UC San Diego
UC San Diego Previously Published Works

Title

Objective, computerized video-based rating of blepharospasm severity

Permalink

Journal

ISSN

Authors

Peterson, David A
Littlewort, Gwen C
Bartlett, Marian S
et al.

Publication Date

DOI

Objective, computerized video-based rating of blepharospasm severity

David A. Peterson, PhD
Gwen C. Littlewort, PhD
Marian S. Bartlett, PhD
Antonella Macerollo, MD
Joel S. Perlmutter, MD
H.A. Jinnah, MD, PhD
Mark Hallett, MD
Terrence J. Sejnowski, PhD

Correspondence to
Dr. Peterson:
dap@salk.edu

ABSTRACT

Objective: To compare clinical rating scales of blepharospasm severity with involuntary eye closures measured automatically from patient videos with contemporary facial expression software.

Methods: We evaluated video recordings of a standardized clinical examination from 50 patients with blepharospasm in the Dystonia Coalition's Natural History and Biorepository study. Eye closures were measured on a frame-by-frame basis with software known as the Computer Expression Recognition Toolbox (CERT). The proportion of eye closure time was compared with 3 commonly used clinical rating scales: the Burke-Fahn-Marsden Dystonia Rating Scale, Global Dystonia Rating Scale, and Jankovic Rating Scale.

Results: CERT was reliably able to find the face, and its eye closure measure was correlated with all of the clinical severity ratings (Spearman ρ = 0.56, 0.52, and 0.56 for the Burke-Fahn-Marsden Dystonia Rating Scale, Global Dystonia Rating Scale, and Jankovic Rating Scale, respectively, all $p < 0.0001$).

Conclusions: The results demonstrate that CERT has convergent validity with conventional clinical rating scales and can be used with video recordings to measure blepharospasm symptom severity automatically and objectively. Unlike EMG and kinematics, CERT requires only conventional video recordings and can therefore be more easily adopted for use in the clinic. *Neurology*® 2016;87:2146–2153

GLOSSARY

BFM = Burke-Fahn-Marsden Dystonia Rating Scale; **CERT** = Computer Expression Recognition Toolbox; **CI** = confidence interval; **GDRS** = Global Dystonia Rating Scale; **ICC** = intraclass correlation; **JRS** = Jankovic Rating Scale; **UCSD** = University of California, San Diego; **WUSM** = Washington University School of Medicine.

Blepharospasm is characterized by loss of voluntary control over orbicularis oculi muscles, causing involuntary eyelid closure. It is one of the most common forms of isolated dystonia and can cause functional blindness, significant social disability, and decreased quality of life. Periodic botulinum neurotoxin injections provide some symptomatic relief, but the development of more effective therapies[1,2] requires sensitive and objective methods to rate symptom severity.[3,4]

Current rating scales such as the Burke-Fahn-Marsden Dystonia Rating Scale (BFM),[5,6] the Global Dystonia Rating Scale (GDRS),[7] the Jankovic Rating Scale (JRS),[8] and the recently developed rating scale for blepharospasm[9] are based on inherently subjective clinician evaluation. This raises concerns about interrater reliability and necessitates a substantial effort to evaluate the reliability of such rating scales. In contrast, common availability of inexpensive digital video cameras enables objective video recording analysis with a rapidly growing suite of contemporary artificial intelligence software. One such implementation is the Computer Expression Recognition Toolbox (CERT).[10–13] CERT combines algorithms from computer

Supplemental data at Neurology.org

vision and machine learning. It automatically finds the face and detects facial "action units" on the basis of the facial action coding system,[14] widely used for coding facial expressions in behavioral sciences.[15] In neurologically normal individuals, CERT demonstrates state-of-the-art performance, discriminating emotions in benchmark datasets[16] consisting of >100 individuals producing >500 sequences of expressions.[11,17]

We evaluated the clinical utility of CERT for measuring blepharospasm severity through 2 objectives: to test convergent validity, i.e., whether CERT measures of involuntary eye closure agree with clinical severity ratings, and to determine the viability of CERT, i.e., the proportion of video frames on which CERT can automatically find the patient's face.

