Automated Knee Osteoarthritis Assessment Increases Physicians' Agreement Rate and Accuracy: Data from the Osteoarthritis Initiative

Stefan Nehrer[1] , Richard Ljuhar[2], Peter Steindl[3], Rene Simon[3], Dietmar Maurer[4], Davul Ljuhar[5], Zsolt Bertalan[2], Hans P. Dimai[4], Christoph Goetz[2], and Tiago Paixao[2]

Abstract

Objective. To assess the impact of a computerized system on physicians' accuracy and agreement rate, as compared with unaided diagnosis. *Methods*. A set of 124 unilateral knee radiographs from the Osteoarthritis Initiative (OAI) study were analyzed by a computerized method with regard to Kellgren-Lawrence (KL) grade, as well as joint space narrowing, osteophytes, and sclerosis Osteoarthritis Research Society International (OARSI) grades. Physicians scored all images, with regard to osteophytes, sclerosis, joint space narrowing OARSI grades and KL grade, in 2 modalities: through a plain radiograph (*unaided*) and a radiograph presented together with the report from the computer assisted detection system (*aided*). Intraclass correlation between the physicians was calculated for both modalities. Furthermore, physicians' performance was compared with the grading of the OAI study, and accuracy, sensitivity, and specificity were calculated in both modalities for each of the scored features. *Results*. Agreement rates for KL grade, sclerosis, and osteophyte OARSI grades, were statistically increased in the aided versus the unaided modality. Readings for joint space narrowing OARSI grade did not show a statistically difference between the 2 modalities. Readers' accuracy and specificity for KL grade >0, KL >1, sclerosis OARSI grade >0, and osteophyte OARSI grade >0 was significantly increased in the aided modality. Reader sensitivity was high in both modalities. *Conclusions*. These results show that the use of an automated knee OA software increases consistency between physicians when grading radiographic features of OA. The use of the software also increased accuracy measures as compared with the OAI study, mostly through increases in specificity.

Keywords

Kellgren-Lawrence, computer aided detection, reader study, artificial intelligence

Introduction

Radiographic classification of osteoarthritis (OA) in the knee has typically been performed using semiquantitative grading schemes,[1] the most widely used of which being the Kellgren-Lawrence (KL) scale,[2] which was recognized by the World Health Organization in 1961 as the standard for clinical studies of OA. The KL grading scheme requires the assessment of presence and severity degree of several individual radiographic features (IRFs), including osteophytes, sclerosis, and joint space narrowing (JSN). These assessments are them summarized into a 5-point scale, reflecting the severity of OA. However, the KL grading scheme has come under criticism for assuming a unique progression mode of OA[3] and for depending on subjective assessments,[4,5] exacerbated by the vague verbal definitions of IRFs at each stage.[6] In order to deal with these issues, the

Osteoarthritis Research Society International (OARSI) proposed a classification system for each of the IRFs supported by a reference atlas, in which canonical examples of the classification of each of the IRFs are depicted.[7]

One of the main purposes of a systematic OA grading scheme, such as the KL and the OARSI systems, is to standardize diagnostic and assessments of OA. However, several

[1]Danube University Krems, Krems, Austria
[2]ImageBiopsy Lab, Vienna, Austria
[3]Landesklinikum Neunkirchen, Neunkirchen, Austria
[4]Medical University of Graz, Graz, Austria
[5]Braincon Technologies, Vienna, Austria

Corresponding Author:
Stefan Nehrer, Danube University Krems, Dr. Karl-Dorrek-Strasse 30, Krems, 3500, Austria.
Email: stefan.nehrer@donau-uni.ac.at

Table 1. Population Demographics of the Individuals Present in the Study.

	Female, *n* (%)	Male, *n* (%)	Total, *n* (%)
Age (years)			
45-49	0 (0.00)	5 (8.77)	5 (4.17)
50-59	20 (31.75)	14 (24.56)	34 (28.33)
60-69	23 (36.51)	15 (26.32)	38 (31.67)
70-79	17 (26.98)	19 (33.33)	36 (30.00)
80-89	3 (4.76)	4 (7.02)	7 (5.83)
Total	63 (100.00)	57 (100.00)	120 (100.00)
Ethnicity			
Asian	0 (0.00)	1 (1.75)	1 (0.83)
Black or African American	15 (23.81)	9 (15.79)	24 (20.00)
Other non-white	1 (1.59)	0 (0.00)	1 (0.83)
White or Caucasian	47 (74.60)	47 (82.46)	94 (78.33)
Total	63 (100.00)	57 (100.00)	120 (100.00)
Body mass index (kg/m^2)			
20-25	12 (19.05)	12 (21.05)	24 (20.00)
25-30	21 (33.33)	27 (47.37)	48 (40.00)
30-35	17 (26.98)	14 (24.56)	31 (25.83)
35-40	12 (19.05)	4 (7.02)	16 (13.33)
40-45	1 (1.59)	0 (0.00%)	1 (0.83)
Total	63 (100.00)	57 (100.00)	120 (100.00)

studies report that the KL grading scheme, as well as the accessory assessments, suffer from subjectivity and low interobserver reliability.[8,9] This leads to differences in assessments of the prevalence of the disease[4] and variability of diagnoses of the same patient. This is especially problematic for the early stages of the disease: Severe forms of OA are easily recognized in radiographs, but its early stages are less consensual.[10] In part this stems from the high degree of subjectivity of the assessments,[11] even with the guidance of the OARSI atlas. This problem has consequences at several levels: In clinical practice, it can lead to misdiagnosis, leading to unnecessary examination procedures or omitted treatment, and psychological stress to the patient.[12] In the context of clinical trials, the variability of assessments can decrease the power to detect statistical effects of the efficacy of treatments[13] and complicate the estimation of prevalence and incidence rates.[14]

