A Biclustering Approach to Symptom Clusters and Subgroup Identification in Non-Hodgkin Lymphoma Survivors

Django de Smet
STUDENT NUMBER: 2019023

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COMMUNICATION AND INFORMATION SCIENCES
MASTER TRACK COMMUNICATION AND INFORMATION SCIENCES, DATA SCIENCE:
BUSINESS & GOVERNANCE
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Dr. Katrijn Van Deun
Prof. Eric Postma

External supervisor:

Dr. Simone Oerlemans (IKNL Eindhoven)

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
May 2019

Preface

First and foremost, I would like to thank my supervisor Dr. Katrijn Van Deun of the Methodology and Statistics department at Tilburg University. When I approached her in December during one of her lectures and asked her if she knew an interesting topic I could work on for my master's thesis she immediately responded with a lot of enthusiasm. I would like to thank her for her insightful comments, helpful feedback, and overall nudges in the right direction. Katrijn, thank you, I could not have imagined a better supervisor to guide me during this demanding, highly educational, and also fun, time.

I would also very much like to express my gratitude to Dr. Simone Oerlemans from the Integraal Kankercentrum Nederland for providing the dataset that was used for this thesis and for her words of encouragement, feedback, and overall knowledge on cancer literature. Above all, I would like to thank her for her ever present willingness to answer all of my numerous questions on everything dataset related.

A big thank you also goes out to my parents, Peter and Lean, for always believing in me, for their unwavering support, and for always being able to lift my spirits when needed. To my brother, Jay, I would like to say thank you for providing me with the opportunity to put my mind to something else for a couple of hours a week in the gym. Finally, I would also like to thank my good friend, Robbert. Being in the same boat as me, we experienced many of the same things and our conversations have helped me a great deal throughout the process of writing this thesis.

A Biclustering Approach to Symptom Clusters and Subgroup Identification in Non-Hodgkin Lymphoma Survivors

Django de Smet

The number of long-term Non-Hodgkin lymphoma (NHL) survivors has increased dramatically due to better treatments, care, and improvements in early detection. Previous research has shown that many of these long-term cancer survivors still experience multiple symptoms, referred to as symptom clusters, as a result of their diagnosis and treatment. These symptom clusters have been shown to negatively impact a survivor's health-related quality of life (HRQOL). Traditional clustering algorithms, such as principal component analysis (PCA), hierarchical cluster analysis (HCA), and factor analysis (FA), have previously been used to identify symptom clusters. However, because these methods can only be used to cluster either variables (symptoms) or observations (survivors), but never both simultaneously, the discovered subgroups and symptom clusters are of limited clinical relevance since the differences between subgroups are not taken into account. Rather, the discovered symptom cluster is expressed over all survivors. Thus, since it is known that there is a fair amount of heterogeneity between patients, these differences between subgroups of patients are not taken into account, thereby decreasing the clinical relevance of the discovered symptom clusters. This study tried to address these shortcomings by investigating the use of biclustering algorithms on symptom data gathered from long-term NHL survivors. In the context of symptom science, biclustering algorithms are capable of simultaneously discovering subgroups and symptom clusters, potentially leading to more clinically relevant results. This study is the first study to use biclustering algorithms to discover subgroups and symptom clusters from symptom data. The results of this study show that the repeated Bimax biclustering algorithm can be used to discover distinct subgroups of survivors that differ with respect to the symptoms they experience. More specifically, nine different subgroups of survivors were found, with each subgroup showing a different symptom profile. Importantly, the results from the repeated Bimax algorithm could be used by clinicians to target specific subgroups and develop treatment plans that are tailored to the symptom profiles of the specific subgroups.

1. Theoretical Background

1.1 Introduction

Previous research on cancer has shown that many cancer patients/cancer survivors experience multiple symptoms, referred to as symptom clusters, as a result of their diagnosis and treatment (Esther Kim et al. 2009). Other studies have also shown that these symptom clusters negatively affect patients' health-related quality of life (HRQOL) and functional status (Leach et al. 2014). However, as of yet, there is no consensus on which statistical methods are most appropriate for symptom cluster identification, leading to

results of different studies being incomparable and to limited clinical relevance of the findings.

A limited number of statistical methods, including Principal Component Analysis (PCA), Factor Analysis (FA), and Hierarchical Cluster Analysis (HCA), has been considered for symptom cluster identification and these methods have significant shortcomings that make them less than optimal for symptom cluster identification. One issue with the aforementioned methods is that some degree of subjectivity is involved with deciding which symptoms to include in a cluster. Ideally, one would want to find discrete symptom clusters and thus find symptoms that are experienced by all members of a particular subgroup. The methods used so far are not the most suitable for this purpose. Another issue is that these methods are unable to directly discover subgroups of patients that differ with respect to symptom cluster experience (i.e., either variables are clustered, or observations, but never both simultaneously). This also means that symptom clusters are expressed over the entire sample. Thus, clinically relevant information on differences between subgroups with respect to symptom profiles is lost.

