Developing and Validating Visual Assessment Tools for Use in Medical Device Trials

Alvin Van Orden[1]
[1]FDA, 10903 New Hampshire Ave White Oak, MD

Abstract
Many outcomes rely on a visual assessment by a physician. For example, aesthetic outcomes and the level of bleeding are often judgments and not strict measurements. To design a clinical trial where the primary outcome is a judgment, not a measurement, it is necessary to develop a visual assessment tool or scale. The purpose of the scale is to allow comparisons in the amount of improvement. The scale must be validated, demonstrating that evaluators can use the scale consistently and that the increments on the scale are meaningful. This talk described the process for how to develop these outcomes and how to present them to the FDA for use in future clinical trials. Many of the visual examples in the talk are not presented in this paper.

Key Words: Scale validation, Interrater reliability, Weighted kappa, Medical Devices

1. Introduction

1.1 What is a Visual Assessment Tool?
A visual assessment tool is a scale that is used by evaluators to systematically determine the severity of a particular condition. These evaluators are typically blinded to the treatment received and the time point in the study. A visual assessment tool is not a Patient Reported Outcome, as it is not designed for use by patients. It is designed to be used by trained clinicians. It usually consists of reference photos and text descriptions. In the formal talk at JSM, many examples of different Visual Assessment Tools were given.

1.2 When Should You Use Visual Assessments?
The goal of this talk is not to encourage the use of visual assessments. If there are other easily accessible options that are more objective, they should be used. For example, if you can count the number of hairs on the head, then this count should be used instead of trying to assess from a photo if the subjects' hair looks thicker. Using a visual assessment should not be thought of as a shortcut. Even after you have validated your scale and trained your evaluators, you will still need to provide photographs to the FDA and show success in other endpoints.

Many device areas use visual assessments because there are no other good alternatives. The most common example is aesthetics. Visual assessments are used in dermal fillers, which have scales for a wide range of indications, and for assessing cellulite. It can also be used in determining the severity of bleeding, where it is impractical to stop and measure the level of bleeding. However, it is

not used in all areas where photos are taken. Another example of where it is not used is in toe nail fungus, where a percent of clear nail is measured.

1.3 Satisfaction as Co-primary

In aesthetic devices, the visual assessment tool is officially the primary endpoint, and satisfaction is secondary, due to concerns about the placebo effect. For a first of a kind indication, the only control is a no-treatment arm, so subjects will know if they are receiving treatment or not. In an unblinded study, it is not appropriate to use patient reported outcomes as a primary endpoint. However, it is understood that devices must show high levels of satisfaction. Aesthetic devices often show around 90% satisfaction. This is not officially co-primary because difficult to agree on the definition of success. We have denied applications of devices that met the primary endpoint, based on a visual assessment, but that had low satisfaction rates. Certainly, below 50% satisfaction is a red flag.

2. Validation

2.1 Live vs. Photographic Evaluation

The question often comes up if it is better to validate a scale using photographs or using live evaluations. The rule is that you should validate the tool in the same way you will use it in the clinical trial. Either method (photographs or live) is acceptable, as long as the scale is validated for how it is going to be used. There are tradeoffs no matter which one you use. Photographs are easier to use in validation, but live evaluations are preferred by experts and are felt to be more accurate because they're 3-D. Thus, companies want to use photographs in validation and live evaluations in the clinical trial, but this is inappropriate.

2.2 Subjects in a Validation Study

There is not a set number of subjects that must be in the validation study. Sample size ranges from 55-120. The validation must test the full range of the scale, from least to most severe. However, there doesn't need to be an equal number in each group. It is understood that there may be a natural distribution of subjects, and the validation study can and should reflect the potential pivotal study. This also applies to demographics. The validation study should anticipate clinical study which will determine the labeling. If the product is going to be labeled for a broad use in terms of race, gender, skin type, baseline severity, etc., then this should be reflected in both the pivotal and validation study.

2.3 Validation Analyses

The main validation analysis is a measure of inter- and intra-rater reliability, typically using the weighted kappa statistic. To calculate these statistics, the validation study will need to include at least 3 raters and each needs to rate each subject more than once. These statistics are not the only important measure. We also like to look at the exact agreement, and we check accuracy over the full range of the scale. To do this, we recommend that companies send in all of the validation data, not just summary statistics.

There are two typical interpretations of weighted kappa that get cited from the literature, as seen in the table below.

Table 1: Interpretations of Weighted Kappa in Literature

Literature	Weighted Kappa Coefficient	Interpretation
Fleiss	< 0.40	Poor agreement
	0.40 – 0.75	Fair to Good agreement
	> 0.75	Excellent agreement
Landis and Koch	< 0.20	Poor agreement
	0.20 – 0.39	Fair agreement
	0.40 – 0.59	Moderate agreement
	0.60 – 0.79	Substantial agreement
	0.80 – 1.0	Almost Perfect agreement

Companies will often site the second paper and say that a weighted kappa above 0.6 is sufficient. We expect the weighted kappa above 0.7 or entire confidence interval above 0.6, but no weighted kappa value is a guarantee that the scale will be considered validated. Experience shows that the 0.7 threshold does not tend to be a problem for meaningful scales and well trained evaluators. Experience also shows that the level of agreement in the pivotal study is often lower than that seen in the validation study, which is why there is a need for stronger agreement in the validation study.

2.4 Showing that every interval change is meaningful

It is also important to show that a 1 point or one interval change on the scale is meaningful. This is typically accomplished by showing pairs of photographs representing 0, 1 or 2 point differences and asking evaluators, "Is there a clinically meaningful difference in these photographs?" If the clinicians consistently agree that a 1 point change is meaningful but a 0 point change is not, then this suggests that each change on the scale denotes a meaningful difference. Please note that photos can be used in this stage even if live evaluations will be done in the clinical trial.

2.5 Outlines of Validation studies

In the JSM presentation, graphics were shown that outlined the validation process, as it was completed for several successful validation studies. While that process has largely already been described here, it is important to note that there should be a gap between the first and second evaluation of the same subject, so that evaluators in the validation study do not simply try to remember what they rated the subject the previous time. When the number of subjects is sufficient, the time between evaluations does not have to be overly burdensome, and some companies have completed the validation study over a long weekend in order to reduce the number of subjects that did not return. More common is a two week gap between the evaluations. It also helps if evaluators are only evaluating the specific part of the body that requires visual assessment and are blinded to the identity of the subject.

References

Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382

Landis, J. R. and Koch, G. G. (1977) "The measurement of observer agreement for categorical data" in *Biometrics*. Vol. 33, pp. 159–174