherche und Analyse von diagnostischen Genauigkeitsstudien) unabhängig voneinander die Titel, Abstracts und Keywords der Studien nach einzuschließenden diagnostischen Genauigkeitsstudien durchsucht. Anschließend wurden die Studien im Volltext erneut unabhän-

gig voneinander bezüglich eines Einschlusses beurteilt. Konnten sich die beiden Auswerterinnen dabei nicht einigen, hat ein dritter Auswerter (B.K., ein Arzt mit acht Jahren Erfahrung in der radiologischen Forschung) die finale Entscheidung getroffen.

Artikel wurden eingeschlossen, wenn eine klinische Studienkohorte (keine Tiere, Feten [in und ex utero], Leichen, Phantome oder Entwicklung von Modellen) untersucht wurde und sie mindestens ein Maß für diagnostische Genauigkeit wie Sensitivität, Spezifität, Likelihood-Ratio, Vorhersagewert, Genauigkeit oder Fläche unter der Grenzwertoptimierungskurve verwendet haben. Ausgeschlossen haben wir systematische Übersichtsarbeiten, Meta-Analysen, Briefe, Richtlinien, Editorials (Leitartikel), Stellungnahmen, Kommentare, klinische Studien und Studien zur Vorhersagegenauigkeit.

2.2 Auswertung

Um die Qualität der Berichterstattung einschätzen zu können, wurden alle eingeschlossenen Studien mit Hilfe der aktuellen STARD-Checkliste analysiert. Die Checkliste umfasst 30 Items (9), wobei wir für diese Analyse Item 11 (Grund für die Wahl des Referenzstandards [wenn Alternativen existieren]) ausgeschlossen haben, da nicht in jeder klinischen Situation Alternativen für den Referenzstandard existieren (26). Den beiden Auswerterinnen war es folglich nicht möglich einzuschätzen, ob Item 11 von den Autor*innen in ihrer Studie nicht berücksichtigt wurde, weil es keine Alternativen zu ihrem gewählten Referenzstandard gab oder weil sie es schlichtweg nicht erwähnt und damit nicht erfüllt haben. Zudem wurde dieser Ansatz schon erfolgreich von Wilczynski et al. verfolgt (27). Folglich haben wir für unsere Arbeit eine Checkliste mit 29 Items verwendet.

Wenn die Studie ein Item der Checkliste erfüllt hat, das heißt, wenn die Beschreibung des geforderten Punktes ausreichend war, haben wir für das jeweilige Item einen Punkt vergeben. War die Beschreibung eines Items hingegen unzureichend oder wurde ein Item in der Studie gar nicht erwähnt, haben wir keinen Punkt vergeben. Hierbei sollte nicht die methodologische Qualität oder die Wahrscheinlichkeit für Bias, sondern lediglich die Qualität der Berichterstattung der Studie bewertet werden (9, 28). Die STARD-Checkliste umfasst auch Items, die sich aus zwei Unterpunkten zusammensetzen (Items 10, 12, 13 und 21) (9). Hierbei haben wir für einen Unterpunkt bei ausreichender

Berücksichtigung jeweils 0,5 Punkte vergeben, damit allen Items die gleiche Gewichtung zufiel.

Alle Auswerter*innen waren dem Publikationsjahr, dem Journal und den Autor*innen der zu analysierenden Studien gegenüber nicht verblindet. Die Auswertung erfolgte durch zwei Auswerterinnen (A.S. und A.T.) unabhängig voneinander. Ihre Ergebnisse wurden miteinander verglichen, um sich auf ein Ergebnis zu einigen. Wenn es zu keiner Einigung kommen konnte, hat ein dritter Auswerter (B.K.) die finale Entscheidung getroffen.

Neben den Punkten der STARD-Checkliste wurden auch das Publikationsjahr (Datum der gedruckten Version), das Studiendesign (Kohorten- versus Fall-Kontroll-Studie), die Methodik der Datensammlung (prospektiv versus retrospektiv) und die Zitierrate der jeweiligen Studie notiert. Die Zitierrate haben wir berechnet, indem wir die Anzahl an Zitaten, die ein Artikel bis zum 30. April 2021 (für Studien, die in *Radiology* veröffentlicht wurden) beziehungsweise bis zum 31. August 2021 (für Studien, die in *European Radiology* veröffentlicht wurden) erfahren hat, durch die Anzahl an Monaten, die seit seiner Veröffentlichung in der gedruckten Fassung vergangen sind, geteilt haben. So hatten Studien, die schon früher veröffentlicht wurden, keinen Vorteil gegenüber den neueren Studien. Als Quelle für die Anzahl an Zitaten diente die Datenbank *Web of Science* von Thomson Reuters.

Vor der eigentlichen Analyse der eingeschlossenen Studien haben beide Auswerterinnen (A.S. und A.T.) vier Studien aus den Jahren 2014 und 2020 der Journale *Radiology* und *European Radiology* zur Übung ausgewertet. So konnten wir sicherstellen, dass sie sich in der Definition aller Items der STARD-Checkliste einig waren.

2.3 Statistische Analyse

Für jede Studie haben wir den STARD-Gesamtscore berechnet, indem wir die Summe aller erfüllten Items berechnet haben. Dabei waren Werte von 0 bis 29 möglich. Da der Gesamtscore die Qualität der Berichterstattung der jeweiligen Studie abbildet, steht ein hoher Wert für eine gute bis sehr gute Qualität, während ein niedrige Wert im Gesamtscore auf Mängel in der Qualität der Berichterstattung hinweist (2).

Außerdem haben wir die Übereinstimmung der beiden Auswerterinnen mit dem Kappa-Koeffizient nach Cohen berechnet. Zur Einordnung des Ergebnisses haben wir uns an den Angaben von Landis und Koch orientiert, die einen Wert von < 0,00 als keine, einen Wert von 0,00 - 0,20 als kaum, einen Wert von 0,21 - 0,40 als geringe, einen Wert von 0,41 - 0,60 als moderate, einen Wert von 0,61 - 0,80 als erhebliche und einen Wert von 0,81 - 1,00 als fast perfekte Übereinstimmung zwischen den Auswerter*innen klassifiziert haben (29).

2.3.1 Erste Publikation

Mit Hilfe des Shapiro-Wilk-Tests haben wir die Daten auf eine Normalverteilung hin geprüft. Darauf konnten wir den Mittelwert ± die Standardabweichung und die Spannweite der erfüllten STARD-Items berechnen. Anhand des Publikationsdatums (Datum der gedruckten Version), des Studiendesigns (Kohorten- versus Fall-Kontroll-Studie), der Methodik der Datensammlung (prospektiv versus retrospektiv) und des Median-Splits der Zitierrate konnten wir die eingeschlossenen Studien jeweils in zwei Gruppen teilen. Um signifikante Unterschiede in diesen Gruppen aufzudecken, haben wir die Daten mit Hilfe des t-Tests untersucht.

Zudem haben wir für jedes Item der STARD-Checkliste die Anzahl der Studien berechnet, die das jeweilige Item erfüllt haben. Dabei konnte jedes Item Werte von 0 bis 114 erzielen und die Ergebnisse konnten wir für die Jahre 2015 und 2019 als Prozentwerte angeben.

Für die Zeit, die die beiden Auswerterinnen für die Analyse gebraucht haben, haben wir den Median und den Interquartilsabstand (IQA) berechnet, da diese Daten nicht normalverteilt sind, wie der Shapiro-Wilk-Test gezeigt hat. p-Werte unter 0,05 galten in dieser Analyse als signifikant. Die statistische Analyse wurde mit IBM SPSS Statistics (Version 27.0.0.0) durchgeführt.

2.3.2 Zweite Publikation

Erneut haben wir für jede Studie den STARD-Gesamtscore berechnet, indem wir die Summe aller erfüllten Items berechnet haben. Hierbei war eine Spannweite von 0 bis 29 möglich. Für diese Publikation haben wir den Median und Interquartilsabstand des STARD-Gesamtscores berechnet.

Mit dem Wilcoxon-Mann-Whitney-Test haben wir den STARD-Gesamtscore zwischen Studien aus dem Jahr 2015 sowie aus 2019 verglichen. Zum einen haben wir bei diesem Vergleich alle Studien miteingeschlossen, zum anderen haben wir diesen Vergleich in den folgenden Untergruppen durchgeführt: prospektive Studien, retrospektive Studien, Kohortenstudien, Fall-Kontroll-Studien, Studien mit einer Zitierrate über dem Median und Studien mit einer Zitierrate unter dem Median. Außerdem haben wir den Wilcoxon-Mann-Whitney-Test genutzt, um den STARD-Gesamtscore von allen Kohortenstudien mit dem aller Fall-Kontroll-Studien zu vergleichen sowie von allen prospektiven Studien mit dem aller retrospektiven Studien und von allen Studien mit einer Zitierrate über dem Median mit dem aller Studien mit einer Zitierrate unter dem Median. Die Effektstärken wurden nach Vargha und Delaney berechnet.

Auch für die zweite Publikation haben wir für jedes Item der STARD-Checkliste die Anzahl der Studien berechnet, die das jeweilige Item erfüllt haben. Jedes Item konnte dabei Werte von 0 bis 66 erzielen. Die Ergebnisse haben wir in Prozent angegeben. p-Werte unter 0,05 galten in dieser Analyse als signifikant. Für die statistische Analyse wurde die Programmiersprache R (Version 4.2.0) genutzt.

2.3.3 Vergleich beider Journale (nicht in den Publikationen enthalten)

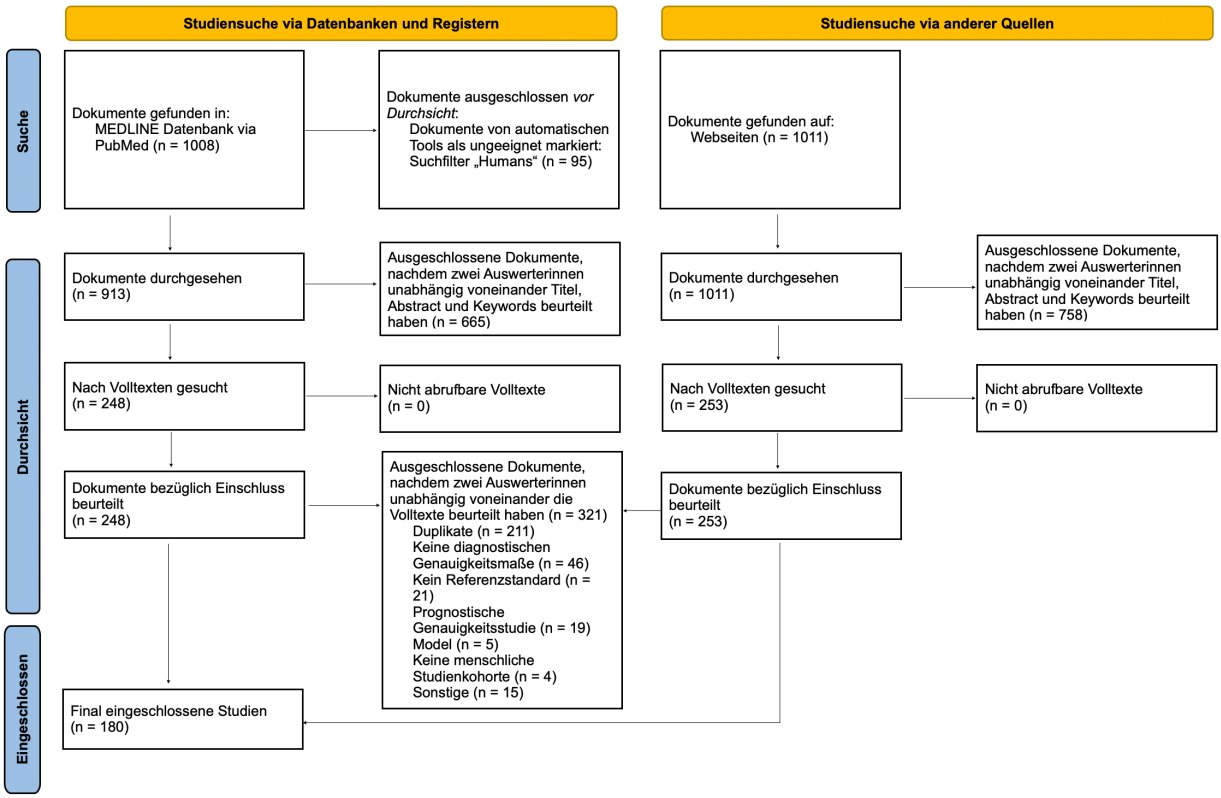
In Einklang mit der aktuellen Promotionsordnung haben wir ergänzend den STARD-Gesamtscore der beiden Journale miteinander verglichen. Diese Analyse ist in keiner der beschriebenen Publikationen enthalten, sondern wurde im Rahmen der Arbeit an dem vorliegenden Manteltext durchgeführt. Dafür haben wir die gesamten Daten mit dem Shapiro-Wilk-Test auf eine Normalverteilung getestet. Anschließend konnten wir den Mittelwert ± die Standardabweichung und die Spannweite der erfüllten STARD-Items aller 180 Studien berechnen. Einen signifikanten Unterschied zwischen beiden Journalen haben wir mit Hilfe des t-Tests untersucht. Dabei galten auch hier Werte unter 0,05 als signifikant. Auch für alle 180 Studien gemeinsam haben wir die Anzahl der Studien berechnet, die ein bestimmtes Item erfüllt haben.