One potential, albeit not practical, solution for the problem of variability of diagnosis would be to have the same radiograph reviewed by several physicians and to have a procedure to determine consensus, as it is done when establishing the gold-standard readings in many clinical studies. This is clearly not a practical solution for clinical practice. However, one way to approach such a problem could be make use of a computer assisted detection system to standardize the readings of the relevant features. Artificial intelligence, and especially deep learning, has proven remarkably efficient at recognizing complex visual patterns. When applied to medical imaging, these systems can provide guidance and recommendations for radiographic assessments to the reader in a robust fashion. These artificial intelligence systems can be trained on the assessments of several clinicians (or the consensus readings after several physicians have reviewed the case) and so incorporate the experience of several clinicians and could potentially simulate a consensus procedure. Here we take this latter approach.

We make use of a computer-assisted detection system (KOALA, IB Lab GmbH) that was trained in a large dataset of radiographs graded for KL and JSN, sclerosis, and osteophyte OARSI grades through a consensus procedure. KOALA makes use of deep learning networks to provide fully automated KL and OARSI grades in the form of a report. Here, we assess how the use of this computer assisted detection system affects physicians' performance in terms of inter-observer variability at assessing KL grade and IRFs, as well as their accuracy performance at detecting several clinically relevant conditions.

Materials and Methods

Subjects

The Osteoarthritis Initiative (OAI) study (https://oai.epi-ucsf.org/) is a large longitudinal study conducted by 5 U.S. institutions. Among other outputs, the study collected knee radiographs of about 4,500 patients, over a period of 8 years. In addition, the study also provided consensus readings for KL grade, as well as JSN, osteophytes, and subchondral sclerosis OARSI grades. These readings were obtained through a consensus reading protocol which included adjudication procedures for discrepancies between readers.

Table 2. Distribution of Kellgren-Lawrence (KL) and Osteoarthritis Research Society International (OARSI) Grades in the Population.

	0	1	2	3	4
KL grade	24	20	35	29	16
Joint space narrowing OARSI grade	47	32	29	16	—
Sclerosis OARSI grade	62	26	27	9	—
Osteophyte OARSI grade	30	43	17	34	—
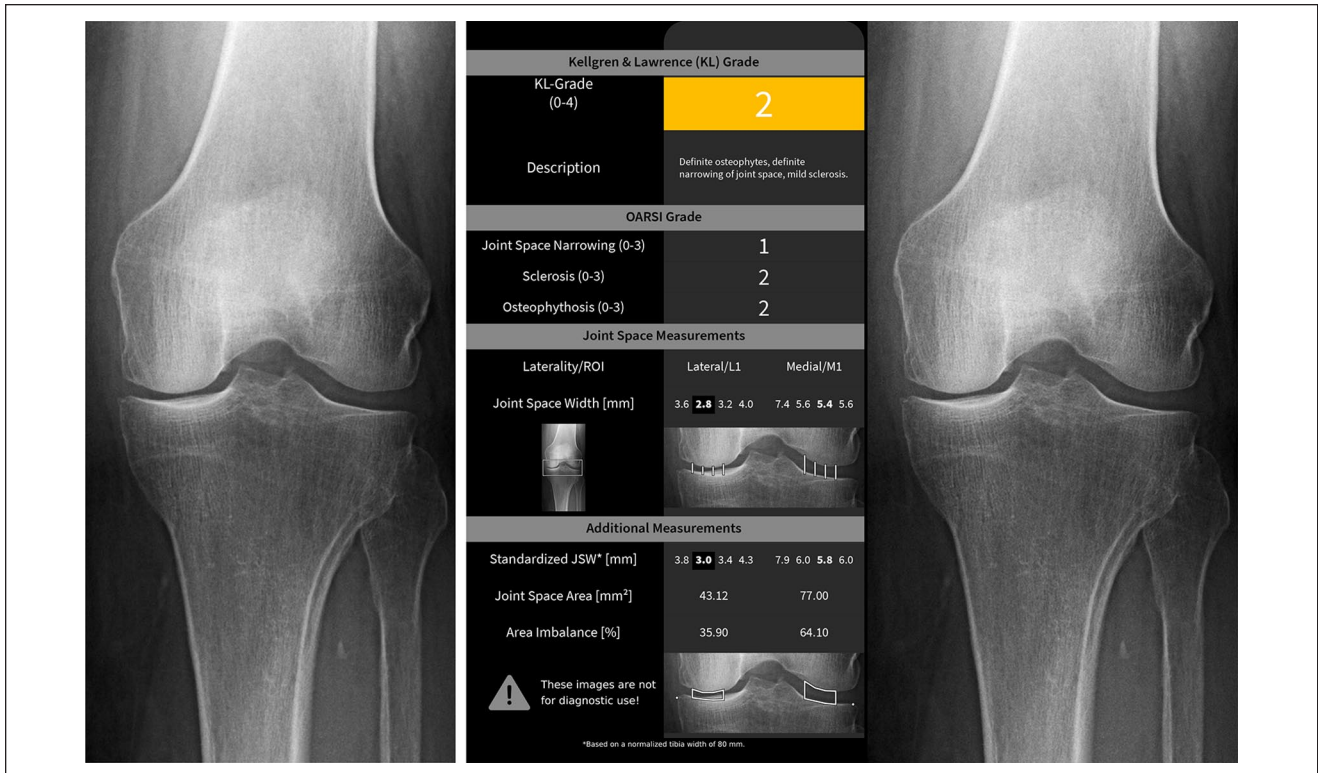
Figure 1. Example of the radiographs presented to the readers. The same knee as presented in the unaided (left) modality and the aided (right) modality.

