Aus der Klinik für Augenheilkunde
der Heinrich-Heine-Universität Düsseldorf
Direktor: Univ.-Prof. Dr. med. Gerd Geerling

Automatisierte Quantifizierung der kornealen Fluoreszeinfärbung bei Keratitis superficialis punctata

Dissertation

zur Erlangung des Grades eines Doktors der Medizin
der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Rashid Joseph Kourukmas

2024

Als Inauguraldissertation gedruckt mit Genehmigung der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

gez.:

Dekan: Prof. Dr. med. Nikolaj Klöcker

Erstgutachter: Univ.-Prof. Dr. med. Gerd Geerling

Zweitgutachter: Univ.-Prof. Dr. med. Orhan Aktas

Widmung

Ich widme diese Arbeit meiner Familie.

I Zusammenfassung deutsch

Die Beurteilung von punktförmigen Hornhautepithelläsionen, Keratitis superficialis punctata (KSP), durch Anfärbung der Augenoberfläche mit fluoreszein-haltigen Augentropfen gehört zu den häufigsten Untersuchungsmethoden in der Augenheilkunde - insbesondere bei der Diagnostik des trockenen Auges und hier insbesondere als primärer Endpunkt vieler klinischer Studien. Durch Fluoreszein werden die kleinen punktförmigen Epithelläsionen unter blauem Licht an der Spaltlampe erkennbar. Da es sich bei der Beurteilung des Schweregrades der KSP um den subjektiven Eindruck des Untersuchers handelt, wurden zwecks Standardisierung verschiedene Klassifikationen entwickelt.

Ziel der hier vorliegenden Arbeit war es, eine bildgestützte, software-assistierte Methode zur Analyse der Fluoreszeinfärbung von Epitheldefekten der Hornhaut zu entwickeln und diese neue Methode mit menschlichen Untersuchern zu vergleichen. Dafür wurden alle roten und blauen Anteile des Bildes eliminiert und der grüne Farbkanal extrahiert. Sättigung und Kontrast wurden entsprechend geläufiger Methoden zur wissenschaftlichen Bildbearbeitung angepasst. Anschließend wurden nur Bildpartikel mit einer spezifischen Größe und Rundheit analysiert und gezählt. Hierzu wurde die Bildanalysesoftware ImageJ eingesetzt und ein Algorithmus entwickelt, über den alle Schritte der Bearbeitung und Analyse in Reihe geschaltet sind und automatisiert ablaufen.

Zunächst wurden von 50 Augen von 50 verschiedenen Patienten mit unterschiedlich ausgeprägter KSP standardisierte Fotoaufnahmen angefertigt. Dazu wurden 2 µl Fluoreszein 2% mittels Eppendorfpipette in den unteren Bindehautsack getropft und nach einer Einwirkzeit von 30 Sekunden mit einer Fotospaltlampe unter Verwendung eines Cobaltblaufilters und vorgeschaltetem Gelbsperrfilter Bilder in 10-facher Vergrößerung aufgenommen. Zwanzig Bilder dienten als Trainings-Set zur Erstellung des Algorithmus. Die anderen 30 Bilder wurden als Test-Set durch den neu erstellten Algorithmus bewertet und zum Vergleich mit menschlichen Untersuchern in einen Online-Fragebogen eingepflegt. Die 30 Test-Bilder wurden von 22 Personen, davon 9 Fachärzte für Augenheilkunde, 11 Assistenzärzte und zwei PJ-Studenten nach dem Oxford-Schema (Grad 0-5) befundet. Nach 6-8 Wochen erfolgte ein zweiter Durchlauf, bei dem dieselben Bilder einmal als Original und einmal horizontal gespiegelt bewertet werden mussten.

Die software-assistierte Analyse mittels des neu erstellten Algorithmus zeigte eine starke signifikante ($p < 0.01$) lineare Korrelation der gezählten fluoreszeingefärbten Epithelläsionen mit dem durch den menschlichen Untersucher am häufigsten gewählten Oxford-Grad (Sr=0,91). Die menschlichen Untersucher zeigten nur eine moderate Intrarater-Reliabilität (K=0,426). Die höchste Interrater-Reliabilität erlangte mit 75,6% der erfahrenste Untersucher mit 29 Jahren klinischer Erfahrung. Die größte Differenz zu den anderen Untersuchern erreichte ein Weiterbildungsassistent mit 2 Jahren klinischer Erfahrung mit nur 25,6% Übereinstimmung. Es konnte gezeigt werden, dass eine software-assistierte automatisierte Quantifizierung der KSP mittels des neu erstellten Algorithmus gute Ergebnisse liefert und für den klinischen Einsatz geeignet ist.

II Zusammenfassung englisch

Evaluation of superficial punctate Keratopathy (SPK) is one of the most common tests in ophthalmology, especially in evaluation of dry eye disease, and a primary endpoint in many clinical trials. Fluorescein eye drops are instilled into the lower fornix to stain the ocular surface and make epithelial lesions visible under illumination with blue light. Since evaluation of SPK is very subjective and can differ from examiner to examiner, numerous classifications have been designed to make evaluation of SPK more consistent.

The aim of this study was to develop a software-assisted method with scientific image processing and analyses for automated evaluation of SPK and compare this new method with human graders of different levels of experience in ophthalmology. The images underwent elimination of the red and blue channels to isolate the green channel. Saturation and contrast were enhanced using common image processing techniques. Then, all particles of certain size and circularity have been counted. Therefore, the scientific image processing/analysis software ImageJ was used and an algorithm was written to run all the steps automatically.

Corneal slit-lamp images of 50 eyes from 50 patients with different severity of SPK were taken under standardized conditions. 2µl of fluorescein 2% were instilled into the lower fornix with an Eppendorf pipette to stain the ocular surface. 30 seconds later corneal images were taken in 10x magnification with a Haag-Streit photo-slitlamp using a yellow-filter. The algorithm was designed using a training set of 20 images. The other 30 images were used as test-set and were evaluated through both, the new software-assisted method and 22 human graders, including 9 specialists in ophthalmology, 11 residents and two students, using the Oxford-Scheme (grade 0-5). After 6-8 weeks the same images were evaluated again by the same graders. All 30 images were mixed into the questionnaire as original image and also horizontally mirrored, to evaluate intrarater-agreement.

The software-assisted analysis using the new algorithm showed strong significant ($p < 0.01$) linear correlation between the counted epithelial lesions to the Oxford-grade which was most frequently chosen by the human graders for each image ($Sr=0.91$). Human graders showed only moderate intrarater-agreement ($K=0.426$). The highest interrater-agreement was seen with 75.6% in the most experienced grader with 29 years of clinical experience. The lowest agreement with the other graders was seen in a resident with 2 years of clinical experience with only 25.6%. Our study showed that automated quantification of SPK is possible and achieves good results. Furthermore, we found high inconsistency in human evaluation of SPK not only regarding interrater-agreement but also intrarater-agreement. Therefore, we think a software-assisted method would be superior for evaluation of SPK, especially for follow-up examinations and in clinical trials.

III Abkürzungsverzeichnis

KSP	Keratitis superficialis punctata
SPK	Superficial punctate keratopathy
MDD	Meibomdrüsendysfunktion
BUT	Break-Up-Time
NIBUT	Nicht-invasive Break-Up-Time
IPL	Intense pulsed light
RGB	Rot-Grün-Blau

IV Inhaltsverzeichnis

1 Einleitung

1.1 Anatomie und Grundlagen des Tränenfilms

Der größte Teil der Tränenflüssigkeit entstammt der Tränendrüse (Glandula lacrimalis), welche sich in der superotemporalen Orbita in der Fossa lacrimalis des Stirnbeins (Os frontale) befindet. Die Tränendrüse wird durch den M. levator palpebrae superioris in eine Pars orbitalis und eine Pars palpebralis unterteilt. Das Sekret der Tränendrüse findet über die Ausführungsgänge im temporal oberen Fornix seinen Weg zur Augenoberfläche. Durch den Lidschlag wird die Tränenflüssigkeit dann über die gesamte Augenoberfläche verteilt.

Während die wässrige Tränenschicht den Mittelteil der Tränenflüssigkeit bildet, wird die darunter liegende Muzinphase wesentlich von den Becherzellen der Bindehaut und den Epithelzellen selbst gebildet. Die Becherzellen liegen verteilt über die gesamte Bindehaut – sowohl in der bulbären, als auch der palpebralen Bindehaut, vermehrt jedoch ihrem medialen Anteil.

Den obersten Teil des Tränenfilms bildet die Lipidschicht. Diese entstammt den Meibom-Drüsen, welche sich im Ober- und Unterlid befinden. Ihre Ausführungsgänge münden an den Lidkanten und geben ein öliges Sekret an die Augenoberfläche ab. Durch die hydrophoben Lipide bildet sich eine Mikrometer-feine Lipidschicht als oberster Teil des Tränenfilms und schützt so den wässrigen Anteil vor Verdunstung.[1] Durch ihre hydrophoben Eigenschaften können die Lipide die Oberflächenspannung reduzieren, wodurch der Tränenfilm stabiler wird. Trotz des geringen Beitrags der Lipidschicht zum Gesamtvolumen, ist diese doch von essentieller Bedeutung für die Funktion des Tränenfilms.

Die Ableitung der Tränenflüssigkeit erfolgt über die beiden Tränenpünktchen im nasalen Lidwinkel. Während der Großteil (ca. 80%) der Tränenflüssigkeit über das untere Tränenpünktchen drainiert wird, laufen über das obere nur etwa 20%. Die Canaliculi lacrimales superior et inferior münden dann in einen gemeinsamen Canaliculus lacrimalis communis, welcher in den Saccus lacrimalis, den Tränensack, mündet. Von dort aus führt der Ductus nasolacrimalis bis in den unteren Nasengang, den Meatus nasi inferior.

1.2 Das trockene Auge

Der Formenkreis des „trockenen Auges“ zählt zu den häufigsten Problemen in der Augenheilkunde und umfasst unterschiedliche Ätiologien.[2, 3] Die Prävalenz liegt zwischen 5% - 50%, wobei Frauen häufiger betroffen sind als Männer. [4] Definiert ist das trockene Auge als eine multifaktorielle Erkrankung der Augenoberfläche, bei der es zu einer Störung der Homöostase des Tränenfilms kommt, welche von okulären Symptomen begleitet wird. Dabei können Tränenfilminstabilität, Hyperosmolarität und Inflammation, sowie eine gestörte Neurotrophie eine Rolle spielen. Es wird ferner zwischen einer hyposekretorischen und hyperevaporativen Form sowie Mischformen unterschieden.[5, 6] Nicht selten sind die Betroffenen wesentlich in ihrer Lebensqualität eingeschränkt. [7]

1.2.1 Hyposekretorisches Sicca-Syndrom

Beim hyposekretorischen Sicca-Syndrom besteht ein quantitativer Tränenmangel.[5] Es wird also zu wenig Tränenflüssigkeit von der Tränendrüse produziert. Gründe für einen quantitativen Tränenmangel können zum Beispiel rheumatologische Erkrankungen wie das Sjögren-Syndrom, chronisch entzündliche Darmerkrankungen, eine Graft-Versus-Host Erkrankung im Rahmen einer Stammzelltransplantation oder eine Schilddrüsenfehlfunktion sein.[5, 8–10] Zudem gibt es Pathologien an der Tränendrüse selbst, welche degenerativ fibrotisch, traumatisch, iatrogen oder tumorbedingt sein können. Eine angeborene Hypoplasie oder Aplasie der Tränendrüse sind seltene Differentialdiagnosen.[11]

1.2.2 Hyperevaporatives Trockenes Auge

Bei der hyperevaporativen Form des Sicca-Syndroms wird zwar eine ausreichende Tränenmenge produziert, die Tränenfilmstabilität ist jedoch reduziert.[5] Die Meibomdrüsen in Ober- und Unterlid produzieren die Lipidschicht, welche die Oberflächenspannung des Tränenfilms reduziert und diesen somit vor Verdunstung schützt. Eine unzureichende Lipidschicht kann häufig auf eine Meibomdrüsendysfunktion (MDD) zurückgeführt werden.

