Digital Biomarkers

Research Reports – Research Article

Digit Biomark 2023;7:63–73
DOI: 10.1159/000531224

Received: January 25, 2023
Accepted: May 16, 2023
Published online: July 28, 2023

Development and Assessment of an Artificial Intelligence-Based Tool for Ptosis Measurement in Adult Myasthenia Gravis Patients Using Selfie Video Clips Recorded on Smartphones

Meelis Lootus[a] Lulu Beatson[a] Lucas Atwood[a] Theo Bourdais[a] Sandra Steyaert[b] Chethan Sarabu[a] Zeenia Framroze[a] Harriet Dickinson[c] Jean-Christophe Steels[d] Emily Lewis[d] Nirav R. Shah[e] Francesca Rinaldo[a]

[a]Sharecare Inc., Atlanta, GA, USA; [b]Stanford University, Center for Bioinformatics Research, Palo Alto, CA, USA; [c]UCB Pharma, Slough, UK; [d]UCB Pharma, Brussels, Belgium; [e]Clinical Excellence Research Center, Stanford University, Palo Alto, CA, USA

Keywords
Decentralized study · Ptosis · Computer vision · Personal smartphones

Abstract

Introduction: Myasthenia gravis (MG) is a rare autoimmune disease characterized by muscle weakness and fatigue. Ptosis (eyelid drooping) occurs due to fatigue of the muscles for eyelid elevation and is one symptom widely used by patients and healthcare providers to track progression of the disease. Margin reflex distance 1 (MRD1) is an accepted clinical measure of ptosis and is typically assessed using a hand-held ruler. In this work, we develop an AI model that enables automated measurement of MRD1 in self-recorded video clips collected using patient smartphones. ***Methods:*** A 3-month prospective observational study collected a dataset of video clips from patients with MG. Study participants were asked to perform an eyelid fatigability exercise to elicit ptosis while filming "selfie" videos on their smartphones. These images were collected in nonclinical settings, with no in-person training. The dataset was annotated by non-clinicians for (1) eye landmarks to establish ground truth MRD1 and (2) the quality of the video frames. The ground truth MRD1 (in millimeters, mm) was calculated from eye landmark annotations in the video frames using a standard conversion factor, the horizontal visible iris diameter of the human eye. To develop the model, we trained a neural network for eye landmark detection consisting of a ResNet50 backbone plus two dense layers of 78 dimensions on publicly available datasets. Only the ResNet50 backbone was used, discarding the last two layers. The embeddings from the ResNet50 were used as features for a support vector regressor (SVR) using a linear kernel, for regression to MRD1, in mm. The SVR was trained on data collected remotely from MG patients in the prospective study, split into training and development folds. The model's performance for MRD1 estimation was evaluated on a separate test fold from the study dataset. ***Results:*** On the full test fold ($N = 664$ images), the correlation between the ground truth and predicted MRD1 values was strong ($r = 0.732$). The mean absolute error was 0.822 mm; the mean of differences was −0.256 mm; and 95% limits of agreement (LOA) were −0.214–1.768 mm. Model performance showed no improvement when test

karger@karger.com
www.karger.com/dib

Correspondence to:
Francesca Rinaldo, francesca.rinaldo@sharecare.com

data were gated to exclude "poor" quality images. ***Conclusions:*** On data generated under highly challenging real-world conditions from a variety of different smartphone devices, the model predicts MRD1 with a strong correlation (r = 0.732) between ground truth and predicted MRD1.

Introduction

Myasthenia gravis (MG) is a rare autoimmune disease that is characterized by muscle weakness and fatigue [1]. The estimated prevalence of MG is 20 per 100,000 in the USA [2]. For most MG patients, autoantibodies against the acetylcholine receptor (AChR) at the neuromuscular junction interfere with neural transmission [3]. Thus, weakness and fatigue worsen with sustained or repetitive muscle contractions.

Most patients (80–85%) who present with ocular symptoms progress to generalized MG [4, 5]. Ocular symptoms of MG manifest when the muscles that control eye movement and eyelid elevation are impacted, affecting one or both eyes. Physiologically, these muscles are susceptible due to their high frequency of synaptic firing [3]. If the muscles controlling eyelid elevation – primarily the superior palpebral levator muscle – fatigue, then eyelid drooping (known as blepharoptosis or ptosis) occurs [6]. This symptom can be elicited clinically using the "eyelid fatigability" test [7, 8]. This noninvasive exercise consists of asking the patient to maintain a prolonged upward gaze before returning to primary gaze (i.e., when the eyes are looking straight ahead), and the resulting change in eyelid positioning is then observed. Traditionally, ptosis has been quantified using the margin reflex distance (MRD), which is measured using a hand-held ruler [9]. To measure the MRD in the clinic, a clinician may direct a pen light at the patient's eye while they are in primary gaze. Margin reflex distance 1 (MRD1) measures the distance in millimeters (mm) between the corneal light reflex and the center of the upper eyelid margin. A normal MRD1 is 4–5 mm, and the difference in MRD1 between a normal and ptotic eyelid is the degree of ptosis [10].

While MRD1 is often assessed by expert clinicians, it can be difficult for nonspecialists to measure reliably and is limited by the presentation of the patient at the time of the evaluation (it does not capture the fluctuation of symptoms over time). This can be particularly true for MG patients, as symptoms may not always be present during an in-person evaluation. Indeed, Guterman et al. [11] describe a case study where diagnosis of seronegative ocular MG was enabled by asking the patient to take daily smartphone selfies to document fluctuating ptosis and eye misalignment.

Image analysis software for measurement of eyelid positioning from patient photographs has been previously developed [12–14]. Bodner et al. [14] reported development of a software algorithm to automate measurement of MRD1 from digital photographs of patients taken in an oculoplastic surgery clinic using a standardized protocol. Several groups have recently reported AI algorithms for recognition of eyelid positioning, specifically the measurement of MRD1 and palpebral fissure height [15–22]. Notably, these studies report the use of datasets acquired by taking patient photographs using commercially available cameras or a smartphone camera on a single device [16, 17], under highly standardized and optimized conditions. Furthermore, these models require the presence of a measurement standard in the photographs (or they assume the average size of the human eye as a measurement standard) and require manual data processing of eyelid morphology prior to data input into the model.