METHODS Patients. We evaluated patients previously recruited into the Dystonia Coalition's multicenter Natural History and Biorepository of Isolated Dystonia cross-sectional study. The biorepository includes a centralized, web-based platform for securely uploading, storing, and serving patient data, including video recordings.[18] The system meets Health Insurance Portability and Accountability Act guidelines for security and has been approved by the Human Research Protection Office at the Washington University School of Medicine (WUSM). The Human Research Protection Office at the University of California, San Diego (UCSD) also approved access to the patient data by the UCSD team, including the ability to observe streamed videos (protocol 111255X). We identified 50 patients with isolated dystonia, including at least blepharospasm with or without lower facial involvement, for subsequent analysis with CERT. Most patients were treated with botulinum toxin, with the last injection at least 2 months before the examination. Our only inclusion criterion was that the GDRS and BFM clinical severity ratings were conducted at the same visit as the video recording (for patient characteristics, see the table).

Table	Patient characteristics (n = 49)
Demographics	
Onset age, y	52 ± 10 (30–71)
Sex, F/M	40/9
Disease duration, y	12 ± 9 (1–41)
Severity ratings	
BFM (live)	2.2 ± 1.1 (0–4)
GDRS (live)	3.8 ± 2.5 (0–10)
BFM (video)	2.2 ± 0.9 (0–4)
GDRS (video)	3.7 ± 1.8 (0–8)
JRS (video)	2.2 ± 0.9 (0–4)

BFM = Burke-Fahn-Marsden Dystonia Rating Scale; GDRS = Global Dystonia Rating Scale; JRS = Jankovic Rating Scale.
Values expressed as mean ± SD (range).

Clinical examination and severity ratings. Patients underwent a standardized clinical examination based in part on a protocol developed by the Dystonia Study Group,[7] modified to accommodate features of many types of dystonia as described by Yan et al.,[18] and recorded with digital video at 30 frames per second. We extracted from the videos the part of the examination protocol focusing on blepharospasm (part I, steps 1–4).[18] The participant is seated in a chair facing the video camera. Feet are resting on the floor, and hands are resting in the lap. The camera is zoomed in to capture the head and shoulders only: (1) at rest, eyes open for 10 seconds; (2) at rest, eyes closed gently for 10 seconds; (3) at rest, after eyes are opened, for another 10 seconds; and (4) forced eyelid closure 3 times, with the effect observed for 5 seconds after each closure.

One patient was excluded from further analysis because she was noted on observation to not fully close her eyes during periods of instructed eye closure.

Patient videos were accompanied by clinical severity scores. In this study, we specifically focused on blepharospasm severity ratings and did not include "duration" factors because of the relatively short observation periods. Site ratings, normally conducted live by clinicians who were familiar with the patient's history, included the upper face item from the BFM[5,6] and GDRS.[7] Three movement disorders experts (M.H., H.J., J.P.), blinded to the live ratings, scored the videos using the BFM, GDRS, and JRS.[8] Interrater agreement was evaluated with the intraclass correlation (ICC), and then the 3 raters' video scores were averaged on each patient for each rating scale.

Computational video analysis with CERT. To mitigate security risks associated with transferring files containing video recordings of patient faces, we developed an iterative video preprocessing procedure between the Dystonia Coalition Biorepository team at WUSM and the primary analysis team at UCSD. First, at WUSM, video files were cropped to include just the relevant part of the recording and saved as a separate file that was also re-encoded for streaming. Second, the UCSD team reviewed the streamed videos to validate proper cropping. Third, the WUSM team manually segmented the videos into 9 periods of the examination protocol, 5 periods when the eyes should be open and 4 periods of instructed eye closures, using open-source video annotation software (Elan,[19] Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, the Netherlands), saving as outputs the time stamps of each segment boundary. The WUSM team also ran the video file through CERT. Both the segment boundary time stamps and CERT outputs, without video or patient-identifying information, were sent to the UCSD team as ASCII files. Fourth, the UCSD team reviewed the Elan output files against the streamed videos to validate the segmenting process.

