Evaluating Corneal Fluorescein Staining Using a Novel Automated Method

Francisco Amparo,1 Haobing Wang,2 Jia Yin,1 Anna Marmalidou,1 and Reza Dana1

1Massachusetts Eye and Ear Infirmary, Department of Ophthalmology, Harvard Medical School, Boston, Massachusetts, United States
2Massachusetts Eye and Ear Infirmary, Peabody Laboratory, Harvard Medical School, Boston, Massachusetts, United States

Correspondence: Reza Dana, Massachusetts Eye and Ear Infirmary, Cornea Service, 243 Charles Street, Boston, MA 02114, USA; reza_dana@meei.harvard.edu.
Submitted: March 9, 2017
Accepted: June 5, 2017

PURPOSE. To evaluate interobserver concordance in measured corneal fluorescein staining (CFS) using the National Eye Institute/Industry (NEI) grading scale and the Corneal Fluorescein Staining Index (CFSi), a computer-assisted, objective, centesimal scoring system.

METHODS. We conducted a study to evaluate CFS in clinical photographs of patients with corneal epitheliopathy. One group of clinicians graded CFS in the images using the NEI while a second group applied the CFSi. We evaluated the level of interobserver agreement and differences among CFS scores with each method, level of correlation between the two methods, and distribution of cases based on the CFS severity assigned by each method.

RESULTS. The level of interobserver agreement was 0.65 ($P < 0.001$) with the NEI, and 0.99 ($P < 0.001$) with the CFSi. There were statistically significant differences among clinicians’ measurements obtained with the NEI ($P < 0.001$), but not with the CFSi ($P = 0.78$). There was a statistically significant correlation between the CFS scores obtained with the two methods ($R = 0.72; P < 0.001$). The NEI scale allocated the majority of cases (65%) within the higher quartile in the scale’s severity (12–15/15). In contrast, the CFSi allocated the majority of cases (61%) within the lower quartile in the scale’s severity (0–25/100).

CONCLUSIONS. The CFSi is easy to implement, provides higher interobserver consistency, and due to its continuous score can discriminate smaller differences in CFS. Reproducibility of the computer-based system is higher and, interestingly, the system allocates cases of epitheliopathy in different severity categories than clinicians do. The CFSi can be an alternative for objective CFS evaluation in the clinic and in clinical trials.

Keywords: corneal fluorescein staining, ocular surface, dry eye, CFS, automated CFS, CFSi

Ocular surface disease is characterized by clinical signs and symptoms that are indispensable for the assessment, staging, clinical management, and follow-up of the condition.1–4 Punctate keratitis (corneal epitheliopathy) is a clinical sign that reflects, physiologically and anatomically, the viability of the corneal epithelium.3,5 Punctate keratitis and other forms of corneal epitheliopathy (e.g., confluent defects) are easily studied and assessed in the clinic through dyes applied onto the ocular surface.3,4 Sodium fluorescein has been used for decades as a biomarker that can be excited with cobalt blue light to depict the degree of corneal epitheliopathy; a technique commonly referred as to corneal fluorescein staining (CFS).3–5

Corneal fluorescein staining is easy to perform in the clinic and is cardinal for the diagnosis, treatment, and follow-up of ocular surface diseases, including dry eye disease.1–5 Although various CFS grading scales have been described, clinicians generally rely on practical ordinal scales, for example, “mild,” “moderate,” “severe” or 0+, 1+, 2+, and so on, or variations thereof. Although used with less frequency, other ordinal CFS scales that are more specific, defining the degree of punctate keratitis, are also available, including the Oxford grading scale and the National Eye Institute/Industry grading scale (NEI).6,7 The Oxford scale relies on a comparative chart to define the degree of CFS, with different images defining six levels of severity (0 [absent] to 5 [severe]).6 The NEI scale relies on a chart that divides the cornea into five sections and assigns a value from 0 (absent) to 3 (severe) to each section, based on the amount, size, and confluence of the punctate keratitis, for a maximum of 15 points.7 Currently, the Oxford and NEI scales are the most commonly used CFS grading scales in clinical trials due to the systematic methodology they utilize and the popularity among clinicians and researchers they have gained over the past years. However, even though they are more systematic, these CFS scales are still observer dependent and subjective, and thus susceptible to high inter- and intraobserver variance.