1.2.3 Sonstige Ursachen

Andere Erkrankungen wie z. B. Allergien, neurodegenerative Veränderungen oder Lidstörung können ein dem trockenen Auge sehr ähnliches klinisches Bild auslösen. Dazu zählt z. B. die neurotrophe Keratopathie, bei der eine Störung der sensiblen Innervation der Hornhaut zu einer Epithelstörung der Hornhaut führt. Pathologien der Lider können ebenfalls ursächlich für ein trockenes Auge sein. Beispielsweise führt eine horizontale Liderschlaffung zu einer Verschlechterung der Benetzungssituation der Augenoberfläche.[12] Konzentriertes Arbeiten am Bildschirm führt zu einer reduzierten Lidschlagfrequenz, was wiederum die Verteilung des Tränenfilm negativ beeinflusst. [13]

Luftfeuchtigkeit und Temperatur wirken sich ebenfalls auf die Benetzung der Augenoberfläche aus. Oft beschreiben Patienten stärkere Symptome in den Wintermonaten. Der Mangel an frischer Luft sowie die trockene Heizungsluft verstärken die Symptome. Auch ein Zusammenhang des trockenen Auges mit Feinstaubbelastung konnte in Tierversuchen nachgewiesen werden. [14]

1.3 Diagnostik des trockenen Auges

Die Diversität im Formenkreis des trockenen Auges spiegelt sich auch in den Untersuchungsmethoden wider. Im Folgenden sind die wesentlichen Untersuchungsmethoden aufgeführt.

1.3.1 Schirmer-Test

Der Schirmer-Test quantifiziert die Tränensekretion und wurde erstmals 1903 von Otto Schirmer beschrieben.[15] Dabei wird für fünf Minuten ein Streifen Filterpapier von fünf mm Breite und 35 mm Länge in den Bindehautsack des Unterlides eingehängt. Das Filterpapier nimmt die Tränenflüssigkeit auf. Über eine Millimeterskala kann der Untersucher ablesen, welche Strecke die Tränenflüssigkeit auf dem Papierstreifen zurückgelegt hat, was wiederum mit der Menge der Tränensekretion korreliert. [16] Ein einheitlicher Wert für die physiologische Tränensekretion ist nicht festgelegt. In der Literatur wird angegeben, dass eine reduzierte Tränensekretion vorliegt, wenn die Tränenflüssigkeit die 10 mm-Markierung nicht erreicht.[17] Eine reduzierte Tränensekretion von ≤ 5 mm zählt außerdem zu den okulären Diagnosekriterien für das primäre Sjögren-Syndrom. [6, 18] Neben dem Schirmer-Test ohne Lokalanästhesie,

welcher die Reizsekretion testet, kann die basale Tränensekretion mit dem Schirmer-Test mit Lokalanästhesie durch vorherige Applikation eines betäubenden Augentropfens getestet werden.[6, 19]

1.3.2 Untersuchungen der Meibomdrüsen und der Lipidschicht

Bei der Meibomdrüsenexpression übt der Untersucher Druck auf die obere oder untere Lidkante aus und beobachtet dabei die Anzahl an exprimierbaren Meibomdrüsen sowie die Menge und Qualität des Meibomdrüsensekrets. Physiologischerweise entleert sich bereits auf leichten Druck das ölige Sekret. Liegt eine MDD vor, so kann das Sekret dickflüssig bis zahnpastaartig verändert sein und zur Obstruktion der Drüse führen.

Die Meibomdrüsen können an den ektropionierten Ober- und Unterlidern mit einer Infrarotkamera dargestellt werden.[20, 21] Hierdurch kann ein degenerativer Verlust der Meibomdrüsen erkannt und quantifiziert werden.[22]

Die Lipidinterferometrie misst die Lipidschichtdicke des Tränenfilms über intermittierende Lichtimpulse. [23, 24] Ist die Lipidschichtdicke reduziert, weist dies auf eine mangelnde Meibomdrüsensekretion hin.

1.3.3 Färbung der Augenoberfläche

Um Veränderungen der Augenoberfläche für den Untersucher sichtbar zu machen, können verschiedene Farbstoffe eingesetzt werden. Am häufigsten und geläufigsten ist die Färbung mittels Fluoreszein.[25] Dabei können Zellschäden des kornealen Epithels sichtbar gemacht werden.[26] Hierfür färbt der Untersucher den Tränenfilm über Augentropfen oder Färbestreifen an. Während intaktes, gesundes Hornhautepithel kein Fluoreszein aufnimmt, färben sich Schäden im Hornhautepithel an. Sowohl deepithelialisierte Areale (zum Beispiel nach superfizieller Keratektomie) als auch noch intakte, apoptotische Epithelzellen nehmen Fluoreszein auf.[27] Durch Beleuchtung unter Verwendung eines Cobalt-Blaufilters fluoresziert der Farbstoff und die angefärbten Areale werden dadurch erkennbar. Liegt ein punktueller Zellschaden am kornealen Epithel vor, so imponiert dieser als superfizielle punktförmige Keratopathie oder Keratitis superficialis punctata (KSP), was auch als Stippung bezeichnet wird. Da es sich teils um mehrere hunderte punktförmige Areale handelt, welche der Untersucher nicht einzeln zählen kann, wurden verschiedene Bewertungssysteme entwickelt, um die Beurteilung der KSP zu vereinfachen und zu vereinheitlichen. Insgesamt gibt es über 40 Bewertungssysteme oder „Scores“ für die Beurteilung der Augenoberfläche, von denen 18 für die Beurteilung von kornealer und/oder konjunktivaler Stippung entwickelt wurden. [28] Zu den Geläufigsten gehört das Oxford-Schema für korneale und konjunktivale Stippung (Abb. 1).[29] Es basiert auf Piktogrammen, die dem Untersucher im Vergleich mit dem klinischen Fluoreszein-Befund helfen sollen, die Ausprägung der

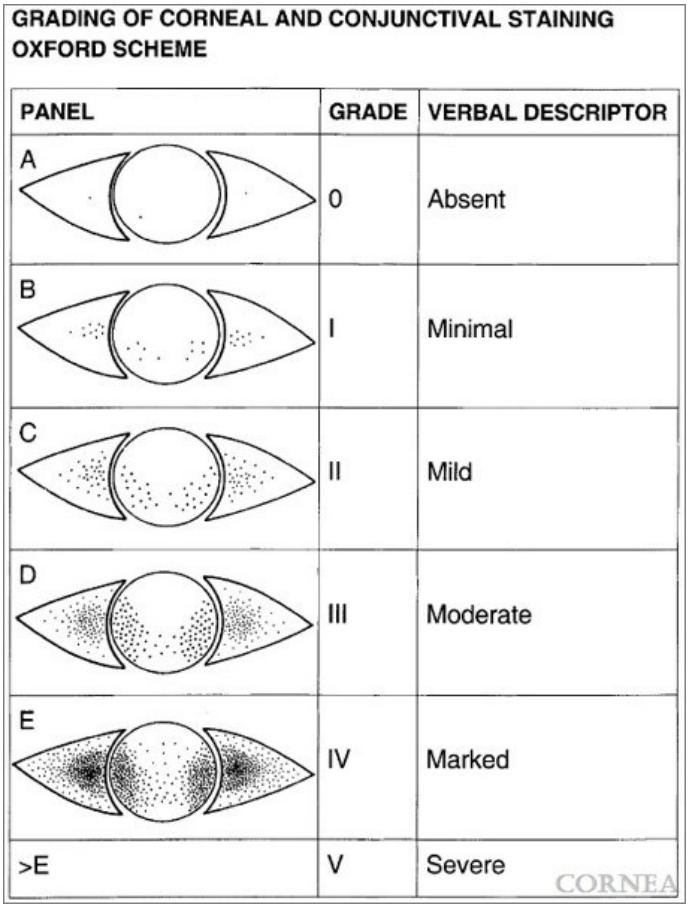

Abb. 1 Oxford-Schema für korneale und konjunktivale Stippung, Bron A. J. et al. Grading Of Corneal and Conjunctival Staining in the Context of Other Dry Eye Tests. Cornea 22(7):p 640-650, October 2003. ©Wolters Kluwer Health, Inc.

Stippung einer Gradeinteilung von 0 bis 5 zuzuordnen. Das Schema wurde so entwickelt, dass die Zahl der Stippung mit jedem Grad logarithmisch steigt. Da hierbei stets der subjektive Eindruck des Untersuchers festgehalten wird, sind derartige Bewertungssysteme anfällig für Fehler.

Die Beurteilung von KSP ist nicht nur in der klinischen Routine ein wichtiger Parameter, sondern wird auch regelmäßig als Endpunkt in klinischen Studien, z. B. in Zulassungsstudien für neue Medikamente eingesetzt.[30] Bei der Verordnung von entzündungshemmenden Ciclosporin-A-Augentropfen ist die korneale Fluoreszeinfärbung außerdem entscheidend für die Indikationsstellung und Befundkontrolle.

Neben Fluoreszein kann auch Lissamingrün und Bengalrosa zur Färbung der Augenoberfläche eingesetzt werden. Lissamingrün stellt insbesondere unter Beleuchtung mittels Rotsperrfilter Zellschäden an der Bindehaut dar und wird in der Regel von den Patienten gut vertragen.[31] Bengalrosa färbt nicht nur generell geschädigte Zellen an, sondern reichert sich insbesondere in den Nuclei der Zellen an, welche nicht durch Komponenten des Tränenfilms wie Albumin oder Mucin geschützt sind.[32] Anders als Fluoreszein und Lissamingrün führt die Färbung mittels Bengalrosa bei vielen Patienten zu vorübergehenden Schmerzen.[33] Diese Unverträglichkeit führte dazu, dass Bengalrosa im Vergleich zu den beiden anderen Farbstoffen heute deutlich seltener und eher bei gezielten Fragestellungen zum Einsatz kommt.

1.3.4 Break-Up-Time

Die Break-Up-Time (BUT) bezeichnet die Zeit in Sekunden, bis der Tränenfilm aufreißt. Hieran kann die Stabilität des Tränenfilms beurteilt werden. Zur besseren Darstellung wird der Tränenfilm in der Regel mit Fluoreszein angefärbt. Der Untersucher misst dann die Zeit, bis ein Aufriss des Tränenfilms an der Spaltlampe erkennbar wird. In dieser Zeit darf kein Lidschlag erfolgen. Reißt der Tränenfilm verfrüht auf, so besteht eine reduzierte BUT, was für eine reduzierte Tränenfilmstabilität spricht. Für gesunde Augen wurde eine BUT von über 10 Sekunden beschrieben.[34] Bei Anfärbung des Tränenfilms mittels Fluoreszein wird dieser Test auch als „invasive“ Break-Up-Time bezeichnet. Zusätzlich zum Einfluss von Fluoreszein auf die Tränenfilmstabilität ist die Konzentration von Fluoreszein bei diesem Verfahren nicht standardisiert, was zu Messungenauigkeiten führen kann. Im Gegensatz dazu wird bei der nicht-invasiven Break-Up-Time (NIBUT)

auf die Anfärbung mittels Fluoreszein verzichtet.[35] Mehrere Hersteller haben Kamerasysteme entwickelt, welche die NIBUT über Placido-basierte Hornhauttopographie oder Interferometrie messen können.[36]

1.3.5 Sonstige diagnostische Methoden

Neben den bereits erwähnten stehen eine Vielzahl weiterer diagnostischer Methoden zur Verfügung wie beispielsweise die Messung des Tränenmeniskus, die Messung der Tränenfilmosmolarität oder die automatisierte Beurteilung des bulbären Rötungsgrades.[37–39] Zusätzlich werden häufig standardisierte Fragebögen zur besseren Beurteilung der Symptome eingesetzt.[40]

1.4 Behandlung des trockenen Auges

Ebenso komplex wie die Pathologie des trockenen Auges ist auch die Therapie. Im folgenden Abschnitt ist deshalb nur ein Teil der Therapiemöglichkeiten aufgeführt.