In this study, we used CERT version 4.4.5, previously available from UCSD for academic use and now available as FACET commercial software from iMotions.com (see the CERT processing pipeline depicted in figure 1). Faces were automatically located in each video frame by searching for standard facial landmarks such as eyes, nose, and mouth using previously published computer vision methods.[17,20] We set an a priori criterion of 80% to be the minimum proportion of face-found frames for reasonable assessment of the patient's eye closure features. The raw output of CERT consisted of a continuous-valued measure of eye closure for each video frame. Because of the nature of the machine learning classifier from which it was derived, it was dimensionless and relative, with more negative values associated with open eyes and more positive with closed eyes. Downstream analysis with the CERT raw measure of eye closure included only the face-found frames and was conducted with MATLAB (The
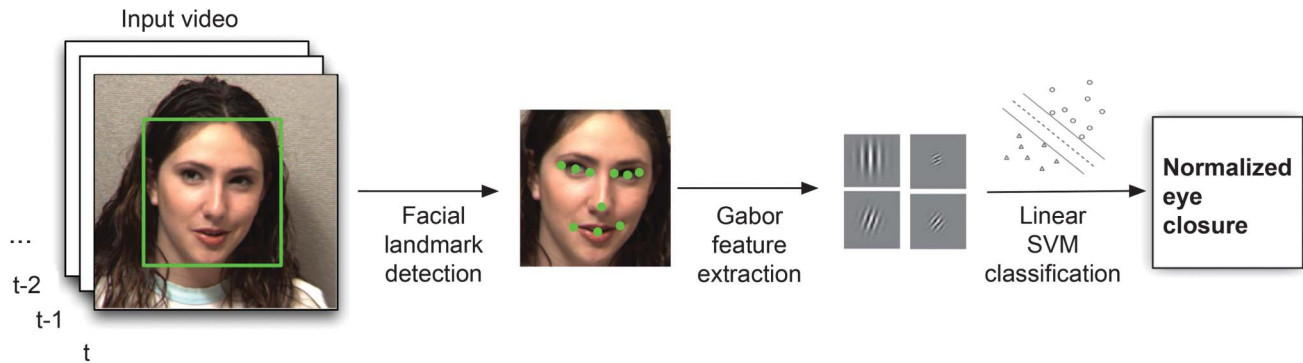
Video segments are analyzed on a frame-by-frame basis. Within the frame image, faces are located with the use of standard facial landmarks. After the face is found, Gabor features are extracted and used in a high-dimensional support vector machine (SVM) large-margin linear classifier, and the distance of the patient's metrics from the hyperplane determines the normalized eye closure magnitude, in arbitrary units (modified from Littlewort et al.,[10] figure 2, with permission).

MathWorks, Inc, Natick, MA). To determine the percentage of time of eyes closed during the periods of instructed eye opening, we needed a way to declare eyes open or closed on each video frame, so we calculated an eye closure threshold for each patient. The video examination protocol included substantial periods of instructed eye closures and eye openings. For this protocol, the distributions of the CERT raw eye closure measure over time would be bimodal for neurologically normal individuals, in which case choosing a threshold between eyes open and eyes closed states would be trivial. In contrast, we hypothesized that the distribution would not be bimodal for patients with blepharo-spasm, so we first evaluated the distributions for bimodality using the Sarle bimodality test (see appendix e-1 at Neurology.org). In the event that >10% of the patients' eye closure distributions were not bimodal, we used a more conventional thresholding method, similar to that used by Bologna et al.[21] Specifically, we calculated the mean of the means of the lowest 5% and highest 5% of samples. Using this per-patient eye closure threshold, we calculated the percentage of time that each patient's eyes were closed when the patient was instructed to have them open, i.e., the number of video frames on which eyes are closed divided by

the total number of video frames from the concatenated eyes open time segments.

Statistical analysis. Correlations were evaluated with the Spearman rank correlation coefficient (ρ) with JMP (SAS Institute Inc, Cary, NC) and characterized with the Cohen effect size conventions.[22] One-sided correlation comparisons were made with Fisher z transforms of the Spearman ρ values, parameterized for dependent and overlapping groups (cocor,[23] comparing correlations, 1.1–2, http://comparingcorrelations.org). We used an α level of 0.05 to determine significance and 95% confidence intervals (CIs).

RESULTS Clinical ratings. The demographics of our patient cohort (table) were consistent with blepharo-spasm epidemiology. Although their severity was, on average, relatively mild (mean BFM = 2.2 for both live and video ratings and GDRS = 3.8 and 3.7 for live and video ratings, respectively), individual patient severity ratings covered the full range of the severity rating scales. Reliability among the 3 video raters was assessed with the ICC, class 2, because all raters rated all participants (Real Statistics Resource Pack software, release 4.5, copyright 2013–2016, Charles Zaiontz, www.real-statistics.com). There was moderate agreement among the video raters on all 3 rating scales, consistent with prior studies using these scales. Specifically, the ICCs (lower–upper CIs) were 0.58 (0.42–0.71) for the BFM, 0.62 (0.44–0.75) for the GDRS, and 0.57 (0.39–0.72) for the JRS. The live and video-based ratings were in general agreement (figure 2) for both the BFM [Spearman $\rho(47) = 0.61$, $p < 0.0001$] and the GDRS [Spearman $\rho(47) = 0.58$, $p < 0.0001$].