Objective methods to assess CFS based on corneal image analysis have been proposed, but none has been implemented for routine CFS assessment in the clinic or in clinical trials.8–11 We proposed that a systematic method that objectively evaluates CFS, minimizes subjective input from human observers, and facilitates comparisons reducing inter- and intrarater variability will represent an important advance for the evaluation of CFS changes over time or after an intervention. We thus developed the Corneal Fluorescein Staining Index (CFSi), a computer-guided method that objectively quantifies CFS in clinical photographs and delivers a continuous, centesimal (0–100) score.

Herein we present the results of a prospective study involving consecutive patients with ocular surface disease and...
A Novel Method to Quantify Corneal Fluorescein Staining

A cornea specialist selected all the photographs considered to have good clinical quality to permit another clinician to score CFS out of the clinic. All the selected photographs were evaluated clinically (NEI system) and with the CFSi, and all the scores obtained with both techniques were included in the analysis.

Clinical Scoring of Corneal Photographs

Four ophthalmologists with formal training in corneal diseases and with particular expertise in dry eye syndromes scored CFS in the set of randomly arranged photographs using the NEI scale. Briefly, the NEI grading scale consists of a grid that divides the corneal area into five sections, each of which is assigned a score between zero and 3 depending of the amount and distribution of CFS; the total CFS score ranges from 0/15 (absence of corneal epitheliopathy) to 15/15 (severe epitheliopathy). All observers scored all the images using the same computer monitor with the same brightness and color settings, the same room illumination, and with no time limit to complete their evaluations. Observers were allowed to have at hand the NEI criteria and guiding charts.

Automated Scoring of Corneal Photographs

Three nonclinicians scored CFS in the same set of randomized photographs using the CFSi. A fourth clinician who scored the photographs using the NEI scale also scored the images with the CFSi, and the scores obtained by this observer were used as control observations to evaluate the correlation between the two CFS scoring methods. A total of four graders scored the photographs with each method (Fig. 1). The process to assess CFS with the CFSi was as follows: The photograph file was loaded into a computer; then the observer selected the corneal area to score with a prebuilt, modifiable region of interest selection tool (to avoid lids, conjunctiva, and other areas not intended for scoring); and finally, a CFSi score expressed in a series of plug-ins for ImageJ (an open-source image processing/analysis software developed by the NIH, Bethesda, MD, USA) that can be executed on standard computers and is both PC and Mac compatible.

Fluorescein Staining Map Generation

At the start of the plug-in, a circle is generated and the user may adjust the circle to fit the corneal size. Before generating a fluorescein map of the cornea, areas with specular reflections are detected by a flooding algorithm. Areas with specular reflections are then excluded from the calculation of the fluorescein staining score, as the specular reflection does not represent the true colors. The plug-in then reads the RGB value of each pixel within the grid and converts the RGB values to

FIGURE 1. Process followed for clinical and automated assessment of corneal fluorescein staining in clinical photographs.
A Novel Method to Quantify Corneal Fluorescein Staining

HSV space, where the value component is given by:

\[ V = \max(R, G, B). \]  

The saturation component is given by:

\[ S = \begin{cases} \frac{\delta}{\max(R, G, B)} & \text{if } \max(R, G, B) \neq 0 \\ 0 & \text{else} \end{cases} \]  

where \( \delta = \max(R, G, B) - \min(R, G, B) \).

