Content-based image retrieval assists radiologists in diagnosing eye and orbital mass lesions in MRI

J. Lorenz Rumberger*, MSc • Winna Lim, MD* • Benjamin Wildfeuer • Elisa B. Sodemann, MD • Augustin Lecler, MD • Simon Stemplinger, Dipl. Inf. • Ahi Sema Issever, MD, PD • Ali R. Sepahdari, MD • Sönke Langner, MD • Dagmar Kainmueller, PhD • Bernd Hamm, MD • Katharina Erb-Eigner#, MD

From the Max-Delbrück Center for Molecular Medicine in the Helmholtz Association, Robert-Rössle-Str. 10, 13125-Berlin, Germany (J.L.R., D.K.), Humboldt University Berlin, Faculty of Mathematics and Natural Sciences, Rudower Chaussee 25, 12489-Berlin, Germany (J.L.R.), Helmholtz Imaging (J.L.R., D.K.), Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Hindenburgdamm 30, 12203 Berlin, Germany (W.L., B.W., E.B.S., S.S., A.S.I., B.H., K.E.E.), Hôpital Fondation Adolphe de Rothschild, 29 Rue Manin, 75019-Paris, France (A.L.), Paris Cité University, 5 Rue Thomas Mann, 75013-Paris, France (A.L.), Diagnostic Neuroradiology, Department of Radiology, Scripps Clinic Medical Group, 10666 North Torrey Pines Rd., La Jolla, CA 92037, USA (A.R.S.), David Geffen School of Medicine, University of California Los Angeles, 10833 Le Conte Ave, Los Angeles, CA 90095, USA (A.R.S.), Greifswald University Medicine, Fleischmannstraße 8, 17475-Greifswald, Germany (S.L.), Potsdam University, Digital Engineering Faculty, Prof.-Dr.-Helmert-Str. 2-3, 14482-Potsdam, Germany (D.K.)

* equal contributions, # corresponding author

Background:

Diagnoses of eye and orbit pathologies by radiological imaging is challenging due to their low prevalence and the relative high number of possible pathologies and variability in presentation, thus requiring substantial domain-specific experience.

Purpose:

This study investigates whether a content-based image retrieval (CBIR) tool paired with a curated database of orbital MRI cases with verified diagnoses can enhance diagnostic accuracy and reduce reading time for radiologists across different experience levels.

Material and Methods:

We tested these two hypotheses in a multi-reader, multi-case study, with 36 readers and 48 retrospective eye and orbit MRI cases. We asked each reader to diagnose eight orbital MRI cases, four while having only status quo reference tools available (e.g. Radiopaedia.org, StatDx, etc.), and four while having a CBIR reference tool additionally available. Then, we analyzed and compared the results with linear mixed effects models, controlling for the cases and participants.

Results:

Overall, we found a strong positive effect on diagnostic accuracy when using the CBIR tool only as compared to using status quo tools only (status quo only 55.88%, CBIR only 70.59%, 26.32% relative improvement, p=.03, odds ratio=2.07), and an even stronger effect when using the CBIR tool in conjunction with status quo tools (status quo only 55.88%, CBIR + status quo 83.33%, 49% relative improvement, p=.02, odds ratio=3.65). Reading time in seconds (s) decreased when using only the CBIR tool (status quo only 334s, CBIR only 236s, 29% decrease, p<.001), but increased when used in conjunction with status quo tools (status quo only 334s, CBIR + status quo 396s, 19% increase, p<.001).

Conclusion:

We found significant positive effects on diagnostic accuracy and mixed effects on reading times when using the CBIR reference tool, indicating the potential benefits when using CBIR reference tools in diagnosing eye and orbit mass lesions by radiological imaging.

Main

Introduction

Abbreviations

CBIR = Content-based image retrieval, SQ = status quo reference tools (Radiopaedia.org, StatDx, etc.), ML = machine learning, ROI = Region of Interest, Infl. & Infect. = Inflammatory and infectious diseases.

Summary

Using a content-based image retrieval tool significantly improved diagnostic accuracy and had mixed effects on reading time for diagnosing MRI exams of patients with eye and orbit pathologies.

Key Results

- Using the CBIR tool alone improved diagnostic accuracy from 55.88% to 70.59% (odds ratio=2.07, p=.03) and decreased reading time from 334s to 236s (p<.001) compared to SQ alone.
- Using CBIR together with SQ tools further increased accuracy to 83.33% (odds ratio=3.65, p=.02) but increased reading time to 396s (p<.001) compared to SQ only.

Inaccurate diagnoses in medical imaging reports are a burden to the patient and the healthcare system (1). Reading MRI scans of patients with eye and orbit diseases poses a particular diagnostic challenge due to the rarity of these lesions. Most radiologists lack profound experience reading these cases or they may find it difficult to recall imaging features from past cases. Radiologists specialized in the eye and orbit area are also rare, thus these cases are often read by general radiologists or neuroradiologists, increasing the probability for diagnostic inaccuracies. Additionally, the high number of distinctive tissue types in the orbit enables a variety of orbital pathologies, increasing the number of possible differential diagnoses to consider.

Although large, multi-center studies describing the diagnostic accuracy of eye and orbital lesions are lacking, it has been reported for lacrimal gland lesions that the degree of correspondence between image-based diagnosis and histopathologic diagnosis is only moderate (Cohen's kappa=0.451, p <.001) (2). Other studies found that diagnostic errors occur at an average rate of 3%-4%, with a 32% retrospective error rate for interpretation of abnormal studies (3). These challenges may delay diagnosis and treatment or expose patients to potentially unnecessary biopsies and treatments, which can cause harm and be costly (1).

