"Science Stays True Here"
Journal of Mathematics and Statistical Science (ISSN 2411-2518, USA), Vol.11, Issue 11, 87-96 | Science Signpost Publishing

Model-Free Organization of Patient Reported Outcomes Data: Geometrical Rep-resentation of the Modified Compartmen-talization Method

Manasi Sheth[1,2] and N. Rao Chaganty[3]

1. *Department of Mathematics, University of Wisconsin - Whitewater, WI, USA*
2. *Department of Computer Science, University of Wisconsin - Whitewater, WI, USA*
3. *Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA, USA*

Received: August 30, 2025 / Accepted: September 28, 2025 / Published: November 25, 2025.

Abstract: There is a recent advancement in the field of mathematics and statistics to understand the geometry or connectedness of the data due to the massive amounts of data being generated. The data provided for analyses are usually very large and need to be organized and minimized in order to make it more useful and meaningful. In biostatistics or medical field, it is important for patients to have access to high-quality, safe and effective and/or efficacious medical products. 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 the patients by incorporating the patient perspective as evidence in the decision-making process, including patient-reported outcomes (PROs). PROs are often relevant in assessing diagnostic evaluations and can be used to capture a patient's everyday experience with a medical product, including experience outside of the clinician's office and the effects of the treatment on a patient's activities of daily living and functionality. In some cases, PRO measures enable us to measure important health status information that cannot yet be detected by other measures, such as pain and mobility. Here, we present the geometrical representations of a novel approach of analyzing PROs using Modified Compartmentalization Method.

Key words: Compartmentalization Method, Patient Reported Outcomes, Geom-etry, Transition Proportion Matrix, Repeated Measures, Longitudinal Data

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. PROs are self-reported by the patients on the physical, mental and functioning health status and are useful to evaluate health conditions. In clinical trials, they are also useful to measure benefit or risk of a medical product such as a novel treatment or medical device.

PROs can provide evidence of the benefit of the medical product 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 statistical measurements are standardized, PROs can decrease random error and bias,

thereby increasing the clarity and precision of the endpoints. In treatment trials, PROs are used for the assessment of symptom improvement or resolution. But for prevention trials, PROs are used to define the 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. 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 a patient perceives 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.

PRO instruments measure concepts ranging from the state of discretized signs or symptoms (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).

Generally, for most health status indicators of interest for which the quantitative differences between levels are unclear or unknown, an ordinal scale is used. Characteristics such as perceived health status, functional independence, mobility, and pain are most appropriately obtained using an ordinal scale. However, the data are rarely analyzed on an ordinal scale. Most of the methods for analyzing ordinal data are done by altering the nature of the data, that is, collapsing the ordinal scale to a dichotomous one or considering it to be continuous. When the ordinality of ordered data is not fully utilized, there is a significant loss of information.

There has been very few papers that discusses methodologies for analyzing patient reported outcomes (see Kucukemiroglu and Sheth (2020 and 2021); Detry and Ma (2016), Prakash et al. (2008)). In this paper, we are proposing to quantify the increment or detriment in the ordinal scale pf the patient reported outcomes that is easily interpretable by the clinicians and epidemiologists so that it is clinically as well as mathematically/statistically meaningful. In this paper, we introduce a novel approach using the Modified Compartmentalization Method is a non-parametric dimension reduction approach to analyzing categorical data for two or more time points or raters. The geometrical implications of this approach will also be discussed.

2. Methodology Section

2.1 Data Format of Patient Reported Outcomes PROs come in different forms and formats, collected either digitally or nondigitally. Usually, patient