From the full set of OAI radiographs that had readings available, 124 individual knee radiographs were randomly selected with probability proportional to the frequency of its KL grade. This procedure ensures that the distribution of KL grades in the sampled set is roughly uniform. The demographic description of the population, corresponding to 120 individuals, is depicted in **Table 1**.

The distribution of KL and OARSI grades, as reported by the consensus readings provided by the OAI study, is presented in **Table 2**.

Computer-Assisted Detection System

The Knee Osteoarthritis Labelling Assistant (IB Lab KOALA, http://www.imagebiopsy.com) is an automated software system that analysis anterior-posterior (AP) knee radiographs for the detection and classification of features relevant for the diagnosis of osteoarthritis. KOALA deploys a series of convolutional neural networks that provide all the readings and measurements that are presented to the user. These deep learning algorithms were trained on data coming from a large longitudinal study that provided radiographs annotated with KL and OARSI grades through a multireader consensus procedure. OARSI grades are obtained solely from the imaging data, without taking into account any other clinical data. KL grade is computed by a network that takes as inputs the several OARSI grades.

Given an AP knee radiograph (either unilateral or bilateral), KOALA produces a standardized report (see **Fig. 1**) in which readings for JSN, sclerosis, and osteophyte OARSI grades are provided. Based on these readings, KOALA also proposes a KL grade for each of the knees in the radiograph.

Table 3. Agreement Rates between Physicians for the Unaided (left) and Aided (right) Modalities.[a]

Score	ICC Unaided (95% CI)	ICC Aided (95% CI)
KL	0.67 (0.59, 0.74)	0.81 (0.76, 0.86)
JSN	0.71 (0.62, 0.78)	0.76 (0.66, 0.83)
Sclerosis	0.41 (0.30, 0.52)	0.60 (0.51, 0.69)
Osteophytes	0.55 (0.24, 0.73)	0.73 (0.61, 0.82)
OA	0.43 (0.33, 0.54)	0.60 (0.51, 0.69)

ICC = intraclass correlation; KL = Kellgren-Lawrence; JSN = joint space narrowing; OA = osteoarthritis.
[a]Confidence intervals calculated according to Shrout and Fleiss (1979).

In addition, KOALA also reports joint space width measurements along the tibiofemoral joint, although these outputs were not used in the present study.

Methods

The readers (3, all with more than 4 years of experience in radiological imaging assessment) underwent a training session, where the structure of the KOALA report was explained, and 3 images were used to exemplify the process. The trainer was familiar with the graphical outputs of KOALA and explained only where to find the relevant information in the graphical outputs of KOALA. He did not interpret any images since the purpose is for readers to make use of their medical expertise.

In the first session, the readers were instructed to rate the set of knees with regard to KL grade (0-4), and JSN, sclerosis, and osteophyte OARSI grades (all 0-3) based solely on their visual inspection of the knee radiograph (the unaided modality). In order to avoid reader fatigues, the readers were allowed to use unlimited time to perform all readings and allowed to make the readings at the most convenient times for them. Readings were performed on normal digital screens.

After a washout period of at least 4 weeks, starting from the time the first sessions was completed, a second session was held where the readers re-scored the same images (presented in a different, random order) presented together with the KOALA report—the aided modality (**Fig. 1**).

Statistical Analysis

Agreement Rates. Agreement rates for the different readings (KL, JSN, sclerosis, and osteophytes) were assessed by intraclass correlation (ICC),[15] assuming random effects for the readers (ICC(2, 1)). Ninety-five percent confidence intervals were calculated according to the original derivations by Shrout and Fleiss.[15] Standard errors of the mean for ICCs were estimated by resampling the observations with replacement (bootstrap) 1000 times. Statistical significance of the difference between aided and unaided modalities was assessed by a *z*-score method.

Accuracy Measures. Performance was quantified by several measures, including accuracy, sensitivity, and specificity for several clinically relevant criteria. For each of the criteria, true positives, true negatives, false positives and false negatives of the readers were calculated against the ground truth (the readings from the OAI study). Specifically, we analyzed the ability to detect

- any abnormality (KL grade > 0)
- osteoarthritis (KL grade > 1)
- any narrowing (JSN > 0)
- any sclerosis (SC > 0)
- severe sclerosis (SC > 1)
- presence of osteophytes (OS > 0)

Standard errors and confidence intervals for sensitivity, specificity, and accuracy were calculated using a normal approximation to the binomial proportional interval.

Receiver Operating Characteristic Curve. In addition to grade recommendations, KOALA also produces a confidence score on the recommendation of the grade. Using these confidence scores, a receiver operating characteristic (ROC) curve can be plotted. The ROC curve quantifies the tradeoffs between true and false positive rates (TPR and FPR, respectively) that are possible. This curve was used to visualize the effect of the use of KOALA on the readers' performance, in terms of changes to their TPR and FPR.

