AMERICAN ACADEMY OF OPHTHALMOLOGY®

Marker-Free, Automated Eyelid Assessment in Thyroid Eye Disease Using Artificial Intelligence: A Multicenter Validation Study

Jae Hoon Moon, MD, PhD,[1,2,3,*] *Jongchan Kim,*[1,*] *Joonhyeon Park, PhD,*[1] *Min Joo Kim, MD, PhD,*[2] *Tae Jung Oh, MD, PhD,*[2] *Joon Ho Moon, MD, PhD,*[2] *Sung Hye Kong, MD, PhD,*[2] *Kyubo Shin, PhD,*[1] *Jaemin Park, MS,*[1] *Jin Sook Yoon, MD, PhD,*[4] *JaeSang Ko, MD, PhD,*[4] *Won Sang Yoo, MD, PhD,*[5] *Raquel Monge Carmona, MD,*[6,7,8] *Marina Soto Sierra, MD,*[6] *Luz María Valverde Cano, MD,*[6] *Tomás Martín Hernández, MD, PhD,*[9] *Mariola Méndez Muros, MD,*[7,9] *Namju Kim, MD, PhD,*[10] *Antonio Manuel Garrido Hermosilla, MD, PhD*[6,7,11,12]

Objective: To validate "Glandy LID," a novel marker-free, deep learning–based software designed for the automated assessment of eyelid morphology using intrinsic corneal diameter for calibration.

Design: Multicenter retrospective study.

Participants: The internal validation cohort consisted of 119 patients with thyroid eye disease (TED) from Seoul National University Bundang Hospital (South Korea). The external validation cohort included 140 patients from Hospital Universitario Virgen Macarena (Spain).

Methods: An artificial intelligence (AI) algorithm segmented eyelid and corneal regions from facial photographs. Eyelid metrics were calculated using population-specific corneal diameters as a reference scale for pixel-to-millimeter conversion. The internal cohort utilized digital single-lens reflex images compared against ground truth derived from physical markers. The external cohort utilized smartphone-captured images compared against clinical medical records to assess clinical generalizability.

Main Outcome Measures: Geometric precision was assessed using Intersection over Union. Accuracy of quantitative measurements (margin reflex distance 1 [MRD1] and 2 [MRD2]) was evaluated using Pearson correlation coefficient (PCC), mean absolute error, and mean absolute percentage error (MAPE).

Results: In the internal validation, the AI system demonstrated high geometric precision (Intersection over Union 0.94). Quantitative measurements showed excellent agreement with the ground truth, achieving a PCC of 0.98 for MRD1 and 0.94 for MRD2, with low MAPEs of 5.69% and 4.57%, respectively. In the external validation using smartphone images without markers, the system maintained high reliability. For MRD1, it achieved a PCC of 0.94 and a MAPE of 9.06%. Although MRD2 showed a slightly higher MAPE (15.07%), likely due to manual measurement variability in clinical records, the correlation remained strong (PCC 0.93).

Conclusions: This AI system provides an accurate and robust automated solution for eyelid measurement without the need for physical reference markers. By overcoming the limitations of manual assessment and marker-based systems, this tool offers a practical and accessible method for objectively monitoring TED in both routine clinical practice and remote health care settings.

Financial Disclosures: Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2026;6:101202*

Supplemental material available at www.ophthalmologyscience.org.

Lid retraction is the most common clinical sign of thyroid eye disease (TED) and serves as a critical indicator of both its activity and severity.[1–3] The standard method for quantitatively assessing upper eyelid retraction is the measurement of the margin reflex distance 1 (MRD1).[4] While standardized protocols for manual measurement can yield low intraobserver and interobserver variability, their reliability is subject to several challenges. These include a learning curve for clinicians and numerous potential confounders, chief among which is the patient's physiological response to light.[5,6] The direct shining of a light to measure MRD1 often causes patients to squint, altering the very measurement being taken. Although mitigation techniques like shining a penlight on the forehead exist, these procedural complexities and inherent potential for variability can lead clinicians to rely on

http://doi.org/10.1016/j.xops.2026.101202
ISSN 2666-9145/26

subjective visual estimation, compromising the reproducibility and accuracy of the assessment.

To address these limitations, artificial intelligence (AI) is being increasingly leveraged, particularly in the field of ophthalmology where its potential is rapidly being recognized.[7–12] "Glandy LID" (THYROSCOPE INC.) is an AI-based medical device software developed to overcome the challenges of manual assessment. From a single frontal facial photograph, it automatically quantifies a comprehensive set of metrics for lid retraction, including not only the conventional MRD1 but also the mid-pupil to lid distance (MPLD), mean upper eyelid height, maximum eyelid height, and the ocular surface area. Furthermore, it provides visualization of the eyelid contour, allowing for a more qualitative assessment of its shape and changes over time. These capabilities are enabled by a deep learning algorithm that precisely segments the eyelid and corneal regions from the image.

This study aims to comprehensively validate the AI software, "Glandy LID," through two distinct validation phases. First, we present the results of a pivotal internal validation (the domestic clinical trial) that served as the basis for regulatory approval by the Korean Ministry of Food and Drug Safety. This phase was designed to rigorously establish the core accuracy and geometric precision of the algorithm within a controlled environment. Second, we present an external validation conducted on an independent Spanish patient cohort to evaluate the software's generalizability across different ethnicities and its robustness in real-world clinical settings using smartphone photography. By combining these two distinct objectives, we seek to establish "Glandy LID" as a reliable and versatile tool for objective TED assessment.

Methods

Study Design and Data Set

This multicenter retrospective study was designed to validate the performance of the AI-based eyelid analysis software, "Glandy LID," by addressing two primary validation objectives. Phase 1 (internal validation) focused on verifying technical precision against a high-resolution digital gold standard (digital single-lens reflex with physical markers) to meet regulatory performance standards. The data set obtained from Seoul National University Bundang Hospital (SNUBH) in Seongnam, Korea, comprised a total of 519 patients diagnosed with, or suspected of having, TED who visited the clinic between August 2013 and July 2023. This SNUBH cohort was divided into a training set of 400 patients, used to develop and optimize the algorithm, and an internal validation set of 119 patients. For the internal validation set, digital facial images were captured using a digital single-lens reflex camera at a 1-m distance with a 5 mm circular reference marker attached to the patient's face for precise ground truth calibration.