Table 1: Cross-tabulation of responses for Treatment A

	Visit 2				
Visit 1	1	2	...	k	Total
1	n_{a11}	n_{a12}	...	n_{a1k}	$n_{a1.}$
2	n_{a21}	n_{a22}	...	n_{a2k}	$n_{a2.}$
.	...	...	...	...	...
k	n_{ak1}	n_{ak2}	...	n_{akk}	$n_{ak.}$
Total	$n_{a.1}$	$n_{a.2}$	...	$n_{a.k}$	$n_{a..}$

characteristics or covariates such as patient demographics, socio-economic factors, etc. are also collected during the study. Patient-reported outcomes can be discrete or continuous. Here, we will concentrate only on the patient - reported outcomes that are of ordinal nature, such as those with Likert scale. The format of the data is usually submitted as follows: Let y_i be the ordinal response for subject i, $i = 1, ..., n$. Let $x_i = (x_{i1}, ..., x_{il})$ be the corresponding set of covariate(s) associated with the subject i. For this paper, we will especially concentrate on the geometrical representations and dimension reduction of the PROs such as severity level of pain when there are repeated measurements over time, i.e. when there are ordinal responses obtained from the patients for multiple visits.

2.2 Methodology and Illustration In this paper, we have proposed a methodology used to analyze repeated measurements of an ordinal PRO variable over time such as pain, nausea, mobility, etc. The basis of this methodology involves calculating the agreement or discrepancy in scores between each consecutive time point using a cross-tabulated matrix. For notation purposes, let $i = 1, ..., n$ represent the subject i; $t = 1, ..., n_t$ represent the time of the PRO evaluation, and x_{it} be the PRO rating for subject i at time t. At each consecutive time point t and $t + 1$, the agreement in PRO ratings are calculated using a cross-tabulation between the ordinal responses. For each matrix comparing each consecutive time point, the proportion of each subjects or patients that have reached consistency (p_{ot}), seen improvement (p_{It}), and and seen deterioration or are in worse condition (p_{Wt}) are calculated. This method is called Modified C ompartmentalization M ethod. T his c ross-tabulation i s c reated for all consecutive time points in the study so that there will be $t - 1$ consecutive comparisons as well as contingency tables and the trend of the three proportion measures are analyzed through time. Note that usually PRO survey data are usually provided in a long form as depicted in Table 1.

Suppose we are evaluating two treatments. For each of the two treatments, we create a cross-tabulated table as shown below in Table 1.As represented in the Table 1, there are k ordinal responses for each visit creating a square cross-tabulation or square matrix of the data with the count or frequency n_{ajk} going from visit 1 to visit 2. Similarly, a table for treatment B can be created for the k ordinal responses with the count or frequency n_{bjk} for visit 1 to visit 2 as shown below in Table 2. Doing this for t number of consecutive visits, we can create $t - 1$ square matrices or cross-tabulations.

Figure 1 represents an example cross-tabulation for ordinal responses for one of the treatments for when there are 5 visits.

Table 2: Cross-tabulation of responses for Treatment B

	Visit 2				
Visit 1	1	2	...	k	Total
1	n_{b11}	n_{b12}	...	n_{b1k}	$n_{b1.}$
2	n_{b21}	n_{b22}	...	n_{b2k}	$n_{b2.}$
.	...	...	...	...	...
k	n_{bk1}	n_{bk2}	...	n_{bkk}	$n_{bk.}$
Total	$n_{b.1}$	$n_{b.2}$	...	$n_{b.k}$	$n_{b..}$

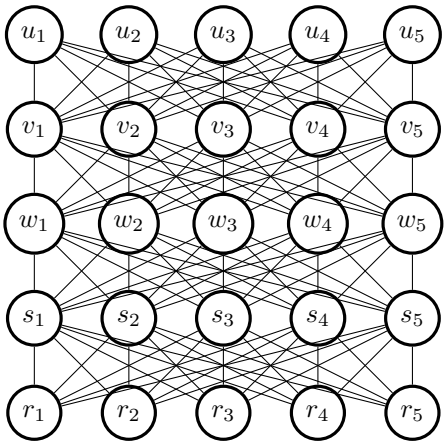


Figure 1: Multipartite Graph using Cross-tabulated Responses.