Results

Agreement between Readers in the 2 Modalities

Agreement rates between the readers were calculated separately for the 2 modalities (aided and unaided) and for the several scores (KL, JSN, sclerosis, and osteophyte). In general, agreement rates between physicians increased for all scores (**Table 3**, **Fig. 1**), except for JSN.

Agreement rates increased 21% for KL grade, 47% for sclerosis OARSI grade, 33% for osteophyte OARSI grade, and 39% for OA diagnosis (KL grade >1) by the use of the computerized detection device. According to proposed guidelines for the interpretation of ICC values,[16] the agreement rate went from "good" to "excellent" for KL grade, and from "fair" to "good" for sclerosis and osteophyte OARSI grades, as well as for the diagnosis of OA (**Fig. 2**).

Readers' Performance in the 2 Modalities

In addition to the readers' agreement rate, we also compared the readers' performance relative to the ground truth (OAI reference standard) by calculating their sensitivity and specificity for the detection of clinically relevant features. In particular we calculated the impact of being presented the KOALA report on
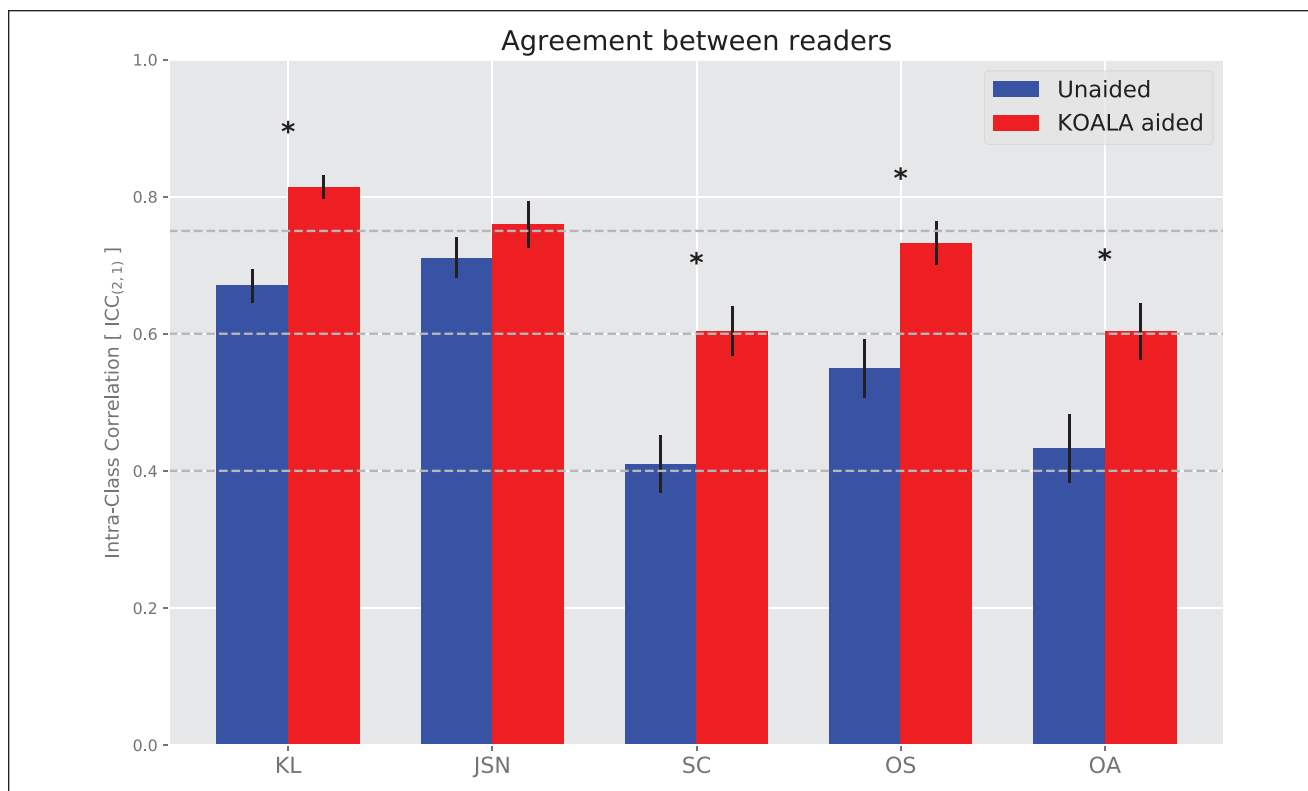
Figure 2. Agreement rates between physicians for the unaided (blue) and aided (red) modalities. Error bars denote the standard error of the intraclass correlation (ICC). Stars indicate statistically significant difference between unaided and aided modalities. KL, Kellgren-Lawrence; JSN, joint space narrowing; SC, sclerosis; OS, osteophyte; and OA, osteoarthritis (KL > 1). Horizontal lines denote the thresholds separating poor, fair, good and excellent agreement, according to Cicchetti.[16]

sensitivity and specificity for any abnormality (KL grade > 0), OA (KL grade >1), any narrowing (JSN > 0), any sclerosis (SC > 0), severe sclerosis (SC > 1), and presence of osteophytes (OS > 0).