Content-based image retrieval (CBIR) allows radiologists to retrieve relevant cases from a curated database with clinical or histopathological validation, based on visual similarity with supplied patient query images. Given the cases and their associated diagnoses retrieved by the CBIR system, radiologists may be able to give better informed and more accurate diagnoses. Previous studies on CBIR showed increases in diagnostic accuracy, particularly for diagnosing interstitial lung diseases on CT scans (9–12). However, these studies often did not compare CBIR with status quo reference tools (e.g. StatDx, radiopaedia.org, etc.) (11,12), and involved a small number of participants, albeit many cases per participants. Notably absent is research on CBIR's effectiveness in challenging MRI diagnoses and other organ systems where retrieval of reference cases can be crucial and time consuming.

Thus, our study seeks to close this gap by evaluating whether a CBIR system can improve diagnostic accuracy and reading time for diagnosing challenging eye and orbital pathologies. We developed a CBIR tool and conducted a retrospective study involving 36 radiologists and 48 orbital MRI cases to assess its effectiveness across a wide range of experience levels and orbital pathologies.

Materials and Methods

Orbital pathologies datasets

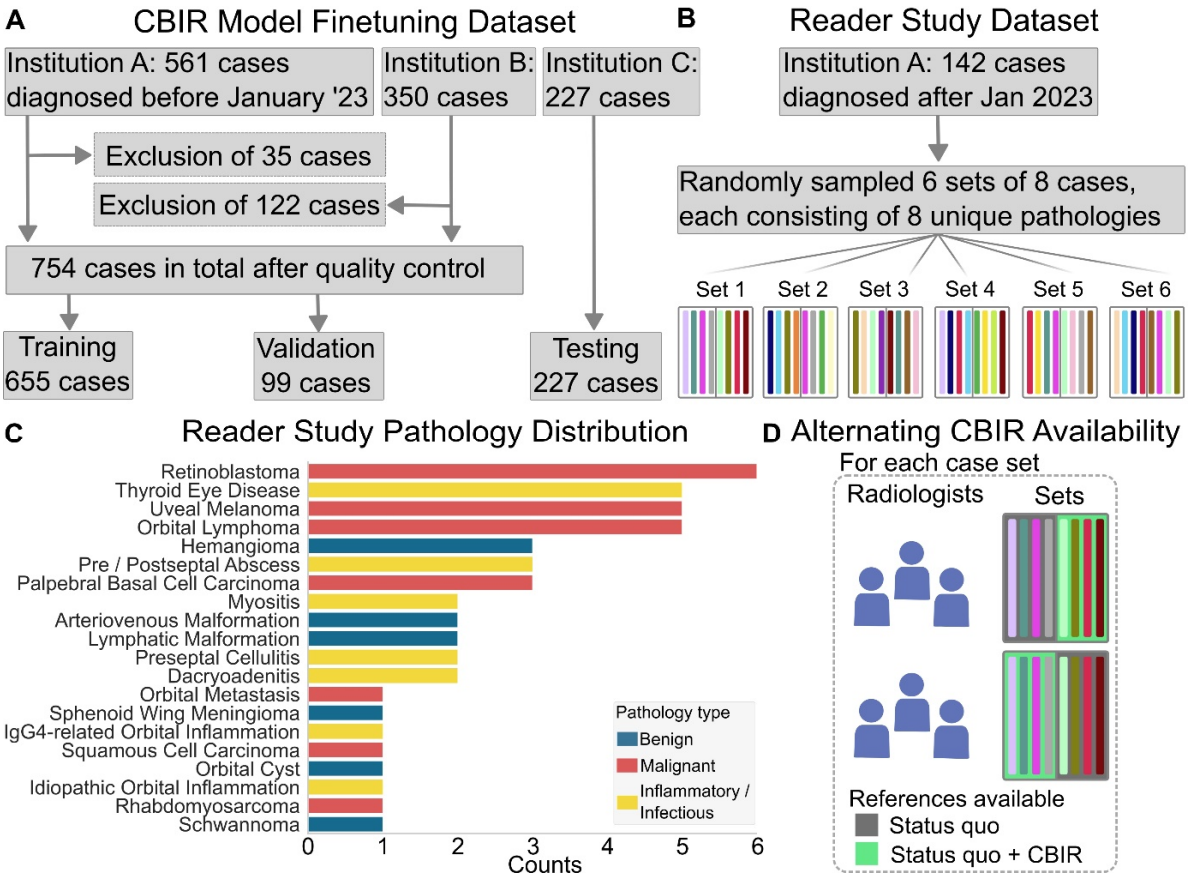

Figure 1: A, for the CBIR model we gathered data from 3 sources and excluded cases based on quality control measures. **B-C**, for the reader study we only used cases from Institution A, diagnosed after the cases in the finetuning dataset with 20 distinct diagnoses of different types. **D**, each set of 8 cases was read by 6 radiologists with alternating availability of the CBIR reference tool for the first or last four cases.

This retrospective study was approved by the institutional review board under ethics application code EA121422 and informed consent was waived. For developing the CBIR machine learning (ML) model and the database, we collected anonymized data from patients with eye and orbit pathologies who were diagnosed between 2012 and 2022 at Institution A, Institution B, and Institution C (**Fig. 1 A**). The inclusion criteria required a clinical or histopathological confirmation of the diagnosis verified through multidisciplinary clinical assessments, visible lesions on the respective MRI scans, scans performed prior to any therapeutic treatment, and sufficient image quality. For the ML model development, 3D regions of interest (ROIs) were annotated around each lesion by three expert radiologists. The following routinely acquired MRI sequences were annotated: T1-weighted spin echo sequences before and after intravenous contrast agent administration, T2-weighted sequences with and without fat suppression, and Fluid-Attenuated Inversion Recovery sequences. Sequences were acquired with a range of different scanners: Siemens (Skyra, Aera, Avanto, Magnetom Amira, Vida), Philips (Ingenia, Intera, Symphony), Toshiba (Titan) and GE (Optima, Signa). Field strength varied between 1.5T and 3T depending on the scanner. Data from Institution A and B was split into training and validation cases, with the validation dataset being constructed by taking 10% of cases of each pathology, to ensure a representative sample. The Institution C dataset was used as an external test dataset.