In the complete multipartite graph depicted in Figure 1, the letters with the subscript, i.e. the nodes, u_1, u_2, u_3, u_4, and u_5 represent the most commonly used ordinal response elements at visit 1. Similarly, the ordinal response elements at visit 2 can be written using the following notations: v_1, v_2, v_3, v_4, and v_5. The notations w_1, w_2, w_3, w_4, and w_5 represent the ordinal response elements at visit 3; s_1, s_2, s_3, s_4, and s_5 represent the ordinal response elements at visit 4; and, r_1, r_2, r_3, r_4, and r_5 represent the ordinal response elements at visit 5. Each of edges represent the possibility of transitioning from one response to the other OR staying consistent in the ordinal response. This is done for several visits.

Note that a "multipartite graph" is a graph where the vertex set can be divided into multiple independent sets, meaning that no edges exist between vertices within the same set, and all edges must connect vertices from different sets; essentially, it's a generalization of a bipartite graph where you can have more than two distinct sets of vertices. A bipartite graph is a graph where the vertices can be divided into two disjoint sets such that all edges connect a vertex in one set to a vertex in another set. There are no edges between vertices in the disjoint sets. By definition, a graph $G = (V, E)$ is called $r - partite$ if V admits a partition into r classes such that every edge has its ends into different classes; vertices in the same partition class must not be adjacent (See Diestel (2020)).

2.3 Dimension Reduction Suppose we look at Table 2 which represents the cross-tabulation of responses for Treatment A. For each matrix or cross-tabulation comparing each consecutive time point(s) of such tables, the proportion of subjects or patients that have reached consistent condition (p_{g0t}), seen improvement (p_{gIt}), and seen no improvement or are in worse condition (p_{gWt}) are calculated using the following dimension reduction technique. A clinically meaningful strategy is to calculate (p_{g0t}) using the diagonal elements of the confusion matrix where the response from visit 1 to visit 2 is the same; (p_{gIt}) is calculated using the upper diagonal elements of the confusion matrix where the response from visit 1 is lower than the response from visit 2; and, (p_{gWt}) using the lower diagonal elements of the confusion matrix where the response from the visit 1 is higher than the response from visit 2. There are several other strategies that can be utilized for computing consistency proportion(s), improvement proportion, and deterioration proportions using this confusion matrix, however for this paper we will utilize this strategy for the illustration purposes.

The above two tables (1 and 2) can be written using conditional probabilities transitioning from one visit to the subsequent visit. Thus, we can write p_{aj1} in terms of conditional probabilities: $p_{a11} = P(T_2 = i|T_1 = i)$, $p_{a21} = P(T_2 > i|T_1 = i)$, and $p_{a31} = P(T_2 < i|T_1 = i)$. Similarly, for treatment B, using the matrix or the cross-tabulation from visit 1 to visit 2, we get the following conditional probabilities: $p_{b11} = P(T_2 = i|T_1 = i)$, $p_{b21} = P(T_2 > i|T_1 = i)$, and $p_{b31} = P(T_2 < i|T_1 = i)$. If there are more treatments, we can easily deduce such formulae for the conditional probabilities.

Therefore, extending this pattern to the visit-to-visit comparison for 5 visits for instance, we get the following conditional probabilities:

- $p_{a12} = P(T_3 = i|T_2 = i)$, $p_{a22} = P(T_3 > i|T_2 = i)$, and $p_{a32} = P(T_3 < i|T_2 = i)$ for matrix with Visit 2 and 3 for Treatment A.
- $p_{b12} = P(T_3 = i|T_2 = i)$, $p_{b22} = P(T_3 > i|T_2 = i)$, and $p_{b32} = P(T_3 < i|T_2 = i)$ for matrix with Visit 2 and 3 for Treatment B.

Table 3: Edges and Nodes of Geometrical Figure 1 for Different Scenarios

No. of Visits	No. of Ordinal Responses	(No. of Nodes, No. of Edges)
3	5	$(15, 50)$
4	5	$(20, 75)$
5	5	$(25, 100)$
6	5	$(30, 125)$