Following this rigorous internal validation, phase 2 (external validation) was conducted to assess the system's clinical generalizability and performance in a variable environment using smartphone-captured images without external markers. For this phase, an initial pool of 1614 images was retrospectively screened from patients diagnosed with, or suspected of having, TED at Hospital Universitario Virgen Macarena in Seville, Spain, between September 2021 and May 2024. After a manual retrospective review, 495 images were excluded due to inappropriate data types (e.g., electronic medical record [EMR] screen captures), nonfrontal poses, or significant processing limitations such as blinking or insufficient contrast. This resulted in a final curated data set of 1119 images from 140 patients. Unlike the internal data set, these images were acquired using various smartphone cameras without reference markers, relying on clinical medical records for ground truth to reflect real-world clinical practice.

This study was approved by the Institutional Review Board (IRB) and Ethics Committees of SNUBH (IRB No. 2107-709-101) and Hospital Universitario Virgen Macarena (IRB No. SICEIA-2025-000843). Informed consent was waived by the IRB due to the retrospective nature of the study. All experiments were performed following the relevant guidelines and regulations. The study adhered to the tenets of the Declaration of Helsinki. All images and clinical data were anonymized prior to analysis.

Ground Truth Establishment

To comprehensively validate the AI system, we established the ground truth (reference standard) using the following two distinct methodologies, tailored to the specific objectives of the internal and external validations:

- Internal validation (precision assessment): for the internal validation at SNUBH, a rigorous reference standard was established to verify the algorithm's geometric accuracy and adherence to regulatory performance standards. Prior to image acquisition, a circular sticker with a diameter of 5 mm was attached to the patient's face. The physical diameter of the sticker was verified using medical calipers to ensure precise pixel-to-millimeter calibration. An experienced oculoplastic surgeon manually annotated the boundaries of the eyelids and cornea on the digital images. Based on these manual annotations and the 5-mm reference scale, the ground truth values for MRD1 and margin reflex distance 2 (MRD2), MPLD at various angles, and ocular surface area were calculated. These values served as the gold standard for comparison with the AI outputs.
- External validation (real-world clinical assessment): for the external validation at Hospital Universitario Virgen Macarena, the study aimed to evaluate the system's performance in a routine clinical environment where reference markers are not typically used. Therefore, images were captured without facial calibration stickers, reflecting standard outpatient photography. The ground truth was defined as the clinical measurements (e.g., MRD1 and MRD2) recorded in the patients' EMR. These values were obtained by ophthalmologists using conventional manual measurement techniques (e.g., millimeter ruler) during clinical examinations. This approach allowed for the assessment of the AI's agreement with standard clinical practice.

Development of AI Algorithm

The "Glandy LID" software was developed to precisely quantify palpebral fissure geometry, including lid contours and associated metrics such as lengths and areas. The core of the system consists of two independent semantic segmentation models based on the Bilateral Segmentation Network architecture, a lightweight framework optimized for real-time, high-resolution semantic segmentation.

To ensure analysis reliability, the software incorporates a built-in quality control module. This module targets images that meet a minimum resolution requirement of 7 megapixels and evaluates image suitability based on the Face Quality Score as defined in International Organization for Standardization 29794-5.[13] Specifically, the system verifies image quality using quantitative

metrics including lighting symmetry, brightness, contrast, Global Contrast Factor, exposure, blur, and sharpness. Images that fail to meet these quality thresholds are automatically rejected prior to segmentation to minimize measurement errors caused by suboptimal acquisition conditions.

Following quality verification, the independent semantic segmentation models delineate the palpebral fissure and the cornea. The corneal segmentation model fulfills two critical functions: first, the segmented corneal contour provides an estimate of the pupil center, which is essential for calculating MRD1, MRD2, and MPLDs; second, the corneal radius derived from the segmentation enables pixel-to-millimeter conversion by applying population-specific average corneal radii stratified by sex and ethnicity.[14,15] Specifically, gender-specific averages were applied for the internal Korean cohort (11.47 mm for males; 11.17 mm for females),[16] while a unified average of 11.86 mm was applied for both males and females in the Spanish cohort.[17]

The palpebral fissure and cornea models were trained independently using a dataset of 1350 facial images, comprising 400 clinical photographs of patients with TED collected at SNUBH and 950 smartphone-based images from adult volunteers (primarily healthy controls) affiliated with THYROSCOPE INC., following a standardized imaging protocol. This hybrid training approach was employed to simultaneously capture specific TED-related pathological features from clinical data and ensure environmental robustness against diverse smartphone imaging conditions. All images were resized to 512 × 512 pixels, with a 448 × 448 random crop used as the final input. For each task, the data set was randomly partitioned into training (85%) and validation (15%) subsets. Training was conducted for up to 200 epochs, selecting the model with the lowest validation loss as the final version. To supervise learning across Bilateral Segmentation Network's three output scales (out, out16, out32), the Online Hard Example Mining Cross-Entropy Loss was computed for each output and summed to form the final objective; this loss function was chosen for its effectiveness in handling small and challenging anatomical boundaries. Finally, to ensure robustness against variations in lighting, camera type, and facial presentation, extensive data augmentation techniques were employed during training, including random resizing with preserved aspect ratio, random cropping, horizontal flipping, random scaling, and color jitter adjustments for brightness, contrast, and saturation.

Statistical Analysis

The performance of the AI algorithm was evaluated by comparing its automated measurements against the established ground truth for each dataset. For the internal validation, the agreement between the AI-predicted values (calibrated using gender-specific corneal diameters) and the precise manual ground truth (calibrated using the 5-mm reference sticker) was assessed using the Pearson correlation coefficient (PCC), mean absolute error (MAE), and mean absolute percentage error (MAPE). Additionally, the geometric accuracy of the segmentation masks for determining the ocular surface area was evaluated using the Intersection over Union. The performance criteria for regulatory approval were defined as a PCC $\geq$0.9, MAPE $\leq$0.1, and Intersection over Union $\geq$0.9. For the external validation, the agreement between the AI-predicted values (calibrated using a unified corneal diameter) and the clinicians' measurements recorded in the EMR was similarly analyzed using the PCC, MAE, and MAPE to verify the system's reliability and precision in a real-world clinical setting. Additionally, a sensitivity analysis was performed using the internal validation data set to quantify the theoretical impact of corneal diameter variations on measurement accuracy (MAE).

