

ARTICLE



Automated extraction of clinical measures from videos of oculofacial disorders using machine learning: feasibility, validity and reliability

Christopher B. Schulz^{1✉}, Holly Clarke¹, Sarith Makuloluwe¹, Peter B. Thomas² and Swan Kang^{1,2}

© The Author(s), under exclusive licence to The Royal College of Ophthalmologists 2023

OBJECTIVES: To determine the feasibility, validity and reliability of automatically extracting clinically meaningful eyelid measurements from consumer-grade videos of individuals with oculofacial disorders.

METHODS: A custom computer program was designed to automatically extract clinical measures from consumer-grade videos. This program was applied to publicly available videos of individuals with oculofacial disorders, and age-matched controls. The primary outcomes were margin reflex distance 1 (MRD1) and 2 (MRD2), blink lagophthalmos, and ocular surface area exposure. Test-retest reliability was evaluated using Bland–Altman analysis to compare the agreement in obtained measures between separate videos of the same individual taken within 48 h of each other.

RESULTS: MRD1 was reduced in individuals with ptosis versus controls (2.2 mm versus 3.4 mm, $p < 0.001$), and increased in individuals with facial nerve palsy (FNP) (3.9 mm, $p = 0.049$) and thyroid eye disease (TED) (4.1 mm; $p = 0.038$). Blink lagophthalmos was increased in individuals with FNP (3.7 mm); $p < 0.001$ and those with TED (0.1 mm, $p = 0.003$) versus controls (0.0 mm). Ocular surface exposure was reduced in individuals with ptosis compared with controls (12.2 mm² versus 13.1 mm²; $p < 0.001$) and increased in TED (13.7 mm²; $p 0.002$). Bland-Altman analysis demonstrated 95% limits of agreement for video-derived measures: median MRD1: -1.1 to 1.1 mm; median MRD2: -0.9 to 1.0 mm; blink lagophthalmos: -3.5 to 3.7 mm; and average ocular surface area exposure: -1.6 to 1.6 mm².

CONCLUSIONS: The presented program is capable of taking consumer grade videos of patients with oculofacial disease and providing clinically meaningful and reliable eyelid measurements that show promising validity.

Eye (2023) 37:2810–2816; <https://doi.org/10.1038/s41433-023-02424-z>

INTRODUCTION

Digitally enabled outpatient care and telemedicine is an important cornerstone in the long-term plan of healthcare providers [1]. There has been a growing trend for the adoption of telemedicine within ophthalmology which predates the global COVID-19 pandemic. This has been driven by increasing demands for service, cost efficiency, patient convenience, and by the availability of modern ophthalmic imaging techniques. Compared with other ophthalmic subspecialties, oculoplastic and adnexal specialists have previously been slower to embrace telemedicine, but during the COVID-19 pandemic there has been a shift in practice for many [2, 3]. As a subspecialty, oculoplastics is likely to lend itself well to consumer-grade video consultation (e.g. using a smartphone or webcam). Indeed, it is estimated that almost 40% of new oculoplastic patient encounters and 60% of follow up appointments might be suitable for remote video consultation [4]. Conducted during the height of the pandemic's first wave, nationwide surveys of oculoplastic surgeons in the UK and the USA demonstrated that 86.6–88.8% of respondents were incorporating telemedicine into their routine clinical practice [5, 6]. Of those, about 85% were using a video-based platform to conduct remote consultation [6]. That said, there appear to be some important

barriers to the widespread adoption of telemedicine in oculoplastics. Two-thirds of oculoplastic consultants remain dissatisfied by the limitations of clinical examination via telemedicine [5] and only 4% of surveyed surgeons felt comfortable proceeding to surgery based on remote consultation [6]. If video-based consultations are to provide the improvements in service delivery that are hoped for, we must strive for a standard of patient assessment that is at least comparable to that which we can provide face to face. A central aspect of the clinical assessment of patients in the oculoplastic clinic is the measurement of various key eyelid parameters, allowing the clinician to evaluate disease severity, progression and response to intervention. Using recent developments in computer vision and deep learning, our aim was to determine the feasibility of obtaining clinically meaningful eyelid measurements from consumer-grade (i.e. non-professional) videos of individuals.

METHODS

Program development

A custom program was developed using Python 3 [7] and OpenCV [8]. This program, known hereafter as 'VALID' (Video Analysis of the eyelids), is

¹Portsmouth Hospitals University NHS Trust, Portsmouth, United Kingdom. ²Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom.