We found that on average readers' accuracy improved significantly for JSN > 0 (0.11), Sclerosis OARSI grade >1 (0.16) and Osteophyte OARSI grade >0 (0.08). These increases were mostly due to an increase in specificity which increased significantly for all criteria, with no significant change in average sensitivity (**Fig. 3**, **Table 4**) across all clinically relevant criteria.

Accuracy Performance by Reader

In order to visualize the effect of the aided modality on individual readers we calculated their true positive rate (TPR = sensitivity) and false positive rate (FPR = 1 − specificity) under the 2 modalities (**Fig. 4**). We find that all readers are affected in qualitatively the same way by KOALA: a reduction in FPR and no or little cost of TPR. Furthermore, we find that for most criteria the readers become more similar, consistent with the observed increase in agreement rate.

Discussion

One of the main findings of this study is that agreement rates between physicians increase when using a computer assisted detection system. In our study, the computer system simply produced a report with proposals for the several grades under study (KL, JSN, sclerosis, and osteophyte OARSI grades) and the physicians still had full access to the radiograph, enabling them to confirm any assessments made by the software. Nevertheless, the agreement rate between physicians increased in the aided modality, showing that the report enables a standardization and homogenization of assessments. In fact, agreement rate improved from "good" to "excellent" for KL grade and from "fair" to "good" for sclerosis and osteophyte OARSI grades, and for diagnosis of OA. It could be argued that this increase in agreement rate follows from a sort of psychological "anchoring" effect,[17] where the suggestion of a number by some external entity would predispose the physicians to make similar assessments. Two facts argue against this. First, these are practicing physicians whose training should enable them to make objective assessments, immune to these psychological effects. Second, and most important, their accuracy, as compared to the consensus readings of the
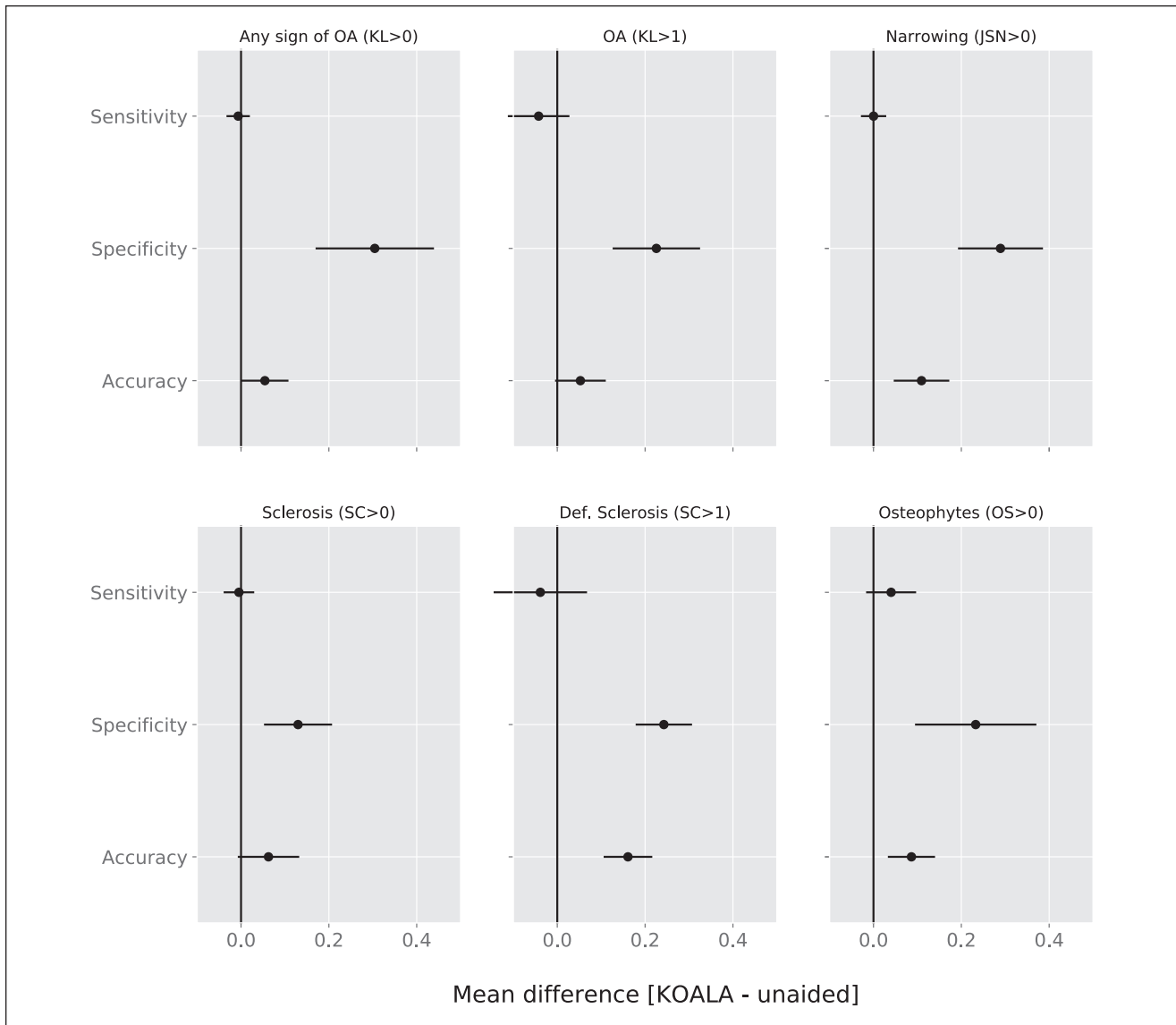
Figure 3. Mean difference in sensitivity, specificity, and accuracy for KL > 0, KL > 1, JSN > 0, sclerosis OARSI grade >0, sclerosis OARSI grade >1, and osteophyte OARSI grade >0. Values to the right of the vertical line at 0 are improvements by the use of KOALA. Error bars denote 95% confidence intervals. KL, Kellgren-Lawrence; JSN, joint space narrowing; OARSI, Osteoarthritis Research Society International.