In this paper, we report the development of a patient-centered algorithm to measure MRD1 from "selfie" videos of MG patients taken using their own smartphones. These real-world conditions for remote data collection bring challenges such as variations in smartphone device and camera quality, patient positioning, camera angle, and lighting. To our knowledge, this is the first algorithm developed and tested for detecting a measure of ptosis on data contributed directly by patients using smartphones and demonstrates the feasibility of assessing MRD1 using automated image analysis on remotely collected video data recorded before and after an eyelid fatigue exercise.

Materials and Methods

Study Design

We conducted a prospective, observational study between October 2020 and July 2021 using fully decentralized methods (ClinicalTrials.gov identifier: NCT04590716). Participants were recruited from across the USA, using ads and social media campaigns to direct interested individuals to a landing page and screening tool for the eligibility criteria. Individuals who completed the screening and were approved to join the study did so by downloading the study mobile phone application (the study "app," available on iOS and Android) and completing a CFR Part 11 compliant eConsent and signature process. Inclusion criteria were a documented diagnosis of MG; ocular or bulbar (speech) symptoms; age 18+; resident in the USA for the duration of the study;

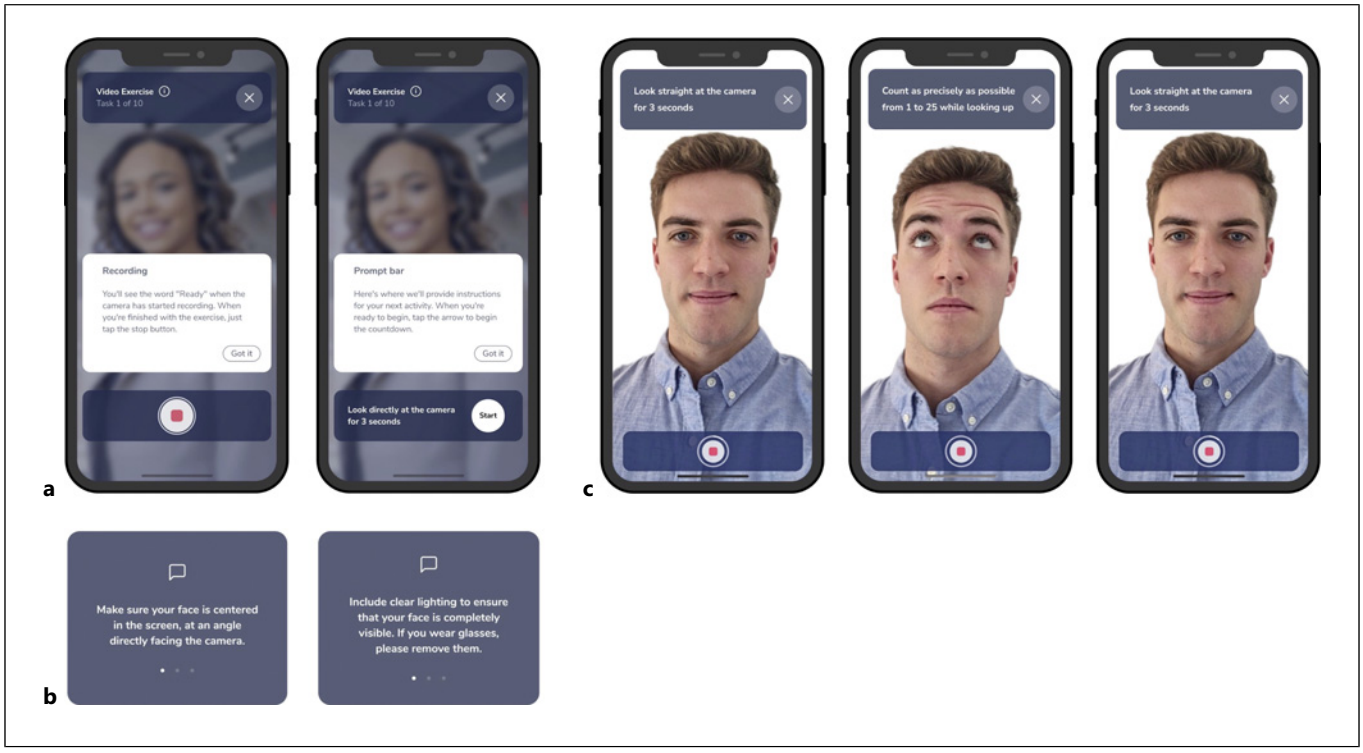


Fig. 1. Video prompts were used to allow participants to conduct the lid fatigability test using their own smartphones. **a** An in-app interface was created for conducting the twice-weekly video exercises. Participants could see prompts at the bottom of the screen and use the interface to record themselves performing the eyelid fatigability exercise. **b** Prior to the start of the video exercise, participants were prompted to center their faces on the screen, move to a well-lit area, and remove their eyeglasses. **c** Using the app interface, the video exercise was administered in three steps. First, the participant was instructed to gaze directly at the smartphone camera for 3 s in order to capture a baseline image ("before" the exercise). Next, the participant was instructed to maintain an upward gaze while counting from 1 to 25 (thus fatiguing the eyelid elevation muscles impacted by MG "during" the applied strain). Finally, the participant was instructed to gaze directly back into the smartphone camera for 3 s, allowing capture of images to assess the level of ptosis elicited ("after" the exercise). The images shown are of a member of the study team and are used with his permission. The images are not from a participant enrolled in the study.

ability to read, understand, and write in English; possession of an internet-enabled smartphone capable of supporting the research app. Participants were compensated with a USD 250 Amazon gift card if they completed the study with an overall adherence to study tasks of 60% or more. Participants who enrolled were asked to confirm their MG diagnosis by uploading the appropriate documentation via the study app. The study was reviewed and approved by a central institutional review board (Salus IRB), protocol number DOC-005-2020.

Selfie Video Collection from Participant Smartphones

Over 3 months, participants used the study app to complete twice-weekly exercises where they recorded selfie video clips (Fig. 1). Prior to completing the video exercises, participants were reminded to go to a well-lit area, remove their eyeglasses, and center their faces on the smartphone screen. To elicit ptosis, participants were prompted to complete the eyelid fatigability test by (1) gazing directly into their smartphone camera for 3 s, (2) sustaining an upward gaze while counting from 1 to 25, and (3) gazing directly back into the camera for 3 s. This task was chosen specifically because the sustained upward gaze fatigues the muscles of eyelid elevation and causes ptosis specifically in MG patients.

