Ophthalmologica

Ophthalmologica , DOI: 10.1159/000527345
Received: February 18, 2022
Accepted: September 29, 2022
Published online: October 10, 2022

Evaluation of an Artificial Intelligence-based Detector of Sub- and Intra-Retinal Fluid on a large set of OCT volumes in AMD and DME

Habra O, Gallardo M, Meyer zu Westram T, De Zanet S, Jaggi D, Zinkernagel M, Wolf S, Sznitman R

ISSN: 0030-3755 (Print), eISSN: 1423-0267 (Online)
https://www.karger.com/OPH
Ophthalmologica

Disclaimer:
Accepted, unedited article not yet assigned to an issue. The statements, opinions and data contained in this publication are solely those of the individual authors and contributors and not of the publisher and the editor(s). The publisher and the editor(s) disclaim responsibility for any injury to persons or property resulting from any ideas, methods, instructions or products referred to the content.

Copyright:
This article is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC) (http://www.karger.com/Services/OpenAccessLicense). Usage and distribution for commercial purposes requires written permission.

© 2022 The Author(s). Published by S. Karger AG, Basel

Research Article

Evaluation of an Artificial Intelligence-based Detector of Sub- and Intra-Retinal Fluid on a large set of OCT volumes in AMD and DME

Oussama Habra[1] *, Mathias Gallardo[2] *, Till Meyer zu Westram[2], Sandro De Zanet[3], Damian Jaggi[1], Martin Zinkernagel[1], Sebastian Wolf[1], Raphael Sznitman[2]

Affiliations:

[1] Department for Ophthalmology, Inselspital, University Hospital, University of Bern, Bern, Switzerland

[2] AIMI, ARTORG Center, University of Bern, Bern, Switzerland

[3] RetinAI Medical AG, Bern, Switzerland

*these authors contributed equally to this work

Running head: Evaluating a retinal fluid detector in AMD and DME

Corresponding Author: Oussama Habra, Department for Ophthalmology, Inselspital, University Hospital of Bern, Freiburgstrasse 18, 3010 Bern, Switzerland. E-mail: oussama.habra@insel.ch. Tel: 0041797147081.

Number of Tables: 2.

Number of Figures: 4.

Word count: 4320.

Keywords: artificial intelligence; intra-and subretinal fluid; optical coherence tomography; biomarker detection; neovascular AMD; DME.

Accepted Manuscript

Abstract
Introduction: In this retrospective cohort study, we wanted to evaluate the performance and analyze the insights of an artificial intelligence (AI) algorithm in detecting retinal fluid in spectral-domain OCT volume scans from a large cohort of patients with neovascular age-related macular degeneration (AMD) and diabetic macular edema (DME).
Methods: A total of 3'981 OCT volumes from 374 patients with AMD and 11'501 OCT volumes from 811 patients with DME, acquired with Heidelberg Spectralis OCT device (Heidelberg Engineering Inc., Heidelberg, Germany) between 2013 and 2021. Each OCT volume was annotated for the presence or absence of intraretinal fluid (IRF) and subretinal fluid (SRF) by masked reading center graders (ground truth). The performance of an already published AI-algorithm to detect IRF, SRF separately and a combined fluid detector (IRF and/or SRF) of the same OCT volumes was evaluated. An analysis of the sources of disagreement between annotation and prediction and their relationship to central retinal thickness was performed. We computed the mean areas under the curves (AUC) and under the precision-recall curves (AP), accuracy, sensitivity, specificity and precision.
Results: The AUC for IRF was 0.92 and 0.98, for SRF 0.98 and 0.99, in the AMD and DME cohort, respectively. The AP for IRF was 0.89 and 1.00, for SRF 0.97 and 0.93, in the AMD and DME cohort, respectively. The accuracy, specificity and sensitivity for IRF was 0.87, 0.88, 0.84, and 0.93, 0.95, 0.93, and for SRF 0.93, 0.93, 0.93, and 0.95, 0.95, 0.95 in the AMD and DME cohort respectively. For detecting any fluid, the AUC was 0.95 and 0.98, the accuracy, specificity and sensitivity was 0.89, 0.93, 0.90 and 0.95, 0.88 and 0.93, in the AMD and DME cohort, respectively. False positives were present when retinal shadow artifacts and strong retinal deformation were present. False negatives were due to small hyporeflective areas in combination with poor image quality. The combined detector correctly predicted more OCT volumes than the single detectors for IRF and SRF, 89.0% versus 81.6% in the AMD and 93.1% versus 88.6% in the DME cohort.
Discussion/Conclusion:
The AI-based fluid detector achieves high performance for retinal fluid detection in a very large dataset dedicated to AMD and DME. Combining single detectors provides better fluid detection accuracy than considering the single detectors separately. The observed independence of the single detectors ensures that the detectors learned features particular to IRF and SRF.

Accepted Manuscript

Introduction

The advent of high-resolution Optical Coherence Tomography (OCT) has led to the identification of biomarkers in diabetic retinopathy (DR) with diabetic macular edema (DME) and neovascular age-related macular degeneration (nAMD), delivering a vast amount of morphological information, which could be used to individualize treatment decisions [1-3]. In nAMD, subretinal fluid (SRF) is associated with a better visual outcome and a lower rate of transformation into geographic atrophy, while intraretinal fluid (IRF) has a poorer visual outcome with increasing rates of macular atrophy [4, 5]. In DME, the size and location of the macular cystoid spaces are predictive in the functional outcomes, large cysts (>200 μm) located in the outer nuclear layer being associated with poorer visual prognosis [6, 7].

However, the necessary human resources and expertise to assess these images are challenged with the increasing number of patients requiring OCT examinations for optimal management of macular diseases [5, 8]. In addition, even in the pivotal clinical trials discrepant qualitative assessment of fluid between ophthalmologists and certified reading center graders have been reported, leading to a significant number of missed treatments [9, 10].

Machine learning has already shown an immense cost-effective potential to assist ophthalmologists in clinical routine, such as detecting biological markers in retinal OCT-scans and predicting the treatment demand for a given patient [11-14]. Most of these works evaluated their proposed methods in controlled data, usually on a single dataset obtained from the same center, a single clinical trial, or using OCT scans with a single disease, selected manually by a specialist [11-13, 15-17]. However, this promising technology requires a stronger assessment of the performances and robustness, evaluating on larger datasets and understanding the discrepancies between algorithmic predictions and human annotations .

The aim of this study was to evaluate such a method in a more thorough way and to have a clearer view of its practical capability. For this, our evaluation relied on a large set of OCTs and two groups of pathologies and was intended to verify that the detectors of IRF and SRF learned specific features to IRF and SRF respectively. Along with reporting the general performance, we analyzed the performances according to central subfield thickness and identified the most important cases where the algorithm fails. The assessed method is a CE-marked deep-learning-based biomarker classifier capable of identifying the presence of SRF and IRF in OCT volumes and comparing it with reading center grader annotations in a systematic way.

Methods

Study Cohort and Evaluation Data

The evaluation was performed using completely anonymized OCTs volumes of two cohorts of patients, respectively including patients with exudative AMD and DME. We refer to them as AMD and DME cohorts. All patients received an anti-VEGF treatment, were treated and monitored between October 2013 and June 2021 and were initially treatment naïve.