Ergebnisse

3.1 Studienselektion

Der Such- und Selektionsprozess der diagnostischen Genauigkeitsstudien ist für die Studien beider Journale gemeinsam im PRISMA-Flussdiagramm in Abbildung 1 dargestellt (30). Die Suche in der Datenbank MEDLINE durch PubMed ergab 719 Treffer für *European Radiology* und 289 Treffer für *Radiology*. Die manuelle Suche auf der *European Radiology* Webseite erzielte hingegen 657 Treffer, während die Suche auf der *Radiology* Webseite 354 Treffer ergab. 1423 Artikel wurden von zwei Auswerterinnen (A.S. und A.T.) unabhängig voneinander nach Durchsicht von Titel, Abstract und Keywords aus verschiedenen Gründen ausgeschlossen. Der häufigste Grund dabei war das Fehlen von diagnostischen Genauigkeitsmaßen (n = 384). Anschließend konnten 211 Duplikate ausgeschlossen werden. Es blieben 290 Artikel für die Durchsicht des Volltextes. Hierbei haben die beiden Auswerterinnen 110 Studien unabhängig voneinander ausgeschlossen. Gründe hierfür waren beispielsweise das Fehlen von diagnostischen Genauigkeitsmaßen (n = 46), von menschlichen Studienkohorten (n = 4), eines Referenzstandards (n = 21), prognostische Genauigkeitsstudien (n = 19) oder die Entwicklung eines Models oder Algorithmus' während der Studie (n = 5). Schließlich erfüllten 180 Artikel die Selektionskriterien und konnten für die Analyse der Qualität der Berichterstattung in unsere Studien eingeschlossen werden.

Abbildung 1: PRISMA-Flussdiagramm der Studienselektion beider Journale (*European Radiology* und *Radiology*)



Anmerkung: Eigene Darstellung der Daten aus Stahl et al., 2023 (21) und Stahl et al., 2023 (22): Ann-Christine Stahl.

Der Median der Zitierrate der eingeschlossenen Studien von *European Radiology* lag bei 0,28 Zitaten pro Monat. In den eingeschlossenen Studien von *Radiology* lag der Median der Zitierrate hingegen bei 0,56 Zitaten pro Monat und war damit doppelt so hoch wie der der Studien aus *European Radiology*. Weitere Studienmerkmale sind für die Studien aus *European Radiology* (erste Publikation) in Tabelle 1 und für die Studien aus *Radiology* (zweite Publikation) in Tabelle 2 jeweils im Vergleich von 2015 zu 2019 dargestellt.

Tabelle 1: Charakteristika der Studien aus *European Radiology* (erste Publikation)

Studiencharakteristika	**2015**	**2019**
Eingeschlossene Studien insgesamt	42 (100%)	72 (100%)
Studiendesign		
Kohortenstudie Fall-Kontroll-Studie	36 (85,7%) 6 (14,3%)	61 (84,7%) 11 (15,3%)
Methodik der Datensammlung		
Prospektiv Retrospektiv	23 (54,8%) 19 (45,2%)	33 (45,8%) 39 (54,2%)
Zitierrate (Median-Split)		
Niedrig (≤ 0,28 Zitate/Monat) Hoch (> 0,28 Zitate/Monat)	25 (59,5%) 17 (40,5%)	32 (44,4%) 40 (55,6%)

Anmerkung: Übersetzt und modifiziert nach Stahl et al., 2023 (21).

Tabelle 2: Charakteristika der Studien aus *Radiology* (zweite Publikation)

Studiencharakteristika	**2015**	**2019**
Eingeschlossene Studien insgesamt	39 (100%)	27 (100%)
Studiendesign		
Kohortenstudie Fall-Kontroll-Studie	34 (87,2%) 5 (12,8%)	23 (85,2%) 4 (14,8%)
Methodik der Datensammlung		
Prospektiv Retrospektiv	23 (59,0%) 16 (41,0%)	11 (40,7%) 16 (59,3%)
Zitierrate (Median-Split)		
Niedrig (< 0,56 Zitate/Monat) Hoch (≥ 0,56 Zitate/Monat)	21 (53,8%) 18 (46,2%)	12 (44,4%) 15 (55,6%)

Anmerkung: Übersetzt und modifiziert nach Stahl et al., 2023 (22).

3.2 Qualität der Berichterstattung diagnostischer Genauigkeitsstudien

3.2.1 Erste Publikation

Der Mittelwert der erfüllten STARD-Items aller 114 eingeschlossenen diagnostischen Genauigkeitsstudien betrug 15,9 ± 2,6 von 29 Items (54,8%; Spannweite 9,5 - 22,5). Die Übereinstimmung der beiden Auswerterinnen beim Bewerten der einzelnen STARD-Items war 86,3%. Der Kappa-Koeffizient nach Cohen betrug 0,58 (95% KI [95% Konfidenzintervall] 0,49 - 0,68) und zeigte damit eine moderate Übereinstimmung zwischen den beiden Auswerterinnen an. Der Median der Auswertungszeit pro Studie betrug 19,5 Minuten (IQA 17,5 - 22).

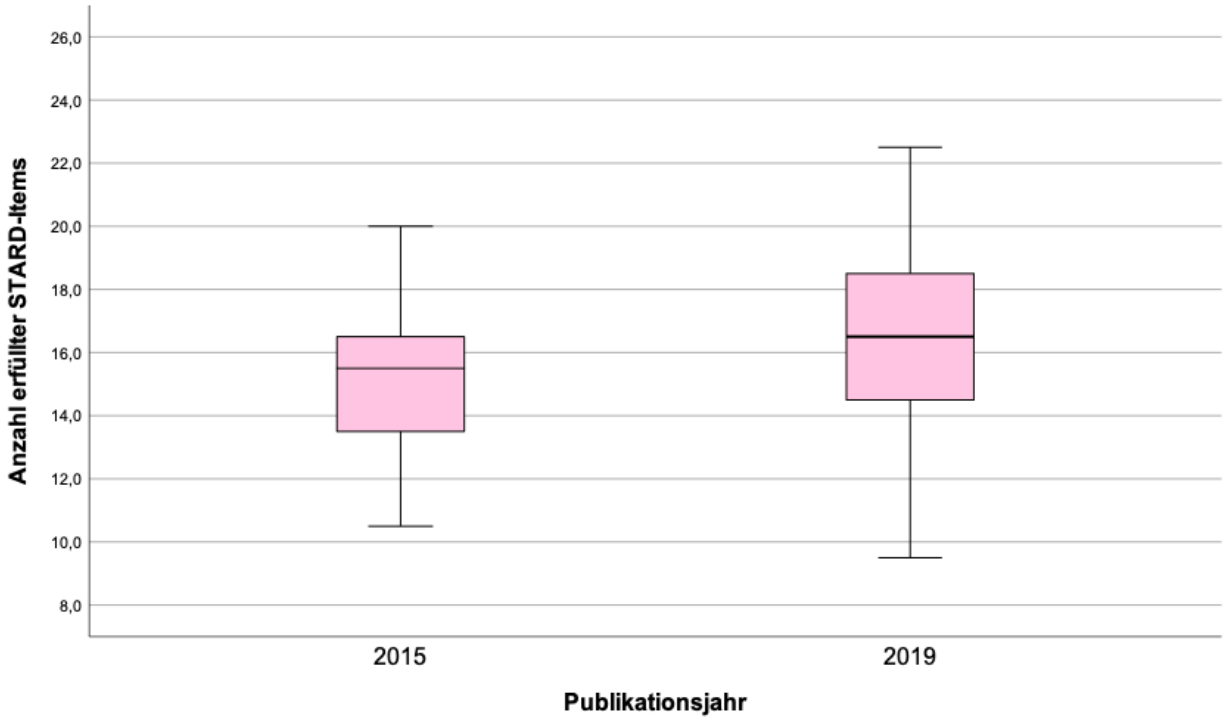
Die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien war 2019 (16,3 ± 2,7) signifikant besser als 2015 (15,1 ± 2,3; $p < 0,02$). Keine signifikanten Unterschiede wurden hingegen in Bezug auf das Studiendesign ($p = 0,13$), die Methodik der Datengewinnung ($p = 0,87$) und die Zitierrate ($p = 0,09$) gefunden. Weitere Details zu den Ergebnissen sind in Tabelle 3 und Abbildung 2 dargestellt.

Tabelle 3: Detaillierte Ergebnisse der Untergruppenanalyse der Studien aus *European Radiology* (erste Publikation)

Untergruppe	**Anzahl erfüllter STARD-Items, Mittelwert ± SD**	**p-Wert**
Publikationsjahr		0,016
2015	15,1 ± 2,3	
2019	16,3 ± 2,7	
Studiendesign		0,129
Kohortenstudie	16,1 ± 2,7	
Fall-Kontroll-Studie	15,0 ± 2,4	
Methodik der Datensammlung		0,865
Prospektiv	15,9 ± 2,9	
Retrospektiv	15,9 ± 2,4	
Zitierrate (Median-Split)		0,094
Niedrig (≤ 0,28 Zitate/Monat)	15,5 ± 2,4	
Hoch (> 0,28 Zitate/Monat)	16,3 ± 2,8	

STARD = Standards for Reporting Diagnostic Accuracy
SD = Standard Deviation (Standardabweichung)
Anmerkung: Übersetzt und entnommen aus Stahl et al., 2023 (21).

Abbildung 2: Boxplot der Ergebnisse der Studien aus *European Radiology* (erste Publikation)



STARD = Standards for Reporting Diagnostic Accuracy
Anmerkung: Übersetzt und modifiziert nach Stahl et al., 2023 (21).