✉email: chrisschulz@doctors.org.uk

Received: 20 October 2022 Revised: 12 December 2022 Accepted: 23 January 2023

Published online: 1 February 2023

provided with an MPEG-4 video as input and is designed to read each consecutive frame of the inputted video. For each frame, a region of interest (ROI) is detected using previously trained and publicly available machine learning models [9, 10]. The ROI detected corresponds to the right and/or left periocular region so long as either one or both are visible in the frame. For each frame, VALID uses a custom-trained deep learning model that is used to predict whether each pixel in the ROI (left and/or right periocular region) belongs to either eyelid skin, bulbar conjunctiva or cornea. The solution for this 'image segmentation' deep learning model is based on the 'U-Net' convolutional neural network architecture [11] and was custom trained on a random sample of 7101 images from a publicly available annotated dataset of eyes [12]. The model was tested on a further 1781 images from the same dataset that had not been used for training the model. As part of this cross-validation exercise the model achieved an accuracy of 98.2%.

Among the eyelid parameters most commonly recorded in the oculoplastic clinics are the margin reflex distance 1 (MRD1) and 2 (MRD2). MRD1 and MRD2 refer to the distance from the corneal light reflex to the central upper eyelid margin and central lower lid margin, respectively. Both are important measures in the assessment of a wide range of oculofacial disease and are potentially feasible to calculate from the segmented video images. To achieve this, several computer vision techniques were applied to the segmented image of each frame. These techniques included contouring of the eyelid margin, pupil identification using circle Hough transformation, and corneal stabilisation and tracking of head position in order to account for partial or complete pupil obscuration (e.g. by blinking), eye ductions and head movement (roll, pitch and yaw), as demonstrated in Fig. 1.

VALID converts MRD1 and MRD2 from pixels to millimetres (mm), in order to aid clinical interpretation. The fixed data point used as a basis for this conversion was an assumed corneal diameter ('white-to-white') of 11.7 mm. This assumption is based on the finding that 'white-to-white' is consistently measured within 0.5 mm of 11.7 mm in adults, and with a standard deviation of less than 0.5 mm [13, 14]. Moreover, in adults, the corneal diameter measurement would not be expected to change significantly within the same individual measured at different timepoints and so is likely to be appropriate in most clinical applications e.g. before and after eyelid surgery, or to monitor disease progression in facial palsy or thyroid eye disease. As a summary measure, the median values for MRD1 and MRD2 across all frames in each video are calculated.

In addition to MRD1 and MRD2, VALID was also designed to automatically calculate blink lagophthalmos (in mm) and average ocular surface area exposure (in mm²). Neither of these indices are reliably or accurately measurable in standard clinical practice, however we believe that both may be potential sensitive measures of dynamic eyelid function. Blink lagophthalmos and ocular surface exposure may be of particular relevance in patients with dry eye syndrome, eyelid malposition, and orbicularis oculi weakness. Blink lagophthalmos was recorded by identifying each blink cycle in a video sequence and measuring the minimum interpalpebral height (combine MRD1 and MRD2) during each blink (Fig. 2). A median value for each detected blink was calculated as the final measure. Average ocular surface exposure was calculated by the summation of the total number of pixels segmented as cornea or bulbar conjunctiva and converting this to mm² according to the aforementioned conversion factor. A mean value of all frames in the video excerpt was calculated as a summary value.

Validity testing

A dataset was gathered using excerpts from videos made publicly available online. All videos were used under the 'fair use' or equivalent exception to copyright law for non-commercial research specific to the video's country of origin and in accordance with the ethical principles outlined in the Declaration of Helsinki. Videos were identified by a strategic search of an online video repository (YouTube™). All videos were downloaded and converted into MPEG-4 format. Included videos were vetted by an oculoplastic specialist for confirmation that the primary subject has a diagnosis of either acute facial nerve palsy (FNP) within 1 week of onset, thyroid eye disease (TED), or blepharoptosis. Other available information (such as other posted videos, audio content or written material available from the same subject) were used in conjunction with clinical judgement to confirm the diagnosis. Videos were required to include at least a 10 s excerpt of the subject talking naturally in a *predominantly* frontal plane

(camera-facing), although brief deviations from this frontal plane were permitted. Videos with more than one subject were excluded. For each individual, age was recorded (or estimated if the exact age could not be confirmed) according to groups 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80+. Age-matched control videos were downloaded of individuals with no known or apparent oculofacial disorder. For each of the videos, our program 'VALID' automatically calculated median MRD1, median MRD2, blink lagophthalmos, and average ocular surface exposure for the side most affected by disease. If both sides were equally affected or for those unaffected controls, measures were calculated for one side selected at random. For each of the test groups, statistical comparison was made using the Wilcoxon rank-sum test for each measure. Bonferroni correction was applied to an intended total alpha level of 0.05, equating to a test threshold of 0.004 for statistical significance.

In some subjects with acute FNP, other videos were available of the same subject prior to the onset of FNP and/or between 5 and 7 months after the onset of FNP. These videos were also downloaded and converted into MPEG-4 format. The Wilcoxon signed-rank test was used to compare the VALID-derived measures at the different time points, with an alpha level of 0.05.