We extracted anonymized OCT volumes from these patients from the database of the Ophthalmology department at the University Hospital of Bern, Bern, Switzerland, which have been imaged with a Spectralis SD-OCT imaging system (Heidelberg Engineering Inc., Heidelberg, Germany) and for which grader annotations were available. Our study focuses on OCT volumes using a 49 B-scans acquisition protocol and a resolution of 496 x 512 pixels per B-scan corresponding to an area of 5.90mm x 5.75mm x 1.92mm centered on the fovea. For each horizontal scan, 9 B-scans were averaged. We did not exclude patients with poor image quality.

In the AMD cohort we collected 3’981 OCT volume scans corresponding to 374 patients (748 eyes) for SRF and 3’619 OCT volume scans from 371 patients (742 eyes) for IRF. In the DME cohort, we collected 11’501 OCT volume scans from 809 patients (1618 eyes) for SRF and 11’499 OCT volume scans related to 811 patients (1’620 eyes) for IRF. The discrepancies in the number of OCT volume scans for IRF and SRF are explained by the fact that only the annotated ones were taken into consideration, ungradable OCT volume scans being excluded.

Annotation Protocol – Reading Center Assessments

A central reading center (Bern Photographic Reading Center, Inselspital, University hospital Bern, department of ophthalmology, Bern, Switzerland) performed an independent, masked review of all OCT volume scans. For each above-mentioned cohort, a grading protocol and grading form were used, containing for instance the definitions of morphological changes/biomarkers. Each OCT volume scan was graded independently by two certified-graders and without access to each other's grading. In case of disagreement among the annotations of both graders, a third evaluation by a senior grader was performed. Details of the annotation protocols and biomarkers definition are available as supplementary material.
Figure 1 reports the distribution of present and absent biomarkers in both studies, as annotated by the graders of the central reading center.

Automatic System for Biomarkers Detection

The biomarker detector we evaluated in this study is the CE-marked Discovery® OCT Biomarker detector (RetinAI AG, Switzerland) the one from Kurmann et al., 2019 [12]. We used the trained model from Kurmann et al., 2019, but evaluated on a different dataset described earlier. This model aims to detect the presence probability of a set of 10 biomarkers at the B-scan level. These include: SRF, IRF, hyperreflective foci (HF), drusen, reticular pseudodrusen (RPD), epiretinal membrane (ERM), geographic atrophy (GA), outer retinal atrophy (ORA), fibrovascular pigment epithelial detachment (FPED) and healthy B-scans (defined by the absence of all the listed biomarkers). This detector relies on a CNN architecture and was trained in a supervised way on B-scans for which the presence of these 10 biomarkers were manually annotated. 23'030 annotated B-scans were used to train this architecture and 1'029 B-scans were used for the evaluation performance reported by the authors. Training pathologies included early, intermediate, and late AMD, as well as DR but without DME. It is important to note that none of the B-scans used for model development, *i.e.,* training and evaluation, are included in the present study. We refer the readers to the paper for the description of these biomarkers, the annotation protocol, and the distribution of each biomarker in the training and testing set [12]. While Kurmann et al. evaluated the detection of 10 biomarkers at the B-scan level, we evaluated the detection of IRF and SRF at the volume level. This is because the 15'482 OCT volumes used in our work are annotated at the volume level for IRF and SRF, as explained in the previous section. As the biomarker detector gives the presence probability of the B-scan level, we used the maximum presence probability of the biomarkers across the whole volume as the final predicted presence probability for the OCT volume.

Evaluation

The evaluation was conducted by comparing the predictions of the biomarker detection model and the annotations performed by the BPRC independently in both cohorts. First, the performance of the model was measured using the area under the receiver operating characteristic (ROC) curve (AUC) with 95% confidence intervals (CI), sensitivities and specificities. ROC curves were computed separately for IRF and SRF.
Second, we computed the Precision and Recall (PR) curves, which are useful in case of distribution imbalance between two classes of a binary model.
Third, we performed an analysis of the sources of disagreement between annotation and prediction. For this purpose, the false positive and false negative rate of each cohort and of each biomarker were calculated. The operating point of the predictive model was set to maximize sensitivity and specificity for each biomarker in both cohorts. Among the false positive and false negative samples, we also identified the main categories of the underlying errors. Some representative cases with strong disagreements between the annotation and the prediction were manually inspected and reported here. This aims to identify the current challenges and evaluate whether these are inherent to the collected data, the acquisition protocol or the evaluated system.
Fourth, the retinal thickness of the false positive and false negative samples was compared to that of true negative and true positive.

Lastly, we wished to analyze the relationship of the detectors of IRF and SRF and the performance of a joint detector of IRF and SRF, referred to as *fluid detector*. We first aimed to identify any covariate between the two single detectors, *i.e.*, another variable/factor, which affects the detection of fluid, and which appears only when the IRF and SRF are considered jointly. Thus, we compared the final predictions, *i.e.*, after setting an operating point, of the single detectors of IRF and SRF to the ones of the fluid detector. The operating point was set to maximize sensitivity and specificity. The output probability of the joint detector is set to the maximum probability between the output of the IRF and SRF detectors.

We used Python 3 and its open-source package Scikit-learn v0.20.3 for the ROC and PR implementation and AUC/AP calculations. Statistical software were used to conduct data analysis (Prism 9.2.0; GraphPad Software, Inc., La Jolla, CA, USA, and IBM, SPSS statistics, Version 21; SPSS Inc, Chicago, IL, USA). To compare retinal thickness values among the subgroups, we performed a Welch's unequal variances t-test, as the groups have different size and variance.

Results

Performance of IRF and SRF Detection

AMD cohort

Predicted and annotated IRF and SRF presence correlated strongly in the AMD cohort. The AUC per OCT-volume scan for IRF detection was 0.92 (CI = 0.91-0.93, n = 3'618) and the AUC for SRF detection per OCT-volume scan was 0.98 (CI = 0.98, n = 3'981). These results are presented in Figure 2A and demonstrate the robust capability of the deep-learning algorithm to accurately predict the presence of IRF and SRF among a significant number of OCT volume scans. The computed precision-recall (PR) curves with the calculated APs are presented in Figure 2B and showed an AP per OCT volume scan of 0.97 (CI = 0.96-0.98) and 0.89 (CI = 0.87-0.90) for SRF and IRF, respectively.

DME cohort

Predicted and annotated IRF and SRF also showed a high level of correlation in the DME cohort. The AUC for IRF (n=11'499) and SRF (n= 11'501) per OCT-volume scan was 0.98 (CI = 0.98-0.99) and 0.99 (CI = 0.98-0.99), respectively. These results are displayed in Figure 2C. The PR curves are presented in Figure 2D and supported our observations with an AP per OCT-volume scan for SRF of nearly 1.00, though a modest decrease in the calculated AP regarding SRF was found, with an AP per OCT-volume scan of 0.93 (CI 0.91-0.94).

Annotation and Prediction Disagreement, Analysis of Representative Cases

Figure 3 summarizes the total number of false positive (FP), false negative (FN), true negative (TN) and true positive (TP) samples, the ground truth being the graders annotation. The threshold value used to assign each OCT volume scan to the categories was calculated with the aim of maximizing sensitivity and specificity for each cohort and each biomarker. The threshold for IRF was 0.946 and 0.985 and for SRF 0.95 and 0.68 in the AMD and DME cohorts, respectively. For instance, an OCT volume scan in the AMD cohort with IRF annotated as present by the grader and with a prediction probability of 0.94 was classified as false negative.