3.2.2 Zweite Publikation

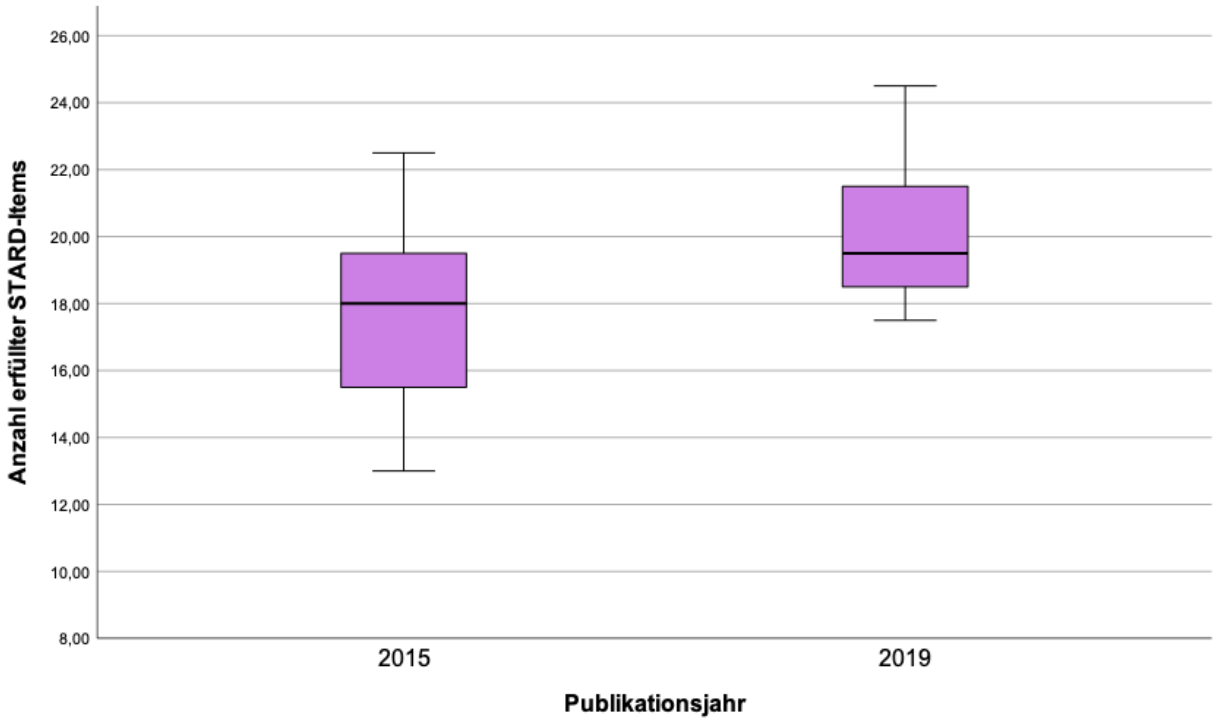
Der Median der erfüllten STARD-Items aller 66 analysierten diagnostischen Genauigkeitsstudien betrug 18,5 von 29 Items (IQA 17,5 - 20,0; Spannweite 13 - 24,5). Beide Auswerterinnen hatten beim Bewerten der STARD-Items eine Übereinstimmung von 85,4%. Der Kappa-Koeffizient nach Cohen betrug 0,70 (95% KI 0,66 - 0,73) und zeigte damit eine erhebliche Übereinstimmung zwischen den Auswerterinnen an.

Die Qualität der Berichterstattung der diagnostischen Genauigkeitsstudien, die im Jahr 2019 veröffentlicht wurden, war signifikant besser als die der diagnostischen Genauigkeitsstudien, die im Jahr 2015 publiziert wurden (19,5 [IQA 18,5 - 21,5] versus 18,0 [IQA 15,5 - 19,5]; 95% KI 1,2 - 3,3; $p < 0,001$; Varghas und Delaneys A = 0,24). Ein signifikanter Unterschied der Mediane im STARD-Gesamtscore zwischen den Jahren 2015 und 2019 wurde ebenfalls in folgenden Untergruppen gefunden: prospektive Studien (3,3; 95% KI 1,6 - 5,0; $p < 0,001$; Varghas und Delaneys A = 0,14), Kohortenstudien

(2,0; 95% KI 0,8 - 3,1; p = 0,002; Varghas und Delaneys A = 0,28), Fall-Kontroll-Studien (4,15; 95% KI 1,0 - 7,3; p = 0,017; Varghas und Delaneys A = 0,0), Studien mit einer Zitierrate über dem Median (2,5; 95% KI 0,9 - 4,1; p = 0,003; Varghas und Delaneys A = 0,23) und Studien mit einer Zitierrate unter dem Median (2,2; 95% KI 0,6 - 3,8; p = 0,008; Varghas und Delaneys A = 0,24). Im Gegensatz dazu zeigte sich bei den retrospektiven Studien kein signifikanter Unterschied in der Qualität der Berichterstattung zwischen den Jahren 2015 und 2019 (1,4; 95% KI -0,1 - 2,9; p = 0,065; Varghas und Delaneys A = 0,36).

Ebenfalls keine signifikanten Unterschiede wurden beim Vergleich des Studiendesigns (p = 0,81; Varghas und Delaneys A = 0,53), der Methodik der Datengewinnung (p = 0,68; Varghas und Delaneys A = 0,47) und der Zitierrate (p = 0,54; Varghas und Delaneys A = 0,54) gefunden. In Abbildung 3 und Tabelle 4 sind weitere Details zu den Ergebnissen aufgeführt.

Abbildung 3: Boxplot der Ergebnisse der Studien aus *Radiology* (zweite Publikation)



STARD = Standards for Reporting Diagnostic Accuracy
Anmerkung: Übersetzt und modifiziert nach Stahl et al., 2023 (22).

Tabelle 4: Zusammenfassung der Ergebnisse der Untergruppenanalyse der Studien aus *Radiology* (zweite Publikation)

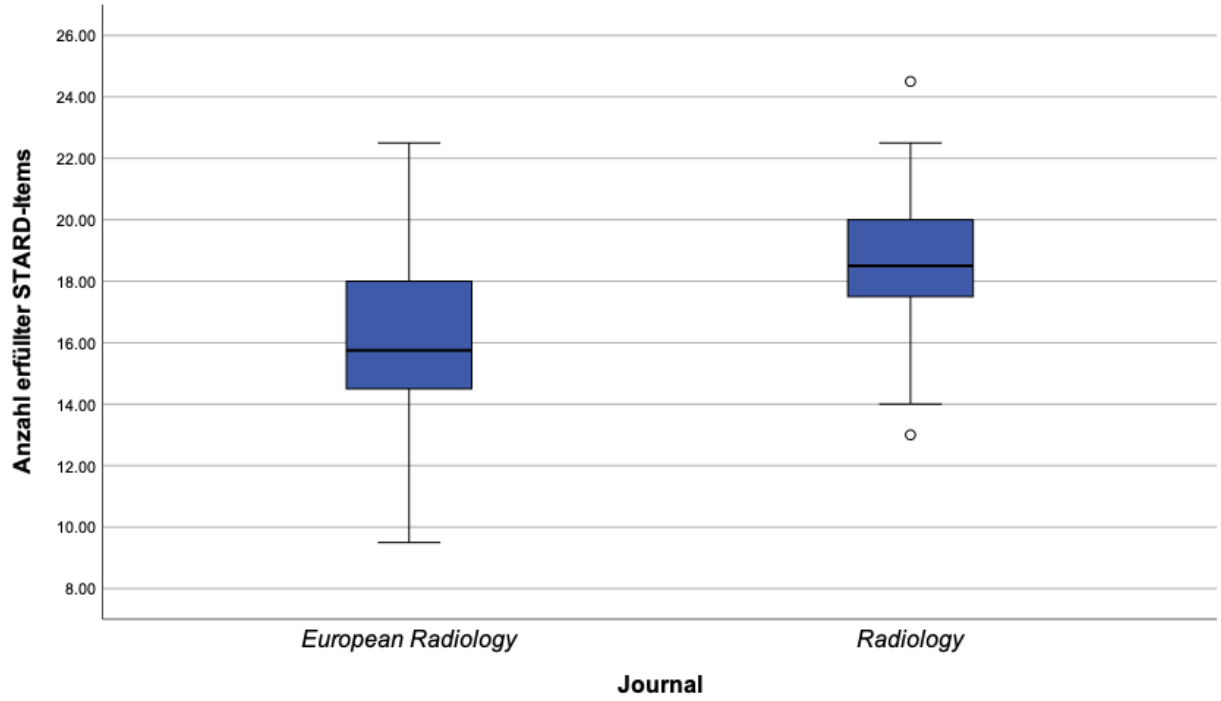
Untergruppe	Zusammenfassung der Ergebnisse
Publikationsjahr	Studien, die in 2019 veröffentlicht wurden, haben mehr Items erfüllt als die Studien, die 2015 veröffentlicht wurden (19,5 [IQA 18,5 - 21,5] versus 18,0 [IQA 15,5 - 19,5]; p < 0,001; Varghas und Delaneys A = 0,24).
Studiendesign	Es wurde kein signifikanter Unterschied des STARD-Gesamtscores in Bezug auf das Studiendesign gefunden (Fall-Kontroll-Studien: 18,5 [IQA 17,5 - 19,0] versus Kohortenstudien: 18,5 [IQA 17,5 - 20,0]; p = 0,81; Varghas und Delaneys A = 0,53).
Methodik der Datensammlung	Es wurde kein signifikanter Unterschied des STARD-Gesamtscores in Bezug auf die Methodik der Datensammlung gefunden (retrospektiv: 18,8 [IQA 17,9 - 19,6] versus prospektiv: 18,5 [IQA 16,8 - 20,0]; p = 0,68; Varghas und Delaneys A = 0,47).
Zitierrate (Median-Split) [a]	Es wurde kein signifikanter Unterschied des STARD-Gesamtscores in Bezug auf die Zitierrate gefunden (niedrig: 18,5 [IQA 18,0 - 20,5] versus hoch: 18,5 [IQA 17,5 - 19,5]; p = 0,54; Varghas und Delaneys A = 0,54).

STARD = Standards for Reporting Diagnostic Accuracy
IQA = Interquartilsabstand
[a] niedrig < 0,56 Zitate/Monat, hoch ≥ 0,56 Zitate/Monat
Anmerkung: Übersetzt und entnommen aus Stahl et al. 2023 (22).

3.2.3 Vergleich beider Journale (nicht in den Publikationen enthalten)

Der Mittelwert des STARD-Gesamtscores aller 180 diagnostischen Genauigkeitsstudien beider Journale zusammen betrug 16,9 ± 2,9 von 29 Items (58,3%; Spannweite 9,5 - 24,5). Die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien war in *Radiology* (18,6 ± 2,4) signifikant besser als in *European Radiology* (15,9 ± 2,6; p < 0,01). Die Ergebnisse sind ebenfalls im Boxplot in Abbildung 4 dargestellt.

Abbildung 4: Boxplot der Ergebnisse der Studien beider Journale im Vergleich



STARD = Standards for Reporting Diagnostic Accuracy
Eigene Darstellung der Daten aus Stahl et al., 2023 (21) und Stahl et al., 2023 (22): Ann-Christine Stahl.