CERT viability and convergent validity with clinical scales. CERT was able to find the face in 100% of the video frames for 46 patients and in 95%, 94%, and 93% for the other 3 patients. Patients exhibited a wide variety of bimodality in the distributions of the CERT raw eye closure measure over all of the

Figure 2   Clinical severity ratings: video vs live ratings
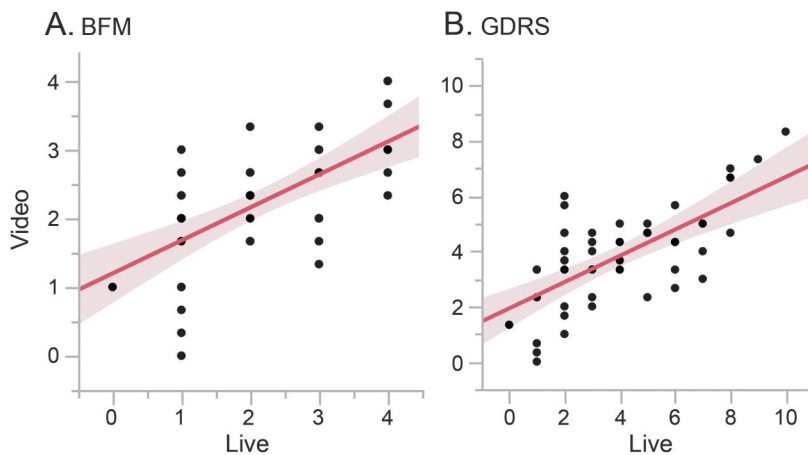
Clinical severity measures based on video vs live ratings for (A) Burke-Fahn-Marsden Dystonia Rating Scale (BFM) and (B) Global Dystonia Rating Scale (GDRS), including linear regression lines and 95% confidence intervals (shaded).

Figure 3   Eye closure thresholding
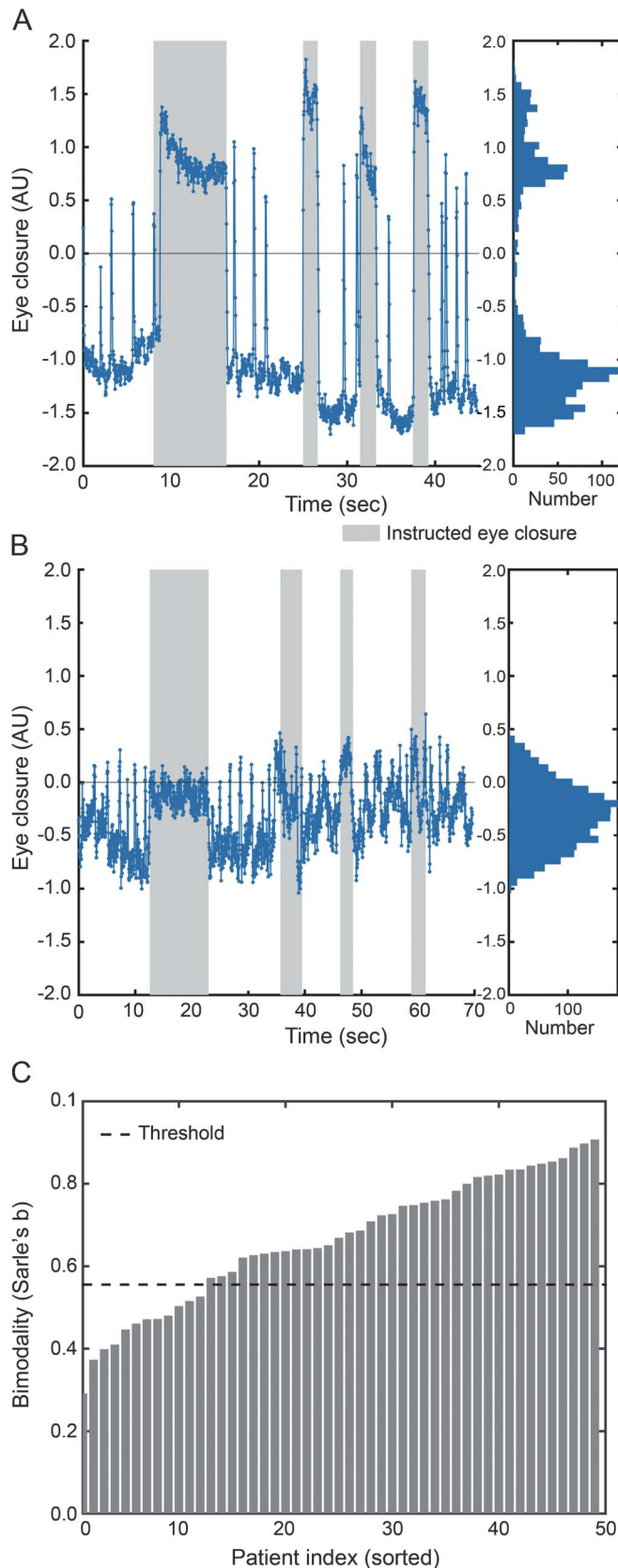
Frame-by-frame continuous-valued Computer Expression Recognition Toolbox measures of eye closure during clinical examination, corresponding distributions over time, and bimodality of the distributions. Two sample patients: (A) one with and (B) one without a clear bimodal distribution in eye closure. (C) Bimodality of patients' eye closure distribution (sorted by bimodality measure). Horizontal line represents threshold above which distribution is conventionally considered bimodal (Sarle b, see appendix e-1). BFM = Burke-Fahn-Marsden Dystonia Rating Scale; GDRS = Global Dystonia Rating Scale; JRS = Jankovic Rating Scale. AU = arbitrary units.