More modern data mining (DM) techniques exist, such as biclustering algorithms, that could provide more clinically relevant results with respect to symptom cluster identification and that address some of the shortcomings of the methods mentioned above. In the context of symptom clusters, biclustering algorithms can be used to find subgroups of patients that share a similar symptom experience (i.e., symptom clusters). Thus, biclustering algorithms are capable of discovering subgroups that are similar on a subset of the variables, whereas traditional clustering algorithms are only capable of discovering patterns (based on all variables) that hold for all observations. This means that bicluster algorithms are potentially capable of providing more clinically relevant findings because symptom clusters can be expressed differently for different patient subgroups. Therefore, in this paper, the use of biclustering algorithms for symptom cluster and subgroup identification in Non-Hodgkin Lymphoma (NHL) survivors will be investigated and compared to symptom clusters generated by PCA and HCA.

In the remainder of this section, I will provide a brief overview of cancer statistics, address different types of cancer, discuss symptoms experienced by long-term cancer survivors, look at symptom cluster consistency, present past and current statistical approaches to symptom cluster identification, introduce the biclustering algorithms that will be used in this paper, and present the research question that will be investigated in this paper. In section 2, I will present the dataset and the algorithms that were used to analyze the data. Subsequently, in section 3, I will address the results obtained from the analyses. Lastly, in section 4, the results of the analyses will be discussed in relation to findings from previous research.

1.2 Cancer Statistics

In the Netherlands, as well as globally, the number of cancer patients increases yearly. (Kankerbestrijding 2011; Bray et al. 2018). In many countries, cancer consistently ranks as one of the leading causes of death for people below the age of 70 (Bray et al. 2018). In the Netherlands, the incidence rate of cancer (i.e., the number of new cases within a specified period of time) is expected to increase dramatically. In a report on cancer statistics in the Netherlands, published by KWF Kankerbestrijding in 2011, the authors predicted that the number of new cancer cases in 2020 will be roughly 123.000, as opposed to 86.800 new cases in 2007 (Kankerbestrijding 2011). Given that cancer is a disease that is much more common in older people, this increase in cancer cases can be attributed to two factors: 1) an increase in the number of senior citizens (i.e. people

over the age of 65), and 2) an increase in the age of death of this group of people (Kankerbestrijding 2011).

The data that will be used for this project comes from patients/survivors diagnosed with NHL, which is a specific type of cancer that encompasses a broad range of subtypes. With respect to NHL, a distinction can be made between indolent NHL and aggressive NHL (see section 1.3 below for details on the differences between these cancer types). The incidence of both indolent NHL and agressive NHL has risen significantly since 1989 (van de Schans et al. 2011; Nederlandse Kankerregistratie).

The relative 5-year survival rate has also improved for many types of cancer due to better treatments, care, and improvements in early detection (Verdecchia et al. 2007). This means that increasingly more patients diagnosed with cancer become long-term cancer survivors. In oncology, patients are considered to be survivors from the first day they are diagnosed with the disease (National Coalition for Cancer Survivorship 2006). For NHL patients specifically, the 5-year overall survival rate has also increased dramatically, with survival rates for NHL patients ranging from 50% to 62% (Nederlandse Kankerregistratie). This increase in survival rate leads to a growing number of long-term NHL survivors (Verdecchia et al. 2007; Engert et al. 2012, 2017). These long-term cancer survivors are at risk of experiencing adverse physical and psychological effects as a result of their medical treatments and their cancer diagnosis. This can lead to symptoms even long after their original diagnosis (Oerlemans et al. 2013; Harrington et al. 2010).

1.3 Types of Lymphoma

Given that the present study will be based on data collected from patients with NHL, a short overview of the different types of lymphoma is given below.

Lymphomas can be be divided in two broad categories: Non-Hodgkin lymphomas and Hodgkin lymphomas (HL). Both HL and NHL are cancers that begin in a subset of white blood cells known as lymphocytes. Three types of lymphocytes can be distinguished: T cells (thymus cells), B cells (bone marrow derived cells), and NK cells (natural killer cells). All of these cells are involved in the immune system and play a major role in helping the body deal with infections and fight off infected cells (Leukemia & Lymphoma Society 2013a). HL is identified by the presence of cancerous Reed-Sternberg cells, which are usually a malignant type of B cell lymphocytes (Leukemia & Lymphoma Society 2013b). NHL, on the other hand, can occur in B cells, T cells, and NK cells (Leukemia & Lymphoma Society 2013c). In short, any lymphoma that is not a Hodgkin lymphoma is classified as a non-Hodgkin lymphoma. Another difference between NHL and HL is that NHL is much more likely to be extra-nodal (i.e., it manifests itself in an area outside of the lymph nodes) (Newton et al. 1997).