Reliability testing

Using the same methods as above, further online videos were amalgamated from individuals who had recorded and posted two separate videos of themselves within 48 h of each other. In order to provide sufficient spread in values, this included individuals with acute onset FNP, as well as individuals with no known diagnosis. Test-retest reliability was evaluated using Bland-Altman analysis to compare the agreement in VALID-derived measures of median MRD1 and MRD2, blink lagophthalmos and average ocular surface exposure. Analysis was conducted on the affected eye in individuals with FNP, and a randomly selected eye in non-affected individuals (controls).

Statistical analysis

All statistical analysis for reliability and validity testing was performed using the R language and environment (version 2021.09.1 Build 372) [15].

RESULTS

The dataset included 77 individuals with FNP, 33 with ptosis, 33 with TED and 65 controls. The age and sex distribution of individuals included is shown in Supplementary Fig. S1 and Fig. S2 respectively. Statistical comparison of VALID-derived measures is shown in Table 1. Visual comparison of the measure distribution for each group is shown in Fig. 3.

The calculated median MRD1 was significantly reduced in individuals with ptosis compared with controls (2.2 mm versus 3.6 mm; $p < 0.001$) and increased in patients with FNP (3.9 mm; $p = 0.049$) and TED (4.1 mm; $p = 0.038$). The median MRD2 was greater in individuals with TED than in controls (6.4 mm versus 5.9 mm; $p < 0.001$). Calculated median MRD2 did not significantly differ in individuals with ptosis or FNP compared with controls. Blink lagophthalmos was significantly increased in individuals with FNP and those with TED (versus controls). Ocular surface exposure was reduced in individuals with ptosis compared with controls and increased in patients with TED.

Of those individuals with acute onset FNP, a subset of 15 had additional videos taken within the 6 months *preceding* the onset of FNP. Both blink lagophthalmos (mean change 5.0 mm; $p < 0.001$) and ocular surface area exposure (mean change 1.1 mm²; $p = 0.04$) were significantly greater after the onset of FNP. There was no significant difference in MRD1 (mean change 0.1 mm; $p = 0.6$) or MRD2 (mean change 0.2 mm; $p = 0.2$). A subset of 27 individuals with acute onset FNP had additional videos available from a timepoint of 5–7 months *following* the onset of FNP. In this subset, a significant decrease was seen in blink lagophthalmos (mean change -2.2 mm; $p = 0.004$), ocular surface area exposure (mean change -0.9 mm²; $p = 0.003$) and MRD1 (mean change -0.4 mm; $p = 0.03$). Figure 4 highlights these trends in a subset of 10 individuals with FNP who had videos available at all three timepoints, thus allowing sequential intra-individual comparisons to be appreciated.