Results

Baseline Characteristics

A total of 252 patients were included in this study, comprising 119 patients in the internal validation set (Korean cohort) and 140 patients in the external validation set (Spanish cohort). The baseline demographic and clinical characteristics of the study populations are summarized in Table 1. The mean age of patients was 44.7 $\pm$ 15.9 years in the internal set and 49.5 $\pm$ 13.4 years in the external set. Both cohorts showed a female predominance (73.1% and 77.1%, respectively), consistent with the known epidemiology of TED. The mean MRD1, serving as a primary indicator of lid retraction, was 3.99 $\pm$ 1.48 mm in Korean cohort and 4.81 $\pm$ 1.30 mm in Spanish cohort.

Validation of AI Performance

The internal validation assessed the algorithm's absolute precision against a rigorous ground truth established by manual annotation and a 5-mm reference marker. The geometric accuracy of eyelid segmentation, measured by the intersection over union, was 0.94, surpassing the performance goal of 0.90. Quantitative measurements also demonstrated high agreement with the ground truth (Table 2 and Fig 1). The PCC were 0.98 for MRD1, 0.94 for MRD2, 0.96 for angle-specific MPLD, and 0.92 for ocular surface area. In terms of error metrics, the system achieved an MAE of 0.20 mm for MRD1, 0.26 mm for MRD2, 0.32 mm for MPLD, and 12.42 mm^2 for ocular surface area. The MAPE for these metrics was consistently low: 5.69% for MRD1, 4.57% for MRD2, 4.96% for MPLD, and 7.12% for ocular surface area. All parameters met the regulatory success criteria of PCC $\geq$0.9 and MAPE $\leq$10%. Furthermore, the Bland−Altman analysis confirmed minimal bias and narrow limits of agreement between the AI predictions and manual measurements for both MRD1 (bias: 0.064 mm) and MRD2 (bias: 0.143 mm), reinforcing the system's reliability (Fig 2).

To evaluate the system's robustness in a real-world clinical environment, an external validation was conducted on the Spanish cohort using smartphone-captured images without calibration stickers. Notably, within the curated external validation dataset (n = 1119) that followed manual prescreening for pose and data type, 100% of the images successfully satisfied the software's automated quality control thresholds for lighting, exposure, and focus. In this analysis, where the AI's corneal-based measurements were compared against clinical values from medical records, the system maintained high accuracy (Table 3 and Fig 3). For MRD1, the analysis yielded a PCC of 0.94, an MAE of 0.40 mm, and a MAPE of 9.06%. Similarly, for MRD2, the system showed a PCC of 0.93, an MAE of 0.79 mm, and a MAPE of 15.07%. The Bland−Altman analysis further visualized the agreement between the AI predictions and clinical measurements (Fig 4). Margin reflex distance 2 was analyzed as a secondary finding, reflecting the challenges of obtaining consistent manual reference standards for the lower eyelid in routine clinical practice.

Table 1. Basal Characteristics of the Subjects from Internal and External Validation Data Set

	Internal Validation Set (Korea)	External Validation Set (Spain)
Number of patients	119	140
Number of ocular images	119	1119
Age (years)		
Mean ± SD	44.7 ± 15.9	49.5 ± 13.4
Range	10 − 83	19 − 83
Sex, n (%)		
Male	32 (26.9%)	32 (22.9%)
Female	87 (73.1%)	108 (77.1%)
Clinical metrics (ground truth)		
MRD1 (mm)	3.99 ± 1.48	4.81 ± 1.30
MRD2 (mm)	5.55 ± 0.92	5.23 ± 0.89
Device used	DSLR	Smartphone

Data are presented as mean ± standard deviation for continuous variables and number (%) for categorical variables.
MRD1 = margin reflex distance 1; MRD2 = margin reflex distance 2; DSLR = digital single-lens reflex; SD = standard deviation.

Discussion

This study successfully developed and validated "Glandy LID," a novel deep learning-based software designed for the automated, objective assessment of eyelid morphology in patients with TED. Our primary objective was to overcome the limitations of conventional manual measurements and existing computer-assisted methods that rely on cumbersome physical markers. The results from the internal validation, conducted on a Korean cohort, demonstrated that the system achieves high geometric precision and measurement accuracy comparable to rigorous manual annotations referenced by a 5-mm sticker. Furthermore, the external validation on a Spanish cohort confirmed the system's robustness and generalizability across different ethnicities and image acquisition environments. Most notably, the algorithm maintained high reliability in a real-world clinical setting using smartphone-captured images without any external calibration markers. These findings collectively suggest that "Glandy LID" serves as a reliable, versatile, and accessible tool for quantifying eyelid retraction, a hallmark of TED severity.

The pursuit of automated eyelid measurement has been a longstanding goal in oculoplastic research, yet previous attempts have often been hindered by procedural constraints. While recent deep learning initiatives have shown promise, the majority of these systems—including those proposed by Shao et al,[8] Lou et al,[18] and Nam et al[19]—necessitate the attachment of a circular reference sticker on the patient's forehead to establish a pixel-to-millimeter ratio. Similarly, smartphone-based algorithms developed by Chen et al[20] and Bodnar et al[21] also rely on fixed-size markers placed on the face. Although these marker-based methods can yield accurate measurements, the requirement for a physical reference object complicates the clinical workflow, introduces potential errors from marker placement or detachment, and severely limits the utility of the software for retrospective analysis of archival photographs where markers are absent. In contrast, the "Glandy LID" system introduces a paradigm shift by utilizing an intrinsic biological reference: the horizontal corneal diameter. By leveraging an advanced segmentation algorithm that precisely detects the corneal limbus, our system automatically calibrates measurements without the need for external artifacts. This "marker-free" approach significantly streamlines the image acquisition process, making it as simple as taking a standard facial photograph. A critical concern with biological calibration is the natural variation in corneal diameter among individuals. However, our results confirm that the software's calibration method based on population-specific corneal diameter averages is sufficiently robust for clinical use. The successful validation on both the Korean data set (using gender-specific averages) and the Spanish dataset (using a unified average) underscores that this intrinsic calibration remains reliable despite ethnic differences in ocular anatomy. To further quantify this robustness, we conducted a sensitivity analysis using the internal validation cohort, where the ground truth was established with a 5-mm physical reference marker for maximum precision (Supplementary Table 1). Our analysis confirmed that the system maintains a high level of reliability, with a predicted MAE below 0.5 mm for corneal diameters ranging from 10.53 to 12.24 mm. This range encompasses the vast majority of the normal adult population, offering higher precision than conventional manual assessment. Furthermore, even in cases of mild microcornea or megalocornea (ranging from 9.95 to 13.13 mm), the predicted MAE remained within 0.75 mm, which is