Annotation and Splitting of Study Data

To establish a labeled ("ground truth") dataset for model evaluation, researchers added manual annotations to video data collected in the study. All annotations were produced with Amazon SageMaker Ground Truth and VGG VIA annotation tools (https://www.robots.ox.ac.uk/~vgg/software/via/). These annotations included (1) landmarks on the eyelid, iris, and pupil (Fig. 2a); (2) image/video quality (Fig. 2b); and (3) the presence of other factors that could adversely affect landmark placement, such as eyeglasses (not shown). The eye landmarks were used to compute the ground truth MRD1, in pixels. The diameter of the iris (in pixels) was used to normalize the MRD1 value. This value was then converted to mm by multiplying by a standard conversion factor, the horizontal visible iris diameter of the human eye (11.8 mm) [22, 23].

To train and ensure fair evaluation on subsets that represented the MG study dataset, we split the data into training, development, and test folds so that (1) no study participants were shared between

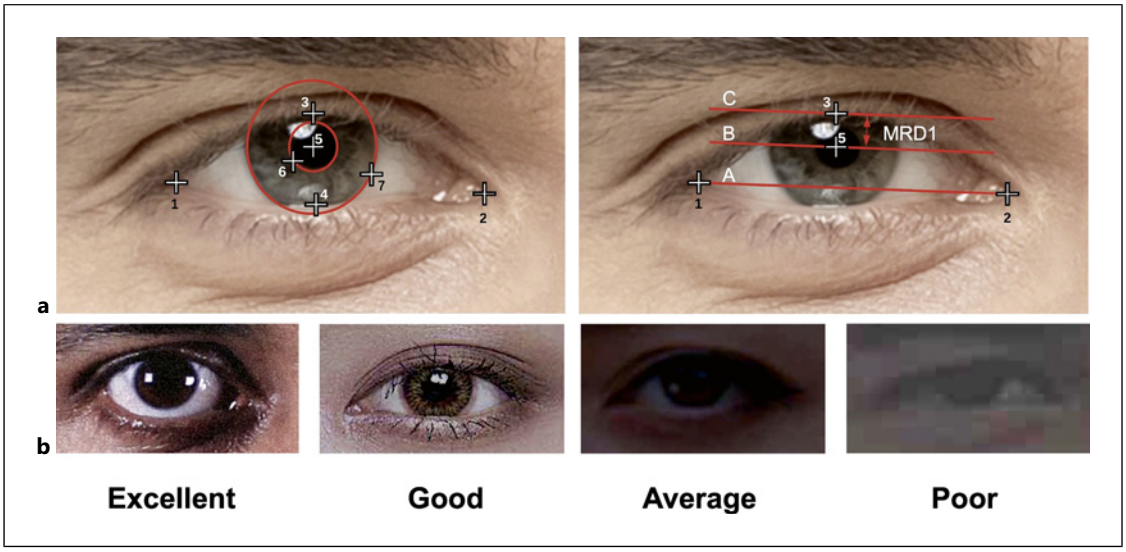


Fig. 2. Manual annotation of study data to create a labeled dataset for model evaluation. The annotations were created using custom-configured Amazon SageMaker Ground Truth and VGG VIA annotation tools. Panel **a** shows representative annotations for landmarks on the upper and lower eyelids, iris, and pupil. Key landmarks annotated were (1) the lateral canthus (corner) of the eye, (2) the medial canthus of the eye, (3) the center of the upper eyelid margin, (4) the center of the lower eyelid margin, (5) the center of the pupil, (6) the edge of the pupil, and (7) the edge of the iris. Ellipses for the pupil and iris circumferences were generated from landmarks 5–7 (shown in red). To calculate the ground truth MRD1, horizontal lines were established between the medial and lateral canthi (horizontal A) and through the points at the center of the pupil and at the upper eyelid margin (horizontals B and C, respectively). Horizontals B and C were established as parallel to horizontal A. The distance between horizontals B and C was the MRD1, in pixels. The diameter of the iris was calculated by taking the distance between landmarks 5 and 7 as the radius of the iris (in pixels); this value was used to normalize the MRD1 value. The resulting normalized value was then converted to mm by multiplying by a standard conversion factor, in this case the horizontal visible diameter of the human iris (11.8 mm) [22, 23]. Panel **b** shows representative annotations for image quality across 4 categories (excellent, good, average, and poor). The images shown in this figure are publicly available and are not from participants enrolled in the clinical study.

the splits; (2) there was a similar distribution of age and sex in each split; (3) there was similar MRD1 distribution in each split. Images of "unusable" quality (in which the participant's face was not in the video frame or images were too dark or blurred to recognize facial landmarks) and where the participants wore eyeglasses were excluded from the folds.

Model Training

We trained a neural network for eye landmark detection consisting of a ResNet50 backbone [24] plus two dense layers of 78 dimensions. Only the ResNet50 backbone was used, discarding the last two layers. The embeddings from the ResNet50 were used as features for a support vector regressor (SVR) using a linear kernel [25], for regression to MRD1, in mm. This approach was selected after exploring an early iteration of the model where the outputs of the two dense layers were used to calculate MRD1 based on landmark locations, using elementary geometry (deriving the features directly). However, the accuracy with this architecture was limited, as shown in online supplementary Figure 1 (for all online suppl. material, see https://doi.org/10.1159/000531224). Training of the ResNet50 backbone and the linear regressor are described in more detail below.

