Compartmentalization of Discrete Repeated Measures in Patient-Reported Outcome (PRO) Questionnaires

Saryet Kucukemiroglu[1] and Manasi Sheth[1]
[1]Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993

Abstract
In a public health regulatory setting, it is important for patients to have access to high-quality, safe, and effective medical devices. It is important to ensure that patients and their care-partners stay at the center of the regulatory decision-making process. It is necessary to partner with patients by incorporating the patient perspective as evidence in the decision-making process, including both patient preference information (PPI) and patient-reported outcomes (PROs). Patient-reported outcomes are often relevant in assessing diagnostic evaluations and can be used to capture a patient's everyday experience with a medical device, including experience outside of the clinician's office and the effects of treatment on a patient's activities of daily living. Furthermore, in some cases, PRO measures enable us to measure important health status information that cannot yet be detected by other measures, such as pain. To be useful to patients, researchers, and decision makers, PROs must undergo a validation process to support the accuracy and reliability of measurements from a device. Here, we present a novel approach for analyzing PROs obtained from two examples of diagnostic medical devices.

Key Words: Patient-reported outcomes, categorical data, medical devices

1. Introduction

A patient-reported outcome (PRO) is defined as “any report of the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else” [1]. PROs are useful to assess health conditions, particularly on the physical, mental, and functioning health status self-reported by the patient. It can also be used to measure treatment benefit or risk of a medical product in clinical trials.

PROs can provide evidence of medical product benefit in that they directly measure how patients feel or function. Most clinician reported outcomes are indirect measures based on clinical observations of physical signs or symptoms at one or only several points in time. Direct measures of patient benefit do not require follow-up studies or other external information to understand how the interventions studied affect patient’s feelings, function or survival. By standardizing measurements, PROs can decrease random error and bias, thereby increasing the clarity and precision of endpoints. In treatment trials, PROs are used for the assessment of symptom improvement or resolution. But for prevention trials, PROs are used to define disease onset (symptoms plus laboratory confirmation) and also for the assessment of the intensity and duration of symptoms once disease occurs and the correlation of protection [2]. PROs can be used in pragmatic studies to evaluate the impact of interventions in a real-world setting.

PRO questionnaires may be included in clinical trials to examine and measure treatment effects that cannot be measured clinically, and only through the patient’s own observations. Through PROs, clinicians can gain an understanding of the patient’s perspective of a medical product and whether patients perceive a treatment as effective. These patient self-assessments can as a result provide valuable information that cannot be obtained accurately through clinical examinations and/or interview questions in a clinical setting [3].

"PRO instruments measure concepts ranging from the state of discrete symptoms or signs (e.g. pain severity or seizure frequency) to the overall state of a condition (e.g. depression, heart failure, angina, asthma, urinary incontinence, or rheumatoid arthritis), where both specific symptoms and the impact of the condition (e.g. on function, activities or feelings) can be measured, to feelings about the condition or treatment (e.g. worry about getting worse, having to avoid certain situations, feeling different from others). PRO concepts can be general (e.g. improvement in physical function, psychological well-being, or treatment satisfaction) or specific (e.g. decreased frequency, severity or how bothersome the symptoms are). PRO concepts can also be generic (i.e. applicable in a broad scope of diseases or conditions as in the case of physical functioning), condition-specific (e.g. asthma-specific), or treatment-specific (e.g. measures of the toxicities of a class of drugs such as interferons or opioids)" [3].

2. Background

2.1 Mean Change from Baseline in PROs

Patient Reported Outcomes (PROs), usually in the form of administered of self-completed questionnaires, are widely used in medical outcome studies. The data produced by these questionnaires are usually of ordinal level of measurement. At the same time, most patient reported outcome studies depend on the calculations of means, standard deviations, change scores, and concepts such as Minimally Important Difference (MID) or effect sizes. However, ordinal levels of measurements do not support the mathematical operations needed to calculate these types of statistical measures. It is known widely that if there are several items or survey questions that are measured using an ordinal scale, then the sum of the scores may not preserve ordinal properties [4, 5]. Despite these constraints, most analyses ignore these limitations and widely report means and MID thereby running the risk of making an incorrect inference from data based upon PROs [5].