video frames (figure 3). Some exhibited clearly bimodal distributions (figure 3A), while others did not (figure 3B). On the basis of the Sarle bimodality test (see appendix e-1), a substantial portion (12 of 49, 24%) had distributions that were not bimodal (figure 3C). Thus, we used the conventional method for thresholding the eye closure output (see Methods). This per-patient threshold was then used to determine whether the eyes were closed on each video frame and the percentage of frames during the instructed "eyes open" segments used to calculate each patient's CERT-based metric: eye closure time (percent). The total duration of the concatenated eyes open segments was 43 ± 8 seconds (range 28–64 seconds). The CERT eye closure time (percent) was correlated with all of the clinical rating scales (figure 4), both for the live [BFM: Spearman $\rho(47) = 0.46$, $p = 0.0008$; GDRS: Spearman $\rho(47) = 0.30$, $p = 0.035$] and for the video-based severity [BFM: Spearman $\rho(47) = 0.56$, $p < 0.0001$; GDRS: Spearman $\rho(47) = 0.52$, $p < 0.0001$; JRS: Spearman $\rho(47) = 0.56$, $p < 0.0001$] ratings.

The correlations between the CERT eye closure time (percent) and the severity ratings were higher for the video-based than for the live severity ratings. This difference ($0.56 > 0.46$) was not significant for the BFM (based on Pearson and Filon $z = 0.953$, $p = 0.170$ and Meng et al $z = 0.926$, $p = 0.177$, difference CI $= -0.151$ to 0.422). The difference ($0.52 > 0.30$) was, however, significant for the GDRS (Pearson and Filon $z = 1.887$, $p = 0.030$ and Meng et al $z = 1.830$, $p = 0.034$), although the difference CI (-0.019 to 0.553) is not consistent with an effect.

DISCUSSION We analyzed the CERT eye closure measures to objectively compute the percent of time that the patients' eyes were closed when the patients were instructed to open them. We found correlations between this measure and all of the clinical severity measures. Thus, the CERT eye closure measure exhibits convergent validity with conventional clinician-scored rating scales. This overarching result held across all 3 clinical rating scales used (the BFM, GDRS, and JRS), as well as the 2 modalities in which they were administered (live and from video observations).