The ResNet50 eye landmark detector was trained on a 5.6-million-frame dataset obtained by combining the following publicly available datasets: TEyeD [26], BUHMAP (Boğaziçi University Head Motion Analysis Project Database) [27], MUCT ("Milborrow/University of Cape Town") [28], Remote [29]. A random subset of 500 images from TEyeD was used for validation during the training process. A summary of the datasets used for training the eye landmark detector model is given in Table 1. The public datasets differed in terms of their annotation: while TEyeD had the full complement of landmarks used in our model, the other datasets had only a subset of these landmarks. To benefit maximally from both the densely annotated TEyeD as well as the less densely annotated MUCT, BUHMAP, and Remote datasets, we trained the ResNet50 using a custom loss function described in further detail in online supplementary information. The loss function penalizes L2 error on landmarks and the difference between the ellipse parameters of the ground truth and model iris and pupil ellipses. The loss was minimized using the Adam optimizer [30]. During model training, the following operations were performed to augment the dataset and improve generalization: (1) random rotation up to +/− 45°, (2) random brightness alteration +\− 20%, (3) random contrast alteration [50%, 200%], and (4) random JPEG quality adjustment [35%, 100%].

Table 1. Summary of the public datasets used to create the training dataset for eye landmark detection

Dataset name	No. of videos	No. of frames	No. of individuals represented	Pupil landmark style	Iris landmark style	No. of Eyelid landmark points per eye	Key variation between video frames
TEyeD	413	5,648,924	39	Ellipse	Ellipse	34	Eyeball and eyelid movement
MUCT	–	3,755	751	Center point	Center point	8	Camera angle
BUHMAP	–	2,800	11	–	–	8	Facial expression, camera angle
Remote	–	445	2	Ellipse	Ellipse	–	Exaggerated facial expressions
Total	413	5,655,924	803				

Using the training and development data folds from the study, we trained a linear regression (we used an SVR algorithm) to predict MRD1 directly. That enabled us to swiftly experiment on the data gathered remotely on participants' smartphones. The SVR used a linear kernel, with hyperparameter values of epsilon = 0.1, C = 0.0001 after tuning was performed on the development fold. Notably, the study data used for training of the linear regressor are different from the data available in the public datasets in two key ways. First, there is a different distribution of ground truth MRD1 values in the study data as compared to the public datasets (online suppl. Fig. 2). This is expected, as MG patients should exhibit smaller MRD1 values (consistent with the presence of ptosis). For the public data, the distribution of MRD1 lies overwhelmingly between 3.5 and 5.5 mm, and there are very few values below 3.5 mm. For the study dataset, the distribution is wider, including many more MRD1 values between 0 and 3.5 mm. In addition, the study dataset reflects a variety of conditions, including different smartphone devices, and variation in user- and environmental-related factors, such as lighting, distance from the smartphone camera, and camera angle.

Model Inference Pipeline

The three-step model inference pipeline is shown in Figure 3. The model takes as input a selfie video clip and generates as output an MRD1 measure, in millimeters. Step 1 leverages a multitask cascaded convolutional neural network for facial feature recognition. It takes as input the first frame from the video clip containing the user's entire face and produces two cropped images, one for each eye. In step 2, a headless ResNet50 architecture is used to embed the eye image into a 2,048-dimensional feature space. Finally, in step 3, regression from the embedding to MRD1 is performed using an SVR.

Model Performance Evaluation

The model's performance was evaluated on the testing subset of the study data. Model performance is reported in terms of mean absolute error (MAE), standard deviation (SD), absolute standard deviation (ASD), and mean squared error (MSE) for MRD1 measurement, in mm. In addition, the Pearson correlation (product moment correlation coefficient) and Bland-Altman analyses were selected for their ability to assess correlation and agreement between model predictions and the ground truth, respectively. Bland-Altman metrics reported are mean of differences (MD), 95% limits of agreement (95% LOA), and the percent of predictions with MD > 2 mm. PCA analysis [31] was performed on the embeddings of the SVR training set to assess collinearity in the embeddings. The number of PCA components was set to 1,063, equal to the number of training samples, and the cumulative explained variance was plotted against the number of principal comments. All statistical analyses were conducted in Python (version 3.8.2).

Results

Study Population Characteristics

The study enrolled 113 participants; 73% (N = 82) completed the full study course (i.e., were not withdrawn). Participants completing the study were in 33 US states, their mean age was 54.9 (range 22–77 years, SD 13.6), and 58.5% were female. Represented racial and ethnic groups were 89% white, 4% black or African American, 2% Hispanic or Latinx, 1% Asian, 1% other, 3% no response.

Study Dataset Characteristics

A total of 909 video recording sessions were collected from 80 individuals across 68 different smartphone devices using iOS (55%) and Android (45%) operating systems. In each video recording session, three separate "clips" were recorded before, during, and after the eyelid fatigability exercise protocol (Fig. 1c). From these, we utilized 1,813 video clips (from before and after the eyelid fatigability exercise) for labeling. The first frame was extracted from each of these clips, and images of both eyes were cropped using a multitask cascaded convolutional neural network detector, which produced 3,626 individual eye images. Across the whole dataset, there

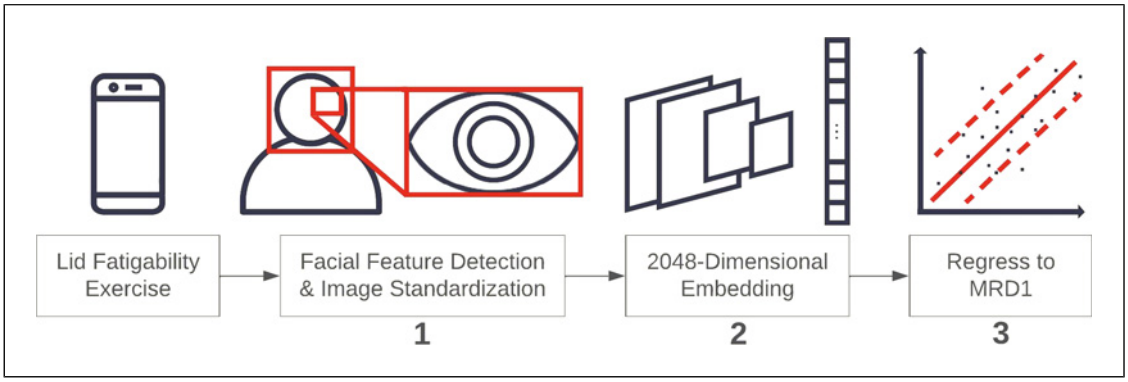


Fig. 3. MRD1 prediction model inference pipeline. The model pipeline consists of three fully automated steps. In step 1, an input video frame is captured, the face and eyes are detected in the frame, and the frame is cropped to isolate the left and right eyes. During step 2, a 2,048-dimensional feature embedding is produced and then ingested by the SVR to predict MRD1 in step 3.

were 170 unusable (4.7%), 716 poor (19.7%), 2,074 average (57.2%), 610 good (16.8%), and 56 perfect quality (1.5%) images (Fig. 2b).

