# Non-Parametric Analysis of Patient-Reported Outcomes Using Compartmentalization Method

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 quite necessary to ascertain that the patients and their care-partners stay at the center of the regulatory decision-making process. In order to do so, it becomes 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 two-stage non-parametric approach for analyzing PROs using compartmentalization method. Two illustrations from diagnostic medical devices are used to test this approach.

**Key Words:** Patient-reported outcomes, categorical data, non-parametric, compartmentalization

## 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 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. If the statistical measurements are standardized, PROs can decrease random error and bias, thereby increasing 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 are sometimes included in clinical trials to examine and measure treatment effects that cannot be measured clinically, and only through the patient's own observations or reporting. Through PROs, clinicians can gain an understanding of the patient's perspective of a medical product (device or drugs) 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].

In this paper, we introduce some of the commonly used methods for analyzing PRO data as well as introduce a novel approach using a two-stage non-parametric analyses using compartmentalization method. The compartmentalization method was introduced previously by Kucukemiroglu and Sheth [8]. The compartmentalization method is a non-parametric approach to analyzing categorical data for two or more time points or raters. This can be further extended to be used with other non-parametric approaches to analyze the data obtained from survey instruments as such. We refer to our previous paper (Kucukemiroglu and Sheth [8]) for a discussion on the commonly used approaches on analyzing PRO data, i.e. mean change from baseline in PROs as well as anchor and distribution based approaches.

## 2. Background

Usually, an ordinal scale is used for a majority of health status indicators of interest 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, even though data are collected on an ordinal scale, they are rarely analyzed using an ordinal scale. 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 addressed as a breakthrough in analyzing ranked outcomes, there is a controversy regarding their use in problems of classification. Several varieties of non-parametric ordinal models are in the development process. 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 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 using Compartmentalization Method that was introduced in Kucukemiroglu and Sheth (2020) [8]. In Section 3, we will present a description of our data structure methodology using Compartmentalization Method and statistical approach using Two-Stage Non-

Parametric Analyses. For further discussion of the background and comparative methods to Compartmentalization Method, please refer to Kucukemiroglu and Sheth (2020).

## 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, $t = 1, \dots, n_t$ 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 so that there will be t - 1 comparisons as well as contingency tables and the trend of the three proportion measures ($p_{Ot}, p_{It}, p_{Wt}$) are analyzed through time.

*Stage 1: Compartmentalization Method/Approach*

Table 1: Confusion matrix comparing agreement at two consecutive timepoints

| | | Week t+1 Rating | | | | |
|---|---|---|---|---|---|---|
| Week t Rating | Severity of pain | 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 $t$ rating to time $t + 1$ 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 $n_{ij}$ indicates the number of subjects in that cell having that particular combination of severity of symptom assessment. For interpreting $n_{ij}$, if $i = 1$ and j = 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 $i - j > 0$ then the patient is improving since their pain level is becoming better; this is indicated in green in the matrix. If $i - j = 0$ or if $i - j < 0$ then the patient has seen no improvement or is in worse condition and this is indicated in orange in the matrix. Thus, the data reduces from 16 individual cells/categories for analyses to three categories. If only two categories are considered (that is, improvement and deterioration), then the data reduces to that of binomial probabilities and can be analyzed using logistic regression. Similarly, one can also attempt to analyze the data using a very similar categorization, i.e. consistent, improvement and deterioration, and similar methods can be applied. In doing so, there is a lesser loss of information compared to when the confusion matrix is reduced to just one statistic such as a mean change or mean difference. One can also compartmentalize using 4 categories such as optimal, consistent, improvement and deterioration and apply two-stage non-parametric approach or ordinal regression methods to analyze the data.

*Stage 2: Sen-Adichie Statistic*

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 non-parametric analyses. The Sen-Adichie statistic is a distribution free procedure used to test for parallelism of two or more regression lines. In our example, this statistic is used to compare each of the proportions in time between two medical devices to determine which device generally improves the quality of life in patients. We test the null hypothesis,

$$H_0: \beta_{Device\ 1} = \beta_{Device\ 2}$$

versus the alternative hypothesis,

$$H_1: \beta_{Device\ 1} \neq \beta_{Device\ 2}$$

For notation purposes, let $k = 1, \dots, d$ where $d$ is the number of devices examined and $t = 1, \dots, n_t$ denote the time of the PRO assessment. For the k[th] line, we observe the value of the k[th] response random variable $Y_k$ that represents one of the three proportion measures $(p_{Ot}, p_{It}, p_{Wt})$ being examined. $Y_k$ is observed at each of the $n_{tk}$ fixed levels, $x_{k1}, \dots, x_{kn_{tk}}$ of the k[th] independent predictor variable $x_k$. Thus for the k[th] line, $k = 1, \dots, d$, we obtain a set of observations $Y_{k1}, \dots, Y_{kn_{tk}}$, where $Y_{kt}$ is the value of the response variable $Y_k$when $x_k = x_{kt}$.