There are over 30 different types of NHL. A common distinction is made between indolent types of NHL and aggressive types of NHL. These terms refer mainly to the pace at which the disease progresses and both types require different treatment approaches. Indolent NHL progresses slowly, symptoms may not appear for quite some time, and treatment is only started when the first symptoms appears. Aggressive NHL, on the other hand, progresses much more rapidly and treatment is usually started immediately after diagnosis. However, the most important difference between indolent and aggressive NHL is that indolent NHL is incurable. Even though patients often live many years with indolent NHL, they can never be declared cancer free, whereas patients with aggressive NHL can often be cured of their disease with proper treatment (Leukemia & Lymphoma Society 2013a).

1.4 Symptoms of Long-Term Cancer Survivors

Luckily, many cancer survivors are able to continue their lives with minimal long-terms effects of the disease and the treatment. However, a smaller but substantial group of survivors experiences symptoms that can impact their lives even years after their treatment has been completed, with the symptoms becoming chronic of nature (Stein, Syrjala, and Andrykowski 2008; Harrington et al. 2010). Some of these symptoms can actually be caused by the treatment itself and will show an immediate impact on HRQOL, while other symptoms may only manifest themselves years later. Thus, it is important to understand the health profiles of these long-term cancer survivors, examine what kind of symptoms these survivors experience, how these symptoms interact, and how these symptoms affect their HRQOL (Leach et al. 2014).

Symptoms are measured in most studies by administering a questionnaire made for use with cancer patients. One example of such a questionnaire is the European Organisation for Research and Treatment of Cancer Quality of Life questionnaire (EORTC QLQ-C30), where different items represent different symptoms and patients are asked to provide a score on a scale from 1 to 4 to indicate the extent to which they experience this symptom (Aaronson et al. 1993). Given that the EORTC QLQ-C30 will also be used in the current study, a more detailed description can be found in section 2.3 below.

Some examples of long-term symptoms include physical and psychological difficulties such as pain, fatigue, problems with infertility, loss of sexual interest, anxiety, and depression (Fobair et al. 1986; Stein, Syrjala, and Andrykowski 2008). Multiple population-based studies have compared long-term cancer survivors with controls without cancer and overall the results indicate that cancer survivors experience more difficulties with physical activities such as doing housework and taking a walk, as well as with participating in social activities, than controls without cancer (Ness et al. 2006; Sweeney et al. 2006).

With respect to NHL specifically, symptoms include, but are not limited to, loss of energy (fatigue), depression, anxiety, problems with infertility, decrease of interest in sexual activity, nausea, vomiting, reduced cognitive functioning, restrictions in role functioning, and increased levels of psychological distress (Fobair et al. 1986; Loge et al. 1999; Hjermstad et al. 2006; Oerlemans et al. 2013). Importantly, these long-term symptoms can greatly affect the HRQOL of long-term NHL survivors.

Recently, more attention has been given to the fact that many survivors experience multiple symptoms. Several studies have looked at the potential interaction between long-term symptoms and how different symptoms may co-occur and form so called symptom clusters that could have a larger effect on HRQOL than the symptoms individually. Treatment of a single symptom could also help to alleviate the adverse effects of the cluster as a whole that this symptom belongs to (Kim et al. 2005). Furthermore, by targeting specific symptom clusters that have the greatest effect on HRQOL, the overall well being of long-term cancer survivors might be improved.

1.5 Symptom Clusters

A relatively recent and promising area of cancer research involves symptom clusters (Miaskowski et al. 2017). Kim, McGuire, Tulman, and Barsevick (2005) provide the following definition of a symptom cluster, which is the most widely cited definition in oncology:

> "A symptom cluster consists of 2 or more symptoms that are related to each other and that occur together. Symptom clusters are composed of stable groups of symptoms, are relatively independent of other clusters, and may reveal specific underlying dimensions of symptoms. Relationships among symptoms within a cluster should be stronger than relationships among symptoms across different clusters. Symptoms in a cluster may or may not share the same etiology" (p. 278).

Some aspects of this definition need clarification. A distinction is made between symptoms being related to each other and symptoms occurring together because concurrence has not yet been sufficiently demonstrated in the literature. This means that, when symptoms are assumed to occur together on the basis of statistical analysis, this does not necessarily imply that these symptoms also presented together at the same time in clinical practice (Kim et al. 2005). Stability, with regards to symptom clusters, should be interpreted as the extent to which symptom clusters are consistent across time, cancer site, and statistical method, which is a somewhat problematic concept (see section 1.5.3 below).

As mentioned in section 1.4, many studies have looked at the characteristics of single symptoms such as nausea, vomiting and fatigue. While knowledge on these single symptoms is important, more recent studies have shown that many cancer patients and survivors experience multiple symptoms and that the negative effect on patient-reported outcomes (such as HRQOL) and functional status of these multiple symptoms is stronger than that of a single symptom (Esther Kim et al. 2009).