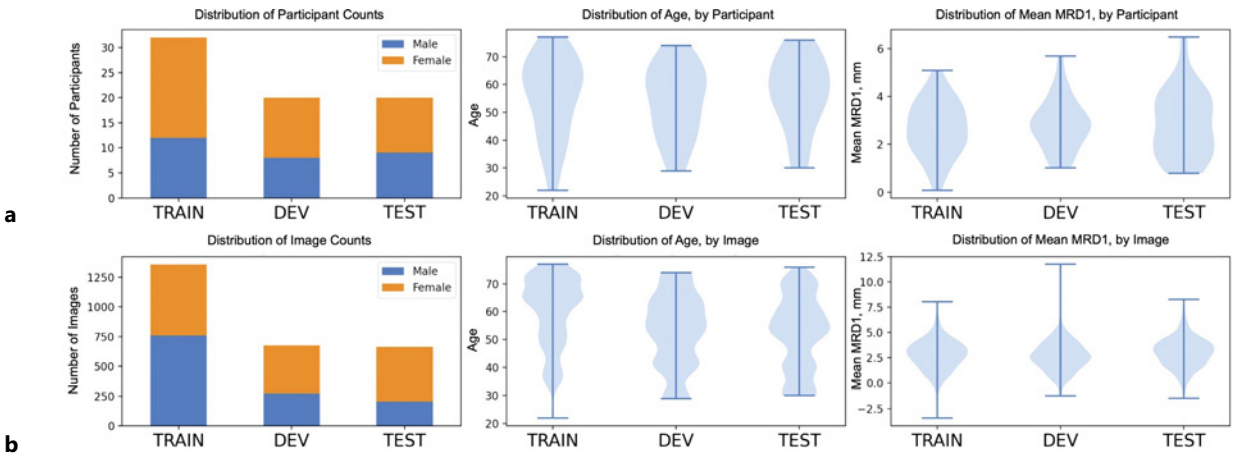
Characteristics of the training, development, and testing folds are shown in Figure 4. Each set achieved a comparable distribution of participants by age, sex, and ground truth MRD1 values. No participants were shared between the sets. The test set (N = 664 images) contained 83 poor (12.5%), 326 average (49.0%), 221 good (33.2%), and 34 perfect (5.1%) quality images. Characteristics of the test set gated for data quality are shown in Figure 4c.

Model Performance for MRD1 Measurement

Figure 5 summarizes the model's performance on the testing fold. For the full testing fold ("poor" to "perfect" quality data, N = 664 images), comparison of the model predictions with ground truth MRD1 values demonstrated a strong correlation (r = 0.736) (Fig. 5a). Bland-Altman analysis demonstrated MD of 0.256 mm, with 95% LOA of −0.214–1.768 mm (Fig. 5b). MAE and MSE were 0.822 mm and 1.133 mm, respectively (Fig. 5c). For the testing fold gated for data quality (excluding "poor" quality data, N = 581 images), performance metrics were r = 0.732, MAE 0.793, MSE 1.044. Bland-Altman MD was 0.214 mm, with 95% LOA of 0.058–1.726 mm. PCA analysis of the embeddings on the training dataset showed that the first principal component covers 19.7%, the first 10 cover 72.3%, the first 100 cover 94.7%, and the first 200 cover 97.8% of the variation (online suppl. Fig. 3). As expected, this suggests that there may be collinearity between features in the data. In addition, as shown in Figure 5, there is some bias in the model output as it overestimates low MRD1 values and underestimates high MRD1 values. We computed the lines between the predicted and ground truth values, where the slopes were 0.4999 and 0.5076 and the intercepts were 1.2524 and 1.2552 for the full and limited test sets, respectively. This suggests that the model tends to underestimate the predicted values compared to the ground truth values. The high intercept values indicate that even for small ground truth values, the predicted values are relatively high.

Impact of Demographic Features on Model Performance

To evaluate model robustness, we investigated the impact of demographic variables on model performance. We measured performance metrics on three different subsets of the test dataset (online suppl. Table 1): females (510 images), individuals older than 35 (556 images), and white participants (461 images). The performance did not vary significantly, with the following MAE values observed for each test subset: full test set (664 images): 0.822 mm, white participants: 0.906 mm, female participants: 0.788 mm, and participants over 35: 0.865 mm. To further explore the generalizability of the model, we retrained the model on images from a subpopulation of white participants only (2,106 images, representing 94% of our entire dataset) and tested the model on a subpopulation of non-white participants (306 images, representing 4% of the dataset). The resulting performance metrics were MAE 0.562, SD 0.744, and ASD 0.487. Bland-Altman MD was −0.003, and MD >2 mm = 0.013. For comparison, the MAE of the model trained and tested on the full dataset was 0.822 mm (Fig. 5). Taken together, the results



c

Data Fold	Participant Count	Image Count	Mean Age, Participants (Std Dev)	Mean Age, Images (Std Dev)	% Female, Participants	% Female, Images	Mean MRD1 (mm), by Participants (Std Dev)	Mean MRD1 (mm), by Images (Std Dev)
TRAIN	33	1,383	55.3 (14.6)	60 (12.7)	60.6%	44.0%	2.66 (1.16)	2.78 (1.51)
DEV	22	685	53.9 (13.0)	53.1 (13.3)	54.5%	59.4%	2.85 (1.11)	2.66 (1.48)
TEST	20	664	56.5 (13.3)	52.1 (13.3)	55%	69.4%	2.85 (1.45)	3.02 (1.52)
Reduced TEST	20	581	56.5 (13.3)	51.6 (13.6)	55%	67.5%	2.84 (1.43)	2.98 (1.47)

Fig. 4. Characteristics of training, development, and test data folds. Each plot in the figure shows the distribution, by fold (training [TRAIN], development [DEV], test [TEST]), of a property (count, age, sex, mean MRD1) of the dataset. Panel **a** shows the characteristics for participants assigned to each fold. Panel **b** shows characteristics of the images in each fold (assigned according to the characteristics of the image's subject). The distributions in panel **b** differ from those in panel **a** because of the difference in the number of images contributed by each participant. Panel **c** shows the summary statistics for each fold. The properties (age, sex, mean MRD1) are similarly distributed between the folds. The proportion of participants in each fold is similar to the proportion of images in each fold – suggesting that the mean number of images contributed by each participant is similar in each fold. The last row of the table also shows the characteristics of the test set, gated for data quality to exclude 83 samples labeled as "poor" image quality (reduced TEST).

of these experiments suggest that on our dataset, there was limited impact of selected demographic variables on model performance.

