Performance and change impact analysis of a commercial artificial intelligence tool for radiographic knee osteoarthritis grading and joint space width measuring

Study acronym: AutoRayValid-RBknee-validation

Study protocol version: This document has been written based on the information contained in the study protocol of AutoRayValid-RBknee study, version 1, dated August 30, 2022[1]

SAP version number with date: Version 1.0 – [2024-mar-13]

SAP revision history, justifications, and timing: SAP version 1

Author/chief investigator: Mathias Willadsen Brejnebøl[1], MD

Signature and date:  13-03-2024

.....

Other contributors:

Mikael Boesen[1], investigator

Kay Geert A. Hermann[2], investigator

Edwin Oei[3], investigator

Huib Ruitenbeek[3], investigator

Katharina Ziegeler[2], investigator

Jacob J. Visser[3], investigator

Anders Lenskjold[1], investigator

Philip Hansen[1], investigator

Janus Uhd Nybing[1], investigator

Affiliations:

1. Department of Radiology, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark
2. Department of Radiology, Charité Universitätsmedizin, Berlin, Germany
3. Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands

Funding:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 954221 for the EIC SME Instrument project AutoRay. The work only reflects the authors' view and the European Commission is not responsible for any use that may be made from the information it contains.

1. ABSTRACT

Background and rationale: Knee osteoarthritis (OA) is a common disease characterized by reduced function, stiffness, and pain. This clinical diagnosis is commonly supported with radiography of the weight-bearing knee. Radiographic features, such as the Kellgren-Lawrence (KL) grading system, are used as eligibility criteria for clinical studies while others, such as the OARSI grades and minimal joint space width, are used as endpoints for structural OA progression. A higher preoperative KL-grade has been correlated with better pain- and functional outcomes after knee arthroplasty surgery. Consequently, the KL-grade is a common requirement for approving knee arthroplasty among American health insurance providers and it is commonly used by orthopedic surgeons as part of determining knee arthroplasty candidacy.

Historically, a radiologist was required to draw on and grade radiographs of the knee to extract the features. With increasing computational power and the increased use of deep convolutional neural networks, off-the-shelf artificial intelligence (AI) tools have become available for automatic extraction of these features. They have received regulatory approval for commercialization but it is apparent that more diligent external validation is required. Finally, as AI tools begin to mature, new versions are released. It is important to assess how these developments change the current performance of the tool.

Objectives: The aim of this analysis is to evaluate the performance of a commercially available AI tool for grading tibiofemoral OARSI grades, KL grades and patellar osteophytes as well as the accuracy of measuring joint space width. Additionally, a change impact analysis will be performed where the performance of the current version of the AI tool will be compared to that of the previous version.

Methods: This study is a secondary analysis of the data from the AutoRayValid-RBknee study, a retrospective observer performance study. It consists of non-fixed-flexion radiographs acquired from the production picture archiving and communications system (PACS) from three European centers. Root mean square error (RMSE) will be used for estimating the accuracy of minimal and fixed-location joint space width (JSW) measurements. Ordinal ROC will be used for estimating ordinal OARSI-grade and the KL-grade classification AUC. Area under the receiver operating curve (AUC) is used for estimating binary OARSI-grade and patellar osteophyte classification performance.

Population:
Patients with knee pain referred for radiography on suspicion of knee osteoarthritis

Index test:
RBknee-2.2.0 (CE version, KL-grading, OARSI grading, patellar osteophytes) and RBknee-fda-1.0.1 (FDA version, Joint Space Width measurement). RBknee-2.1.0 (CE version, KL-grading, OARSI grading, patellar osteophytes) will be used to perform the change impact analysis of advancing product development.

Reference test:
For all discrete variables, the reference value will be the majority vote, arbitrated by consensus where grades differ by 2 or more. The readers will be three board-certified musculoskeletal radiologists with substantial clinical and research experience. For continuous variables, annotation will be done by a single radiologist trained in the task. The annotations will be reviewed by a board-certified musculoskeletal radiologist with substantial clinical and research experience.

Further statistical details
Sample size: Not applicable as this is a secondary analysis.

Framework: This is a diagnostic test accuracy study assessing the performance of a commercially available AI tool for radiographic evaluation of knee osteoarthritis according to established grading systems. Additionally, change impact analysis will be performed where multiple versions of the AI tool are available.