In a review of the literature, Esther Kim et al. (2009) found that approximately 40% of cancer patients experienced multiple symptoms, emphasizing the need for research on symptom clusters. In a study on HRQOL and functional status in elderly cancer patients, Cheng and Lee (2011) found that a symptom cluster of pain, fatigue, insomnia, and mood disturbance explained between 8,7% and 52,9% of the variance in functional status and HRQOL. This again highlights the clinical importance of gaining an understanding of symptom clusters because it could lead to improvements in treatment that might positively affect patients' HRQOL and functional status (Dodd, Miaskowski, and Paul 2001).

Symptom clusters have also been found in other chronic conditions such as HIV disease, heart disease, chronic obstructive pulmonary disease (COPD), osteoarthritis, rheumatoid arthritis, and end-stage renal disease (Namisango et al. 2015; Moens et al. 2015; Breland et al. 2015; Park and Larson 2014; Jenkins and McCoy 2015; Bender et al. 2008; Jurgens et al. 2009; Amro et al. 2014; Lee and Jeon 2015).

1.5.1 Symptom Co-occurrence vs. Symptom Clusters. While it seems straightforward to assume that multiple symptoms occurring together have a stronger effect on HRQOL than a single symptom, a crucial distinction has to be made between multiple symptoms simply occurring together and multiple symptoms occurring together and forming a symptom cluster. For example, a cancer patient could experience both pain and nausea but the mere fact that these symptoms co-occur for this particular patient does not necessarily mean that these symptoms also form a symptom cluster. In order for symptoms that co-occur to be considered a symptom cluster, they will also need to be related more strongly to each other than to other symptoms that are not considered part of the cluster and they may share a common underlying (hidden) etiology. This implies that the effect of a symptom cluster on HRQOL should be very different from the effect that the symptoms would have if they were assumed to be simply occurring together, without any relation between them or common underlying construct. One

might think of interaction effects in this context where the effect on HRQOL of a certain symptom (e.g., vomiting) in a cluster differs based on the intensity of another symptom in the cluster (e.g., nausea). Research on symptom clusters, as opposed to research on symptoms simply occurring together, is important because this allows for a better identification of patient subgroups that differ with respect to symptom cluster experience, a better understanding of the relationship between symptoms that make up a cluster, a better understanding of the impact of multiple symptoms on HRQOL, and it could provide knowledge on underlying constructs / common etiology underlying a specific cluster (Barsevick 2007).

1.5.2 Approaches to Symptom Cluster Formation. Two approaches to symptom cluster identification can be seen in the literature. One approach involves symptom cluster identification, *"a priori"*, through qualitative research. In this context, *"a priori"* means that the symptom clusters are identified beforehand. This involves the creation of symptom clusters on the basis of clinical evaluations and through patient interviews (e.g., Molassiotis et al. (2011)). The other approach involves symptom cluster identification *"de novo"*, or empirically, meaning that symptom clusters are derived through quantitative research using methods such as PCA, FA, and HCA (Cheung, Le, and Zimmerman 2009). At first glance, these methods appear to be suitable for symptom cluster identification because they seem to capitalize on the theoretical and clinical assumption that related symptoms cluster together. However, as we will see in section 1.6.1, there are some compelling reasons as to why these methods might not necessarily be the most suitable for symptom cluster identification.

One issue with qualitative identification of symptom clusters is that it is difficult to determine when a cluster can be considered complete and how many clusters there should be (Barsevick 2016). The empirical, or *"de novo"*, approach aims to resolve this issue by using statistical methods whose symptom cluster formations can be evaluated using objective measures. Nevertheless, as will be further explained in section 1.6 below, the problem of selecting the optimal number of symptoms to form a cluster and selecting the optimal number of clusters still remains.

There are still many issues regarding symptom cluster research in general. For instance, there is no consensus on what exactly constitutes a symptom cluster. Some studies argue for a minimum of two symptoms, while others argue for a minimum of three symptoms in order to count as a cluster (Barsevick 2016). Multiple studies have found symptom clusters that consisted of only two symptoms, thus lending support to the inclusion of symptom pairs. For example, Jiménez et al. (2011) found a gastrointestinal cluster consisting of two symptoms, namely nausea and vomiting. Other examples include an anxiety-depression symptom cluster, which was found to be consistent across different statistical approaches (Chen et al. 2012b). Even though a symptom cluster was first formulated in the field as consisting of three or more symptoms that are interrelated, because there are studies showing the advantage of including symptom pairs (i.e., a two-symptom cluster), the definition by Kim et al. (2005) shown above will be used in this paper (Dodd, Miaskowski, and Paul 2001).

Another ongoing issue related to symptom cluster research concerns symptom cluster consistency. There seem to be highly variable results regarding symptom cluster consistency across time as well as across different statistical methods and different cancer sites. These aspects of symptom cluster research will be discussed next.