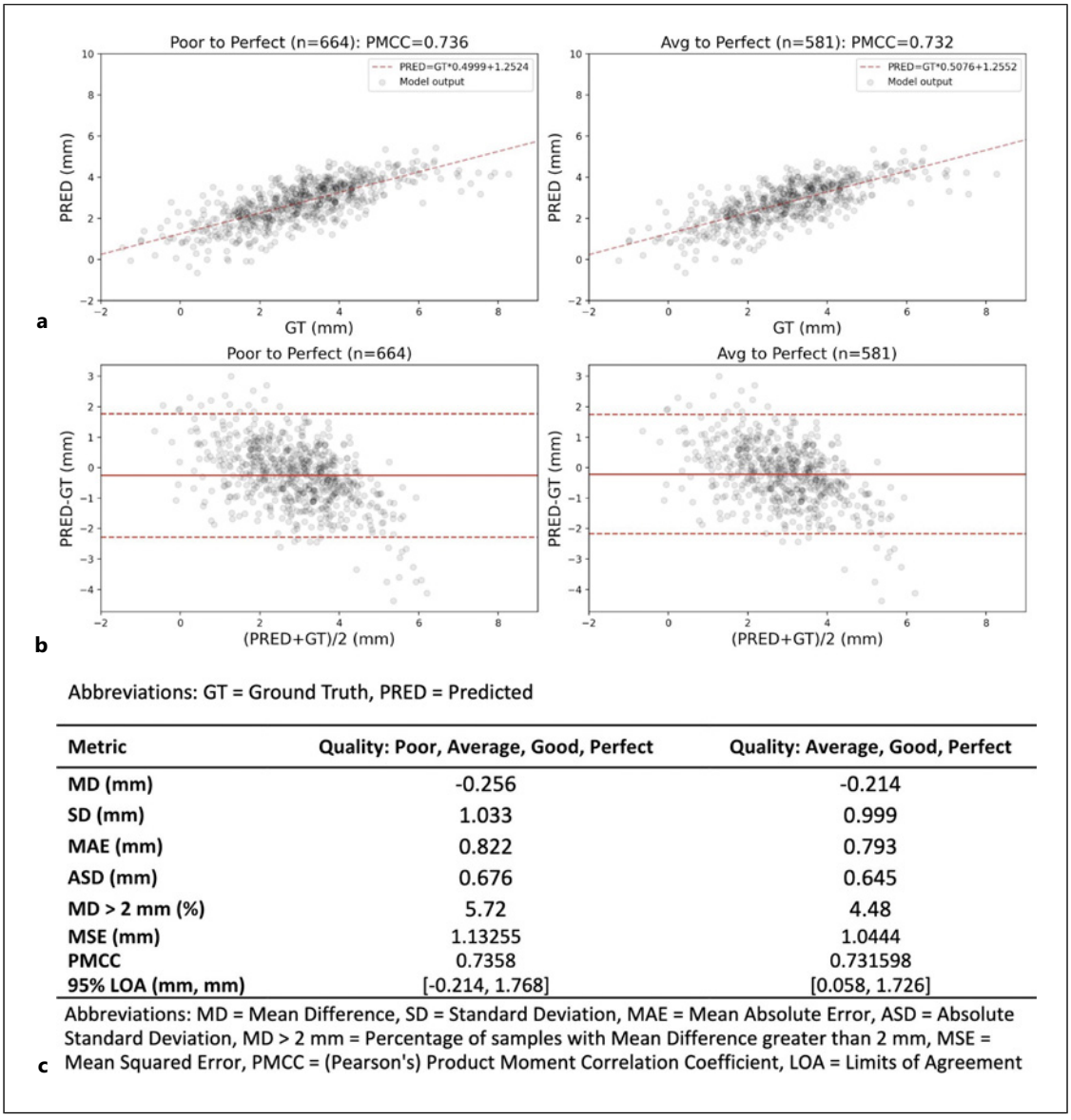
Impact of Participant Data Contribution on Model Performance

Online supplementary Figure 4 shows the distribution of the number of video samples contributed per participant in the training dataset, indicating that approximately half of the participants assigned to the training dataset contributed the majority of the data used to train the linear regressor (SVR). To evaluate potential bias due to an unbalanced training set, we trained a new version of the SVR on an equal number of images (two) per participant. The resulting performance metrics were MAE 0.935, SD 1.162, and ASD 0.758. Bland-Altman MD was −0.312, and MD >2 mm = 0.098. This suggests that the unbalanced training set may have had some impact on model performance; however, further experiments are required as the balanced training set is likely too small to allow for generalization to previously unseen data.

Discussion

Key Contributions and Comparison with Prior Work

Prior work has demonstrated the feasibility of producing automated ptosis measures using AI models. However, these studies utilize high-quality datasets obtained under tightly controlled conditions which leverage digital cameras, specially trained technicians in oculoplastic and ophthalmology clinics, and images obtained from eye



Abbreviations: GT = Ground Truth, PRED = Predicted

Metric	Quality: Poor, Average, Good, Perfect	Quality: Average, Good, Perfect
MD (mm)	-0.256	-0.214
SD (mm)	1.033	0.999
MAE (mm)	0.822	0.793
ASD (mm)	0.676	0.645
MD > 2 mm (%)	5.72	4.48
MSE (mm)	1.13255	1.0444
PMCC	0.7358	0.731598
95% LOA (mm, mm)	[-0.214, 1.768]	[0.058, 1.726]

c Abbreviations: MD = Mean Difference, SD = Standard Deviation, MAE = Mean Absolute Error, ASD = Absolute Standard Deviation, MD > 2 mm = Percentage of samples with Mean Difference greater than 2 mm, MSE = Mean Squared Error, PMCC = (Pearson's) Product Moment Correlation Coefficient, LOA = Limits of Agreement

