Stockholm University

Department of Mathematics
Master program in Biostatistics

Self-reported Gastrointestinal Symptom Clusters and Factor Models in Survivors of Gynecological Cancer

A thesis submitted in fulfillment of the degree requirements
for the degree of M.Sc. in Biostatistics

AUTHOR:

Mohammad Masud Pervez

SUPERVISORS:

Martin Sköld
Associate professor
Department of Mathematics
Stockholm University

Tommy Nyberg
Statisticians
Department of Oncology-Pathology
Karolinska Institute

December 2013

Stockholm University

Department of Mathematical Statistics
Master program in Biostatistics

Self-reported Gastrointestinal Symptom Clusters and Factor Models in Survivors of Gynecological Cancer

A thesis submitted in fulfillment of the degree requirements
for the degree of M.Sc. in Biostatistics

AUTHOR:

(Mohammad Masud Pervez)

SUPERVISORS:

(Martin Sköld)

(Tommy Nyberg)

December 2013

This thesis is dedicated to my Family.
For their boundless love, support and encouragement.

Acknowledgments

First of all, I am grateful to almighty Allah who blessed me to complete my thesis eventually from various obstacles in my life. This thesis paper is a requirement for the completion of my Master degree at Stockholm University and therefore place a huge achievement in my life. It has been kept extreme cared through to end with the support and encouragement of many people including my friends, well wishers, colleagues and various institutions. My sincere gratitude to all of those people who helped me to make this thesis possible.

My deepest appreciation to my both supervisors, Martin sköld and Tommy Nyberg, for their patient guidance and enormous support in this study. Their valuable comments and suggestions on the manuscripts were very beneficial for the development of the report and keeping my progress on schedule. It was a great pleasure to work together with some expertise like them in where my research knowledge was also broaden. Without their proper guidance and assistance this master thesis would not have been possible. I present my highest gratitude from my heart to them.

I would like to express my special thank to professor Gunnar Steineck of Clinical Cancer Epidemiology at Karolinska Institute (KI), who let me to do my research study and this thesis within his department. His inspiration and valuable support helped me a lot to continue my work gently. He set me there as one of his research group member and introduced myself very warmly to other members at the department. He is a very good-hearted person and always kept his kind wishes to me. His contribution in my life is unforgettable and should express my heartfelt gratefulness to him.

I would also like to extend my sincere appreciation to adjunct professor Elisabeth Åvall Lundqvist, specialist on gynecologic oncology at Karolinska Institute, for her valuable and constructive suggestions from her clinical aspect during the formulation and development of the hypothetical clinical classification for this thesis. Her willingness to give her time so generously has been very much appreciated.

I wish to thank to all other members in the Department of Clinical Cancer Epidemiology at KI, for their assistance and encouragement during my staying there. My gratitude also to all the academic staffs in my own department, Department of Mathematics at Stockholm University, for their warm wishes and compassionate support to my study.

My grateful thanks are also extended to my two very good friends; Mohammed Ferdous-Ur Rahman and Saad Ahmed Salman, for their enthusiastic encouragement and useful critiques of this research work.

I also express my sense of gratitude to one and all who, directly or indirectly, have lent their helping hand for successfully completion of this study.

I wish to thank the **R** statistical software, which is a free and open source system and available to install on the Internet, used for core data analysis of this research study. The R package **Clues**, developed by Chang et al. (2010), was used for cluster comparison and determining optimal number of clusters. Therefore, the credit should give all the personnel for their sincere contribution and valuable technical support relating to this software.

Finally, I have to pay my high regards to my parents for their love and prayers throughout my life to fulfill my dreams. My brothers and other family members who supported me always with their kind wishes, deserve my wholehearted thanks as well.

Masud Pervez

Abstract

The main aim of this study is to verify empirical techniques for clustering symptoms experienced in gynecological cancer patients, by comparing the results with a clustering based on clinical experience. Data are taken from a survey of symptoms experienced by 516 cancer patients who received radiation therapy at Karolinska University Hospital in Stockholm or at Jubileumskliniken at Sahlgrenska University Hospital in Gothenburg between $1991-2003$. The study finds that Wards clustering algorithm based on the phi correlation matrix gives the closest agreement out of several empirical clustering combinations, as measured by adjusted Rand index. An alternative approach based on factor analysis is also considered.

Key Words

Symptom clusters; cancer survivors; clinical classification; adjusted Rand index; exploratory factor analysis.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Introduction

Gynecological cancer is one of the most common types of cancers among women in Sweden. Five major gynecologic cancers are: ovarian cancer,cervical cancer, uterine cancer, vulvar cancer and endometrial cancer, and affects female's life, reproductive system, sexuality and intimacy. According to Swedish National Board of Health and Welfare (2008), in Sweden, about 2700 new gynecological cancer cases were diagnosed annually [2]. The risk of gynecologic cancer increases with age and the majority of the patients are diagnosed at their age of 60 years or older [3]. Early detection and diagnosis of the disease and increased treatment effectiveness may improve the mortality rate and lead to the cancer survivors to live for several years after treatment. So, an in depth knowledge about the treatment and treatment related symptoms (knowledge about the disturbed physiological functions, i.e., pathophysiologies) are necessary.

Radiation therapy is one of the widely used method for treating gynecological diseases and often treat in combination with other cancer treatments. Pelvic radiotherapy can make major injury on the anorectal region of the survivors. So, cancer patients may have experience a wide variety of treatment related adverse side-effects which usually refer as *symptoms*, that can arise long after their treatment and make negative impact on their usual life style. In cancer research, the assessment of symptoms can be determined by using various instruments/tools. Dodd et al [4] defines *symptom* as a subjective experience of a patient that reflects changes in the individuals' biopsychosocial functioning, sensations, or cognition. In this study, *symptom* is referred by the cancer survivors' self-reported responses to a questionnaire. Earlier study shows that the most common gastrointestinal symptoms for radiotherapy are: diarrhea(loose

stools), flatulence, defecation urgency, abdominal pain and bloating, fecal leakage [3].

Effective symptom management strategies can play an important role for the improvement of the cancer patients' quality of life. Symptom management research is a versatile field which mainly focused on evaluating multiple symptoms and to give care for the improvement of patients quality of life from their serious or life-threating diseases. The objective of symptom management is to prevent or treat the symptoms of the disease as early as possible, their side effects caused by treatment of the disease, and psychological, social, and spiritual problems related to the disease or its treatment [5]. An important area related to the aspect of symptom management research is identifying the nature of clinically significant clusters of symptoms and their associated prevalence rate [6]. In the area of intervention studies symptom cluster plays a vital role. Kim et al [7] defined *symptom clusters* based on a concept analysis as a group of two or more symptoms that are related to each other and are relatively independent of the other symptom clusters. In the clustering symptoms, there exist high relationships among symptoms within a cluster than the relationships among symptoms across different clusters. The etiology of the clustering symptoms in a group may or may not be the same [8]. Several theoretical frameworks are available in literature to symptom cluster, but the method which can provide optimal clusters empirically is still not clear [9]. Some of those approaches group symptoms by considering the occurrence or experiences of symptoms by the patients (refer as *symptom* cluster), and some approaches group patients or individuals by considering the probability of getting a symptom (refer as *patient* cluster). The study focused to cluster gynecologic cancer treatment related symptoms by using two widely used multivariate techniques *cluster analysis* and *factor analysis*, which are also very important tools in oncology research for clustering symptoms. In this study we consider the method of *cluster analysis* to identify groups of symptoms by analyzing patients' experience data and alternatively the method of *factor analysis* was used to discover underlying factors of the study symptoms.

Clustering is essentially about exploring natural groups in data that are meaningful, useful or both. In oncology, this technique is important for understanding the experience of treatment related effects occurring in various chronic illness. Clustering methods are used extensively in situations where there exits no pre-specified and well-defined groups in data. The attributes of the data are then used with the assistance of clustering techniques to assign elements into artificial groups. So, this statistical technique is very appealing for medical research. Ideally, symptoms are not independent entities and they are associated with each others in some way. Some symptoms are likely to stay together in a cluster and some shows opposite tendency. The meaning of relation can be expressed in various manners. It may be observed that the relation of symptoms to each other can be related to some biological mechanisms of the disease such as high inflammation. Another way to look at the relations is through the weights in which they are reported by the patients. For example, the symptoms reported with high frequency by the individuals are clustered together than the symptoms reported with lower frequency.

Ideally, many clustering methods have been developed based on the aim and goal of particular research question, and each having certain advantages and disadvantages. One of the most widely used technique for clustering is agglomerative hierarchical cluster analysis, which is a set of nested clusters that creates a tree of the data by progressively adding similar groups of elements. The advantages of the technique are that it considers a measure of similarity (or dissimilarity) among the elements and a clustering algorithm. For these advantages and well-known characteristics, the study executed this clustering technique for determining clusters of symptoms that are clinically similar to each other.

Alternatively, another multivariate data-driven method factor analysis is considered for cluster symptoms. This technique determines factors or clusters that are related to multiple symptoms and the symptoms may appear meaningfully to more than one symptom factor [10]. In factor analysis, *Exploratory* factor analysis is the simplest one and probably the widely used technique that can provide informative results of the data [11]. The results of factor solution are not unique and therefore the interpretation can varies. However, the optimal decision depends on the choices of following attributes: sample size, the number of factors, and the method of factor rotation techniques [12].

This master thesis has used data collected from a long-term gynecological cancer study, on survivors after radiation therapy in Stockholm and Gothenburg, Sweden. The data was collected by the Division of Clinical Cancer Epidemiology at the Department of Oncology and Pathology of Karolinska Institute, Stockholm, Sweden and the Division of Clinical Cancer Epidemiology at Sahlgrenska Academy of Gothenburg University,Gothenburg. The aim of the main study was to assess the quality of life of the cancer survivors and to identify which symptoms they may suffer as a treatment related side effects from their cancer treatments for radiotherapy and or combinations with other therapy. Data was collected by a survey questionnaire sent to gynecological cancer survivors in 2006 that had been treated with pelvic radiotherapy between 1991-2003. In this study we basically focused our interest to the scientific questions: how many classes of the symptoms are available in the questionnaire for the gastrointestinal symptoms from a clinical point of view and what are the underlying constraints for these classifications for measuring the symptoms? Hierarchical cluster and factor analysis are essentially about discovering answers to such questions.

1.2 Background of the Study

In literature for classifying cancer symptoms, different approaches were applied, but not found any intensive work on the area of gynecological cancer using hierarchical clustering and factor analysis techniques. Some study also proposed symptom clusters based on previous empirical research study. One of the main objective of this study is to determine clusters among the gastrointestinal symptoms by using hierarchical clustering in where we consider a distance metric that can yield a closest clusterings in compare to an external clusterings. As an alternative approach for identifying factor models of the symptoms we also consider a factor

analysis technique. Some background studies of these two techniques used extensively on several areas are discussed below:

The background of the primary study of this thesis is available on Dunberger et al. [3] Ph.D. research study. They investigated long-lasting gastrointestinal symptoms and assess the quality of life of the cancer survivors after pelvic radiotherapy.

Bender et al. (2005) [13] used hierarchical cluster analysis on the symptoms experienced by breast cancer patients and identified three symptoms clusters corresponding to three different phases. Their suggested clusters are fatigue, perceived cognitive impairment, and mood symptoms cluster.

In [14], Wilmoth et al. performed a clustering technique on breast cancer related symptoms after chemotherapy and proposed that fatigue, weight gain, and altered sexuality are three symptom clusters for a patient. They also suggested that every cluster has a significant impact on the cancer survivors quality of life and considering them as a group can magnifies their affect.

In [15], Eisen et al. (1998) suggested a clustering algorithm that can use to study genome-wide expression from a hybrid DNA microarray data. The designed of the study was based on microarrays for developing yeast genome and a human fibroblast cell line. The authors proposed an hierarchical clustering algorithm based on the average criterion and used the correlation coefficient as a dissimilarity measure suggested by Sokal and Michener [16]. In this study, they used the dissimilarity matrix to cluster the genes and not on the samples as the samples were measured on an ordered list for which the clustering was useless.

In 2010, Wentzensen et al. [17] suggested a hierarchical clustering technique on human papilloma viruses (responsible for anogenital cancers) genotype. Their proposed hierarchical clustering system used complete linkage algorithm and Euclidean distance metric as a dissimilarity measure. They clustered both their referred disease combinations and HPV genotypes simultaneously and created dendrograms to visualized the clustering results. They ended up with the conclusion of four major disease clusters and three major groups of HPV genotypes.

Guillaud et al. [18] studied in the evaluation of optimal technologies for the screening and detecting on cervical neoplasia an early emerging of cervical cancer. They performed numerical histo-pathological analysis of biopsies from their 1800 patients. On that study, the authors performed linear discriminant analysis to assess the diagnostic information in three different sets of features on a cell-by-cell and sample-by-sample basis. Their selected feature values and summary scores were used to evaluate intra- and inter-observer variability.

A specifically designed clustering algorithm using gene expression data for breast cancer can be found in [19]. In their study, the authors performed an unsupervised two way cluster analysis independently: clustering of gene and clustering of tumor, using an hierarchical agglomerative clustering technique. For the clustering of gene, the pairwise similarity/ dissimilarity were measured on the basis of tumor expression ratio measurements to all tumors,

and for clustering of tumor, pairwise similarity/ dissimilarity were measured based on expression ratio to all significant genes.

Perou et al. [20] followed the similar method as Eisen et al. [15] for clustering genes using their expression profiles of normal breast tissue, breast cancer bulk tissue, and breast cancer cell lines. Using the known or suspected sequences of genes involved in cancer, the authors presented their own microarrays data for clustering. Some of their suggested gene clusters are co-regulated.

1.3 Objective

The study considers long-lasting gastrointestinal gynecological cancer survivors self-reported symptoms and their classes. The specific objectives for this thesis work are:

- determining whether a group of questions measure the same symptom by the responses or they measure separately.
- exploratively identifying the clusters that are most similar to hypothesized clinical clusterings by comparing various empirical clustering results.
- finding the differences between the most similar clustering and the hypothetical clustering by observing in which variables are to be assigned to which classes.
- identifying the method of estimating pairwise dissimilarity for such a study.
- implementing exploratory factor analysis for explaining the correlations among the gastrointestinal symptoms by considering a smaller number of unobserved latent constructs or factors.
- labeling the latent constructs or factors.

CHAPTER 2

Method

2.1 Study Sample

This paper has analyzed data collected for a long-term gynecological cancer survivors quality-of-life study after radiation therapy in Stockholm and Gothenburg, Sweden. The primary study of this thesis was carried out to the gynecological cancer survivors who were received pelvic radiation therapy (RT) only or as part of other combination therapy: Operation (Op), Brachytherapy (Br) and Chemotherapy (Ch), which gives a total of $2^3 = 8$ treatment combinations, during their treatment period [21]. The cancer survivors were treated in their pelvic region $2 - 10$ years earlier. The selected survivor cohort were then asked to fill in a study-specific questionnaire consisting of 351 questions pertaining symptoms from various regions of human functions, such as the gastrointestinal region, urinary bladder, genitals, pelvic bones, abdomen and legs, as well as symptoms on psychological behavior and their quality-of-life and social functioning [3].

The questionnaire was sent to gynecological cancer survivors in 2006, that had been treated with pelvic radiation therapy between $1991 - 2003$ at Karolinska University Hospital in Stockholm or at Jubileumskliniken at Sahlgrenska University Hospital in Gothenburg. The questionnaire was developed following interviews with 26 cancer survivors where the women were asked to describe their symptoms in their own words, and the questions of the questionnaire were formulated to follow the wording used by the interview participants. Out of 1800 identified cancer patients, 789 met the eligibility criteria of being alive and free from tumor recurrence in 2006, being born 1927 or later (i.e. less than 80 years of age), and being able to read and understand Swedish. Of the 789 invited, 616(78%) agreed to participate and returned a completed questionnaire. The regional ethics committees in Stockholm and

Gothenburg approved of the project. More details of the questionnaire development and data collection has been published on Gail et al [3].

2.2 Study Data

In this study, we excluded 11 survivors from our primary data that had a stoma since individuals with stomata are unable to have fecal leakage symptoms. Also the number of missing data per question among the patients lies between 0.5% and 3% for the study 37 gastrointestinal symptom questions of the questionnaire, makes a total of 89 participants (15% of 605) that have at least one missing value. The study analyzes complete cases, omitting the missing observations and that yields a number of 516 individuals. The reason behind is, the amount of missing value is low and also in our assumption the alternative of imputing data may risk to create clusters that are not there to start with.

Details about demographic and treatment characteristics for the cancer survivors are shown in Table A.1 on page 68.

2.3 Gastrointestinal Self-reported Symptoms

From the large questionnaire we have focused our this study only on the "Bowel section", which involves questions about bowel functions as a side effect after giving the radiotherapy and or surgery treatment. The dataset contains a sample of 516 completed survivor information with an ID number for each patient, and a total of 37 questions that all have various response alternatives. The responses are typically stored as integers on different scales, all of which are at least ordinal (for example, 0="Not applicable" if a question has such an alternative, otherwise in the order that they're stated in the questionnaire, i.e. 1="No, never", 2="Yes, occasionally", $\cdots$, 6="Yes, at least once a day"). In the study, the term "*symptom*" represents cancer survivors' self-reported responses of the questionnaire.

The questions used from the gastrointestinal section of the questionnaire for performing hierarchical clustering are presented in Table A.2 on page 69.

The response scales for all of the questions are described shortly below:

N38, N44, N52, N53, N60, N62, N65, N67, N70, N73, N77, N81, N82, N84, N85, N88, N91, N92, N93, N94, N95, N96, N97, N98, N114, N137 : Have you had [$\cdots$ this symptom $\cdots$], in the last six months.

1= No,
2= Yes, Occasionally
3= Yes, at least once a month
4= Yes, at least once a week

5= Yes, at least three times a week
6= Yes, at least once a day

N48, N74, N139, N140, N141: Have you had [$\cdots$ this symptom $\cdots$], in the last six months.
0= Not Applicable, Haven't had [$\cdots$ this symptom $\cdots$]
1= No,
2= Yes, less than half of occasions
3= Yes, more than half of occasions
4= Yes, at every occasion

N49: Have you had ability to push out feces, in the last six months. 0= Not Applicable, I haven't had any need
1= No ability
2= Small ability
3= Moderate ability
4= Great ability

N75: How long have you been able to keep the stools at the urgency, the last six months
0= Not Applicable, I havn't had stool efforts
1= Less than 1 minute,
2= Between 1 and 5 minutes
3= Between 5 and 10 minutes
4= Between 10 and 30 minutes
5= 30 minutes or more

N76, N99, N102: Have you had [$\cdots$ this symptom $\cdots$], in the last six months

0= Not Applicable
1= No,
2= Yes, Occasionally
3= Yes, at least once a month
4= Yes, at least once a week
5= Yes, at least three times a week
6= Yes, at least once a day

N138: how strong is the pain in the abdomen was when it was at its worst, the last six months (Visual Digital Scale/ Numeric Rating Scale)
1= No pain

2=· · · · · ·
3=· · · · · ·
4=· · · · · ·
5=· · · · · ·
6=· · · · · ·
7= Worst imaginable pain

It is obvious that the questionnaire has different formatting of variables, so there are some additional things we have been considering for the ease of our analysis:

- All "Not applicable"'s are handled as a "No" (i.e., 0→1).
- **N49** (ability to push feces): inverse scale, lower values mean more severe symptom (opposite to most other questions). 0="not applicable" should probably not be considered a 1="no" for this question. We have changed "0" to the maximum response 4, and that the scale is inverted (so that 1→ 4, $\cdots$, 4→1).
- **N75** (ability to hold feces): inverse scale, lower values mean more severe symptom (opposite to most other questions). 0="not applicable" should probably not be considered a 1="no" for this question. We have changed "0" to the maximum response 5, and that the scale is inverted (so that 1→ 5, $\cdots$, 5→1).

Figure B.1 on page 76 presents the marginal distributions of our 37 categorical gastrointestinal variables which illustrates the proportion of counts for each category next to each other for ease comparison. The height of each bar is the respondents proportionate response to a particular category of a variable. We can also observe that the marginal distributions of the variables are heterogeneous, most likely reflecting that questions asked concerns both rare and common symptoms.

2.4 Hypothetical Cluster

In clustering, it is often motivating to compare the empirical clustering results with some external criteria. Based on this, we have tried to formulate an external *clinical symptoms classification* for the given 37 gastrointestinal questions of the questionnaire with the help of a clinical expertise in where the questions that they would believe are related to each other without looking at the data. Elisabeth Åvall Lundqvist, specialist on gynecologic oncology at Karolinska Institute and who is also responsible for the development of the questionnaire of the primary study, formulate a hypothetical clinical classification [Figure 2.4] from her clinical point of view for these questions based on long term side effects of cancer survivors. The objective of clinical classification is that to compare it with the empirical hierarchical clustering results and to evaluate a clustering result of the data out of our many combinations of hierarchical clusterings which is appropriate from mathematically and clinically.