For the reader study, data with similar characteristics, but diagnosed after January 2023 were collected at Institution A. The dataset included 28 pathologies. Six sets of eight cases were randomly sampled for the reader study, such that each set consisted of cases with eight distinct pathologies without repetition (**Fig 1. B**), resulting in the pathology distribution shown in **Fig 1. C**. The 48 sampled patients had an average age of 43±24 years and 48% were female.

Content-Based Image Retrieval Tool

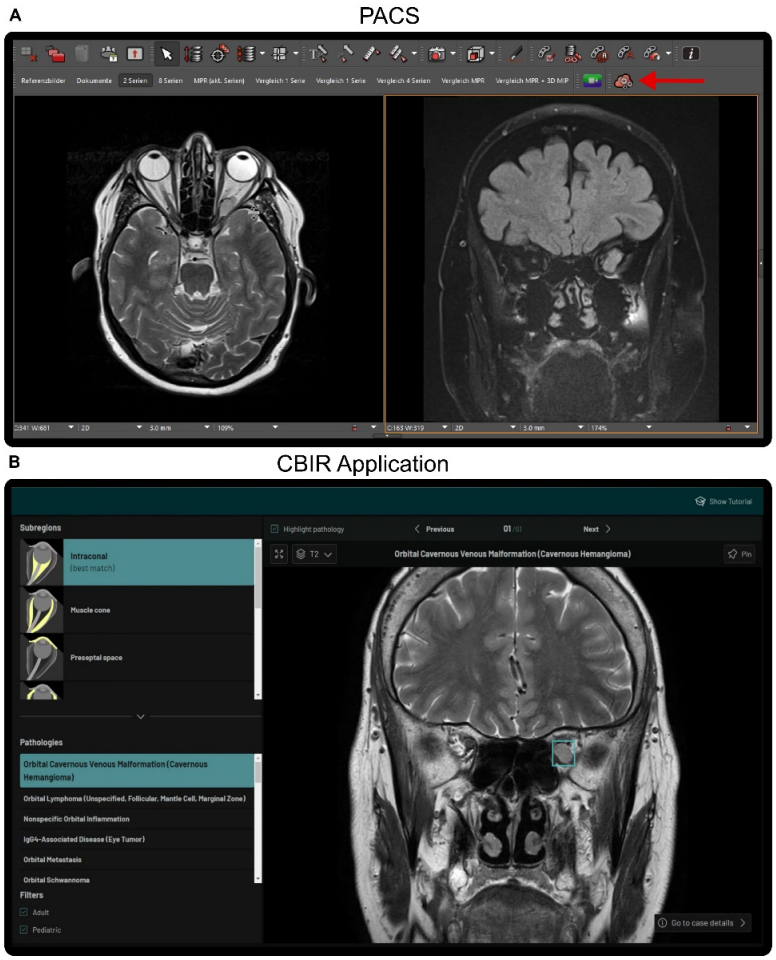

Figure 2: A, the PACS viewer environment, with the button starting the CBIR tool highlighted with a red arrow. **B**, the CBIR web application with the search results for the slice shown on the right in **A**.

The CBIR tool is seamlessly integrated into the PACS viewer and accessible to eligible radiologists with one click on a dedicated button in the PACS. To use the CBIR tool, users navigate to a sequence slice where the pathology is clearly visible, then click on the button which opens the web application that shows a range of pathologies, sorted by image similarity (**Fig. 2. B**). The user interface enables exploration of several cases across 77 verified eye and orbit pathologies in seven anatomical subregions (preseptal space, globe, optic nerve, intraconal, extra ocular muscles, extraconal, lacrimal gland, subperiosteal space and bony orbit). The CBIR algorithm employs an ML model that compares the uploaded radiology sequence slice with those in the database, ranking them by similarity. The algorithm is based on the DinoV2 framework (13,14), whose pre-trained checkpoint was further trained on publicly available radiology datasets (15), then finetuned on an image-retrieval objective (16) on the CBIR fine-tuning dataset (**Fig. 1 A**). The ML model was developed using PyTorch (version 2.3.0) and Python (version 3.10).

Table 1: Study Participant demographics

A) Demographic	No. of participants
Female	15 (41.67)
B) Medical Role	
Resident	16 (44.44)
Board-certified	10 (27.78)
Senior	10 (27.78)
C) Tenure	
0 - 5 years	16 (44.44)
6 - 10 years	10 (27.78)
11 - 15 years	4 (11.11)
>15 years	6 (16.67)
D) Prior exp. in orbital MRI	
No exp.	7 (19.44)
Little exp.	21 (58.33)
Sufficient exp.	8 (22.22)

Note. — Data is presented as number of participants with percentages in parentheses.

Study population

The study was conducted in March and April 2024 at Institution A. Eligible for the study were radiologists with experience in reading MRI exams. 36 radiologists were randomly recruited for the study, who covered a representative cross section of the department (**Tab. 1 A**), working in a range of medical roles (**Tab. 1 B**) and having varying job tenure (**Tab.1 C**). Prior experience in reading orbital MRI cases was low (**Tab. 1 D**), with 28 of 36 participants having either no or little prior experience.