Table 2. Performance Metrics for Internal Validation (Korean Cohort)

	PCC	MAE (mm or mm^2)	MAPE (%)
MRD1	0.98	0.20	5.58
MRD2	0.94	0.26	4.59
MPLD (angle-specific)	0.96	0.32	4.98
Ocular surface area	0.92	12.42	7.12

PCC = Pearson correlation coefficient; MAE = mean absolute error; MAPE = mean absolute percentage error; MRD1 = margin reflex distance 1; MRD2 = margin reflex distance 2; MPLD = mid-pupil lid distance.

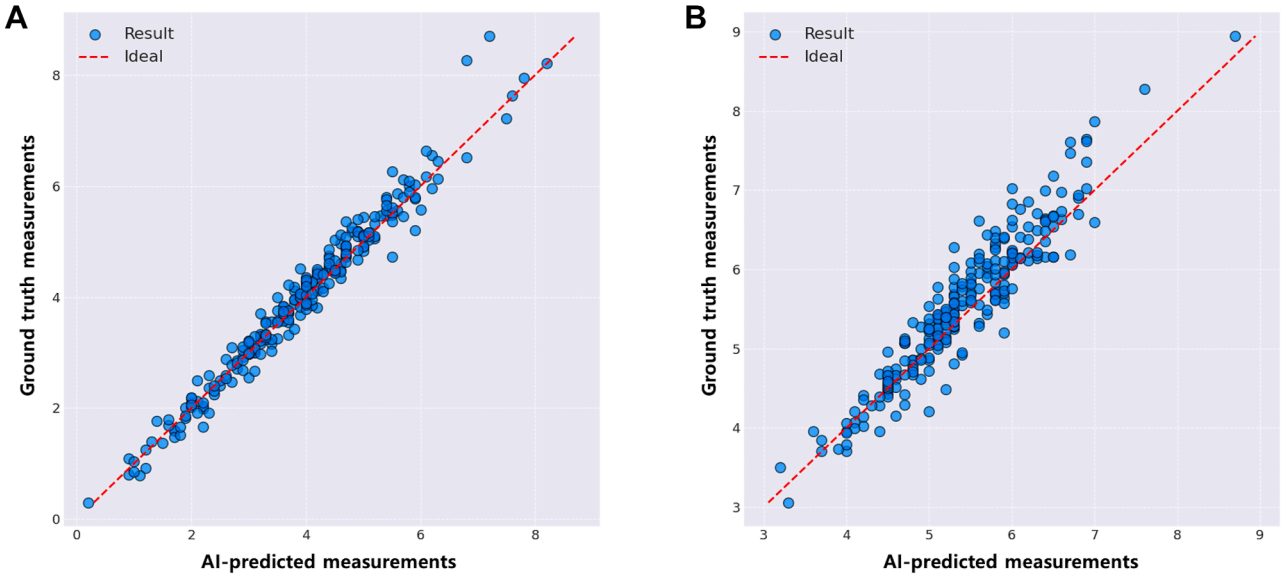

Figure 1. Correlation between AI-predicted measurements and ground truth values in the internal validation cohort. Scatter plots illustrating the linear relationship between the AI-predicted measurements and the manual ground truth measurements for (**A**) margin reflex distance 1 and (**B**) margin reflex distance 2. The red dashed line represents the ideal 1:1 reference line. The blue circles represent individual data points. AI = artificial intelligence.

sufficient for clinically meaningful longitudinal tracking. Furthermore, the system's performance on the Spanish data set, which comprised images taken with various smartphone cameras under uncontrolled lighting conditions, demonstrates that our algorithm is resilient to the variations in image quality and resolution that are typical in mobile health applications. This robustness positions "Glandy LID" as a superior alternative to previous systems, offering enhanced usability without compromising accuracy.

The clinical utility of "Glandy LID" lies in its ability to provide a comprehensive topographic analysis of the eyelid, extending beyond simple metric calculation. By visualizing the entire eyelid contour and quantifying it using radial MPLD, the software enables clinicians to objectively detect subtle shape deformities, such as the "lateral flare" characteristic of TED (Fig 5). This automated visualization helps overcome the interobserver variability inherent in manual measurements using a ruler. Furthermore, the applicability of this technology extends to a broader spectrum of eyelid disorders, including blepharoptosis (ptosis). While previous studies have demonstrated the utility of AI for evaluating ptosis severity and surgical outcomes,[18,20] their reliance on physical markers has been a barrier to routine adoption. By offering a marker-free, smartphone-compatible solution, "Glandy LID" facilitates

