Journal of Dentistry 74 (2018) S27–S33

Contents lists available at ScienceDirect

Journal of Dentistry

journal homepage: www.elsevier.com/locate/jdent

Digital epidemiology – Calibrating readers to score dental images remotely

Bruce A. Dye[a,*], Michaela Goodwin[b], Roger P. Ellwood[b], Iain A. Pretty[b]

[a] National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD, United States
[b] University of Manchester Colgate Palmolive Dental Health Unit, Manchester, UK

ARTICLE INFO

Keywords:
Inter-examiner reliability
Epidemiology
Dental public health
Kappa
Data quality
Training
Dental fluorosis
NHANES

ABSTRACT

Background: There is growing interest to use digital photographs in dental epidemiology. However, the reporting of procedures and metric-based performance outcomes from training to promote data quality prior to actual scoring of digital images has not been optimal.

Methods: A training study was undertaken to assess training methodology and to select a group of scorers to assess images for dental fluorosis captured during the 2013–2014 National Health and Nutrition Examination Survey (NHANES). Ten examiners and 2 reference examiners assessed dental fluorosis using the Deans Index (DI) and the Thylstrup-Fejerskov (TF) Index. Trainees were evaluated using 128 digital images of upper anterior central incisors at three different periods and with approximately 40 participants during two other periods. Scoring of all digital images was done using a secured, web-based system.

Results: When assessing for nominal fluorosis (apparent vs. non-apparent), the unweighted Kappa for DI ranged from 0.68 to 0.77 and when using an ordinal scale, the linear-weighted kappa for DI ranged from 0.43 to 0.69 during the final evaluation. When assessing for nominal fluorosis using TF, the unweighted Kappa ranged from 0.67 to 0.89 and when using an ordinal scale, the linear-weighted kappa for TF ranged from 0.61 to 0.77 during the final evaluation. No examiner improvement was observed when a clinical assessment feature was added during training to assess dental fluorosis using TF, results using DI was less clear.

Conclusion: Providing examiners theoretical material and scoring criteria prior to training may be minimally sufficient to calibrate examiners to score digital photographs. There may be some benefit in providing an in-person training to discuss criteria and review previously scored images. Previous experience as a clinical examiner seems to provide a slight advantage at scoring photographs for DI, but minimizing the number of scorers does improve inter-examiner concordance for both DI and TF.

1. Introduction

Epidemiology uses systematic approaches to study the distribution of diseases and adverse health conditions in populations. This systematic approach is typically influenced by a number of factors including the use of a relevant study design to test a hypothesis or address an objective and the use of appropriate data collection methodology to maximize internal validity. To achieve internal validity, we want to collect data aiming for precision, accuracy, and reproducibility. Precision occurs when repeated measurements on the same item become very close to each other and accuracy occurs when a measurement made is as close to a standard or to the "truth" as possible. Reproducibility is the ability to achieve the same results and is very important in epidemiology, especially from a "systems" perspective when multiple data collection teams or locations are employed. When the data collection process is functioning well across teams for example, precision and accuracy are high, strengthening our confidence that the data collection system is obtaining reliable data.

There are number of approaches used to minimize bias in data collection to strengthen a study's validity. The use of reliable instrumentation, well-defined coding protocols and methodology, clear recording and data-entry procedures, and technology all can contribute to enhancing data reliability. After considerable effort is invested in epidemiologic examinations on developing a systematic data collection strategy, many studies fall short when it comes to a plan to maximize examiner or reader performance. Consequently, many studies have a robust systematic data collection plan that minimizes examiner or reader training. Under-valuing training can have unintended consequences that can negatively impact study findings. When a study employs a single examiner or reader, misclassification and low internal consistency could occur. Misclassification can produce random or systematic bias and weak internal consistency will generate poor intra-

* Corresponding author at: 31 Center Drive Suite 5B55 Bethesda, MD, 20892-2190, United States.
E-mail addresses: Bruce.dye@nih.gov (B.A. Dye), michaela.goodwin@manchester.ac.uk (M. Goodwin), roger.p.ellwood@manchester.ac.uk (R.P. Ellwood), Iain.A.Pretty@manchester.ac.uk (I.A. Pretty).

https://doi.org/10.1016/j.jdent.2018.04.020
Received 23 February 2018; Accepted 24 April 2018
0300-5712/ Published by Elsevier Ltd.

examiner reliability. In addition for the potential for misclassification and low internal consistency to exist when multiple examiners or readers are used, weak external consistency, which leads to poor inter-examiner reliability, can occur.

Because one of the most significant contributors to data problems in epidemiologic studies can be a result of examiner error, a robust examiner training plan demonstrating knowledge and skill (calibration) prior to data collection and during data collection is critical. The standard metric for assessing examiner performance has been the Kappa Statistic (k) when assessing the level of agreement for categorical items between examiners. The Kappa Statistic was introduced in 1960 by Cohen to adjust for the possibility of the agreement observed between examiners could occur by chance alone [1]. Conceptually, the k is considered to be a more conservative measure of examiner reliability compared to a basic percent-agreement calculation.