Figure 2.4 on the next page shows our proposed hypothetical clinical clustering of the cancer survivor questions on the focus of 37 gastrointestinal symptoms. At the beginning of our grouping we considered that each original question of the questionnaire measure a particular symptom and construct a single cluster. We grouped them then based on their clinical association and formed sixteen large and small clusters as shown in the picture. Later on, we realized that this forming still bearing very large clusterings as compared to the number of questions, so we collapsed those sixteen clusters more compactly in where symptoms we thought to be most associated to each other. In this way we have found eight smaller number of clusters which are marked clearly on the picture also. It is assumed, the symptoms in the same cluster are strongly related to each other than the symptoms of other clusters. As the formulated grouping is hypothetical, so the distances from one group to another is arbitrary. As an example, *Anal pain* and *Hard stools* are two separated groups and they are staying far away in the figure, but this doesn't mean that these two groups have large distance compare to other groups. Intuitively they are not so closely correlated as like they are with their nearest other groups. In the middle of the figure (all "green" and "blue" colors) actually represents a very big group about the abdominal questions to which we can refer as *abdominal pain and flatulence group* and the small groups within this big group are showing very close association, i.e.,the small groups are correlated and that some of the groups are combinations of two or more. Loose stools is a very well known side effects after radiotherapy and it's very common among the survivors [3]. In the figure *Loose stools* is on a separate group and have strong association with *Loose stools leakage, Solid stools leakage, defecation urgency, urgency with leakage* and *return to bathroom within an hour* groups, i.e., large association with the leakage questions groups. *Anal pain* group is separated from the *Leakage of blood or mucus* groups, which is in turn separated from all the other abdominal groups of questions, solid stools leakage and hard stools.

From the above hypothetical clinical consideration, all the gastrointestinal questions were grouped separately later on into eight clusters for comparison. Table 2.1 represents these hypothetical eight groups or clusters.

Table 2.1: Hypothetical eight clusters

Cluster No.	**Cluster Name**	**Cluster Size**	**Variable Names**
1	Leakage of blood and mucus	8	N62, N65, N91, N92, N95, N96, N99, N102
2	Abdominal pain and flatulence	9	N60, N84, N85, N88, N137, N138, N139, N140, N141
3	Defecation Urgency	5	N53, N73, N74, N75, N77
4	Leakage of loose stools and urgency	6	N76, N81, N82, N93, N97, N114
5	Anal pain	2	N67, N70
6	Loose stools	1	N38
7	Leakage of solid stools	2	N94, N98
8	Hard stools	4	N44, N48, N49, N52
Total		**37**	

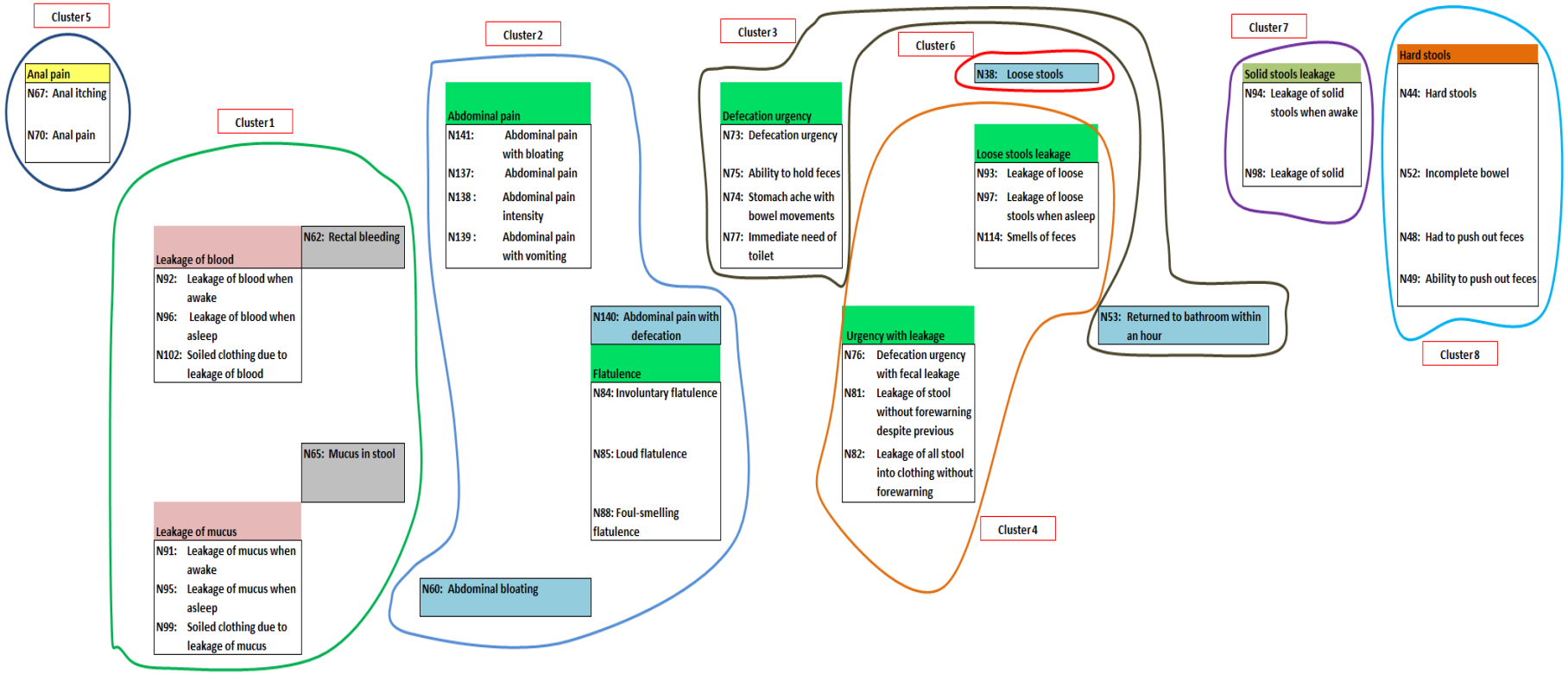

Figure 2.1: Hypothetical clinical clustering

2.5 Basic steps and Numerical methods for hierarchical clustering analysis

Often clustering is used for clustering individuals, but in this study we consider methods for clustering variables rather than individuals. Clustering individuals usually attempt to identify relatively homogeneous groups of cases (individuals) where as clustering variables identifies a set of non-overlapping homogeneous grouping variables. Variable clustering can be used for estimating collinearity, redundancy, and for separating variables into similar clusters that can be interpret as a single variable and this reflects to data reduction.

In this study, for grouping variables, the basic steps involve in a hierarchical cluster analysis are:

- Select a measure of similarity or dissimilarity between variables.
- Choosing a appropriate clustering algorithm.
- Deciding the number of clusters. and
- Validate and interpret the cluster solution.

Suppose, we have an $n \times p$ multivariate data matrix, $\mathbf{X}$, containing the individuals responses at the row and at column describing each variable to be clustered; that is,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

The generic entry x_{ij} in the $\mathbf{X}$ matrix gives the value of the j^{th} variable on individual i. Our main interest center on clustering the variables which define the columns of the above data matrix $\mathbf{X}$.

Generally, the characteristics of the variables in matrix $\mathbf{X}$ can be a mixture of continuous, ordinal and/or categorical, and often some entries can be missing, while we have a data set where variables are in ordinal and with missing observations. cluster analysis techniques usually begin by converting this matrix $\mathbf{X}$ into an $p \times p$ matrix of inter-object *similarities*, *dissimilarities* or *distances* which we generally call *proximity* measures. Using the proximity matrices and with selected hierarchical clustering algorithms, a clustering process can be obtained which generates similar clusters/groups of variables from the data set and where the number of groups g is smaller than or equal to p (i.e., $g \leq p$). Thus cluster analysis summarize data redundancy by reducing the information on the whole set.

2.6 Data Dichotomization and Standardization

Usually, the raw data or the original measurements are not directly used for cluster analysis. Therefore, arrangement of data is necessary and important for cluster analysis. Preparing data involves some way of transformation, such as dichotomization, standardization or normalization.

2.6.1 Data Dichotomization

Since our data have varying response scales and not identically distributed, one way to simplify the data is to dichotomize at the median to get similar distributions for the variables. We split the variable at the median of the data and not of the items on the scale, as that is conceptually the middle response of the data. For example, in our study question N38, where this loose stools question ranges from 1 to 6 and the median of this question is 3. We transformed the variable to a new dichotomous variable, *zero* and *one*, by categorizing each subject of the variable as either have low loose stools scores (1 through 3(i.e.,median value)) or have high loose stools scores (4 through 6) respectively. Mathematically we can write it in the following way.

Let, $x_1, x_2, ..., x_p$ be the variables in the dataset where $j = 1, 2, ..., p$. Also let, M_j be the median response of the variable j. We dichotomize the i^{th} observations of the j^{th} variable x_{ij} based on the following criterion:

$$x^*_{ij} = \begin{cases} 0 & ; \text{ if } x_{ij} \leq M_j \\ 1 & ; \text{ if } x_{ij} > M_j \end{cases} \quad (i = 1, ..., n \quad and \quad j = 1, ..., p) \tag{2.1}$$

where, x^*_{ij} entities will formulate a new binary dataset with approximately the same marginal distributions of the variables, but slightly differs from the formulations of many methods for dichotomization in which "0" and "1" usually signifies absence or presence of symptom for a particular question.

As for many ties observations at the median values of the original variables in our data, the dichotomization can not be done uniformly. We assumed that the losing of information of the data with this dichotomizing is minor and could use this binary dataset for our further cluster analysis.

Figure B.2 on page 77 shows the marginal distributions for our new dichotomize variables. The figure reveals that though the dichotomization has done based on the median, the shape of the marginals (binary categories) are not evenly distributed for many of the variables. This is because there are many observations lie on the median value for that particular variable and that makes the data skewed instead of symmetric even after the dichotomization.

2.6.2 Data Standardization

Generally, data standardization concentrates on variables and it makes data dimensionless. When there are unequal scales of the variables in a dataset then it is meaningful to convert the data to some standard indices. By standardization we may loose all original information about the location and scale of the data, but it is useful to standardize variables in cases when using Euclidean distance as for dissimilarity measure and also where the data is sensitive to the differences in the original scales of the variables [22]. There are various ways to standardize variables and the appropriate method depends on the dataset and the particular field of study. In this study we recalculate each variable by using the following equation:

$$x^*_{ij} = \frac{x_{ij}}{\max_i(x_{ij})} \quad (i = 1, ..., n \quad and \quad j = 1, ..., p) \tag{2.2}$$

where x^*_{ij} denotes the recalculated value, and $\max_i(x_{ij})$ is the maximum observed value of the jth variable.

After standardization of data, all variables transformed to a unique closed scale [0,1] with the same ordering and consistency as the source data had. That is, the minimum value means low severity of a symptom and maximum value means high severity of the symptom for a question and can be comparable with other variables. This method of standardization allows variables to have different means and standard deviations but equal ranges.

2.6.3 Normalization of Rank Transform Data

Another way to handle ordinal variables with different distributions and value ranges is to transform them into quantitative variables taking values in the interval [0,1] using their ranks. This will make the distributions more similar, but relies on the strong assumption that the minimum and maximum values of one variable are comparable to the minimum and maximum values of another variable. After that the usual distance methods (such as, Euclidean distance , correlation coefficient, etc.) for quantitative variables can be used to calculate distance by treating the ordinal variables as quantitative variables from the rank normalize data.

The transformation can be done on the following ways:

- Rank the values of the original variable from t = 1 to n.
- For ties observations, used average of ranks.
- The rank entities are normalized into standardized value of zero to one [0,1] by [23]:

$$x^*_{ij} = \frac{t-1}{T-1} \tag{2.3}$$

where x^*_{ij} denotes the normalized value, t is the rank of the observed value for a variable and T is the maximum rank of the variable after averaging ranks for ties observations.

2.7 Notation

To describe the dissimilarity measure between variables and clusters we used following some notations in our subsequent sections:

Let,

n	= number of cases/observations.
p	= number of questions/variables.
$\mathbf{x}$	$= (x_1, x_2, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$,
	be two vectotrs of variables which denote two survey questions.
x_{ij}	= value of the ith observation and jth variable in the data matrix,
	where, $i = 1, \cdots, n$ and $j = 1, \cdots, p$.
x^*_{ij}	= value for the ith observation and jth variable of the
	transformed data. where, $i = 1, \cdots, n$ and $j = 1, \cdots, p$.
w_i	= weight for the jth objects, where $j = 1, \cdots, n$.
$d(\mathbf{x}, \mathbf{y})$	= the distance between variables $\mathbf{x}$ and $\mathbf{y}$.
$s(\mathbf{x}, \mathbf{y})$	= the similarity between variables $\mathbf{x}$ and $\mathbf{y}$.
n_c	= number of concordant pairs between variables $\mathbf{x}$ and $\mathbf{y}$.
n_d	= number of discordant pairs between variables $\mathbf{x}$ and $\mathbf{y}$.
G_k	$= \{y_1, y_2, \cdots, y_r\}$denotes cluster k of size r.
r_k	= number of elements in cluster G_k.

2.8 Dissimilarity measures between variables

In hierarchical clustering, it is required to define a measure of dissimilarity between sets of variables in order to decide how clusters or variables will be merged together by agglomerative method, or how a large cluster will be splited out into smaller clusters by divisive method. A wide variety of distance and similarity measures are proposed based on the nature and characteristics of the data. Moreover, there are no general theoretical guidelines for selecting a measure for any given application. Different methods in defining the distance between two data points can lead to different clustering results. Ideally, field knowledge should be used to guide the formulation of a suitable distance measure. However, for the questions in this study, sufficient field knowledge about this study-specific questionnaire is lacking, and we take a more explorative approach where we can consider a number of different distance measure. In this study we consider the following common distance methods.

2.8.1 Distances for numerical data

In clustering, Euclidean distance and Pearson correlation distances are probably the most commonly used measures for numerical data [24].

2.8.1.1 Euclidean Distance

This is the fundamental distance measure between two points and are also known as the "straight line distance". It is defined as the positive square root of the sum of squares of the differences between corresponding points of two data sets. Given two n-dimensional data vectors $\mathbf{x}$ and $\mathbf{y}$, the standard Euclidean distance is defined by

$$d_{euc}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T} \tag{2.4}$$

where x_i and y_i are the value of the ith observations of $\mathbf{x}$ and $\mathbf{y}$, respectively.

This distance is sensitive to the differences of the measure variables with their magnitude or scale [24]. Therefore it is often necessary to standardize the variables beforehand to compute this distance.

The squared Euclidean distance, which is the square of the standard Euclidean distance, often use in situations when we would like to put higher weights to the objects which have greater distances from each other. So, in the cases when distances only have to be compared then we usually use the squared Euclidean distance instead of the standard Euclidean distance. The general formula is,

$$d_{sqeuc}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n}(x_i - y_i)^2 = (\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T \tag{2.5}$$

2.8.1.2 Pearson's correlation distance

Pearson Correlation measures the similarity with direction (shape) between two variables. Given two n-dimensional data vectors $\mathbf{x}$ and $\mathbf{y}$, the Pearson correlation distance is defined by

$$d_{pearson}(\mathbf{x}, \mathbf{y}) = 1 - r \tag{2.6}$$

where r denotes Pearson product-moment correlation coefficient,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2 \sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2}}$$

where x_i and y_i are the ith attributes of $\mathbf{x}$ and $\mathbf{y}$, respectively and $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the corresponding vector means.

It is to consider that the ranges of Pearson correlation is $[-1, 1]$ and that the $1-r$ lies between [0, 2]. A convenient measure between 0 and 1 is $\frac{1}{2}(1-r)$ in where the low distance signifies for positive correlation.

2.8.2 Distances for binary data

Out of many proposed methods for the dichotomize categorical variables, we used *Jaccard's distance* and *phi correlation coefficient* as a distance measure. *Jaccard's distance* usually focus on the asymmetric information in the binary variables. Where the asymmetric means the value of present (1) and absent (0) do not carry equal information. On the basis of our binary data, we made our first believe that the co-absences (0 and 0) in both variables are not important and counting the values may have no meaningful contribution to the distance measure. Based on this believe, we used Jaccard's distance as a measure of dissimilarity between two binary variables excluding all co-absences information. In contrary, *phi correlation coefficient* considers symmetric information in our binary variables i.e., equal weight has given to presences and absences cases. This *phi* method is binary analogue of the usual Pearson correlation coefficient.

2.8.2.1 Jaccard's distance

Given two n-dimensional *binary* data vectors $\mathbf{x}$ and $\mathbf{y}$, and the counts of presences (1) and absences (0) in the pairs are represented in a 2×2 contingency table 2.2 as below.

Table 2.2: Table for counts of binary samples.

		Vector y		
		1	0	
Vector x	1	m	r	m+r
	0	p	q	p+q
		m+p	r+q	m+r+p+q

where,

m = number of presences (1 and 1) on both vectors.

r = number of mismatches where $\mathbf{x}$ has value 1 but $\mathbf{y}$ has value 0.

p = number of mismatches where $\mathbf{x}$ has value 0 but $\mathbf{y}$ has value 1.

q = number of absences (0 and 0) on both vectors.

$w = m + r + p + q$ = total counts for both vectors.

Jaccard's distance (dissimilarity measure) is therefore defined by

$$d_{jac}(\mathbf{x}, \mathbf{y}) = 1 - s_{jac}(\mathbf{x}, \mathbf{y}) \tag{2.7}$$

where, $s_{jac}(\mathbf{x}, \mathbf{y})$ is the Jaccard's similarity coefficient and can be defined by

$$s_{jac}(\mathbf{x}, \mathbf{y}) = \begin{cases} 1 & ; \text{ if m=r=p=0} \\ \frac{m}{m+r+p} & ; \text{ otherwise} \end{cases} \tag{2.8}$$

which ranges in $[0, 1]$.

Thus the Jaccard's distance becomes,

$$d_{jac}(\mathbf{x}, \mathbf{y}) = 1 - s_{jac}(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & ; \text{ if m=r=p=0} \\ \frac{r+p}{m+r+p} & ; \text{ otherwise} \end{cases} \tag{2.9}$$

From the above equations we can say that if our two study symptoms (variables) have sets of cases with many co-presences and few mismatches (regardless of co-absences) then their dissimilarity measure (distance) will be minimum.

2.8.2.2 Phi correlation coefficient

For the categorical binary variables, the correlation coefficient also can be used to estimate the dissimilarity between two objects. Given two n-dimensional data vectors $\mathbf{x}$ and $\mathbf{y}$, the *Phi* distance can be defined by

$$d_{\phi}(\mathbf{x}, \mathbf{y}) = 1 - r_{\phi}(\mathbf{x}, \mathbf{y}) \tag{2.10}$$

where $r_{\phi}(\mathbf{x}, \mathbf{y})$ is defined as the binary correlation coefficient and can be represent by

$$r_{\phi}(\mathbf{x}, \mathbf{y}) = \frac{mq - rp}{\sqrt{(m+r)(p+q)(m+p)(r+q)}} \tag{2.11}$$

where m, r, p and q represent similarly as above in Table 2.2.

Note that the ranges of binary correlation coefficient, $r_{\phi}(\mathbf{x}, \mathbf{y})$ is $[-1, 1]$ and that the $1 - r_{\phi}(\mathbf{x}, \mathbf{y})$ lies between 0 (when $rp = 0$, no mismatches) to 2 (when $mq = 0$, no matches). A convenient measure between 0 and 1 is $\frac{1}{2}(1 - r_{\phi}(\mathbf{x}, \mathbf{y}))$ in where the low distance signifies for positive correlation.

2.8.3 Gower's similarity/dissimilarity coefficient

Generally, we consider Gower's similarity or dissimilarity measure for mixed data types, i.e., a dataset having mixture of binary, nominal, ordinal and continuous variables and therefore is one of the most popular measures of proximity. Gower's original method was proposed to estimate the distances between pair of cases by considering variables. Therefore for the purpose of our this study we slightly modified the Gower's original method by considering to estimate the distances between variables instead of cases/objects. In this case we have only ordinal variables and we focused our attention on specifying the contribution made to measure the dissimilarity between variables done by a single object/case. Overall distance between a pair of variables is then obtained by summing weighted such contributions over all the objects.

Given two n-dimensional data vectors $\mathbf{x}$ and $\mathbf{y}$, the Gower's dissimilarity Coefficient is defined by

$$d_{gower}(\mathbf{x}, \mathbf{y}) = \sqrt{1 - S_{gower}(\mathbf{x}, \mathbf{y})} \tag{2.12}$$

where, $S_{gower}(\mathbf{x}, \mathbf{y})$ is the Gower's similarity coefficient defined by

$$S_{gower}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} w(x_k, y_k) s(x_k, y_k)}{\sum_{k=1}^{n} w(x_k, y_k)} \tag{2.13}$$

in where, $s(x_k, y_k)$ denotes the similarity provided by the kth object, and

for ordinal variables,

$$w(x_k, y_k) = \begin{cases} 1 & ; \text{ if the data points } \mathbf{x} \text{ and } \mathbf{y} \text{ are comparable for the } k\text{th object. i.e., for an object/survivor who answered a constant value for all the variables has } R_k = 0, \text{ if not comparable and hence discarded those objects.} \\ 0 & ; \text{ otherwise} \end{cases}$$

Also for ordinal variables the value of $s(x_k, y_k)$ can be computed as

$$s(x_k, y_k) = 1 - \frac{|x_k - y_k|}{R_k} \tag{2.14}$$

where R_k is the range, which represent the difference between the maximum and the minimum values of object k present in all samples under consideration.