By including PROs in a study, one can obtain ordinal estimates of a patient's own assessment of certain symptoms or attributes measured. In the analysis, appropriate nonparametric statistics should be performed. However, if the data is not ordinal, the data may convert to an ordinal scale using a Rasch model that is based on the Additive Conjoint Measurement theory [5].

2.2 Anchor and Distribution Based Approaches

Even though the anchor-based approach is simplistic and widely used for the purpose of the minimally important difference (MID), one cannot make a distinction between an improvement and deterioration. Small sample sizes are generally a major limitation when using anchor categories. If the number of patients is substantially small for each category, then the resulting MID is not reliable and the estimator is not as robust. The occurrence of missing data can also bias the MID analysis [6].

It is important to determine if the interpretation of patient reported outcome is used to make an inference for an individual or a specific patient population. In the population perspective, even small differences may be substantial or significant as a large number of individuals may get affected. However, at an individual level, large differences may need to be substantial in order to see the effect. Another important determination remains in the fact that the degree of change that is needed to stimulate clinicians to consider a clinical intervention.

The first set of limitations of anchor-based methods is that the application of different anchors or anchor types may produce different values of MCID. Another potential limitation is "a potential discordance of defined MCID values based on whether data collection of the anchor was prospective versus retrospective, and possibility that the MCID as determined by anchor-based methods falls within the instrument's random variation, and the susceptibility of some ratings to recall bias" [7].

The first set of limitations of distribution-based methods include that the application of various distribution-based approaches will result in different definitions of MCID. More importantly, distribution-based methods are limited by their ability to define only a minimal value below which the outcome measure occur only due to a measurement error and it doesn't provide any information on clinical importance. Thus, the importance of MCID, ability to define the clinical importance of a given change in the outcome scores separate from their statistical significance, is lost when utilizing anchor and distribution-based methods. [7].

2.3 Ordinality of Data

Usually, a majority of health status indicators of interest are measured on an ordinal scale for which the quantitative differences between levels is unclear or unknown. Attributes such as perceived health status, functional independence, mobility, and pain are most appropriately captured using an ordinal scale. However, while data may be collected using an ordinal scale, they are rarely analyzed as such. A majority of the methods for analyzing ordinal data are done by altering the nature of the data: collapsing the ordinal scale to a dichotomous one, treating it as a nominal, or considering it to be continuous.

There is a significant loss of information when the ordinality of ranked data is not fully utilized. Chi-square tests of trend, t-tests, analysis of variance, and analysis of covariance are also usually utilized in the analysis of ordinal data. However, they require that the ordinal categories and therefore, the distances between them, be quantified and treated as continuous. In general, the classification used to quantify the ordinal categories can have a substantial effect on the generalizations made from the results and can produce misleading results.

Ordinal regression methods have been developed theoretically and presented in the statistical literature including recent work on sample size estimation and models for dependent observations. While, in most cases they are hailed as a breakthrough in analyzing ranked outcomes, there is a controversy regarding their use in problems of classification. Several varieties of ordinal models have been developed. The purpose of this paper is to motivate the use of these models by presenting the methodology in a form that is readily useable by the statisticians, epidemiologists and the clinical researchers.

In this paper, we are proposing to quantify the increment or detriment in the ordinal scale that is easily interpretable by clinicians and epidemiologists. In Section 3, we will present a description of our data structure and statistical approach.

3. Methodology

In this research, we propose a methodology used to analyze repeated measurements of an ordinal PRO variable over time. The basis of this methodology involves calculating the agreement in scores at each consecutive timepoint using a confusion matrix. For notation purposes, let $i = 1, \dots, n$ represent the subject number, $\mathrm{t} = 1, \dots, n_t$ represent the time of PRO assessment, and x_{it} be the PRO rating for subject i at time t. At each consecutive timepoint t and $t + 1$, the agreement in PRO ratings are calculated using a confusion matrix. For each matrix comparing each consecutive time point, the proportion of subjects that have reached optimal condition (p_{Ot}), seen improvement (p_{It}), and seen no improvement or are in worse condition (p_{Wt}) are calculated. This confusion matrix is created for all consecutive time points in the study and the trend of the three proportion measures (p_{Ot}, p_{It}, p_{Wt}) are analyzed through time.