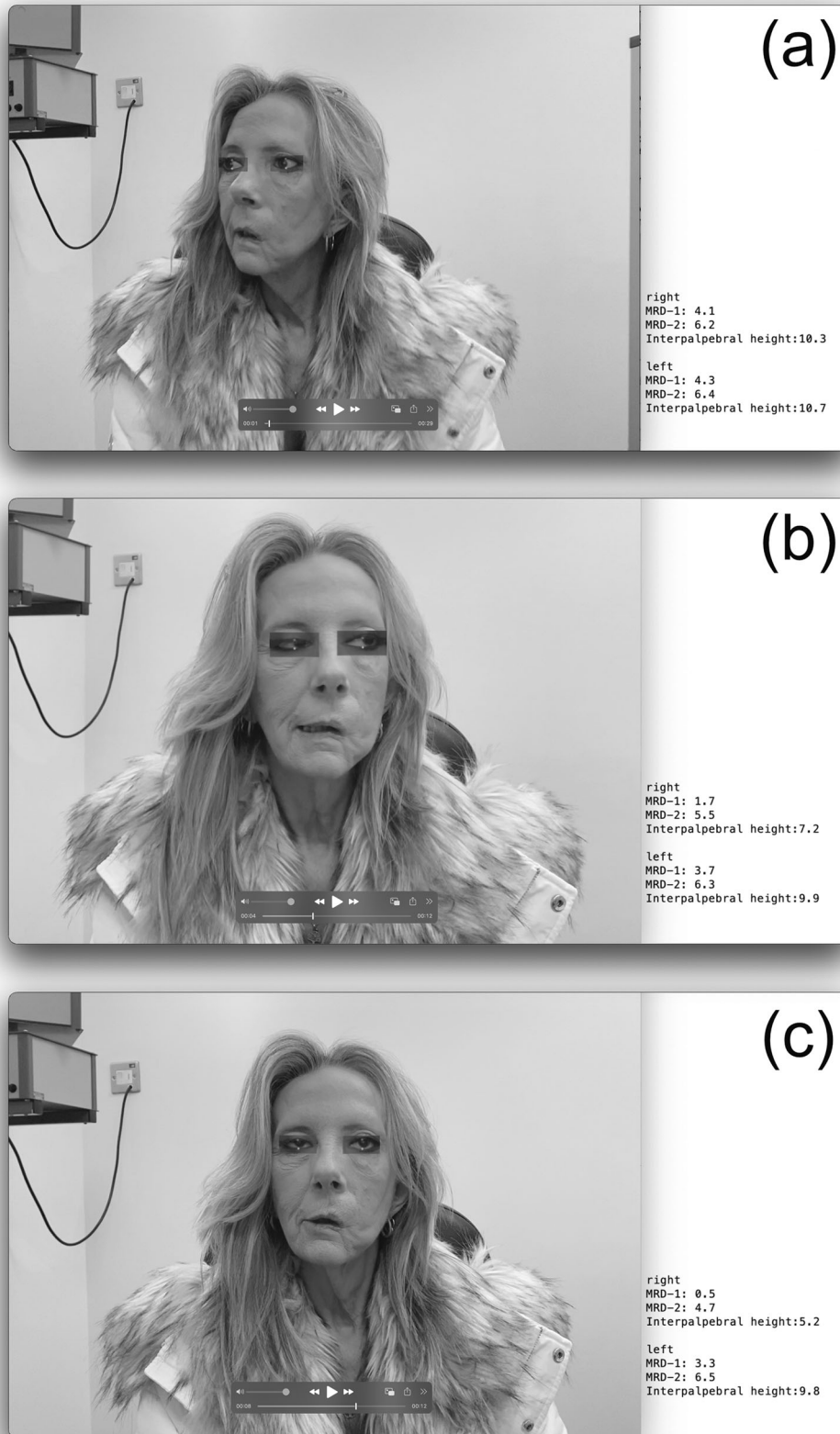


Fig. 1 Series of video still images demonstrating live eyelid parameter calculation. A series of screen captures from a single video clip of an individual with left facial nerve palsy demonstrating the calculation of MRD1 (Margin Reflex Distance 1) and MRD2 (Margin Reflex Distance 2) in real-time on a frame-by-frame basis (panels a–c). The predicted image segmentation mask overlies the region of interest (the right and left eyes). The program is designed to handle head movement (a) and eye ductions (b). Consent given from patient for use and publication of images.