Reader evaluation

In total 36 participants each diagnosed a set of eight cases only based on the MRI scans (**Fig. 1 B**), four with and four without the CBIR tool available. Other status quo reference tools like radiopaedia.org, StatDx or Google were available throughout the study. Half of the participants had the CBIR tool available for the first four cases, whereas the other half for the last four cases. Each individual case was read by six participants with alternating availability of the CBIR tool (**Fig. 1 D**). Before the participants read cases with the CBIR tool, they went through a

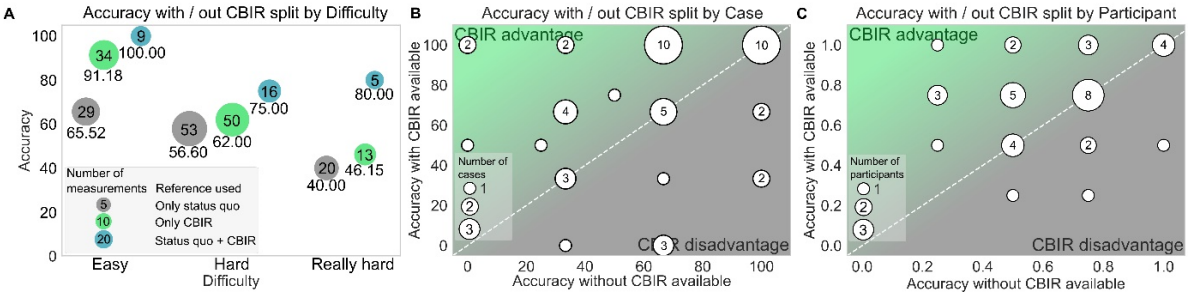

Figure 3: A, diagnostic accuracy averaged over individual cases that readers perceived as easy, hard or really hard. **B-C**, diagnostic accuracy with CBIR available (Y-axis) and without CBIR available (X-axis) averaged over individual cases (**B**) and over individual study participants (**C**). Dots above the white isoline indicate higher accuracy with the CBIR tool than without and vice versa. Dot-size indicates the number of measurements (**A**), of cases (**B**) and participants (**C**).

Table 2: Summary Statistics

	No CBIR	CBIR
A) General		
Reading time [s]	260±228	257±193
Reading time with reference tool [s]	336±230	272±193
Reference tool use	101 (70.14)	133 (92.36)
Accurate diagnoses	91 (63.19)	106 (73.61)
B) Confidence		
Really low confidence	13 (9.03)	6 (4.17)
Low confidence	25 (17.36)	26 (18.06)
Sufficient confidence	89 (61.81)	91 (63.19)
High confidence	17 (11.81)	21 (14.58)
C) Difficulty		
Really easy	0 (0.00)	0 (0.00)
Easy	54 (37.50)	51 (35.42)
Hard	68 (47.22)	70 (48.61)
Really hard	21 (14.58)	18 (12.50)
Not stated	1 (0.07)	5 (3.47)
D) Reference tools used by participants		
CBIR tool	0 (0.00)	132 (91.67)
Radiopaedia	86 (59.72)	26 (18.06)
Google	55 (38.19)	15 (10.42)
StatDx	13 (9.03)	2 (1.39)
Pubmed	10 (6.94)	0 (0.00)
Others	3 (2.08)	1 (0.69)
E) Reference categories		
No reference used	43 (29.86)	11 (7.64)
Only status quo	101 (70.14)	1 (0.69)
Only CBIR	0 (0.00)	102 (70.83)
CBIR + status quo	0 (0.00)	30 (20.83)

Note. — Unless otherwise stated data is presented as numbers with percentages relative to the total number of measurements per treatment phase in parentheses. Participants were allowed to use multiple reference tools, so the relative numbers in D) add up to more than 100%.

short tutorial and were allowed to test the tool by diagnosing a case with a pathology not present in the reader study dataset. In addition, they were allowed to ask questions of the experimenter regarding the CBIR tool. Cases were read on radiology workstations within a standard PACS environment. After each case, the participants were asked to give their diagnosis in free-text form, rate the perceived difficulty, provide their confidence level in the diagnosis, and the reference tools that they used. A person (either anonymous author B, C or D) instructing the participants and taking time measurements was in the room during the session. After the measurements were completed, an eye and orbit radiology specialist with over 15 years of expertise (anonymous author F) with access to additional clinical information on each case, assessed the diagnoses given by the participants in a fully blinded manner. The evaluation was based on the criterion that the diagnosis was sufficiently correct to ensure the accurate administration of downstream treatment, meaning only clinically significant errors were counted as being incorrect. This assessment considers that the classification of orbital lesions can vary among centers and countries, thus diagnostic accuracy should not be judged merely on technical correctness, but on its clinical impact on patient management and outcomes.

Statistical Analysis

Prior to commencement of the study, a power analysis was conducted to determine the number of participants required to detect significant effects ($p<.05$) for the endpoints. We reviewed effect sizes from comparable studies (9–11,17) and calculated that a sample size of 36 participants and 48 cases, resulting in 288 measurements in total, would allow us to detect effects down to an effect size of Cohen's D 0.6 at 80% statistical power. We split reference usage into four categories: no reference used, only status quo (only SQ) used, only CBIR used, or both status quo and CBIR used (SQ+CBIR). We analyzed the

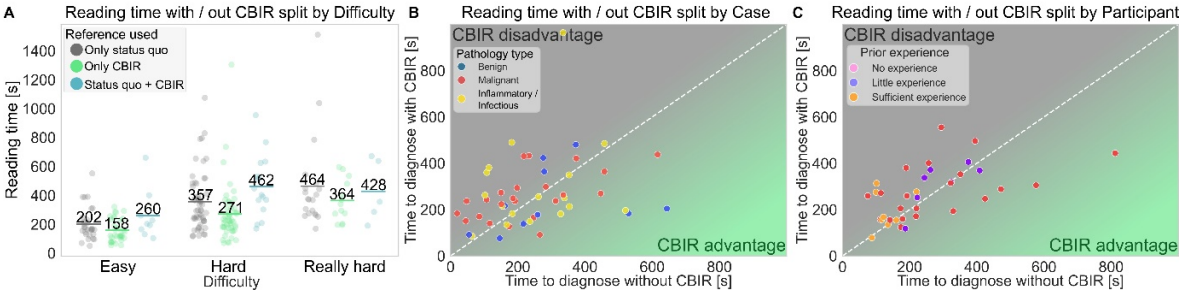

Figure 4: A, reading time split by perceived difficulty and use of the CBIR tool with averages overlayed. **B-C**, reading time with CBIR available (Y-axis) and without CBIR available (X-axis) split by cases (**B**) and study participants (**C**). Dots below the white isoline indicate a lower reading time with the CBIR tool than without and vice versa, dots on the isoline.