1.4.1 Tränenersatzmittel

Tränenersatzmittel bestehen in der Regel aus Wasser und z. B. Hyaluronsäure. Sie sollen die benetzende Funktion des Tränenfilms ersetzen.

Augensalben mit Dexpanthenol oder Vitamin-A-Zusätzen werden zum Schutz der Augenoberfläche nur bei schwerem trockenem Auge oder zur Nacht eingesetzt und benetzen diese länger, sind aber sichteinschränkend.

Neben den freiverkäuflichen Tränenersatzmitteln kommen bei besonders schweren Formen des trockenen Auges mit drohendem Verlust der Sehschärfe auch autologe Serumaugentropfen zum Einsatz.[41] Die im Serum enthaltenen biologischen Faktoren ersetzen die nutritive Funktion des Tränenfilms. Bestehen etwaige Kontraindikationen wie beispielsweise Infektionserkrankungen, oder ist eine Blutabnahme bei reduziertem Allgemeinzustand oder Anämie nicht zumutbar, so werden alternativ auch Human-Albumin-Augentropfen verwendet.

1.4.2 Lidrandhygiene und Lidrandmassage

Durch Säuberung der Lidkanten und Massage der Meibomdrüsen soll einer MDD entgegengewirkt werden.[2] Dabei werden die Augenlider beispielsweise mit warmen Kompressen erwärmt, wodurch die Konsistenz des Meibomsekrets flüssiger wird.[42] Durch mechanischen Druck in Richtung der Lidkanten kann das Sekret aus der Drüse heraus und auf die Augenoberfläche gelangen und somit zur Tränenfilmstabilität beitragen.

1.4.3 Okklusion der Tränenpünktchen

Ziel der Okklusion der Tränenpünktchen ist es, den Abfluss der Tränenflüssigkeit zu verringern und somit das Volumen der Tränenflüssigkeit auf der Augenoberfläche zu erhöhen.[43] Dabei werden Kunststoffokklusive in das zu verschließende Tränenpünktchen eingesetzt. Je nach Schweregrad können ein oder auch beide Tränenpünktchen jeden Auges okkludiert werden. Neben Fremdkörpergefühl durch Reiben auf der Bindehaut kann es durch unzureichenden Abfluss der Tränenflüssigkeit auch zu Epiphora kommen.[44] Nicht selten kommt es nach einer gewissen Zeit zum Verlust der Okklusive. Bereits nach 3 Monaten sind 36% der eingesetzten Okklusive verloren.[45] Zu den chirurgischen Verfahren gehört auch die dauerhafte Okklusion der Tränenpünktchen durch Verödung.[46] Hiernach ist eine spätere Rekonstruktion kompliziert, weswegen diese Therapieoption nur in bestimmten Fällen zum Einsatz kommt.

1.4.4 Sonstige

Neben den hier abgehandelten gibt es eine Vielzahl weiterer Therapien. Dazu gehören die Meibomdrüsensondierung, die technisch assistierte Lidkantenmassage (LipiFlow®-System, Firma TearScience®, Morrisville, NC, USA), die technisch assistierte Lidrandreinigung (Blephex™, Firma BlephEx® LLC., Franklin, TN, USA) und Behandlungen mit „intense pulsed light“ (IPL, E-Eye®, Firma E-Swin, Houdan, Frankreich).[47–50] Nicht bei allen diesen Therapieansätzen ist die Wirksamkeit ausreichend wissenschaftlich belegt. In besonders schweren Fällen kann auch eine systemische Immunsuppression notwendig sein.

1.4.5 Ziele der Therapie

Ziel der oben genannten Therapieansätze ist stets, eine verbesserte Benetzung und Versorgung der Augenoberfläche zu gewährleisten. Je nachdem welche Form des trockenen Auges vorliegt, können reine Schmerzlinderung, eine Verbesserung der Sehschärfe oder – in schweren Fällen – der Erhalt des Auges das Ziel der Therapie sein. Da die zurzeit verfügbaren diagnostischen Methoden teils von der subjektiven Einschätzung des Untersuchers abhängen und andere wiederum in der Messung stark schwanken können, ist ein echter Therapieerfolg oft schwierig zu erkennen. Objektive und möglichst reproduzierbare diagnostische Methoden sind daher essentiell, um das therapeutische Ergebnis beurteilen zu können.

1.5 Automatisierte Bildanalyse

1.5.1 Grundlagen digitale Bilder

Ein digitales Bild ist aus einzelnen Bildpunkten, sogenannten Pixeln, zusammengesetzt, wobei jedes Pixel eine bestimmte Intensität wiedergibt. Während die Anzahl der Pixel für die Auflösung eines Bildes entscheidend ist, so ist die Zahl der Intensitätsstufen für den Kontrast ausschlaggebend. Mit zwei Intensitätsstufen kann man dementsprechend nur schwarze (0) und weiße (1) Pixel darstellen. Jedes Pixel verfügt demnach über eine Informationstiefe von einem Bit. Ein Bit bezeichnet die kleinste mögliche Recheneinheit. Das Wort setzt sich aus dem Englischen „binary digit“ zusammen. Solche schwarz-weiß-Bilder werden als binäre Bilder oder 1-Bit Bilder bezeichnet. Kommen nun weitere Intensitätsstufen hinzu, so erhöht sich der Informationsgehalt pro Pixel und Zwischenstufen zwischen Schwarz und Weiß (0 oder 1) werden möglich. Dadurch können auch Grautöne dargestellt werden. Die Bit-Tiefe eines Bildes entscheidet darüber, wie viele Intensitätsstufen darstellbar sind. Ein 2-Bit Bild kann dementsprechend (2^2 =) 4 verschiedene Intensitäten wiedergeben. Am geläufigsten ist heutzutage das 8-Bit Bild, bei dem (2^8 =) 256 mögliche Intensitäten zur Verfügung stehen.

1.5.2 RGB-Bilder

Die gebräuchlichste Variante von digitalen Farbbildern sind sogenannte RGB-Bilder. RGB steht hierbei für die drei Kanäle Rot, Grün und Blau, aus denen ein jedes RGB-Bild zusammengesetzt ist. Jeder Kanal für sich besteht aus einem Raster einzelner Pixel, wobei

die Intensität des Pixels die Intensität der entsprechenden Farbe an diesem definierten Bildpunkt wiedergibt. Ein roter Pixel hat dementsprechend im Rot-Kanal eine hohe Intensität, während derselbe Bildpunkt im Grün-Kanal und im Blau-Kanal eine niedrige Intensität hat. Geht man von dem geläufigen 8-Bit Format für jeden Kanal aus, so sind pro Farbkanal (Rot, Grün, Blau) 256 verschiedene Intensitätsstufen möglich. Ein RGB-Bild mit 8 Bit pro Kanal hat dementsprechend eine Farbtiefe von (3 x 8 =) 24 Bit. Dadurch sind (2^{24} =) 16 Millionen Kombinationen pro Pixel möglich.

1.5.3 Bildbearbeitung und Bilderkennung

Die Bildbearbeitung im wissenschaftlichen Sinne dient nicht etwa der Verschönerung von Bildern zu Darstellungszwecken, sondern der Vorbereitung der Daten auf eine wissenschaftliche Bildanalyse. Wird die Spanne der möglichen Intensitäten in einem Bild beispielsweise nicht voll ausgeschöpft, so können die Intensitäten weiter voneinander verteilt werden, wodurch eine bessere Abgrenzung durch die Analyseverfahren möglich wird. Ferner kann durch gezieltes Eliminieren von Bildrauschen das Bild von unbedeutenden Artefakten gereinigt werden, welche sonst in der Bildanalyse stören könnten.

Die Bilderkennung bezeichnet die Fähigkeit einer Software, bestimmte Objekte in einem Bild zu identifizieren. Hierzu werden beispielsweise Größe und Form der zu erkennenden Objekte vorgegeben.

2 Ziele der Arbeit

Ziel der Arbeit war es, eine automatisierte und objektive Quantifizierung der kornealen Fluoreszeinfärbung zu entwickeln und die Genauigkeit und Reproduzierbarkeit der Methode mit menschlichen Untersuchern zu vergleichen.

3 Publizierte Originalarbeit:

Automated vs. human evaluation of corneal staining,
R. Kourukmas, M. Roth, G. Geerling, Graefe's Archive for Clinical and Experimental Ophthalmology, Volume: 260, 2605–2612 (2022)

Graefe's Archive for Clinical and Experimental Ophthalmology
https://doi.org/10.1007/s00417-022-05574-0

CORNEA

Automated vs. human evaluation of corneal staining

R. Kourukmas[1] · M. Roth[1] · G. Geerling[1]

Received: 7 August 2021 / Revised: 26 December 2021 / Accepted: 21 January 2022

Abstract

Background and purpose Corneal fluorescein staining is one of the most important diagnostic tests in dry eye disease (DED). Nevertheless, the result of this examination is depending on the grader. So far, there is no method for an automated quantification of corneal staining commercially available. Aim of this study was to develop a software-assisted grading algorithm and to compare it with a group of human graders with variable clinical experience in patients with DED.

Methods Fifty images of eyes stained with 2 µl of 2% fluorescein presenting different severity of superficial punctate keratopathy in patients with DED were taken under standardized conditions. An algorithm for detecting and counting superficial punctate keratitis was developed using ImageJ with a training dataset of 20 randomly picked images. Then, the test dataset of 30 images was analyzed (1) by the ImageJ algorithm and (2) by 22 graders, all ophthalmologists with different levels of experience. All graders evaluated the images using the Oxford grading scheme for corneal staining at baseline and after 6–8 weeks. Intrarater agreement was also evaluated by adding a mirrored version of all original images into the set of images during the 2nd grading.

Results The count of particles detected by the algorithm correlated significantly ($n = 30$; $p < 0.01$) with the estimated true Oxford grade (Sr = 0,91). Overall human graders showed only moderate intrarater agreement ($K = 0{,}426$), while software-assisted grading was always the same ($K = 1{,}0$). Little difference was found between specialists and non-specialists in terms of intrarater agreement ($K = 0{,}436$ specialists; $K = 0{,}417$ non-specialists). The highest interrater agreement was seen with 75,6% in the most experienced grader, a cornea specialist with 29 years of experience, and the lowest was seen in a resident with 25,6% who had only 2 years of experience.

Conclusion The variance in human grading of corneal staining - if only small - is likely to have only little impact on clinical management and thus seems to be acceptable. While human graders give results sufficient for clinical application, software-assisted grading of corneal staining ensures higher consistency and thus is preferrable for re-evaluating patients, e.g., in clinical trials.

Keywords Cornea · Dry eye disease · Grading · Image analysis

Key messages

What is known

- Corneal fluorescein-staining is one of the most important diagnostic tests in dry eye disease.
- Human grading of medical images is known to be subjective.
- So far, there is no method for an automated quantification of corneal staining commercially available.