The correlations between CERT and the clinical severity ratings varied slightly, depending on whether the ratings were live or based on video review. This is expected, given the moderate but imperfect correlation between the live and video-based rating modalities,

Figure 4 Convergent validity between Computer Expression Recognition Toolbox (CERT) and clinical severity ratings
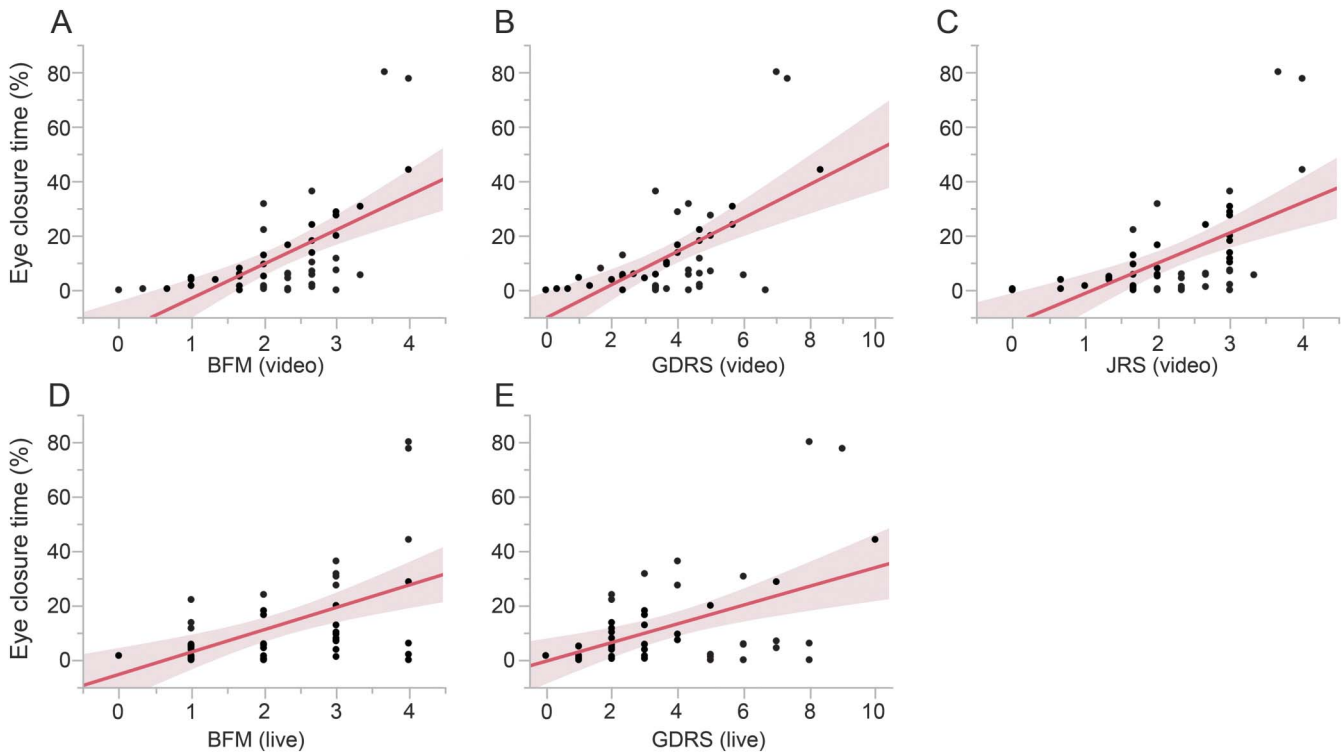
Convergent validity between percent time of eye closure based on CERT and clinical severity ratings, separately for video (A–C) and live (D and E) clinical rating modalities. Lines are linear regressions, and shaded regions show the 95% confidence intervals.

a discrepancy that has been widely acknowledged[24] but not systematically investigated. There was a trend toward the CERT eye closure measure correlating better with video than live ratings, reaching significance for some but not all statistical comparisons. CERT and the video raters may detect eye closures that are missed by a live rater who may not direct complete attention to observing the patient when simultaneously recording scores. Perhaps it is more important that the video raters evaluate the same video that CERT analyzed, which may enhance convergence.

CERT was robust with respect to variable head movement and variable lighting conditions during the video recording, reliably registering the face an average of 99.6% of the video frames across all patients. The lowest percentage of face-found frames, at 93%, far exceeded our a priori threshold of 80% for retaining videos for further analysis. This is particularly encouraging for the relevance of CERT for patients with cranial dystonia with or without comorbid cervical dystonia because about one-fourth of the patients in this study also had head tremor. It also supports the broader applicability of CERT for analyzing video recordings from a conventional clinical setting because no special efforts were made to optimize lighting conditions in the present study.