Confidence intervals and P values: All 95% confidence intervals and P values will use an alpha of 5%.

Multiplicity: No explicit multiplicity correction will be performed. Instead a hierarchical approach will be taken based on tabular order of the tested hypotheses.

Statistical software: R version 4.2.2 (or newer).

2. ELABORATIONS ON OUTCOMES AND DATA

Data management:
Values outputted by the AI tool will be compared to the reference standard. Additionally, subgroup analyses will be based on image conformity and acquisition type.

Continuous variables:

Minimal Joint Space Width (mJSW):

Medial (mmJSW) and lateral (lmJSW) compartments.

If bone-on-bone configuration is true for the compartment, mJSW is 0. Otherwise, it is calculated as the smallest distance between Femoral Condyle and the Tibial Plateau annotations for each compartment.

Fixed-location Joint Space Width:

Fixed-location Joint Space Width will be acquired for the medial (mfJSW) and lateral (lfJSW) compartments as described by Neumann and Duryea et al [2–4].

Discrete variables:

OARSI grades:

On the frontal image. Joint space narrowing (ordinal grades 0 = normal joint space, 3 = more than 2/3 narrowed) for the medial and lateral compartments, osteophytes (ordinal grades 0-3) for the medial and lateral femur and tibia and tibial eminence, subchondral sclerosis (yes/no) for the medial and lateral femur and tibia[5].

Kellgren-Lawrence grade:

On the frontal image (ordinal grades 0 = no osteoarthritis, 4 = severe osteoarthritis)[6].

Patellar osteophytes:

On the lateral image. Proximal and distal patellar osteophytes (yes/no).

Image nonconformity (not outputted by the index test):

Frontal images with rotation, angulation and/or inadequate for medial or lateral estimation (all binary). The groups are non-exclusive. An image can be both rotated and angled while only the lateral compartment is inadequate.

Data validation:
All variables used in the analyses, including the derived variables, will be checked for missing values, outliers, and inconsistencies.

Data template:
Based on this SAP, the statistical analyst will develop a tailored data template illustrating the data structure required for the statistical analyses.

3. OUTLINE

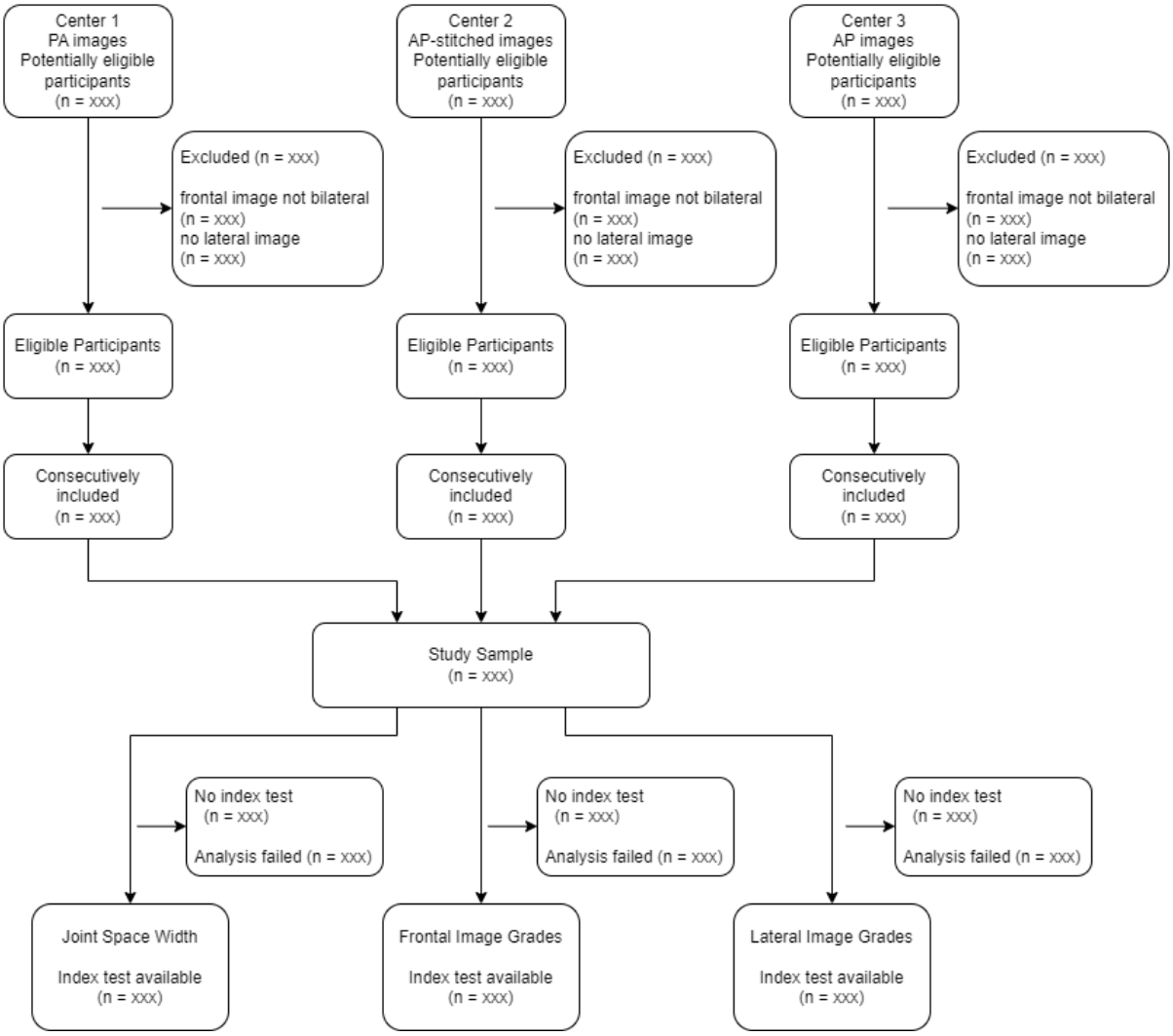
One-against-all AUC for the ordinary variables and mean absolute error for the continuous variables will be presented in manuscript body only.