- $p_{a13} = P(T_4 = i|T_3 = i)$, $p_{a23} = P(T_4 > i|T_3 = i)$, and $p_{a33} = P(T_4 < i|T_3 = i)$ for matrix with Visit 3 and 4 for Treatment A.
- $p_{b13} = P(T_4 = i|T_3 = i)$, $p_{b23} = P(T_4 > i|T_3 = i)$, and $p_{b33} = P(T_4 < i|T_3 = i)$ for matrix with Visit 3 and 4 for Treatment B.
- $p_{a14} = P(T_5 = i|T_4 = i)$, $p_{a24} = P(T_5 > i|T_4 = i)$, and $p_{a34} = P(T_5 < i|T_4 = i)$ for matrix with Visit 4 and 5 for Treatment A.
- $p_{b14} = P(T_5 = i|T_4 = i)$, $p_{b24} = P(T_5 > i|T_4 = i)$, and $p_{b34} = P(T_5 < i|T_4 = i)$ for matrix with Visit 4 and 5 for Treatment B.

Figure 2 represents an example of the conditional probabilities for one of the treatments for when there are 5 visits. In this multigraph graph depicted in Figure 2, the nodes T_1, T_2, T_3, T_4, and T_5 represent the classes or levels of the responses of the 5 visits for the patients in each treatment group. The three lines connecting each set of consecutive are the above conditional probabilities for consistency, improvement, and deterioration in the proportions transitioning from one visit to the other.

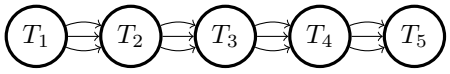


Figure 2: Multipartite Graph of Conditional Probabilities.

As it turns out that one creates multipartite graphs for each possible scenario such as Figure 1, then there are a simple formulae to compute the number of edges and number of nodes for the graphs. Similarly, one can also create multipartite graphs for each possible scenario such as Figure 2, and obtain a simple formulae to compute the number of edges and the number of nodes for the graphs. Below are the two tables that show these patterns:

From Table 4 one can deduce that the formulae for computing the number of nodes is the number of visits times the number of possible ordinal responses; and for computing the number of edges is the (number of visits -1) times the number of possible ordinal responses times 5.

From Table 5 one can deduce that the formulae for computing the number of nodes is the number of visits; and for computing the number of edges is the (number of visits -1) times 3.

Thus for 4 visits, we can write the transition proportional matrices for the

Table 4: Edges and Nodes of Geometrical Figure 2 for Different Scenarios

No. of Visits	No. of Ordinal Responses	(No. of Nodes, No. of Edges)
3	5	$(3, 6)$
4	5	$(4, 9)$
5	5	$(5, 12)$
6	5	$(6, 15)$

two treatments as follows: $M_A =$

$$\begin{pmatrix} p_{a11} & p_{a21} & p_{a31} \\ p_{a12} & p_{a22} & p_{a32} \\ p_{a13} & p_{a23} & p_{a33} \end{pmatrix}$$

and $M_B =$

$$\begin{pmatrix} p_{b11} & p_{b21} & p_{b31} \\ p_{b12} & p_{b22} & p_{b32} \\ p_{b13} & p_{b23} & p_{b33} \end{pmatrix}$$

Similarly for 5 visits, we can write the transition proportional matrices for the two treatments as follows: $M_A =$

$$\begin{pmatrix} p_{a11} & p_{a21} & p_{a31} \\ p_{a12} & p_{a22} & p_{a32} \\ p_{a13} & p_{a23} & p_{a33} \\ p_{a14} & p_{a24} & p_{a34} \end{pmatrix}$$

and $M_B =$

$$\begin{pmatrix} p_{b11} & p_{b21} & p_{b31} \\ p_{b12} & p_{b22} & p_{b32} \\ p_{b13} & p_{b23} & p_{b33} \\ p_{b14} & p_{b24} & p_{b34} \end{pmatrix}$$

Thus, if we obtain these geometrical representations for 4 visits, we can get a 3 x 3 transitional proportion matrix for each of the treatments; for 5 visits, we get a 4 x 3 transitional proportional matrix for each of the treatments; for 6 visits, we can get a 5 x 3 transitional proportional matrix for each of the treatments; and so on.

2.4 Method of Analyses As seen above, if we obtain these geometric representations for k visits, then we get a $(k-1)$ x 3 transitional proportional matrix. For square transition proportional matrices, we can employ Markov Chains to fit the probabilistic distributions.