Table 1: Confusion matrix comparing agreement at two consecutive timepoints

		Week t+1 Rating				
Week t Rating		1	2	3	4	Total
	1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
	2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
	3	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
	4	n_{41}	n_{42}	n_{43}	n_{44}	$n_{4.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{..}$

If $i = 1$ and j= 1 then the patient is optimal
If $i - j > 0$ then the patient is improving
If $i - j = 0$ or if $i - j < 0$ then patient has seen no improvement or is in worse condition

Table 1 shows an example of a confusion matrix comparing time $\boldsymbol{t}$ rating to time $\boldsymbol{t+1}$ rating for patients in a study. We assume a PRO health quality variable such as pain is evaluated on an ordinal scale where 1 indicates no pain or optimal condition and 4 indicates severe pain that the patient is experiencing. This data matrix can be created for any ordinal scoring of pain or other symptom assessment. Each cell in the confusion matrix $\boldsymbol{n_{ij}}$ indicates the number of subjects in that cell. For interpreting $\boldsymbol{n_{ij}}$, if $\boldsymbol{i = 1}$ and j $\boldsymbol{= 1}$ then the patient is in optimal condition since the patient indicated no pain for two consecutive time periods. The optimal condition in the matrix is shown in yellow. If $\boldsymbol{i - j > 0}$ then the patient is improving since their pain level is becoming better; this is indicated in green in the matrix. If $\boldsymbol{i - j = 0}$ or if $\boldsymbol{i - j < 0}$ then the patient has seen no improvement or is in worse condition and this is indicated in orange in the matrix. The proportion of subjects that have reached optimal condition, seen improvement, and seen no improvement or have experienced worse condition are then compared by analyzing the trend of these three proportions in time using ordinary linear regression.

4. Results

We implemented this methodology in simulated data of patients with benign prostate hyperplasia which is a condition in which the flow of urine is blocked due to an enlarged prostate. We assume that this is a randomized, double blinded study comparing device 1 and device 2 that are indicated to treat this condition in newly enrolled patients with benign prostate hyperplasia. Approximately 1700 subjects were randomized 1:1 to device 1 and 2. PRO data of pain and sleep assessment are collected from patients where these variables are rated on a scale from 1-4 where 1 indicates no pain or that the patient had no trouble sleeping and 4 indicates extreme pain or that the patient had trouble sleeping. This PRO data is collected from each subject every 4 weeks and the PRO data was simulated over a 2-year period.

Figure 1: Differences in the PRO sleep variable between two medical devices evaluated over two years

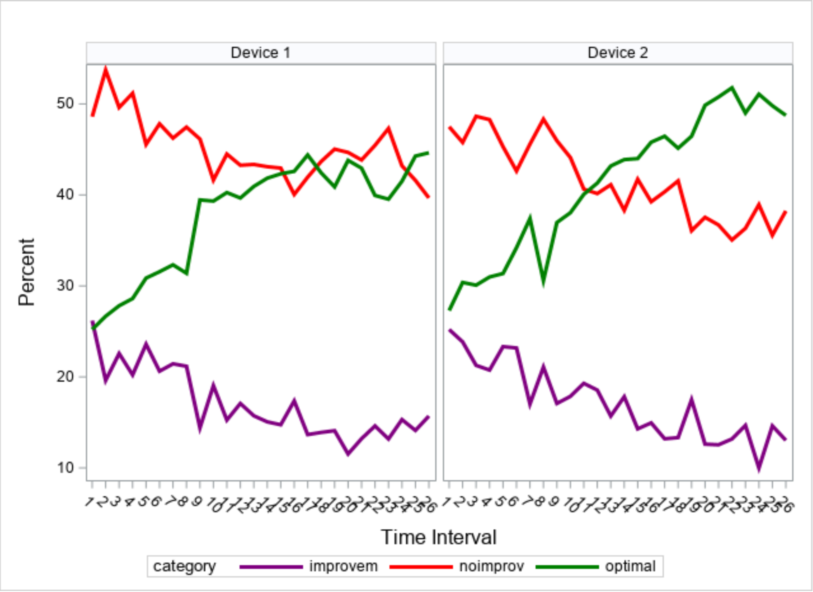

Table 2: Evaluation of the differences in sleep condition between two medical devices