So, the Gower's distance can be simplified as

$$d_{gower}(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\frac{|x_k - y_k|}{R_k}} \qquad ;iff \quad w(x_k,y_k) = 1\forall k \quad and \quad \sum_{k=1}^{n} w(x_k,y_k) = n \tag{2.15}$$

where n is the number of objects and R_k is the range for kth object. The ranges of Gower's distance is [0, 1].

2.8.4 Distances for Rank Data

To deal with pure rank of our ordinal data, we used Kendall tau, Spearman footrule, and Goodman-Kruskal gamma coefficient as our distances measure.

2.8.4.1 Kendall-Tau rank distance

The Kendall tau rank distance is a metric that consider the pairwise disagreements and counts the total number of discordant pairs between two ranking lists.

Given two n-dimensional data vectors $\mathbf{x}$ and $\mathbf{y}$, the normalized Kendall tau distance can be represented by

$$d_{kendall}(\mathbf{x},\mathbf{y}) = \frac{n_d}{\frac{1}{2}n(n-1)} \in [0,1] \tag{2.16}$$

where, n_d denotes the number of discordant pairs between $\mathbf{x}$ and $\mathbf{y}$ and $n_d \in [0, \frac{1}{2}n(n-1)]$, and, $\frac{1}{2}n(n-1)$ denotes the kendall's distance normalizing factor, which defines the total number of pairs of items in the two ordered list and in where n is the list size and is thus the upper limit of n_d.

The value $d_{kendall}(\mathbf{x},\mathbf{y}) = 0$ means that the two measure ranking list are identical where as $d_{kendall}(\mathbf{x},\mathbf{y}) = 1$ means they are in opposite order.

For the computation of Kendall's tau distance, it is simple to verify that

$$d_{kendall}(\mathbf{x},\mathbf{y}) = \frac{1}{2}(1-\tau) \tag{2.17}$$

where τ is Kendall's tau rank correlation coefficient and can be defined by

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{2.18}$$

where,
n_c = total counts for the pairs of items ranked in the same order on both variables (i.e., number of concordant pairs for two rank list), and
n_d = total counts for the pairs of items ranked differently on both variables (i.e., number of discordant pairs for two rank list).

Therefore,

$$n_c = |\{(i, i') : i < i', (\mathbf{x}_i < \mathbf{x}_{i'} \quad \text{and} \quad \mathbf{y}_i < \mathbf{y}_{i'}) \\ \text{or} \quad (\mathbf{x}_i > \mathbf{x}_{i'} \quad \text{and} \quad \mathbf{y}_i > \mathbf{y}_{i'})\}|$$

and

$$n_d = |\{(i, i') : i < i', (\mathbf{x}_i < \mathbf{x}_{i'} \quad \text{and} \quad \mathbf{y}_i > \mathbf{y}_{i'}) \\ \text{or} \quad (\mathbf{x}_i > \mathbf{x}_{i'} \quad \text{and} \quad \mathbf{y}_i < \mathbf{y}_{i'})\}|$$

2.8.4.2 Footrule distance

Footrule distance is defined as an absolute distance between two rank list, i.e., it computes a total element-wise displacement from a identity combination. This method is similar to the city block distance or Manhattan distance that used for numerical data, but in where Footrule distance is used for rank data. This measure is also named as "Spearman footrule distance".

Given two n-dimensional data vectors $\mathbf{x}$ and $\mathbf{y}$, the Spearman Footrule distance can be defined by

$$d_{footrule}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} \left| \rho(\mathbf{x}_k) - \rho(\mathbf{y}_k) \right| = \sum_{k=1}^{n} \left| \rho_{kj} - \rho_{kj'} \right| \tag{2.19}$$

where $\rho(\mathbf{x}_k)$ denotes ranking of variable $\mathbf{x}$ for individual k.

The distance matrix is always positive definite.

2.8.4.3 Goodman and Kruskal's gamma

Goodman and Kruskal's gamma usually measure a rank correlation statistic, i.e., it assess the strength of association of the orderings of the data when ranked by each of their entities. The method is relevant and good in the situation where there are many ties or zeroes in the data and the variables are in ordinal level.

Given two n-dimensional data vectors $\mathbf{x}$ and $\mathbf{y}$, the Goodman and Kruskal's gamma distance

measure can be defined by

$$d_{gamma}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(1 - \gamma) \tag{2.20}$$

where, γ is the Goodman and Kruskal's measure and defined as,

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

where, n_c and n_d were defined earlier.

For the tied observations between two ordered list, the method drop them from the calculation, i.e., no adjustment for ties observations.

The value of Goodman and Kruskal's gamma ranges from -1 (perfect negative association) to $+1$ (perfect positive agreement) and a value of zero indicates the absence of association. Hence, the range of gamma distance becomes [0, 2] and for the convenience of our computation we divide this range by 2 to make it [0, 1].

2.9 Dissimilarity measures between clusters

Hierarchical cluster analysis is a set of nested partitions and follows two types of algorithms: agglomerative (bottom-up clustering) or divisive (top-down clustering). In bottom-up clustering algorithm, each element is considered as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters which have minimum distance. The process of merging clusters is repeated until all clusters have been assembled into a one big cluster that contains all data points. In top-down (divisive) clustering algorithm, the mechanism starts in a opposite way by considering all data points in a single cluster and then successively splitting into minor clusters based on their distances. For the both scenario, it is necessary to compute the distance between two clusters.

In practice the agglomerative hierarchical clustering is the most widely used method and we consider it for our this study as well. In agglomerative hierarchical clustering, the clustering methods differ based on the way to define the distance between two clusters. Lance and Williams [25] proposed a general linear relation to calculate the distance between two groups/cluster and have been using most widely in the practice of hierarchical clusters.

2.9.1 Lance and Williams Dissimilarity measure

In all hierarchical agglomerative clustering algorithm, it is required to calculate the inter group dissimilarities between points of new cluster formed by two clusters and existing clusters. Lance and Williams (1966) [25] suggested a general formula which is known as the Lance-Williams combinatorial dissimilarity formula, that measure the differences between

new clusters and other existing clusters, based on the distances prior to merge of the new cluster. If two clusters G_i and G_j with r_i and r_j elements respectively have been merged to form a new cluster G_k with $r_k(= r_i + r_j)$ elements, then the distance $D(.,.)$ between the new cluster G_k and any existing cluster G_h in the space is given by the following recurrence formula:

$$\begin{aligned} D(G_h, G_k) &= D(G_h, G_i \cup G_j) \\ &= \alpha_i D(G_h, G_i) + \alpha_j D(G_h, G_j) + \beta D(G_i, G_j) \\ &+ \gamma |D(G_h, G_i) - D(G_h, G_j)| \end{aligned} \tag{2.21}$$

where the parameters $\alpha_i, \alpha_j, \beta$ and γ in equation (2.21) determine the nature of the clustering strategy. The hierarchical clustering algorithms with various inter-cluster dissimilarities can be achieved by choosing suitable values of the above Lance-Williams parameters, which are presented in Table 2.3.

Table 2.3: Some chosen values commonly used for the Lance-Williams parameters in hierarchical clustering.

Hierarchical Algorithm	α_i	α_j	β	γ
Single Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Group Average	$\frac{r_i}{r_i+r_j}$	$\frac{r_j}{r_i+r_j}$	0	0
Ward's Method	$\frac{r_h+r_i}{\Delta_{ijh}}$	$\frac{r_h+r_j}{\Delta_{ijh}}$	$-\frac{r_h}{\Delta_{ijh}}$	0

where, r_i is the number of elements in cluster G_i and $\Delta_{ijh} = r_i + r_j + r_h$.

Unlike to the general clustering algorithms, there exists also some algorithms specific to individual hierarchical agglomerative clustering methods which are known as linkage criterion. These criterion represent the differences between sets of elements, that is a function of the distances between elements in pairs. Some most widely used linkage methods as well as used for this study are discussed below.

2.9.1.1 Single linkage method

Single linkage is the simplest and oldest of the conventional agglomerative methods. In the single linkage method, it consider the smallest distance between two elements each of which is in one of the two groups and merges those closest elements as a new cluster. Therefore, this algorithm is also known as the nearest-neighbor method or the minimum method.

Let $G = \{y_1, y_2, \cdots, y_r\}$ and $G' = \{z_1, z_2, \cdots, z_s\}$ are two nonempty, non-overlapping clusters

of size r and s, respectively. So, the single linkage method can be defined as:

$$D(G, G') = \min_{\mathbf{y}\in G, \mathbf{z}\in G'} d(\mathbf{y}, \mathbf{z}) \tag{2.22}$$

where $d(.,.)$ denotes the distance function between data points from which the proximity matrix is computed.

Equation (2.22) is equivalent to and easy to compute from the general Lance-Williams equation (2.21).

Suppose, we have three clusters of elements G_i, G_j and G_h. Then the dissimilarity between cluster G_h and cluster G_k ($G_k = G_i \cup G_j$) can be measured from equation (2.21) as follows:

$$\begin{aligned} D(G_h, G_k) &= D(G_h, G_i \cup G_j) \\ &= \frac{1}{2}D(G_h, G_i) + \frac{1}{2}D(G_h, G_j) - \frac{1}{2}|D(G_h, G_i) - D(G_h, G_j)| \\ &\Big[\text{when} \quad \alpha_i = \alpha_j = \frac{1}{2}; \quad \beta = 0 \quad and \quad \gamma = -\frac{1}{2} \quad \text{in equation (2.21)}\Big] \\ &= min\{D(G_h, G_i), D(G_h, G_j)\} \end{aligned} \tag{2.23}$$

where $D(.,.)$ is a distance function between two groups.

Figure 2.2(a) shows a graphical representation of the single-linkage method for two groups, in where it consider the shortest distance between two closest points in different subsets of points.

2.9.1.2 Complete linkage

Complete linkage algorithm usually suggests for the needed of intense grouping and is the exact opposite of the preceding, in which the dissimilarity between two clusters defines by the distance between two furthest pair of elements, one from each group.

Let $G = \{y_1, y_2, \cdots, y_r\}$ and $G' = \{z_1, z_2, \cdots, z_s\}$ are two nonempty, non-overlapping clusters of size r and s, respectively. So, the complete linkage method can be defined as:

$$D(G, G') = \max_{\mathbf{y}\in G, \mathbf{z}\in G'} d(\mathbf{y}, \mathbf{z}) \tag{2.24}$$

where $d(.,.)$ denotes the distance function between data points from which the proximity matrix is computed.

Equation (2.24) is equivalent to and easy to compute from the general Lance-Williams equation (2.21).

Suppose, we have three clusters of elements G_i, G_j and G_h. Then the dissimilarity between cluster G_h and cluster G_k $(G_k = G_i \cup G_j)$ can be measured from equation (2.21) as follows:

$$\begin{aligned} D(G_h, G_k) &= D(G_h, G_i \cup G_j) \\ &= \frac{1}{2}D(G_h, G_i) + \frac{1}{2}D(G_h, G_j) + \frac{1}{2}|D(G_h, G_i) - D(G_h, G_j)| \\ &\left[\text{when} \quad \alpha_i = \alpha_j = \frac{1}{2}; \quad \beta = 0 \quad and \quad \gamma = \frac{1}{2} \quad \text{in equation (2.21)}\right] \\ &= max\{D(G_h, G_i), D(G_h, C_j)\} \end{aligned} \tag{2.25}$$

where $D(.,.)$ is a distance function between two groups.

Figure 2.2(b) shows a graphical representation of the complete-linkage method for two groups, in where it consider the longest distance between two furthest points in different subsets of points.

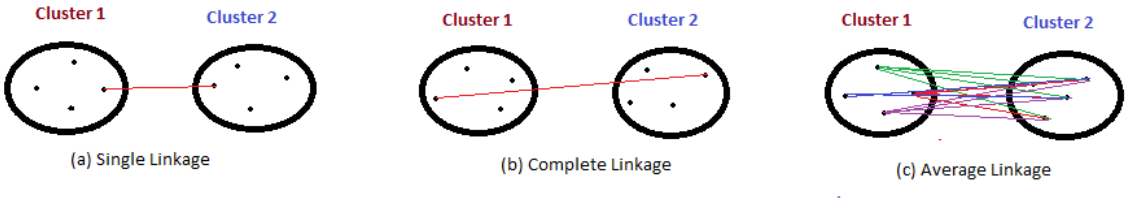

Figure 2.2: Representation of inter-cluster dissimilarity

2.9.1.3 Group Average

The distance between two clusters can also be measured as the average distance between elements from the first cluster and elements from the second cluster. Such a method is known as the *group average clustering algorithm*. The averaging is considered over all pairs of elements in both groups.

Let $G = \{y_1, y_2, \cdots, y_r\}$ and $G' = \{z_1, z_2, \cdots, z_s\}$ are two nonempty, non-overlapping clusters of size r and s, respectively. So, the group average method can be defined as:

$$D(G, G') = \frac{1}{r \times s} \sum_{\mathbf{y} \in G, \mathbf{z} \in G'} d(\mathbf{y}, \mathbf{z}) \tag{2.26}$$

where $r = |G|$ and $s = |G'|$ are the numbers of elements in clusters G and G' respectively. and
$d(.,.)$ denotes the distance function between data points from which the proximity matrix is computed where $y \in G, z \in G'$.

Equation (2.26) is equivalent to and easy to evaluate from the general Lance-Williams equation (2.21).

Suppose, we have three clusters of elements G_i, G_j and G_h. Then the dissimilarity between cluster G_h and cluster G_k ($G_k = G_i \cup G_j$) can be measured from equation (2.21) as follows:

$$\begin{aligned} D(G_h, G_k) &= D(G_h, G_i \cup G_j) \\ &= \frac{|G_i|}{|G_i| + |G_j|} D(G_h, G_i) + \frac{|G_j|}{|G_i| + |G_j|} D(G_h, G_j) \\ &= \frac{r_i}{r_k} D(G_h, G_i) + \frac{r_j}{r_k} D(G_h, G_j) \end{aligned} \tag{2.27}$$

when $\alpha_i = \frac{|G_i|}{|G_i|+|G_j|} = \frac{r_i}{r_k}$, $\alpha_j = \frac{|G_j|}{|G_i|+|G_j|} = \frac{r_j}{r_k}$; $\beta = 0$ *and* $\gamma = 0$ in equation (2.21).

where $|G_i| = r_i$ is the number of elements in cluster G_i and $|G_i| + |G_j| = r_i + r_j = r_k$ and $D(.,.)$ is the distance function between two clusters.

2.9.1.4 Ward's Minimum-Variance criterion

Ward Jr. (1963) [26] and Ward Jr. and Hook (1963) [27] proposed an agglomerative clustering algorithm which is established on the error sum of squares (E) criterion and can be defined as the sum of squared distances of each data points from the mean of it's assigned cluster. At each stage of the merging, the aim of Ward's method is to find compact, spherical clusters by minimizing the increase in the total within-group error sum of squares. Hence, this method is also referred to as the *Ward's minimum variance criterion* [24]. At the beginning of the algorithm E is set to 0 as each elements is then appear in its own cluster.

Suppose there are h singletons clusters $G_1, G_2, \cdots, G_h$, then the Ward's minimum variance criterion can be expressed as:

$$E = \sum_{j=1}^{h} E(G_j) \tag{2.28}$$

which is the total within-cluster sum of squares and where

$$E(G_j) = \sum_{l=1}^{r_j} \sum_{m=1}^{p_m} \Big(x_{jlm} - \bar{x}_{jm} \Big)^2 \tag{2.29}$$

in which $\bar{x}_{jm} = (\frac{1}{r_j}) \sum_{l=1}^{r_j} x_{jlm}$; is the mean of the j^{th} cluster for the m^{th} variable, and, x_{jlm} be the score on the m^{th} variable ($m = 1, 2, \cdots, p$) for the l^{th} object ($l = 1, 2, \cdots, r_j$) in the j^{th} cluster ($j = 1, 2, \cdots, k$).

In matrix notation we can write the equation (2.29) as follows:

Let we have a cluster G with data points $\{z_1, z_2, \cdots, z_s\}$ of size $s = |G|$, the error sum of squares for the cluster G can be computed by:

$$\begin{aligned} E(G) &= \sum_{\mathbf{z}\in G} (\mathbf{z} - \bar{G})(\mathbf{z} - \bar{G})^T \\ &= \sum_{\mathbf{z}\in G} \left(\mathbf{z}\mathbf{z}^T - \bar{G}\mathbf{z}^T - \mathbf{z}\bar{G}^T + \bar{G}\bar{G}^T\right) \\ &= \sum_{\mathbf{z}\in G} \mathbf{z}\mathbf{z}^T - \bar{G}\Big(\sum_{\mathbf{z}\in G} \mathbf{z}\Big)^T \\ &= \sum_{\mathbf{z}\in G} \mathbf{z}\mathbf{z}^T - \frac{1}{s}\Big(\sum_{\mathbf{z}\in G} \mathbf{z}\Big)\Big(\sum_{\mathbf{z}\in G} \mathbf{z}\Big)^T \\ &= \sum_{\mathbf{z}\in G} \mathbf{z}\mathbf{z}^T - s\bar{G}\bar{G}^T \end{aligned} \tag{2.30}$$

where $\bar{G}$ is the mean of cluster G, that is,

$$\bar{G} = \frac{1}{s}\sum_{\mathbf{z}\in \mathbf{G}} \mathbf{z} \tag{2.31}$$

At each step of clustering the pair of clusters are merged whose cluster distance is minimum. Therefore, one can implement this method by finding the pairwise clusters at every stage that accounts for the minimum increase in total within-cluster variance after fusion.

Performing hierarchical clustering with Ward's criterion above is equivalent to and easy to evaluate from the general Lance-Williams equation (2.21).

Suppose, we have three clusters of elements G_i, G_j and G_h. Then the dissimilarity between cluster G_h and cluster G_k ($G_k = G_i \cup G_j$) can be measured from equation (2.21) as follows:

$$\begin{aligned} D(G_h, G_k) &= D(G_h, G_i \cup G_j) \\ &= \frac{|G_h| + |G_i|}{\Delta_{ijh}} D(G_h, C_i) + \frac{|G_h| + |G_j|}{\Delta_{ijh}} D(G_h, G_j) \\ &- \frac{|G_h|}{\Delta_{ijh}} D(G_i, G_j) \end{aligned} \tag{2.32}$$

when $\alpha_i = \frac{|G_h|+|G_i|}{\Delta_{ijh}} = \frac{r_h+r_i}{\Delta_{ijh}}$; $\alpha_j = \frac{|G_h|+|G_j|}{\Delta_{ijh}} = \frac{r_h+r_j}{\Delta_{ijh}}$; $\beta = -\frac{|G_h|}{\Delta_{ijh}} = -\frac{r_h}{\Delta_{ijh}}$ *and* $\gamma = 0$ in equation (2.21).

where $\Delta_{ijh} = |G_h| + |G_i| + |G_j| = r_h + r_i + r_j$.

During the process of clustering, the dissimilarity matrix is updated with equation (2.32) and the two clusters with minimum distance will be fused together. Then the increase in error

sum of squares, E, of the matrix is computed by:

$$\Delta E_{h(ij)} = \frac{1}{2} D(G_h, G_i \cup G_j) \tag{2.33}$$

For example, suppose we have squared Euclidean distance as a distance matrix and suppose groups G_i and G_j are chosen to be merged to make a new cluster G_k, i.e., $G_k = G_i \cup G_j$. Then the increase in E is

$$\begin{aligned}
\Delta E_{ij} &= E(G_k) - E(G_i) - E(G_j) \\
&= \left(\sum_{\mathbf{z} \in G_k} \mathbf{z}\mathbf{z}^T - |G_k| \bar{G}_k \bar{G}_k{}^T \right) - \left(\sum_{\mathbf{z} \in G_i} \mathbf{z}\mathbf{z}^T - |G_i| \bar{G}_i \bar{G}_i{}^T \right) \\
&- \left(\sum_{\mathbf{z} \in G_j} \mathbf{z}\mathbf{z}^T - |G_j| \bar{G}_j \bar{G}_j{}^T \right) \qquad \Big[\text{From equation (2.30)}\Big] \\
&= |G_i| \bar{G}_i \bar{G}_i{}^T + |G_j| \bar{G}_j \bar{G}_j{}^T - |G_k| \bar{G}_k \bar{G}_k{}^T
\end{aligned}$$

where $\bar{G}_k, \bar{G}_i$ and $\bar{G}_j$ are the means of clusters G_k, G_i and G_j, respectively.

2.9.2 Dendogram

To visualize the clustering results various versions of the dendrogram have been proposed and widely used. The main objective of this visualization is a tree based structure yielded by a hierarchical algorithm which is easy to understand the similarities/dissimilarities between elements. It demonstrates both the cluster and the sub-cluster relationships and also the order in which the clusters are merged for agglomerative algorithm or split for divisive case. Usually a dendrogram is read from left to right.

In a dendrogram, the data points are represented along the bottom and referred to as *external/terminal nodes*, and clusters to which the data belong are formed by joining with the *internal nodes*. The name of objects added to the terminal nodes are known as *labels*. The *height* of the dendrogram shows the distance between objects or clusters. Therefore, the small values of the height indicates a high similarity between the objects and the large values of the height indicates more distance. This can be shown in the figure 2.3. The coalition between each internal node and the height can be explain mathematically by the following condition:

$$h(Q) \leq h(R) \Leftrightarrow Q \subseteq R$$

for every subsets of objects Q and R and $Q \cap R \neq \Phi$,
Here, $h(Q)$ and $h(R)$ represent the heights of Q and R, respectively .