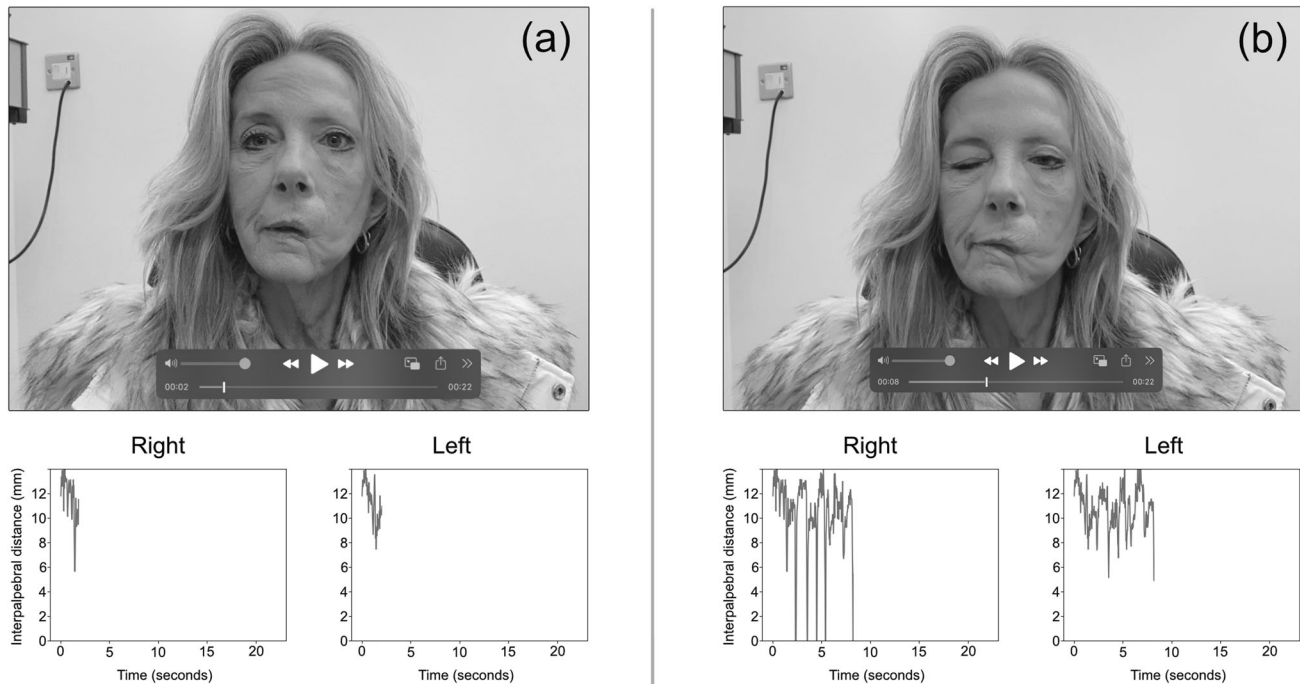


Fig. 2 Video still images demonstrating live tracking of interpalpebral height. Two sequential screen captures (**a** and **b**) from a video of an individual with left facial nerve palsy. A live tracker of interpalpebral height (IPH) for right and left eye is provided underneath each screen capture. Full blinks can be observed in the right eye, with 5 mm blink lagophthalmos seen in the left eye (**b**). Consent given from patient for use and publication of images.

Table 1. Comparison of automated eyelid measures, grouped by disorder.

Eyelid parameter	Control group (n = 65)	Facial nerve palsy (n = 77)		Ptosis (n = 33)		Thyroid eye disease (n = 33)	
	Median (IQR)	Median (IQR)	P value	Median (IQR)	P value	Median (IQR)	P value
MRD1 (mm)	3.6 (2.8–4.2)	3.9 (3.3–4.4)	0.049	2.2 (1.6–3.5)	<0.001*	4.1 (3.4–4.6)	0.038
MRD2 (mm)	5.9 (5.4–6.2)	5.8 (5.4–6.2)	0.90	5.7 (5.4–6.1)	0.37	6.4 (6.0–6.9)	<0.001*
Blink Lagophthalmos (mm)	0.0 (0.0–1.1)	3.7 (0.0–5.7)	<0.001*	0.0 (0.0–4.1)	0.06	0.0 (0.0–0.54)	0.003*
Ocular Surface Area Exposure (mm ²)	13.1 (12.3–13.8)	13.3 (12.1–14.1)	0.49	12.2 (11.5–13.0)	<0.001*	13.7 (12.9–14.5)	0.002*

Wilcoxon test used for comparison with control group.

MRD1 Margin Reflex Distance 1, MRD2 Margin Reflex Distance 2.

*P value < alpha level of 0.004 (after Bonferroni correction of alpha < 0.05).

With regards to test-retest reliability, we obtained videos of 32 individuals affected by acute onset facial nerve palsy and 34 non-affected individuals where two separate videos were recorded within 48 h of each other. Bland–Altman analysis demonstrated the 95% limits of agreement for median MRD1: -1.1 to 1.1 mm; median MRD2: -0.9 to 1.0 mm; blink lagophthalmos: -3.5 to 3.7 mm; and average ocular surface area exposure: -1.6 to 1.6 mm². Bland–Altman plots are shown in Supplementary Fig. S3.

DISCUSSION

This study demonstrates the potential feasibility of obtaining computer-derived clinically meaningful eyelid measurements from unconstrained digital videos. Using our pilot computer program (VALID) individuals with FNP were found to have greater blink lagophthalmos versus controls. Individuals with ptosis were found to have reduced MRD1 and reduced ocular surface exposure versus controls. Furthermore, subjects with TED were

found to have a greater MRD2, blink lagophthalmos, and average ocular surface exposure versus controls. All of these findings are in-keeping with expected clinical manifestations of such oculo-facial disorders and offer evidence of construct validity. It suggests that our program is able to detect differences between groups of patients and may be valuable for research purposes. Moreover, our solution appears to be able to detect differences in eyelid parameters within the same individual over time. Taken in combination with the test-retest reliability of these automated measures as reported in this study, we suggest that VALID measures taken within the same individual could be reliably used in sequence. This would be particularly useful in evaluating the disease course of an individual patient, or their response to treatment. That said, it is not yet known how well these automated measures agree with manually derived measures taken in clinic by a specialist. It is reassuring that the test-retest reliability in this study is approximately equivalent to the agreement found between two independent clinicians