The present study has a few noteworthy limitations. First, severity is not uniformly distributed. In

the case of the GDRS, for example, the preponderance of patients were rated 3 to 4, with relatively few rated in the 1 to 2 and 8 to 10 ranges. Although this likely represents the larger population of patients diagnosed with blepharospasm, it diminishes our ability to test convergent validity among rating methods at the low and high extremes of severity. Post hoc review of the 2 notable outliers, patients with 80% and 78% eye closure, suggests that these were legitimate measures of severe blepharospasm and not CERT processing artifacts. This emphasizes limitations in the linearity of the clinical rating scales at the high end. Second, the present study included 49 patients. A larger cohort may increase statistical power and include more severe patients, depending on the distribution in the population and any sampling bias. Third, the video observation period may be too brief, potentially providing insufficient duration of instructed eyes open periods to permit detection of low but still pathologic amounts of eye closure. The floor effect apparent in the distributions of the CERT eye closure measure (see figure 4), in which several patients exhibited little if any eye closure during the instructed eyes open periods, supports this possibility. In this study, the cumulative duration of instructed eyes open periods averaged 43 seconds, whereas other studies used 120 seconds of video to evaluate blink rate in blepharospasm.[25,26] Nevertheless,

the brief instructed eye open periods in the present study were the same for CERT and for the video reviewers, so both approaches had the same data on which to base their analyses. Indeed, briefer observation periods, if sufficient to assay severity, could contribute to clinical efficiency. All 3 of these limitations could be addressed by CERT application to a larger study incorporating a broader distribution of severity and longer observation periods.

Our results further highlight concerns about the reliability of inherently subjective methods for evaluating blepharospasm severity and support the need for more objective measures of motor symptoms in dystonia.[4,27] The 3 video reviewers, after conducting a joint practice session to work toward consensus on how to apply the rating scales immediately before their independent ratings, exhibited only moderate agreement (ICC ≈ 0.6) on the severity ratings across these patients. The video raters also exhibited only moderate correlation (Spearman $\rho = 0.58–0.61$) with the live raters. These interrater results are consistent with those of previous reports using the BFM and GDRS for blepharospasm.[7] Objective measures can mitigate the intrarater and interrater confounds inherent in human assessments. Indeed, deterministic algorithms such as the CERT eye closure measure give intrarater and interrater variability of zero. Regardless, no gold standard for comparison exists for any of these measures. This is a concern for future rating scale development, whether based on human subjective ratings or automated objective ratings. We expect that increased integration between clinical and computational experts will provide stronger measures than either approach in isolation.

As used in the present study, CERT provides a partially automated means for assessing symptom severity in blepharospasm. Further automation could greatly alleviate the labor-intensive nature of human-based video review, which can be tedious and error prone and can require the valuable and limited resource of blepharospasm experts. The core CERT algorithm can already be used in a real-time mode. Automating the input and output processes would enable a real-time, end-to-end video processing capability. This would facilitate the translation of CERT from research to routine clinical use. Although the focus of the present study was severity, future studies including a healthy control group could be conducted to see whether the CERT measures could reliably detect the presence of blepharospasm. Thus, future development of CERT could include an Internet-based service to provide objective screening and severity measures in near-real time while complying with appropriate Health Insurance Portability and Accountability Act and Institutional Review Board requirements.

Using <2 minutes of video recorded during a standard clinical examination, we have demonstrated that CERT can objectively measure eye closure in blepharospasm. No separate technology or procedure is required during the examination, and the video can be analyzed offline. Thus, a CERT-based objective measure can supplement traditional subjective ratings of blepharospasm severity with minimal additional burden in the clinical setting.