Let we have a set of p variables $x_1, x_2, \cdots, x_p$ and the pairwise distances of these variables can be represent in a $p \times p$ matrix of **D**.

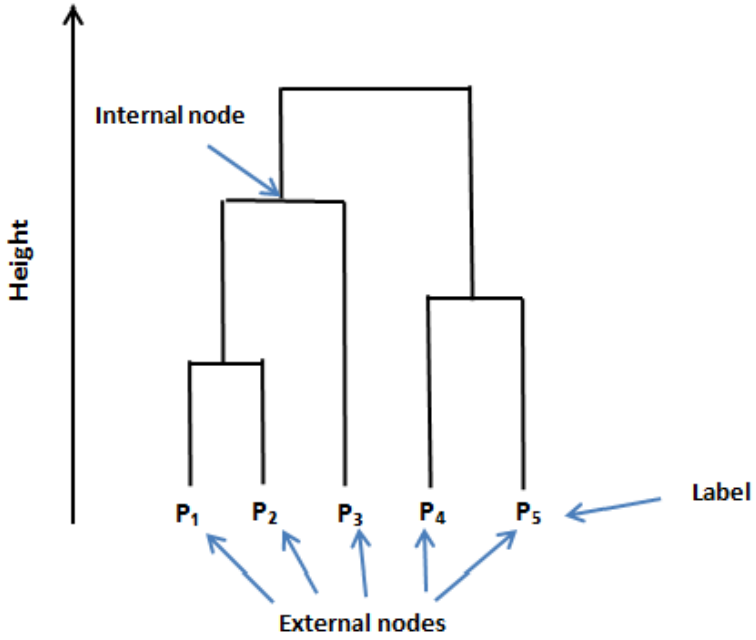

Figure 2.3: Visualization of five variables in a dendrogram

2.10 Method for Comparing Partitions

2.10.1 Adjusted Rand Index

A difficulty often arise during clustering when it comes to make a decision about the optimal number of clusters. There are many procedures available in literature for this purpose based on the *external* and *internal* criteria of the clusterings. However, for the objective of this study we carried out *external criteria* approach to evaluate the results of our clustering algorithms and to make a decision about the optimal cluster numbers. This criterion suggests that the evaluation is done by comparing a pre-specified arrangement of a data set during clustering with reflects to the intuitive structure on it.

"Rand index" proposed by Rand [28] and "Adjusted Rand index" indices are probably the most popular and frequently used measures for the cluster validation with *external criteria*

approach. Both of these methods compare pairwise cluster partitions agreement, in where one partition is obtained by the clustering procedure and the other is defined by some external criteria of the data, i.e., the comparison is based on comparing pairs of objects concerning their group attributes.

Suppose we have a set of p variables $S = \{x_1, x_2, \cdots x_p\}$ and also consider $Q = \{q_1, q_2, \cdots q_C\}$ and $T = \{t_1, t_2, \cdots t_R\}$ represent two different partitions of the same items in S such that $\cup_{k=1}^{C} q_k = S = \cup_{l=1}^{R} T_l$ and $q_k \cap q_{k'} = \emptyset = t_l \cap t_{l'}$ for $1 \leq k \neq k' \leq C$ and $1 \leq l \neq l' \leq R$. Let Q be the partition of the data done by some external criteria with subset C and T is the partition of a clustering result with subset R. The group overlap between Q and T can represent in a contingency matrix $[m_{cr}]$ shown in Table 2.4 where each entities m_{cr} denotes the number of variables that is common between partitions q_k and t_l i.e., $m_{cr} = |q_k \cap t_l|$, and $m_{c.}$ and $m_{.r}$ are the marginal totals of q_c and t_r respectively.

Table 2.4: Contingency tables for pair of observations between Q and T.

Partition	*Group*	T						Total
		t_1	t_2	$\cdots$	t_r	$\cdots$	t_R	
Q	q_1	m_{11}	m_{12}	$\cdots$	m_{1r}	$\cdots$	m_{1R}	$m_{1.}$
	q_2	m_{21}	m_{22}	$\cdots$	m_{2r}	$\cdots$	m_{2R}	$m_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	q_c	m_{c1}	m_{c2}	$\cdots$	m_{cr}	$\cdots$	m_{cR}	$m_{c.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	q_C	m_{C1}	m_{C2}	$\cdots$	m_{Cr}	$\cdots$	m_{CR}	$m_{C.}$
Total		$m_{.1}$	$m_{.2}$	$\cdots$	$m_{.r}$	$\cdots$	$m_{.R}$	$m_{..} = p$

For a dataset with p variables there will be $\binom{p}{2}$ pairs of possible combinations which we can define in the following four ways:

n_{11} = number of pairs of variables in S that are in the same set in Q and in the same set in T ;
n_{10} = number of pairs of variables in S that are in the same set in Q and in different sets in T ;
n_{01} = number of pairs of variables in S that are in different sets in Q and in the same set in T;
n_{00} = number of pairs of variables in S that are in different sets in Q and in the different sets in T.

The quantities n_{11} and n_{00} measure the agreements, and n_{10} and n_{01} measure the disagreements between two partitions.

From the above we can construct an alternative 2×2 contingency Table 2.5 as:

Table 2.5: Contingency table for agreements and disagreements.

Partition Q	T		Total
	Pairs in same set	Pairs in different sets	
Pairs in same set	n_{11}	n_{10}	$n_{11} + n_{10}$
Pairs in different sets	n_{01}	n_{00}	$n_{01} + n_{00}$
$Total$	$n_{11} + n_{01}$	$n_{10} + n_{00}$	p

Hence, the formula for the Rand Index can be represented by:

$$Rand\ Index = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}} \tag{2.34}$$

The range of Rand index is $[0, 1]$. It takes the value of 1 when the two comparing sets of partitions are perfectly agreed, and 0 when no pair of objects occur in the same group or in different groups in both partitions, i.e. $n_{11} = n_{00} = 0$. This scenario usually happens when one partition holds a single cluster of the data objects while the other partition contains only of clusters of single objects. Thus, this method provides higher weights to those pair of objects that are classified together and apart in both clustering partitions. However, though it is desirable, the expected value of the Rand index between two random labeling won't yield zero values, or at least a constant value. Therefore, the similarity index takes its upper limit of unity when the number of clusters becoming larger. To overcome this problem and also for the purpose of this study we implemented Adjusted Rand Index, proposed by Hubert and Arabie (1985) [29], as our measure of agreement between the external criteria of the data and the clustering results. The Adjusted Rand index has been suggested as the index of choice for assessing agreement between two random cluster labelings, in a study comparing the performance of several agreement indices with different numbers of clusters [30, 31].

The Adjusted Rand index is a modification of the Rand index that is corrected for the chance grouping of objects and by considering general hypergeometric distribution of this randomness. This means the two cluster partitions, say Q and T, are chosen at random in where the number of elements in both partitions are fixed. The suggested modified version of the Rand index is:

$$Adjusted\ Index = \frac{\text{Index} - \text{Expected Index}}{\text{Maximum Index} - \text{Expected I1ndex}} \tag{2.35}$$

Therefore, the Adjusted Rand index proposed by Hubert and Arabie can be written in more specifically as:

$$\textit{Adjusted Rand Index} = \frac{\sum_{c,r}\binom{m_{cr}}{2} - \Big[\sum_c \binom{m_{c.}}{2}\sum_r \binom{m_{.r}}{2}\Big]/\binom{p}{2}}{\frac{1}{2}\Big[\sum_c \binom{m_{c.}}{2} + \sum_r \binom{m_{.r}}{2}\Big] - \Big[\sum_c \binom{m_{c.}}{2}\sum_r \binom{m_{.r}}{2}\Big]/\binom{p}{2}} \tag{2.36}$$

where m_{cr}, $m_{c.}$, $m_{.r}$ are values obtained from the above contingency table.

The Index above can yield a value between -1 and $+1$ and equals to its expected value when it takes the value 0. So, the maximum value of the index indicates strong similarity between two partitions.

2.11 Exploratory Factor Analysis

Exploratory Factor Analysis is a multivariate statistical technique used to discover the underlying structure of a relatively large set of variables. As one of the technique of factor analysis, the aim of exploratory factor analysis is to reveal the underlying relationships among measured variables and reduce the dimensionality of the data. Ideally, this method is used in situations where there is no a priori hypothesis about the factors or structure of measured variables and also before applying *confirmatory factor analysis* for developing an index. The exploratory factor analysis technique follows the principle of common factor model in where the measured/observed variables are expressed as a function of common factors, unique factors, and errors of measurements. According to MacCallum [32], in a analysis the accuracy of the exploratory factor procedures can be measured when every factor is described by numerous indicators/observed variables and there should be at least 3 to 5 indicators on each factor. In this study our objective is to identify the numbers of underlying common factors that is responsible for the correlation among symptoms. We performed exploratory factor analysis in parallel with the cluster analysis, because factor analysis uses correlations among variables to search for common clusters. Hence, to identify number of clusters of inter-correlated variables and nature of latent constructs or factors that describe our *phi* correlation structure we executed the factor analysis methodology. For the simplicity of this study we assumed that the underlying factors of the indicator variables (symptoms) are independent (i.e., orthogonal). For this study the basic steps involved to use the factor analysis are:

- Use of *phi* correlation matrix for the analysis.
- Estimate communalities with principal component method.
- Decide on the number of factors to be retained.
- Factor rotation (Varimax).
- Interpretation of results (i.e., factor loadings).

2.11.1 The Orthogonal Factor Model

The goal of factor analysis is to interpret the effect of observed variables in a data matrix $\mathbf{X}$ using fewer random variables, are called factors. These factors are underlying constructs and referred as latent variables or unobserved factors which share common characteristics of the observed $x \in \mathbb{R}^p$.

Suppose we have $\tilde{X}$ of p observed variables, $\tilde{X} = (x_1, x_2, \cdots, x_p)^T$. The factor analysis model describe each of the observed variables as a linear combination of underlying common factors $l_1, l_2, \cdots, l_g$ associated with an error term that is unique and account for that particular

variable only. For each observed $\tilde{X} = (x_1, x_2, \cdots, x_p)^T$, we can write

$$x_j = \sum_{k=1}^{g} \lambda_{jk} l_k + \varepsilon_j + \mu_j; \qquad j = 1, 2, \cdots, p \tag{2.37}$$

Ideally, the choice of g must be substantially lower than p. In equation 2.37, the coefficients λ_{jk} denote *pattern or factor loadings* and considered as weights of the factors, μ_j denotes the mean of variable j and the term ε_j denotes jth error term of variable x_j which is referred as *unique factor*.

Then in matrix notation the equations can be written as:

$$\tilde{\mathbf{X}} = \Lambda \mathbf{L} + \boldsymbol{\varepsilon} + \boldsymbol{\mu} \tag{2.38}$$

where, $\tilde{\mathbf{X}}$ is a p-dimensional random vector, $\tilde{\mathbf{X}} = (x_1, x_2, \cdots, x_p)^T$
$\boldsymbol{\mu}$ is a p-dimensional mean vector of $\tilde{\mathbf{X}}$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^T$
$\mathbf{L}$ is a g-dimensional vector of the g factors, $\mathbf{L} = (l_1, l_2, \cdots, l_g)^T$
$\boldsymbol{\varepsilon}$ is a p-dimensional vector of specific factors $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)^T$ and
Λ is a $(p \times g)$ matrix of factor loadings,

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1g} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pg} \end{pmatrix}$$

For our factor models 2.37 and 2.38, we assume that the factors are not correlated with the unique factors (error components), and the means and variances of the random variables and factors are zero and one, respectively.

$$\left.\begin{array}{lll} i) & E(l_k) = 0, \quad var(l_k) = 1 \quad \text{and} \quad cov(l_k, l_j) = 0, k \neq j \\ ii) & E(\mu_j) = 0, \quad var(\mu_j) = 1 \quad \text{and} \quad cov(\mu_j, \mu_h) = 0, j \neq h \\ iii) & E(\varepsilon_j) = 0, \quad var(\varepsilon_j) = \psi_j \quad \text{and} \quad cov(\varepsilon_j, \varepsilon_h) = 0, j \neq h \\ iv) & cov(\varepsilon_j, l_k) = 0 \quad \forall i, j \end{array}\right\} \; k = 1, 2, \cdots, g \text{ and } j = 1, 2, \cdots, p$$

where $\Psi = (\psi_1, \psi_2, \cdots, \psi_p)^T$ referred as *error componest or unique variances.*

2.11.2 Graphical representation

Figure 2.4 illustrates a general pattern of factor structure among five hypothetical indicators/variables (A, B, C, D and E) and a common factor or unobserved construct (l).

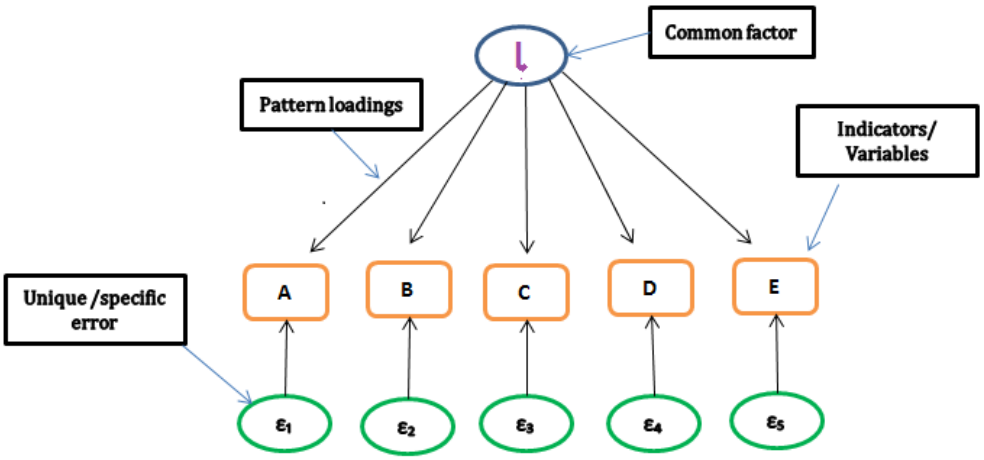

Figure 2.4: Relationship among five indicators and a common factor.

Some general terms use in a one factor analysis are:

- The total variance of any random variable can be partitioned into two components: (a) variance that is common with the factor (l) and can be obtained from the square of the patter loadings. The part of this variance is referred to as communality. In a factor model it is ideal to consider that the greater the communality the better the measure is and vice versa. and

  b) variance that is common with the unique factor (ε) and can be obtained from (the variance of the variable − the communality). This part of the variance is referred to as the unique or specific or error variance.

- The relationship (coefficient) between the common factor (l) and a random variable (x_j) is called *pattern loading*.

- The simple correlation between any variable (x_j) and the factor (l) is called *structure loadings*. So, pattern loadings and structure loadings are usually the same.

- The square of the structure loadings is referred as the *shared variance* between the variable and the factor.

- The correlation between any two indicators/variables is given by the product of their respective pattern loadings.

2.11.3 Estimation of Factor Loadings

Principal Components Method

There are various approaches available that can be used to estimate the factor loadings and its communalities. Principal component factoring method (PCFM) is one of a widely used technique in exploratory factor analysis for extracting the factors. PCFM yields equal eights to all measured variables and seeks a linear combination of the variables such that the maximum possible variance is extracted from the variables. It then removes this variance and seeks a second linear combination which explains the maximum proportion of the remaining variance, and so on. Successive factoring continuing until there is no further meaningful variance left. The factors obtained by this method are orthogonal (uncorrelated). PCFM analyzes total variance explained by a factor as a combination of common and unique variances.

Recall our factor model defined on equation 2.38 as

$$\tilde{\mathbf{X}} = \mathbf{\Lambda L} + \boldsymbol{\varepsilon} + \boldsymbol{\mu} \tag{2.39}$$

In practice, our goal is to estimate $\hat{\Lambda}$ of the factor loadings Λ and estimates $\hat{\Psi}$ of the specific variances Ψ. The PCFM decomposes the correlation matrix $\mathbf{R}$ or the sample covariance matrix $\mathbf{S}$ of the random variables $\tilde{X}$. In this study we used *phi* correlation matrix for the decomposition and estimation.

In the above model, we assume that the factors are not correlated with the unique factors (error components), and the means and variances of the random variables and factors are zero and one, respectively. Based on these assumptions, the correlation matrix $\mathbf{R}$ of the variables can be obtained as

$$\begin{aligned} E(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T) &= E[(\mathbf{\Lambda L} + \boldsymbol{\varepsilon} + \boldsymbol{\mu})(\mathbf{\Lambda L} + \boldsymbol{\varepsilon} + \boldsymbol{\mu})^T] \\ &= E[(\mathbf{\Lambda L} + \boldsymbol{\varepsilon} + \boldsymbol{\mu})(\mathbf{\Lambda}^T\mathbf{L}^T + \boldsymbol{\varepsilon}^T + \boldsymbol{\mu}^T)] \\ &= E(\mathbf{\Lambda L L}^T\mathbf{\Lambda}^T) + E(\boldsymbol{\varepsilon\varepsilon}^T) + E(\boldsymbol{\mu\mu}^T) \\ => \mathbf{R} &= \mathbf{\Lambda\Phi\Lambda}^T + \mathbf{\Psi} \end{aligned} \tag{2.40}$$

where $\mathbf{\Phi}$ is the correlation matrix of the factors, and $\mathbf{\Psi}$ is the diagonal matrix containing the unique variances. The communalities can be estimated from the diagonal of the $\mathbf{R} - \mathbf{\Psi}$ matrix. In a factor model $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are considered as model parameters. The goal of our factor analysis is to estimate these parameter matrices by using the correlation matrix.

For an orthogonal factor model, $\mathbf{\Phi} = 0$, and hence equation 2.40 becomes

$$\mathbf{R} = \mathbf{\Lambda}\Lambda^T + \mathbf{\Psi} \tag{2.41}$$

and with the estimated parameter values the equation becomes

$$\mathbf{R} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}} \tag{2.42}$$

In the PCFM estimation method, we ignore $\hat{\Psi}$ and factor $\mathbf{R}$ into $\mathbf{R} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T$.

Suppose, the eigenvalues of the $\mathbf{R}$ matrix are $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \cdots, \zeta_p)$ with the corresponding eigenvectors $\mathbf{\Gamma} = (\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \cdots, \boldsymbol{\gamma_p})$.

Hence from equation 2.42 the jth unique variance can be obtained as

$$\hat{\psi_j} = r_{jj} - \sum_{k=1}^{g} \hat{\lambda_{jk}}^2 \tag{2.43}$$

and the jth communality is the sum of squares of the jth row of $\hat{\boldsymbol{\lambda}}$ and can be represented as

$$\hat{h_j}^2 = \sum_{k=1}^{g} \hat{\lambda_{jk}}^2 \tag{2.44}$$

The kth eigenvalue of $\mathbf{R}$ is the sum of squares of the kth column of $\hat{\mathbf{\Lambda}}$ as $\sum_{j=1}^{p} \hat{\lambda_{jk}}^2$.

The variance of the jth variable is partitioned into two component parts as:

$$\begin{aligned} r_{jj} &= \text{communality} + \text{unique variance} \\ &= \hat{h_j}^2 + \hat{\psi_j} \\ &= (\hat{\lambda_{j1}}^2 + \hat{\lambda_{j2}}^2 + \cdots + \hat{\lambda_{jg}}^2) + \hat{\psi_j} \end{aligned} \tag{2.45}$$

So the kth factor contributes $\hat{\lambda_{jk}}^2$ to r_{jj} and the contribution to the total sample variance , $tr(R) = r_{11} + r_{22} + \cdots + r_{pp}$ is,

$$\text{Variance accounted by } k\text{th factor} = \sum_{j=1}^{p} \hat{\lambda_{jk}}^2 = \hat{\lambda_{1k}}^2 + \hat{\lambda_{2k}}^2 + \cdots + \hat{\lambda_{pk}}^2 \tag{2.46}$$

which is the sum of squares of factor loadings in the kth column of $\hat{\boldsymbol{\lambda}}$ and this is equal to the kth eigenvalue, ζ_k.

Therefore the proportion of total sample variance accounted by the kth factor is estimated as,

$$\frac{\sum_{j=1}^{p} \hat{\lambda_{jk}}^2}{tr(\mathbf{R})} = \frac{\zeta_k}{p} \tag{2.47}$$

where p is the number of variables.

2.11.4 Deciding the Number of Factors

A number of approaches have used here for determining the appropriate number of factors or components to retain. The proposed criteria are:

2.11.4.1 Kaiser's eigenvalue-greater-than-one rule

The eigenvalue-greater-than-one rule proposed by Kaiser (1960) is often used by researcher in practice. According to this rule, only the factors that have eigenvalues greater than one are retained for interpretation. However, many studies have criticized on this approach and argued that this method overestimate the correct number of factors when taking decision on it [33].