To understand the sources of disagreement, 15% of false negative and false positive samples of each biomarker were randomly selected and manually analyzed. Figure 4 illustrates eight representative sources of discrepancies. In the AMD cohort, cases 1 and 3 present false positive results for IRF and SRF, respectively. Cases 2 and 4 display false negative results for IRF and SRF, respectively. Regarding the DME cohort, cases 5 and 7 show false positive samples and cases 6 and 8 false negative examples for IRF and SRF, respectively.

Relation to central subfield thickness of the retina

We compared the mean central subfield thickness (CST) value of the false negative OCT volume scans to that of true positive, and as well as the false positives to true negatives the results are displayed in Table 1. The definition of the CST was the average value of the retinal thickness in the central-1mm ETDRS region, where the retina is defined as the space between the ILM and Bruch's membrane.

Accepted Manuscript

Independence of the single detectors and analysis of a combined fluid detector

Table 2 compares the number of OCTs correctly predicted (true positives and true negatives) by the single detectors of IRF and SRF, when combined into a single fluid detector and reports the number of OCTs correctly predicted by the combination of the single detectors and the fluid detector, as well as the accuracy, specificity, sensitivity and precision. The proportion of OCT volume scans annotated with fluid (IRF and/or SRF) was 54.64% in the AMD cohort and 89.05% in the DME cohort.

We note, in general, that there were very high percentages of correctly predicted samples across the different detectors, especially in the DME cohort. Checking independence of the two single detectors led to consider the product of the probabilities of correct detection by the IRF and SRF detectors (rows 1 and 2 in Table 2), respectively and to compare the product to the probability of correct detection of IRF and SRF simultaneously. We used the detector accuracy as the probability of correct detection. We then obtained a probability of 0.8684x0.9384=0.8149 of correct detection of any fluid when considering the two single detectors separately and a probability of 0.8159 of correct detection for the combined fluid detector. Such a small difference supported the independence between the two single detectors.

We report very good performances of the fluid detection (IRF and SRF) with an AUC of 0.95 and 0.98 for AMD and DME, respectively, and an average precision of 0.97 and 1.00 for AMD and DME, respectively. In Table 2, we observe that the fluid detector presented a higher accuracy than the combination of the two single detectors. A more precise analysis of the performance of the single detectors in comparison to the combined detector is presented in the supplementary material.

Discussion

IRF and SRF detection

The presence or absence of pathological fluid in exudative retinal diseases is of great importance to evaluate disease activity. Our experimental results demonstrate that the evaluated deep learning algorithm can predict the presence of both major features of exudative disease activity, IRF and SRF, on a large dataset of OCT volumes of patients with AMD and DME with a very high degree of accuracy. In addition, since the testing data differ from the training data, this study demonstrates a powerful external validation of an AI-based algorithm. These excellent results are supported by the computed area under the precision-recall curves (Figure 2), which do not consider the true negative classes. However, we noticed a slight decrease in the AP of IRF (0.89) in the AMD cohort and SRF (0.93) in the DME cohort as compared to their respective AUC. The accuracy, sensitivity, specificity, and precision (Table 2) for IRF were superior in the DME cohort (0.93, 0.93, 0.95 and 0.99) than in the AMD cohort (0.87, 0.84, 0.88 and 0.76). However, this trend was reversed for SRF, with sensitivity and precision of 0.93 and 0.88 in the AMD cohort versus 0.87 and 0.83 in the DME cohort. The accuracy and specificity for SRF in the AMD cohort remained slightly lower than in the DME cohort (0.95 and 0.93 versus 0.95 and 0.95, respectively). The distribution of the presence probabilities of the biomarkers can explain these discrepancies, as the positive class for IRF in the AMD and SRF in the DME cohort was lower than SRF and IRF in the AMD and DME cohorts, respectively (Figure 1). An alternative explanation for the slightly lower performance for IRF in the AMD cohort is the relatively high false positive rate of 8.18 % and false negative rate of 4.98 % (Figure 3a), suggesting a greater challenge for the evaluated system in discriminating IRF from degenerative small cystoid fluid or retinal shadow artifacts. A diffuse non-cystoid fluid accumulation with clearer borders could also be considered as an additional factor, as this kind of fluid is more difficult to distinguish in OCT volume scans of low quality. The slight improvement in performance in the DME cohort may also lie in the intrinsic characteristic of retinal lesions architecture, as AMD lesions tend to have more pigment epithelial detachments (PEDs), which sometimes distort heavily the image, complicate focusing in the context of a thicker retina and expand into areas that might be prone to IRF [7, 18].

Our results are also in line with the results of recent works on automated detection and/or quantification of retinal fluid, which demonstrated high accuracy in detecting retinal fluid such as IRF and SRF [19, 20], detecting any retinal fluid [21-23] or giving a binary yes-no classifier of the presence of an exudative disease [24-26]. In our experiment, we focused on the accuracy of IRF and SRF detection as well as the ability of the evaluated algorithm to discriminate between these two features, since IRF and SRF are associated with variable visual outcomes [4, 7, 27]. The developed method of Schlegl et al. [19] showed excellent results on a testing set containing 1'200 OCT volume scans from 400 patients with AMD, DME and retinal vein occlusion (RVO), of which 65 and 100 contained IRF and 69 and 11 contained SRF in AMD and DME, respectively, and were performed with the a Spectralis device as the present study. However, a decrease in the sensitivity for SRF in eyes with DME was noticed by the authors and explained by the fact that the algorithm was solely trained on AMD and RVO cases and SRF is scarce in DME patients. Lu et al. [20] obtained promising similar results concerning SRF in DME on testing their method on an unseen dataset containing 750 B-scan images. Our results suggest that these challenges can be overcome with a larger sample size and a sufficient amount of positive classes, and showed that training an algorithm on a set with different biomarker distribution among multiple exudative diseases does not necessarily reduce the detection performance. Furthermore, testing a deep-learning approach on a set with entire OCT volume scans (in contrast to individual B-scans) may be beneficial in the perspective of a potential upcoming generalized clinical application, allowing the training of an algorithm to come as close as possible to routine conditions.

Discrepancies between algorithmic predictions and human annotations

In recent years, an increasing number of research groups focused their works on automated detection and segmentation of retinal fluid, comparing their results to expert human graders [28, 29]. In this way, Kurmann et al. [12] provided some additional insights on the biomarker detector for which we propose a more thorough evaluation in this paper. We focused in this paper on a qualitative and quantitative analysis of cases with disagreement to unveil the limitations of the system predictions.

Explanation of the major sources of errors

IRF was mistakenly predicted when retinal shadow artifacts caused by anterior hyperreflective opacities were present (Figure 4, case 1) as well as in cases of strong retinal alteration with poor acquisition quality and/or more segmentations failures, where the low contrast resulted in a darkening of the neurosensory retina, which in turn were falsely interpreted as fluid (case 5). The mean CST of the false positive volume scans regarding IRF in the AMD cohort was higher than the one for true negatives (p=<0.0001), which supports our previous finding. However, these observations were not applicable to the false positive samples for IRF in the DME cohort, where no significant difference with the true negatives was detected (p=0.559). In addition, the false positive rate of IRF in DME (0.52%) was lower than in AMD (8.18%). This could be explained by the very small sample size of the false positive class for IRF in the DME cohort, as nearly 90% of the DME dataset was annotated with IRF present (Figure 1) and because low acquisition quality represents the principal factor, leading to false positive results in this cohort (Figure 4, case 5).