Note that when there are lesser time points, such as $t = 1$ or $t = 2$, we are not as computationally challenged as we would be if there are more than 3 visits. For less than three points, the steps to obtain the significance are pretty straightforward as following: (a) obtain cross-tabulation for ordinal responses from visit 1 to visit 2 and/or visit 2 to visit 3; (b) calculate the conditional probabilities; (c) test significance between the two visits and/or two treatments using any standardized test as long as the assumptions are met.

Let $\boldsymbol{p} = (\boldsymbol{p_1}', \boldsymbol{p_2}', ..., \boldsymbol{p_s}')'$ denote the vector sr x 1 of observed proportions of all s subpopulations. The mean of $\boldsymbol{p}$ is $E(\boldsymbol{p}) = \boldsymbol{\pi}$, where $\boldsymbol{\pi} = (\boldsymbol{\pi_1}', \boldsymbol{\pi_2}', ..., \boldsymbol{\pi_s}')'$ and the covariance matrix of $\boldsymbol{p}$ is the matrix $\boldsymbol{V_{srxsr}}$.

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V_1} & \boldsymbol{0_{rxr}} & ... & \boldsymbol{0_{rxr}} \\ \boldsymbol{0_{rxr}} & \boldsymbol{V_2} & ... & \boldsymbol{0_{rxr}} \\ ... & ... & ... & ... \\ \boldsymbol{0_{rxr}} & \boldsymbol{0_{rxr}} & ... & \boldsymbol{V_s} \end{pmatrix}$$

Let $n++ = \sum n_{i+}$. If $n++$ goes to infinity and n_{i+}/n_{++} goes to p_i, then the asymptotic distribution of $\boldsymbol{p}$ is approximately $MVN(\boldsymbol{\pi_{sr}}, \boldsymbol{V})$.

- Let $\boldsymbol{F(\pi)} = (\boldsymbol{F_1(\pi)}', \boldsymbol{F_2(\pi)}', ..., \boldsymbol{F_u(\pi)}')$ be a vector of u linearly independent response functions of interest. Each of the functions is required to have continuous partial derivatives through order 2.
- Let $\boldsymbol{F(p)} = (\boldsymbol{F_1(p)}', \boldsymbol{F_2(p)}', ..., \boldsymbol{F_u(p)}')$ denote the corresponding sample response functions, and let $\boldsymbol{Q} = \partial F/\partial \pi$ be the u x sr matrix of partial derivatives evaluated at the sample proportions $\boldsymbol{p}$. Then, $\boldsymbol{F(p)}$ is approximately distributed as normal distribution with mean $\boldsymbol{F(\pi)}$ and variance $\boldsymbol{V_F} = \boldsymbol{Q}\hat{V}\boldsymbol{Q}'$.
- In particular, $\boldsymbol{F(\pi)} = \boldsymbol{A\pi}$, where A is a matrix of known constants, is appropriate when the response functions are linear functions of the underlying probabilities.