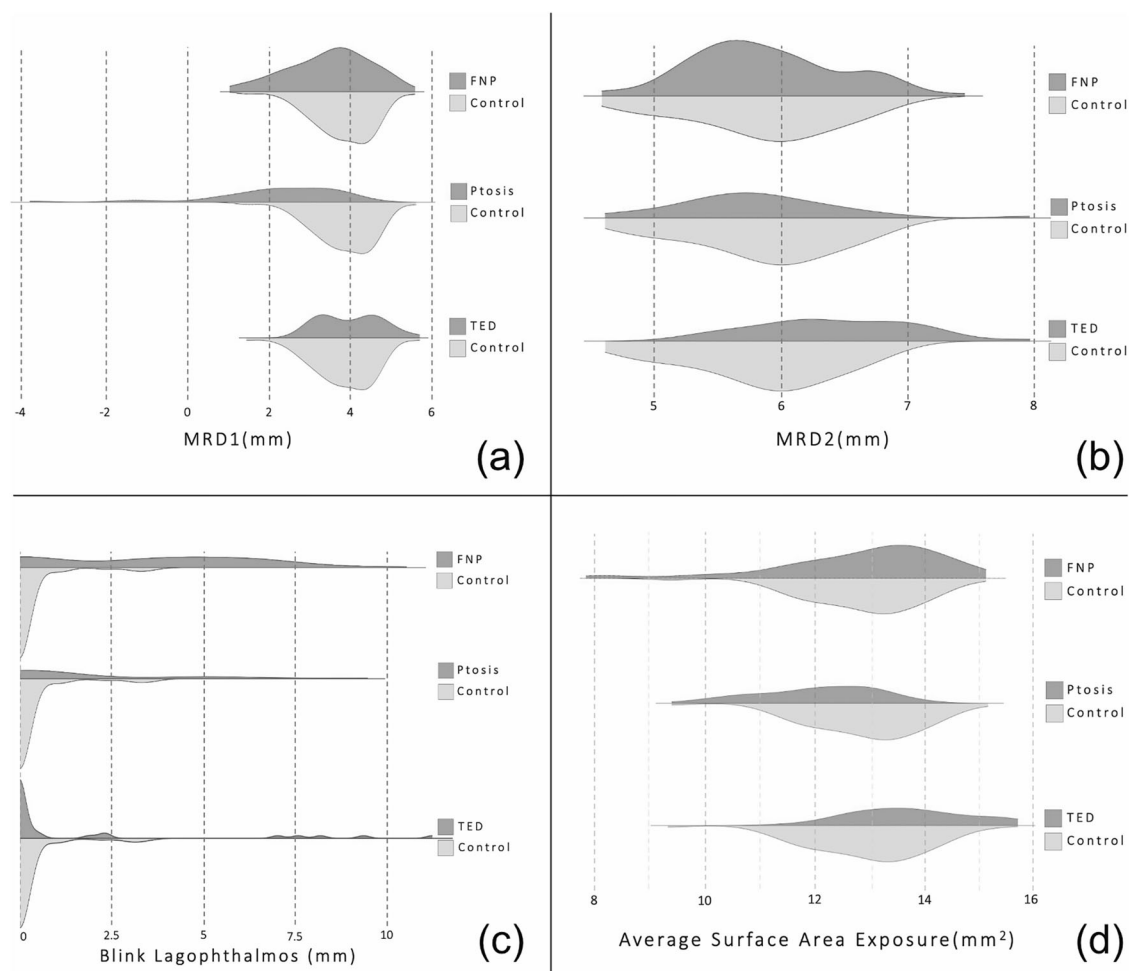


Fig. 3 Distribution of automated eyelid parameters in oculofacial disorders versus controls. Comparing the distribution of (a) MRD1 (Margin Reflex Distance 1); (b) MRD2 (Margin Reflex Distance 2); (c) blink lagophthalmos; and (d) average surface area exposure between controls (in light grey) and individuals with facial nerve palsy (FNP), ptosis and TED (thyroid eye disease).

measuring MRD1 in clinic [16]. It will be important to better explore the agreement between automated and clinically derived measures.

It is important to acknowledge that this study did not directly recruit patients with oculofacial disorders when making comparisons in measurements. Rather, it relied on publicly sourced videos of individuals with self-reported disorders, that were corroborated by a clinician based on the best available information. As such, the encouraging findings reported in this proof-of-concept study can now pave the way for further development and validation in a clinical population.

Artificial intelligence has previously been used to automatically derive MRD1 [17–19] and MRD2 [19] from still images. To date, these solutions have all relied on machine learning models that are trained on standardised and professionally sourced images in the frontal plane. In contrast, the deep learning model used within VALID is trained on a dataset of low-resolution images of mixed quality, with no standardisation in head or eye positioning. As such, this model lends itself much more favourably to video footage provided by a patient's device (webcam or smartphone) and thus video consultation, which has gained traction during the current pandemic [20]. Another important consideration is that still images taken by an untrained user only provide a single snapshot of a patient's oculofacial status. It is well known that subjects may involuntarily activate certain periorcular muscle groups as a reflex response to being photographed. Therefore, video footage used by VALID is more likely to capture a patient in

a natural state that is representative of their everyday and dynamic oculofacial status.

Eyelid function is a dynamic process, and therefore lends itself more to video assessment versus photography. Blink lagophthalmos is an important dynamic parameter which is difficult to accurately measure in clinic or from still images. By using video footage, our software was able to elicit blink lagophthalmos values, which were significantly greater in individuals with FNP and TED when compared with controls. This data could be used both to stratify risk of exposure keratopathy at initial presentation using designated acute FNP pathways, and to monitor recovery in some individuals. In our small cohort of individuals with FNP with follow-up data, most eyelid parameters had recovered at 6 months from onset. However, there appears to be a residual deficit in blink lagophthalmos that can be appreciated objectively using the presented method, and correlates well with anecdotal clinical experience. Thus, the example of blink lagophthalmos highlights the possible advantages of automated eyelid measurements over still image analysis, and in some instances may even also supplement face-to-face examination.