The vast majority of missed IRF was attributed to the presence of small hyporeflective areas in combination with poor image quality and resulting lower contrast (cases 2 and 6). This observation is supported by the comparison of the CST of each positive and negative class. Regarding IRF in both cohorts, the mean CST of the false negative samples was significantly smaller than the one of the true positive volume scans (AMD; p=0.0048, DME; p=<0.0001). A previous study comparing the performance of an AI-algorithm to retinal specialists [27], showed that the estimated fluid volume in the false negative cases was significantly lower to that of the true positives. The clinical implication of this observation is questionable, as the presence of peripheral small degenerative cystoid fluid does not always lead to a modification of the therapy management in DME and as degenerative cystoid fluid does not necessarily respond to anti-VEGF therapy [31,35] . Therefore, the ability to detect

small amounts of fluid does not necessarily provides an advantage to an automated method, but underlines that an algorithm can be considered as a second grader in the scope of clinical trials. The algorithm tended also to overestimate the presence of SRF, even if the homogeneous hyporeflective area was less than 100µm in a horizontal extent (case 3). This situation occurred mostly in cases of strong retinal deformation with retinal shadowing (case 7), but did not result in a significant difference in the scope of the CST concerning the false positive and true negative samples in the AMD cohort (p=0.112). In the DME cohort, SRF was falsely predicted as present in cases with strong retinal deformation (case 7), which resulted in a slightly thicker CST in the false positive subgroup compared to the true negatives (p=0.0088).

Concerning SRF in both cohorts, the mean CST value of the false negatives did not differ significantly from the one of the true positives (AMD; p=0.2526, DME; p=0.185). This could be explained either in the light of the relatively small sample size or in the context of extended retinal alterations within the false negative samples, causing shadow artifacts and/or low-quality acquisition (cases 4 and 8). Further, a marked difference was observed between the false negative rates of SRF in the AMD (2.54%) and DME cohorts (0.39%), which could be explained by the biomarker distribution, given that only 8.5% of the DME set was annotated with SRF versus 34.9% of the AMD set (Figure 1).

This non-exhaustive analysis of cases provides an interesting insight into the characteristics of the volume scans where the algorithm encountered difficulties, suggesting that clinical relevant areas in an OCT scan are considered as critical in the classification task of a deep-learning algorithm, which correlate with previous studies [27,30]. We must add that the standard deviation values were relatively high, indicating a certain dispersion around the mean, which could negatively influence our findings. Furthermore, the relatively small sample size of the false classes emphasizes the robustness of the evaluated system but limits the interpretation of Table 1.

Independence of the IRF and SRF detectors and analysis of a combined fluid detector

We found, overall, a high diagnostic accuracy, specificity, sensitivity and precision in detecting fluid on a large set of OCT volume scans of patients with DME and AMD (Table 2). These metrics are slight higher than the one reported by the recent published prospective study [22]. The observed independence of the single detectors is a positive characteristic, which ensures in some ways that the detectors of IRF and SRF learned features particular to IRF and SRF respectively. The higher accuracy of the fluid detector compared to the two single detectors suggests that combining single detectors in the case of fluid provides better fluid detection accuracy than considering separately the single detectors. These observations support the development of an AI-based assistance for clinicians with the intention of minimizing missed cases of fluid detection and thus improving the visual outcome. A combined fluid detector cannot discriminate the fluid type but would be relevant in such a context. Another interesting application would be in clinical trial settings with large numbers of OCT volume scans requiring grading, where such a detector could be used as a supportive tool to reduce the burden on expert graders by not sending for annotations the OCTs where no fluid is detected. This would allow a better allocation of workload of the human graders and could improve the performance of reading centers. Lastly, such an algorithm could be implemented in clinical routine as a supportive tool to screen patients with active exudative macular disease, especially in settings where ophthalmologists are not readily available.

Limitations and future works

A major limitation of our study is its retrospective design and resulting lack of data and possible selection bias, as the OCT volume scans were completely anonymized. Second, our data was obtained from devices from a same manufacturer, which severely limits the generalization of the evaluated system. Third, our dataset comprised OCT volume scans with a single disease per OCT. The performance of the tested algorithm could be lower in the real-world setting, especially in presence of two or more distinct pathologies. Fourth, as we considered the grader's annotations as ground truth, possible human errors might have appeared and led to incorrect results. However, the excellent correlation between annotations and prediction in the present study and the low probabilities of three expert-graders being mistaken, somewhat mitigates this limitation. Fifth, the

evaluated system faces difficulties in discriminating IRF from intraretinal degenerative cysts, which can be tackled by a larger collection of this biomarker to improve the capability to distinguish cysts. However, we obtained an overall small false positive rate, indicating an acceptable global. In order to generalize the application in clinical routine, further evaluation using real-life data is needed. A continuation of this work could focus on the quantification of fluid and testing and validating the evaluated AI-based algorithm on different OCT devices and others biomarkers. We were able to illustrate the ability of such an algorithm to learn specific features about IRF and SRF, and its robustness according to many CST, *i.e.* in different pathomorphological stages in DME and AMD. This study presents some interesting insights to understand the limitations of an automated method, which appears to approach the human levels in grading difficult OCT volume scans. Further works are needed to understand to which extend an algorithm transposes bias, especially human biases and how this affects its performance. Similarly, future work could include saliency maps combined with different grading methods in order to improve the interpretability of an automated method.

Statements

Ethics

This study includes human subjects. The Study was performed in accordance with International Conference on Harmonization Good Clinical Practice guidelines and the tenets of the Declaration of Helsinki. The study protocol was reviewed and approved by the Ethics Committee of the Canton of Bern (Kantonale Ethikkomission Bern), Switzerland, approval number 2021-01280, and written informed consent to participate in the study was obtained from participants.

Conflict of Interest Statement

The authors have no conflicts of interests regarding this study to declare. We refer the readers to the ICMJE Form for Disclosure of Potential Conflicts of Interest of each author.

Funding Sources

There was no financial support to be declared for this study.

Author Contributions

Mathias Gallardo and Oussama Habra contributed equally in research design, data acquisition, research execution, analysis, interpretation and manuscript preparation. Till Meyer zu Westram contributed in Data acquisition, research execution and data analysis and interpretation. Sandro De Zanet contributed in Data acquisition and research execution. Damian Jaggi contributed in Data acquisition and interpretation. Martin S. Zinkernagel and Sebastian Wolf contributed in research design, data analysis and interpretation. Raphael Sznitman contributed in research design, data analysis, interpretation and manuscript preparation.

Data Availability Statement

The data in this study was obtained from the Bern Photographic Reading Centre (BPRC) where restrictions may apply. Such dataset may be requested from the Bern Photographic Reading Centre (BPRC), bprc@insel.ch, Universitätsklinik für Augenheilkunde, Inselspital, Freiburgstrasse, 3010 Bern.

References

1. Yu S, Rückert R, Munk MR. Treat-and-extend regimens with anti-vascular endothelial growth factor agents in age-related macular degeneration. Expert Review of Ophthalmology. 2019;14(6):287-307.