OAI study, increases, indicating that this increase in agreement rate is driven by more accurate assessments and not some form of the "anchoring" effect.

Our results show that the increase in accuracy of physicians, when aided by the computer assisted detection system KOALA, is mostly driven by an increase in specificity. This reveals that physicians, when unaided by KOALA, tend to err on the side of false positives. A bias toward false positives can lead to unnecessary interventions or examinations costing time, money, and causing discomfort and anxiety on the patient. In particular, the improvements in specificity reported here allow physicians to better recognize the early stages of

OA. Our results suggest that a computer assisted detection system, such as KOALA, can improve the standard of care by decreasing the rate of false positives. For the detection of OA, as a particular example, the improvement in specificity reported here (0.65 for the unaided modality vs. 0.88 for the aided modality) means that only 12% of patients would be falsely diagnosed when using KOALA versus 35% when using only plain radiographs to perform the diagnosis. This represents 20% less patients that are subjected to further, potentially expensive or invasive, examinations or that are being unnecessarily prescribed drugs. Moreover, this decrease in false positives is certainly important in the

Table 4. Average Accuracy, Sensitivity, and Specificity, Including 95% Confidence Intervals, of the Readers in Both Modalities for a Number of Clinically Relevant Criteria, Using as Ground Truth the Readings Provided by the Osteoarthritis Initiative Study.

	Unaided			Aided		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
JSN > 0	0.65 (0.60, 0.70)	0.97 (0.95, 0.99)	0.11 (0.06, 0.17)	0.76 (0.72, 0.80)	0.97 (0.95, 0.99)	0.41 (0.32, 0.49)
KL > 0	0.80 (0.76, 0.84)	0.97 (0.95, 0.99)	0.10 (0.03, 0.18)	0.86 (0.82, 0.89)	0.97 (0.94, 0.99)	0.40 (0.29, 0.52)
KL > 1	0.76 (0.72, 0.81)	0.83 (0.77, 0.87)	0.65 (0.58, 0.73)	0.82 (0.78, 0.85)	0.78 (0.73, 0.83)	0.88 (0.82, 0.93)
KL > 2	0.83 (0.79, 0.87)	0.68 (0.60, 0.76)	0.92 (0.88, 0.95)	0.91 (0.88, 0.94)	0.81 (0.74, 0.87)	0.96 (0.94, 0.98)
OS > 0	0.76 (0.71, 0.80)	0.84 (0.79, 0.88)	0.50 (0.39, 0.61)	0.84 (0.80, 0.88)	0.88 (0.84, 0.91)	0.73 (0.63, 0.82)
OS > 1	0.79 (0.75, 0.83)	0.65 (0.58, 0.73)	0.88 (0.84, 0.92)	0.84 (0.81, 0.88)	0.71 (0.64, 0.77)	0.94 (0.90, 0.97)
SC > 0	0.54 (0.49, 0.59)	0.97 (0.95, 0.99)	0.11 (0.07, 0.15)	0.60 (0.55, 0.65)	0.97 (0.94, 0.99)	0.24 (0.18, 0.30)
SC > 1	0.72 (0.67, 0.76)	0.80 (0.73, 0.88)	0.68 (0.62, 0.74)	0.88 (0.84, 0.91)	0.77 (0.68, 0.84)	0.92 (0.89, 0.96)

JSN = joint space narrowing; KL = Kellgren-Lawrence; OS = osteophytes; SC = sclerosis.

context of drug clinical trials: wrongly diagnosing patients at the baseline of the study will certainly decrease the observed effect of the drug, since a high fraction of individuals which are in fact healthy would be accounted as disease individuals for whom the drug had no effect. Importantly, this decrease in the false positive rate does not come at a cost in sensitivity, since on average sensitivity is not affected for any of the clinical criteria studied here.

Our study included only 3 physicians, which could hinder its generalizability. Because of this, the type of ICC we used to quantify agreement rates considers the reader as a random effect. This corresponds to interpreting the pool of readers as a sample of a larger population, allowing us to generalize the results to a broader population. Nevertheless, it is unlikely that the number of readers in practical applications, such as longitudinal studies or clinical trials, is much larger than this. As an example, in the OAI study, the largest longitudinal study for knee OA, radiographs were read by a minimum of two and a maximum of 3 readers, depending on discrepancies. Furthermore, the effects of KOALA on specificity are large and consistent enough between physicians, suggesting that the effect is not an artifact of the sample of physicians studied here. It should also be noted that KOALA did not have an effect on sensitivities, mostly because sensitivities were already extremely high for this pool of physicians.