Comparison of AUC for the current and previous AI tool versions for image nonconformity and acquisition type will be presented in manuscript body only.

The anticipated (predefined) outline of the manuscript is illustrated below.

Figure 1. Flow diagram

Anticipated plot design, illustrating potential reasons for exclusion:



PA: posteroanterior; AP: anteroposterior;

Table 1. Characteristics in the AutoRayValid-RBknee population

Characteristics	Distance	0 / no (n =)	1 / yes (n =)	2 (n =)	3 (n =)	4 (n =)	Total (n =)
Female sex, no (%)							
Age							
Kellgren-Lawrence, no (%)							
Minimal JSW, mm							
Medial							
Lateral							
Fixed JSW, mm							
Medial							
Lateral							
OARSI JSN, no (%)							
Medial							
Lateral							
OARSI Osteophytes, no (%)							
Femur Medial							
Femur Lateral							
Tibia Medial							
Tibia Lateral							
Tibia Eminence							
OARSI Subchondral Sclerosis, no (%)							
Femur Medial							
Femur Lateral							
Tibia Medial							
Tibia Lateral							
Patellar Osteophytes, no (%)							
Proximal							
Distal							
Analysis Error, no (%)							
Frontal view (CE)							
Frontal view (FDA)							
Lateral view (CE)							
Inadequacies, no (%)							
Rotated							
Angled							
Medial inadequate							
Lateral inadequate							