Test of $\beta_{Device1}=\beta_{Device2}$				
Condition	Parameter Estimate	Standard Error	t Value	p-value
Optimal	0.281	0.100	2.81	0.0072
Improve	-0.072	0.076	-0.95	0.3482
Worse	-0.208	0.083	-2.49	0.0161

*T-test was used to test the equality of regression slopes for each sleep condition between device 1 and 2

Figure 2: Differences in the PRO pain variable between two medical devices evaluated over two years

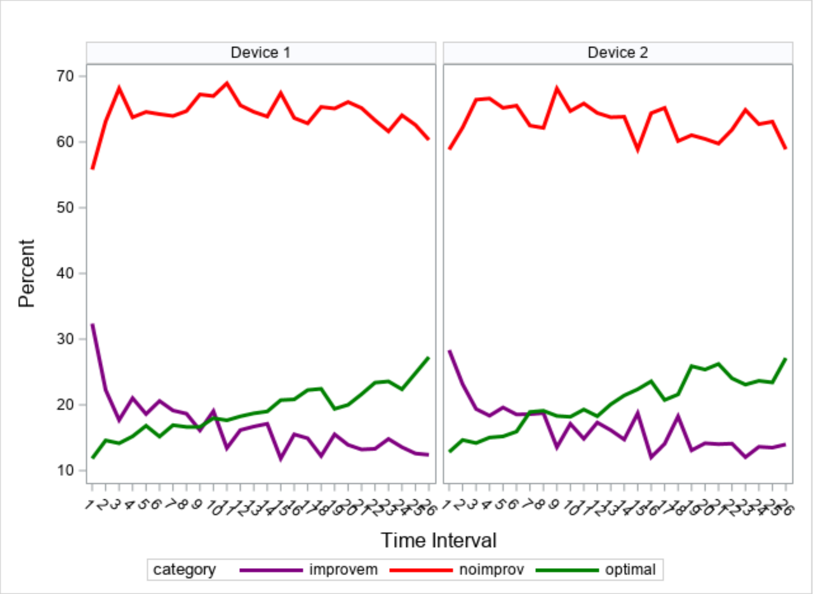

Table 3: Evaluation of the differences in pain condition between two devices

Test of $\beta_{Device1}=\beta_{Device2}$				
Condition	Parameter Estimate	Standard Error	t Value	p-value
Optimal	0.041	0.049	0.84	0.4039
Improve	0.069	0.096	0.72	0.4729
Worse	-0.101	0.095	-1.06	0.2957

*T-test was used to test the equality of regression slopes for each sleep condition between device 1 and 2

Figure 1 and Table 2 display the results when examining the proportions of the PRO sleep variable between devices 1 and 2 over a two-year time period. In the graph, green curve represents optimal condition, purple curve represents improvement, and red curve represents no improvement or worsening condition. Based on the graphs, the proportion for optimal condition increases in time for both devices; however, it appears that there is more improvement when patients use device 2 than device 1 since the slope appears larger. This can be verified when testing the equality of the slopes where the p-value for the test is 0.0072 which may indicate device 2 is better in improving sleep for patients than device 1. The proportion for improvement declines in time for both devices since one would expect the treatment to convert more patients to optimal condition in time. The proportion for no improvement or worsening condition declines in time for both devices which is expected with an effective treatment. When testing the equality of the slopes for no improvement, the slopes for devices 1 and 2 are significantly different which may also indicate device 2 is better in improving sleep in patients than device 1.

Figure 2 and Table 3 display the results when examining the proportions of the PRO pain variable between devices 1 and 2 over a two-year time period. Based on the results of the graph and testing of slopes there was no significant difference in pain improvement over time for subjects using device 1 vs device 2.