1.5.3 Symptom Cluster Consistency. Many studies have looked at symptom cluster consistency across different statistical methods, across different cancer sites, and over time.

Because of differences in statistical methods, symptom cluster definitions, assessment tools, cancer sites, and patient populations, the results of different studies on symptom clusters are difficult to compare. Based on the literature that will be discussed below, it seems that much effort is put towards trying to find consistency with regards to symptom cluster formation. However, it is important to consider whether symptom cluster consistency can ever be expected and whether this is the most pressing issue to address. Below, literature on symptom cluster consistency across different statistical methods, across different cancer sites, and over time, is discussed. After that, in section 1.5.7, I will offer a critical reflection on the results from these studies and discuss why symptom cluster consistency might not necessarily be expected and what a more clinically useful approach to symptom cluster research might be.

1.5.4 Symptom Cluster Consistency Across Statistical Methods. Dong et al. (2016) compared symptom cluster formation across different statistical methods (PCA, FA, and HCA), and different cancer sites and they found four symptom clusters that were consistent across methods and cancer sites: an emotional cluster (tense-worry-irritable-depressed), a fatigue-pain cluster, a nausea-vomiting cluster, and a cognitive cluster (concentration-memory). This finding is interesting because it seems to indicate that similar symptom clusters occur, regardless of cancer site and regardless of statistical method. On the other hand, Chen et al. (2012b) found no consistency in the symptom cluster formations of PCA, FA, and HCA across different cancer sites. However, the authors note that the symptom cluster findings of PCA agreed more closely with the symptom cluster findings of HCA than the symptom cluster findings of both of these methods did with FA. This is surprising since there are studies that argue that the outcomes of PCA and FA usually show a high degree of similarity (see section 1.6) (Velicer and Jackson 1990). In a study on patients with different types of advanced cancer, Cheung, Le, and Zimmerman (2009) also found that symptom cluster formations generated with PCA varied by primary cancer site (i.e., different symptom clusters were found for different types of cancer).

With regards to symptom cluster consistency across different methods but in one cancer site, Henoch, Ploner, and Tishelman (2009) looked at the consistency of the symptom cluster formations in a homogeneous sample of lung cancer patients generated by Pearson correlations, HCA, FA, and Cronbach alphas (i.e., a measure of internal consistency). They found multiple clusters that were consistent across methods: a pain cluster (pain-nausea-bowel issues-appetite loss-fatigue), a mood cluster (mood-outlook-concentration), and a respiratory cluster (breathing, cough). As a secondary goal of their study, Henoch, Ploner, and Tishelman (2009) also compared the symptom cluster results of two different instruments, the Symptom Distress Scale (SDS) and the EORTC QLQ-C30, which were also found to be consistent (McCorkle and Young 1978). On the other hand, there are also many studies that found no consistency in symptom cluster formations in the same cancer site but across different statistical methods. For example, Chen et al. (2012a) found no consistency in cluster composition between the symptom cluster findings of PCA, FA, and HCA in patients with bone metastases. For all studies discussed above, HCA was used to cluster variables (symptoms) rather than observations (patients).

1.5.5 Symptom Cluster Consistency in the Same Cancer Sites. In a literature review of symptom clusters in patients with breast cancer, Nguyen et al. (2011) compared the symptom cluster outcomes of five different studies. The results of this literature review showed that there were no clusters that were completely similar across the

five different studies, although some common individual symptoms were identified. Nguyen et al. (2011) note that this disparity in symptom cluster formation of the studies under investigation is likely due to differences in symptom assessment tool, statistical method, and patient population. This highlights the importance of identifying a unified approach to symptom cluster research in order to find clinically relevant results.

Literature reviews conducted by Chen et al. (2011) on symptom clusters in patients with lung cancer and by Thavarajah et al. (2012) on symptom clusters in patients with metastatic cancer showed results similar to those of Nguyen et al. (2011) mentioned above. Only a cluster consisting of nausea and vomiting was consistently identified in two out of the five studies that were investigated by Chen et al. (2011). No other clusters were consistently identified across different studies. Thavarajah et al. (2012) looked at eight studies involving patients with metastatic cancer and found no symptom clusters that consistently identified for all of the studies.

Similar to Nguyen et al. (2011), Chen et al. (2011) and Thavarajah et al. (2012) emphasize that symptom cluster research shows a lot of promise, but significant hurdles regarding methodology need to be overcome in order for the symptom cluster findings to be clinically meaningful. If the wide variation in symptom clusters found in different studies in the same populations is attributable to differences in statistical methods, then it is difficult to paint a clear picture of symptom clusters in the same cancer sites because of these incongruent findings.