effect of the CBIR tool on diagnostic accuracy using a logistic mixed effects model, treating individual participants and cases as random effects, and including reference usage, medical roles, tenure, and interaction terms as fixed effects. For analyzing the effect of the CBIR tool on reading times, we employed a linear mixed effects model with the same random and fixed effects. Reading times were log-transformed, to meet the distributional assumption of the model. We excluded fixed effects via a backwards elimination process based on the Akaike information criterion (18,19). The residuals of the mixed effects models were visually examined to check if all assumptions were met (20). Statistical analysis and data visualization was performed using R (version 4.3.3) and Python (version 3.10) by anonymous author A.

Results

Participants spent on average 01h:03m:57s $\pm$ 35m:31s in total on the tutorial, reading the cases and providing the measurements. When not accounting for the reference tools that the participants actually used but only for the ones that were available in the respective study phase, reading times stayed approximately constant (no CBIR 260s, CBIR 257s p=.09), while during the CBIR phase participants used reference tools more often (no CBIR 70.14%, CBIR 92.36%) and had a higher diagnostic accuracy (no CBIR 63.19%, CBIR 73.61% p=.049) (**Tab. 2 A)**. In addition, having the CBIR tool available slightly increased confidence in the diagnoses (**Tab. 2 B**). No trend is visible on the perceived difficulty of the cases over the study phases (**Tab. 2 C**). Without the CBIR tool available, most participants used radiopaedia.org and Google for finding reference cases, whereas with the CBIR tool available, participants used considerably fewer other reference resources (**Tab. 2 D**). Participants often used only the CBIR tool when it was available and only used additional status quo reference tools in 20.83% of the cases (**Tab. 2 E**). In the following sections, the impact on the diagnostic accuracy and reading times of using only status quo (only SQ) reference tools, only the CBIR reference tool (only CBIR), and using both in conjunction (SQ+CBIR) are analyzed.

Impact of CBIR Usage on Diagnostic Accuracy

Diagnostic accuracy significantly improved overall from 55.88% with status quo reference tools only, to 70.59% when using the CBIR tool only (odds ratio=2.07, p=.03) and to 83.33% when using the CBIR tool in conjunction with status quo tools (odds ratio=3.65, p=.02), which constitutes a 26.32% and a 49.12% relative improvement over the status quo (**Tab. 3 F**).

At the case level, accuracy increased on average with CBIR usage in 21 cases, stayed constant for 18 cases and decreased for 9 cases (**Fig. 3 B** cases above, on and below the isoline). Accuracy declined with increased perceived difficulty independent of reference tool use, but using the CBIR tool retained a higher accuracy across increasing difficulty levels (**Fig. 3 A**, **Tab. 3 A**). We found an increase in diagnostic accuracy from 65.52% with status quo tools only, to 91.18% with the CBIR tool only, a 39% relative increase (p=.02) for ‘easy’ cases. For ‘hard’ and ‘really hard’ cases, we did not find evidence for varying effects (**Tab 3 B**). Stratified by pathology type, the highest increase in accuracy was observed for inflammatory and infectious diseases (only SQ 55.56%, only CBIR 77.78% p=.055, SQ+CBIR 81.82% p=.11), albeit not significant.

Table 3: Diagnostic Accuracy with/out CBIR

Characteristics	Only SQ	Only CBIR	P value	SQ+CBIR	P value
A) Difficulty					
Easy	65.52 (19/29)	91.18 (31/34)	.02*	100.00 (9/9)	.99
Hard	56.60 (30/53)	62.00 (31/50)	.47	75.00 (12/16)	.23
Really hard	40.00 (8/20)	46.15 (6/13)	.86	80.00 (4/5)	.13
B) Pathology Type					
Infl. & Infect.	55.56 (20/36)	77.78 (28/36)	.055	81.82 (9/11)	.11
Benign	43.48 (10/23)	44.44 (8/18)	.93	83.33 (5/6)	.12
Malignant	62.79 (27/43)	75.00 (36/48)	.22	84.62 (11/13)	.23
C) Medical Role					
Resident	62.26 (33/53)	79.49 (31/39)	.09	80.95 (17/21)	.11
Board-certified	59.09 (13/22)	53.13 (17/32)	.62	75.00 (3/4)	.62
Senior	40.74 (11/27)	77.42 (24/31)	.01*	100.00 (5/5)	.99
D) Prior Experience					
No exp.	52.00 (13/25)	77.27 (17/22)	.10	100.00 (6/6)	.11
Little exp.	57.38 (35/61)	69.81 (37/53)	.15	79.17 (19/24)	.058
Sufficient exp.	56.25 (9/16)	66.67 (18/27)	.79	-	
C) Tenure					
0-5 years	62.26 (33/53)	79.49 (31/39)	.10	80.95 (17/21)	.11
6-10 years	41.67 (10/24)	61.76 (21/34)	.15	100.00 (4/4)	.90
11-15 years	55.56 (5/9)	50.00 (6/12)	.64	-	
>15 years	56.25 (9/16)	82.35 (14/17)	.16	80.00 (4/5)	.44
F) Overall					
All	55.88 (57/102)	70.59 (72/102)	.03*	83.33 (25/30)	.02*

Note. — Statistics of diagnostic accuracy in percent are shown for measurements where only status quo reference tools were used (Only SQ), where only the CBIR tool was used (Only CBIR) and where both were used (SQ+CBIR). Total numbers on parentheses. P values indicate significant differences to reference level 'Only status quo' and were calculated using logistic mixed effects models with individual readers and patients as random effects.