2. Wong DT, Berger AR, Bourgault S, Chen J, Colleaux K, Cruess AF, et al. Imaging Biomarkers and Their Impact on Therapeutic Decision-Making in the Management of Neovascular Age-Related Macular Degeneration. Ophthalmologica Journal international d'ophtalmologie International journal of ophthalmology Zeitschrift fur Augenheilkunde. 2021;244(4):265-80.

3. Flores R, Carneiro Â, Vieira M, Tenreiro S, Seabra MC. Age-Related Macular Degeneration: Pathophysiology, Management, and Future Perspectives. Ophthalmologica Journal international d'ophtalmologie International journal of ophthalmology Zeitschrift fur Augenheilkunde. 2021;244(6):495-511.

4. Jaffe GJ, Ying GS, Toth CA, Daniel E, Grunwald JE, Martin DF, et al. Macular Morphology and Visual Acuity in Year Five of the Comparison of Age-related Macular Degeneration Treatments Trials. Ophthalmology. 2019;126(2):252-60.
5. Schmidt-Erfurth U, Waldstein SM. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. Progress in retinal and eye research. 2016;50:1-24.
6. Markan A, Agarwal A, Arora A, Bazgain K, Rana V, Gupta V. Novel imaging biomarkers in diabetic retinopathy and diabetic macular edema. Therapeutic advances in ophthalmology. 2020;12:2515841420950513.
7. Zur D, Iglicki M, Busch C, Invernizzi A, Mariussi M, Loewenstein A. OCT Biomarkers as Functional Outcome Predictors in Diabetic Macular Edema Treated with Dexamethasone Implant. Ophthalmology. 2018;125(2):267-75.
8. Bourne RRA, Flaxman SR, Braithwaite T, Cicinelli MV, Das A, Jonas JB, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. The Lancet Global health. 2017;5(9):e888-e97.
9. Martin DF, Maguire MG, Fine SL, Ying GS, Jaffe GJ, Grunwald JE, et al. Ranibizumab and bevacizumab for treatment of neovascular age-related macular degeneration: two-year results. Ophthalmology. 2012;119(7):1388-98.
10. Monés J, Singh RP, Bandello F, Souied E, Liu X, Gale R. Undertreatment of Neovascular Age-Related Macular Degeneration after 10 Years of Anti-Vascular Endothelial Growth Factor Therapy in the Real World: The Need for A Change of Mindset. Ophthalmologica Journal international d'ophtalmologie International journal of ophthalmology Zeitschrift fur Augenheilkunde. 2020;243(1):1-8.
11. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine. 2018;24(9):1342-50.
12. Kurmann T, Yu S, Márquez-Neila P, Ebneter A, Zinkernagel M, Munk MR, et al. Expert-level Automated Biomarker Identification in Optical Coherence Tomography Scans. Scientific reports. 2019;9(1):13605.
13. Gallardo M, Munk MR, Kurmann T, De Zanet S, Mosinska A, Karagoz IK, et al. Machine learning can predict anti-VEGF treatment demand in a Treat-and-Extend regimen for patients with nAMD, DME and RVO associated ME. Ophthalmology Retina. 2021.
14. Rêgo S, Dutra-Medeiros M, Soares F, Monteiro-Soares M. Screening for Diabetic Retinopathy Using an Automated Diagnostic System Based on Deep Learning: Diagnostic Accuracy Assessment. Ophthalmologica Journal international d'ophtalmologie International journal of ophthalmology Zeitschrift fur Augenheilkunde. 2021;244(3):250-7.
15. Moraes G, Fu DJ, Wilson M, Khalid H, Wagner SK, Korot E, et al. Quantitative Analysis of OCT for Neovascular Age-Related Macular Degeneration Using Deep Learning. Ophthalmology. 2021;128(5):693-705.
16. Schmidt-Erfurth U, Bogunovic H, Sadeghipour A, Schlegl T, Langs G, Gerendas BS, et al. Machine Learning to Analyze the Prognostic Value of Current Imaging Biomarkers in Neovascular Age-Related Macular Degeneration. Ophthalmology Retina. 2018;2(1):24-30.
17. Lee CS, Baughman DM, Lee AY. Deep learning is effective for the classification of OCT images of normal versus Age-related Macular Degeneration. Ophthalmology Retina. 2017;1(4):322-7.
18. Mrejen S, Sarraf D, Mukkamala SK, Freund KB. Multimodal imaging of pigment epithelial detachment: a guide to evaluation. Retina (Philadelphia, Pa). 2013;33(9):1735-62.
19. Schlegl T, Waldstein SM, Bogunovic H, Endstraßer F, Sadeghipour A, Philip AM, et al. Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning. Ophthalmology. 2018;125(4):549-58.
20. Lu D, Heisler M, Lee S, Ding GW, Navajas E, Sarunic MV, et al. Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. Medical image analysis. 2019;54:100-10.

21. Chakravarthy U, Goldenberg D, Young G, Havilio M, Rafaeli O, Benyamini G, et al. Automated Identification of Lesion Activity in Neovascular Age-Related Macular Degeneration. Ophthalmology. 2016;123(8):1731-6.
22. Keenan TDL, Clemons TE, Domalpally A, Elman MJ, Havilio M, Agrón E, et al. Retinal Specialist versus Artificial Intelligence Detection of Retinal Fluid from OCT: Age-Related Eye Disease Study 2: 10-Year Follow-On Study. Ophthalmology. 2021;128(1):100-9.
23. Mantel I, Mosinska A, Bergin C, Polito MS, Guidotti J, Apostolopoulos S, et al. Automated Quantification of Pathological Fluids in Neovascular Age-Related Macular Degeneration, and Its Repeatability Using Deep Learning. Transl Vis Sci Technol. 2021;10(4):17.
24. Sidibé D, Sankar S, Lemaître G, Rastgoo M, Massich J, Cheung CY, et al. An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images. Computer methods and programs in biomedicine. 2017;139:109-17.
25. Rim TH, Lee AY, Ting DS, Teo K, Betzler BK, Teo ZL, et al. Detection of features associated with neovascular age-related macular degeneration in ethnically distinct data sets by an optical coherence tomography: trained deep learning algorithm. British Journal of Ophthalmology. 2021;105(8):1133.
26. Li F, Chen H, Liu Z, Zhang X, Wu Z. Fully automated detection of retinal disorders by image-based deep learning. Graefe's archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie. 2019;257(3):495-505.
27. Sophie R, Lu N, Campochiaro PA. Predictors of Functional and Anatomic Outcomes in Patients with Diabetic Macular Edema Treated with Ranibizumab. Ophthalmology. 2015;122(7):1395-401.
28. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018;172(5):1122-31.e9.
29. Wu J, Philip A-M, Podkowinski D, Gerendas BS, Langs G, Simader C, et al. Multivendor Spectral-Domain Optical Coherence Tomography Dataset, Observer Annotation Performance Evaluation, and Standardized Evaluation Framework for Intraretinal Cystoid Fluid Segmentation. J Ophthalmol. 2016;2016:3898750-.

Figure and table legends

Fig. 1. Distribution of intraretinal fluid (IRF) and subretinal fluid (SRF) in the Age related macular degeneration (AMD) and Diabetic macular edema (DME) cohorts, as annotated by the experts of the Bern Photographic Reading Center (BPRC). The "n" value indicates the total number of annotated (*i.e.*, present/absent) OCT volume scans for IRF and SRF in both cohorts. In the DME cohort, we included 11'499 and 11'501 annotated OCT volumes scans for IRF and SRF, respectively. In the AMD cohort, we included 3'618 and 3'981 annotated OCT volume scans for IRF and SRF, respectively.