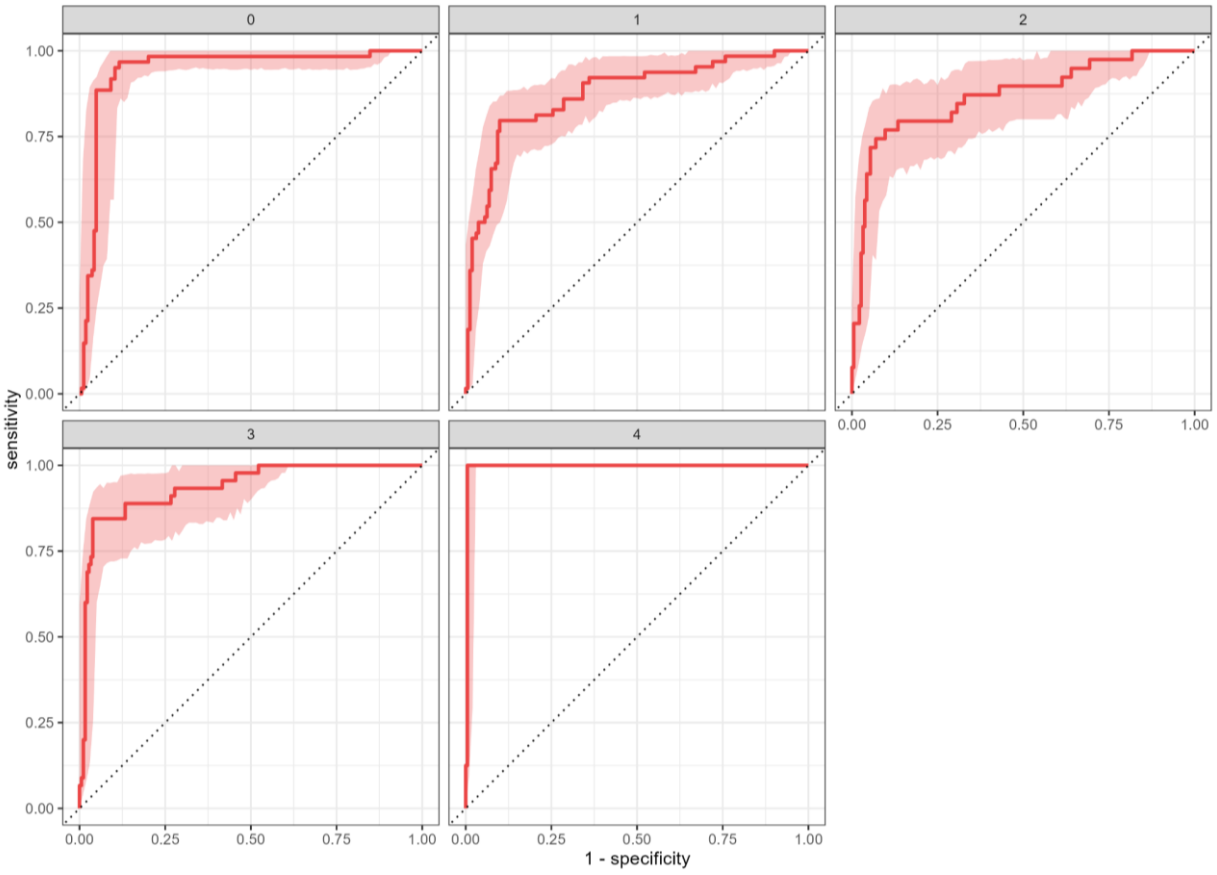
Values are mean (SD) unless otherwise stated in the table.
JSW, Joint Space Width; OARSI, Osteoarthritis Research Society International; JSN, Joint Space Narrowing; CE, *conformité européenne*; FDA, Food and Drug Association

Further statistical information related to Table 1:
Data will be presented as means with standard deviations (SD) when normally distributed or as medians with interquartile range in case of skewed data. Dichotomous and categorical data will be presented as absolute counts and proportions.

Figure 2. Receiver Operating Curve for classifying radiographic knee osteoarthritis (MockUp)

Hypothetical ROC for the AI tool:

The FPR-TPR relationship of the AI tool for one against all KL-grading is plotted for various prediction thresholds. The solid red line is the ROC for the AI tool and the light red bands are the 95% confidence interval. The x-axis is the TPR/specificity and the y-axis is the FPR/ 1- sensitivity.



Further statistical information related to Figure 2:

TPR, true positive rate; FPR, false positive rate. The diagonal dashed line from (0, 0) to (1, 1) indicates classification performance equal to chance.

Table 2. Performance of the AI tool

Characteristics	Current	Previous	P
Minimal JSW (RMSE)			
Medial		N/A	
Lateral		N/A	
Fixed JSW (RMSE)			
Medial		N/A	
Lateral		N/A	
Kellgren-Lawrence grade (AUC)			
OA diagnosis (KL >= 2) (AUC)			
OARSI JSN (AUC):			
Medial			
Lateral			
OARSI Osteophytes (AUC)			
Femur Medial			
Femur Lateral			
Tibia Medial			
Tibia Lateral			
Tibia Eminence			
OARSI Subchondral Sclerosis (AUC)			
Femur Medial			
Femur Lateral			
Tibia Medial			
Tibia Lateral			
Patellar Osteophytes (AUC)			
Proximal			
Distal			