What is new

- We found only moderate intra- and interrater agreement in grading superficial punctate keratopathy.
- Experience in ophthalmology seems to have only little impact on intrarater agreement.
- Software-assisted evaluation of superficial punctate keratopathy is possible and works satisfyingly however it is not yet commercially available.

Extended author information available on the last page of the article

Published online: 31 March 2022

Introduction

Fluorescein staining is one of the most important diagnostic tests for clinical and research purposes in dry eye disease (DED) [1]. While more and more examinations are being assisted by computers (optical coherence tomography, corneal topography, wavefront analyses) in the last decades, objective methods for an automated quantification of corneal staining have been developed, but are not yet commercially available [2–5]. The aim of our study was to examine if software-assisted grading is superior to human grading in accuracy and consistency. The problem of high intra- and interrater error in human grading of medical images is a known problem in ophthalmology and other fields of medicine [6–9]. In particular human grading of corneal staining with different scores is known to be subjective and lacks reproducibility [10]. There are 41 different grading scales to evaluate the ocular surface in humans, of which 18 are for grading corneal and/or conjunctival staining [11]. The choice of the grading scale has effect on both sensitivity and consistency. While fewer steps within a grading system lead to good repeatability, they mostly lack sensitivity [12–14]. For this reason, a grading system with 0–100 steps was developed [15]. Higher numbers of possible grades on the other hand tend to produce inconsistent results and might be biased by the well-known problem that human graders tend to choose numbers that can be divided by five more often than others what again reduces the amount of steps and therefore the sensitivity [15, 16]. The Oxford scale consisting of grades from 0 to 5 is one of the most commonly used grading scales for corneal staining. Considering the diverse nature of superficial punctate keratitis, a 0–5 gradation seems relatively coarse. Therefore, an automated grading system which is not limited to a specific scale would be favorable.

Material and methods

Acquisition of corneal images

Images of 50 eyes with different grades of dry eye disease were taken under standardized conditions. Two microliter of 2% fluorescein were instilled into the lower fornix with a 2 µl Eppendorf Pipette. After 30 s, the images were acquired with a Canon camera model DS126251 attached to a Haag-Streit photo slit lamp model 900.8.2.0165 with diffuse lightning, yellow filter and 10×enlargement in a completely dark room, and saved in red–green–blue-format (RGB). Twenty of these images were used as a training set to develop the algorithm and 30 as training dataset for comparison with the human graders.

ImageJ algorithm for automated quantification of corneal staining

ImageJ, the most common software for image analysis and processing in biological research, was used for automated quantification of corneal staining [17]. All functions used for preprocessing and analyzing the images are commonly used in scientific image analysis. “Auto-threshold” was used to detect particles by separating the images into a foreground and background depending on differences of intensity. The background was eliminated, leaving the foreground with the “objects of interest” for quantification. Many different auto-threshold methods are available. Comparing the different methods using the training dataset, we found best conformity with “triangle-white” to isolate and count particles without having a large number of false positives from artifacts [18]. Next, to exclude possible artifacts like tear film or mucus, size and circularity of the “objects of interest”, i.e., positive epithelial staining, had to be specified. Following repetitive assessment using the training-dataset, particles bigger than 200 pixels or with circularity below 0,7 were eliminated. After defining those prerequisites, a macro was developed, that executed the following steps, when the cornea was marked manually as region of interest (ROI).

Preprocessing: The green channel of the RGB image was isolated and transformed into 8 bit format. With ImageJ embedded automatic contrast enhancement, the distribution of intensities became wider for better separation. Convoluted background subtraction with a radius of 14 pixels and a Gaussian blur with a sigma of 2 pixels were used to generate a so-called pseudo-background (Fig. 1) which was then subtracted from the main image to remove artifacts and background structures such as the iris, pupil, or tear film artifacts.

Analysis: Auto-threshold triangle-white technique was used to isolate particles from remaining noise. Then, a binary mask was created, showing only two intensities (1 = positive staining; 0 = no staining) (Fig. 2). Finally, the number of particles with the defined size and circularity was counted. Execution of this macro takes approximately 3–5 s per image on an average desktop computer. For the exact script of the macro, see Supplements.

Human grading of corneal staining

A cohort of 22 graders, 9 board certified ophthalmologists, 11 residents, and 2 medical students with less than 1 year of experience of the Department of Ophthalmology, University Hosptial Düsseldorf, were asked to grade the full test set. Grading was performed according to the Oxford classification for corneal staining under standardized conditions with a tablet computer (Samsung® Galaxy

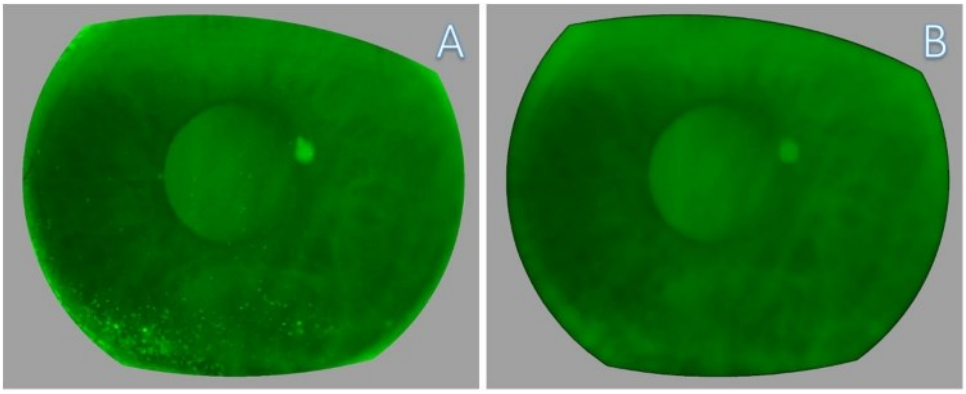

Fig. 1 **A** Original image and **B** artificial pseudo-background

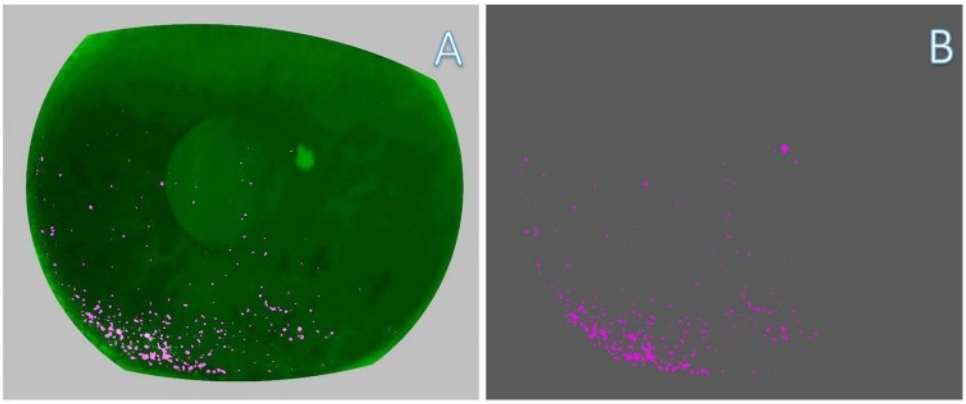

Fig. 2 **A** Detected corneal staining overlay and **B** particle mask. Brightness was adjusted for illustration

Tab S2) with a high-quality display and full brightness in a completely dark room [19].

The Oxford scheme with sample graphics was displayed to the graders throughout the entire grading process on the same screen, below the image that was to be graded. After 6–8 weeks, all participants graded the identical 30 images twice again, once as original and once mirrored horizontally, without previously being informed about this second grading and the fact that the identical images were used and had been mirrored. Software results were then compared to human grading. There was no time limit for the graders to complete the grading, but the full test set had to be graded in a single episode.

Statistics

SPSS version 27 (IBM, USA, NY, Armonk) was used for statistical analysis. Cohens-Kappa and Fleiss-Kappa were used to evaluate intra- and interrater reliability. For interpretation of *K*-values, Landis and Koch Table were used. The software-assisted evaluation (measured in number of particles) was compared to the most frequent picked Oxford grade (estimated true) using Spearman's rank correlation. A *p*-value below 0,05 was regarded as statistically significant.

Results

Interrater agreement

In the first grading episode, human interrater agreement was $K = 0{,}462$ and thus moderate between all graders. The result was the same, when all human ratings were analyzed together, i.e., all gradings from the first and second round of gradings ($K = 0{,}426$). Table 1 shows the deviation for every grader from the estimated true Oxford grade. The highest agreement with the estimated true Oxford grade was seen in the most experienced grader in 75,56% of all cases. Deviation by more than one Oxford grade from the estimated true Oxford grade was seen in 18 of 22 graders. While there was a maximal deviation of 4 Oxford grades in one case of a non-specialist, deviation of 3 Oxford grades was seen in three participants (7 cases). Resident number 8 showed a deviation of three Oxford grades in three cases

Table 1 Distribution of deviation from the estimated true Oxford grade for every grader in percentage

Grader	Deviation of grading from estimated true Oxford scale				
	0	1	2	3	4
Specialist 1	75,56%	22,22%	1,11%	-	-
Specialist 2	73,33%	26,67%	-	-	-
Specialist 3	42,22%	48,89%	4,44%	-	-
Specialist 4	54,44%	32,22%	4,44%	-	1,11%
Specialist 5	70,00%	30,00%	-	-	-
Specialist 6	71,11%	28,89%	-	-	-
Specialist 7	50,00%	37,78%	4,44%	1,11%	-
Specialist 8	67,78%	27,78%	2,22%	-	-
Specialist 9	72,22%	23,33%	2,22%	-	-
Resident 1	64,44%	26,67%	4,44%	-	-
Resident 2	62,22%	33,33%	2,22%	-	-
Resident 3	58,89%	34,44%	3,33%	-	-
Resident 4	71,11%	26,67%	1,11%	-	-
Resident 5	61,11%	34,44%	2,22%	-	-
Resident 6	67,78%	30,00%	1,11%	-	-
Resident 7	66,67%	28,89%	2,22%	-	-
Resident 8	73,33%	26,67%	-	-	-
Resident 9	25,56%	46,67%	8,89%	3,33%	-
Resident 10	57,78%	35,56%	3,33%	-	-
Resident 11	45,56%	40,00%	5,56%	1,11%	-
Student 1	25,56%	54,44%	6,67%	2,22%	-
Student 2	73,33%	24,44%	1,11%	-	-

and selected the estimated true Oxford grade only in 25,6% of all cases, as one of the students.

Intrarater agreement

All *K*-values for every grader are listed in Table 2. Specialists and non-specialists showed "moderate" intrarater agreement, but specialists were slightly more consistent than non-specialists ($K=0{,}436$ specialists; $K=0{,}417$ non-specialists). The most experienced grader, a cornea specialist with circa 30 years of experience was most consistent in his grading with an "almost perfect" agreement of $K=0{,}831$ for grading the native and the mirrored images in the second grading episode. The lowest intrarater agreement was found in a resident with $K=0{,}155$ ("slight agreement") for re-evaluation after 6–8 weeks. Figure 3 shows the total intrarater agreement in relation to years of experience in ophthalmology.

Automated grading

Software-assisted grading was identical after 6 weeks and not affected by mirroring the images. The count of particles detected by the algorithm correlated significantly ($n=30$; $p<0.01$) with the estimated true Oxford grade ($Sr=0{,}91$) (see Fig. 4).