2.11.4.2 Cattell's Scree test

Another popular method proposed by Cattell (1966) is based on the scree test, which is a graphical representation of the eigenvalues. According to this method, the eigenvalues are plotted in descending order and then joined with a line. Later, the plot is examined to determine the point at which the graph reflects a significant drop or break,i.e. the point where the line cuts off or have much smaller slope. The scree plot is a two dimensional plot in where factors represents on the x-axis and eigenvalues on the y-axis. The idea behind this rule is that this cut point distinguish the major and important factors from the other minor factors and one should select the factors before this point. Hence, by this method for choosing the number of factors requires a kind of subjective judgment.

2.11.4.3 Variance Criterion

This method consider the proportion of variance accounted for retaining a factor. Therefore, the maximum number of factors (g) is achieved if the proportion of variance accounts for a predetermined percentage level, say (75%), of the total variation ($tr(\mathbf{R})$) [34].

2.11.4.4 Parallel Analysis

Horn (1965) [35] suggested a method for determining the number of factors based on Monte Carlo simulation technique which is known as *Parallel analysis* method. The idea behind this method is to generate a number of correlation matrices of random variables considering the same number of variables and sample size as the original data set have. Then the average of the eigenvalues from these various random correlation matrices are compared with the eigenvalues estimated from the original data correlation matrix. The comparison is made on the basis that the first real data eigenvalue is compared with the first randomly generated eigenvalue, the second eigenvalue is compared with the second randomly generated eigenvalue,

and so on. Factors that have original eigenvalues greater than the randomly generated parallel average eigenvalues are retained and the real data eigenvalues which are less than or equal to the parallel average random eigenvalues are considered as sampling errors [36].

2.11.4.5 Root Mean Square Error Residuals

We can use the residual matrix to examine and estimate the square root of the average squared values of the off-diagonal elements. This measure is known as the root mean square error residual (RMSER) and has been suggested as a factor retention criterion. The RMSER of the rasidual matrix can be define as

$$RMSER = \sqrt{\frac{\sum_{j=1}^{p}\sum_{h=1}^{p} res_{jh}^2}{p(p-1)/2}} \tag{2.48}$$

This measure estimates the differences between the factor model and the data per degree of freedom for the model. The value of $RMSER$ suggested to be small for obtaining a good factor structure. A value less than 0.05 can be considered as a good fit and values greater than 0.10 considered as poor fit [34].

2.12 Significant Factor Loadings

For a meaningful interpretation of a factor, a decision is needed regarding factor loadings which are considered to be significant. In practice, the decision can be yielded by taking into account on various criterion including the number of variables, that makes the interpretation of factor loadings.

At a first glance, the most frequently used rule of thumb was given by Hair et al (1998) [1] on making a preliminary investigation of the factor loading matrix. The rule is not based on any mathematical or statistical proposition, instead it relates more to assess practical significance. Ideally, a significant factor loading defines the responses of the variables that are influenced by the underlying construct. The rule suggested various cut-offs of factor loadings going from ± 30 to ± 75, respective of sample size. Table2.6 shows that the higher loadings are required when the sample size is small and needs a minimum sample size of 350 to achieve the minimal level of significance (0.30) for variables. The authors also suggested that the large absolute value of the factor loadings are considered to more practically significant and important in interpreting the loading matrix.

Field (2009) [37] suggested another guidelines for choosing significant loadings from a loading matrix. He advocates that without considering the sample size, if a factor has four or more loadings of at least 0.6 is considered as reliable.

Table 2.6: Hair et al (1998) [1] suggested thresholds for significant factor loadings respective of sample size.

Sample Size	Thresholds for Significant Factor Loading
50	0.75
60	0.70
70	0.65
85	0.60
100	0.55
120	0.50
150	0.45
200	0.40
250	0.35
350	0.30

Based on these above guidelines and considering the sample size of 516 for this study, we have considered lower loadings of ± 0.30 as a significance level in our factor loading interpretation.

2.13 Factor Rotation

After obtaining the initial factor solutions, one is interested to rotate the loadings. The goal of factor rotation is to find simpler factor structure that can use to make interpretation of the resulting factors easily and meaningfully and to determine the appropriate number of factors. It is a way of maximizing high loadings and minimizing low loadings so that the simplest possible structure is achieved. To accomplish this and for the simplicity of this study, orthogonal rotations are done using the varimax procedure. Here we only focus on the orthogonally rotated solutions as they can produce more simplified factor structures from a large amount of data. The methodology behind this technique has described below:

Varimax Orthogonal Rotation

In an orthogonal factor model it is assumed that the factors are uncorrelated, i.e. $\Phi = \mathbf{I}$. Orthogonal rotation approach involves to introduce a transformation matrix (orthogonal matrix) $\mathbf{G}$, which can use to estimate the new rotated loading matrix as $\hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}\mathbf{G}$ and $\mathbf{R} = \hat{\mathbf{\Lambda}}^* \hat{\mathbf{\Lambda}}^{*^T}$.

$\hat{\mathbf{\Lambda}}^*$ matrix holds the rotated pattern loadings in where the variance accounted for could be measure by the sum of squares of each column of $\hat{\mathbf{\Lambda}}^*$ and the communalities by the sum of squares of each row of it.

Varimax rotation proposed by Kaiser (1958) [38] is probably the most commonly used orthogonal rotation technique for obtaining the transformation matrix ($\mathbf{G}$). The objective of this rotation is to determine this transformation matrix in such a way that for any given factor there will have some variables that will load very highly on it and some variables will load low or very low on it. This is achieved by maximizing the sum of the variances of the

squared loadings $\hat{\lambda}_{jk}$ in each column of $\hat{\mathbf{\Lambda}}$ [1], subject to the constraint that the communality of each variable is unchanged.

Let the simplicity of a factor can be defined as the variance of its squared loadings as

$$\begin{aligned} V_s &= \frac{\sum_{j=1}^{p}(\lambda_{js}^2 - \lambda_{.s})^2}{p} \\ &= \frac{p\sum_{j=1}^{p}(\lambda_{js}^2)^2 - (\sum_{j=1}^{p}\lambda_{js}^2)^2}{p^2} \end{aligned} \tag{2.49}$$

where V_s is the variance of the communalities of the variables within factor s and $\lambda_{.s}$ is the average squared loadings for factor s.

Therefore, the total variance for all the factors can be obtained as

$$\begin{aligned} V &= \sum_{s=1}^{g} V_s \\ &= \sum_{s=1}^{g}\left(\frac{p\sum_{j=1}^{p}(\lambda_{js}^2)^2 - (\sum_{j=1}^{p}\lambda_{js}^2)^2}{p^2}\right) \end{aligned} \tag{2.50}$$

For fixed number of variables, maximizing equation 2.50 is equal to maximize

$$V = \sum_{s=1}^{g}\left[p\sum_{j=1}^{p}\left(\lambda_{js}^2\right)^2 - \Big(\sum_{j=1}^{p}\lambda_{js}^2\Big)^2\right]\Big/p^2 \tag{2.51}$$

Hence, the orthogonal matrix, $\mathbf{G}$, is obtained when equation 2.51 is maximized, based on the limitation that the communality of each variable remains the same. .

Data analysis was done by using *R* statistical software for *Cluster Analysis* and *SAS 9.2* (SAS Institute, Inc., Cary, North Carolina) for *Factor Analysis*.

[1] This mean, when the loadings in a column of the loading matrix $\hat{\mathbf{\Lambda}}$ are almost equal, then the variance will be close to 0. Since the squared loadings technique suggests 0 and 1, the variance will tend to approach a maximum. Thus the varimax rotation tries to make the loadings either large or small for simpler interpretation.

CHAPTER 3

Analysis

3.1 Demographic and Clinical Characteristics

Table A.1 on page 68 lists some demographic and clinical characteristics of the study gynecologic cancer patients who were diagnosed and treated. The table shows that most of the patients were high school or vocational educated, old pensioner, married and living together, Swedish born, and came from big cities, such as, Stockholm, Göteborg or Malmö. The mean age of the survivors at the time of questionnaire is 64 years (SD= 10.5) with a range 28 to 79 years. From the table it is observed that Endometrial cancer-$C54.9$ is the most common (about 59%) gynecologic cancer among the study cancer survivors in Sweden. Also the Medical journal from where the patients were selected reports that very few of the patients received radiation therapy (RT) alone as for their cancer treatments. Almost all of the study patients received radiotherapy as a combination with other treatment. Among the treatment combinations, the combination is high (56%) for the cohort who received radiation therapy (RT), Operation (Op) and Brachytherapy (Br) altogether for their cancer treatment.

3.2 Frequency and Co-Occurrence of self-reported gastrointestinal symptoms

In the questionnaire the symptoms are measured by various ordinal scales. Survivors who are reported other than 1 of the scale, carry some kind of symptoms which is either severe or not. At a first glance, it could be interesting to look at the data with a distribution that represents the number of symptoms present and reported by each patient. For this reason we converted

our original categorical data to a binary data with "Yes" and "No" in where the term "Yes" represents the presence of symptoms reported at least 1 or more of the original scale by the participants and the term "No" represents the absence of symptom reported 0 of the original scale by the participants. The reason for this dichotomization is to make the different uneven categories for various questions to a equal scale for comparison and interpretation whether the survivors are experienced in a symptom or not. From this consideration we can observe that respondents of the cancer survivors contributed on an average of 13.4 (standard deviation 7.5) gastrointestinal symptoms with a median value of 13. Summary statistics of the distribution of self-reported symptoms for 516 survivors are presented on Table 3.1.

Table 3.1: Summary of number of self-reported gastrointestinal symptoms of the questionnaire.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	8	13	13.4	19	36

Figure 3.1 present a bar-plot of this distribution in where the horizontal axes represents the number of symptoms and the vertical axes represents the count of participants reported for each number.

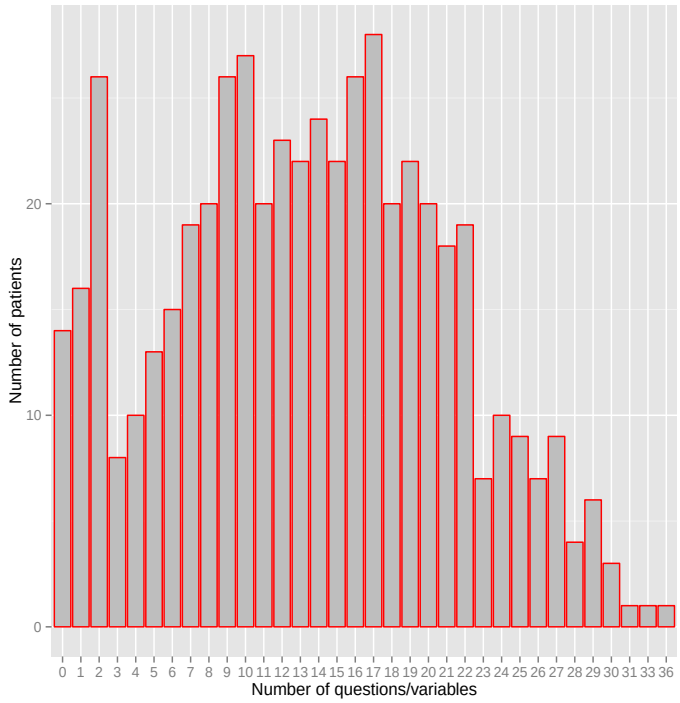

Figure 3.1: Barplot for Number of questions answered other than "no" in terms of the number of patients.

The figure exhibits that most of the patients seem to experience at least some degree of gastrointestinal symptoms, with no clear separation between a low-frequency and high-frequency symptom group. Of the 516 cancer survivors, 14 survivors (2.7 %) reported "no" symptom

at all and about 279 (54%) survivors reported they have experienced with thirteen or more concurrent symptoms. From the bar-plot on Figure 3.1 and the table on Table 3.1, it has been clear that 75% of the binary data is scattered within $0 - 19$ number of symptoms. Besides, a few (approx. 22%) survivors have reported that they received $20 - 37$ symptoms for their treatment. Moreover, there are only six patients out of 516 who have experienced more than thirty symptoms.

From the dichotomized data we construct the Table A.3 on page 70 which describes the frequency for each symptom reported by the cancer survivors, their percentage of occurrence, median, mean of original scores with corresponding standard deviations. Based on this table, the five most frequently reported gastrointestinal symptoms by the cancer survivors are N38:Loose stools, N73:Defecation urgency, N53:Returned to bathroom within an hour, N75:Ability to hold feces and N84:Involuntary flatulence. Symptoms' mean of original scores ranges from 3.08 for N38: loose stools to 1.01 for N96: leakage of blood when asleep. The five most symptoms irritating scores reported by the cancer survivors are N38: loose stools, N75: ability to hold feces, N73: defecation urgency, N138: abdominal pain intensity and N53: returned to bathroom within an hour.

3.3 Cluster Analysis

The hierarchical cluster analysis assign our multiple self-reported symptoms into distinct groups or clusters. The survey data used to analyze and identify significant groups among the questions that may define a clinically important (but previously unknown) relationship, but this is speculative. We used agglomerative hierarchical clustering to identify clusters and also used our hypothesized clinical clustering to validate the clustering results with adjusted Rand index method.

In this study our objective is to evaluate a clustering structure exploratively with many distance methods that can give us a clustering results which is close enough to the hypothesized clinical clustering. For deciding and choosing the closest distance method we used adjusted Rand index for our two clustering agreement measurement.

Figure 3.2 depicts the clustering agreements between hypothetical clinical clusterings and our empirical clusterings concerning with *phi* correlation as distance measure and Ward linkage as clustering algorithm. The horizontal axis represents the number of clusters and the vertical axis represents the corresponding adjusted rand index value calculated from two defined partitions. The value of the line starts with zero for one cluster and end with zero again for 37 clusters. The figure exhibits that the value of adjusted Rand index increases (i.e. the agreement between hypothesized and empirical clusterings become closer) as the number of clusters for the empirical clustering increases, and the agreemented adjusted rand index get its pick (0.73) when the number of clusters for the empirical cluster is nine. Afterwards, the line gradually goes down as the number of clusters increases more and when there are 37

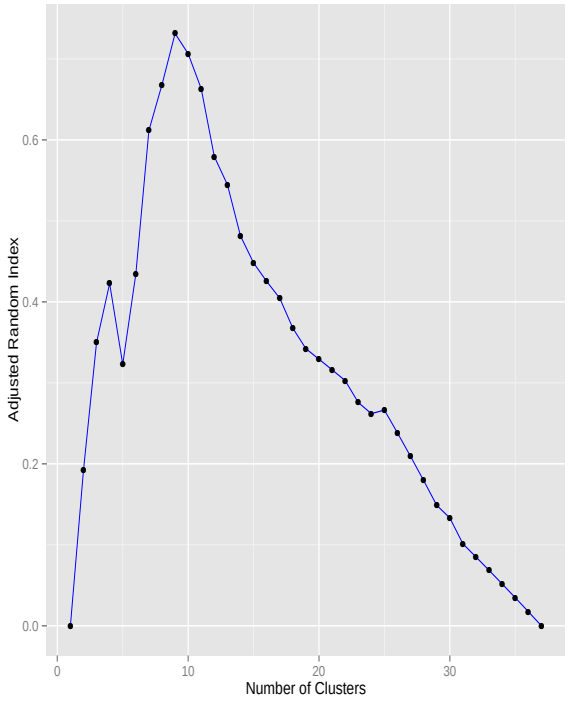

Figure 3.2: Agreement between clinical clustering and empirical ward clustering with *phi* correlation distance for binary data.

clusters for the empirical clusters (i.e., each element is considered as a single cluster under study) then the agreement with adjusted Rand index between the two partitions (empirical vs hypothetical) becomes zero.

This study is particularly involved in comparing various distance measures that can fit our survey data precisely and can classify symptoms into hierarchical groups in compare to the clinical groupings. Table A.4 on page 71 represents a number of combinations of different methods we used and to compare them with our clinical clusterings in order to get some feel of stability of the results under various clustering algorithms. Each of these methods has its' own objectives to use them in any scientific research problem. For example, for the binary data, the Jaccard's coefficient excludes elements that are both zero or absent where as the phi coefficient is the binary equivalent of the Pearson correlation coefficient and use all present and absent elements information. Most combinations of clustering algorithms and distance metrics showing meaningful and high values with agreement to the clinical clustering, while a few combinations showing poor agreements. The calculated adjusted Rand index for various approaches is presented at the last column of the table. From our mathematical and statistical point of view, the method for which the adjusted Rand index obtained its' highest value, represents the method which can make our clusterings closer to the clinical clusterings. We can observe that for our dichotomize (binary) data, the use of *phi* correlation as similarity measure among treatment effects and the use of Ward clustering algorithm provides maximum adjusted Rand index value 0.73 with nine clusters. Therefore we consider this clustering method as our suggested method for this particular validation study and will

continue our further research and cluster interpretations with this distance method which are also one of the main interesting parts of this study.

From Table A.4 on page 71 we can also observe that there are some distance methods for which complete linkage algorithm process well, for some methods average linkage showing good results and also for some distance methods Ward method provides good clusterings compare to others. It is also visible that none of the distance method perform well with the single linkage clustering algorithm.

3.3.1 Correlation Matrix

The *phi* correlation distance measure gives the clustering results which is closest to our hypothetical clusterings. Figure 3.3 presents a heat-map of the selected correlation matrix in different colors. The correlation plot indicates the strength of correlation with color-coded squares, so that more highly positively correlated treatment effects are appeared with darker green color and negatively correlated treatment effects are presented with darker red color . The figure also reveals that most of the correlations among the self-reported symptoms lie within the range $(0, 0.25)$.

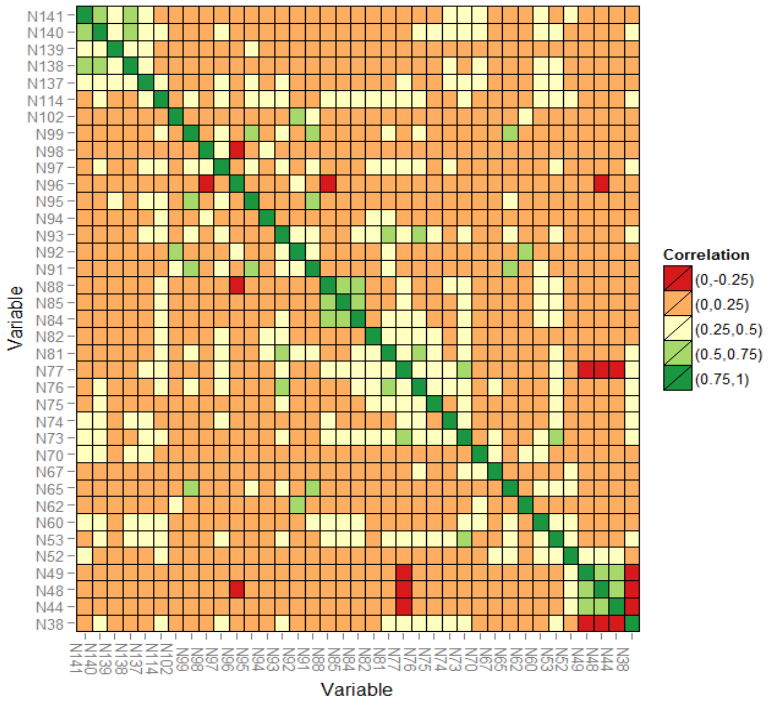

Figure 3.3: Heat-map for *phi* correlation matrix

The results of the clusterings can also be visualized using a Matrix Tree plot. Figure 3.4 shows another color coded heat-map for the correlation matrix as before but with a dendrogram added to the left side on it. In the figure the dark green color represents high positive correlation and dark red color represents negative correlation among the various symptoms.

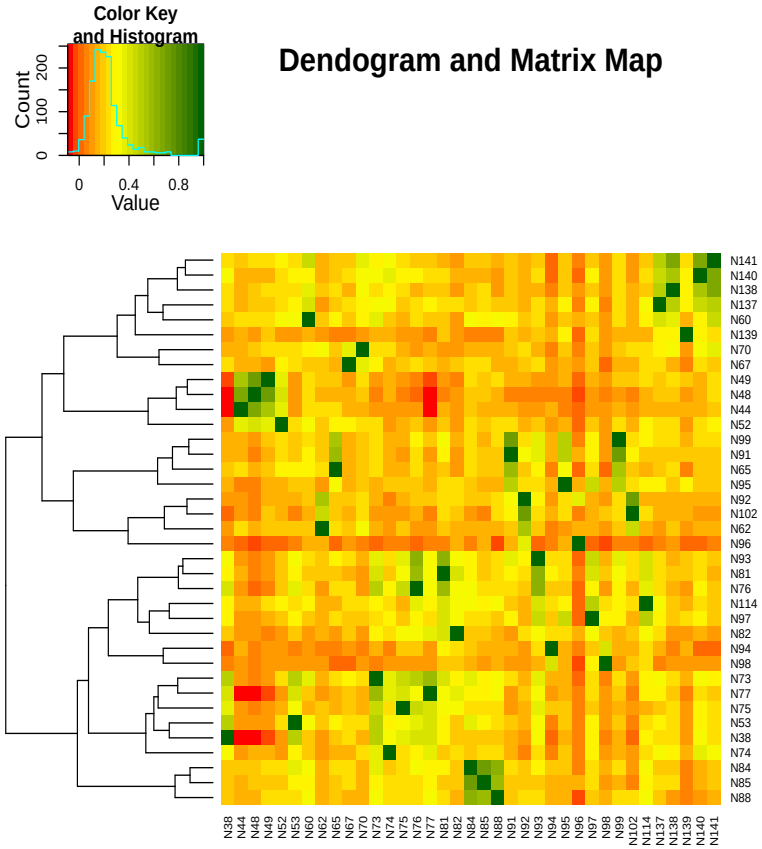

Figure 3.4: Heat-map for correlation matrix with dendrogram

From Figure 3.3 and Figure 3.4, some negative relations among symptoms are visible. It has been observed that the symptoms, N77:Immediate need of toilet and N38:Loose stools, are negatively associated with the symptoms N44:Hard stools, N48:Hard to push out feces and N49:Ability to push out feces. Besides, rare symptom N96:Leakage of blood when asleep, is negatively associated with the other symptoms N48:Hard to push out feces, N88:Foul-smelling flatulence and N98:Leakage of solid stools when asleep. Although the values of these correlations are small, they could be interesting for further research.