At the participant level, diagnostic accuracy increased on average for 15 study participants, stayed constant for 16 and decreased for 5 (cf. **Fig. 3 C**, participants above, on and below the isoline). Accuracy of participating senior radiologists improved with the CBIR tool (only SQ 40.74%, only CBIR 77.42% p=.01), whereas accuracy of resident and board-certified radiologists did not vary significantly (**Tab. 3 C**). Diagnostic accuracy improved the most for participants with no experience (only SQ 52%, only CBIR 77.27% p=.10, SQ+CBIR 100% p=.11) and those with little experience (only SQ 57.38%, only CBIR 69.81% p=.15, SQ+CBIR 79.17% p=.058), albeit not significantly (**Tab. 3 D**). Accuracy showed a positive trend for all tenure levels, except for the 11-15 years tenure level where it showed a decreasing trend (only SQ 55.56%, only CBIR 50% p=.64, **Tab. 3 C**).

Impact of CBIR Usage on Reading Time

Reading time decreased by 29% when using only the CBIR tool compared to only status quo tools (only SQ 334s, only CBIR 236s p<.001). In contrast, reading time increased by 19% when using CBIR in conjunction with status quo tools (only SQ 334s, SQ+CBIR 396s p<.001, **Tab. 4 F**).

At the case level, reading time decreased when using only the CBIR tool and increased when using it together with SQ tools, for hard cases (only SQ 357s, only CBIR 271s p=.002, SQ+CBIR 462s p=.03, **Fig. 4 A, Tab. 4 A**). In addition, we found evidence for a similar effect for malignant lesions (only SQ 314s, only CBIR 207s p<.001, SQ+CBIR 365s p=.045) and a decrease in reading times for inflammatory and infectious lesions when using only the CBIR tool (only SQ 338s, only CBIR 226s p=.005, **Fig 4 B**, **Tab. 4 B**).

At the participant level, resident radiologists benefitted the most from the CBIR tool (only SQ 417s, only CBIR 276s p<.001, **Tab. 4 C**). In addition, the decrease in reading time was the strongest for participants with little experience (only SQ 377s, only CBIR 236s p<.001, **Fig. 4 C**, **Tab. 4 D**). Reading times among participants of different tenure levels decreased the most for the 0-5 years of tenure group, with a relative decrease of 31%

Table 4: Reading time with/out CBIR

Characteristics	Only SQ	Only CBIR	P value	SQ+CBIR	P value
A) Difficulty					
Easy	202 (113)	158 (78)	.14	260 (173)	.01*
Hard	357 (206)	271 (198)	.002*	462 (212)	.03*
Really hard	464 (317)	364 (144)	.41	428 (215)	.94
B) Pathology Type					
Infl. & Infect.	338 (201)	226 (112)	.005*	389 (242)	.10
Benign	363 (240)	335 (304)	.40	476 (278)	.34
Malignant	314 (250)	207 (126)	<.001*	365 (163)	.045*
C) Medical Role					
Resident	417 (278)	276 (232)	<.001*	441 (223)	.046*
Board-certified	208 (96)	228 (136)	.34	205 (64)	.79
Senior	273 (112)	195 (90)	.08	360 (188)	.58
D) Prior Experience					
No exp.	313 (160)	288 (181)	.75	393 (140)	.19
Little exp.	377 (263)	236 (186)	<.001*	397 (232)	.054
Sufficient exp.	204 (113)	194 (122)	.50	-	
E) Tenure					
0-5 years	417 (278)	276 (232)	<.001*	441 (223)	.049*
6-10 years	237 (104)	210 (126)	.58	408 (179)	.29
11-15 years	187 (89)	250 (133)	.32	-	
>15 years	286 (114)	188 (74)	.17	198 (57)	.73
F) Overall					
All	334 (230)	236 (172)	<.001*	396 (215)	<.001*

Note. — Statistics of reading time in seconds are shown for measurements where only status quo reference tools were used (Only SQ), where only the CBIR tool was used (Only CBIR) and where both were used (SQ+CBIR). Standard deviations on parentheses. P values indicate significant differences to reference level 'Only status quo' and were calculated using linear mixed effects models with individual readers and patients as random effects.

(only SQ 417s, only CBIR 276s p<.001), while they showed an increase when both CBIR and SQ tools were used together (only SQ 417s, SQ+CBIR 444s p=.049, **Tab. 4 E**).

Discussion

Our results indicate a significant positive impact on diagnostic accuracy with high effect sizes when using CBIR for characterizing various orbital lesions. Furthermore, we found evidence for a decrease in reading times when using only the CBIR tool but an increase in reading time when using CBIR in conjunction with status quo tools.

Our measured diagnostic accuracy of 55.88% with status quo reference tools only is comparable to other studies that assessed accuracy for orbital lesions (2,21). However, our measured status quo accuracy is considerably higher than status quo measurements of most studies that analyzed the effect of CBIR on interstitial lung disease diagnostics in chest CT. There, the reported diagnostic accuracies range between 35% (22) and 46.1% (9), except for (12) who reported 30% for novice and 60.7% for resident readers. The positive effect of the CBIR tool on diagnostic accuracy is comparable to the effects reported in Choe et al. (9) (without CBIR 46.1%, with CBIR 60.9%), but more moderate than the ones reported in other studies (12,22). In general, the measured diagnostic accuracy in our and other studies might underestimate the true diagnostic accuracy in the clinic, as only limited patient history and no laboratory data, nor reports from other sub-specialties were available to the participants.