3.3 Häufigkeit der Verwendung der einzelnen Items

3.3.1 Erste Publikation

Tabelle 5 zeigt, wie häufig die jeweiligen Items in den Studien aus *European Radiology* berücksichtigt wurden. Die Unterschiede in der Berücksichtigung der einzelnen STARD-Items waren mit einer Spannweite von 1% bis 100% sehr groß. Sechs Items wurden von nur sehr wenigen Studien verwendet (< 20%): Die Studienziele und Hypothesen (Item 4), die Information darüber, wie mit unklaren Ergebnissen des Indextests oder des Referenzstandards umgegangen wurde (Item 15), die beabsichtigte Stichprobengröße und wie sie ermittelt wurde (Item 18), die Kreuztabelle der Ergebnisse des Indextests und des Referenzstandards (Item 23) und irgendwelche unerwünschten Ereignisse bei der Durchführung des Indextests oder Referenzstandards (Item 25). Item 28 (Registrierungsnummer und Name des Registers) wurde nur von einer der eingeschlossenen diagnostischen Genauigkeitsstudien berücksichtigt.

Im Gegensatz dazu wurden die folgenden zwei Items von allen untersuchten Studien wiedergegeben: Item 2 (strukturierte Zusammenfassung des Studiendesigns, der Methodik, Ergebnisse und Schlussfolgerungen) und Item 30 (Finanzierungsquellen und andere Unterstützung). Weitere Items wurden ebenfalls häufig, also von > 80% der analysierten Studien, berücksichtigt: Item 1 (Kennzeichnung der Studie als diagnostische Genauigkeitsstudie durch Verwendung mindestens eines Genauigkeitsmaßes), Item 3 (wissenschaftlicher und klinischer Hintergrund, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests), Item 5 (ob die Datenerhebung vor oder nach der Durchführung des Indextests und Referenzstandards geplant wurde), Item 7 (auf welcher Basis potenziell einzuschließende Studienteilnehmer*innen gefunden wurden), Item 10a (ausreichend detaillierte Beschreibung des Indextests, um Wiederholungen zu ermöglichen), Item 21a (Verteilung der Schweregrade der Krankheit bei Studienteilnehmer*innen mit dem untersuchten Zustand) und Item 27 (Implikationen für die Praxis, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests).

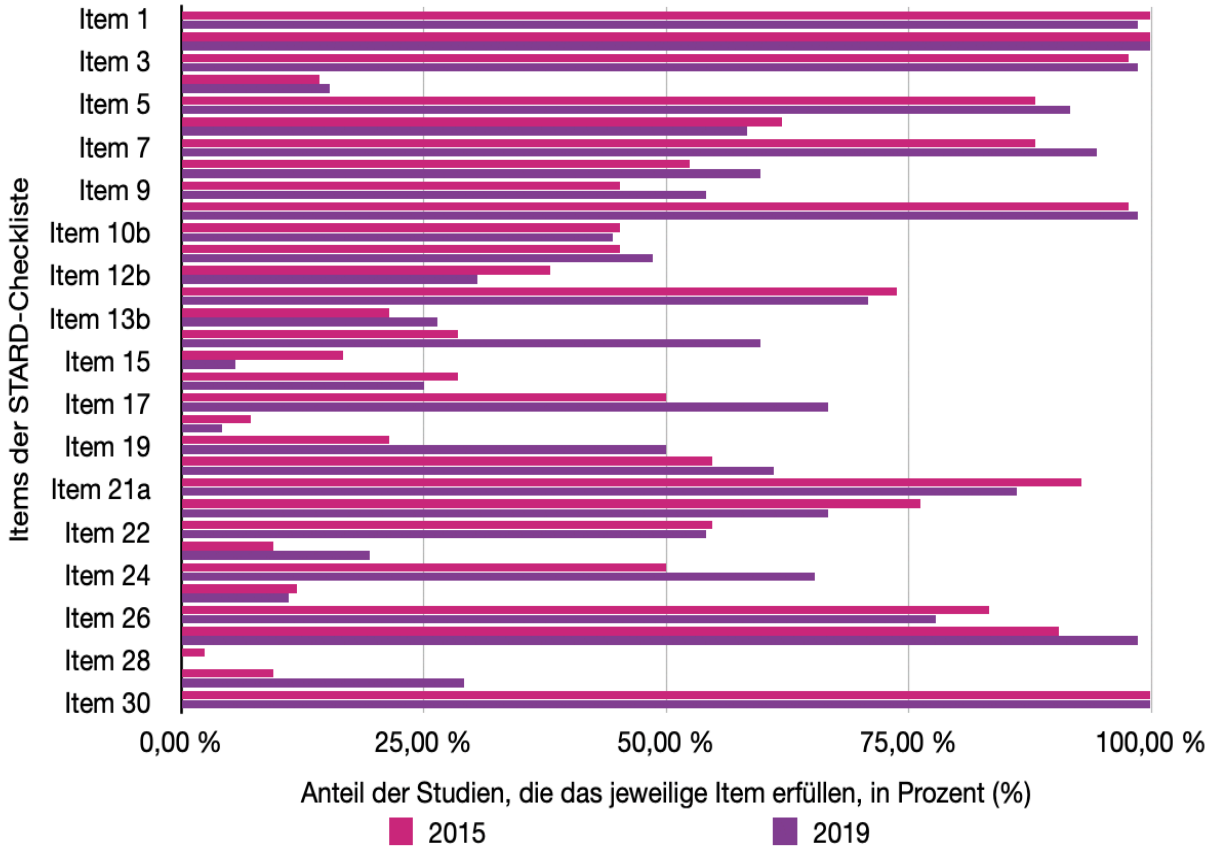
Allerdings gibt es Unterschiede in der Verwendung der einzelnen Items zwischen den Jahren 2015 und 2019, die in Abbildung 5 dargestellt sind. Die meisten Items wurden 2019 häufiger berücksichtigt als 2015. Für die folgenden Items war dieser Unterschied besonders groß und ist damit erwähnenswert: Methoden für die Kalkulation oder den Vergleich der diagnostischen Genauigkeitsmaße (Item 14; 29% versus 60%), die Verwendung eines Flussdiagramms (Item 19; 21% versus 50%), die Kreuztabelle der Ergebnisse des Indextests und des Referenzstandards (Item 23; 10% versus 19%) und der Zugang zum vollständigen Studienprotokoll (Item 29; 10% versus 29%).

14 Items wurden jedoch im Jahr 2015 häufiger wiedergegeben als im Jahr 2019. Dabei sind vor allem folgende Items hervorzuheben: Item 12b (vorab festgelegte Definition von und Gründe für Cut-off-Werte und Ergebniskategorien für den Referenzstandard; 38% versus 31%), Item 15 (wie mit unklaren Ergebnissen des Indextests oder des Referenzstandards umgegangen wurde; 17% versus 6%), Item 18 (beabsichtigte Stichprobengröße und wie sie ermittelt wurde; 7% versus 4%), Item 21a (Verteilung der Schweregrade der Krankheit bei Studienteilnehmer*innen mit dem untersuchten Zu-

stand; 93% versus 86%) und Item 21b (Verteilung alternativer Diagnosen bei Studienteilnehmer*innen ohne den untersuchten Zustand; 76% versus 67%).

Items, die den Indextest betreffen, wurden generell häufiger in den Studien wiedergegeben als Items, die sich auf den Referenzstandard beziehen (Items 10, 12 und 13).

Abbildung 5: Häufigkeiten der Verwendung der einzelnen STARD-Items der Studien aus *European Radiology* (erste Publikation)



STARD = Standards for Reporting Diagnostic Accuracy
Anmerkung: Eigene Darstellung der Daten aus Stahl et al., 2023: Ann-Christine Stahl (21).

3.3.2 Zweite Publikation

In Tabelle 5 ist dargestellt, wie häufig die jeweiligen Items in Studien, die in *Radiology* publiziert wurden, berücksichtigt wurden. Auch hier war die Spannweite sehr groß (5% - 100%). Vier Items wurden nur von sehr wenigen (< 20%) Studien berücksichtigt: Item

18 (beabsichtigte Stichprobengröße und wie sie ermittelt wurde), Item 23 (Kreuztabelle der Ergebnisse des Indextests und des Referenzstandards), Item 25 (irgendwelche unerwünschten Ereignisse bei der Durchführung des Indextests oder Referenzstandards) und Item 28 (Registrierungsnummer und Name des Registers).

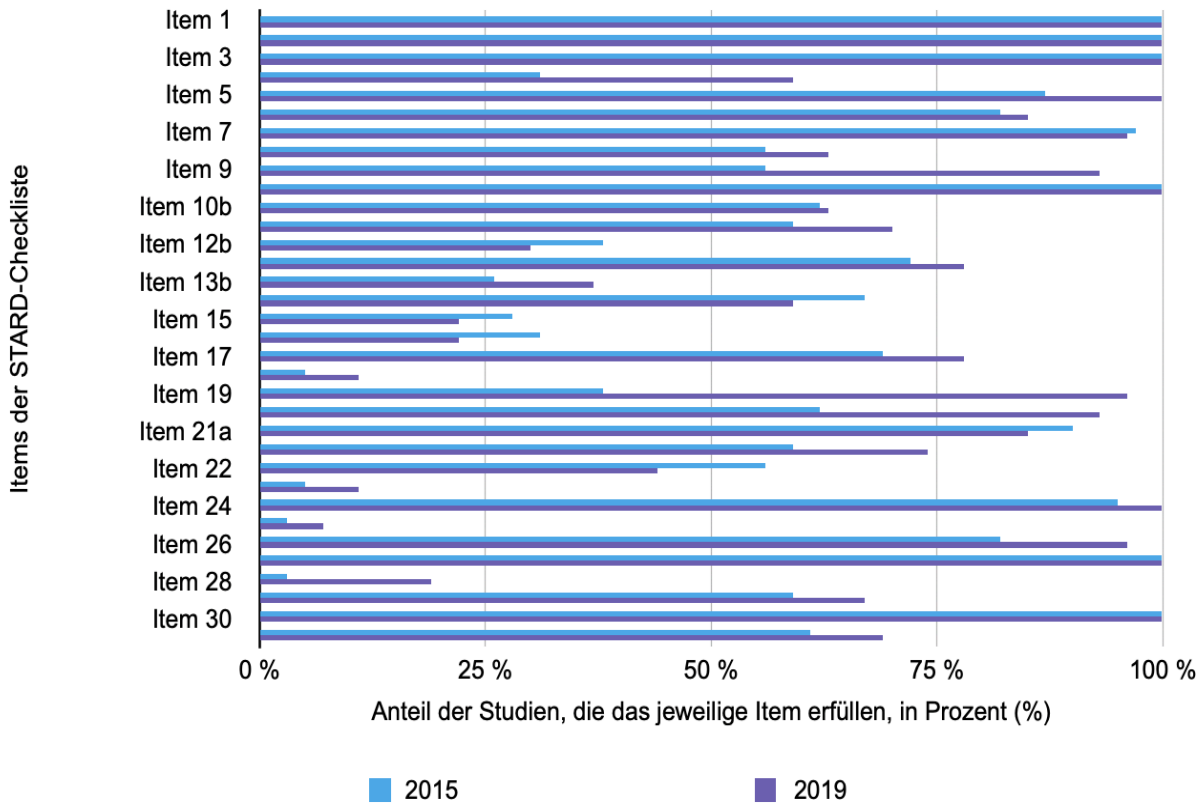
Item 1 (Kennzeichnung der Studie als diagnostische Genauigkeitsstudie durch Verwendung mindestens eines Genauigkeitsmaßes), Item 2 (strukturierte Zusammenfassung des Studiendesigns, der Methodik, Ergebnisse und Schlussfolgerungen), Item 3 (wissenschaftlicher und klinischer Hintergrund, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests), Item 10a (ausreichend detaillierte Beschreibung des Indextests, um Wiederholungen zu ermöglichen), Item 27 (Implikationen für die Praxis, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests) und Item 30 (Finanzierungsquellen und andere Unterstützung) wurden hingegen von allen 66 aus *Radiology* eingeschlossenen Studien berücksichtigt. Die folgenden sechs Items wurden ebenfalls von vielen (> 80%) Studien erwähnt: Item 5 (ob die Datenerhebung vor oder nach der Durchführung des Indextests und Referenzstandards geplant wurde), Item 6 (Auswahlkriterien), Item 7 (auf welcher Basis potenziell einzuschließende Studienteilnehmer*innen gefunden wurden), Item 21a (Verteilung der Schweregrade der Krankheit bei Studienteilnehmer*innen mit dem untersuchten Zustand), Item 24 (Kalkulationen der diagnostischen Genauigkeit und ihrer Exaktheit) und Item 26 (Studienlimitationen, einschließlich Quellen für mögliche Bias, statistischer Unsicherheit und Generalisierbarkeit).