Fig. 5. Model performance for MRD1 prediction on the test fold of the study dataset. Panel **a** shows the predicted versus ground truth MRD1 for the full test fold and for the test fold gated for data quality (excluding "poor" quality images). Panel **b** shows Bland-Altman plots indicating the mean of differences (MD, solid line) and the limits of agreement (LOA, dotted lines) for MRD1 predictions and ground truth on the full test fold and the gated test fold. Panel **c** summarizes the relevant performance metrics for the model on each test dataset. Note that there are some negative MRD1 values in the dataset. Negative MRD1 values can occur in cases of severe ptosis, where the center of the iris is concealed by the drooping upper eyelid margin. In these cases, the center of the pupil can be inferred by the visible amount of boundary of the iris.

image atlases [13, 16, 17, 19]. Other groups have tested the performance of models on data obtained from smartphones, though only a single device was used [16, 17]. In contrast to prior work, our model's performance was evaluated on a real-world dataset acquired remotely by asking study participants with MG to record selfie videos on their smartphones. While Qin et al. [18] also describe a similar approach, they do not provide any results to support the validation of model performance under real-world conditions. The samples collected in our study represent the challenging and hugely variable conditions present in real-world data. Our model predicts MRD1

within the context of variations in focal distance, lighting, background, and image quality. Notably, the study included 60 devices, many of which present variations in camera quality.

Furthermore, the images used for model evaluation did not include placement of a scale or object of known size during image acquisition (such as a paper ruler on the participant's face). The model performs well considering the lack of a reference standard in the images, demonstrating the potential for further development of an autonomous MRD1 prediction model using data collected remotely. We chose to use the visible iris diameter as a conversion metric, which has been used in similar work by others [22]. In our exploration of data quality's impact on model performance, we found no increase in predictive accuracy on images of "average" quality or better (Fig. 5). Despite variability in data quality, the performance of our model on this challenging dataset demonstrates a promising solution for remote MRD1 measures.

Limitations

Our decentralized approach to data collection offers strengths and limitations. Enabling the patient to self-capture images can result in a more patient-friendly tool, which can be used by participants at times more suitable to them. On the other hand, the recording conditions lack the control afforded under laboratory or standardized conditions, and the domain presents high variability in the quality of devices involved and the conditions under which the video recordings take place. In addition to device quality implications, patients may not adhere to the eyelid fatigability exercise. In future work, the model pipeline could be further refined by adding controls for the automated estimation of video quality during the measurements using real-time detection of lighting, camera angle and rotation, or eye gaze direction.

Principal component analysis demonstrated that there may be collinearity between some of the features in the data and thus high correlation between some features (online suppl. Fig. 3). In addition, the current model does have some bias, with a tendency to underestimate high MRD1 values and overestimate low MRD1 values. One of the primary challenges for developing the model was the limited availability of data from MG patients (a rare disease), who exhibit the pathology of interest (ptosis). To address the lack of available data for model development, we leveraged public datasets to assemble a large dataset of eye images that was used to train the ResNet50 backbone (Table 1). However, this dataset contained primarily samples with corresponding MRD1 values that were in the normal range (4–5 mm in healthy adults). In contrast, the dataset collected in our study and used to train the linear regressor (SVR) had only a limited set of samples from participants with MG, with increased representation of lower MRD1 values due to the fact that MG patients exhibit ptosis (online suppl. Fig. 2). While the observed bias was significantly reduced from the first iteration of the model (online suppl. Fig. 1) through the use of a linear regressor, some bias still remains, and we plan to address this in future work. First, this could be addressed by implementing more stringent quality controls for participant data collection, as discussed above. In addition, we plan to conduct additional future studies with a larger number of patients in order to gather more data for model training and validation that is representative of the intended clinical context for patients with MG.

We also examined the impact of demographic features such as sex, age, and race on model performance. Testing the model on gated subsets of the test dataset reflecting these variables did not significantly impact model performance. Similarly, training on data only from white participants and testing the model on a test dataset of non-white participants also indicates that the model may be able to generalize more broadly, even though the diversity of participants of our current dataset is limited. Future work will aim to further examine model generalizability by enriching the diversity of participants represented in new data collected for model training and refinement.

Conclusion

Our work demonstrates the feasibility of automated ptosis assessment from frames of video data collected remotely over a broad range of smartphones. The model developed holds promise as a patient-centric tool for objective, remote measurement of this important MG symptom.

Acknowledgments

The authors would like to thank the Sharecare Smart Omix team, who developed the research app with which the study was conducted. This includes Scott Cressman, Thomas Noulelis, Yuri Sabach, Steven Chand, Daniel Dresser, Jeffrey Kim, Gilbert Kwan, Michael Lay, Anton Bielousov, Rohan Jahagirdar, Aria Vaghefmoghaddam, Savita Singha, Vijay Sivaji, and Nicholas Bernhardt-Lanier. We would like to thank Gabriel Zaccak and Akio Yoshimoto for critical reading of the manuscript. Finally, we would also like to express our deep thanks to all patients who participated in the study.

Statement of Ethics

This study protocol was reviewed and approved by Salus IRB, protocol number DOC-005-2020. Written informed consent was obtained from all participants prior to their enrollment and participation in the study. In addition, written informed consent was obtained from a member of a study team to use the images shown in Figure 1.

Conflict of Interest Statement

M.L., T.B., L.B., S.S., and N.R.S. are former employees and stockholders of Sharecare, Inc. Z.F., C.S., L.A., and F.R. are employees and shareholders of Sharecare, Inc. J.C.S. and E.L. are employees and shareholders of UCB Pharma. H.D. is a former employee and shareholder in UCB Pharma.

Funding Sources

The study was funded by UCB Pharma and conducted in collaboration with Sharecare, Inc. UCB consulted with Sharecare on the study objectives and design. UCB was not involved in the data collection, model development, or evaluation.