There are, of course, other important eyelid measurements used in oculoplastic assessments (e.g., exophthalmometry, levator function and eyelid laxity), which are not addressed in this study. There is still a very long way to go if telemedicine is going to fully satisfy the needs of our daily clinical practice. However, this proof-of-concept study offers promise that, with targeted application, such developments could now be within reach. Deep learning

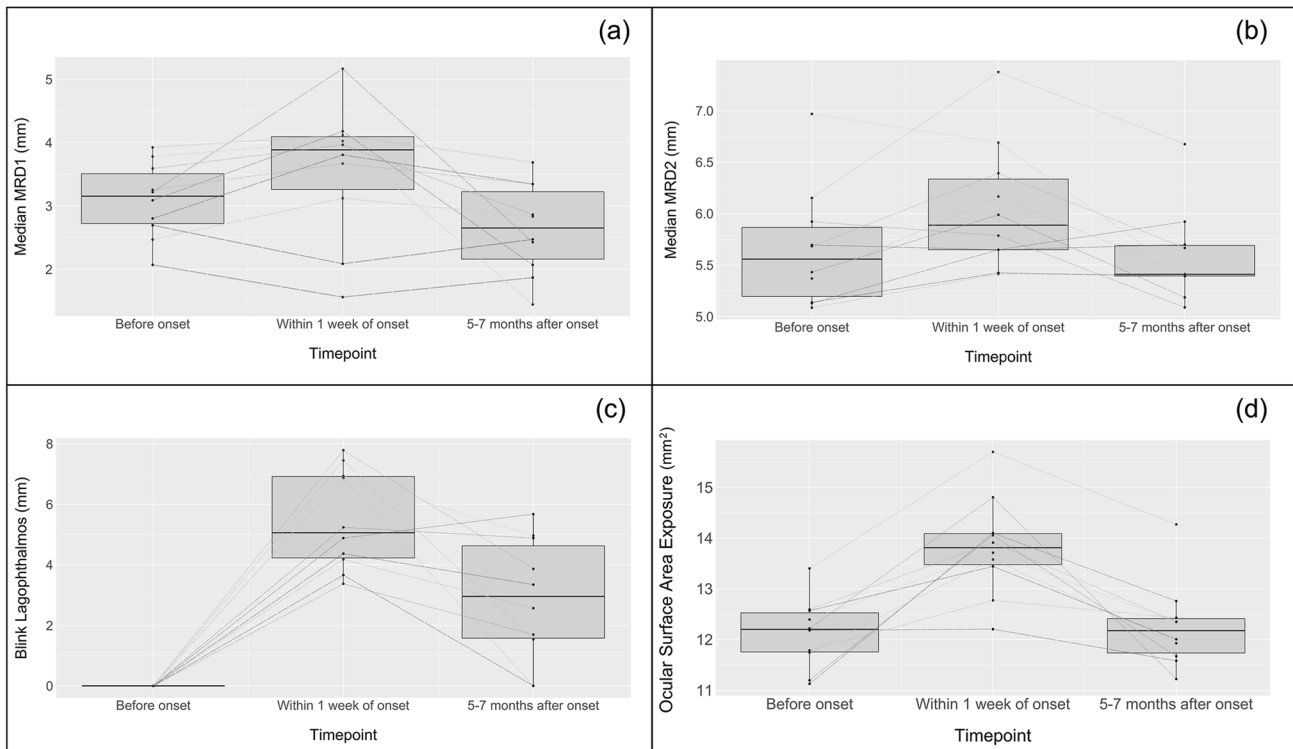


Fig. 4 Boxplots demonstrating the change in eyelid parameters in 10 individuals before the onset of FNP (Facial Nerve Palsy), within 1 week of onset and at 6 months (± 1 month) after onset. Automated eyelid measures include (a) Median MRD1 (Margin Reflex Distance 1); (b) Median MRD2 (Margin Reflex Distance 2); (c) Blink Lagophthalmos; (d) Ocular Surface Area Exposure.

models and computer vision algorithms might soon offer other useful adjuncts to the oculoplastic service, for example in monitoring clinical activity of TED, or triaging patients with FNP and periocular skin lesions. Digitally enabled outpatient care is viewed by many as one of the central pillars of a sustainable future healthcare service. This study represents one of many first steps towards this vision.

Summary

What was known before

- Telemedicine and mobile health monitoring is likely to play a key role in the delivery of future healthcare services. Automated eyelid measures have been captured from photos of patients using machine learning techniques.