1.5.6 Symptom Cluster Consistency over Time. Because of the increase in long-term cancer survivors it is important to examine symptom clusters over time and investigate whether they change after some time since the original diagnosis has passed. Identifying changes in symptom clusters over time could lead to improvements in treatments because this would enable practitioners to be more responsive to a patient's changing needs in the treatment trajectory.

In a study on breast and prostate cancer patients undergoing radiation therapy, Kim et al. (2009) investigated symptoms clusters at three different points in time using FA: at the middle of radiation therapy, at the end, and one month after completion. They found three clusters that were relatively stable across the different time points, although no clusters were completely identical at each time point, especially with regards to symptom severity. Some symptoms, such as pain, increased in severity across time while the occurrence rate decreased. Multiple other studies also found no change in symptom cluster formation, thus arguing for relatively stable symptom clusters over time (Gift et al. 2003; Kim et al. 2008). However, it should be noted that Kim et al. (2008) indicate that the clustering stability decreased as the time since completion of treatment increased. There are also many studies reporting no consistency in symptom cluster formation over time (Ahlberg, Ekman, and Gaston-Johansson 2005; Hadi et al. 2008a,b).

In conclusion, there seem to be inconsistent findings regarding symptom cluster consistency over time, with some studies reporting relatively stable symptom clusters over time and other studies reporting little to no consistency in symptom cluster formation. Also important to keep in mind is that the findings of these studies are difficult to compare because of differences in methodology, patient population, and the time span that was chosen for analysis.

1.5.7 Critical Reflection on Symptom Cluster Consistency. An open question in symptom cluster research is whether symptom clusters are consistent across statistical methods. With respect to homogeneous samples (i.e., patients with the same cancer type and the same disease stage), one would ideally like to see the same symptom cluster

formations across different statistical methods (Henoch, Ploner, and Tishelman 2009). However, as will be discussed in section 1.6, some methods might be more suitable for symptom cluster identification than others, even though they might produce similar clusters. With respect to heterogeneous samples (i.e., patients with different cancer types and at different disease stages), consistency might not necessarily be expected because different cancer types and different disease stages can lead to very different symptom experiences. Also, most of the studies discussed above looked at symptom clusters in patients that are either still in treatment, or shortly after treatment. Much less is known about symptom clusters in patients several years after their original diagnosis.

The fact that some studies have found no symptom cluster consistency between methods in patient populations that are homogeneous with respect to cancer site and disease stage might indicate that the results of the methods used are unstable. This might be due to differences between subgroups of patients which are not accurately captured by PCA, FA, and HCA because these methods cluster either variables or observations and thus do not take potential differences between patient subgroups into account. This is indicative of a major flaw in the current research on symptom clusters: the methods that are used for symptom cluster identification do not fit well with the most clinically relevant goal of symptom cluster identification, namely to find symptom clusters that take into account the variation between patient subgroups. Thus, perhaps the goal should not be to try to prove consistency between methods but rather to investigate which method works best with respect to the clinical goal of symptom cluster identification and to consider the possibility that the current methods are not necessarily the most suited for the task at hand. This could lead to more clinically meaningful symptom clusters which better capture variation between patient subgroups and which might have stronger predictive capabilities with respect to HRQOL. Even if consistency could be found, for example, between the clustering of PCA and FA, if both are less-then-optimal methods for symptom cluster identification then this consistency has little clinical value. Furthermore, it should be noted that PCA, FA, and HCA are based on different definitions of what a symptom cluster is. PCA does not assume an underlying construct, whereas FA does, and HCA can be used to cluster either variables or observations. Thus, a valid question to ask is whether consistency between the results of these methods can be expected, and, more importantly, whether the methods are appropriate to begin with.

Differences in symptom cluster formation between different cancer sites seem to indicate that different patients with different kinds of cancer experience different symptom clusters. Also, differences in symptom cluster formation over time for patients with the same cancer type indicate that a patient's symptom cluster experience changes over time. Both of these results point to a heterogeneity of symptom cluster experience over time and across cancer site/type (Cheung, Le, and Zimmerman 2009; Chen et al. 2012b). This is perhaps not surprising since it seems intuitive that a patient's symptom experience changes over time and that different cancer types cause different symptoms. Critically, it should be noted that the results of these studies are difficult to compare because of differences in statistical methods, disease stage, patient population, assessment tools, etc. However, similar to what was mentioned above, the statistical methods used in these studies may not be the best suited methods for symptom cluster identification and it should be considered whether proving consistency is the best goal to strive towards given that the optimal method for symptom cluster identification has not been found yet.

Taken together, the results on symptom cluster consistency across statistical methods discussed above exemplify the lack of consensus regarding the best statistical

approach to finding symptom clusters and thus far, much of the research has focused on comparing the results of three different methods (PCA, FA, and HCA), with virtually no interest being shown in investigating the use of more advanced DM clustering methods that might be better suited for the goals of symptom cluster research. Therefore, one of the aims of this thesis is to investigate the use of DM biclustering algorithms for symptom cluster identification. These biclustering algorithms might be better suited for finding symptom clusters, while taking differences between patient subgroups into account, because these algorithms work by clustering both variables and observations simultaneously. Thus, these algorithms address some shortcomings of the more traditional methods and they might provide more clinically meaningful symptom clusters (see section 1.8 below for more information on biclustering).