Previously, other automated systems were introduced for the grading of knee OA.[18,19] Unlike IB Lab's KOALA, these systems provide only a black-box prediction of the KL grade, without any of the OARSI scores that help justify it. Since no reader study was conducted with these other solutions for automated KL grading it is impossible to know how they affect reader performance. However, it is conceivable that the extra transparency that these extra scores provide helps the reader (1) understand the KL grading proposal by the software and (2) judge reliability by judging its consistency with the other assessments. Furthermore, we have shown that the physicians have a propensity for false

positives. These extra scores likely play a role in the increase in specificity we observed, since they explicitly identify the subfeatures responsible for the more subtle features of OA, which are often a source of interobserver variability.[9,10]

One interesting finding of the present study is that the performance of physicians was consistently superior to KOALA's performance in the aided modality, even though it is worse than KOALA's in the unaided modality (**Fig. 4**). This suggests that physicians do not simply accept KOALA's recommendations when grading, as in this case their performance would be the same as KOALA. Instead, it suggests that the physicians learn canonical examples of specific grades from KOALA, improving their performance even beyond KOALA's performance. Informal conversations with some of the readers indicated that this is true, especially for the scores of the IRFs sclerosis and osteophytes. One interesting possibility then is that this type of software can be used as a training tool for junior physicians. Similar approaches[20] have been reported to have a positive effect on reliability.

Artificial intelligence promises to revolutionize radiology. Our study highlights that these software systems are not meant to replace radiologists but instead to support and enhance radiologists' performance in the clinical practice. That said, automated assessment systems such as KOALA could be used to quickly assess and grade large numbers of radiographs, especially in the context of clinical studies that often require detailed assessments, for example, to determine radiographic inclusion/exclusion criteria. Furthermore, the increased consistency between readers obtained when using KOALA will certainly improve reliability of measurements, by decreasing the effect of interobserver variability.

Conclusion

In conclusion, our study suggests that the use of a computer-assisted detection system, such as KOALA, improves both agreement rate and accuracy when assessing radiographic
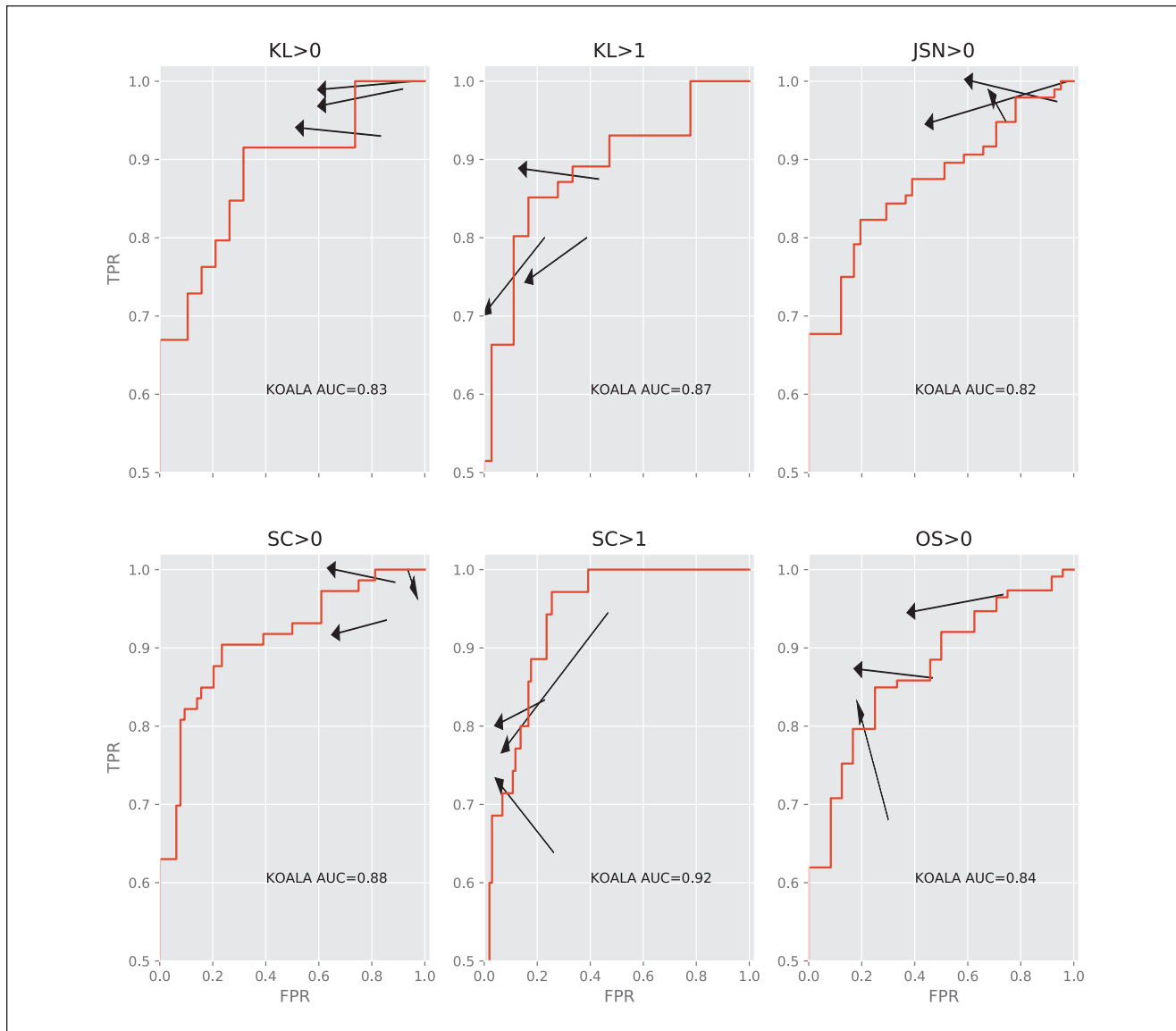
Figure 4. Changes to the true positive rate (*y*-axis, TPR) and false positive rate (*x*-axis, FPR) for each individual reader for KL > 0, KL > 1, JSN > 0, sclerosis OARSI grade >0, sclerosis OARSI grade >1, and osteophyte OARSI grade >0. Red line denotes the ROC curve of KOALA of the dataset. Arrows point from the unaided to the aided modality. Arrows pointing upward and left are absolute improvements in detection ability. Note that even though some arrows point downward and left, the improvement in FPR is greater than the loss in TPR, representing a net increase in accuracy. KL, Kellgren-Lawrence; JSN, joint space narrowing; OARSI, Osteoarthritis Research Society International; ROC, receiver operating characteristic; AUC, area under the ROC curve.