3.3.2 Clusters of Self-reported Symptoms

The results of our hierarchical clustering by phi correlation distance method can be visualized graphically with a dendrogram, a tree structure of the symptoms and presents which elements are group together or clustered to each other. The dendrogram in Figure 3.5 suggests nine empirical clusters of symptoms in comparison with the clinical hypothetical clusterings. N84:"Involuntary flatulence", N85:"Loud flatulence" and N88:"Foul-smelling flatulence" symptoms are clustered into a cluster that we have named *flatulence*. N38: "Loose stools", N73: "Defecation urgency", N75:"Ability to hold feces", N74:"Stomach ache with bowel move-

ments" and N77:"Immediate need of toilet" symptoms are merged into a cluster which have been named as *loose stools and defecation urgency*. N94:"Leakage of solid stools when awake" and N98:"Leakage of solid stools when asleep" symptoms form a cluster which have been named *leakage of solid stools*. N76:"Defecation urgency with fecal leakage", N81:"Leakage of stool without forewarning despite previous defecation", N82:"Leakage of all stool into clothing without forewarning", N93:"Leakage of loose stools when awake", N97:"Leakage of loose stools when asleep" and N114:"Smells of feces" symptoms are merged into a cluster called *leakage of loose stools*. N91:"Leakage of mucus when awake", N95:"Leakage of mucus when asleep", N99:"Soiled clothing due to leakage of mucus" and N65:"Mucus in stool" symptoms are merged into a single cluster which have been referred as *leakage of mucus*. N92:"Leakage of blood when awake", N96:"Leakage of blood when asleep", N102:"Soiled clothing due to leakage of blood" and N62:"Rectal bleeding" symptoms are merged into a cluster which we call *leakage of blood and rectal bleeding*. N44:"Hard stools", N52:"Incomplete bowel emptying", N48:"Hard to push out feces" and N49:"Ability to push out feces" symptoms are merged into a single cluster which we have named as *hard stools*. N67:"Anal itching" and N70:"Anal pain" symptoms are merged into cluster which we refer to *anal pain*, and the last cluster of the dendrogram that we have named as *abdominal pain and bloating*, have involved with the symptoms N141:"Abdominal pain with bloating", N137:"Abdominal pain", N138:"Abdominal pain intensity", N139:"Abdominal pain with vomiting", N140:"Abdominal pain with defecation", and N60:"Abdominal bloating". The naming of the clusters are done here subjectively and from clinical perspectives.

The dendrogram divides the whole set of symptoms into two different parts. On the right part of the figure is holding the groups: Leakage of mucus, Leakage of blood, Hard stools, Anal pain, and Abdominal pain. On the other hand, the left part of the dendrogram is holding the groups: Flatulance, Urgency leakage, Leakage of solid stools, and Leakage of loose stools. The groups within a part are closer to each other than the groups on the other part. The dendrogram roughly agrees with our hypothetical clustering.

Table 3.2 shows the number of clusters with their sizes, cluster members and name of different pathophysiologies of the 37 gastrointestinal self-reported symptoms.

Table 3.2: Table of nine clusters with their pathophysiology names and cluster members

Cluster No.	**Cluster Name**	**Cluster Size**	**Variable Names**
1	Loose stools and defecation urgency	6	N74, N38, N53, N75, N73, N77
2	Hard stools	4	N52, N44, N48, N49
3	Abdominal pain and bloating	6	N139, N138, N140, N141, N60, N137
4	Leakage of blood and rectal bleeding	4	N96, N62, N92, N102
5	Leakage of mucus	4	N95, N65, N91, N99
6	Anal pain	2	N67, N70
7	Leakage of loose stools	6	N76, N81, N93, N82, N97, N114
8	Flatulence	3	N88, N84, N85
9	Leakage of solid stools	2	N94, N98
Total		**37**	

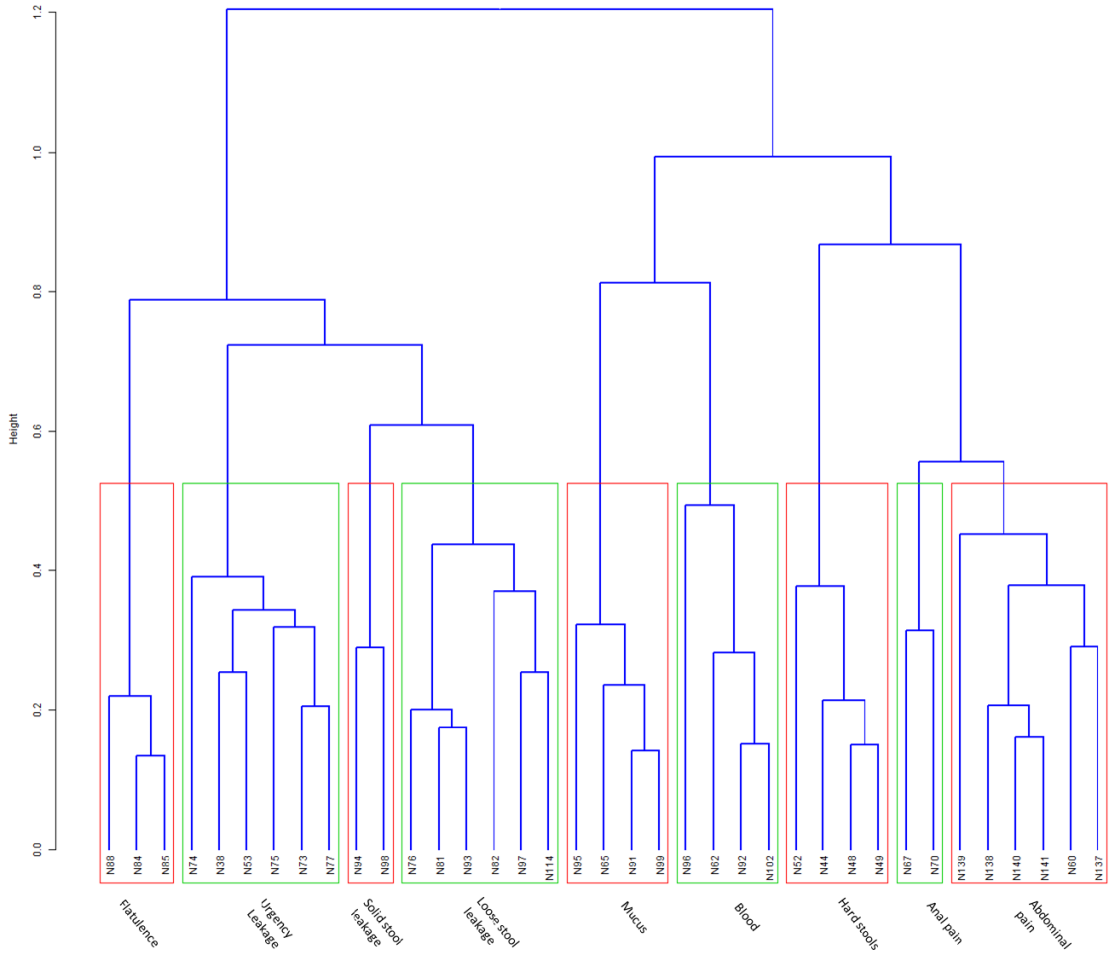

Figure 3.5: Clustering dendograms for Pearson phi method with Ward Clustering algorithm

3.3.3 Comparison between clinical and empirical clusterings

It is interesting and one of our main objective is to compare our empirical clusterings with our formulated hypothetical clinical clusterings. When we make the comparison between Table 2.1 on page 11 and Table 3.2 on the previous page, they actually seem not too far from each other. This is not strange or unexpected as the criterion for this clustering was that the empirical should be similar to the clinical clustering. We could observe that the four clusters "Leakage of solid stools", "Hard stools", "Anal pain" and "Leakage of loose stools and urgency" are separated same way in both clustering. N38:"Loose Stools", which was an separate group on the hypothetical clustering, was merged with the group "Defecation urgency" on the empirical clustering. "Abdominal pain and flatulence" separated to two clusters in the empirical clustering, but they were within a big cluster referred as abdominal group in the hypothetical clustering. In hypothetical clustering, one of the big groups "Leakage of blood and mucus", become well separated in our empirical clustering and turn to two distinct clusters "Leakage of blood and rectal bleeding" and "Leakage of mucus".

3.4 Exploratory Factor Analysis

Based on the interest to exploratively identify a factor structure of the symptoms of cancer survivors, we next conducted a exploratory factor analysis. The principal objective of this idea is to group together all those symptoms which are highly correlated with each other and extract factors representing the consequences. The factor analysis executed in this study is based upon analyzing the *phi* correlation matrix obtained from the earlier section of cluster analysis.

The analysis consists of 37 cancer survivors self-reported symptoms with 516 observations. The *phi* correlation matrix was analyzed using **SAS** with principal components factorization method (PCFM). Results from the factor analysis shows that the Kaiser's measure for overall sampling adequacy is equal to 0.8699 which is higher than the proposed level 0.80 by kaiser and Rice [39]. This high value suggests that the correlation matrix (*phi* correlation) is appropriate for factoring.

The determination of the number of factors needed to explain the correlation among symptoms is heuristic. However, in this study we have obtained this by several approaches: table of eigenvalues, proportion of variance explained by eigenvalues, scree plot and parallel analysis plot.

Table 3.3 shows the result of first 11 largest eigenvalues computed from the sample correlation matrix and their corresponding proportion of variances explained. The first eigenvalue to account for the largest proportion of the total variance (23%), the second eigenvalue to account for the second largest proportion of the remaining variance (7%), and so on. The table also suggests that 10 factors have eigenvalue greater than one and all these factors to account for 65% of the total variance.

Table 3.3: Eigenvalues of the Correlation Matrix: Total = 37 Average = 1

No.	Eigenvalue	Difference	Proportion(variance)	Cumulative
1	8.73405489	5.95286856	0.2361	0.2361
2	2.78118633	0.47354291	0.0752	0.3112
3	2.30764342	0.23822073	0.0624	0.3736
4	2.06942269	0.32125269	0.0559	0.4295
5	1.74816999	0.06610791	0.0472	0.4768
6	1.68206209	0.27061031	0.0455	0.5222
7	1.41145178	0.29056911	0.0381	0.5604
8	1.12088267	0.07120744	0.0303	0.5907
9	1.04967522	0.04518012	0.0284	0.619
10	1.00449511	0.03665996	0.0271	0.6462
11	0.96783515	0.07114527	0.0262	0.6723

*10 factors will be retained by the MINEIGEN criterion.

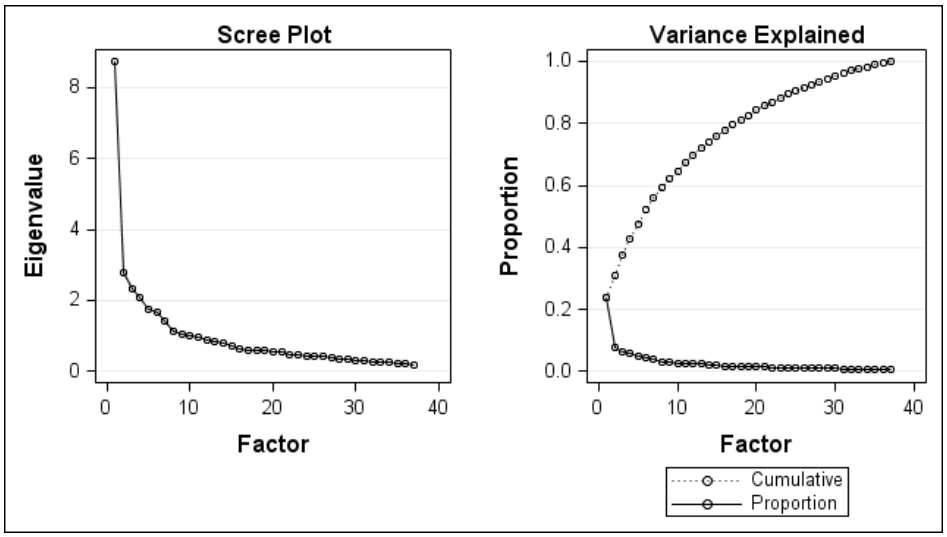

Figure 3.6: Scree plot for the eigenvalues of the correlation matrix

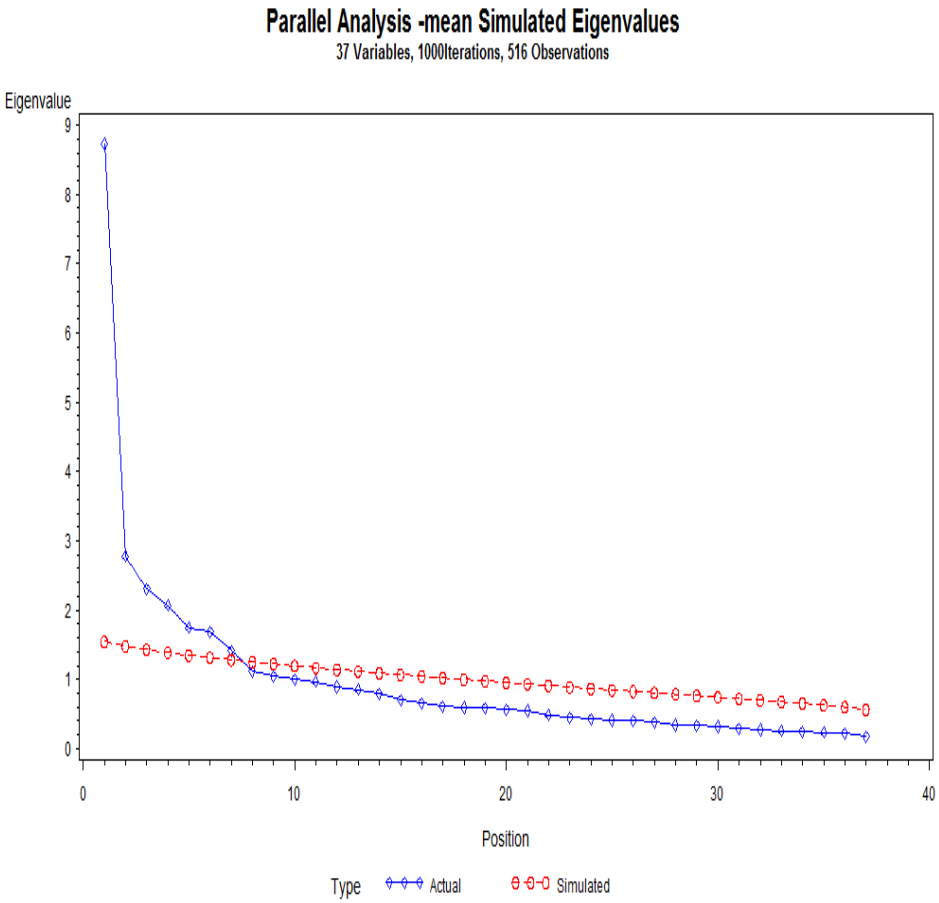

Figure 3.7: Parallel analysis plot

The scree plot and the parallel analysis plot of the correlation matrix are given in Figure 3.6 and Figure 3.7, respectively which search for significant break on the plot. These plots with an elbow in the seven largest eigenvalue indicates that a 7 factors should be retained.

Although the Kaiser-Guttmann's criterion (eigenvalue-greater-than-one rule) suggests that 10 factors should be retained [Table 3.3]. However, these procedures are arbitrary and has been criticized by many researchers [34]. The seven retained factors to account for 56.04% of the total variance in the data and we therefore suggest a seven factors solution are acceptable for this data.

A structure matrix of pattern loadings from factor analysis can be obtained in a $p \times g$ form, which shows the correlations between the symptoms and their latent factors and where p is the number of symptoms and g is the number of factors retained . Table A.5 on page 72 shows the initial unrotated factor structure matrix, which consists of the correlations between the 37 symptoms and the seven retained factors.

For this study we assumed, the underlying factors for measuring the treatment-related symptoms are orthogonal (uncorrelated) and performed varimax orthogonal rotation technique for obtaining interpretability factor solutions. The pattern loadings after rotation are reported on Table A.6 on page 73, leadings to meaningful factor interpretations.

Figure 3.8: Figure for initial and rotated pattern loadings in terms of first two factors.

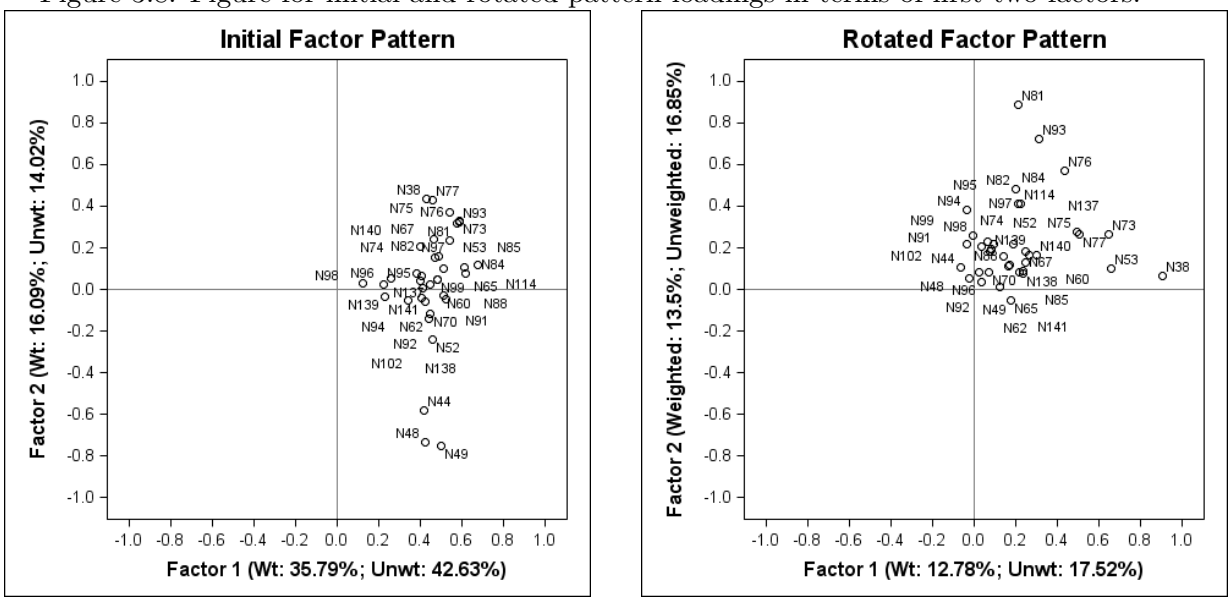

Figure 3.8 exhibits an example of pattern loadings in terms of first two factors before and after rotation. Rotation technique makes the factor structure more interpretable by increasing the number of pattern loadings close to 1, 0 or -1.

The overall root mean square error residuals (RMSER) is equal to 0.047. The value is considered to be small and hence indicating that the factor solution is good.

3.4.0.1 Interpretation of Factors

In Table A.5 on page 72 and Table A.6 on page 73, the significant loadings are marked with a symbol "*". For identifying the significant loadings on factors we used the suggested threshold, loadings greater than the absolute value of 0.30. The extracted seven initial factors can be described as follows in where the arrangement of symptoms are considered on the basis of their highest loadings to a factor.

- **Factor 1** is identified with the subsets of all the symptoms except symptoms N96:Leakage of blood when asleep, and N98:Leakage of solid stools when asleep. In this factor all significant loadings have high positive correlations and can be considered as *general factor* of the symptoms.
- **Factor 2** consists of symptoms N48:Had to push out feces, N49:Ability to push out feces, N44:Hard stools, and N52:Incomplete bowel emptying, which have high positive loadings, respectively, and with symptoms N77:Immediate need of toilet, N38: Loose stools, N76: Defecation urgency with fecal leakage, N73:Defecation urgency, and N93:Leakage of loose stools when awake, which have high negative loadings, respectively.
- **Factor 3** is identified with the subset of symptoms N102:Soiled clothing due to leakage of blood, N92:Leakage of blood when awake, N91:Leakage of mucus when awake, N99:Soiled clothing due to leakage of mucus, and N95:Leakage of mucus when asleep, which have high positive loadings, respectively, and with symptoms N141:Abdominal pain with bloating, N140:Abdominal pain with defecation, N138:Abdominal pain intensity, and N137:Abdominal pain, where they have high negative loadings, respectively.
- **Factor 4** consists of symptoms N84:Involuntary flatulence, N85:Loud flatulence, N88:Foul-smelling flatulence, N48:Had to push out feces, N44:Hard stools, and N49: Ability to push out feces, where they have high positive loadings, respectively, and with symptoms N140:Abdominal pain with defecation, N138:Abdominal pain intensity, N95: Leakage of mucus when asleep, N141:Abdominal pain with bloating, N91:Leakage of mucus when awake, and N99:Soiled clothing due to leakage of mucus, which have high negative loadings, respectively.
- **Factor 5** consists of the subset of symptoms N92:Leakage of blood when awake, N62:Rectal bleeding, N96:Leakage of blood when asleep, and N102:Soiled clothing due to leakage of blood, where they have high positive loadings, respectively, and with symptoms N97:Leakage of loose stools when asleep, and N94: Leakage of solid stools when awake, where they have high negative loadings, respectively.
- **Factor 6** is identified with symptoms N82:Leakage of all stool into clothing without forewarning, N98:Leakage of solid stools when asleep, and N102:Soiled clothing due to leakage of blood, which have high positive loadings, respectively, and with symptoms N65:Anal leakage of mucus, N91:Leakage of mucus when awake, N99:Soiled clothing

due to leakage of mucus, and N84:Involuntary flatulence, which have high negative loadings, respectively.