Epidemiologic assessments typically use some form of a visual examination of a participant to make a direct measure or observation. For example, when assessing for periodontal disease an examiner would use a particular periodontal probe and make measures at specific sites around each tooth or when assessing for dental caries the examiner would directly observe the condition of each tooth and record the observation following study protocols. However, when an examiner or reader observes an image, the observation is not based on an indirect examination. Regardless of if data is collected through a “direct’ or an ‘indirect’ examination, the same concerns regarding maximizing internal validity and minimizing examiner error are present. Therefore, rigorous training protocol should be part of any reading scheme of images or digital photographs in epidemiologic studies.

However, it is frequently not clear in the literature that a robust training and evaluation protocol has been incorporated into most studies using digital photography to assess for intraoral conditions. The majority of studies using digital photography were designed to ascertain the validity of the photographic assessment method to the clinical examination method [2–13]. Overall, investigators generally concluded that the photographic method detected significantly more oral disease or conditions than the clinical examination. More importantly, most studies concluded that the photographic method for detecting oral diseases and conditions was a reliable method based on the intra- and inter-examiner statistics calculated. What is clear from the literature is that there has been (1) inconsistency in reporting metric-based performance outcomes from training (inter-ex or intra-ex reliability) prior to actual scoring of digital images, (2) inconsistency of procedurals used to train the photographic scorers, and (3) a focus on the validity assessment of the clinical-photographic methodology and not if the actual training methodologies used to prepare the digital image scorers were valid.

Large scale epidemiologic studies using clinical examination methodologies are resource intensive and logistically complicated. If digital epidemiology has the potential to replace or supplement direct examinations in studies like national examination surveys, we need to demonstrate that examiners can score photographs producing reliable data. To accomplish this, we need to better understand what training techniques should be implemented to maximize examiner performance when scoring digital images and what procedures should be applied to demonstrate that examiner training has achieved adequate, ongoing calibration. Therefore, the aim of this study was to evaluate how to train examiners to assess digital photographs.

2. Methods

2.1. Background

In 2014, a training study was undertaken in Newcastle, UK to assess training methodology and to select a core group of photographic scorers to assess images for dental fluorosis captured during the 2013–2014 National Health and Nutrition Examination Survey (NHANES). The NHANES is a cross-sectional survey designed to monitor the health and nutritional status of the civilian, non-institutionalized U.S. population. The survey consists of interviews conducted in participants’ homes and standardized physical examinations in mobile examination centers (MEC). Additional information on NHANES can be located at http://www.cdc.gov/nchs/nhanes.htm. Digital photographs were taken using a new dual imaging system developed to nearly simultaneously take a quantitative light induced fluorescence (QLF) image and a polarized white light image. Photographs can be difficult to standardize in an epidemiological setting and the potential effects of specular reflection can make reading them difficult. This new system was designed to minimize those challenges. Additional details on the design and performance of this imaging system are available elsewhere [14,15]. Examples of images captured are in the Appendix.

2.2. Protocol

The presence of dental fluorosis was determined by the modified Deans Index (DI) and by the Thylstrup-Fejerskov (TF) Index [16–18]. For purposes of this training study, the same direct examination protocols (DI) used to assess dental fluorosis on NHANES was used [19,20]. Additionally, both the DI and TF criteria were not altered when scoring photograph images. Therefore, when examiners evaluated participants or digital images, the same criteria were applied. Criteria for both scoring methods are shown in the Appendix. Ten examiners with differing experiences in assessing dental fluorosis were selected by the principal investigator (IP). Some of the examiners had participated in previous national surveys assessing dental fluorosis and/or had prior public health research experience involving dental fluorosis. Additionally, two individuals were chosen as reference examiners, one each for DI and TF. Both reference examiners had extensive experience in assessing and training other examiners in their respective dental fluorosis indices. The ten examiners were equally divided into two separate groups with half receiving training and scoring using DI and half for assessing dental fluorosis using TF. Overall, 12 examiners participated in this training study.

This training study had eight components with five of them representing periods when examiner performance was evaluated. Trainees were evaluated using 128 digital images at three different periods and with approximately 40 participants during the other two periods. The sequence of activities and assessment periods are outlined in the Appendix. Examiners were provided with clinical examination scoring criteria for either DI or TF along with some theoretical information pertaining to dental fluorosis prior to training. Examiners were then asked to score 2 maxillary central incisors from a set of 128 digital images. Examiners were later brought together for in-person training, which took place from Friday 26th September to Monday 29th September 2014 in Newcastle (UK). Each reference examiner provided an in-person Power Point review of approximately 100 images. The purpose of this discussion was to reach consensus of scoring using either the DI or TF dental fluorosis criteria. Following this discussion, examiners scored a second set of images (128 maxillary central incisors). During day two and three, trainees and reference examiners examined children age 12–15. All examiners assessed the same set of children on day two (n = 44) and day three (n = 42). For this clinical exercise activity, all examiners including the reference examiners, assessed each participate in a random order. For day four, both reference examiners provided a second review of approximately 100 images. Afterwards, examiners were asked to score 2 maxillary central incisors from a set of 128 digital images. For all photographic scoring activities, all trainees including the reference examiners, scored images on their personal laptops independent of each other at prescribed times.

Scoring of all digital images was done using a secured, web-based system. Trainees logged into the system using 2-level authentication procedures. Once an individual was recognized as a registered user, the system provided only one option to score each image using TF or DI