The effect of CBIR on reading time is mixed in the literature. Haubold et al. (22) find an increase in reading time by 22% (p<0.001) which moderates to 7% after readers become more familiar with the software, whereas Röhrich et al. (10) find a decrease by 31.3% (p<0.001). In our study we found a significant 29% decrease in reading times when using only the CBIR tool, and a significant 19% increase when SQ+CBIR tools were used for diagnosing eye and orbit mass lesions. Other studies did not analyze whether the CBIR tool was used in conjunction with other tools, thus the two opposing effects could be conflated. However, our study may have overestimated reading times with the CBIR tool, since participants only read four cases having the CBIR tool available, thus they only had limited time to get used to the software and it might be lower under routine conditions.

In other studies, participants were required to read 54 (10) or more cases in total (9,17,22), which allows readers to become more familiar with the software but severely limits the total number of study participants

that could be included to 8 (9,10) or less (17,22). In our study, the low number of cases per participant allowed us to include 36 participants with considerable differences in experience and tenure, which better accounts for the heterogenous effects that AI assistance can have on radiologists (23). In addition, this and other studies (9,10) compared CBIR usage with status quo reference tools, whereas others compared CBIR assistance to no assistance at all (17,22), which may lead to different interpretations of the impact of CBIR on outcome variables.

This study has two main limitations. While we included a diverse range of cases and participants, the small sample size still limits the generalizability of our findings. Further studies will expand to a larger and more geographically diverse participant and case pool, ideally involving participants from multiple medical centers, which would provide more robust data and would allow for more granular sub-group analyses. Another concern is the potential for the CBIR tool to negatively influence radiologists by retrieving confusing or irrelevant cases, which was not evaluated. Given that 5 of 36 participants and 9 of 48 cases had lower diagnostic accuracy with the CBIR tool available than without, it is crucial to assess if there exist underlying systematic factors, either radiologist-specific or case-specific, that may lead to this disparate impact.

In conclusion, adopting CBIR in routine diagnostic workflows for eye and orbital mass lesions could have a substantial positive impact on radiological decision making and thus patient outcomes. However, more work is needed to assess the benefits of CBIR tools in other organ systems and imaging modalities. We plan to continue developing and refining the CBIR tool, expanding it to other organ systems and testing it in future studies.

Acknowledgements

We thank our participants for their time and dedication. We are grateful for the support from the Charité Institute for Biometrics for giving advice on the statistical analysis. K.E.E., W.L., S.S. and J.L.R. received funding from the Digital Health Accelerator of the Berlin Institute of Health. K.E.E. received support from Stiftung Charité. J.L.R. received support from the IFI program of the German Academic Exchange Service (DAAD).

Author Contributions

J.L.R. and K.E.E. conceived the study. W.L., A. L., B.W. and A.-S.I. retrieved the data. W.L. and S.-S.I. annotated the data. J.L.R and S.S. developed and deployed the software together with external service providers. W.L., B.W. and E.B.S conducted the experiments with the participants. J.L.R. did the statistical analysis. J.L.R. and K.E.E. wrote the manuscript. All authors revised the manuscript.

References

1. Newman-Toker DE, Nassery N, Schaffer AC, et al. Burden of serious harms from diagnostic error in the USA. BMJ Quality & Safety. BMJ Publishing Group Ltd; 2024;33(2):109–120.

2. Macedo S. Reliability of magnetic resonance imaging as a diagnostic tool for lacrimal gland tumors and predictors of a correct image-based diagnosis [PhD Thesis]. Charité-Universitätsmedizin Berlin; 2022.

3. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights into imaging. Springer; 2017;8:171–182.

4. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS medicine. Public Library of Science San Francisco, CA USA; 2018;15(11):e1002686.

5. Hirsch L, Huang Y, Luo S, et al. Radiologist-level performance by using deep learning for segmentation of breast cancers on MRI scans. Radiology: Artificial Intelligence. Radiological Society of North America; 2021;4(1):e200231.

6. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. Nature Publishing Group UK London; 2023;616(7956):259–265.

7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. Nature Publishing Group UK London; 2023;620(7972):172–180.

8. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877–1901.

9. Choe J, Hwang HJ, Seo JB, et al. Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest CT. Radiology. Radiological Society of North America; 2022;302(1):187–197.

10. Röhrich S, Heidinger BH, Prayer F, et al. Impact of a content-based image retrieval system on the interpretation of chest CTs of patients with diffuse parenchymal lung disease. European Radiology. Springer; 2023;33(1):360–367.

11. Haubold J, Zeng K, Farhand S, et al. AI co-pilot: content-based image retrieval for the reading of rare diseases in chest CT. Scientific Reports. Nature Publishing Group UK London; 2023;13(1):4336.

12. Pogarell T, Bayerl N, Wetzl M, et al. Evaluation of a novel content-based image retrieval system for the differentiation of interstitial lung diseases in CT examinations. Diagnostics. MDPI; 2021;11(11):2114.

13. Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:230407193. 2023;

14. Darcet T, Oquab M, Mairal J, Bojanowski P. Vision Transformers Need Registers. arXiv:2309.16588. 2023.

15. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards generalist foundation model for radiology. arXiv preprint arXiv:230802463. 2023;

16. Deng J, Guo J, Xue N, Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 4690–4699.

17. Pogarell T, Bayerl N, Wetzl M, et al. Evaluation of a novel content-based image retrieval system for the differentiation of interstitial lung diseases in CT examinations. Diagnostics. MDPI; 2021;11(11):2114.