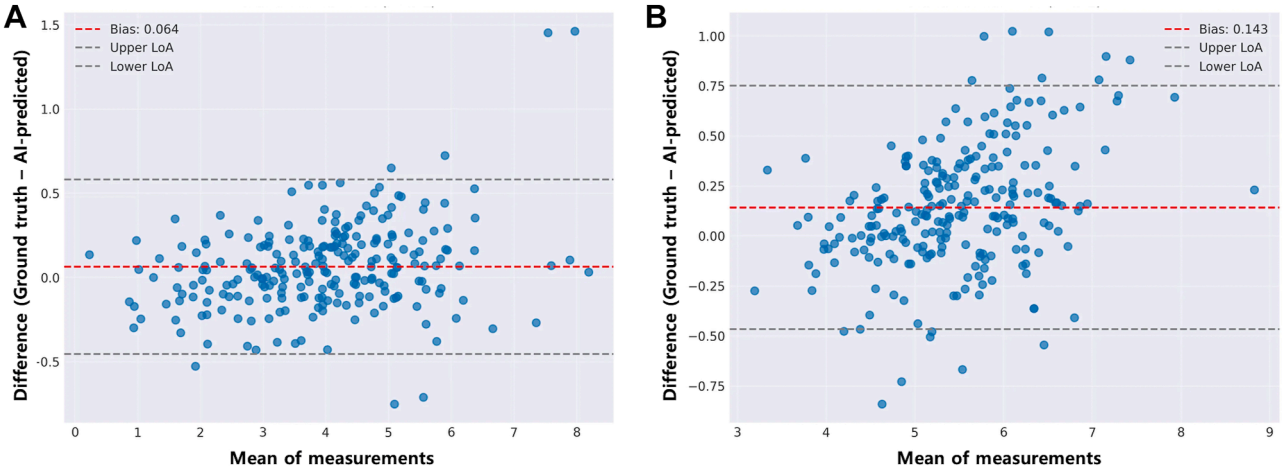

Figure 2. Bland–Altman plots of agreement between AI-predicted and ground truth values in the internal validation cohort. Plots displaying the difference between the manual ground truth and AI-predicted values against the mean of the two measurements for (**A**) margin reflex distance 1 and (**B**) margin reflex distance 2. The red dashed line indicates the mean bias (mean difference), while the gray dashed lines represent the 95% LoA (mean bias ± 1.96 standard deviations). AI = artificial intelligence; LoA = limits of agreement.

Table 3. Performance Metrics for External Validation (Spanish Cohort)

	PCC	MAE (mm)	MAPE (%)
MRD1	0.94	0.40	9.06
MRD2	0.93	0.79	15.07

PCC = Pearson correlation coefficient; MAE = mean absolute error; MAPE = mean absolute percentage error; MRD1 = margin reflex distance 1; MRD2 = margin reflex distance 2.

seamless integration into general oculoplastic practice, enabling the objective evaluation of ptosis and postoperative monitoring without the logistical burden of physical calibration.

The compatibility of our algorithm with smartphone imagery creates a foundation for telemedicine and comprehensive disease management.[22] Patients can capture facial images at home, allowing for longitudinal self-monitoring of disease progression, which is particularly valuable for the chronic course of TED. Beyond its immediate application, future research may explore the potential synergy of integrating "Glandy LID" with other AI-based assessment tools, such as previously reported systems for automated exophthalmometry and clinical activity score prediction.[9–11,23] Such a combined approach could eventually support a more comprehensive digital evaluation of both disease activity and severity. Further studies are warranted to determine how integrated AI-driven toolkits might effectively complement existing clinical grading scales and contribute to more standardized, data-driven management of TED.

Despite these promising results, this study has limitations that warrant consideration. First, while our sensitivity analysis confirmed that the corneal diameter-based calibration is robust across a broad physiological range (9.95–13.13 mm), it inherently relies on the assumption of a stable reference. Therefore, automated analysis should be supplemented or replaced by manual measurement in cases of: (1) extreme anatomical outliers, such as severe microcornea or megalocornea falling outside the validated range of 9.95 to 13.13 mm; (2) advanced ocular surface diseases or corneal opacities that obscure the corneal limbus; and (3) significant eyelid deformities or trauma that distort the palpebral fissure beyond the patterns represented in our training data set. Future iterations of the software could incorporate additional facial landmarks or iris diameter estimation to further refine this calibration. Second, regarding image quality, extreme lighting conditions, deep shadows, or low resolution remain potential challenges. To address this, "Glandy LID" incorporates an internal quality control module that automatically filters out suboptimal images prior to analysis. While 100% of the images in our external validation successfully passed this automated quality check, it is important to note that these samples had already undergone manual retrospective preselection for pose and basic visibility. This implies that in real-world clinical use, successful assessment is contingent upon capturing photographs that meet these baseline quality criteria. Third, in the external validation, the MAPE for both MRD1 (9.06%) and MRD2 (15.07%) was higher than in the internal validation. This general increase in error reflects the fundamental difference between the reference standards used; while the internal validation relied on a subpixel digital gold standard, the external validation used clinical EMR records based on manual ruler measurements, which are typically limited to 0.5-mm increments. The inherent subjectivity and lower resolution of manual measurement led to a higher MAPE across all metrics in a real-world setting. Specifically, the even-higher MAPE observed

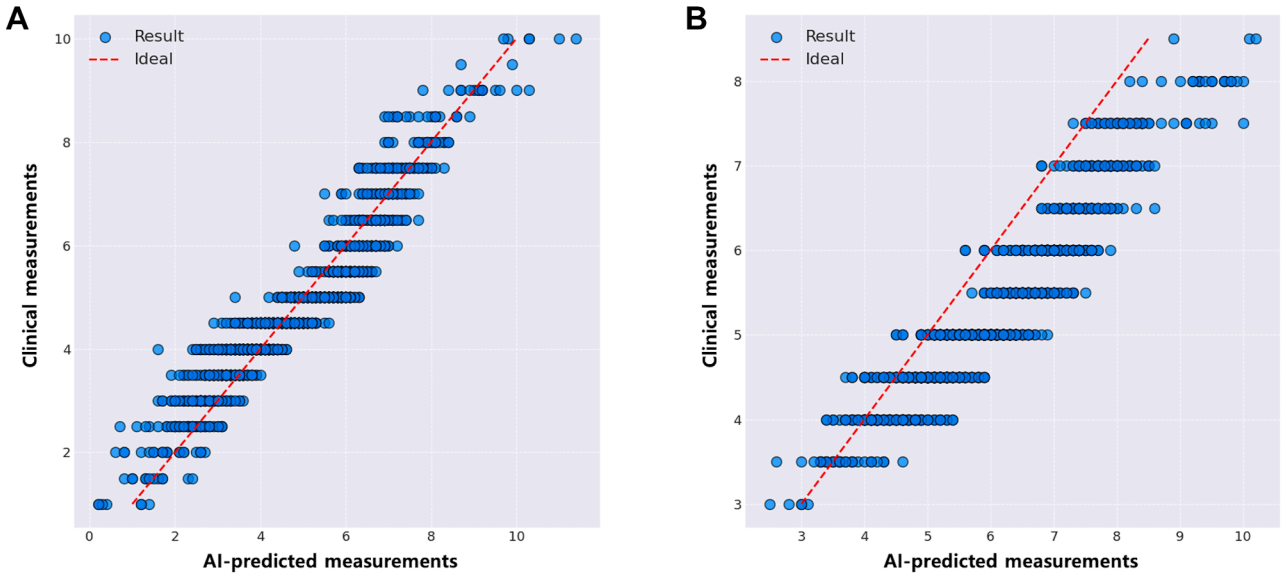

Figure 3. Correlation between AI-predicted measurements and clinical ground truth in the external validation cohort. Scatter plots illustrating the relationship between the AI algorithm's automated measurements and the clinical measurements for (**A**) margin reflex distance 1 and (**B**) margin reflex distance 2. The red dashed line represents the ideal 1:1 reference line. The blue circles represent individual data points. AI = artificial intelligence.