Discussion

Experience is a well-known denominator of grading precision in ophthalmology [20]. Although only little difference was found between specialists and non-specialists overall, Fig. 1 shows that lower intrarater agreement is seen in the less experienced graders. Nevertheless, our results show that even highly experienced graders are not as consistent as a software-assisted grading method. Highest inconsistency (low intrarater agreement) was found when pictures were regraded after 6–8 weeks ($K=0{,}461$). This temporal intrarater agreement is also known as temporal drift or grade-regrade-agreement [21]. Ebenezer et al. have worked on grading of retinopathy of prematurity and used a particular temporal drift sample of 25 images that were regraded at three different points of time and found strong variety in intrarater agreement over time ranging from 0.57 to 0.94 [22]. This variation over time in human grading especially might become a problem in study settings, where reliable data needs to be gathered.

Nichols et al. investigated repeatability of several DED parameters at two time points including only one grader and found poor to moderate intrarater agreement for corneal fluorescein staining [23].

Rasmussen et al. investigated human grading of corneal and conjunctival staining on the slit lamp in 11 physicians with van Bijstervald score (vBS) and the ocular staining score (OSS) and found moderate to good intrarater agreement with intraclass correlation coefficient (ICC) of 0.77 for the vBS and 0.74 for the OSS [10]. It should be mentioned that the study mainly focused on the comparison between vBS and OSS; thus, only a small number of individuals (20 out of total 994) were invited for a second examination, and only nine were re-evaluated by the same physician.

Beyond the limited intrarater agreement, in a real-world setting of a busy clinic with rising number of follow-up visits, a patient is likely to be examined by different individuals adding interrater error. Unlike intrarater agreement, interrater agreement is difficult to investigate because there is no certainty about the true Oxford grade of an image. Furthermore, Fleiss-Kappa measures whether the grading is identical between the two time points but do not quantify a possible deviation. While grading corneal staining referring to the Oxford scheme is a method of comparing a slit lamp image with a graphic scheme, it can be assumed that the true Oxford grade for a picture is the one that was most frequently picked. Therefore, modus was chosen for the estimated true Oxford grade in our study. Rasmussen et al. still

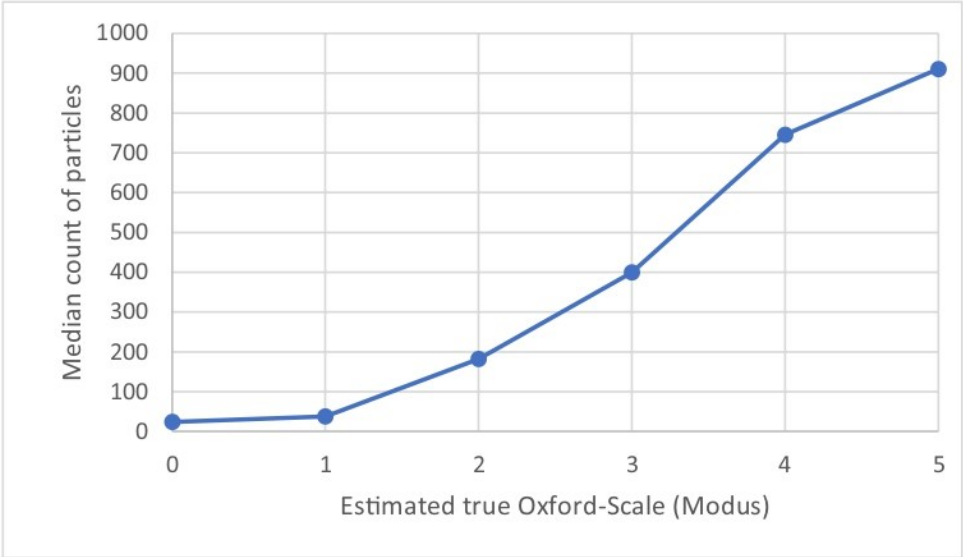

Fig. 4 Median count of particles detected by software in dependence to the estimated true Oxford scale (modus) for all images

was not calculated for our cohort, because it is more suitable to assess agreement between two, but not multiple graders. Amparo et al. tested interrater agreement in four clinicians grading 61 images using the National Eye Institute/Industry (NEI) grading scale and gathered ICC of only 0.65, what is considered moderate interrater agreement [25]. For better comparison, we calculated ICC from our cohort and gathered 0.994 what is considered excellent reliability. We want to put this value into perspective as we found only moderate agreement with Fleiss-Kappa and high deviation of 3 or 4 Oxford grades could be found between graders for both, specialists, and non-specialists. This difference can be explained by the fact that Fleiss-Kappa, as mentioned earlier, only measures whether the same grade has been chosen and ICC also respects the level of disagreement between two Oxford grades. Therefore, we think neither ICC nor Fleiss-Kappa solemnly can represent the true agreement between graders, as it is necessary for clinical practice or study settings.

While time between two gradings can especially influence the intrarater-results, there are some parameters that might influence both intra- and interrater grading. First, conditions in real live grading could vary depending on possible fluctuation in light situations or use of different slit lamp settings. Second, also the patient reported symptoms or conjunctival hyperaemia might bias a human grader in his evaluation. In addition, there may be differences in grading photographed slit lamp images versus live grading [26].

In contrast to human grading, a computer-based evaluation is not bound to a specific scale but simply counts predefined affected areas. Our algorithm showed proper correlation with the Oxford grades (Sr = 0,91; $n = 30$; $p < 0.01$). The previously mentioned groups have developed similar algorithms for corneal staining and compared the results to human grading. Rodriguez et al. used an algorithm programmed with OpenCV© (Open Source Computer Vision Library) and focused on the inferior corneal staining as region of interest and used the Ora Calibra Staining Scale®, a logarithmic scale for the number of counted particles [24]. The software-based grading was compared to human grading results. In their study, the agreement between human and software-assisted grading was high ($R = 0.89$) [24].

Amparo et al. analyzed the complete corneal area similar to us but used the National Eye Institute/Industry (NEI) grading scale and compared the results of human grading with those of an algorithm programmed in ImageJ [25]. They reported a significant correlation between their software-assisted method and human grading ($R = 0.72$) [25].

Chun et al. compared the grading of two independent clinicians using the Oxford scheme and the National Eye Institute/Industry (NEI)-recommended guidelines to a software-assisted method programmed in Microsoft Visual C++ and Open CV©. They achieved high correlation between the software-based grading and both human grading scores (Oxford scheme: $R = 0.85$; NEI: $R = 0.903$) [27].

While the above-mentioned groups have achieved similar results to our cohort, the main difference and novelty in our study are the large number of human graders with different levels of experience using the grade-regrade method that allows the best possible comparison between them. Overall, as shown in our work and the other studies mentioned above, software-based grading achieves sufficient results with precision at least as accurate in comparison to human grading. Comparison between the different groups in case of precision of the algorithm is difficult because there is difference in the selected region of interest and the chosen grading score. Although software-assisted grading might be challenged, e.g., by confluent staining, by refining the techniques

Table 2 Intrarater agreement: kappa values for all graders

Grader	Years of Experience	Intrarater Agreement κ Total	Intrarater Agreement κ After 6 Weeks	Intrarater Agreement κ Same Session Mirrored
Clinic Director	29	0,595	0,461	0,831
Specialist 1	23	0,614	0,544	0,812
Specialist 2	18	0,455	0,635	0,612
Specialist 3	12	0,592	0,667	0,617
Specialist 4	10	0,537	0,533	0,647
Specialist 5	8	0,555	0,506	0,560
Specialist 6	8	0,472	0,471	0,578
Specialist 7	7	0,570	0,466	0,670
Specialist 8	6	0,732	0,791	0,751
Resident 1	5	0,626	0,620	0,690
Resident 2	4	0,532	0,632	0,462
Resident 3	4	0,337	0,453	0,304
Resident 4	4	0,595	0,576	0,578
Resident 5	3	0,541	0,487	0,540
Resident 6	3	0,595	0,633	0,627
Resident 7	3	0,703	0,689	0,686
Resident 8	3	0,523	0,380	0,747
Resident 9	2	0,205	0,155	0,630
Resident 10	1	0,634	0,637	0,598
Resident 11	1	0,377	0,416	0,217
Student 1	0	0,413	0,385	0,357
Student 2	0	0,508	0,460	0,613
Specialists Total		0,436		
Non-specialist Total		0,417		

Intrarater agreement Kappa values. Color-marked level of agreement from dark to bright (Landis and Koch): „slight" 0,00-0,20 marked red, „fair" 0,21-0,40 marked brown, „moderate" 0,41-0,60 marked yellow, „substantial" 0,61-0,80 marked blue, „almost perfect" 0,81-1,00 marked green

Almost Perfect	0,81-1,00
Substantial	0,61-0,80
Moderate	0,41-0,60
Fair	0,21-0,40
Slight	0,00-0,20

found significant variation between human graders, although all participants had undergone a particular training before. As every participant graded a different subject, the results are difficult to compare to our cohort study.

Rodriguez et al. found a mean concordance correlation coefficient (CCC) of 0.882 between three human graders grading 54 images in the Ora Calibra Fluorescein Staining Scale, what can be considered good reliability [24]. CCC

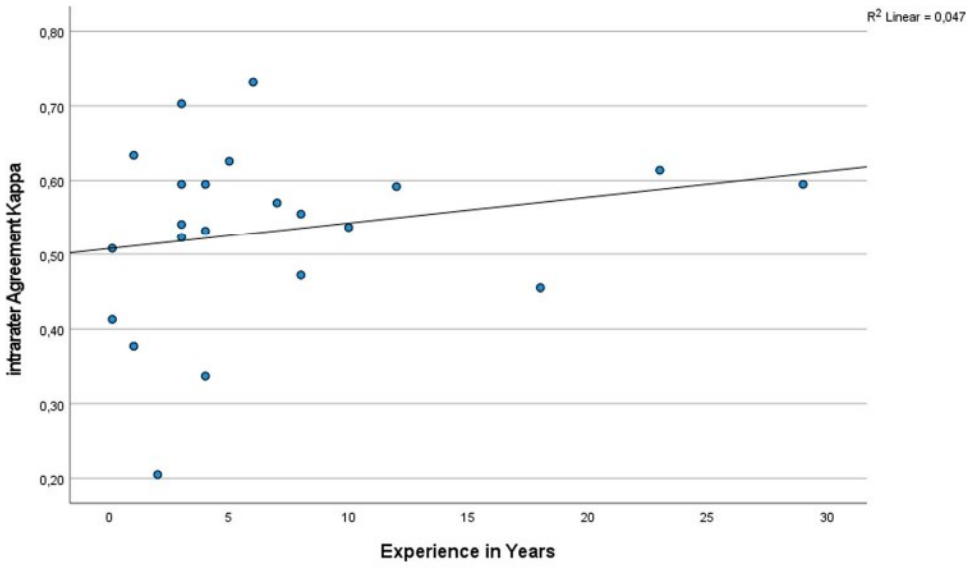

Fig. 3 Total intrarater agreement with experience in years

and algorithms, and, e.g., by application of deep learning, we think it will still be superior to a pictogram-based grading system in the future. Besides precision, another advantage of a computer-based evaluation is its consistency. Similar to the results of the grading of corneal staining, software-assisted grading of ocular redness or conjunctival lissamaine green staining has been shown to be superior compared to human evaluation [16, 28–30].

Also clinical studies would probably benefit from such a technique, because corneal fluorescein staining often is considered an important endpoint, e.g., in the SANSIK study, a multicenter phase III study for cyclosporine-A eye drops [31]. Especially in a multicenter setup with numerous graders, a more objective method would be favorable. Furthermore such a method could be used for a more exact (sub-)staging of several corneal conditions, e.g., like neurotrophic keratopathy [32]. While the benefit of a software-assisted grading system in clinical studies is obvious, it should be noted that in a clinical setting small deviation in the grading of corneal staining is acceptable and does probably rarely lead to changes in treatment. A deviation of 3–4 Oxford grades as we found in some cases though cannot be considered as negligible.