The hue component is given by:

\[ H = \begin{cases} 0 & \text{if } \delta = 0 \\ \frac{60 \times (\frac{\delta}{\max(R, G, B)}) + 360}{360} & \text{if } \delta \neq 0 \text{ and } \max(R, G, B) = R \\ \frac{60 \times (\frac{\delta}{\max(R, G, B)}) + 360}{360} & \text{if } \delta \neq 0 \text{ and } \max(R, G, B) = G \\ \frac{60 \times (\frac{\delta}{\max(R, G, B)}) + 360}{360} & \text{otherwise} \end{cases} \]  

Then the hue value is mapped from \([0°, 360°)\) to \([0, 1]\) using a parabola curve. Based on multiple experimental data, we found that for the range of \([200°, 220°)\), the linear function is a better marker than a parabolic function to avoid confusion between the blue (from the light source utilized to excite fluorescein) and green colors (fluorescence from epithelial defects). The final fluorescence (greenness) value is calculated as:

\[ \text{greenness} = H \times S \times V \]  

At this point the greenness score falls within the range of \([0, 1]\), as the \(H\), \(S\), and \(V\) values fall all within the range of \([0, 1]\). A zero score means there is absolutely no greenness in the corneal selection, and a score of 1 means the entire selection is positive (preset threshold) for green. When calculating the percentage of the fluorescent area, the fluorescence threshold is applied; that is, if the greenness score of a pixel is above the threshold, then it is regarded as fluorescein staining. This threshold was incorporated to the algorithm to avoid false positives triggered by a stained tear film or other artifacts, and was selected after careful review of multiple fluorescent maps by an experienced clinician.

**Data and Statistical Analysis**

We evaluated the level of absolute agreement among the scores obtained with the NEI scale by the clinicians and among the raters using the CFSi scale with the intraclass coefficient of correlation in a 2-way random model. We assessed the presence of significant differences among the mean CFS scores obtained by observers with each technique using the analysis of variance (ANOVA) for repeated measures for the CFSi continuous scores and the Friedman’s test for the ordinal scores obtained with the NEI scale. We evaluated the correlation between the CFS scores obtained by the control observer using the two different methods. Finally, we evaluated the distribution of CFS severity scores within the studied cohort.

**RESULTS**

We obtained 61 corneal images with punctate keratitis in patients with ocular surface disease. The mean CFS scores obtained by the clinical observers using the NEI scale were as follows: observer A, 13.8/15 (range, 5–15); observer B, 11.5/15 (5–15); observer C, 10.7/15 (2–15); and observer D, 12.7/15 (3–15). The CFS scores obtained by the observers using the computer-assisted system were observer E, 21.3 (range, 1–69); observer F, 21.4 (5–69); observer G, 21.3 (0.5–68); and observer D, 21.2 (5–68). We found a statistically significant correlation between the scores obtained by the control observer using the two different methods (\(R = 0.72; P < 0.001\)).

The intraclass coefficient of correlation for absolute agreement among clinical observers using the NEI scale was 0.65 (\(P < 0.001\); 95% confidence interval [CI] 0.482–0.775) and 0.99 (\(P < 0.001\); 95% CI 0.997–0.999) for the observers using the CFSi. The Friedman’s test reported statistically significant differences among the mean CFS scores obtained by clinical observers with using the NEI scale (\(P < 0.001\)). The ANOVA test for repeated measures did not report significant differences among the mean scores obtained by the observers using the CFSi method (\(P = 0.78\)).

Finally, we analyzed the distribution of the CFS scores obtained with both methods across the population, we found that when using the NEI scale, 4 cases were allocated within the scale’s lowest (first) severity quartile (0–3/15), 24 cases in the second quartile (4–7/15), 57 cases in the third quartile (8–11/15), and 159 cases in the highest (fourth) severity quartile (12–15/15). Using the CFSi scale, 150 cases were allocated within the scale’s lowest severity quartile (0–25/100), 77 cases in the second quartile (26–50/100), and 17 cases in the third quartile (51–75/100). None of the cases was allocated in the highest severity (fourth) quartile (76–100/100).