1.6 Statistical Methods Used for Symptom Cluster Identification

As mentioned above, one of the limitations in symptom cluster research is the lack of consensus on which statistical techniques are most appropriate for identifying symptom clusters. Some of the traditional statistical techniques that have been used most frequently for symptom cluster identification include PCA, FA, and HCA. Below, an overview of these methods will be given.

PCA is a dimension reduction method that creates so-called principal components that aim to capture the variation that is present in the data (Jolliffe and Cadima 2016). Each of the constructed components is a linear combination of all of the original features. In other words, the goal of PCA is to reduce a set of p observed features to a set of m new features where $m < p$. Most of the time, the first few principal components capture most of the variation present in the data, thus leading to dimension reduction (James et al. 2013; Jolliffe and Cadima 2016). If we envision that the original features are the individual symptoms, then the principal components that are constructed by PCA can be interpreted as symptom clusters where multiple individual symptoms are loaded onto a single principal component (cluster). Thus, PCA is one way of generating symptom clusters from symptom data. A clear example of PCA applied to symptom data can be found in (Fan, Hadi, and Chow 2007).

The most used form of FA is common FA. This is also a dimension reduction method that is very comparable to PCA in the sense that the goal of common FA is also to reduce a set of p observed features to a set of m new features where $m < p$. However, in contrast to PCA, in FA there is the assumption of an underlying latent variable that FA aims to model. For example, we may not be able to directly model motivation (i.e., we cannot capture it with one variable), but we can measure several variables that we think relate to motivation (e.g., "I always work hard", "I find it important to get good grades") and subsequently use these new variables as a substitute for the latent (invisible) variable, motivation. FA can then be used to find the optimal weights between the latent variable and the observed variables. Thus, the major difference between FA and PCA is that in FA we assume that some underlying construct is causing the responses we see on the observed variables, whereas in PCA no such assumption is made.

One study has pointed out that PCA might not be the most suitable method for finding symptom clusters. Skerman, Yates, and Battistutta (2009) claim that, conceptually, only FA and HCA are appropriate methods for discovering symptom clusters. According to Skerman, Yates, and Battistutta (2009) PCA is not appropriate because there is no assumption with PCA that an underlying construct is at the basis of a cluster and causes the clustering of particular symptoms to occur. However, Velicer and Jackson (1990), in a study comparing FA and PCA, conclude that the choice of method will have

no significant effect on the conclusions derived from the outcomes of these methods because of the high degree of similarity between the outcomes of FA and PCA. This statement has to be treated with caution, because, as mentioned in section 1.5.4 above, Chen et al. (2012b) found the symptom cluster findings of PCA to be more in agreement with the findings of HCA than the symptom cluster findings of FA. Nevertheless, as mentioned above, it should be noted that FA and HCA might also not be ideal methods for symptom cluster identification in this context.

Another statistical method that is often used in symptom cluster identification is HCA. HCA can be used to group observations or variables that are similar to other members of their group (cluster) and are most dissimilar to members of others groups based on some distance measure. This dissimilarity can be measured in multiple ways but the most commonly used distance measure is the Euclidian distance. Agglomerative (bottom-up) HCA starts off with every variable (or observation) in its own cluster. At every next step, it joins together the most similar clusters until all of the variables (or observations) are in one single cluster. The result of an HCA is a dendogram, which is often compared to an upside-down tree. At the bottom of this dendogram are the "leaves", which refer to the individual variables, and at the top of the dendogram is the "trunk", which refers to the single cluster that contains all of the variables (James et al. 2013).

1.6.1 Shortcomings of PCA, FA, and HCA. All of these statistical methods have some practical disadvantages that could potentially lead to a less than optimal clustering of variables. Furthermore, there are important theoretical considerations with respect to symptom clusters that indicate that some of these methods might not be the most suitable for symptom cluster research.

One of the shortcoming of PCA and FA, with respect to symptom science, is that PCA and FA create latent variables (components) that are linear combinations of all of the original variables. This means that the loadings for the individual variables on a particular latent variable will be non-zero, potentially making it very difficult to interpret the results (Zou, Hastie, and Tibshirani 2006). Thus, these methods do not fit in well with the theoretical assumption that symptom clusters are discrete and that individual symptoms belong to a single symptom cluster. Instead, with PCA and FA, every single symptom is loaded onto every symptom cluster. As a result, the number of symptoms to include in a cluster needs to be determined ad hoc. This mostly involves researchers selecting a threshold for the loadings on a particular cluster; symptoms with loadings above the threshold will be included in the cluster, symptoms with loadings smaller than the threshold will be dropped from the cluster. An example of this approach can be seen in Fan, Hadi, and Chow (2007). The downside of this approach is that one can never be certain as to what the best threshold is, meaning that symptom clusters constructed this way will have some degree of subjectivity to them as a result of the researcher's decision regarding the threshold.