Conclusion

High inter- and intrarater bias has been seen in human grading of corneal fluoresceine staining. While accuracy of human grading may be considered sufficient, it lacks intra- and interrater consistency. Although the measured inconsistency is likely to have little impact on clinical management and outcome, an objective method would be beneficial for study settings and development of more precise staging schemes of anterior eye diseases.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00417-022-05574-0.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Not applicable.

Code availability Algorithm available in Supplements.

Declarations

Conflict of interest/Competing interests. The authors declare no competing interests.

References

1. Sullivan BD, Crews LA, Messmer EM et al (2014) Correlations between commonly used objective signs and symptoms for the diagnosis of dry eye disease: clinical implications. Acta Ophthalmol 92:161–166. https://doi.org/10.1111/aos.12012
2. Pellegrini M, Bernabei F, Moscardelli F, et al (2019) Assessment of corneal fluorescein staining in different dry eye subtypes using digital image analysis. Transl Vis Sci Technol 8: https://doi.org/10.1167/tvst.8.6.34
3. Aumann S, Donner S, Fischer J, Müller F (2019) Optical coherence tomography (OCT): principle and technical realization. In: Bille JF (ed). Cham (CH), pp 59–85
4. Fan R, Chan TC, Prakash G, Jhanji V (2018) Applications of corneal topography and tomography: a review. Clin Experiment Ophthalmol 46:133–146. https://doi.org/10.1111/ceo.13136
5. Carones F (2004) Diagnostic use of ocular wavefront sensing. Ophthalmol Clin North Am 17(129–33):v. https://doi.org/10.1016/j.ohc.2004.02.007
6. Krause J, Gulshan V, Rahimy E et al (2018) Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 125:1264–1272. https://doi.org/10.1016/j.ophtha.2018.01.034
7. Mucci B, Murray H, Downie A, Osborne K (2013) Interrater variation in scoring radiological discrepancies. Br J Radiol 86:1–5. https://doi.org/10.1259/bjr.20130245
8. Liu Y, Gadepalli K, Norouzi M, et al (2017) Detecting Cancer metastases on gigapixel pathology images. 1–13
9. Lin H, Chen H, Graham S et al (2019) Fast ScanNet: fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection. IEEE Trans Med Imaging 38:1948–1958. https://doi.org/10.1109/TMI.2019.2891305
10. Rasmussen A, Stone DU, Kaufman CE et al (2019) Reproducibility of ocular surface staining in the assessment of Sjögren syndrome–related keratoconjunctivitis sicca: implications on disease classification. ACR Open Rheumatol 1:292–302. https://doi.org/10.1002/acr2.1033
11. Eaton JS, Miller PE, Bentley E et al (2017) Slit lamp-based ocular scoring systems in toxicology and drug development: a literature survey. J Ocul Pharmacol Ther 33:707–717. https://doi.org/10.1089/jop.2017.0021
12. Bailey IL, Bullimore MA, Raasch TW, Taylor HR (1991) Clinical grading and the effects of scaling. Invest Ophthalmol Vis Sci 32:422–432
13. Sparrow NA, Frost NA, Pantelides EP, Laidlaw DA (2000) Decimalization of the oxford clinical cataract classification and grading system. Ophthalmic Epidemiol 7:49–60
14. Sook Chun Y, Park IK (2014) Reliability of 4 clinical grading systems for corneal staining. Am J Ophthalmol 157:1097–1102. https://doi.org/10.1016/j.ajo.2014.02.012
15. Woods J, Varikooty J, Fonn D, Jones LW (2018) A novel scale for describing corneal staining. Clin Ophthalmol 12:2369–2375. https://doi.org/10.2147/OPTH.S178113
16. Fieguth P, Simpson T (2002) Automated measurement of bulbar redness. Invest Ophthalmol Vis Sci 43:340–347

17. Schindelin J, Arganda-Carrera I, Frise E, et al (2009) Fiji - an open platform for biological image analysis. Nat Methods 9: https://doi.org/10.1038/nmeth.2019.Fiji
18. Zack GW, Rogers WE, Latt SA (1977) Automatic measurement of sister chromatid exchange frequency. J Histochem Cytochem Off J Histochem Soc 25:741–753. https://doi.org/10.1177/25.7.70454
19. Bron AJ, Evans VE, Smith JA (2003) Grading of corneal and conjunctival staining in the context of other dry eye tests. Cornea 22:640–650. https://doi.org/10.1097/00003226-200310000-00008
20. Andersson S, Heijl A, Bengtsson B (2011) Optic disc classification by the Heidelberg Retina Tomograph and by physicians with varying experience of glaucoma. Eye 25:1401–1407. https://doi.org/10.1038/eye.2011.172
21. Danis RP, Domalpally A, Chew EY et al (2013) Methods and reproducibility of grading optimized digital color fundus photographs in the age-related eye disease study 2 (AREDS2 Report Number 2). Investig Ophthalmol Vis Sci 54:4548–4554. https://doi.org/10.1167/iovs.13-11804
22. Daniel E, Quinn GE, Hildebrand PL et al (2015) Validated system for centralized grading of retinopathy of prematurity: telemedicine approaches to evaluating acute-phase retinopathy of prematurity (e-ROP) study. JAMA Ophthalmol 133:675–682. https://doi.org/10.1001/jamaophthalmol.2015.0460
23. Nichols KK, Mitchell GL, Zadnik K (2004) The repeatability of clinical measurements of dry eye. Cornea 23:272–285. https://doi.org/10.1097/00003226-200404000-00010
24. Rodriguez JD, Lane KJ, Ousler GW et al (2015) Automated grading system for evaluation of superficial punctate keratitis associated with dry eye. Investig Ophthalmol Vis Sci 56:2340–2347. https://doi.org/10.1167/iovs.14-15318
25. Amparo F, Wang H, Yin J, et al (2017) Evaluating corneal fluorescein staining using a novel automated method. Invest Ophthalmol Vis Sci 58:BIO168–BIO173. https://doi.org/10.1167/iovs.17-21831
26. Sorbara L, Peterson R, Schneider S, Woods C (2015) Comparison between live and photographed slit lamp grading of corneal staining. Optom Vis Sci Off Publ Am Acad Optom 92:312–317. https://doi.org/10.1097/OPX.0000000000000496
27. Chun YS, Yoon WB, Gi Kim K, Ki Park I (2014) Objective assessment of corneal staining using digital image analysis. Investig Ophthalmol Vis Sci 55:7896–7903. https://doi.org/10.1167/iovs.14-15618
28. Amparo F, Yin J, Di Zazzo A et al (2017) Evaluating changes in ocular redness using a novel automated method. Transl Vis Sci Technol 6:13. https://doi.org/10.1167/tvst.6.4.13
29. Bunya VY, Chen M, Zheng Y et al (2017) Development and evaluation of semiautomated quantification of lissamine green staining of the bulbar conjunctiva from digital images. JAMA Ophthalmol 135:1078–1085. https://doi.org/10.1001/jamaophthalmol.2017.3346
30. Peterson RC, Wolffsohn JS (2007) Sensitivity and reliability of objective image analysis compared to subjective grading of bulbar hyperaemia. Br J Ophthalmol 91:1464–1466. https://doi.org/10.1136/bjo.2006.112680
31. Leonardi A, Van Setten G, Amrane M et al (2016) Efficacy and safety of 0.1% cyclosporine A cationic emulsion in the treatment of severe dry eye disease: a multicenter randomized trial. Eur J Ophthalmol 26:287–296. https://doi.org/10.5301/ejo.5000779
32. Dua HS, Said DG, Messmer EM et al (2018) Neurotrophic keratopathy. Prog Retin Eye Res 66:107–131. https://doi.org/10.1016/j.preteyeres.2018.04.003

Authors and Affiliations

R. Kourukmas[1] · M. Roth[1] · G. Geerling[1]

✉ R. Kourukmas
rashid.kourukmas@med.uni-duesseldorf.de

[1] Department of Ophthalmology, Heinrich-Heine University Düsseldorf, Moorenstr. 5 40225, Düsseldorf, Germany

4 Diskussion

Es konnte gezeigt werden, dass eine automatisierte Quantifizierung von kornealer Fluoreszeinfärbung mittels etablierter Bildanalysetechniken möglich ist und zufriedenstellende Ergebnisse liefert. Statistisch zeigt sich nur eine moderate Intrarater- und Interraterreliabilität für korneale Fluoreszeinfärbung bei menschlichen Untersuchern. Hierbei ist zu beachten, dass ein gewisses Maß an Ungenauigkeit zu vernachlässigen ist, da im klinischen Alltag kleinere Abweichungen selten zu einer Änderung der Therapie oder des Procedere führen. Es konnten jedoch auch zwischen verschiedenen Untersuchern Abweichungen von bis zu 4 Graden auf der Oxford-Skala festgestellt werden, was den Bereich der vernachlässigbaren Variabilität überschreitet.

Bezüglich der Intraraterreliabilität ist zu betonen, dass sich insbesondere beim erneuten Bewerten eines Bildes nach 6-8 Wochen – also der Retest-Reliabilität - nur eine geringe Übereinstimmung fand (K=0,461). Diese Abweichung, bedingt durch zeitlichen Abstand zwischen zwei Messungen wird auch als „temporal drift“ bezeichnet.[51] Dieses Phänomen konnte auch bereits bei der Beurteilung von Fotoaufnahmen des hinteren Augenabschnittes in der Diagnostik der Frühgeborenenretinopathie nachgewiesen werden. Es fanden sich starke Unterschiede in der Intraraterreliabilität von 0,57 bis 0,94. [52] In der klinischen Routine ist bei ambulanten Patienten mit einem Sicca-Syndrom eine Verlaufskontrolle nach 6-8 Wochen gängig, sodass diese geringe Retest-Reliabilität menschlicher Untersucher klinisch zum Tragen kommt. Dabei ist zu bedenken, dass es sich in unserem Fragebogen lediglich um isolierte Fotoaufnahmen handelte. In der Praxis ist die korneale Fluoreszeinfärbung nur ein Teilaspekt der Diagnostik. Symptome des Patienten, bulbäre Rötung, vorherige Therapieintensivierung und viele andere Aspekte können die Einschätzung des Untersuchers beeinflussen. Wenn also ein Patient über eine deutliche Beschwerdelinderung berichtet, nachdem er die vom Augenarzt verordnete Therapie konsequent eingehalten hat, könnte dies auch zu einer milderen Einschätzung der kornealen Fluoreszeinfärbung durch den Untersucher führen.

Derartige kognitive Verzerrungen sind in der Kognitionspsychologie bekannt. Das Phänomen, dass Menschen nach Informationen suchen, die die eigene Hypothese stützen, wird als Bestätigungsfehler (englisch confirmation bias) bezeichnet und kann auch bei Ärzten in der Diagnosefindung nachgewiesenermaßen zu Fehleinschätzungen führen. [53, 54] Insbesondere in der Beurteilung von medizinischen Bildern sind Bestätigungsfehler nachgewiesen. [55] Inwieweit dieser Effekt bei der Bewertung von

KSP eine Rolle spielt, bedarf weiterer Untersuchungen. Eine automatisierte, software-assistierte Methode ist hingegen gänzlich frei von diesen Effekten.