Values will be reported as least squared means (standard error) unless noted otherwise in the table. Current refers to the newest version of the tested AI tool and previous refers to the second-to-newest version, where applicable. OA, Osteoarthritis; OARSI, Osteoarthritis Research Society International; mmJSW, medial minimal Joint Space Width; mfJSW, medial fixed-location Joint Space Width; lmJSW, lateral minimal Joint Space Width; lfJSW, lateral fixed-location Joint Space Width; RMSE, root mean squared error; ord.acc, accuracy when the reference standard is ordinal; AUC, area under the receiver operating curve.

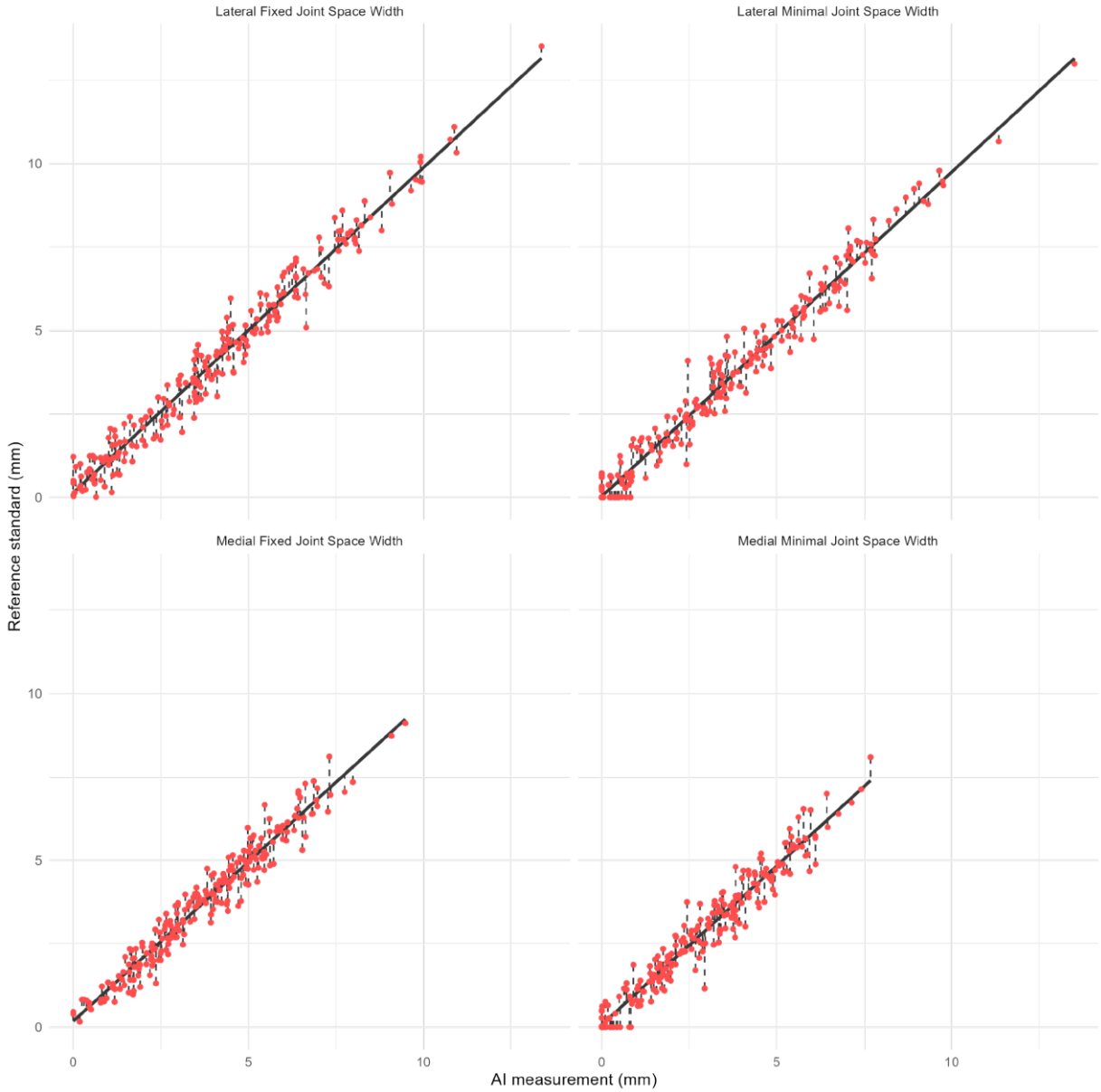
Further statistical information related to Table 2:

For continuous variables, diagnostic accuracy will be estimated as the root mean squared error since large deviations should be penalized most. For ordinal variables, performance will be estimated as the AUC and compared across groups using the ordinal ROC method as proposed by Obuchowski et al. For binary variables, performance will be estimated as the AUC of the ROC and groups will be compared using the DeLong method.

Figure 3. Root mean square error of the AI tool in predicting the correct joint space width (MockUp)

Hypothetical RMSE plot for the AI tool:

Plots of the root mean squared error for the AI tool's prediction of mmJSW, lmJSW, mfJSW, and lfJSW compared to the reference standard. The solid line is the linear model of reference value as a function of the AI tool predicted value. The red points are the actual AI tool predictions and the corresponding dashed lines link these predictions to the fitted line. The abscissa is the AI tool prediction in mm and the ordinate is the reference standard in mm.



Further statistical information related to Figure 3:

mmJSW, medial minimal joint space width; mlJSW, medial lateral joint space width; mfJSW, medial fixed-location joint space width; lfJSW, lateral fixed-location joint space width;

Table 3. Performance of the AI tool for image nonconformity subgroups

Characteristics	Rotated	Angulated	Medial Inadequate	Lateral inadequate	Optimal
Minimal JSW (RMSE)					
Medial					
Lateral					
Fixed JSW (RMSE)					
Medial					
Lateral					
Kellgren-Lawrence grade (AUC)					
OA diagnosis (KL >= 2) (AUC)					
OARSI JSN (AUC):					
Medial					
Lateral					
OARSI Osteophytes (AUC)					
Femur Medial					
Femur Lateral					
Tibia Medial					
Tibia Lateral					
Tibia Eminence					
OARSI Subchondral Sclerosis (AUC)					
Femur Medial					
Femur Lateral					
Tibia Medial					
Tibia Lateral					
Patellar Osteophytes (AUC)					
Proximal					
Distal					

Values will be reported as least squared means (standard error) unless noted otherwise in the table.
OA, Osteoarthritis; OARSI, Osteoarthritis Research Society International; mmJSW, medial minimal Joint Space Width; mfJSW, medial fixed-location Joint Space Width; lmJSW, lateral minimal Joint Space Width; lfJSW, lateral fixed-location Joint Space Width; RMSE, root mean squared error; ord.acc, accuracy when the reference standard is ordinal; AUC, area under the receiver operating curve.

Further statistical information related to Table 3:
For continuous variables, diagnostic accuracy will be estimated as the root mean squared error since large deviations should be penalized most. For ordinal variables, performance will be estimated as the AUC and compared across groups using the ordinal ROC method as proposed by Obuchowski et al. For binary variables, performance will be estimated as the AUC of the ROC and groups will be compared using the DeLong method.

Table 4. Performance of the AI tool for acquisition type subgroups

Characteristics	PA	AP-stitched	AP
Minimal JSW (RMSE)			
Medial			
Lateral			
Fixed JSW (RMSE)			
Medial			
Lateral			
Kellgren-Lawrence grade (AUC)			
OA diagnosis (KL >= 2) (AUC)			
OARSI JSN (AUC):			
Medial			
Lateral			
OARSI Osteophytes (AUC)			
Femur Medial			
Femur Lateral			
Tibia Medial			
Tibia Lateral			
Tibia Eminence			
OARSI Subchondral Sclerosis (AUC)			
Femur Medial			
Femur Lateral			
Tibia Medial			
Tibia Lateral			
Patellar Osteophytes (AUC)			
Proximal			
Distal			

Values will be reported as least squared means (standard error) unless noted otherwise in the table.
OA, Osteoarthritis; OARSI, Osteoarthritis Research Society International; mmJSW, medial minimal Joint Space Width; mfJSW, medial fixed-location Joint Space Width; lmJSW, lateral minimal Joint Space Width; lfJSW, lateral fixed-location Joint Space Width; RMSE, root mean squared error; ord.acc, accuracy when the reference standard is ordinal; AUC, area under the receiver operating curve; PA: posteroanterior; AP: anteroposterior;

Further statistical information related to Table 4:
For continuous variables, diagnostic accuracy will be estimated as the root mean squared error since large deviations should be penalized most. For ordinal variables, performance will be estimated as the AUC and compared across groups using the ordinal ROC method as proposed by Obuchowski et al. For binary variables, performance will be estimated as the AUC of the ROC and groups will be compared using the DeLong method.