Fig. 2. Receiver operating characteristic (ROC) and precision-recall (PR) curves on detection performance of intraretinal fluid (IRF) and subretinal fluid (SRF). **First row:** Age-related macular degeneration (AMD). **Second row:** Diabetic macular edema (DME). **A and C**, ROC curves. **B and D**, PR curves. The area under the curve (AUC) and area under the precision-recall curve (AP) with the confidence intervals as measures of general performance are specified in parentheses.

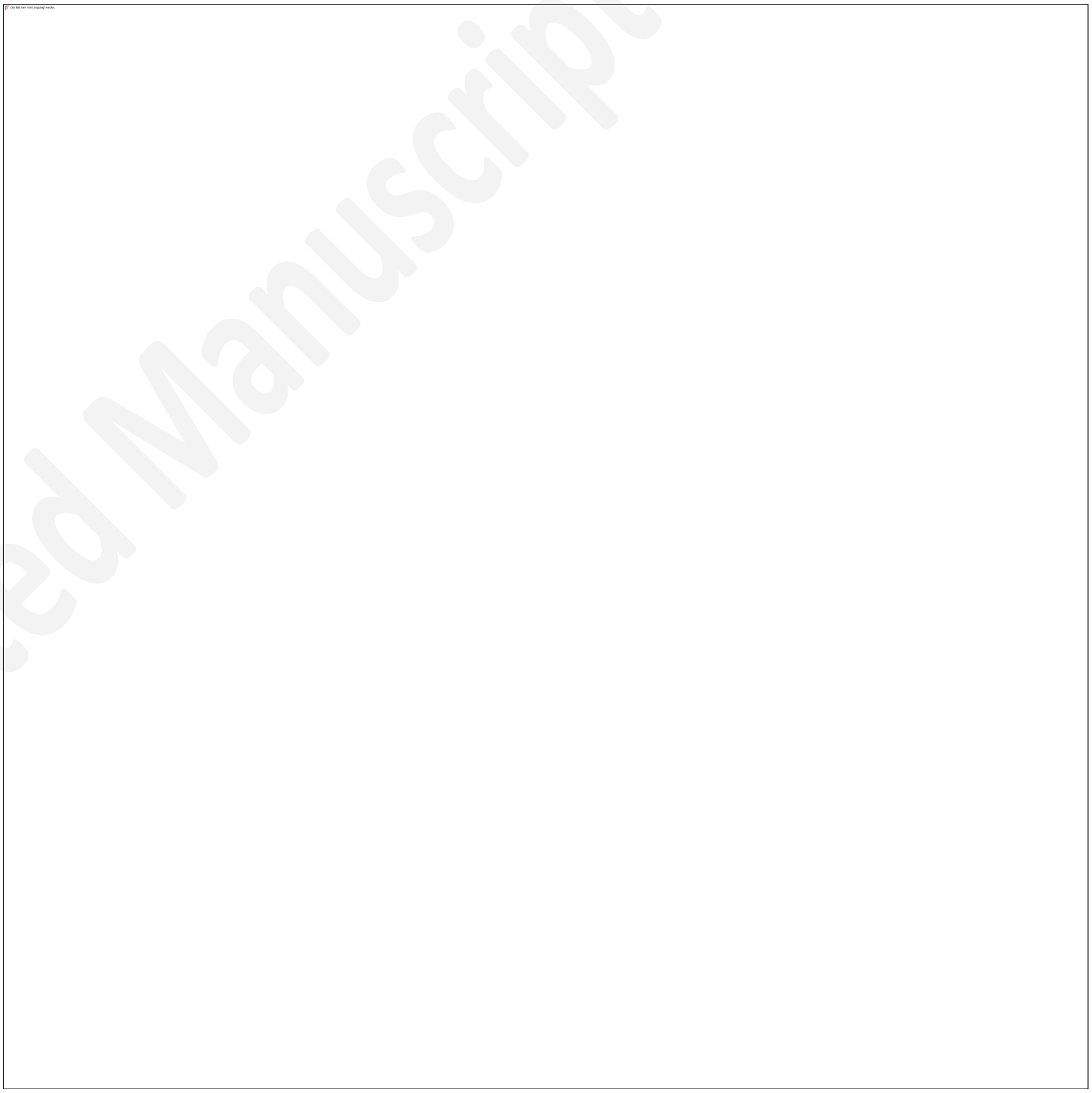
Fig. 3. Confusion matrix on detection performance (prediction, x-axis) of intraretinal fluid (IRF) and subretinal fluid (SRF), the samples divided in four categories based on the grader's annotations (ground truth, y-axis); true negative (TN), false positive (FP), false negative (FN) and true positive (TP). **First row:** Age-related macular degeneration (AMD). **Second row:** Diabetic macular edema (DME). **A and C**, IRF. **B and D**, SRF. The threshold was set to maximize sensitivity and specificity for each group and each biomarker. The horizontal bottom bar gives the color code from the minimum (purple) to the maximum (yellow) sample numbers for the four categories.

Fig. 4. Eight OCT volumes with discrepancies between the annotation and prediction of intraretinal fluid (IRF) and subretinal fluid (SRF). The study cohort, evaluated fluid, grader annotation, prediction of the evaluated system with presence probability in parenthesis and source of discrepancy are displayed above the OCT scan image. For each case, in the bottom left of the OCT scan image is the image-index reported (out of 49 B-scans in a volume). The location of the error source is framed in red with a zoomed view. The computed presence probability for both detectors per OCT B-scan image over the whole OCT volume-scan (total 49 images) is presented underneath the OCT scan image. The red-box corresponds to the image-index of the displayed OCT images. The presence probability is illustrated with a color code, ranging from purple (0%, absent) to yellow (100%, present).