Die nur moderate Interraterreliabilität (K=0,426) zeigt erneut die Grenzen der Beurteilung von KSP durch menschliche Untersucher. Eine ähnlich niedrige Interraterreliabilität wurde bereits für die Beurteilung der konjunktivalen Epitheliopathie mittels Lissamingrün-Färbung (K=0,3–0,36) beschrieben.[31] Die statistische Erfassung der Interraterreliabilität ist schwierig. Bei der Interpretation von Fleiss-Kappa ist zu beachten, dass nur erfasst wird, wie oft zwei Untersucher in ihrem Ergebnis übereinstimmen oder voneinander abweichen. Wie weit jedoch die Abweichung der beiden Untersucher im Einzelnen ist, also wie viele Oxford-Grade die beiden Untersucher voneinander abweichen, wird nicht erfasst. Wir haben deshalb für jedes Bild den am häufigsten gewählten Oxford-Grad als den geschätzten „wahren Oxford-Grad“ errechnet – also den statistischen Modus. Dabei zeigten sich teils Abweichungen von 3, bis sogar 4 Oxford Graden. In einer großen Klinik ist es wahrscheinlich, dass ein Patient in der Verlaufskontrolle von mehreren Personen untersucht wird. Daher sollte die nur moderate Interraterreliabilität der Fluoreszeinfärbung stets bedacht werden. Ein software-assistiertes System wäre frei von den Fehlern dieser subjektiven Einschätzung.

Ein weiterer Vorteil einer software-basierten Quantifizierung der Fluoreszeinfärbung ist, dass auch kleinere Unterschiede in der Intensität von KSP erfasst werden können, welche bei der Oxford-Skala noch innerhalb desselben Schweregrades liegen. Bei der Entwicklung eines Scoring-Systems hat die Anzahl der im System enthaltenen Grade direkten Einfluss auf Spezifität und Sensitivität.[56–58] Eine hohe Anzahl an Graden, wie zum Beispiel ein 0 – 100-System, erfasst auch sehr kleine Unterschiede. Limitiert sind diese Systeme jedoch dadurch, dass menschliche Grader sich nachgewiesenermaßen auf Grade festlegen, die ein Vielfaches von 5 sind. Damit werden von den prinzipiell 101 Graden effektiv deutlich weniger genutzt. Auch für die Beurteilung von KSP wurde ein solches Scoring-System von Woods et al. entwickelt, wobei die Autoren eben diese Problematik in ihrer Arbeit diskutieren.[59] Ein Scoring-System mit wenigen Graden hingegen, wie zum Beispiel ein 0–3-System, ist weniger sensitiv für kleinere Abweichungen, bleibt aber deutlich spezifischer. Der Vorteil einer software-assistierten Methode ist, dass alle Partikel im Bild gezählt werden und nicht unbedingt einem Score zugeteilt werden müssen, wodurch ein hohes Maß sowohl an Sensitivität als auch Spezifität erreicht werden kann. Selbstverständlich müsste bei der Entwicklung eines

solchen Systems die Messgenauigkeit reevaluiert werden, um den Bereich der natürlichen Messschwankungen zu definieren.

Das Problem der unzureichenden Intrarater- und Interraterreliabilität besteht auch bei der Befundung radiologischer und histologischer Bilder, sodass in diesen Bereichen ebenfalls bezüglich automatisierter Bildanalyse geforscht wird. [60–62]

In Notfallsituationen wie bei Polytraumata spielt der Zeitfaktor bei der Befundung radiologischer Bilder eine wichtige Rolle. Der Arbeitsgruppe Kim et al. gelang es, mit künstlicher Intelligenz einen Pneumothorax in Röntgenbildern zu erkennen.[63] Mit einer Sensitivität von 69,2% und Spezifität von 97,8% beim Erkennen des Pneumothorax war der Algorithmus den erfahrenen Radiologen zwar unterlegen, jedoch wird hier die schnelle Verfügbarkeit in Notfallsituationen für eine eventuelle automatisierte Auswertung betont. Die enorme Menge an standardisierten, weltweit verfügbaren radiologischen Bildern macht das Feld der Radiologie besonders ansprechend für die Entwicklung von künstlicher Intelligenz.[64]

Insbesondere bei der Frage nach Metastasen in exzidiertem Gewebe müssen die histologischen Schnitte mit größter Sorgfalt von den Pathologen einzeln untersucht werden. Die Arbeitsgruppe um Yun Liu konnte ein künstliches neuronales Netzwerk so anlernen, dass es bei Patienten mit Mammakarzinom 92,4% der Metastasen erkennt.[61] Dabei wird für die Befundung über den Pathologen eine Sensitivität von 73,2% angegeben. Die Arbeitsgruppe betont allerdings auch, dass das künstliche neuronale Netzwerk pro Bild 8 falsch-positive Befunde gestellt hat. Hieran zeigt sich, dass software-assistierte Analysen – zum jetzigen Zeitpunkt – nur eine Hilfestellung in der Diagnostik darstellen und nicht das Urteil eines menschlichen Untersuchers ersetzen können.

Während das im Rahmen dieser Arbeit entwickelte Bildanalysesystem bereits sehr spezifisch zwischen KSP und Artefakten unterscheiden kann, wäre eine weitere Verbesserung der Genauigkeit über die Verwendung von künstlicher Intelligenz bzw. neuronalen Netzwerken möglich und für ein marktreifes System wünschenswert. Für die Beurteilung von Fotoaufnahmen des hinteren Augenabschnittes sind solche Systeme mit künstlicher Intelligenz bereits in Entwicklung. Dabei sind jedoch oft sehr große Datenmengen erforderlich, um die gewünschte Sensitivität und Spezifität zu erreichen. Für ein von Ling Dai et al. entwickeltes Deep-Learning System namens DeepDR zur Diagnostik der diabetischen Retinopathie wurden mehr als 400.000 Bilder im Trainings-Set verwendet.[65] In der Anwendung erkennt DeepDR Veränderungen des

Augenhintergrundes, die mit einer milden nicht-proliferativen diabetischen Retinopathie einhergehen mit hoher Präzision und erzielt eine Sensitivität von 88,8% und eine Spezifität von 83,9%. Bezüglich Fotoaufnahmen der Hornhaut zeigte sich in der hier geschilderten Arbeit, dass die meisten Artefakte bei der Darstellung von KSP durch Partikel im Tränenfilm entstehen. Hier wäre die Akquise von mehreren Bildern jeweils vor und nach einem Lidschlag eine Möglichkeit, um die Spezifität zu verbessern. Bewegliche Partikel wie Sekret, Schminkreste oder Zellschrott auf der Augenoberfläche, welche sonst fälschlicherweise als KSP gewertet werden, könnten so erfolgreich herausgerechnet werden.

Verschiedene Geräte mit Fotofunktion stehen in der Augenheilkunde bereits zur Verfügung und könnten für ein entsprechendes automatisiertes Verfahren nachgerüstet werden. Da die Untersuchung an einem separaten Gerät jedoch zeitaufwändiger und im klinischen Alltag eventuell zu unpraktisch sein könnte, wäre ein in die Spaltlampe integriertes System zu favorisieren. Ein Kameramodul mit Verknüpfung zu einer entsprechenden Software am Arbeitsplatz-PC und sofortiger Analyse wäre nach aktuellem Stand der Technik umsetzbar.

5 Schlussfolgerungen

Die menschliche Bewertung von kornealer Fluoreszeinfärbung ist nicht nur untersucherabhängig, sondern variiert z. B. über die Zeit auch stark innerhalb des einzelnen Untersuchers. Es konnte gezeigt werden, dass eine automatisierte, softwaregestützte Quantifizierung der kornealen Fluoreszeinfärbung umsetzbar ist und zufriedenstellende Ergebnisse erzielt. Insbesondere im Rahmen von klinischen Studien bzw. für die verlässliche Beurteilung von klinischen Befunden im zeitlichen Verlauf wäre eine automatisierte und damit objektive Methode zur Beurteilung von Epitheldefekten der Hornhaut daher sehr sinnvoll.

6 Literatur- und Quellenverzeichnis

1. Bai Y, Nichols JJ (2017) Advances in thickness measurements and dynamic visualization of the tear film using non-invasive optical approaches. Prog Retin Eye Res 58:28–44. https://doi.org/10.1016/j.preteyeres.2017.02.002

2. Craig JP, Nichols KK, Akpek EK, et al (2017) TFOS DEWS II Definition and Classification Report. Ocular Surface 15:276–283.

3. Spaniol K (2018) Erkrankungen der Lider und Augenoberfläche. Kompass Ophthalmologie 4:176–177. https://doi.org/10.1159/000494608

4. Craig JP, Nelson JD, Azar DT, et al (2017) TFOS DEWS II Report Executive Summary. Ocul Surf 15:802–812. https://doi.org/10.1016/j.jtos.2017.08.003

5. The Definition and Classification of Dry Eye Disease: Report of the Definition and Classification Subcommittee of the International Dry Eye Workshop (2007) DEWS Definition and Classification

6. Craig JP, Nichols KK, Akpek EK, et al (2017) TFOS DEWS II Definition and Classification Report. Ocul Surf 15:276–283. https://doi.org/10.1016/j.jtos.2017.05.008

7. Le Q, Zhou X, Ge L, et al (2012) Impact of dry eye syndrome on vision-related quality of life in a non-clinic-based general population. BMC Ophthalmol 12:1. https://doi.org/10.1186/1471-2415-12-22

8. Ogawa Y, Okamoto S, Wakui M, et al (1999) Dry eye after haematopoietic stem cell transplantation. Br J Ophthalmol. 83(10):1125-1130. https://doi.org/10.1136/bjo.83.10.1125

9. Knox DL, Schachat AP, Mustonen E (1984) Primary, Secondary and Coincidental Ocular Complications of Crohn's Disease. Ophthalmology 91:163–173. https://doi.org/10.1016/S0161-6420(84)34322-6

10. Alanazi SA, Alomran AA, Abusharha A, et al (2019) An assessment of the ocular tear film in patients with thyroid disorders. Clinical Ophthalmology 13:1019–1026. https://doi.org/10.2147/OPTH.S210044

11. Chapman DB, Shashi V, Kirse DJ (2009) Case report: Aplasia of the lacrimal and major salivary glands (ALSG). Int J Pediatr Otorhinolaryngol 73:899–901. https://doi.org/10.1016/j.ijporl.2009.03.004

12. Chhadva P, McClellan AL, Alabiad CR, et al (2016) Impact of eyelid laxity on symptoms and signs of dry eye disease. Cornea 35:531–535. https://doi.org/10.1097/ICO.0000000000000786

13. Skotte JH, Nøjgaard JK, Jørgensen L v, et al (2007) Eye blink frequency during different computer tasks quantified by electrooculography. Eur J Appl Physiol 99:113–119. https://doi.org/10.1007/s00421-006-0322-6

14. Song SJ, Hyun S-W, Lee TG, et al (2020) New application for assessment of dry eye syndrome induced by particulate matter exposure. Ecotoxicology and Environmental Safety 205:111125. https://doi.org/10.1016/j.ecoenv.2020.111125

15. Schirmer, Otto. (1903) Studien zur Physiologie und Pathologie der Tränenabsonderung und Tränenabfuhr. Albrecht von Graefes Archiv für Ophthalmologie 56:197-291.

16. Holly FJ, Lamberts DW, Esquivel ED (1982) Kinetics of capillary tear flow in the Schirmer strip. Curr Eye Res 2:57–70. https://doi.org/10.3109/02713688208998380

17. Serin D, Xafak Karslıoglu S, Kıyan A, Alagöz G (2007) A Simple Approach to the Repeatability of the Schirmer Test Without Anesthesia Eyes Open or Closed? Cornea, 26(8), 903–906. https://doi.org/10.1097/ICO.0b013e3180950083.