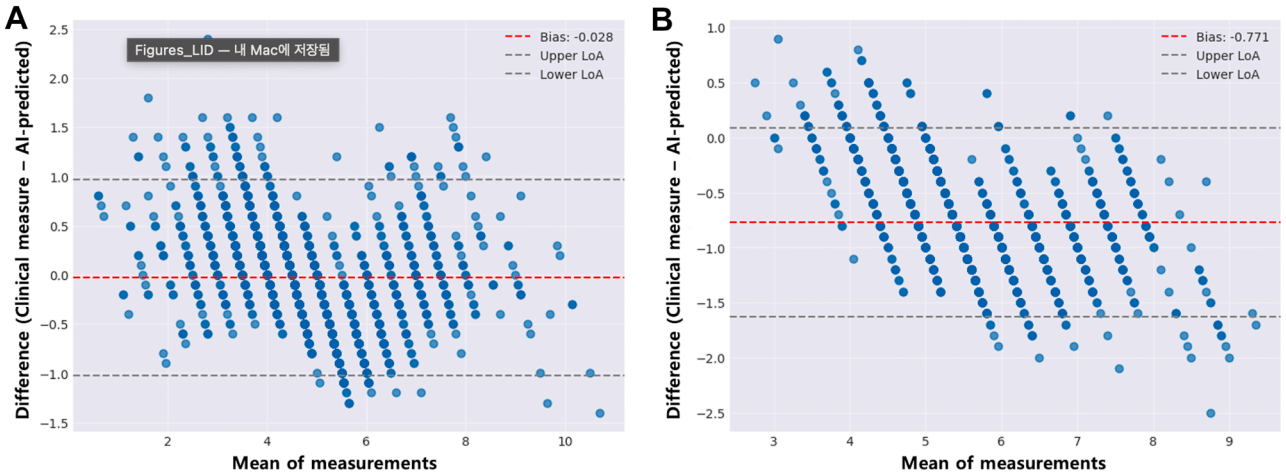

Figure 4. Bland–Altman plots of agreement in the external validation cohort. Plots displaying the difference between the clinical ground truth and AI-predicted values against the mean of the two measurements for (**A**) margin reflex distance 1 and (**B**) margin reflex distance 2. The red dashed line indicates the mean bias (mean difference), while the gray dashed lines represent the 95% LoA (mean bias ± 1.96 standard deviations). AI = artificial intelligence; LoA = limits of agreement.

for MRD2 (15.07%) likely stems from the additional difficulty in accurately localizing the lower eyelid margin due to factors such as lower eyelash interference.[8,18,20] Consequently, the elevated MAPE for MRD1, and particularly MRD2 in the external validation, should be interpreted as highlighting the inherent limitations and variability of manual measurement rather than indicating an algorithmic failure. The clinical acceptability of the

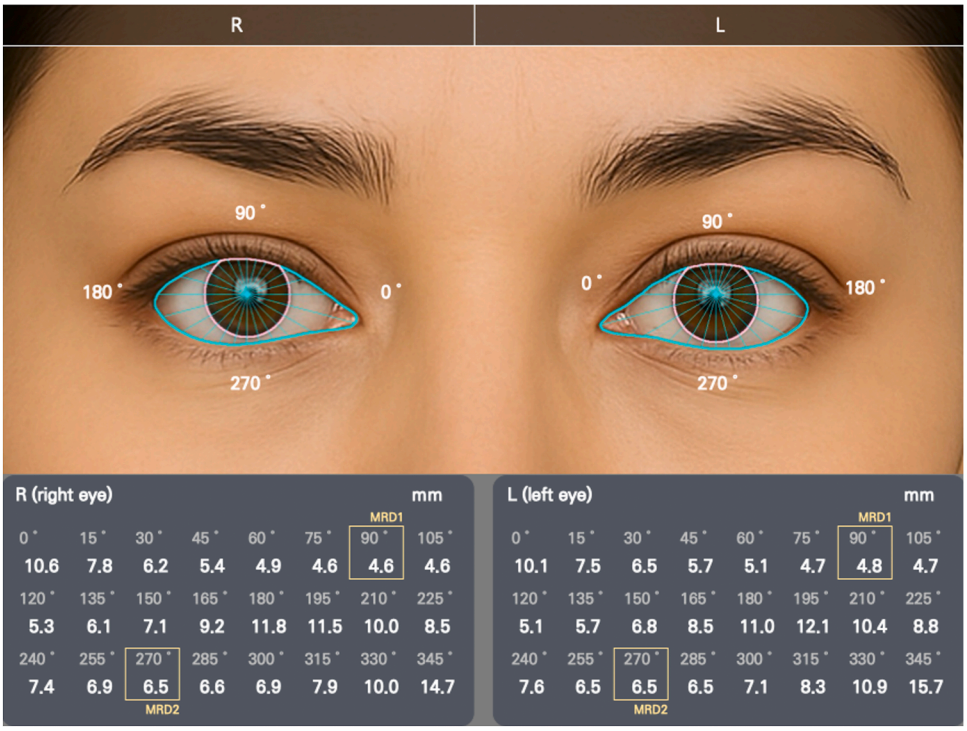

Figure 5. Visualization of comprehensive eyelid analysis using "Glandy LID". Representative output from the AI software demonstrating the automated segmentation and quantitative analysis of eyelid morphology. The upper panel shows the original facial photograph of a virtual person with overlaid segmentation masks delineating the eyelid margins (cyan lines) and radial measurement axes (cyan rays) originating from the pupil center. The lower panel displays the corresponding quantitative metrics, including MRD1 and MRD2 and radial mid-pupil lid distance measured at 15-degree intervals. Note: This figure features a virtual subject for illustrative purposes only and was not part of the quantitative validation data sets. MRD1 = margin reflex distance 1; MRD2 = margin reflex distance 2.

observed 0.4-mm MAE for MRD1 is supported by established management thresholds in TED. Considering that surgical intervention for eyelid retraction is generally indicated at 2 mm[1,24] and treatment success is defined as an interocular difference of <1 mm,[25,26] an MAE of 0.4 mm provides sufficient precision for therapeutic decision-making. Finally, the retrospective nature of this study limits our ability to assess the software's direct impact on clinical decision-making and patient outcomes. Prospective studies are needed to evaluate how the integration of this AI tool into routine practice influences treatment choices and efficiency.