REFERENCES

1. Price KM, Ramey NA, Richard MJ, Woodward DJ, Woodward JA. Can methylphenidate objectively provide relief in patients with uncontrolled blepharospasm? A pilot study using surface electromyography. Ophthal Plast Reconstr Surg 2010;26:353–356.
2. Kranz G, Shamim EA, Lin PT, Kranz GS, Voller B, Hallett M. Blepharospasm and the modulation of cortical excitability in primary and secondary motor areas. Neurology 2009;73:2031–2036.
3. Wabbels B, Jost WH, Roggenkämper P. Difficulties with differentiating botulinum toxin treatment effects in essential blepharospasm. J Neural Transm 2011;118:925–943.
4. Peterson DA, Berque P, Jabusch HC, Altenmüller E, Frucht SJ. Rating scales for musician's dystonia: the state of the art. Neurology 2013;81:589–598.
5. Burke RE, Fahn S, Marsden CD, Bressman SB, Moskowitz C, Friedman J. Validity and reliability of a rating-scale for the primary torsion dystonias. Neurology 1985;35:73–77.
6. Fahn S. Assessment of the primary dystonias. In: Munsat TL, ed. Quantification of Neurologic Deficit: Butterworths; 1989:241–270.
7. Comella CL, Leurgans S, Wuu J, Stebbins GT, Chmura T; Dystonia Study Group. Rating scales for dystonia: a multicenter assessment. Mov Disord 2003;18:303–312.
8. Jankovic J, Orman J. Botulinum-a toxin for cranial-cervical dystonia: a double-blind, placebo-controlled study. Neurology 1987;37:616–623.
9. Defazio G, Hallett M, Jinnah HA, et al. Development and validation of a clinical scale for rating the severity of blepharospasm. Mov Disord 2015;30:525–530.
10. Littlewort G, Whitehill J, Wu T, et al. The Computer Expression Recognition Toolbox (CERT). In: IEEE International Conference on Automatic Face and Gesture Recognition. 2011;298–305.
11. Bartlett MS, Littlewort G, Frank MG, Lainscsek C, Fasel I, Movellan J. Automatic recognition of facial actions in spontaneous expressions. J Multimedia 2006;1:22–35.
12. Bartlett MS, Hager JC, Ekman P, Sejnowski TJ. Measuring facial expressions by computer image analysis. Psychophysiology 1999;36:253–263.
13. Donato G, Bartlett MS, Hager JC, Ekman P, Sejnowski TJ. Classifying facial actions. IEEE Trans Pattern Anal Mach Intell 1999;21:974–989.
14. Ekman P, Friesen W. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto: Consulting Psychologists Press; 1978.
15. Ekman P, Rosenberg EL. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). 2nd ed. Oxford: Oxford University Press; 2005.
16. Tian YL, Kanade T, Cohn JF. Recognizing upper face action units for facial expression analysis. Proc CVPR IEEE 2000:294–301.
17. Bartlett MS, Littlewort G, Vural E, et al. Insights on spontaneous facial expressions from automatic expression measurement. In: Curio C, Bülthoff HH, Giese MA, editors. Dynamic Faces: Insights From Experiments and Computation. Cambridge: MIT Press; 2011.
18. Yan L, Hicks M, Winslow K, et al. Secured web-based video repository for multicenter studies. Parkinsonism Relat Disord 2015;21:366–371.

19. Lausberg H, Sloetjes H. Coding gestural behavior with the NEUROGES-ELAN system. Behav Res Methods 2009; 41:841–849.
20. Fasel I, Fortenberry B, Movellan J. A generative framework for real time object detection and classification. Comput Vis Image Und 2005;98:182–210.
21. Bologna M, Fasano A, Modugno N, Fabbrini G, Berardelli A. Effects of subthalamic nucleus deep brain stimulation and l-dopa on blinking in Parkinson's disease. Exp Neurol 2012;235:265–272.
22. Cohen JA. Power primer. Psychol Bull 1992;112:155–159.
23. Diedenhofen B, Musch J. Cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 2015;10:e0121945.
24. Vidailhet M, Vercueil L, Houeto JL, et al. Bilateral deep-brain stimulation of the globus pallidus in primary generalized dystonia. N Engl J Med 2005;352:459–467.
25. Bentivoglio AR, Daniele A, Albanese A, Tonali PA, Fasano A. Analysis of blink rate in patients with blepharospasm. Mov Disord 2006;21:1225–1229.
26. Ferrazzano G, Conte A, Fabbrini G, et al. Botulinum toxin and blink rate in patients with blepharospasm and increased blinking. J Neurol Neurosurg Psychiatry 2015; 86:336–340.
27. Lunardini F, Maggioni S, Casellato C, Bertucco M, Pedrocchi ALG, Sanger TD. Increased task-uncorrelated muscle activity in childhood dystonia. J Neuroeng Rehabil 2015;12:52.