Abbildung 6 zeigt die Häufigkeit der Verwendung der einzelnen STARD-Items im Vergleich zwischen den Jahren 2015 und 2019. Die meisten Items (61%; 20/33) wurden im Jahr 2019 häufiger wiedergegeben als im Jahr 2015. Die Verbesserung war besonders groß (> 20 Prozentpunkte) für folgende Items: Item 4 (Studienziele und Hypothesen), Item 9 (ob die Studienteilnehmer*innen eine konsekutive, zufällige oder willkürliche Stichprobe gebildet haben), Item 19 (Verwendung eines Flussdiagramms) und Item 20 (grundlegende demographische und klinische Merkmale der Studienteilnehmer*innen). 7 Items wurden jedoch im Jahr 2015 häufiger berücksichtigt als im Jahr 2019: Item 7 (auf welcher Basis potenziell einzuschließende Studienteilnehmer*innen gefunden wurden), Item 12b (vorab festgelegte Definition von und Gründe für Cut-off-Werte und Er-

gebniskategorien für den Referenzstandard), Item 14 (Methoden für die Kalkulation oder den Vergleich der diagnostischen Genauigkeitsmaße), Item 15 (wie mit unklaren Ergebnissen des Indextests oder des Referenzstandards umgegangen wurde), Item 16 (wie mit fehlenden Daten des Indextests oder des Referenzstandards umgegangen wurde), Item 21a (Verteilung der Schweregrade der Krankheit bei Studienteilnehmer*innen mit dem untersuchten Zustand) und Item 22 (Zeitintervall und irgendwelche klinischen Interventionen zwischen dem Indextest und dem Referenzstandard).

Auch hier wurden Items, die den Indextest betreffen, generell häufiger in den Studien wiedergegeben als Items, die sich auf den Referenzstandard beziehen (Items 10, 12 und 13).

Abbildung 6: Häufigkeiten der Verwendung der einzelnen STARD-Items der Studien aus *Radiology* (zweite Publikation)



STARD = Standards for Reporting Diagnostic Accuracy
Anmerkung: Übersetzt und modifiziert nach Stahl et al., 2023 (22).

3.3.3 Vergleich beider Journale (nicht in den Publikationen enthalten)

Tabelle 5 zeigt für alle eingeschlossenen Studien zusammen und nach den beiden Journalen getrennt, wie häufig die jeweiligen Items in den Studien berücksichtigt wurden. Die meisten Items (79%; 26/33) wurden in den Studien, die in *Radiology* veröffentlicht wurden, häufiger als in den Studien aus *European Radiology* berücksichtigt. Besonders groß (≥ 20 Prozentpunkte) und damit erwähnenswert war der Unterschied für die folgenden Items: Item 4 (Studienziele und Hypothesen; 15% versus 42%), Item 6 (Auswahlkriterien; 60% versus 83%), Item 9 (ob die Studienteilnehmer*innen eine konsekutive, zufällige oder willkürliche Stichprobe gebildet haben; 51% versus 71%), Item 19 (Verwendung eines Flussdiagramms; 39% versus 62%), Item 24 (Kalkulationen der diagnostischen Genauigkeit und ihrer Exaktheit; 60% versus 97%) und Item 29 (wo das vollständige Studienprotokoll eingesehen werden kann; 22% versus 62%).

Fünf Items wurden aber auch in Studien aus *European Radiology* häufiger erwähnt: Item 21a (Verteilung der Schweregrade der Krankheit bei Studienteilnehmer*innen mit dem untersuchten Zustand; 89% versus 88%), Item 21b (Verteilung alternativer Diagnosen bei Studienteilnehmer*innen ohne den untersuchten Zustand; 70% versus 65%), Item 22 (Zeitintervall und irgendwelche klinischen Interventionen zwischen dem Indextest und dem Referenzstandard; 54% versus 52%), Item 23 (Kreuztabelle der Ergebnisse des Indextests und des Referenzstandards; 16% versus 8%) und Item 25 (irgendwelche unerwünschten Ereignisse bei der Durchführung des Indextests oder Referenzstandards; 11% versus 5%).

Tabelle 5: Qualität der Berichterstattung der einzelnen Items der STARD-Checkliste in Prozent (%)

Abschnitt und Item-Nr.	**Beschreibung des Items**	**Alle**	**European Radiology**	**Radiology**
Titel oder Zusammenfassung				
1	Kennzeichnung der Studie als diagnostische Genauigkeitsstudie durch Verwendung mindestens eines Genauigkeitsmaßes (wie Sensitivität, Spezifität, Vorhersagewerte oder AUC)	99	99	100

Tabelle 5: Qualität der Berichterstattung der einzelnen Items der STARD-Checkliste in Prozent (%)

Zusammenfassung				
2	Strukturierte Zusammenfassung des Studiendesigns, der Methodik, Ergebnisse und Schlussfolgerungen	100	100	100
Einleitung				
3	Wissenschaftlicher und klinischer Hintergrund, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests	99	98	100
4	Studienziele und Hypothesen	25	15	42
Methodik				
5	Ob die Datenerhebung vor (prospektiv) oder nach (retrospektiv) der Durchführung des Indextests und Referenzstandards geplant wurde	91	90	92
6	Auswahlkriterien	68	60	83
7	Auf welcher Basis potenziell einzuschließende Studienteilnehmer*innen gefunden wurden (wie Symptome, Ergebnisse vorheriger Tests, Einschluss im Register)	94	92	97
8	Wo und wann potenziell einzuschließende Studienteilnehmer*innen gefunden wurden (Setting, Ort und Zeit)	58	57	59
9	Ob die Studienteilnehmer*innen eine konsekutive, zufällige oder willkürliche Stichprobe gebildet haben	58	51	71

Tabelle 5: Qualität der Berichterstattung der einzelnen Items der STARD-Checkliste in Prozent (%)

10	a) Ausreichend detaillierte Beschreibung des Indextests, um Wiederholungen zu ermöglichen b) Ausreichend detaillierte Beschreibung des Referenzstandards, um Wiederholungen zu ermöglichen	99 51	98 45	100 62
12	a) Vorab festgelegte Definition von und Gründe für Cut-off-Werte und Ergebniskategorien für den Indextest b) Vorab festgelegte Definition von und Gründe für Cut-off-Werte und Ergebniskategorien für den Referenzstandard	53 34	47 33	64 35
13	a) Ob klinische Informationen und Ergebnisse des Referenzstandards den Durchführer*innen/Auswerter*innen des Indextests zugänglich waren b) Ob klinische Informationen und Ergebnisse des Indextests den Auswerter*innen des Referenzstandards zugänglich waren	73 27	72 25	74 30
14	Methoden für die Kalkulation oder den Vergleich der diagnostischen Genauigkeitsmaße	54	48	64
15	Wie mit unklaren Ergebnissen des Indextests oder des Referenzstandards umgegangen wurde	16	10	26
16	Wie mit fehlenden Daten des Indextests oder des Referenzstandards umgegangen wurde	27	26	27
17	Irgendwelche vorab festgelegten Varianzanalysen der diagnostischen Genauigkeit	65	61	73
18	Beabsichtigte Stichprobengröße und wie sie ermittelt wurde	6	5	8
Ergebnisse				

Tabelle 5: Qualität der Berichterstattung der einzelnen Items der STARD-Checkliste in Prozent (%)

19	Verwendung eines Flussdiagramms	48	39	62
20	Grundlegende demographische und klinische Merkmale der Studienteilnehmer*innen	64	59	74
21	a) Verteilung der Schweregrade der Krankheit bei Studienteilnehmer*innen mit dem untersuchten Zustand b) Verteilung alternativer Diagnosen bei Studienteilnehmer*innen ohne den untersuchten Zustand	88 68	89 70	88 65
22	Zeitintervall und irgendwelche klinischen Interventionen zwischen dem Indextest und dem Referenzstandard	53	54	52
23	Kreuztabelle der Ergebnisse des Indextests (oder ihrer Verteilung) und des Referenzstandards	13	16	8
24	Kalkulationen der diagnostischen Genauigkeit und ihrer Exaktheit (wie z.B. 95% Konfidenzintervalle)	73	60	97
25	Irgendwelche unerwünschten Ereignisse bei der Durchführung des Indextests oder Referenzstandards	9	11	5
Diskussion				
26	Studienlimitationen, einschließlich Quellen für mögliche Bias, statistischer Unsicherheit und Generalisierbarkeit	83	80	88
27	Implikationen für die Praxis, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests	97	96	100
Andere Informationen				
28	Registrierungsnummer und Name des Registers	4	1	9

Tabelle 5: Qualität der Berichterstattung der einzelnen Items der STARD-Checkliste in Prozent (%)

29	Wo das vollständige Studienprotokoll eingesehen werden kann	37	22	62
30	Finanzierungsquellen und andere Unterstützung; Rolle der Finanzierer*innen	100	100	100

STARD = Standards for Reporting Diagnostic Accuracy
Item-Nr. = Itemnummer
AUC = Area under Curve (Fläche unter der Grenzwertoptimierungskurve)
Anmerkung: Übersetzt und modifiziert nach Stahl et al., 2023 (21) und Stahl et al., 2023 (22).

Diskussion

4.1 Kurze Zusammenfassung der Ergebnisse

Die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien lässt mit durchschnittlich 16,9 ± 2,9 erfüllten Items von 29 möglichen Items (58,3%; Spannweite 9,5 - 24,5) in zwei führenden Journalen der radiologischen Literatur weiterhin Spielraum für Verbesserungen. Auch wenn in beiden Journalen eine Verbesserung über die Zeit beziehungsweise mit den geänderten Einreichungsrichtlinien festgestellt werden konnte (*European Radiology:* 16,3 ± 2,7 versus 15,1 ± 2,3; $p < 0,02$; *Radiology:* 19,5 [IQA 18,5 - 21,5] versus 18,0 [IQA 15,5 - 19,5]; $p < 0,001$; Varghas und Delaneys A = 0,24), weisen einige Bereiche weiterhin große Mängel auf.

Zudem konnten wir feststellen, dass die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien in *Radiology* signifikant besser war als die von in *European Radiology* veröffentlichten Studien (18,6 ± 2,4 versus 15,9 ± 2,6; $p < 0,01$).