In conclusion, this study demonstrates the feasibility and potential clinical utility of "Glandy LID," a marker-free AI system for automated eyelid measurement. Key applications include telemedicine-based self-monitoring and serving as a screening tool for clinicians managing Graves disease to facilitate the early detection of eyelid changes and timely referrals to ophthalmology specialists. In professional clinical settings, the software offers an objective alternative to manual measurements, enabling the quantification of subtle changes and complex eyelid contour deformities. Ultimately, this AI-driven approach represents a significant step toward more objective, accessible, and standardized management of TED.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work the authors used Gemini (Google) in order to improve the clarity, readability, and logical flow of the English text, particularly in the Abstract and Methods sections. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Footnotes and Disclosures

Originally received: December 22, 2025.
Final revision: April 11, 2026.
Accepted: April 15, 2026.
Available online: April 22, 2026. Manuscript no. XOPS-D-25-01141.

[1] THYROSCOPE INC., Ulsan, Republic of Korea.

[2] Department of Internal Medicine, Seoul National University Bundang Hospital and Seoul National University College of Medicine, Seongnam, Republic of Korea.

[3] Center for Artificial Intelligence in Healthcare, Seoul National University Bundang Hospital, Seongnam, Republic of Korea.

[4] Department of Ophthalmology, Severance Hospital, Institute of Vision Research, Yonsei University College of Medicine, Seoul, Republic of Korea.

[5] Department of Internal Medicine, Dankook University College of Medicine, Cheonan, Republic of Korea.

[6] Department of Ophthalmology, Virgen Macarena University Hospital, Seville, Spain.

[7] RICORS-REI, Institute of Health Carlos III, Madrid, Spain.

[8] Department of Physics of Condensed Matter, Optics Area, University of Seville, Reina Mercedes S/N, Seville, Spain.

[9] Department of Endocrinology and Nutrition, Virgen Macarena University Hospital, Seville, Spain.

[10] Department of Ophthalmology, Seoul National University Bundang Hospital and Seoul National University College of Medicine, Seongnam, Republic of Korea.

[11] Department of Surgery, Medical School, University of Seville, Seville, Spain.

[12] Institute of Biomedicine of Seville, IBiS / Virgen Macarena University Hospital / CSIC / University of Seville, Seville, Spain.

*J.H.M. and J.K. contributed equally to this paper.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

Jae H.M.: Other financial or nonfinancial interests — Chief Technology Officer at THYROSCOPE INC.; Stock or stock options — THYROSCOPE INC.

K.S.: Stock or stock options — THYROSCOPE INC.

Jaemin P.: Other financial or nonfinancial interests — Chief Technology Officer at THYROSCOPE INC.; Stock or stock options — THYROSCOPE INC.

J. Ko: Consultant - THYROSCOPE INC.; Stock or stock options — THYROSCOPE INC.

W.S.Y.: Consultant - THYROSCOPE INC.; Stock or stock options — THYROSCOPE INC.

N.K.: Consultant — THYROSCOPE INC.; Stock or stock options — THYROSCOPE INC.

This work was supported by the Korea Innovation Foundation (INNOPOLIS) grant funded by the Korea government (MSIT) (RS-2025-25443592).

The Article Publishing Charge fee for our article to publish as Open Access is being paid by the THYROSCOPE INC.

HUMAN SUBJECTS: Human subjects were included in this study. This study was approved by the Institutional Review Board (IRB) and Ethics Committees of SNUBH (IRB No. 2107-709-101) and Hospital Universitario Virgen Macarena (University Hospital Virgen Macarena (IRB No. SICEIA 2025-000843). Informed consent was waived by the IRB due to the retrospective nature of the study. All experiments were performed following the relevant guidelines and regulations. The study adhered to the tenets of the Declaration of Helsinki. All images and clinical data were anonymized prior to analysis.

No animal subjects were used in this study.

Author Contributions:

Conception and design: J. H. Moon, J. Kim, N. Kim, Garrido Hermosilla

Analysis and interpretation: J. H. Moon, J. Kim, Joonhyeon Park, Shin, N. Kim, Garrido Hermosilla

Data collection: J. H. Moon, J. Kim, M.J. Kim, Oh, Joon H. Moon, Kong, Carmona, Sierra, Valverde Cano, Martin Hernández, Mendez Muros, N. Kim, Garrido Hermosilla

Obtained funding: J. H. Moon, Jaemin Park

Overall responsibility: J. H. Moon, J. Kim, Joonhyeon Park, M.J. Kim, Oh, Joon H. Moon, Kong, Shin, Jaemin Park, Yoon, Ko, Yoo, Carmona, Sierra, Valverde Cano, Martin Hernández, Mendez Muros, N. Kim, Garrido Hermosilla

Abbreviations and Acronyms:
AI = artificial intelligence; **EMR** = electronic medical record; **IRB** = Institutional Review Board; **MAE** = mean absolute error; **MAPE** = mean absolute percentage error; **MPLD** = mid-pupil to lid distance; **MRD1** = margin reflex distance 1; **MRD2** = margin reflex distance 2; **PCC** = Pearson correlation coefficient; **SNUBH** = Seoul National University Bundang Hospital; **ROC** = receiver operating characteristic; **TED** = thyroid eye disease.

Keywords:
Artificial intelligence, Deep learning, Telemedicine, Eyelid retraction, Thyroid eye disease.

Correspondence:
Namju Kim, MD, PhD, Professor, Department of Ophthalmology, Seoul National University Bundang Hospital, Seoul National University College of Medicine, 82, Gumi-ro 173 Beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do 13620, Republic of Korea. E-mail: resourceful@hanmail.net; and Antonio Manuel Garrido Hermosilla, MD, PhD, Associate Professor of Ophthalmology, Department of Surgery, Medical School, University of Seville, Seville, Spain; Department of Ophthalmology, Virgen Macarena University Hospital, 3 Doctor Fedriani Avenue, Seville 41009, Spain. E-mail: gaherfamily@hotmail.com.