Author Contributions

All authors provided critical revisions of the manuscript and final approval of the version to be published. In addition to this, each individual author contributed as listed. M.L., T.B., L.B., L.A., and F.R. contributed to study conceptualization, design, data acquisition and annotation, model development and evaluation, as well as drafting and revisions of the manuscript. S.S., C.S., N.R.S., and Z.F. contributed to study conceptualization, design, management, and manuscript revisions. H.D., E.L., and J.C.S. contributed to study conceptualization and manuscript drafting and revision.

Data Availability Statement

The datasets to support model development are publicly available, as listed in Table 1. The data collected during the decentralized study to support the model training and evaluation are available from Sharecare, but due to the sensitive and highly identifiable nature of the selfie videos, these cannot be shared. Further inquiries can be directed to the corresponding author.

References

1 Catalin J, Silviana J, Claudia B. Clinical presentation of myasthenia gravis. Thymus; 2020 May 24.
2 Phillips LH. The epidemiology of myasthenia gravis. Semin Neurol. 2004 Mar;24(1):17–20.
3 Nair AG, Patil-Chhablani P, Venkatramani DV, Gandhi RA. Ocular myasthenia gravis: a review. Indian J Ophthalmol. 2014 Oct;62(10):985–91.
4 Wang L, Zhang Y, He M. Clinical predictors for the prognosis of myasthenia gravis. BMC Neurol. 2017 Apr;17(1):77.
5 Conti-Fine BM, Milani M, Kaminski HJ. Myasthenia gravis: past, present, and future. J Clin Invest. 2006 Nov;116(11):2843–54.
6 Oyster CW. The human eye: structure and function. Sunderland, MA: Sinauer Associates; 2006.
7 Toyka KV. Ptosis in myasthenia gravis: extended fatigue and recovery bedside test. Neurology. 2006 Oct;67(8):1524.
8 Yoganathan K, Stevenson A, Tahir A, Sadler R, Radunovic A, Malek N. Bedside and laboratory diagnostic testing in myasthenia. J Neurol. 2022 Jun;269(6):3372–84.
9 Putterman AM. Margin reflex distance (MRD) 1, 2, and 3. Ophthal Plast Reconstr Surg. 2012 Jul;28(4):308–11.
10 Sruthi R, Pauly M. Ptosis: evaluation and management. Kerala J Ophthalmol. 2019 Jan;31(1):11.
11 Guterman EL, Botelho JV, Horton JC. Diagnosis of tensilon-negative ocular myasthenia gravis by daily selfie. J Neuro Ophthalmol. 2016 Sep;36(3):292–3.
12 Coombes AG, Sethi CS, Kirkpatrick WN, Waterhouse N, Kelly MH, Joshi N. A standardized digital photography system with computerized eyelid measurement analysis. Plast Reconstr Surg. 2007 Sep;120(3):647–56.
13 Chun YS, Park HH, Park IK, Moon NJ, Park SJ, Lee JK. Topographic analysis of eyelid position using digital image processing software. Acta Ophthalmol. 2017 Nov;95(7):e625–32.
14 Bodnar ZM, Neimkin M, Holds JB. Automated ptosis measurements from facial photographs. JAMA Ophthalmol. 2016 Feb;134(2):146–50.
15 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 Oct;8(10):e3089.
16 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 Oct;230:285–96.
17 Chen HC, Tzeng SS, Hsiao YC, Chen RF, Hung EC, Lee OK. Smartphone-based artificial intelligence–assisted prediction for eyelid measurements: algorithm development and observational validation study. JMIR Mhealth Uhealth. 2021 Oct 8;9(10):e32444.
18 Qin S, Ng G, Lo H, Li Y, Cuneo A, Preston M, et al. Application for measuring eyelid weakness in individuals with myasthenia gravis. IEEE global humanitarian technology conference (GHTC). Ieeexplore. ieee.org; 2021. pp. 39–42.
19 Lou L, Yang L, Ye X, Zhu Y, Wang S, Sun L, et al. A novel approach for automated eyelid measurements in blepharoptosis using digital image analysis. Curr Eye Res. 2019 Oct;44(10):1075–9.
20 Hung JY, Perera C, Chen K-W, Myung D, Chiu H-K, Fuh C-S, et al. A deep learning approach to identify blepharoptosis by convolutional neural networks. Int J Med Inform. 2021 Apr;148:104402.
21 Cao J, Lou L, You K, Gao Z, Jin K, Shao J, et al. A novel automatic morphologic analysis of eyelids based on deep learning methods. Curr Eye Res. 2021 Oct;46(10):1495–502.
22 Nallabothula MP, Aleem A, Setabutr P, Hallak J, Yi D. AutoPtosis: a dual model system for rapid and automatic detection of ptosis. Invest Ophthalmol Vis Sci. 2021 Jun;62(8):2160.
23 Aguirre GK. A model of the entrance pupil of the human eye. J Vis. 2019 Jun 27;19(8):85–0.
24 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE. 2016. Available from: http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.
25 Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004 Aug;14(3):199–222.
26 Fuhl W, Kasneci G, Kasneci E. TEyeD: over 20 million real-world eye images with pupil, eyelid, and iris 2D and 3D segmentations, 2D and 3D landmarks, 3D eyeball, gaze vector, and eye movement types. IEEE international symposium on mixed and augmented reality (ISMAR). Ieeexplore. ieee.org; 2021. pp. 367–75.

27 Aran O, Ari I, Guvensan A, Haberdar H, Kurt Z, Turkmen I, et al. A database of non-manual signs in Turkish sign language. IEEE 15th signal processing and communications applications. ieeexplore.ieee.org; 2007. pp. 1–4.
28 Milborrow S, Morkel J, Nicolls F. The MUCT landmarked face database. Pattern Recognition Association of South Africa. 2010;201(0). Available from: http://www.dip.ee.uct.ac.za/~nicolls/publish/sm10-prasa.pdf.
29 Fuhl W, Geisler D, Santini T, Rosenstiel W, Kasneci E. Evaluation of state-of-the-art pupil detection algorithms on remote eye images. Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: adjunct. New York, NY, USA: Association for Computing Machinery; 2016. p. 1716–25.
30 Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [csLG]. 2014 Dec Available from: http://arxiv.org/abs/1412.6980.
31 Abdi H, Williams LJ. Principal component analysis. Wires Comp Stat. 2010 Jul;2(4):433–59.