Items, die besonders von den Mängeln an der Qualität der Berichterstattung betroffen sind, sind die Folgenden: Item 15 (wie mit unklaren Ergebnissen des Indextests oder des Referenzstandards umgegangen wurde), Item 18 (beabsichtigte Stichprobengröße und wie sie ermittelt wurde), Item 23 (Kreuztabelle der Ergebnisse des Indextests und des Referenzstandards), Item 25 (irgendwelche unerwünschten Ereignisse bei der Durchführung des Indextests oder Referenzstandards) und Item 28 (Registrierungsnummer und Name des Registers).

Die Items 1 (Kennzeichnung der Studie als diagnostische Genauigkeitsstudie durch Verwendung mindestens eines Genauigkeitsmaßes), 2 (strukturierte Zusammenfassung des Studiendesigns, der Methodik, Ergebnisse und Schlussfolgerungen), 3 (wissenschaftlicher und klinischer Hintergrund, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests), 10a (ausreichend detaillierte Beschreibung des Indextests, um Wiederholungen zu ermöglichen), 27 (Implikationen für die Praxis, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests) und 30 (Finanzierungsquellen und andere Unterstützung) wurden hingegen von allen oder fast allen eingeschlossenen Studien erfüllt.

Außerdem ist zu erwähnen, dass wir in beiden Journalen keinen signifikanten Unterschied in der Qualität der Berichterstattung in Bezug auf das Studiendesign (Kohorten- versus Fall-Kontroll-Studie), die Methodik der Datengewinnung (prospektiv versus retrospektiv) und die Zitierrate gefunden haben.

4.2 Interpretation der Ergebnisse

In beiden Journalen konnte über die Zeit beziehungsweise mit den geänderten Einreichungsrichtlinien eine Verbesserung der Qualität der Berichterstattung diagnostischer Genauigkeitsstudien festgestellt werden. *European Radiology* hat sich dabei um durchschnittlich 1,2 Items verbessert (16,3 versus 15,1) und *Radiology* um 1,5 Items (19,5 versus 18,0). Die prozentuale Verbesserung war in beiden Journalen folglich nahezu identisch (*European Radiology*: 7,9%; *Radiology*: 8,3%). Die ergänzende Empfehlung zur Verwendung der STARD-Checkliste in den Einreichungsrichtlinien von *European Radiology* hat also zu einer ähnlichen Verbesserung geführt wie die Verpflichtung zur Verwendung der STARD-Checkliste von *Radiology*. Dabei ist es jedoch wichtig zu erwähnen, dass wir bei den diagnostischen Genauigkeitsstudien aus *European Radiology* den Mittelwert betrachtet haben, während wir bei den Studien aus *Radiology* den Median berechnet und verglichen haben.

Die Tatsache, dass die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien in *Radiology* signifikant besser war als die von in *European Radiology* veröffentlichten Studien, lässt sich durch zwei Faktoren erklären:
Zum einen hat *Radiology* die Berücksichtigung der STARD-Checkliste im Januar 2016 verpflichtend gemacht, während *European Radiology* die Verwendung seit Mitte 2017 lediglich empfiehlt, und zum anderen ist der Impact Factor von *Radiology* im Vergleich zu *European Radiology* deutlich höher (29.146 [2021] (31) versus 7.034 [2021] (32)).

Interessant und überraschend ist auch, welche Items in den beiden Journalen besonders selten berücksichtigt wurden. Denn einige der betroffenen Items - wie beispielsweise Item 25 (irgendwelche unerwünschten Ereignisse bei der Durchführung des Indextests oder Referenzstandards) - können teilweise in einem kurzen Satz wiedergegeben werden.

4.3 Einbettung der Ergebnisse in den bisherigen Forschungsstand

Vergleichen wir unsere Studienergebnisse mit denen anderer Arbeiten, die die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien zum Thema haben, lässt sich festhalten, dass wir einen ähnlichen Mittelwert erfüllter STARD-Items gefunden haben wie auch Michelessi et al. im Jahr 2017 (16,8 ± 3,1 [54,2%]), Hong et al. im Jahr 2018 (16,6 ± 2,2 [55,3%]) und Wright et al. im Jahr 2021 (17,8 ± 3,1 [52,4%]) (14, 15, 33). Michelessi et al. haben sich mit der Qualität der Berichterstattung von 106 diagnostischen Genauigkeitsstudien beschäftigt, die zwischen 2003 und 2014 veröffentlicht wurden und die die Diagnostik des Glaukoms zum Thema hatten (14). Dabei haben sie die aktuelle STARD-Checkliste aus dem Jahr 2015 und den gleichen dichotomen Bewertungsmodus wie wir verwendet, bei dem nur zwischen suffizient wiedergegebenen Items und nicht suffizient wiedergegebenen Items unterschieden wird (14). Das macht einen Vergleich zu unseren Ergebnissen problemlos möglich. Ebenfalls unseren Ergebnissen entsprechend, haben Michelessi et al. eine leichte Verbesserung der Qualität der Berichterstattung über die Zeit gefunden (14). Hong et al. wiederum haben die Qualität der Berichterstattung von 142 diagnostischen Genauigkeitsstudien untersucht, die im Jahr 2016 in verschiedenen radiologischen Journalen - darunter auch *European Radiology* und *Radiology* - veröffentlicht wurden (15). Dafür haben sie ebenfalls die aktuelle STARD-Checkliste verwendet, sich im Gegensatz zu unserer Arbeit jedoch gegen einen dichotomen Bewertungsmodus entschieden und die einzelnen Items mit „yes“, „no“ oder „not applicable“ bewertet (15). Bei der anschließenden statistischen Auswertung wurden Items, die als „not applicable“ eingestuft wurden, jedoch wie erfüllte Items behandelt und erzielten somit einen Punkt (15). Dieses Vorgehen könnte das Ergebnis von Hong et al. im Vergleich zu unserem nach oben verzerrt haben. Sie haben zudem festgestellt, dass diagnostische Genauigkeitsstudien aus Journalen, die die Verwendung der STARD-Checkliste empfehlen oder dazu verpflichten, eine signifikant bessere Qualität der Berichterstattung aufweisen also solche aus Journalen, die von einer Empfehlung absehen (15). Auch wir konnten feststellen, dass eine ergänzende Empfehlung oder Verpflichtung in den Einreichungsrichtlinien von *European Radiology* beziehungsweise *Radiology* zu einer signifikanten Verbesserung der Qualität der Berichterstattung geführt hat. Auch Journale mit einem höheren Impact Factor wiesen in der Analyse von Hong et al. eine qualitativ bessere Berichterstattung diagnostischer Genauigkeitsstudien auf (15). Außerdem konnten sie wie wir keinen signifikanten Unterschied in Bezug

auf das Studiendesign (Kohorten- versus Fall-Kontroll-Studie) feststellen (15). Wright et al. haben die Qualität der Berichterstattung von lediglich 26 diagnostischen Genauigkeitsstudien untersucht, die als Referenzen der „Quality Improvement Guidelines for Diagnostic Arteriography“ der Gesellschaft für interventionelle Radiologie zu finden waren (33). Dabei haben sie sich ebenfalls für eine dichotome Bewertungsweise entschieden, jedoch - im Gegensatz zu unserem Vorgehen - Subitems als volle Items gewertet, weswegen ein höherer STARD-Gesamtscore von maximal 34 Punkten möglich war (33).

Eine deutlich schlechtere Qualität der Berichterstattung lassen die aktuellen Studienergebnisse von Nassar et al. vermuten (11,5 ± 2,5 [33,8%]) (34). Sie haben 106 diagnostische Genauigkeitsstudien, die Screeninginstrumente bei Verdacht auf Depressionen untersucht haben, bewertet (34). Auch hier wurden die Subitems als volle Items betrachtet, sodass ein STARD-Gesamtscore von maximal 34 Punkten möglich war (34). Ursächlich für den im Vergleich zu unseren Ergebnissen niedrigeren durchschnittlichen STARD-Gesamtscore ist vor allem die Tatsache, dass sich hier für die vier Bewertungskategorien „adequately reported“, „partially reported“, „inadequately or not reported” und „not applicable“ entschieden wurde (34). Die Auswertung erfolgte für die beschriebenen Bewertungskategorien getrennt, sodass eine durchschnittliche Anzahl von 10,1 ± 2,5 Items als „inadequately reported“ und von 8,6 ± 2,1 Items als „partially reported“ eingestuft wurde (34). Eine Angabe der durchschnittlichen Anzahl an Items, die mit „not applicable“ bewertet wurden, erfolgte in der Publikation von Nassar et al. leider nicht (34). Einen ähnlich niedrigen STARD-Gesamtscore von 11,2 ± 2,7 (37,3%) und damit eine schlechtere Qualität der Berichterstattung fanden auch Jang et al. in 66 diagnostischen Genauigkeitsstudien, die zwischen 2012 und 2018 in *Annals of Laboratory Medicine* veröffentlicht wurden (35). Sie haben teilweise auch Items, die von den STARD-Autor*innen als komplettes Item veröffentlicht wurden, für ihre Bewertung in Subitems untergliedert (35). Diesen wurden darauf fraktionierte Punkte zugeschrieben, sodass das gesamte Item wiederum maximal einen Punkt erzielen konnte (35). Beispielsweise wurde Item 4 (Studienziele und Hypothesen) in Item 4a (Studienziele) und Item 4b (Hypothesen) aufgeteilt und jeweils mit maximal 0,5 Punkten bewertet (35). So war ein maximaler STARD-Gesamtscore von 30 möglich (35). Dieses Vorgehen macht eine detailliertere Analyse der Qualität der Berichterstattung möglich (35). Wir wollten uns jedoch

möglichst an die original STARD-Checkliste und die Vorgaben ihrer Autor*innen halten, sodass wir uns gegen die beschriebene Methodik entschieden haben. Auf jeden Fall kann dieser methodische Unterschied den im Vergleich zu unserer Arbeit niedrigeren STARD-Gesamtscore von Jang et al. erklären. Darüber hinaus lassen sich die geringere Anzahl an untersuchten Studien und das andere Themengebiet als Ursachen für den niedrigeren STARD-Gesamtscore nennen.

Eine bessere Qualität der Berichterstattung mit durchschnittlich 20,0 ± 2,1 (74,1%) erfüllten Items von 27 möglichen Items hingegen fanden Choi et al. in 63 diagnostischen Genauigkeitsstudien, die zwischen 2011 und 2015 im *Korean Journal of Radiology* veröffentlicht wurden (13). Ihr methodisches Vorgehen entspricht mit einer dichotomen Bewertung der einzelnen Items weitestgehend dem unseren (13). Allerdings haben sie für ihre Analyse die Items 28, 29 und 30 ausgeschlossen (13). Item 28 und 29 wurden in unserer Arbeit nur von 4% und 37% der eingeschlossenen diagnostischen Genauigkeitsstudien erfüllt, weswegen ein Ausschluss dieser Items den besseren STARD-Gesamtscore von Choi et al. erklären kann, auch wenn Item 30 von allen unseren eingeschlossenen Studien erfüllt wurde. Des Weiteren haben Choi et al. - unseren Studienergebnissen entsprechend - keinen signifikanten Unterschied der Qualität der Berichterstattung bezüglich der Zitierrate oder des Studiendesigns (Kohorten- versus Fall-Kontroll-Studie) gefunden (13). Dies entspricht auch den Ergebnissen von Hogan et al., die ebenfalls keinen signifikanten Unterschied in Bezug auf die Anzahl an Zitaten der jeweiligen Artikel gefunden haben (17). Allerdings haben Hogan et al. nicht die Zitierrate, also die Anzahl an Zitaten eines Artikels geteilt durch die Anzahl an Monaten, die seit seiner Veröffentlichung vergangen sind, verwendet, sondern die absolute Anzahl an Zitaten, die von dem Einfluss der Zeit seit Veröffentlichung nicht bereinigt ist (17). Ebenfalls eine Übereinstimmung mit unseren Ergebnissen haben Prager et al. im Jahr 2020 gefunden, indem sie auch keinen signifikanten Unterschied in der Qualität der Berichterstattung diagnostischer Genauigkeitsstudien in Bezug auf die Methodik der Datengewinnung (prospektiv versus retrospektiv) feststellen konnten (19). Gegenstand ihrer Analyse waren 74 diagnostische Genauigkeitsstudien, die zwischen 2016 und 2019 zum Thema „Point-of-Care Ultraschall“ veröffentlicht wurden (19). Sie fanden zudem einen recht hohen STARD-Gesamtscore mit durchschnittlich 19,7 ± 2,9 erfüllten Items von 30 möglichen Items (65,7%) (19).