What this study adds

- Using this computer vision model, it is feasible to automatically capture clinically relevant eyelid measures from videos. When tested on videos of individuals with known Oculofacial disorders, the captured measurements demonstrate promising reliability and validity.

DATA AVAILABILITY

The datasets generated and analysed during this study are available from the corresponding author on reasonable request.

REFERENCES

1. NHS England. NHS long term plan. NHS England, 2019. <https://www.england.nhs.uk/long-term-plan/>; Accessed 10 September 2021.
2. Chen EM, Andoh JE, Nwanyanwu K. Socioeconomic and demographic disparities in the utilization of telemedicine for ophthalmic care during the COVID-19 pandemic. *Ophthalmology*. 2021;129:15–25.
3. Portney DS, Zhu Z, Chen EM, Steppe E, Chilakamarri P, Woodward MA, et al. COVID-19 and use of teleophthalmology (CUT Group): trends and diagnoses. *Ophthalmology*. 2021;128:1483–5.
4. Jamison A, Diaper C, Drummond S, Tejwani D, Gregory ME, Cauchi P, et al. Telemedicine in oculoplastics: the real-life application of video consultation clinics. *Ophthalmic Plast Reconstr Surg*. 2021;37:S104–S108.
5. Kang S, Raja L, Sim DA, Thomas PBM, Ezra DG. Telemedicine in oculoplastic and adnexal surgery: clinicians' perspectives in the UK. *Br J Ophthalmol*. 2022;106:1344–9.
6. Homer NA, Epstein A, Shore JW, Somogyi M. Oculoplastic surgeons' surgical, clinical, and management experiences during the COVID-19 Crisis. *Am J Cosmet Surg*. 2021;38:205–10.
7. Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley, CA: CreateSpace; 2009.
8. Bradski G. The OpenCV Library. *Dr Dobb's J Softw Tools*. 2000;120:122–5.
9. V Kazemi and J Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1867–74.
10. King DE. Dlib-ml: a machine learning toolkit. *J Mach Learn Res*. 2009;10:1755–8.
11. Ronneberger, O, Fischer, P, & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Cham: Springer; 2015, pp. 234–41.
12. Luo B, Shen J, Wang Y, Pantic M. The ibug eye segmentation dataset. Pirovano E, Graversen E, editors. 2018 Imperial College Computing Student Workshop, ICCSW 2018. London, United Kingdom: Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik. Vol. 66 of OASICS, 2018; pp. 7:1–7:9 <https://doi.org/10.4230/OASICS.ICCSW.2018.7>.
13. Bergmanson JP, Martinez JG. Size does matter: what is the corneo-limbal diameter? *Clin Exp Optom*. 2017;100:522–8.
14. Gharaee H, Abrishami M, Shafiee M, Ehsaei A. "White-to-white corneal diameter: normal values in healthy Iranian population obtained with the Orbscan II". *Int J Ophthalmol*. 2014;7:309–12.

15. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. <https://www.R-project.org/>.
16. Boboridis K, Assi A, Indar A, Bunce C, Tyers AG. Repeatability and reproducibility of upper eyelid measurements. *Br J Ophthalmol*. 2001;85:99–101.
17. Thomas PBM, Gunasekera CD, Kang S, Baltrusaitis T. An artificial intelligence approach to the assessment of abnormal lid position. *Plast Reconstr Surg Glob Open*. 2020;8:e3089.
18. Aleem A, Nallabothula MP, Setabutr P, Hallak JA, Yi D. AutoPtosis. arXiv preprint arXiv:2106.03905. 2021 Jun 7. Available at: <https://doi.org/10.48550/arXiv.2106.03905>.
19. Van Brummen A, Owen JP, Spaide T, Froines C, Lu R, Lacy M, et al. PeriorbitAI: artificial intelligence automation of eyelid and periorbital measurements. *Am J Ophthalmol*. 2021;230:285–96.
20. Kang S, Thomas PBM, Sim DA, Parker RT, Daniel C, Uddin JM. Oculoplastic video-based telemedicine consultations: Covid-19 and beyond. *Eye*. 2020;34:1193–5.

ACKNOWLEDGEMENTS

We gratefully acknowledge the open source communities associated with the programming languages and libraries cited in the manuscript. There are no funding sources to disclose for this work.

AUTHOR CONTRIBUTIONS

CS, SM, PT and SK responsible for study design. CB and HC responsible for data collection and analysis, as well as manuscript preparation. All authors approved the manuscript in its final form.

COMPETING INTERESTS

The authors declare no competing interests.

CONSENT FOR PUBLICATION

Patient consent is given for use and publication of images (Figs. 1 and 2).

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41433-023-02424-z>.

Correspondence and requests for materials should be addressed to Christopher B. Schulz.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.