References

1. Bartalena L, Kahaly GJ, Baldeschi L, et al. The 2021 European Group on Graves' orbitopathy (EUGOGO) clinical practice guidelines for the medical management of Graves' orbitopathy. *Eur J Endocrinol*. 2021;185:G43−G67. https://doi.org/10.1530/EJE-21-0479.
2. Yoon JS, Kikkawa DO. Thyroid eye disease: from pathogenesis to targeted therapies. *Taiwan J Ophthalmol*. 2022;12:3−11. https://doi.org/10.4103/tjo.tjo_51_21.
3. Bahn RS. Graves' ophthalmopathy. *N Engl J Med*. 2010;362:726−738. https://doi.org/10.1056/NEJMra0905750.
4. Putterman AM, Urist MJ. Muller muscle-conjunctiva resection. Technique for treatment of blepharoptosis. *Arch Ophthalmol*. 1975;93:619−623. https://doi.org/10.1001/archopht.1975.01010020595007.
5. Boboridis K, Assi A, Indar A, et al. Repeatability and reproducibility of upper eyelid measurements. *Br J Ophthalmol*. 2001;85:99−101. https://doi.org/10.1136/bjo.85.1.99.
6. Nemet AY. Accuracy of marginal reflex distance measurements in eyelid surgery. *J Craniofac Surg*. 2015;26:e569−e571. https://doi.org/10.1097/SCS.0000000000001304.
7. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167−175. https://doi.org/10.1136/bjophthalmol-2018-313173.
8. Shao J, Huang X, Gao T, et al. Deep learning-based image analysis of eyelid morphology in thyroid-associated ophthalmopathy. *Quant Imaging Med Surg*. 2023;13:1592−1604. https://doi.org/10.21037/qims-22-551.
9. Park J, Yoon JS, Kim N, et al. Deep learning−driven exophthalmometry through facial photographs in thyroid eye disease. *Ophthalmol Sci*. 2025;5. https://doi.org/10.1016/j.xops.2025.100791.
10. Moon JH, Shin K, Lee GM, et al. Machine learning-assisted system using digital facial images to predict the clinical activity score in thyroid-associated orbitopathy. *Sci Rep*. 2022;12:22085. https://doi.org/10.1038/s41598-022-25887-8.
11. Shin K, Choung H, Lee MJ, et al. A preliminary evaluation of the diagnostic performance of a smartphone-based machine learning-assisted System for evaluation of clinical activity Score in digital images of thyroid-associated orbitopathy. *Thyroid*. 2024;34:744−752. https://doi.org/10.1089/thy.2023.0621.
12. Oh Y, Koh G, Kwak J, et al. TAOD-Net: automated detection and analysis of thyroid-associated orbitopathy in facial imagery. *Comput Ind Eng*. 2025;203:111024. https://doi.org/10.1016/j.cie.2025.111024.
13. ISO/IEC JTC 1/SC 37 Biometrics. *ISO/IEC TR 29794-5:2010 Information technology — Biometric sample quality — Part 5: Face image data*. Geneva, Switzerland: International Organization for Standardization; 2010.
14. Rufer F, Schroder A, Erb C. White-to-white corneal diameter: normal values in healthy humans obtained with the Orbscan II topography system. *Cornea*. 2005;24:259−261. https://doi.org/10.1097/01.ico.0000148312.01805.53.
15. Hashemi H, KhabazKhoob M, Yazdani K, et al. White-to-white corneal diameter in the Tehran Eye Study. *Cornea*. 2010;29:9−12. https://doi.org/10.1097/ICO.0b013e3181a9d0a9.
16. Kwak NH, Kim MS, Kim JH. A statistical study on the corneal diameter. *J Korea Ophth Soc*. 1988;29:245−249.
17. Martinez-de-la-Casa M, Matilla M, Garcia-Bella J, et al. Analysis of ocular biometric characteristics in a Mediterranean population. *Arch Soc Esp Oftalmol (Engl Ed)*. 2025;100:775−780. https://doi.org/10.1016/j.oftale.2025.10.009.
18. Lou L, Cao J, Wang Y, et al. Deep learning-based image analysis for automated measurement of eyelid morphology before and after blepharoptosis surgery. *Ann Med*. 2021;53:2278−2285. https://doi.org/10.1080/07853890.2021.2009127.
19. Nam Y, Song T, Lee J, Lee JK. Development of a neural network-based automated eyelid measurement system. *Sci Rep*. 2024;14:1202. https://doi.org/10.1038/s41598-024-51838-6.
20. Chen HC, Tzeng SS, Hsiao YC, et al. Smartphone-Based artificial intelligence-assisted prediction for eyelid measurements: algorithm development and observational validation Study. *Jmir Mhealth Uhealth*. 2021;9:e32444. https://doi.org/10.2196/32444.
21. Bodnar ZM, Neimkin M, Holds JB. Automated ptosis measurements from facial photographs. *JAMA Ophthalmol*. 2016;134:146−150. https://doi.org/10.1001/jamaophthalmol.2015.4614.
22. Li JO, Liu H, Ting DSJ, et al. Digital technology, telemedicine and artificial intelligence in ophthalmology: a global perspective. *Prog Retin Eye Res*. 2021;82:100900. https://doi.org/10.1016/j.preteyeres.2020.100900.
23. Mourits MP, Koornneef L, Wiersinga WM, et al. Clinical criteria for the assessment of disease activity in Graves' ophthalmopathy: a novel approach. *Br J Ophthalmol*. 1989;73:639−644. https://doi.org/10.1136/bjo.73.8.639.
24. Burch HB, Perros P, Bednarczuk T, et al. Management of thyroid eye disease: a consensus statement by the American thyroid Association and the European thyroid Association. *Thyroid*. 2022;32:1439−1470. https://doi.org/10.1089/thy.2022.0251.
25. Pinas D, R OBDK, Wubbels RJ, et al. Results of surgical correction of upper eyelid retraction in Graves' Orbitopathy. *Acta Ophthalmol*. 2021;99:e608−e613. https://doi.org/10.1111/aos.14622.
26. Halsoy K, Hope AM, Foshaug C, et al. Long-term outcome after correction of lower eyelid retraction with donor sclera in thyroid eye disease. *Plast Reconstr Surg Glob Open*. 2026;14:e7455. https://doi.org/10.1097/GOX.0000000000007455.