SUPPLEMENTARY MATERIAL

The anticipated (predefined) supplementary material of the manuscript is illustrated below.

Supplementary file 1. Protocol[1]

Supplementary file 2. This SAP

Supplementary file 3. Primary study publication

4. REFERENCES

1. Brejnebøl MW, Egnell L, Lundemann M, et al (2022) Protocol for the AutoRayValid-RBknee study: A retrospective, multicenter, fully-crossed, multi-reader, multi-case study investigating the effect of a knee osteoarthritis severity classification model on reader diagnostic accuracy. bioRxiv

2. Neumann G, Hunter D, Nevitt M, et al (2009) Location specific radiographic joint space width for osteoarthritis progression. Osteoarthritis Cartilage 17:761–765. https://doi.org/10.1016/j.joca.2008.11.001

3. Duryea J, Neumann G, Niu J, et al (2010) Comparison of radiographic joint space width with magnetic resonance imaging cartilage morphometry: analysis of longitudinal data from the Osteoarthritis Initiative. Arthritis Care Res 62:932–937. https://doi.org/10.1002/acr.20148

4. Benichou OD, Hunter DJ, Nelson DR, et al (2010) One-year change in radiographic joint space width in patients with unilateral joint space narrowing: data from the Osteoarthritis Initiative. Arthritis Care Res 62:924–931. https://doi.org/10.1002/acr.20149

5. Altman RD, Gold GE (2007) Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis Cartilage 15:1–56. https://doi.org/10.1016/j.joca.2006.06.017

6. Kellgren JH, Lawrence JS (1957) Radiological assessment of osteo-arthrosis. Ann Rheum Dis 16:494–502. https://doi.org/10.1136/ard.16.4.494

7. Bossuyt PM, Reitsma JB, Bruns DE, et al (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 351:h5527. https://doi.org/10.1136/bmj.h5527

8. Cohen JF, Korevaar DA, Altman DG, et al (2016) STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 6:e012799. https://doi.org/10.1136/bmjopen-2016-012799

5. SAP REPORTING GUIDELINE

This SAP has been reported according to the items recommended in STARD guidelines by Cohen et al.[7] Explanation and elaboration of the items are available in the appendix paper.[8]

The guideline checklist and motivation is reproduced below.

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	
	4	Study objectives and hypotheses	
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	
Participants	**6**	Eligibility criteria	
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	
	8	Where and when potentially eligible participants were identified (setting, location and dates)	
	9	Whether participants formed a consecutive, random or convenience series	
Test methods	**10a**	Index test, in sufficient detail to allow replication	
	10b	Reference standard, in sufficient detail to allow replication	
	11	Rationale for choosing the reference standard (if alternatives exist)	
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	
	15	How indeterminate index test or reference standard results were handled	
	16	How missing data on the index test and reference standard were handled	
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	
	18	Intended sample size and how it was determined	
RESULTS			
Participants	**19**	Flow of participants, using a diagram	
	20	Baseline demographic and clinical characteristics of participants	
	21a	Distribution of severity of disease in those with the target condition	
	21b	Distribution of alternative diagnoses in those without the target condition	
	22	Time interval and any clinical interventions between index test and reference standard	
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	
	25	Any adverse events from performing the index test or the reference standard	

DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	
	27	Implications for practice, including the intended use and clinical role of the index test	
OTHER INFORMATION			
	28	Registration number and name of registry	
	29	Where the full study protocol can be accessed	
	30	Sources of funding and other support; role of funders	

STARD 2015

AIM

STARD stands for "Standards for Reporting Diagnostic accuracy studies". This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

Explanation

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition.** This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test.** A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or "2x2" table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on http://www.equator-network.org/reporting-guidelines/stard.