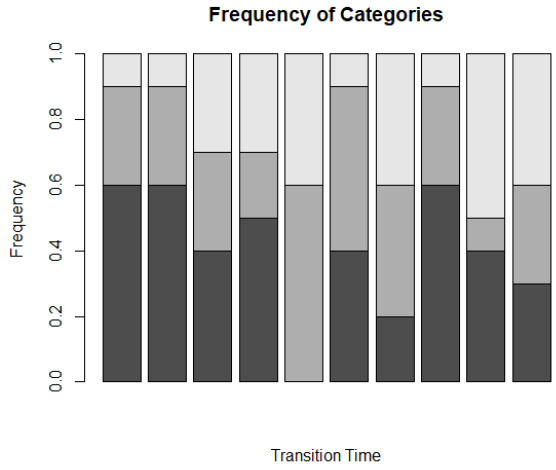
3. Results and Discussion

3.1 Results

For illustration purposes, suppose that there is a randomized, double-blinded, controlled study comparing medical product A and medical product B in newly enrolled patients with benign prostate hyperplasia. This is a condition in which the flow of urine is blocked due to an enlarged prostate. Let's assume that approximately 100 subjects were randomized 1 : 1 to either of the medical products. PRO data was collected for these subjects at eleven time points. Evaluation of the effectiveness of the two medical products are performed using the patient reported outcomes such as pain variables that are rated from a scale of 1 to 5. Assume that the simulated patient reported outcome data was collected from each subject every 4 week for 11 weeks. As you can see that the observed frequencies or proportions for medical product A in Figure 3 are not consistent across different transition time periods, so a clinical researcher may conclude that pain levels may not remain the same for a certain treatment over time for medical product A, but for B the pain levels remain the same over a period of 11 weeks.

3.2 Discussion The proposed methodology of compartmentalization method is a more useful and systematic approach to analyze the consistency, improvement or deterioration of the consistency of the patient's well being over time. It provides more interpretable results when working with categorical outputs from patient reported outcomes and obtain three proportional measures compared to commonly used interval scale results (converts ordinal scale to a continuous). Therefore, there is much less loss of information since the proposed method assumes three different classifications from the matrix structure, whereas the

Figure 3: Observed Proportion Plot for Pain variable Product A



mean change from baseline results in one statistic. Moreover, with this method, the results are more clinically as well as statistically meaningful.

4. Conclusions

There is a dire need to understand the structure of the data especially since large and big data are generated thanks to every changing technology. Understanding the data can not help in providing the appropriate method to analyze, but also reduce time and effort put into the data architecture.

The cross-tabulated matrix of the data provides the discrete choice pairs with the transition probability distribution, where as the graphs provides the dimension reduction logic on the transition probability distribution in our proposed method. For the two distinct choice pairs, a path or edge is determined between them. These choices represent the choice made by the patient in determining the best or worst or consistent transition probabilities when dimension reduction is used. Given the changing nature of responses, this proposal includes a time-dependent approach to a response or choice pairs.

There are several characterizations that can be made from the direction this research can take. For example, the random walk option is another matrix option.

Acknowledgment

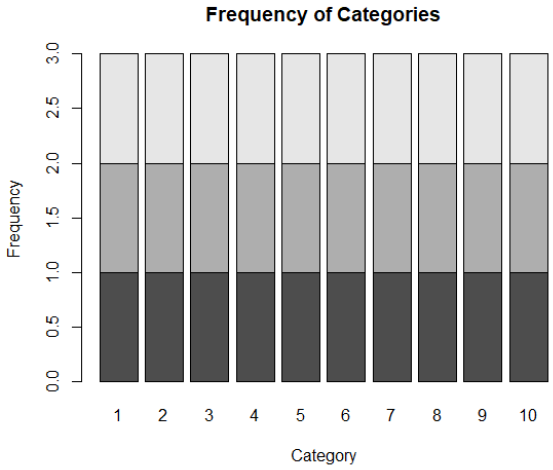
Dr. Manasi Sheth would like to thank Thomas Drucker and Ruth Whitmore for motivating her in the writing of this manuscript. Dr. Sheth would also like to thank Dr. David Bellhouse for his insights into this manuscript.

References

1. S. Kucukemiroglu and M. Sheth, Nonparametric Analysis of Patient Reported Outcomes using Compartmentalization Method, Joint Statistical Meeting Proceedings, 2021, p. 574-581.

2. S. Kucukemiroglu and M. Sheth, Compartmentalization of Discrete Repeated Measures in Patient-Reported Outcomes Questionnaires, In JSM Proceedings, Biometrics Section, pp.2593-2599.

Figure 4: Observed Proportion Plot for Pain variable Product B



3. M.A. Detry and Y. Ma, Analyzing Repeated Measures using Mixed Models, JAMA (2016), 315(4), pp. 407-408.

4. A. Prakash, R.C. Risser, and C.H. Mallinckrodt, The impact of the analytical method on the interpretation of results in longitudinal clinical trials, International Journal of Clinical Practice (2008), 62(8), pp. 1147-1158.

5. R. Diestel, Graph theory, Springer Texts (2020), 6th Edition