To construct the Sen-Adichie V statistic, we first align each of two regression samples. Let $\bar{\beta}$ be the pooled least squares estimator for the common slope $\beta$ under the null hypothesis $H_0$ given by

$$\bar{\beta} = \frac{\sum_{k=1}^{d} \sum_{t=1}^{n_{tk}} (x_{kt} - \bar{x}_k) Y_{kt}}{\sum_{k=1}^{d} \sum_{t=1}^{n_{tk}} (x_{kt} - \bar{x}_k)^2},$$

where

$$\bar{x}_k = \sum_{k=1}^{n_{tk}} \frac{x_{kt}}{n_{tk}}, \ for\ k = 1, \dots, d$$

For each of the d regression samples, the aligned observations are computed as

$$Y^*_{kt} = \left(Y_{kt} - \bar{\beta} x_{kt}\right), \quad k = 1, \dots, d \quad t = 1, \dots, n_{tk}$$

The aligned observations $Y^*_{kt}$ are ordered from least to greatest separately within each of the two regression samples. Let $r^*_{kt}$ denote the rank of $Y^*_{kt}$ in the joint ranking of the aligned observations $Y^*_{k1}, \dots, Y^*_{kn_{tk}}$ in the k[th] regression sample. The Sen-Adichie V statistic can then be computed as

$$V = 12 \sum_{k=1}^{d} \left[\frac{T^*_k}{C_k}\right]^2$$

where

$$T_k^* = \sum_{t=1}^{n_{tk}} [(x_{kt} - \bar{x}_k) r_{kt}^*] / (n_{tk} + 1) \quad k = 1, \ldots, d$$

and

$$C_k^2 = \sum_{t=1}^{n_{tk}} (x_{kt} - \bar{x}_k)^2 \qquad k = 1, \ldots, d$$

The Sen-Adichie V statistic is compared to the chi-square distribution with d-1 degrees of freedom. At the significance level $\propto$, if $V \geq X^2_{d-1,\alpha}$, the null hypothesis $H_0$ is rejected, otherwise it is not rejected [9].

## 4. Illustration

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

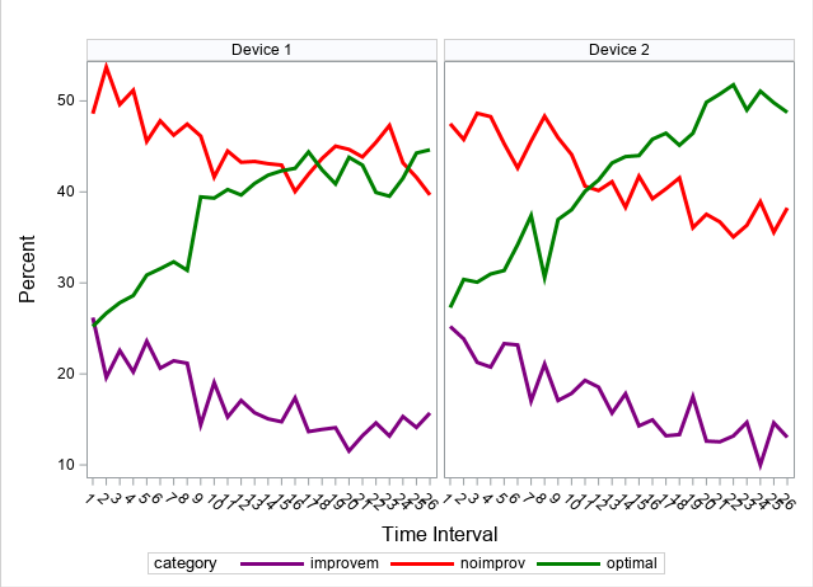


Table 2: Evaluation of the differences in sleep condition between two medical devices

| Test of $\beta_{Device1}=\beta_{Device2}$ | | | |
|---|---|---|---|
| Condition | V | df | p-value |
| Optimal | 6.974 | 1 | 0.0083 |
| Improve | 0.800 | 1 | 0.3712 |
| Worse | 4.556 | 1 | 0.0328 |