- **Factor 7** is identified with symptoms N98:Leakage of solid stools when asleep, N85:Loud flatulence, and N94:Leakage of solid stools when awake, which have high positive loadings, respectively, and with symptom N76:Defecation urgency with fecal leakage, where it has high negative loading.

From the complex interrelationships outlined above in factors, it seems complicated to interpret them by labeling. However, by following the varimax rotation the loadings have changed quite a lot to a simpler structure and is possible to obtain meaningful interpretation about the factors. Therefore, the rotated factors can be labeled as follows in where the arrangement of symptoms are considered on the basis of their highest loadings to a factor.

- **Factor 1** is identified with the subsets of symptoms N76:Defecation urgency with fecal leakage, N73:Defecation urgency, N38:Loose stools, N93:Leakage of loose stools when awake, N75:Ability to hold feces, N77:Immediate need of toilet, N81:Leakage of stool without forewarning despite previous defecation, N53:Returned to bathroom within an hour, N82:Leakage of all stool into clothing without forewarning, N74:Stomach ache with bowel movements, N67:Anal itching, and N114:Smells of feces. Therefore, this factor may be named "Loose stools and defecation urgency".

- **Factor 2** consists with symptoms N141: Abdominal pain with bloating, N138:Abdominal pain intensity, N140:Abdominal pain with defecation, N137:Abdominal pain, N60:Abdominal bloating, N70:Anal pain, N74:Stomach ache with bowel movements, and N139:Abdominal pain with vomiting. This factor may be labeled "Abdominal pain".

- **Factor 3** is identified with the subset of symptoms N91: Leakage of mucus when awake, N99:Soiled clothing due to leakage of mucus, N65:Anal leakage of mucus, N95:Leakage of mucus when asleep, N93:Leakage of loose stools when awake, and N97:Leakage of loose stools when asleep. This factor may be named "Leakage of mucus" .

- **Factor 4** consists with symptoms N48:Had to push out feces, N44:Hard stools, N49: Ability to push out feces, and N52:Incomplete bowel emptying. This factor may be named "Hard stools" .

- **Factor 5** is consisted with the subset of symptoms N85:Loud flatulence, N84: Involuntary flatulence, N88:Foul-smelling flatulence, and N60:Abdominal bloating. This factor may be named "Flatulence" .

- **Factor 6** is identified with symptoms N94:Leakage of solid stools when awake, N98:Leakage of solid stools when asleep, N97:Leakage of loose stools when asleep, N114:Smells of feces, N81:Leakage of stool without forewarning despite previous defecation, N82:Leakage of all stool into clothing without forewarning, N93:Leakage of loose stools when awake, N139:Abdominal pain with vomiting, and N95:Leakage of mucus when asleep. This factor is very tricky to categorize since it seems to load on variables all over the place,

but the factor does seem to capture severe leakage symptoms like leakage of solid stools and leakage when asleep. Therefore, the factor can be named as "Severe leakage".

- **Factor 7** is identified with symptoms N92:Leakage of blood when awake, N102:Soiled clothing due to leakage of blood, N62:Rectal bleeding, N96:Leakage of blood when asleep, and N70:Anal pain, . This factor may be named "Bleeding" .

CHAPTER 4

Discussion & Conclusion

4.1 Discussion

To our knowledge, this is the first study that investigate gastrointestinal symptom clusters on the survivors of Gynecologic cancer. Symptom cluster research is still in its early stage, and the statistical techniques used to analyze the data varies from study to study.

This study focuses on validating empirical clusterings based on clinical perspectives. As we seek for a clusterings that is closest to the hypothetical clinical clusterings, so, the clusterings empirically we obtained from the data (*phi* correlation and Ward minimum variance) indicate that a hypothetical clinical mechanism is associated to link the symptoms together within a cluster and separated the clusters from each. Discovering clinical mechanism of the symptoms is essentially important for symptom cluster research because such behavior of the symptoms suggest that treatment for one symptom may be effective in treating all symptoms in the cluster [40].

We have used several dissimilarity measures and all of these methods have their own advantages and disadvantages. Euclidean distance applied in this study uses the actual ratings between two symptoms. As we have some rare cases and uneven scaling in some symptoms, this method affects the distance measure highly and gives very poor agreement in compare with the clinical clustering. On the other hand, for non-normal and ordinal data, we have used some nonparametric distance methods such as Kendall's tau, Spearman's footrule, and Goodman and Kruskal's gamma. In our data as we have many ties, Goodman and Kruskal's gamma measure explains better output results than the other two nonparametric methods, where Kendall's tau and Spearman's footrule are useful for a dataset with few ties in the observations.

Different linkage algorithms were implemented in combination with various distance measures in clustering. The study found that dichotomized data clustering with *phi* correlation matrix and ward minimum variance criterion gave the closest agreement with the clinical clustering, as measured by adjusted Rand index. The complete explanation for this behavior is unknown. Moreover, it could be informative to compare distance method and clustering algorithms for symptom clusters in cancer and non-cancer groups to observe the changes of the clusters differ by diagnosis. Such an investigation might provide a clearer explanation for relationships between *after treatment* and *current symptoms* in gynecological cancer survivors. Future studies also should examine the possible mechanisms underlying symptom clusters to better understand how and when symptoms interact.

As a data reduction technique, alternatively we also performed an exploratory factor analysis with principal component factoring method (PCFM) to construct scales of the symptoms. We have assumed that the underlying factors are uncorrelated in nature as we are primarily more interested in the generalizability of our results. These scales are also identified based on clinical perspectives as we have used *phi* correlation matrix for our factor extraction. Future studies should examine by considering non-orthogonal (correlated) factors and a confirmatory factor analysis technique to examine better symptom measurement tools for old adult gynecologic cancer survivors.

The study finds that varimax rotation transformed all the symptoms initial *negative* high loadings to *positive* loadings and that gave us a simple factor structure for meaningful interpretation. To us, the idea behind this mechanism is unclear.

Generally, factor analysis is sensitive to outlying cases and as the factor solutions are not unique, several solutions could be possible which can alter our factor interpretations too.

4.2 Limitations

The first limitation to this study is the use of cancer survivors' self-reported data that can be associated with reported bias and weakened correlations. To overcome this limitation, future studies of symptom clusters would be useful if the underlying biological mechanisms of symptom clusters are to be revealed in oncology.

The second limitation is that the sample for this study considered only women survivors who had been diagnosed and treated for their gynecological cancer in only Stockholm and Gothenburg. Participants are basically aging old women, pensioner and somewhat highly educated. Hence, the results of this study may not be generalizable to control women, to survivors with other types of cancer, or to other gynecological survivors living on other cities in Sweden.

Finally, in this study, only gastrointestinal section of the Gynecological cancer survivors' survey questionnaire was examined for clusters and factor analysis. Urinary, diet and other

chronic diseases are also common in Gynecologic Cancer survivors. They may affect the perceived level of symptom clustering on survivors' gastrointestinal function and can also alter our identified clusters or can create more additional clusters. Therefore, future research can be done for "clustering" with the combination of urinary and diet related symptoms. Research can also be performed by "Factor analysis" within each cluster i.e., how many underlying symptoms that have been measured with each cluster.

4.3 Conclusion

The aim of this study is to validate a study-specific questionnaire with a hypothesized clinical clustering. For the symptom cluster, we used hierarchical clustering and factor analysis, two most appealing statistical tools in oncology research. Among all gastrointestinal symptoms, "loose stools" is the most common symptom and has an high irritating scores on the scale reported by the cancer survivors. In our various searching results, the dichotomized data to cluster using *phi* distance and Ward's minimum variance method gives the most similar clusters in comparison to the hypothetical clustering for which we obtained highest value 0.73 of adjusted Rand index. From our exploratory and clinical clusterings, we have found that the four symptom clusters "Leakage of solid stools", "Hard stools", "Anal pain" and "Leakage of loose stools and urgency" are separated same way in both clusterings. The empirical results also show that there are two clusters merged to a single cluster and two separated as comparing to the clinical clustering. Alternatively, our exploratory factor analysis suggests to retain 7 factors and the rotated factors are labeled as: Loose stools and defecation urgency, Abdominal pain, Leakage of mucus, Hard stools, Flatulence, Severe leakage, and Bleeding. This study could provide a scientific basis and new directions for better symptom management strategies and intervention.

Bibliography

[1] J. F. Hair, R. E. Anderson, R. L. Tatham, and C. William, *Multivariate data analysis*. Upper Saddle River, NJ: Prentice Hall, ISBN:0130329290, 1998.

[2] "The National Board of Health and Welfare. Cancer Incidence in Sweden," 2008.

[3] G. Dunberger, H. Lind, G. Steineck, A.-C. Waldenström, T. Nyberg, M. Al-Abany, U. Nyberg, and E. Åvall-Lundqvist, "Self-reported symptoms of faecal incontinence among long-term gynaecological cancer survivors and population-based controls," *European Journal of Cancer*, vol. 46, no. 3, pp. 606–615, 2010.

[4] M. Dodd, S. Janson, N. Facione, J. Faucett, E. S. Froelicher, J. Humphreys, K. Lee, C. Miaskowski, K. Puntillo, S. Rankin, *et al.*, "Advancing the science of symptom management," *Journal of advanced nursing*, vol. 33, no. 5, pp. 668–676, 2001.

[5] National Cancer Institute, "symptom management." `http://www.cancer.gov/dictionary?cdrid=269453/`.

[6] C. Miaskowski, M. Dodd, and K. Lee, "Symptom clusters: the new frontier in symptom management research," *JNCI Monographs*, vol. 2004, no. 32, pp. 17–21, 2004.

[7] H.-J. Kim, D. B. McGuire, L. Tulman, and A. M. Barsevick, "Symptom clusters: concept analysis and clinical implications for cancer nursing," *Cancer nursing*, vol. 28, no. 4, pp. 270–282, 2005.

[8] A. M. Barsevick, "The elusive concept of the symptom cluster," in *Oncology Nursing Forum*, vol. 34, pp. 971–980, Onc Nurs Society, 2007.

[9] A. M. Barsevick, K. Whitmer, L. M. Nail, S. L. Beck, and W. N. Dudley, "Symptom cluster research: conceptual, design, measurement, and analysis issues," *Journal of pain and symptom management*, vol. 31, no. 1, pp. 85–95, 2006.

[10] H. M. Skerman, P. M. Yates, and D. Battistutta, "Identification of cancer-related symptom clusters: an empirical comparison of exploratory factor analysis methods," *Journal of pain and symptom management*, vol. 44, no. 1, pp. 10–22, 2012.

[11] T. Asparouhov and B. Muthén, "Exploratory structural equation modeling," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 16, no. 3, pp. 397–438, 2009.

[12] K. Y. Hogarty, C. V. Hines, J. D. Kromrey, J. M. Ferron, and K. R. Mumford, "The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination," *Educational and Psychological Measurement*, vol. 65, no. 2, pp. 202–226, 2005.

[13] C. M. Bender, F. S. Ergÿn, M. Q. Rosenzweig, S. M. Cohen, and S. M. Sereika, "Symptom clusters in breast cancer across 3 phases of the disease," *Cancer nursing*, vol. 28, no. 3, pp. 219–225, 2005.

[14] M. C. Wilmoth, E. A. Coleman, S. C. Smith, and C. Davis, "Fatigue, weight gain, and altered sexuality in patients with breast cancer: exploration of a symptom cluster," in *Oncology nursing forum*, vol. 31, pp. 1069–1075, Onc Nurs Society, 2004.

[15] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, 1998.

[16] R. R. Sokal, "A statistical method for evaluating systematic relationships," *Univ Kans Sci Bull*, vol. 38, pp. 1409–1438, 1958.

[17] N. Wentzensen, L. E. Wilson, C. M. Wheeler, J. D. Carreon, P. E. Gravitt, M. Schiffman, and P. E. Castle, "Hierarchical clustering of human papilloma virus genotype patterns in the ascus-lsil triage study," *Cancer research*, vol. 70, no. 21, pp. 8578–8586, 2010.

[18] M. Guillaud, D. Cox, A. Malpica, G. Staerkel, J. Matisic, D. Van Niekirk, K. Adler-Storthz, N. Poulin, M. Follen, and C. MacAulay, "Quantitative histopathological analysis of cervical intra-epithelial neoplasia sections: methodological issues," *Analytical Cellular Pathology*, vol. 26, no. 1, pp. 31–43, 2004.

[19] L. J. van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[20] C. M. Perou, S. S. Jeffrey, M. Van De Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. Lee, *et al.*, "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers," *Proceedings of the National Academy of Sciences*, vol. 96, no. 16, pp. 9212–9217, 1999.

[21] H. Lind, A. Waldenström, G. Dunberger, M. al Abany, E. Alevronta, K. Johansson, C. Olsson, T. Nyberg, U. Wilderäng, G. Steineck, *et al.*, "Late symptoms in long-term gy-

naecological cancer survivors after radiation therapy: a population-based cohort study," *British journal of cancer*, vol. 105, no. 6, pp. 737–745, 2011.

[22] G. W. Milligan and M. C. Cooper, "A study of standardization of variables in cluster analysis," *Journal of Classification*, vol. 5, no. 2, pp. 181–204, 1988.

[23] Kardi Teknomo, "Similarity Measurement." `http://people.revoledu.com/kardi/tutorial/Similarity/Normalized-Rank.html/`.

[24] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*, vol. 20. Society for Industrial and Applied Mathematics, 2007.

[25] G. N. Lance and W. T. Williams, "A general theory of classificatory sorting strategies 1. hierarchical systems," *The computer journal*, vol. 9, no. 4, pp. 373–380, 1967.

[26] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[27] J. H. Ward Jr and M. E. Hook, "Application of an hierarchial grouping procedure to a problem of grouping profiles.," *Educational and Psychological Measurement*, 1963.

[28] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[29] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[30] G. W. Milligan and M. C. Cooper, "A study of the comparability of external criteria for hierarchical cluster analysis," *Multivariate Behavioral Research*, vol. 21, no. 4, pp. 441–458, 1986.

[31] J. M. Santos and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *Artificial Neural Networks–ICANN 2009*, pp. 175–184, Springer, 2009.

[32] R. C. MacCallum, "The need for alternative measures of fit in covariance structure modeling," *Multivariate Behavioral Research*, vol. 25, no. 2, pp. 157–162, 1990.

[33] J. C. Hayton, D. G. Allen, and V. Scarpello, "Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis," *Organizational research methods*, vol. 7, no. 2, pp. 191–205, 2004.

[34] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research.," *Psychological methods*, vol. 4, no. 3, p. 272, 1999.

[35] J. L. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika*, vol. 30, no. 2, pp. 179–185, 1965.

[36] B. P. O'connor, "Spss and sas programs for determining the number of components using parallel analysis and velicer's map test," *Behavior research methods, instruments, & computers*, vol. 32, no. 3, pp. 396–402, 2000.

[37] A. Field, *Discovering statistics using SPSS*. Sage publications, 2009.

[38] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, vol. 23, no. 3, pp. 187–200, 1958.

[39] C. D. Dziuban and E. C. Shirkey, "When is a correlation matrix appropriate for factor analysis? some decision rules.," *Psychological Bulletin*, vol. 81, no. 6, p. 358, 1974.

[40] C. Miaskowski and B. E. Aouizerat, "Is there a biological basis for the clustering of symptoms?," in *Seminars in oncology nursing*, vol. 23, pp. 99–105, Elsevier, 2007.

[41] W. Williams, H. Clifford, and G. Lance, "Group-size dependence: a rationale for choice between numerical classifications," *The Computer Journal*, vol. 14, no. 2, pp. 157–162, 1971.

[42] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[43] K. Y. Yeung and W. L. Ruzzo, "Details of the adjusted Rand index and clustering algorithms, supplement to the paper "An empirical study on principal component analysis for clustering gene expression data"," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[44] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, 2001.

[45] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*. John Wiley & Sons, Ltd, 5 ed., 2011.

[46] A. Gordon, *Classification*. Chapman&Hall, CRC, Boca Raton, FL, 2 ed., 1999.

[47] T. Lim and H. Khoo, "Sampling properties of gower's general coefficient of similarity," *Ecology*, pp. 1682–1685, 1985.

[48] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, pp. 857–871, 1971.

[49] J. Newsom, "A quick primer on exploratory factor analysis," *Retrieved August*, vol. 24, p. 2009, 2005.

[50] S. Pavoine, J. Vallet, A.-B. Dufour, S. Gachet, and H. Daniel, "On the challenge of treating various types of variables: application for improving the measurement of functional diversity," *Oikos*, vol. 118, no. 3, pp. 391–402, 2009.

[51] J. M. Fowler, K. M. Carpenter, P. Gupta, D. M. Golden-Kreutz, and B. L. Andersen, "The gynecologic oncology consult: Symptom presentation and concurrent symptoms of depression and anxiety," *Obstetrics and gynecology*, vol. 103, no. 6, p. 1211, 2004.

[52] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[53] R. T. Ng, J. Sander, M. C. Sleumer, *et al.*, "Hierarchical cluster analysis of sage data for cancer profiling," in *Proceedings of BIOKDD 2001 Workshop on Data Mining in Bioinformatics*, pp. 65–72, 2001.

[54] R. A. Roiland and S. M. Heidrich, "Symptom clusters and quality of life in older adult breast cancer survivors," in *Oncology Nursing Forum*, vol. 38, pp. 672–680, Onc Nurs Society, 2011.

[55] A. G. Gift, A. Jablonski, M. Stommel, and C. William Given, "Symptom clusters in elderly patients with lung cancer," in *Oncology Nursing Forum*, vol. 31, pp. 203–212, Onc Nurs Society, 2004.

[56] S. L. Maliski, L. Kwan, D. Elashoff, and M. S. Litwin, "Symptom clusters related to treatment for prostate cancer," in *Oncology nursing forum*, vol. 35, pp. 786–793, Onc Nurs Society, 2008.

[57] R. S. Longman, A. A. Cota, R. R. Holden, and G. C. Fekken, "Pam: A double-precision fortran routine for the parallel analysis method in principal components analysis," *Behavior Research Methods, Instruments, & Computers*, vol. 21, no. 4, pp. 477–480, 1989.

[58] B. Smith, C. W. Hoge, T. C. Smith, T. I. Hooper, E. J. Boyko, G. D. Gackstetter, C. A. LeardMann, M. L. Kelton, and P. D. Bliese, "Exploratory factor analysis of self-reported symptoms in a large, population-based military cohort," 2010.

[59] A. B. Costello, "Getting the most from your analysis," *Pan*, vol. 12, no. 2, pp. 131–146, 2009.

[60] F. Chang, W. Qiu, R. H. Zamar, R. Lazarus, and X. Wang, "Clues: an r package for nonparametric clustering based on local shrinking," *Journal of Statistical Software*, vol. 33, no. 4, pp. 1–16, 2010.

[61] D. P. Farrington and R. Loeber, "Some benefits of dichotomization in psychiatric and criminological research," *Criminal Behaviour and Mental Health*, vol. 10, no. 2, pp. 100–122, 2000.