Zusammenfassend lässt sich also hervorheben, dass die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien in der radiologischen Literatur oftmals besser ist als die Qualität der Berichterstattung diagnostischer Genauigkeitsstudien in anderen medizinischen Fachbereichen.

Wenn wir die Qualität der Berichterstattung einzelner Items betrachten, lässt sich festhalten, dass in der Analyse von Zheng et al., in der 45 diagnostische Genauigkeitsstudien aus vier verschiedenen Journalen der Labormedizin bewertet wurden, ebenfalls die Items 15 (wie mit unklaren Ergebnissen des Indextests oder des Referenzstandards umgegangen wurde), 18 (beabsichtigte Stichprobengröße und wie sie ermittelt wurde), 25 (irgendwelche unerwünschten Ereignisse bei der Durchführung des Indextests oder Referenzstandards) und 28 (Registrierungsnummer und Name des Registers) besonders schlecht abgeschnitten haben (18). Zudem wurde Item 29 (wo das vollständige Studienprotokoll eingesehen werden kann) in der Analyse von Zheng et al. von nur 9% der eingeschlossenen Studien wiedergegeben (18). Auch bei uns war das Abschneiden dieses Items mit 37% unzureichend.

Wie auch in unserer Analyse wurden bei Choi et al. die Items 1 (Kennzeichnung der Studie als diagnostische Genauigkeitsstudie durch Verwendung mindestens eines Genauigkeitsmaßes), 2 (strukturierte Zusammenfassung des Studiendesigns, der Methodik, Ergebnisse und Schlussfolgerungen), 3 (wissenschaftlicher und klinischer Hintergrund, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests) und 27 (Implikationen für die Praxis, einschließlich des beabsichtigten Nutzens und der klinischen Rolle des Indextests) von besonders vielen Studien erfüllt (13). Item 30 (Finanzierungsquellen und andere Unterstützung), das bei uns ebenfalls sehr gut abschnitt, wurde in der Studie von Choi et al. - wie beschrieben - leider ausgeschlossen (13).

4.4 Stärken und Schwächen der eigenen Studien

Die beiden Studien, die die Grundlage dieses Manteltextes bilden, wurden methodisch sehr ähnlich konzipiert und weisen dadurch gemeinsame Stärken und Schwächen auf, die im Folgenden adressiert werden:

Wir haben MEDLINE via PubMed mit einer für diagnostische Genauigkeitsstudien von Devillé et al. konzipierten und validierten Suchstrategie systematisch durchsucht (25). Allerdings weist diese Suchstrategie eine Sensitivität von 80,0% auf (25), sodass die Möglichkeit besteht, dass wir durch diese Suche nicht alle einzuschließenden Studien gefunden haben. Diesem methodischen Mangel sind wir begegnet, indem wir die Webseiten der Journale *European Radiology* und *Radiology* jeweils händisch mit dem Suchbegriff „Diagnostic accuracy studies“ durchsucht haben. Zudem wurde die Suche von zwei Auswerterinnen (A.S., eine fortgeschrittene Medizinstudentin mit drei Jahren Erfahrung in der Recherche und Analyse von diagnostischen Genauigkeitsstudien sowie A.T., eine Zahnärztin mit einem Jahr Erfahrung in der Recherche und Analyse von diagnostischen Genauigkeitsstudien) unabhängig voneinander durchgeführt und Unstimmigkeiten über den Ein- sowie Ausschluss einzelner Studien wurden mit der Unterstützung eines dritten Auswerters (B.K., ein Arzt mit acht Jahren Erfahrung in der radiologischen Forschung) behoben. So konnten wir sicher sein, keine wichtigen einzuschließenden diagnostischen Genauigkeitsstudien für unsere Analyse zu verpassen. Darüber hinaus ist die Spezifität der Suchstrategie von Devillé et al. mit 97,3% hingegen nahezu perfekt (25).

Die Bewertung der eingeschlossenen Studien ist nicht frei von Subjektivität. Dem haben wir versucht entgegenzuwirken, indem wir zum einen die veröffentlichten Erklärungen und Erläuterungen der Autor*innen der STARD-Checkliste genutzt haben, um die einzelnen Items für den Kontext unserer Analyse zu definieren. Zum anderen haben wir insgesamt vier Studien aus den Jahren 2014 und 2020, die in *European Radiology* und *Radiology* veröffentlicht wurden, mit der STARD-Checkliste vor der eigentlichen Analyse ausgewertet. So konnten wir Unstimmigkeiten zwischen den Auswerter*innen bezüglich der Definition einzelner Items bereits im Vorfeld diskutieren, sodass unsere eigentliche Analyse dadurch nicht beeinflusst werden konnte. Außerdem wurde auch die Analyse der eingeschlossenen diagnostischen Genauigkeitsstudien mit der STARD-Checkliste durch zwei unabhängige Auswerterinnen (A.S. und A.T.) durchgeführt, um eine Verzerrung der Ergebnisse durch eine subjektive Bewertung zu reduzieren. Auch an dieser Stelle half ein dritter Auswerter (B.K.) bei Uneinigkeit zwischen den beiden Auswerterinnen.

Des Weiteren haben wir uns dazu entschieden, für erfüllte Items einen Punkt zu vergeben und für Items, die nur eine unzureichende Berücksichtigung fanden, keinen Punkt zu vergeben. Damit haben wir einen anderen Ansatz als beispielsweise Zafar et al. verfolgt, die in ihrer Studie die einzelnen Items mit „vollständig erfüllt", „teilweise erfüllt" und „nicht erfüllt" bewertet haben (36). Dadurch können Items, die mehrere Punkte umfassen, differenzierter bewertet werden. Das betrifft zum Beispiel Item 8 (wo und wann potenziell einzuschließende Studienteilnehmer*innen gefunden wurden), das die Angabe von dem Setting, dem Ort und der Zeit der Selektion der Studienteilnehmer*innen erforderlich macht. Wir haben uns dennoch für eine dichotome Bewertung der Items der STARD-Checkliste entschieden, weil wir zum einen der Meinung sind, dass alle einzelnen Vorgaben eines Items wiedergegeben und erfüllt werden sollten. Zum anderen gibt es aber auch Items, die nur vollständig oder gar nicht erfüllt werden können und bei denen eine dritte Bewertungsoption dementsprechend nicht zielführend wäre. Ein Beispiel dafür ist Item 19 (Verwendung eines Flussdiagramms). Darüber hinaus erleichtert uns unser Vorgehen den Vergleich mit den Ergebnissen anderer wichtiger Studien zu diesem Thema, die sich ebenfalls für eine dichotome Bewertungsweise entschieden haben und unser Vorgehen bleibt auch näher an der Intention der Autor*innen der STARD-Checkliste (2).

In unserer Analyse haben wir uns ausschließlich auf zwei Jahre und zwei Journale konzentriert, da beide Journale bedeutende Quellen der radiologischen sowie wissenschaftlichen Literatur sind und beide Journale innerhalb unseres Untersuchungszeitraums Änderungen in ihren Einreichungsrichtlinien bezüglich der Verwendung der STARD-Checkliste vorgenommen haben. *European Radiology* hat die Empfehlung zur Verwendung der STARD-Checkliste 2017 ergänzt, während *Radiology* die Verwendung 2016 verpflichtend gemacht hat (20, 37). Dennoch könnte die Fokussierung auf lediglich zwei Journale die externe Validität unserer Studienergebnisse einschränken. Um diesem Problem entgegenzuwirken, haben wir das Thema unserer eingeschlossenen Studien im Gegensatz zu anderen Analysen nicht enger eingegrenzt.

Um dennoch weitere Aussagen treffen zu können, haben wir in dem vorliegenden Manteltext ergänzend die in *European Radiology* veröffentlichten diagnostischen Genauigkeitsstudien mit denen aus *Radiology* verglichen. Diese Analyse liefert weitere Informa-

tionen bezüglich der Qualität der Berichterstattung diagnostischer Genauigkeitsstudien und ist in keiner der beschriebenen Publikationen vorgenommen worden. Damit bildet sie eine spezifische Stärke des vorliegenden Manteltextes.

Des Weiteren haben wir Item 11 (Grund für die Wahl des Referenzstandards [wenn Alternativen existieren]) für diese Auswertung ausgeschlossen. Denn wenn Item 11 in einer Studie nicht wiedergegeben wurde, war es für uns nicht möglich einzuschätzen, ob Item 11 in einer Studie von den Autor*innen nicht berücksichtigt wurde, weil es keine Alternativen zu dem gewählten Referenzstandard gab oder weil es lediglich nicht erwähnt und damit nicht erfüllt wurde. Zudem entspricht dieses Vorgehen dem Ansatz von Wilczynski et al. (27). Dennoch können dadurch - abhängig von dem Abschneiden von Item 11 - die Ergebnisse unserer beiden Studien beeinflusst worden sein, was die Generalisierbarkeit unserer Aussagen einschränken kann. Leider kann die Richtung der Beeinflussung unserer Studienergebnisse durch dieses Vorgehen nur schwer eingeschätzt werden, da das Abschneiden von Item 11 in Studien anderer Autor*innengruppen sehr heterogen war (13, 33, 34, 38).

4.5 Implikationen für zukünftige Forschung

Zudem ist es wichtig zu betonen, dass wir uns in unseren Studien lediglich mit der Qualität der Berichterstattung auseinandergesetzt haben und keine Aussagen zu der methodologischen Qualität treffen können. Weitere aktuelle Analysen, die die methodologische Qualität diagnostischer Genauigkeitsstudien beurteilen, sind notwendig. Dafür steht beispielsweise das QUADAS-Tool zur Verfügung, das für systematische Übersichtsarbeiten entwickelt wurde, um unter anderem das Risiko für Bias in diagnostischen Genauigkeitsstudien einschätzen zu können (28). 2011 wurde eine aktualisierte Version des QUADAS-Tools veröffentlicht (39).

Schlussfolgerungen

Folgende Aussagen können wir nach unseren Analysen treffen:

1. Die Qualität diagnostischer Genauigkeitsstudien in der radiologischen Literatur ist nach wie vor moderat, schneidet jedoch im Vergleich zu anderen medizinischen Fachbereichen oftmals besser ab.
2. Eine ergänzende Empfehlung oder Verpflichtung zu der Verwendung der STARD-Checkliste hat in zwei führenden Journalen der radiologischen Literatur zu einer Verbesserung der Qualität der Berichterstattung geführt.
3. Auch war die Qualität der Berichterstattung in einem Journal, in dem die Verwendung der STARD-Checkliste verpflichtend ist, besser als in einem, das die Verwendung der STARD-Checkliste lediglich empfiehlt.