5. Discussion

The adequacy of a PRO instrument as a measure to support medical product claims depends on its development history and demonstrated measurement properties. Industries are encouraged to identify all endpoints (primary as well as secondary) early in product development, before studies are initiated, to provide the basis for product approval or claim substantiation. This will allow “appropriate time for PRO instrument identification, modification or if necessary, new instrument development” [1].

Physical, emotional, economic and cognitive strain on patients can drastically impact the quality and quantity of PRO data. The frequency and the timing of PRO assessments in a protocol and the severity of the illness or toxicity of the treatment under study determines the extent of the respondent burden [1]. The duration of the study must be adequate to support the proposed claim and to assess a durable outcome in the disease or condition being studied. A PRO instrument could be the primary endpoint measure of the study as a co-primary endpoint measure of the study, a co-primary endpoint measure in conjunction with other objective or physician-rated measurements, or a secondary endpoint measure whose analysis would be considered according to a hierarchical sequence.

Because each PRO item or domain often can represent an endpoint that could imply a distinct claim on its own, thorough planning is recommended to avoid substantial increase in Type I error from multiple endpoints. If it is important in a study to demonstrate that PROs have the same indicator effect as other measures of treatment benefit, then statistical procedures can be considered to minimize the impact of multiple endpoint comparisons.

This proposed methodology can be useful in analyzing improvement, deterioration, and optimization of patients’ well-being and functionality over time. Compared to other methodologies mentioned, this methodology may provide more interpretable results when analyzing collective data from categorical PROs. Only three categories of improvement, no improvement or worse condition, and optimal condition are used to analyze patients' conditions throughout time whereas other methods use interval scale analysis in interpreting results. There is also less loss of information since this methodology assumes three different classifications or statistics from the matrix structure whereas a mean change from baseline analysis results in one statistic for analysis. With three different statistics, one can analyze and pinpoint easily in time when most patients in a study have overall seen improvement, have reached optimal condition, or have seen worsening of a condition. This can give more information for researchers on the effectiveness of a medical product through time which can aid in determining how long patients can see a benefit in treatment. This methodology does not require one to make any assumptions.

In future work, we plan to develop a multi-level non-parametric methodology taking into account repeated measurements from subjects and comparing this to established parametric methods. In addition, instead of testing differences in regression slopes, testing differences in average proportions may be applied to determine which device provides better treatment. Lastly, we plan to extend this work to multivariate responses to examine the effect of a combination of PRO variables collectively on a treatment.

Acknowledgements and Disclosure

The authors would like to acknowledge the FDA CDRH Division of Clinical Evidence and Analysis for allowing us the opportunity to work on this research. We would like to note that the views, findings, and conclusions in this report are those of the authors and do not necessarily represent those of the FDA.

References

1. FDA Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. 2009. Available at: http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf. Accessed September 11, 2020.
2. Powers III, J. H., Howard, K., Saretsky, T., Clifford, S., Hoffmann, S., Llorens, L., & Talbot, G. (2016). Patient-reported outcome assessments as endpoints in studies in infectious diseases. *Clinical Infectious Diseases*, *63*(suppl_2), S52-S56.
3. US Department of Health and Human Services FDA Center for Drug Evaluation and Research, FDA Center for Biologics Evaluation and Research, & FDA Center for Devices and Radiological Health. (2006). Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health and Quality of Life Outcomes*, *4*, 1-20.
4. Forrest M, Andersen B; Ordinal Scale and statistics in medical research, BMJ 1986, 202, 537 – 538.
5. Horton, M., & Tennant, A. (2011). Patient Reported Outcomes: misinference from ordinal scales?. *Trials*, *12*(S1), A65.
6. Ousmen, A., Touraine, C., Deliu, N., Cottone, F., Bonnetain, F., Efficace, F., Bredart, A., Mollevi, C. & Anota, A. (2018). Distribution-and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. *Health and quality of life outcomes*, *16*(1), 228.
7. Rai, S. K., Yazdany, J., Fortin, P. R., & Aviña-Zubieta, J. A. (2015). Approaches for estimating minimal clinically important differences in systemic lupus erythematosus. *Arthritis research & therapy*, *17*(1), 143.