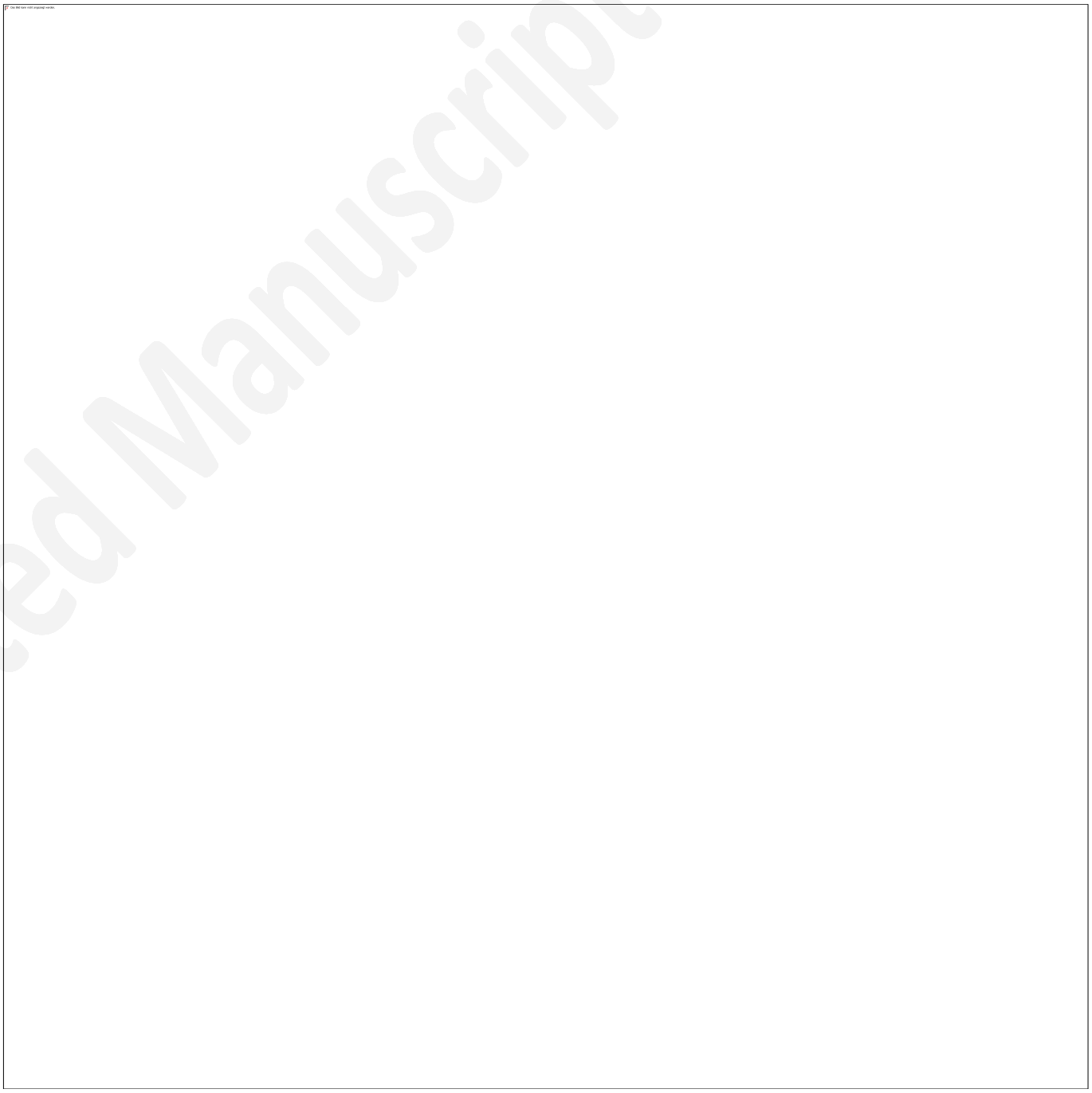
Table 1. AMD = age-related macular degeneration; DME = diabetic macular edema; IRF = intraretinal fluid; SRF = subretinal fluid. The "n" value indicates the number of each OCT volume scan per disease/group. We collected the mean of central subfield thickness of the false negative, true positive, false positive and true negative OCT volume scans. The corresponding standard deviations are specified in parentheses. The "p" value represents the statistical significance of the mean differences between the false negative and true positive as well as false positive and true negative. We used a two-sided Welch's unequal variances t-test with a level of significance alpha of 0.05.

Table 2. AMD = age-related macular degeneration; DME = diabetic macular edema; IRF = intraretinal fluid; SRF = subretinal fluid. For both cohorts, we collected the threshold for fluid detection, accuracy, specificity, sensitivity, precision and the number of OCTs correctly predicted by the single detector of IRF, the single detector of SRF, the combination of the two single detectors, the detector of fluid, *i.e.,* IRF and SRF, and the number of OCTs simultaneously well predicted by the fluid detector and the combination of the two single detectors.