18. Akaike H. Information theory and an extension of the maximum likelihood principle. Selected papers of hirotugu akaike. Springer; 1998. p. 199–213.

19. Heinze G, Wallisch C, Dunkler D. Variable selection–a review and recommendations for the practicing statistician. Biometrical journal. Wiley Online Library; 2018;60(3):431–449.

20. Singer JM, Rocha FM, Nobre JS. Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. International Statistical Review. Wiley Online Library; 2017;85(2):290–324.

21. Duron L, Heraud A, Charbonneau F, et al. A magnetic resonance imaging radiomics signature to distinguish benign from malignant orbital lesions. Investigative Radiology. LWW; 2021;56(3):173–180.

22. Haubold J, Zeng K, Farhand S, et al. AI co-pilot: content-based image retrieval for the reading of rare diseases in chest CT. Scientific Reports. Nature Publishing Group UK London; 2023;13(1):4336.

23. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nature Medicine. Nature Publishing Group US New York; 2024;1–13.

Supplement

Supplementary Table 1: Odds Ratios for Diagnostic Accuracy with CBIR

Characteristics	Only CBIR	P value	SQ+CBIR	P value
A) Difficulty				
Easy	5.69 (1.25 – 25.96)	.02*	3.89 (0.00 – Inf)	.99
Hard	1.36 (0.59 – 3.11)	.47	2.27 (0.60 – 8.61)	.23
Really hard	1.15 (0.26 – 5.08)	.86	7.19 (0.58 – 89.71)	.13
B) Pathology Type				
Infl. & Infect.	2.97 (0.97 – 9.02)	.055	4.38 (0.73 – 26.13)	.11
Benign	0.94 (0.24 – 3.64)	.93	6.87 (0.59 – 79.51)	.12
Malignant	1.88 (0.69 – 5.13)	.22	2.90 (0.52 – 16.18)	.23
C) Medical Role				
Resident	2.46 (0.86 – 7.04)	.09	2.87 (0.78 – 10.58)	.11
Board-certified	0.73 (0.21 – 2.55)	.62	1.97 (0.14 – 28.14)	.62
Senior	4.69 (1.40 – 16.33)	.01*	Inf (0.00 – Inf)	.99
D) Prior Experience				
No exp.	3.30 (0.70 – 12.03)	.10	Inf (0.00 – Inf)	.99
Little exp.	1.82 (0.80 – 4.17)	.15	3.16 (0.96 – 10.37)	.058
Sufficient exp.	1.21 (0.31 – 4.77)	.79		
C) Tenure				
0-5 years	2.38 (0.85 – 6.66)	.10	2.89 (0.79 – 10.55)	.11
6-10 years	2.43 (0.73 – 8.05)	.15	Inf (0.00 – Inf)	.90
11-15 years	0.63 (0.09 – 4.25)	.64		
>15 years	3.42 (0.61 – 19.31)	.16	2.80 (0.20 – 38.99)	.44
F) Overall				
All	2.07 (1.08 – 3.95)	.03*	3.65 (1.21 – 3.95)	.02*

Note. — Odds ratios of diagnostic accuracies for measurements where only the CBIR tool was used (Only CBIR) and when it was used together with status quo tools (SQ+CBIR), relative to reference level only status quo (Only SQ). Odds ratios and P values were calculated by using logistic mixed effects models with individual readers and patients as random effects. 95%- Wald confidence interval reported in parenthesis.

Supplementary Table 2: Adjusted impact of CBIR usage on Reading time

Characteristics	Only CBIR	P value	SQ+CBIR	P value
A) Difficulty				
Easy	0.84 (0.68 – 1.05)	.14	1.58 (1.12 – 2.23)	.01*
Hard	0.76 (0.64 – 0.89)	.002*	1.33 (1.04 – 1.71)	.03*
Really hard	0.88 (0.65 – 1.19)	.41	1.02 (0.66 – 1.56)	.94
B) Pathology Type				
Infl. & Infect.	0.73 (0.58 – 0.90)	.005*	1.33 (0.95 – 1.85)	.10
Benign	0.88 (0.64 – 1.18)	.40	1.25 (0.80 – 1.97)	.34
Malignant	0.71 (0.58 – 0.86)	<.001*	1.37 (1.01 – 1.85)	.045*
C) Medical Role				
Resident	0.58 (0.47 – 0.70)	<.001*	1.28 (1.00 – 1.61)	.046*
Board-certified	1.15 (0.87 – 1.53)	.34	1.08 (0.61 – 1.94)	.79
Senior	0.80 (0.63 – 1.02)	.08	1.14 (0.72 – 1.83)	.58
D) Prior Experience				
No exp.	0.95 (0.71 – 1.27)	.75	1.36 (0.87 – 2.11)	.19
Little exp.	0.64 (0.53 – 0.76)	<.001*	1.26 (0.99 – 1.58)	.054
Sufficient exp.	0.90 (0.67 – 1.21)	.50		
E) Tenure				
0-5 years	0.58 (0.48 – 0.71)	<.001*	1.27 (1.00 – 1.61)	.049*
6-10 years	0.93 (0.72 – 1.19)	.58	1.33 (0.80 – 2.25)	.29
11-15 years	1.24 (0.82 – 1.90)	.32		
>15 years	0.78 (0.55 – 1.10)	.17	0.91(0.54 – 1.53)	.73
F) Overall				
All	0.75 (0.66 – 0.86)	<.001*	1.32 (1.07 – 1.61)	<.001*

Note. — Exponential of the regression coefficients of log(reading time) in seconds are shown for measurements where only the CBIR tool was used (Only CBIR) and where it was used together with status quo tools (SQ+CBIR). P values indicate significant differences to reference level 'Only status quo' and were calculated using linear mixed effects models with individual readers and patients as random effects. 95%- Wald confidence interval reported in parenthesis.