* Sen-Adichie V statistic was used to test the equality of 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

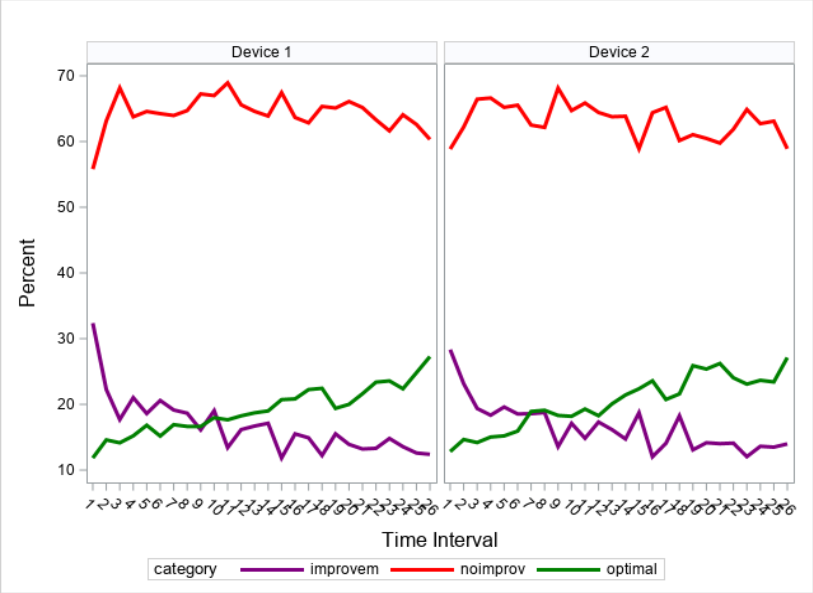


Table 3: Evaluation of the differences in pain condition between two medical devices

| Test of $\beta_{Device1}=\beta_{Device2}$ | | | |
|---|---|---|---|
| Condition | V | df | p-value |
| Optimal | 0.273 | 1 | 0.6017 |
| Improve | 2.202 | 1 | 0.1378 |
| Worse | 1.410 | 1 | 0.2351 |

* Sen-Adichie V statistic was used to test the equality of slopes for each pain 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 using the Sen-Adichie V statistic where the p-value for the test is 0.0083 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

In order to ensure the adequacy of a PRO instrument as a measure to support medical product claims, it is necessary to ensure that there is appropriate development history and demonstrated measurement

properties. Sponsors 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].

The quality and quantity of PRO data depends on the physical, emotional, economic and cognitive strain on patients. 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 studies. A PRO instrument could be the primary endpoint measure of the study as a co-primary endpoint measure of the study, in conjunction with other objective or physician-rated measurements, or a secondary endpoint measure whose analyses would be considered according to a hierarchical sequence of the endpoints being claimed in the labeling.

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 worsening of 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.

The two-stage non-parametric approach is useful and efficient, as it makes fewer assumptions and their applicability is much wider than the corresponding parametric methods. In particular, they may be applied in situations where less is known about the application in question. Also, due to the reliance on fewer assumptions, non-parametric methods are more robust. Furthermore, these methods are more simplistic. In certain cases, even when the use of parametric methods is justified, non-parametric methods may be easier to use. Due both to this simplicity and to their greater robustness, non-parametric methods usually leave less room for improper use and misunderstanding. The only disadvantage is that they might have less power, and hence, they may require a much larger sample size to draw conclusions with the same degree of confidence. However, there is much less loss of information since the proposed two-stage methodology assumes three different classifications from the matrix structure whereas the mean change from baseline results in one statistic, it doesn't capture what happens in each category of PRO. Furthermore, instead of testing differences in regression slopes, testing differences in average proportions may be applied to determine which product provides better treatment.

## Acknowledgements and Disclosure

The authors would like to acknowledge the FDA CDRH Division of Clinical Evidence and Analysis (DCEA) 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.
8. Kucukemiroglu, S., Sheth, M. Compartmentalization of Discrete Repeated Measures in Patient-Reported Outcomes Questionnaires, In JSM Proceedings, Biometrics Section, Alexandria, VA: American Statistical Association. 2593 – 2599.
9. Hollander, M., Wolfe, D. A. & Chicken. E. (2013). Nonparametric statistical methods (Vol. 751). John Wiley & Sons.