APPENDIX A

Tables

Table A.1: Patients characteristics

Characteristic	$\bar{X}$	**SD**
Age(years at questionnaire)	64	10.5
Characteristic	**n**	**%**
Education		
Elementary school or equivalent	146	28
High school, vocational or equivalent	201	39
University or College	169	33
Occupation		
Disability/Sick pensioner	46	9
Employed	185	36
Jobseekers	10	2
Long-term sick, more than 1 month	9	2
Non-working, home workers	11	2
Old pensioner	249	48
Short-term sick leave, less than 1 month	2	0
Student	4	1
Living arrangement		
Living alone with partners "living apart"	32	6
Living alone without partner	126	24
Married or cohabiting	295	57
Widow	63	12
Born in Sweden or abroad		
Abroad	87	17
Sweden	429	83
Residential area		
Big town (Stockholm, Göteborg or Malmö)	312	60
Country (single neighboring houses)	45	9
Small place,Small town or metropolitan	159	31
Gluten intolerance		
No	508	98
Yes	8	2
Diabetes		
No	476	92
Yes	40	8
Diagnosis		
Cancer of the fallopian tube-C57.0	12	2
Cervical cancer-C53.9	118	23
Endometrial cancer-C54.9	303	59
Ovarian cancer-C56.9	42	8
Uterine sarcoma-C49.5	25	5
Vaginal cancer-C52.9	12	2
Vulvar cancer-C51.9	4	1
Types of cancer treatment		
RT	2	0
RT+Br	23	4
RT+Br+Ch	19	4
RT+Ch	7	1
RT+Op	38	7
RT+Op+Br	289	56
RT+Op+Br+Ch	84	16
RT+Op+Ch	54	10

N= 516 (after omitting 89 missing data)
Note. Because of rounding, percentages may not total 100

Table A.2: List of variables from gastrointestinal area of the survey questionnaire

VarList	QuesNo	Variable
1	N38	Loose stools
2	N44	Hard stools
3	N48	Had to push out feces
4	N49	Ability to push out feces
5	N52	Incomplete bowel emptying
6	N53	Returned to bathroom within an hour
7	N60	Abdominal bloating / feeling of bloating
8	N62	Rectal bleeding
9	N65	Mucus in stool
10	N67	Anal itching
11	N70	Anal pain
12	N73	Defecation urgency
13	N74	Stomach ache with bowel movements
14	N75	Ability to hold feces
15	N76	Defecation urgency with fecal leakage
16	N77	Immediate need of toilet
17	N81	Leakage of stool without forewarning despite previous defecation
18	N82	Leakage of all stool into clothing without forewarning
19	N84	Involuntary flatulence
20	N85	Loud flatulence
21	N88	Foul-smelling flatulence
22	N91	Leakage of mucus when awake
23	N92	Leakage of blood when awake
24	N93	Leakage of loose stools when awake
25	N94	Leakage of solid stools when awake
26	N95	Leakage of mucus when asleep
27	N96	Leakage of blood when asleep
28	N97	Leakage of loose stools when asleep
29	N98	Leakage of solid stools when asleep
30	N99	Soiled clothing due to leakage of mucus
31	N102	Soiled clothing due to leakage of blood
32	N114	Smells of feces
33	N137	Abdominal pain
34	N138	Abdominal pain intensity
35	N139	Abdominal pain with vomiting
36	N140	Abdominal pain with defecation
37	N141	Abdominal pain with bloating

Table A.3: Percentage of self-repoted symptoms' occurrence and their original mean scores

Symptom	n	Occurance (%)	Median	Scale Range	Original Mean Score	SD
N38	419	81.20	3	1-6	3.08	1.68
N73	389	75.39	2	1-6	2.73	1.60
N53	365	70.74	2	1-6	2.51	1.55
N75	364	70.54	3	1-5	2.81	1.40
N84	355	68.80	2	1-6	2.40	1.53
N60	346	67.05	2	1-6	2.43	1.49
N77	340	65.89	2	1-6	2.27	1.45
N88	328	63.57	2	1-6	2.31	1.51
N85	307	59.50	2	1-6	2.16	1.44
N138	275	53.29	2	1-7	2.54	1.79
N74	274	53.10	2	1-4	1.85	0.98
N137	273	52.91	2	1-6	1.93	1.25
N52	258	50.00	1.5	1-6	1.93	1.31
N76	257	49.81	1	1-6	1.74	0.98
N48	219	42.44	1	1-4	1.53	0.71
N44	214	41.47	1	1-6	1.62	0.95
N141	214	41.47	1	1-4	1.69	0.96
N140	178	34.50	1	1-4	1.60	0.95
N67	171	33.14	1	1-6	1.52	0.94
N93	165	31.98	1	1-6	1.51	0.98
N49	164	31.78	1	1-4	1.41	0.67
N81	161	31.20	1	1-6	1.46	0.86
N65	138	26.74	1	1-6	1.48	0.98
N70	111	21.51	1	1-6	1.30	0.69
N62	91	17.64	1	1-6	1.26	0.70
N114	91	17.64	1	1-6	1.24	0.62
N91	75	14.53	1	1-6	1.23	0.69
N99	75	14.53	1	1-6	1.19	0.57
N82	57	11.05	1	1-5	1.16	0.52
N97	56	10.85	1	1-6	1.13	0.45
N139	50	9.69	1	1-4	1.15	0.49
N94	39	7.56	1	1-6	1.11	0.48
N92	35	6.78	1	1-6	1.11	0.49
N102	27	5.23	1	1-6	1.07	0.33
N95	26	5.04	1	1-5	1.07	0.35
N98	13	2.52	1	1-6	1.03	0.22
N96	5	0.97	1	1-6	1.01	0.10

Table A.4: Optimal number of clusters and adjusted rand index for using different distance methods on various data types with linkage criterion.

No.	Data Used	Scale	Distance Method	Algorithm	Optimal no of Clusters	Adjusted Rand Index (Max.)
1	Raw data	Ordinal	Pearson correlation	Single	14	0.45
	Raw data		Pearson correlation	Complete	11	0.55
	Raw data		Pearson correlation	Average	12	0.60
	Raw data		Pearson correlation	Ward	10	0.55
2	Dichotomize Data	Binary	Jaccard's	Single	19	0.32
	Dichotomize Data		Jaccard's	Complete	17	0.46
	Dichotomize Data		Jaccard's	Average	16	0.39
	Dichotomize Data		Jaccard's	Ward	14	0.41
3	Dichotomize Data	Binary	Phi correlation	Single	20	0.36
	Dichotomize Data		Phi correlation	Complete	11	0.60
	Dichotomize Data		Phi correlation	Average	11	0.66
	Dichotomize Data		Phi correlation	Ward	9	0.73
4	Standardize Data	Numerical	Euclidean	Single	22	0.21
	Standardize Data		Euclidean	Complete	9	0.28
	Standardize Data		Euclidean	Average	21	0.22
	Standardize Data		Euclidean	Ward	3	0.35
5	Standardize Data	Numerical	Pearson correlation	Single	14	0.45
	Standardize Data		Pearson correlation	Complete	11	0.55
	Standardize Data		Pearson correlation	Average	12	0.60
	Standardize Data		Pearson correlation	Ward	10	0.55
6	Raw data	Ordinal	Gower	Single	17	0.24
	Raw data		Gower	Complete	16	0.28
	Raw data		Gower	Average	18	0.26
	Raw data		Gower	Ward	9	0.30
7	Normalized Rank	Numerical	Euclidean	Single	16	0.33
	Normalized Rank		Euclidean	Complete	5	0.33
	Normalized Rank		Euclidean	Average	14	0.38
	Normalized Rank		Euclidean	Ward	6	0.37
8	Rank Data	Ordinal	Kendall Tau	Single	15	0.43
	Rank Data		Kendall Tau	Complete	11	0.55
	Rank Data		Kendall Tau	Average	11	0.61
	Rank Data		Kendall Tau	Ward	9	0.64
9	Rank Data	Ordinal	Spearman Footrule	Single	13	0.33
	Rank Data		Spearman Footrule	Complete	8	0.37
	Rank Data		Spearman Footrule	Average	8	0.37
	Rank Data		Spearman Footrule	Ward	6	0.35
10	Rank Data	Ordinal	Pearson correlation	Single	16	0.42
	Rank Data		Pearson correlation	Complete	11	0.63
	Rank Data		Pearson correlation	Average	12	0.43
	Rank Data		Pearson correlation	Ward	9	0.64
11	Rank Data	Ordinal	Goodman and Kruskal's gamma	Single	15	0.38
	Rank Data		Goodman and Kruskal's gamma	Complete	8	0.69
	Rank Data		Goodman and Kruskal's gamma	Average	8	0.64
	Rank Data		Goodman and Kruskal's gamma	Ward	11	0.57

Table A.5: Initial loadings/correlations of the symptoms on 7 retained factors.

List	QuesNo	Symptoms	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	N38	Loose stools	0.48455*	-0.42115*	-0.24945	-0.10599	0.07954	-0.03291	-0.21274
2	N44	Hard stools	0.34003*	0.60353*	-0.00292	0.31866*	-0.18778	0.06634	-0.2569
3	N48	Had to push out feces	0.32619*	0.66184*	-0.12218	0.35053*	-0.17631	0.08805	-0.2257
4	N49	Ability to push out feces	0.37479*	0.61586*	-0.00872	0.30707*	-0.15166	0.05326	-0.16827
5	N52	Incomplete bowel emptying	0.46364*	0.3497*	-0.04125	0.17244	-0.05393	-0.02801	-0.22154
6	N53	Returned to bathroom within an hour	0.56317*	-0.21761	-0.20726	0.07139	0.13426	-0.1136	-0.16538
7	N60	Abdominal bloating	0.54965*	0.19513	-0.17797	-0.00329	0.14326	-0.15071	0.14277
8	N62	Rectal bleeding	0.39965*	0.23379	0.28460	-0.07855	0.43871*	0.2629	-0.16991
9	N65	Anal leakage of mucus	0.4839*	0.16825	0.22929	-0.1898	-0.02791	-0.5019*	-0.22357
10	N67	Anal itching	0.42624*	0.02023	-0.06096	0.03609	0.17728	0.09424	-0.14385
11	N70	Anal pain	0.45153*	0.21507	-0.12394	-0.15087	0.17937	0.13201	0.00103
12	N73	Defecation urgency	0.63102*	-0.33569*	-0.22635	0.07229	0.09695	-0.05146	-0.22717
13	N74	Stomach ache with bowel movements	0.48435*	-0.096	-0.28507	-0.05227	-0.0047	0.03284	0.07202
14	N75	Ability to hold feces	0.50229*	-0.27614	-0.11468	0.06812	0.05208	0.1553	-0.22042
15	N76	Defecation urgency with fecal leakage	0.59425*	-0.39487*	0.03313	0.08293	-0.15288	0.0362	-0.29937*
16	N77	Immediate need of toilet	0.50963*	-0.46658*	-0.15778	0.06861	0.15961	0.01372	-0.0027
17	N81	Leakage of stool without forewarning despite previous defecation	0.60233*	-0.28759	0.21513	0.07173	-0.2508	0.16914	-0.16438
18	N82	Leakage of all stool into clothing without forewarning	0.45258*	-0.24117	0.15283	0.09707	-0.07289	0.32418*	0.02512
19	N84	Involuntary flatulence	0.55742*	-0.06669	-0.00894	0.50624*	0.18997	-0.30172*	0.2624
20	N85	Loud flatulence	0.50173*	-0.0403	-0.07483	0.47224*	0.26775	-0.25677	0.34887*
21	N88	Foul-smelling flatulence	0.51194*	-0.03469	0.03472	0.46976*	0.20674	-0.26354	0.26791
22	N91	Leakage of mucus when awake	0.52005*	0.11403	0.42892*	-0.32194*	-0.08448	-0.44829*	-0.00811
23	N92	Leakage of blood when awake	0.44196*	0.11459	0.48086*	-0.14547	0.53248*	0.27731	-0.01518
24	N93	Leakage of loose stools when awake	0.62361*	-0.30276*	0.20523	-0.00423	-0.22985	0.08153	-0.23661
25	N94	Leakage of solid stools when awake	0.29926*	-0.06986	0.24088	0.23836	-0.31194*	0.28023	0.31069*
26	N95	Leakage of mucus when asleep	0.47678*	0.03765	0.30673*	-0.32941*	-0.20126	-0.2004	0.22925
27	N96	Leakage of blood when asleep	0.15506	0.03823	0.17256	-0.12484	0.38319*	0.10637	0.13374
28	N97	Leakage of loose stools when asleep	0.55289*	-0.12899	0.14547	-0.06909	-0.35131*	0.08711	0.15981
29	N98	Leakage of solid stools when asleep	0.26812	-0.01742	0.19618	0.13289	-0.25398	0.32305*	0.37624*
30	N99	Soiled clothing due to leakage of mucus	0.53917*	0.08013	0.36245*	-0.29571*	-0.1649	-0.44661*	-0.01736
31	N102	Soiled clothing due to leakage of blood	0.37991*	0.16942	0.49874*	-0.12338	0.32417*	0.31447*	0.03095
32	N114	Smells of feces	0.58136*	-0.08685	0.0792	0.13231	-0.24352	0.09934	0.14309
33	N137	Abdominal pain	0.56353*	0.11756	-0.32145*	-0.20137	-0.05572	0.05577	0.19787
34	N138	Abdominal pain intensity	0.51921*	0.24113	-0.39718*	-0.3432*	0.04277	0.0656	0.17829
35	N139	Abdominal pain with vomiting	0.32992*	0.13617	-0.04725	-0.24773	-0.24167	0.18027	0.18176
36	N140	Abdominal pain with defecation	0.54494*	0.10868	-0.43064*	-0.35703*	-0.04621	0.1261	0.04956
37	N141	Abdominal pain with bloating	0.54211*	0.27839	-0.46826*	-0.32273*	0.0368	0.04788	0.14513

Table A.6: Rotated (Varimax) loadings of the symptoms on 7 retained factors.

List	QuesNo	Symptoms	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
1	N38	Loose stools	0.6665*	0.247	0.0697	-0.1351	0.0748	-0.0659	-0.0082
2	N44	Hard stools	0.0014	0.0881	0.0664	0.8139*	0.0521	0.089	0.0422
3	N48	Had to push out feces	-0.0344	0.1677	-0.021	0.8554*	0.0835	0.0606	0.0018
4	N49	Ability to push out feces	-0.0332	0.1427	0.0777	0.7809*	0.1165	0.1111	0.071
5	N52	Incomplete bowel emptying	0.1905	0.1744	0.154	0.554*	0.1251	0.0208	0.0896
6	N53	Returned to bathroom within an hour	0.5684*	0.2317	0.1034	0.068	0.2686	-0.0682	0.0447
7	N60	Abdominal bloating	0.1376	0.4437*	0.1898	0.1895	0.3619*	0.0007	0.1297
8	N62	Rectal bleeding	0.154	0.0912	0.0722	0.2309	0.0082	-0.0161	0.7079*
9	N65	Anal leakage of mucus	0.1921	0.0915	0.7164*	0.2243	0.1149	-0.1552	0.0873
10	N67	Anal itching	0.3298*	0.1893	0.0173	0.1782	0.116	-0.0088	0.2406
11	N70	Anal pain	0.147	0.4305*	0.0589	0.1866	0.0615	0.0087	0.298*
12	N73	Defecation urgency	0.7187*	0.2298	0.0789	0.0423	0.2287	-0.0198	0.031
13	N74	Stomach ache with bowel movements	0.3392*	0.422*	0.0269	0.0227	0.1695	0.104	-0.013
14	N75	Ability to hold feces	0.6071*	0.1438	-0.0287	0.0671	0.0708	0.0969	0.1134
15	N76	Defecation urgency with fecal leakage	0.7381*	0.0146	0.1862	0.0769	0.0429	0.2095	-0.0025
16	N77	Immediate need of toilet	0.6071*	0.176	-0.0106	-0.1917	0.2882	0.0878	0.0748
17	N81	Leakage of stool without forewarning despite previous defecation	0.5909*	-0.0064	0.223	0.1185	-0.0108	0.4405*	0.0915
18	N82	Leakage of all stool into clothing without forewarning	0.4067*	0.0422	-0.0169	0.0173	0.0456	0.4404*	0.2074
19	N84	Involuntary flatulence	0.2472	0.0475	0.1232	0.1455	0.8038*	0.1389	0.0496
20	N85	Loud flatulence	0.1662	0.1192	0.0364	0.0869	0.8193*	0.1112	0.0838
21	N88	Foul-smelling flatulence	0.1926	0.0322	0.1168	0.1359	0.7545*	0.142	0.0976
22	N91	Leakage of mucus when awake	0.0948	0.1048	0.8419*	0.0565	0.0913	0.0696	0.1846
23	N92	Leakage of blood when awake	0.1353	0.0406	0.1583	0.0374	0.0709	0.0992	0.8719*
24	N93	Leakage of loose stools when awake	0.6359*	0.0109	0.3096*	0.1006	-0.0363	0.3455*	0.0838
25	N94	Leakage of solid stools when awake	0.0617	-0.0268	0.0005	0.0816	0.133	0.6722*	0.0296
26	N95	Leakage of mucus when asleep	0.0325	0.2414	0.6207*	-0.0658	0.05	0.3165*	0.1139
27	N96	Leakage of blood when asleep	-0.0287	0.0942	0.0282	-0.1205	0.1125	-0.0104	0.4573*
28	N97	Leakage of loose stools when asleep	0.2874	0.2128	0.3045*	0.037	0.04	0.5302*	-0.0095
29	N98	Leakage of solid stools when asleep	-0.0206	0.0775	-0.0289	0.0298	0.0948	0.644*	0.0833
30	N99	Soiled clothing due to leakage of mucus	0.143	0.1288	0.8214*	0.0722	0.0865	0.1005	0.0918
31	N102	Soiled clothing due to leakage of blood	0.0439	0.0194	0.164	0.0962	-0.0087	0.2356	0.7402*
32	N114	Smells of feces	0.3188*	0.1787	0.1737	0.1485	0.1979	0.4834*	0.0185
33	N137	Abdominal pain	0.1837	0.6496*	0.1102	0.0927	0.1299	0.1635	0.0258
34	N138	Abdominal pain intensity	0.0938	0.7746*	0.0984	0.0985	0.0538	0.0273	0.0968
35	N139	Abdominal pain with vomiting	0.0106	0.4062*	0.1426	0.0698	-0.124	0.3291*	0.0302
36	N140	Abdominal pain with defecation	0.2635	0.744*	0.0727	0.0847	-0.0556	0.0521	0.0299
37	N141	Abdominal pain with bloating	0.1128	0.8168*	0.0823	0.1564	0.0674	-0.0128	0.061

APPENDIX B

Figures

Figure B.1: Marginal distributions of variables for survivor data.

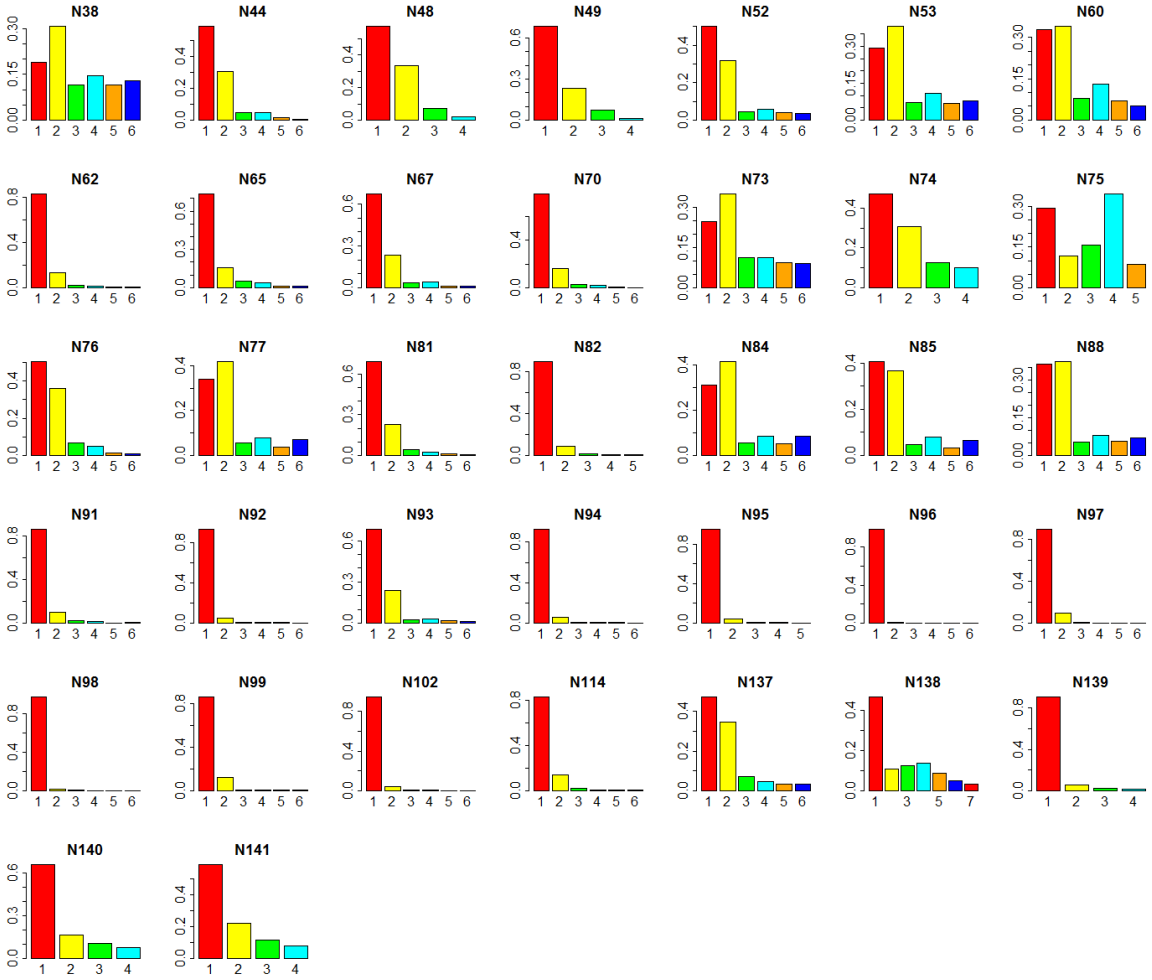

Figure B.2: Marginal distributions for binary variables of the transformed 37 gastrointestinal questions.