features relevant for the diagnosis of knee osteoarthritis. These improvements in physician performance and reliability come without trade-offs in terms of accuracy. These results argue for the use of this type of software as a way to improve the standard of care when diagnosing knee osteoarthritis.

Author Contributions

All authors contributed to the conception and design of the study and gave final approval of the version to be submitted. TP and SN drafted the article and all other authors revised it critically for important intellectual content. CG acquired the data and TP performed data analysis.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this

ORCID iD

Stefan Nehrer https://orcid.org/0000-0001-8008-2226

References

1. Braun HJ, Gold GE. Diagnosis of osteoarthritis: imaging. Bone. 2012;51(2):278-88. doi:10.1016/j.bone.2011.11.019

2. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 1957;16(4):494-502.

3. Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin Orthop Relat Res. 2016;474(8):1886-93. doi:10.1007/s11999-016-4732-4

4. Culvenor AG, Engen CN, Øiestad BE, Engebretsen L, Risberg MA. Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. Knee Surg Sports Traumatol Arthrosc. 2015;23(12):3532-9. doi:10.1007/s00167-014-3205-0

5. Wright RW; MARS Group. Osteoarthritis Classification Scales: interobserver reliability and arthroscopic correlation. J Bone Joint Surg Am. 2014;96(14):1145-51. doi:10.2106/JBJS.M.00929

6. Schiphof D, Boers M, Bierma-Zeinstra SMA. Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis. Ann Rheum Dis. 2008;67(7):1034-6. doi:10.1136/ard.2007.079020

7. Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis Cartilage. 2007;15(Suppl A):A1-A56. doi:10.1016/j.joca.2006.11.009

8. Günther KP, Sun Y. Reliability of radiographic assessment in hip and knee osteoarthritis. Osteoarthritis Cartilage. 1999;7(2):239-46. doi:10.1053/joca.1998.0152

9. Damen J, Schiphof D, Wolde ST, Cats HA, Bierma-Zeinstra SMA, Oei EHG. Inter-observer reliability for radiographic assessment of early osteoarthritis features: the CHECK (cohort hip and cohort knee) study. Osteoarthritis Cartilage. 2014;22(7):969-74. doi:10.1016/j.joca.2014.05.007

10. Hart DJ, Spector TD. Kellgren & Lawrence grade 1 osteo-phytes in the knee—doubtful or definite? Osteoarthritis Cartilage. 2003;11(2):149-50. doi:10.1053/JOCA.2002.0853

11. Gossec L, Jordan JM, Mazzuca SA, Lam MA, Suarez-Almazor ME, Renner JB, et al. Comparative evaluation of three semi-quantitative radiographic grading techniques for knee osteoarthritis in terms of validity and reproducibility in 1759 X-rays: report of the OARSI-OMERACT task force. Osteoarthritis Cartilage. 2008;16(7):742-8. doi:10.1016/j.joca.2008.02.021

12. Bálint G, Szebenyi B. Diagnosis of osteoarthritis. Guidelines and current pitfalls. Drugs. 1996;52(Suppl 3):1-13.

13. Sadler ME, Yamamoto RT, Khurana L, Dallabrida SM. The impact of rater training on clinical outcomes assessment data: a literature review. Int J Clin Trials. 2017;4(3):101. doi:10.18203/2349-3259.ijct20173133

14. Marshall DA, Vanderby S, Barnabe C, MacDonald KV, Maxwell C, Mosher D, et al. Estimating the burden of osteoarthritis to plan for the future. Arthritis Care Res (Hoboken). 2015;67(10):1379-86. doi:10.1002/acr.22612

15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8. doi:10.1037/0033-2909.86.2.420

16. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 1994;6(4):284-90. doi:10.1037/1040-3590.6.4.284

17. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science. 1974;185(4157):1124-31. doi:10.1126/science.185.4157.1124

18. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep. 2018;8(1):1727. doi:10.1038/s41598-018-20132-7

19. Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. J Digit Imaging. 2019;32(3):471-7. doi:10.1007/s10278-018-0098-3

20. Hayes B, Kittelson A, Loyd B, Wellsandt E, Flug J, Stevens-Lapsley J. Assessing radiographic knee osteoarthritis: an online training tutorial for the Kellgren-Lawrence Grading Scale. MedEdPORTAL. 2016;12:10503. doi:10.15766/mep_2374-8265.10503