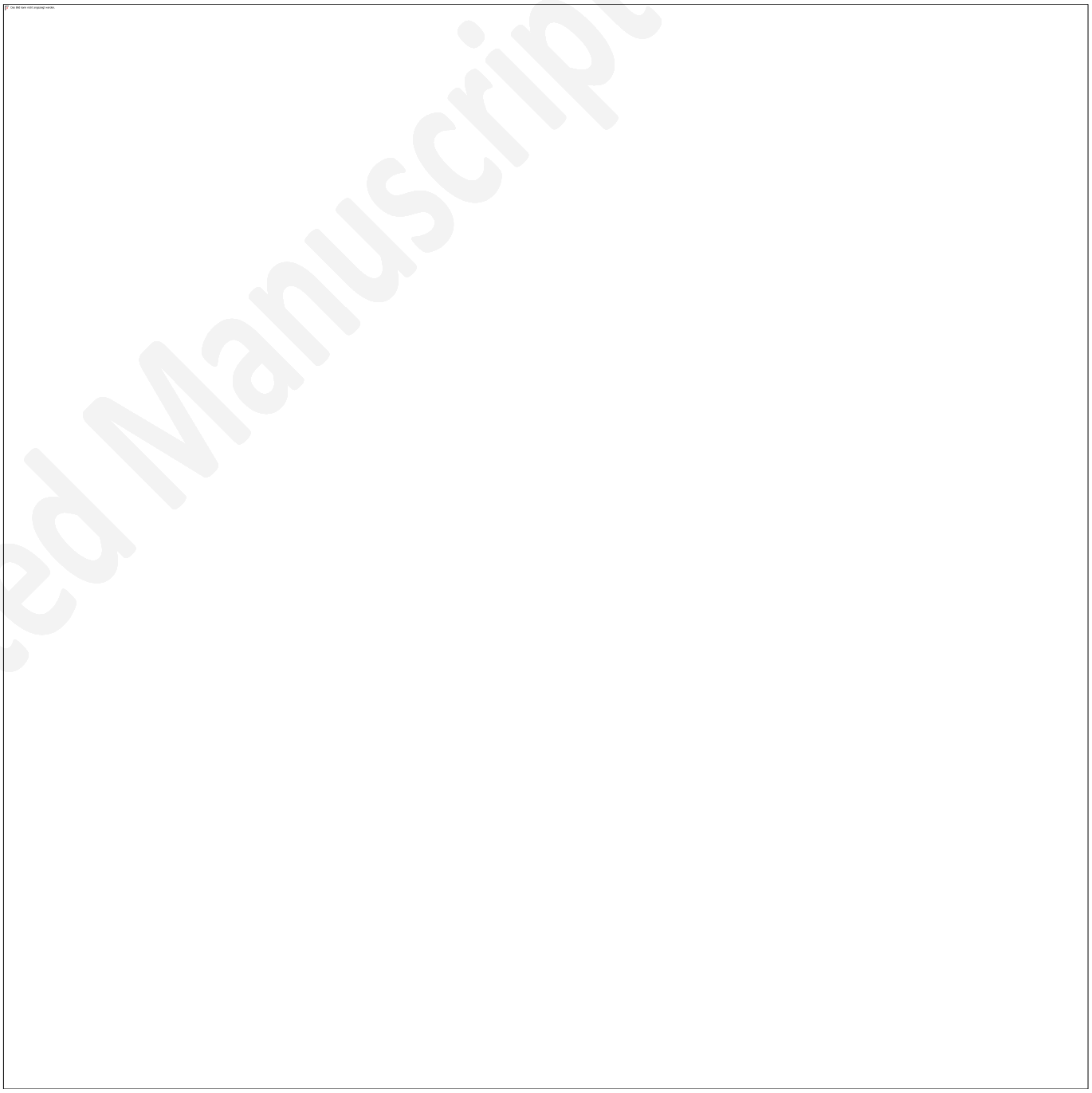
Accepted Manuscript



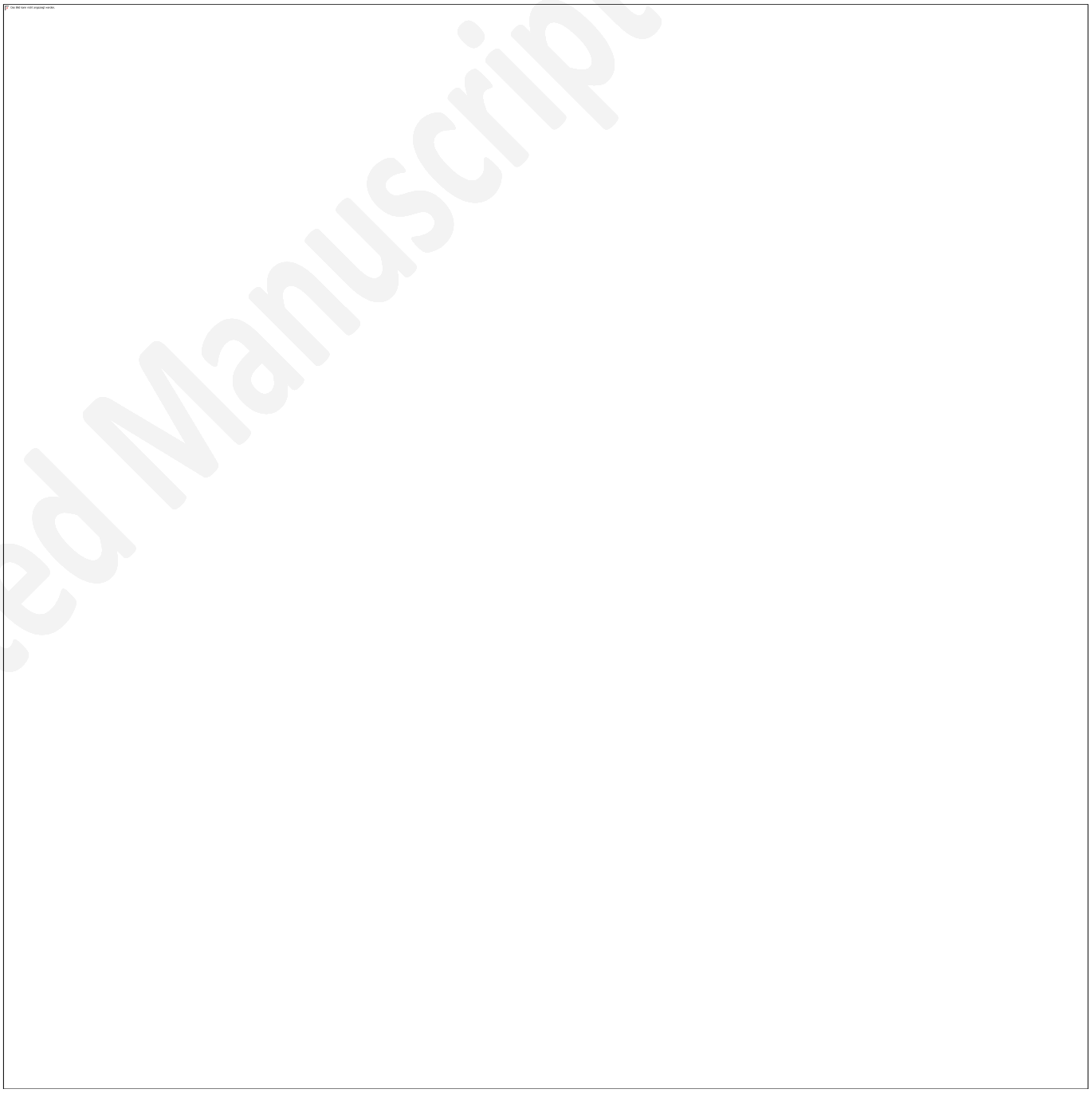
cepted Manuscript



cepted Manuscript



cepted Manuscript



cepted Manuscript

Annotated Central subfield thickness (CST)

		False negative		*True positive*			*False positive*		*True negative*		
Fluid type		mean CST (SD)	n	mean CST (SD)	n	p-value	mean CST (SD)	n	mean CST (SD)	n	p-value
IRF	AMD	323.3µm (135.1)	180	357.7µm (108.9)	930	0.0048	358.8µm (121.3)	295	311.1µm (95.8)	2205	<0.0001
	DME	281.4µm (42.1)	736	354.9µm (112.6)	9503	<0.0001	278.9µm (36.5)	60	276.0µm (35.3)	1200	0.56
SRF	AMD	442.9µm (257.6)	101	413.1µm (149.3)	1285	0.25	309.3µm (83.9)	189	299.3µm (73.9)	2398	0.11
	DME	525.7µm (183.4)	125	502.6µm (162.7)	855	0.18	351.3µm (124.5)	170	325.9µm (84.9)	10351	0.0088

Table header: Table 1. Relation between central subfield thickness and classification

Table 1. AMD = age-related macular degeneration; DME = diabetic macular edema; IRF = intraretinal fluid; SRF = subretinal fluid. The “n” value indicates the number of each OCT volume scan per disease/group. We collected the mean of central subfield thickness of the false negative, true positive, false positive and true negative OCT volume scans. The corresponding standard deviations are specified in parentheses. The “p“ value represents the statistical significance of the mean differences between the false negative and true positive as well as false positive and true negative. We used a two-sided Welch's unequal variances t-test with a level of significance alpha of 0.05.

ted Manuscript

Table header: Table 2. Single fluid detector vs. combined fluid detector

		Threshold	Accuracy	Specificity	Sensitivity	Precision	#OCTs correctly predicted
AMD	**IRF detector**	0.946	0.87 (0.86-0.88)	0.88 (0.87-0.89)	0.84 (0.81-0.86)	0.76 (0.73-0.78)	3142 (86.8%)
	SRF detector	0.966	0.93 (0.92-0.93)	0.93 (0.92-0.94)	0.93 (0.91-0.94)	0.87 (0.86-0.88)	3395 (93.8%)
	IRF detector + SRF detector *OCTs correctly predicted by both detectors at the same time*						**2952 (81.6%)**
	Combined detector	0.97	0.89 (0.88-0.90)	0.9 (0.88-0.91)	0.88 (0.87-0.90)	0.91 (0.90-0.92)	**3220 (89.0%)**
DME	**IRF detector**	0.985	0.93 (0.92-0.94)	0.95 (0.94-0.96)	0.93 (0.92-0.93)	0.99 (0.99-0.99)	10702 (93.1%)
	SRF detector	0.68	0.95 (0.94-0.96)	0.95 (0.95-0.96)	0.95 (0.94-0.96)	0.65 (0.63-0.67)	10959 (95.3%)
	IRF detector + SRF detector *OCTs correctly predicted by both detectors at the same time*						**10189 (88.6%)**
	Combined detector	0.985	0.93 (0.92-0.94)	0.95 (0.94-0.96)	0.93 (0.92-0.4)	0.99 (0.99-0.99)	**10703 (93.1%)**

Table 2. AMD = age-related macular degeneration; DME = diabetic macular edema; IRF = intraretinal fluid; SRF = subretinal fluid. For both cohorts, we collected the threshold for fluid detection, accuracy, specificity, sensitivity, precision and the number of OCTs correctly predicted by the single detector of IRF, the single detector of SRF, and the combination of the two single detectors, the detector of fluid, *i.e.,* IRF and SRF.