18. Vitali C Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. Annals of the rheumatic diseases, 61(6), 554–558. https://doi.org/10.1136/ard.61.6.554

19. Jordan A, Baum J (1980) Basic Tear Flow: Does It Exist? Ophthalmology 87:920–930. https://doi.org/10.1016/S0161-6420(80)35143-9

20. Arita R, Itoh K, Inoue K, Amano S (2008) Noncontact Infrared Meibography to Document Age-Related Changes of the Meibomian Glands in a Normal Population. Ophthalmology 115:911–915. https://doi.org/10.1016/j.ophtha.2007.06.031

21. Finis D, Ackermann P, Pischel N, et al (2015) Evaluation of Meibomian Gland Dysfunction and Local Distribution of Meibomian Gland Atrophy by Non-contact Infrared Meibography. Curr Eye Res 40:982–989. https://doi.org/10.3109/02713683.2014.971929

22. Machalińska A, Zakrzewska A, Safranow K, et al (2016) Risk Factors and Symptoms of Meibomian Gland Loss in a Healthy Population. J Ophthalmol. 7526120. https://doi.org/10.1155/2016/7526120

23. DOANE MG (1989) An Instrument for In Vivo Tear Film Interferometry. Optometry and vision science: official publication of the American Academy of

Optometry, 66(6), 383–388.
https://doi.org/10.1097/00006324-198906000-00008

24. Arita R, Fukuoka S, Morishige N (2017) Functional Morphology of the Lipid Layer of the Tear Film. Cornea. 36 Suppl 1, S60–S66.
https://doi.org/10.1097/ICO.0000000000001367

25. Sullivan BD, Crews LA, Messmer EM, et al (2014) Correlations between commonly used objective signs and symptoms for the diagnosis of dry eye disease: Clinical implications. Acta Ophthalmol 92:161–166.
https://doi.org/10.1111/aos.12012

26. Korb DR, Herman JP, Finnemore VM, et al (2008) An Evaluation of the Efficacy of Fluorescein, Rose Bengal, Lissamine Green, and a New Dye Mixture for Ocular Surface Staining. Eye & contact lens, 34(1), 61–64.
https://doi.org/10.1097/ICL.0b013e31811ead93

27. Bandamwar KL, Papas EB, Garrett Q (2014) Fluorescein staining and physiological state of corneal epithelial cells. Contact Lens and Anterior Eye 37:213–223.
https://doi.org/10.1016/j.clae.2013.11.003

28. Eaton JS, Miller PE, Bentley E, et al (2017) Slit Lamp-Based Ocular Scoring Systems in Toxicology and Drug Development: A Literature Survey. Journal of Ocular Pharmacology and Therapeutics 33:707–717.
https://doi.org/10.1089/jop.2017.0021

29. Bron AJ, Evans VE, Smith JA (2003) Grading of corneal and conjunctival staining in the context of other dry eye tests. Cornea 22:640–650.
https://doi.org/10.1097/00003226-200310000-00008

30. Leonardi A, Van Setten G, Amrane M, et al (2016) Efficacy and safety of 0.1% cyclosporine A cationic emulsion in the treatment of severe dry eye disease: A multicenter randomized trial. Eur J Ophthalmol 26:287–296.
https://doi.org/10.5301/ejo.5000779

31. Hamrah P, Alipour F, Jiang S, et al (2011) Optimizing evaluation of Lissamine Green parameters for ocular surface staining. Eye 25:1429–1434.
https://doi.org/10.1038/eye.2011.184

32. Feenstra RPG, Tseng SCG (1992) Comparison of Fluorescein and Rose Bengal Staining. Ophthalmology 99:605–617.
https://doi.org/10.1016/S0161-6420(92)31947-5

33. Doughty MJ, Lee C-A, Ritchie S, Naase T (2007) An assessment of the discomfort associated with the use of rose bengal 1% eyedrops on the normal human eye: a comparison with saline 0.9% and a topical ocular anaesthetic. Ophthalmic and Physiological Optics 27:159–167.
https://doi.org/10.1111/j.1475-1313.2006.00456.x

34. Lee JH, Kee CW (1988) The significance of tear film break-up time in the diagnosis of dry eye syndrome. Korean J Ophthalmol 2:69–71. https://doi.org/10.3341/kjo.1988.2.2.69

35. Mengher LS, Pandher KS, Bron AJ (1986) Non-invasive tear film break-up time: sensitivity and specificity. Acta Ophthalmol 64:441–444. https://doi.org/10.1111/j.1755-3768.1986.tb06950.x

36. Bandlitz S, Peter B, Pflugi T, et al (2020) Agreement and repeatability of four different devices to measure non-invasive tear breakup time (NIBUT). Contact Lens and Anterior Eye 43:507–511. https://doi.org/10.1016/j.clae.2020.02.018

37. Mainstone JC, Bruce AS, Golding TR (1996) Tear meniscus measurement in the diagnosis of dry eye. Curr Eye Res 15:653–661. https://doi.org/10.3109/02713689609008906

38. Potvin R, Makari S, Rapuano CJ (2015) Tear film osmolarity and dry eye disease: A review of the literature. Clinical Ophthalmology 9:2039–2047. https://doi.org/10.2147/OPTH.S95242

39. Fieguth P, Simpson T (2002) Automated Measurement of Bulbar Redness. Investigative ophthalmology & visual science, 43(2), 340–347.

40. Schiffman RM, Christianson MD, Jacobsen G, et al (2000) Reliability and Validity of the Ocular Surface Disease Index. Archives of Ophthalmology 118:615–621. https://doi.org/10.1001/archopht.118.5.615

41. Geerling G, MacLennan S, Hartwig D (2004) Autologous serum eye drops for ocular surface disorders. British Journal of Ophthalmology 88:1467–1474.

42. Korbmacher JP, Geerling G (2021) Dry eye therapy. Spektrum der Augenheilkunde 35:177–194. https://doi.org/10.1007/s00717-021-00497-3

43. Tost FHW, Geerling G (2008) Plugs for Occlusion of the Lacrimal Drainage System. Developments in ophthalmology, 41, 193–212. https://doi.org/10.1159/000131090

44. Tai M-C, Banu Cosar C, Cohen EJ, et al (2002) The Clinical Efficacy of Silicone Punctal Plug Therapy. Cornea, 21(2), 135–139. https://doi.org/10.1097/00003226-200203000-00001

45. Balaram M, Schaumberg DA, Dana MR (2001) Efficacy and Tolerability Outcomes After Punctal Occlusion With Silicone Plugs in Dry Eye Syndrome. American journal of ophthalmology, 132(4), 600–601. https://doi.org/10.1016/s0002-9394(01)01001-7

46. Yaguchi S, Ogawa Y, Kamoi M, et al (2012) Surgical management of lacrimal punctal cauterization in chronic GVHD-related dry eye with recurrent punctal

plug extrusion. Bone Marrow Transplant 47:1465–1469. https://doi.org/10.1038/bmt.2012.50

47. Magno M, Moschowits E, Arita R, et al (2021) Intraductal meibomian gland probing and its efficacy in the treatment of meibomian gland dysfunction. Surv Ophthalmol 66:612–622.

48. Lane SS, Dubiner HB, Epstein RJ, et al (2012) A new system, the LipiFlow, for the treatment of meibomian gland dysfunction. Cornea, 31(4), 396–404. https://doi.org/10.1097/ICO.0b013e318239aaea

49. Moon SY, Han SA, Kwon HJ, et al (2021) Effects of lid debris debridement combined with meibomian gland expression on the ocular surface MMP-9 levels and clinical outcomes in moderate and severe meibomian gland dysfunction. BMC Ophthalmol 21. https://doi.org/10.1186/s12886-021-01926-2

50. Egri S, van Hollebecke I, Guindolet D, et al (2021) Efficacité de la lumière pulsée dans le traitement des sécheresses oculaires sévères par dysfonctionnement meibomien. J Fr Ophtalmol 44:169–175. https://doi.org/10.1016/j.jfo.2020.04.061

51. Danis RP, Domalpally A, Chew EY, et al (2013) Methods and reproducibility of grading optimized digital color fundus photographs in the Age-Related Eye Disease Study 2 (AREDS2 Report Number 2). Invest Ophthalmol Vis Sci 54:4548–4554. https://doi.org/10.1167/iovs.13-11804

52. Daniel E, Quinn GE, Hildebrand PL, et al (2015) Validated System for Centralized Grading of Retinopathy of Prematurity: Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) Study. JAMA Ophthalmol 133:675–682. https://doi.org/10.1001/jamaophthalmol.2015.0460

53. Elston DM (2020) Confirmation bias in medical decision-making. J Am Acad Dermatol 82:572. https://doi.org/10.1016/j.jaad.2019.06.1286

54. Sibbald M, Panisko D, Cavalcanti RB (2011) Role of clinical context in residents' physical examination diagnostic accuracy. Med Educ 45:415–421. https://doi.org/10.1111/j.1365-2923.2010.03896.x

55. Itri JN, Patel SH (2018) Heuristics and Cognitive Error in Medical Imaging. American Journal of Roentgenology 210:1097–1105. https://doi.org/10.2214/AJR.17.18907

56. Bailey IL, Bullimore MA, Raasch TW, Taylor HR (1991) Clinical grading and the effects of scaling. Invest Ophthalmol Vis Sci 32:422–432

57. Sparrow NA, Frost NA, Pantelides EP, Laidlaw DA (2000) Decimalization of The Oxford Clinical Cataract Classification and Grading System. Ophthalmic Epidemiol 7:49–60

58. Sook Chun Y, Park IK (2014) Reliability of 4 clinical grading systems for corneal staining. Am J Ophthalmol 157:1097–1102. https://doi.org/10.1016/j.ajo.2014.02.012

59. Woods J, Varikooty J, Fonn D, Jones LW (2018) A novel scale for describing corneal staining. Clin Ophthalmol 12:2369–2375. https://doi.org/10.2147/OPTH.S178113

60. Krause J, Gulshan V, Rahimy E, et al (2018) Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. Ophthalmology 125:1264–1272. https://doi.org/10.1016/j.ophtha.2018.01.034

61. Liu Y, Gadepalli K, Norouzi M, et al (2017) Detecting Cancer Metastases on Gigapixel Pathology Images. MICCAI Tutorial. 1–13 https://doi.org/10.48550/arXiv.1703.02442

62. Lin H, Chen H, Graham S, et al (2019) Fast ScanNet: Fast and Dense Analysis of Multi-Gigapixel Whole-Slide Images for Cancer Metastasis Detection. IEEE Trans Med Imaging 38:1948–1958. https://doi.org/10.1109/TMI.2019.2891305

63. Kim D, Lee JH, Kim SW, et al (2022) Quantitative Measurement of Pneumothorax Using Artificial Intelligence Management Model and Clinical Application. Diagnostics 12. https://doi.org/10.3390/diagnostics12081823

64. Hosny A, Parmar C, Quackenbush J, et al (2018) Artificial intelligence in radiology. Nat Rev Cancer 18:500–510

65. Dai L, Wu L, Li H, et al (2021) A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat Commun 12:3242. https://doi.org/10.1038/s41467-021-23458-5

7 Anhang

Macro zur Bearbeitung und Analyse der Bilder:

```
run("Delete Slice", "delete=channel");

run("Stack to Images");

rename("blue");

selectWindow("blue");

close();

rename("green");

run("8-bit");

run("Enhance Contrast...", "saturated=0.01");

run("Convoluted Background Subtraction", "convolution=Gaussian radius=14");

run("Gaussian Blur...", "sigma=2");

run("Auto Threshold", "method=Triangle white");

setOption("BlackBackground", true);

run("Convert to Mask");

run("Analyze Particles...", "size=0-200 pixel circularity=0.7-1.00 show=[Masks] display exclude summarize in_situ");
```

Danksagung

Ich danke Gott, meiner Familie und meinen Betreuern, Freunden und Kollegen, die mich bei dieser Arbeit unterstützt haben.