**DISCUSSION**

Objective CFS scoring using the CFSi in clinical images was feasible and convenient. Observers did not report difficulties or required assistance using the CFSi, which is a fundamental aspect to consider when assessment of CFS is performed a posteriori in large studies or multicenter trials. We found a clear correlation between CFS scores obtained with the two different methods. However, it is important to note that such association reflects only that scores were allocated (ranked) in similar order but meaningful differences in CFS score magnitude were present. The level of interobserver consistency found with the NEI scale (0.65) is considered to be fair to moderate, in contrast to the CFSi showing a level of consistency (0.99) considered excellent.14,15

The CFSi objectively evaluates the amount of fluorescein staining avoiding subjective input from a clinician, which is influenced by the size, number, confluence, or localization of CFS. Although these punctate keratitis characteristics are important for clinical judgment and management, in multicenter studies or clinical trials, CFS scales with low levels of consistency can compromise the reliability of the results. The CFSi linearly assigns higher scores to larger CFS areas and lower scores to areas without CFS, avoiding exponential changes arising from subjective judgment. CFS grading scales such as the NEI and Oxford attempt to address two needs simultaneously: One is giving weight in the severity scale to clinically important features, and the other is being less subjective by providing more quantifiable features in their system. Nevertheless, our data show that these scales lack consistency and are highly variable, limiting further development in the study of the ocular surface.

Ordinal scales are by definition characterized by nonclear intervals between values. Ordinal CFS scales often assign very high scores to cases considered clinically severe, even when a limited area of the cornea is affected, and by doing so preclude the assignment of higher scores as CFS increases. This phenomenon often leads to arbitrarily using the same “severe” score in cases with different degrees of CFS (Fig. 2). In fact, in some cases with very severe CFS (e.g., one that has obtained the maximum score) there is no room to increase the score if the condition worsens (Fig. 2), but the same applies if the
condition improves but not sufficiently to be reflected by the immediate lower value in the CFS scale (i.e., from 15/15 to 14/15) (Figs. 2, 3). Another example of the limitations of ordinal scales presents when cases rated with severe scores improve and the scale cannot reflect the improvement since the improvement is modest and there is not an appropriate value in between, or the case still meets criteria to be rated with the highest CFS score, creating a “plateau” effect (Fig. 4). In this study clinical observers mostly assigned severe scores, allocating 65% of the cases in the NEI scale upper quartile of severity (score 12-15/15); should CFS worsen it would be difficult to reflect that change in many of these cases. Other examples of ordinal scale limitations include variability in the perception or interpretation of guidelines creating a “fickle-grid” effect (Fig. 5), or judgment bias when cases are evaluated next to each other and the scores are influenced by observer’s previous decisions.

The CFSi objectively quantifies the number of fluorescence-positive pixels in the corneal area, thus reflecting even minimal CFS fluctuations regardless of the initial severity. In contrast to the NEI scale, the CFSi allocated most measurements (61%) in the lower quartile of a 100-point scale, while clinicians

![Figure 2](image1.png)  
**Figure 2.** Three different cases depicting a maximum corneal fluorescein staining score (15/15) according to the NEI/Industry grading system despite clear differences in the amount of punctate keratitis among them.

![Figure 3](image2.png)  
**Figure 3.** Examples of cases with marked CFS differences that clinicians graded with the maximum score (15/15) following the NEI/Industry grading system guidelines. (A, C) The clinical images captured in the clinic. (B, D) The area selected by the observer and the score and analysis obtained with the CFSi.
allocated only 2% of the measurements in the lower quartile (NEI 0–3/15). This does not signify that the CFSi underestimates (or overestimates) CFS scores as compared to the NEI scale, but only that the measurements, although correlated, are different in nature. CFS values and severity ranges are different for both grading systems, and each CFS score is inherent to the way it is calculated and allocated within its own severity scale; therefore CFS values should be considered only in the context of the correspondent CFS grading system. A limited number of studies have explored the applicability of CFS image analysis in the clinical setting, mostly correlating objective image-analysis techniques to ordinal scales used during evaluation of contact lens-related complications.\textsuperscript{8–11} These studies show that in general, CFS scores based on image analysis are significantly correlated with clinical scores, independently of the clinical CFS scale used. One report showed that levels of interobserver