Herausgeber*innen können folglich einen großen Beitrag zur Verbesserung der Qualität der Berichterstattung leisten, wenn sie validierte Checklisten für Einreichungen in ihren Journalen verpflichtend machen. Aber auch Autor*innen sollten von sich aus an eine Verwendung einer entsprechenden Checkliste denken, zumal viele Items mit wenig Worten berücksichtigt werden können.

Literaturverzeichnis

1. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ. 2002; 324(7336):539.
2. Bossuyt P, Reitsma J, Bruns D, Gatsonis C, Glasziou P, Irwig L, Moher D, Rennie D, De Vet H, Lijmer J. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. Annals of internal medicine. 2003; 138:W1-12.
3. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med. 1981; 94(4 Pt 2):557-92.
4. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Bmj. 2003; 326(7379):41-4.
5. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ. 1986; 134(6):587-94.
6. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making. 1991; 11(2):88-94.
7. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of Variation and Bias in Studies of Diagnostic Accuracy: A Systematic Review. Annals of Internal Medicine. 2004; 140(3):189-202.
8. Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006; 174(4):469-76.
9. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ : British Medical Journal. 2015; 351:h5527.
10. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HC, Bossuyt PM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016; 6(11):e012799.
11. Smidt N, Rutjes AWS, van der Windt DAWM, Ostelo RWJG, Reitsma JB, Bossuyt PM, Bouter LM, de Vet HCW. Quality of reporting of diagnostic accuracy studies. Radiology. 2005; 235(2):347-53.
12. Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, Bossuyt PMM. Reporting Diagnostic Accuracy Studies: Some Improvements after 10 Years of STARD. Radiology. 2014; 274(3):781-9.
13. Choi YJ, Chung MS, Koo HJ, Park JE, Yoon HM, Park SH. Does the Reporting Quality of Diagnostic Test Accuracy Studies, as Defined by STARD 2015, Affect Citation? Korean J Radiol. 2016; 17(5):706-14.
14. Michelessi M, Lucenteforte E, Miele A, Oddone F, Crescioli G, Fameli V, Korevaar DA, Virgili G. Diagnostic accuracy research in glaucoma is still incompletely reported: An application of Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015. PloS one. 2017; 12(12):e0189716-e.
15. Hong PJ, Korevaar DA, McGrath TA, Ziai H, Frank R, Alabousi M, Bossuyt PMM, McInnes MDF. Reporting of imaging diagnostic accuracy studies with focus on

MRI subgroup: Adherence to STARD 2015. Journal of Magnetic Resonance Imaging. 2018; 47(2):523-44.
16. Zarei F, Zeinali-Rafsanjani B. Assessment of Adherence of Diagnostic Accuracy Studies Published in Radiology Journals to STARD Statement Indexed in Web of Science, PubMed & Scopus in 2015. J Biomed Phys Eng. 2018; 8(3):311-24.
17. Hogan KO, Fraga GR. Compliance With Standards for STARD 2015 Reporting Recommendations in Pathology. American Journal of Clinical Pathology. 2020; 154(6):828-36.
18. Zheng FF, Shen WH, Gong F, Hu ZD, Lippi G, Šimundić AM, Bossuyt PMM, Plebani M, Zhang K. Adherence to the Standards for Reporting of Diagnostic Accuracy Studies (STARD): a survey of four journals in laboratory medicine. Ann Transl Med. 2021; 9(11):918.
19. Prager R, Bowdridge J, Kareemi H, Wright C, McGrath TA, McInnes MDF. Adherence to the Standards for Reporting of Diagnostic Accuracy (STARD) 2015 Guidelines in Acute Point-of-Care Ultrasound Research. JAMA Netw Open. 2020; 3(5):e203871.
20. Levine D, Kressel HY. Radiology 2016: The Care and Scientific Rigor Used to Process and Evaluate Original Research Manuscripts for Publication. Radiology. 2015; 278(1):6-10.
21. Stahl AC, Tietz AS, Kendziora B, Dewey M. Has the STARD statement improved the quality of reporting of diagnostic accuracy studies published in European Radiology? Eur Radiol. 2023; 33(1):97-105.
22. Stahl AC, Tietz AS, Dewey M, Kendziora B. Has the quality of reporting improved since it became mandatory to use the Standards for Reporting Diagnostic Accuracy? Insights Imaging. 2023; 14(1):85.
23. Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, Stewart L. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Syst Rev. 2012; 1:2.
24. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine. 2009; 6(7):e1000097.
25. Devillé WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol. 2000; 53(1):65-9.
26. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HCW, Bossuyt PMM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016; 6(11):e012799-e.
27. Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study. Radiology. 2008; 248(3):817-23.
28. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Medical Research Methodology. 2003; 3(1):25.
29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159-74.
30. Rethlefsen ML, Page MJ. PRISMA 2020 and PRISMA-S: common questions on tracking records and the flow diagram. J Med Libr Assoc. 2022; 110(2):253-7.
31. RSNA. About Radiology. 2022.

https://pubs.rsna.org/page/radiology/about Accessed June 04, 2023.
32. Springer. European Radiology. 2022. https://www.springer.com/journal/330 Accessed June 04, 2023.
33. Wright B, Howard B, Wayant C, Vassar M. STARD Adherence in an Interventional Radiology Guideline for Diagnostic Arteriography. Clin Med Res. 2021; 19(1):26-31.
34. Nassar EL, Levis B, Neyer MA, Rice DB, Booij L, Benedetti A, Thombs BD. Transparency and completeness of reporting of depression screening tool accuracy studies: A meta-research review of adherence to the Standards for Reporting of Diagnostic Accuracy Studies statement. Int J Methods Psychiatr Res. 2023; 32(1):e1939.
35. Jang MA, Kim B, Lee YK. Reporting Quality of Diagnostic Accuracy Studies in Laboratory Medicine: Adherence to Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015. Ann Lab Med. 2020; 40(3):245-52.
36. Zafar A, Khan GI, Siddiqui MA. The quality of reporting of diagnostic accuracy studies in diabetic retinopathy screening: a systematic review. Clin Exp Ophthalmol. 2008; 36(6):537-42.
37. Dewey M, Levine D, Bossuyt PM, Kressel HY. Impact and perceived value of journal reporting guidelines among Radiology authors and reviewers. European Radiology. 2019; 29(8):3986-95.
38. Smith DW, Gandhi S, Dahm P. The reporting quality of studies of diagnostic accuracy in the urologic literature. World J Urol. 2019; 37(5):969-74.
39. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MMG, Sterne JAC, Bossuyt PMM, the Q-G. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Annals of Internal Medicine. 2011; 155(8):529-36.

Eidesstattliche Versicherung

„Ich, Ann-Christine Stahl, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: *„Qualität der Berichterstattung diagnostischer Genauigkeitsstudien in der radiologischen Literatur“/„Quality of reporting of diagnostic accuracy studies published in radiological medical journals“* selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autor*innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem Erstbetreuer, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

Anteilserklärung an den erfolgten Publikationen

Ann-Christine Stahl hatte folgenden Anteil an den folgenden Publikationen:

Publikation 1: Ann-Christine Stahl, Anne-Sophie Tietz, Benjamin Kendziora, Marc Dewey, Has the STARD statement improved the quality of reporting of diagnostic accuracy studies published in *European Radiology?*, European Radiology, 2023
Beitrag im Einzelnen:

1. Vollständige Studienselektion als eine der beiden unabhängigen Auswerterinnen. Daraus ist Abbildung 1 dieser ersten Publikation entstanden.
2. Analyse aller eingeschlossenen Studien mit Hilfe der aktuellen STARD-Checkliste als eine der beiden unabhängigen Auswerterinnen.
3. Aufbereitung der Daten für die statistische Analyse.
4. Statistische Analyse der gewonnen Daten unter Supervision von Dr. Benjamin Kendziora.
5. Eigenständiges Erstellen der Tabellen 1, 2, 3 und A1 sowie der Abbildung 2 dieser ersten Publikation.
6. Ausführliche Literaturrecherche zum aktuellen Forschungsstand der Qualität diagnostischer Genauigkeitsstudien.
7. Verfassen des Manuskriptes unter Supervision von Dr. Benjamin Kendziora und Prof. Marc Dewey. Anne-Sophie Tietz hat die finale Version des Manuskriptes gelesen und erklärte sich mit ihr einverstanden.
8. Durchführen des Submission-Prozesses.
9. Ausführliche Bearbeitung der Revision unter Supervision von Dr. Benjamin Kendziora.

Publikation 2: Ann-Christine Stahl, Anne-Sophie Tietz, Marc Dewey, Benjamin Kendziora, Has the quality of reporting improved since it became mandatory to use the Standards for Reporting Diagnostic Accuracy?, Insights into Imaging, 2023
Beitrag im Einzelnen:

1. Vollständige Studienselektion als eine der beiden unabhängigen Auswerterinnen.
2. Analyse aller eingeschlossenen Studien mit Hilfe der aktuellen STARD-Checkliste als eine der beiden unabhängigen Auswerterinnen.
3. Bereitstellung der Daten für die statistische Analyse.
4. Eigenständiges Erstellen der Abbildung im Graphical Abstract.
5. Intensive Überarbeitung des von Anne-Sophie Tietz entworfenen Manuskriptes unter Supervision von Dr. Benjamin Kendziora und Prof. Marc Dewey.
6. Durchführen des Submission-Prozesses.
7. Ausführliche Bearbeitung der Revision unter Supervision von Dr. Benjamin Kendziora.

_____

Unterschrift, Datum und Stempel des erstbetreuenden Hochschullehrers

_____

Unterschrift der Doktorandin

Druckexemplare der Publikationen

Erste Publikation

Stahl AC, Tietz AS, Kendziora B, Dewey M. Has the STARD statement improved the quality of reporting of diagnostic accuracy studies published in *European Radiology*? *Eur Radiol 2023;* 1: 97-105
https://doi.org/10.1007/s00330-022-09008-7

Zweite Publikation

Stahl AC, Tietz AS, Dewey M, Kendziora B. Has the Quality of Reporting Improved since it Became Mandatory to Use the Standards for Reporting Diagnostic Accuracy? *Insights Imaging 2023;* 1: 85
https://doi.org/10.1186/s13244-023-01432-7

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Komplette Publikationsliste

1. **Stahl AC**, Tietz AS, Kendziora B, Dewey M. Has the STARD statement improved the quality of reporting of diagnostic accuracy studies published in *European Radiology*? *Eur Radiol 2023;* 1: 97-105
IF 2022: 5.9
DOI: 10.1007/s00330-022-09008-7

2. **Stahl AC**, Tietz AS, Dewey M, Kendziora B. Has the Quality of Reporting Improved since it Became Mandatory to Use the Standards for Reporting Diagnostic Accuracy? *Insights Imaging 2023;* 1: 85
IF 2022: 4.7
DOI: 10.1186/s13244-023-01432-7

Danksagung

Mein besonderer Dank gilt Herrn Prof. Dewey, der für mich - auch über diese Dissertation hinaus - ein großartiger Mentor war und ist. Darüber hinaus möchte ich mich bei allen Koautor*innen und dem ganzen Team Dewey für die wunderbare Zusammenarbeit und Unterstützung bedanken.
Mein Dank gilt auch meiner Familie und meinen Freund*innen, die jederzeit ein offenes Ohr für mich hatten.