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{The plateau effect is a major limitation of the CFS NEI/Industry grading system. Cases with severe corneal epitheliopathy that improve, but not sufficiently to reach the immediate lower value in the CFS scale, do not reflect a reduction in CFS scores \textit{(green arrow)}. Similarly, cases that have reached the maximum CFS score cannot reflect score increments if the epitheliopathy later worsens \textit{(red arrow)}. CFS, corneal fluorescein staining.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{The fickle-grid effect consists of a variable perception of the position or size of the NEI/Industry grading system grid during corneal fluorescein staining assessment using the slit lamp.}
\end{figure}
agreement with image analysis are approximately seven times better than with subjective clinical matching scales.\textsuperscript{9} Evidence also shows that depending on the algorithm used for CFS image analysis, repeatability of CFS scores can vary, and the correlation with clinical observations can change dramatically (from positive to negative).\textsuperscript{9} Another study compared a CFS image analysis scoring technique with the Oxford and NEI scales and found good levels of correlation with both.\textsuperscript{8}

To date investigators have focused mainly on the correlation between CFS scores obtained with image analysis and ordinal clinical scores, but an intuitive continuous scale has not consistently been pursued. We developed a novel method that combines objective CFS assessment and an intuitive continuous centesimal scale. While a continuous scale can fit an unlimited number of values, in contrast, the use of subjective CFS scales with arbitrary intervals can be one of the reasons multiple clinical trials continue to fail in detecting changes in CFS. In the series of cases we studied we observed that the subjective scales presented a clear trend to assign a high degree of severity, while the objective scale did exactly the opposite. This might be related to the fact that the cases in the current study fit into what is “typically” considered as severe dry eye, but the same could have happened if patients with “typically” mild disease had been evaluated (allocating the majority in the lower quartile). In both cases, detecting subtle changes in CFS, either worsening in severe cases or improvement in mild cases, would be challenging with an ordinal scale. It is important to note that application of an image-based CFS scoring method has some challenges, mostly related to the achievement of a standardized and repeatable image acquisition process. However, this aim is not insurmountable and can be accomplished by using a defined protocol and with some practice. Specifically, it is critical to minimize variables such as illumination intensity, slit-lamp use technique, fluorescein application, or camera settings. In some cases with severe ocular surface disease, clinical image acquisition may present some challenges, and thus clinical and research personnel must be familiar with the protocols and ready to perform a procedure that is as quick and comfortable as possible. In any case, these are limitations that are present in virtually all scenarios where ocular surface image-based analysis is conducted.

Corneal fluorescein staining is an essential biomarker to understand ocular surface pathology; it is fundamental to guide treatment, evaluate its efficacy, and develop new therapies, and it is a cornerstone for regulatory purposes. The CFSi is an alternative way to assess CFS that is easy to implement and has a higher level of interobserver consistency than an ordinal scale such as the NEI/Industry grading scale. Additionally, our data provide new perspectives suggesting that objective continuous CFS scales are more accurate in detecting differences in measured CFS. Objective continuous scales could lead to an alternative way of corneal epitheliopathy profiling in ocular surface disease and to some paradigm changes, perhaps by showing some interventions to be more effective than previously thought or by improving understanding of the real relationship between clinical signs and patients’ symptoms.

Acknowledgments

The Massachusetts Eye and Ear holds intellectual property pertaining to the use of the technology mentioned in this manuscript. Francisco Amparo and Reza Dana are listed as co-inventors of this technology.

Disclosure: F. Amparo, None; H. Wang, None; J. Yin, None; A. Marmalidou, None; R. Dana, None

References