Another disadvantage that applies to both PCA and FA, as well as HCA, is that all these methods cluster either variables or observations, but never both simultaneously. While symptom cluster identification and patient subgroup identification are important approaches in their own right, combining both of these approaches to identify patients subgroups with a similar symptom cluster experience may be more clinically relevant because it is likely that different subgroups of patients experience different symptom clusters (Barsevick 2016). Especially with regards to long-term cancer survivors, it is likely that many survivors do not longer experience any symptoms or experience some symptoms. Thus, interest should then be in finding subgroups of survivors that do

experience symptoms and the particular symptoms they experience could be different for each subgroup. PCA, FA, and HCA, are unable to produce these kinds of results.

To illustrate the shortcomings of subgroup identification (without simultaneously finding symptom clusters) consider the paper by Kim et al. (2012). In this paper, the authors wanted to find subgroups of breast cancer patients that differed in their symptom experience on an empirically identified psychoneurologic symptom cluster using cluster analyses. One shortcoming is immediately evident, namely that the symptom cluster under investigation has to be determined beforehand, either empirically or qualitatively. Only after selecting a symptom cluster to investigate, can subgroups of patients be identified on the basis of their different scores on the symptoms in the cluster. This highlights the main issue with regular cluster analysis: subgroups are created by considering all of the variables (Chia and Karuturi 2010). Therefore, symptom clusters have to be determined *"a priori"* before subgroups can be identified. The shortcomings of this approach can be addressed by using modern methods, such as biclustering algorithms, that work by simultaneously clustering observations and variables and thus require no *"a priori"* identification of symptom clusters to discover patient subgroups (Chia and Karuturi 2010).

1.7 Data Mining Clustering Algorithms & Symptom Clusters

1.7.1 Previous Research on DM Clustering Algorithms & Symptom Clusters. Because of advances in computing power, ML and DM algorithms have seen an increase in popularity recently. A major advantage of these techniques is that they are extremely flexible and thus capable of fitting very complex data. Previously, these techniques were computationally expensive and required vast amounts of data in order to discover structure in the data. Nowadays, computational issues are less pressing and generally more data is becoming available in the field of cancer research as well. There is an abundance of DM clustering algorithms available that do not have the same limitations as the more traditional statistical methods mentioned above, because of their flexibility and their capability of discovering hidden patterns in the data. The potential use of these DM algorithms in symptom cluster research will be discussed below.

As mentioned before, symptom clusters have only been determined using traditional statistical techniques such as PCA, FA, and HCA. To my knowledge there is only one study that has looked at DM/ML clustering algorithms on cancer symptoms data (Papachristou et al. 2016). In this study, the performance of several DM clustering algorithms was compared with latent class analysis (LCA) on a cancer symptoms data set. Performance was evaluated by the Silhouette coefficient index. The results showed that the performance of at least one DM algorithm, k-Modes, was as good as the clustering produced by LCA. An important difference between these clustering methods and PCA, FA, and HCA is that the former methods cluster patients whereas the latter methods cluster symptoms. Thus, in order to identify symptom clusters from patient clusters an extra step is needed to analyze the symptom profiles of the patients in the different clusters and to examine whether different patient clusters also differ on symptom intensity. However, there are many more DM (bi)clustering algorithms available that cluster variables instead of observations, or that cluster both variables and observations simultaneously and are thus more suitable for symptom cluster identification and can be more readily compared with the symptom cluster findings of PCA, FA, and HCA. Because one of the goals of the current study is to investigate the use of DM algorithms for symptom cluster identification, these methods will be described in more detail in the section below.

1.7.2 DM Clustering Approaches. There are three different approaches to clustering that can be considered for the current project. All of the packages discussed below were made for the R programming language (R Core Team 2017). The first approach involves the `clValid` package (Brock et al. 2008).

The `clValid` package contains functions that allow for evaluating the results of a clustering analysis. The package contains nine different clustering algorithms that the user can select for evaluation. The validation measures that are offered in the package include internal measures (connectivity, silhouette width, and Dunn index) and stability measures (average proportion of non-overlap, average distance, and figure of merit). The outcome of the clValid cluster validation procedure is an overview of the optimal number of clusters for each included clustering algorithm, ranked according to the scores on the internal and validation measures. Thus, a major advantage of this package is that all cluster solutions are automatically evaluated which allows for proper comparison between algorithms.

A disadvantage of the `clValid` package is that some algorithms cluster observations (survivors) rather than variables (symptoms). This makes it harder to compare the results of these algorithms to the results obtained with PCA, FA, and HCA, which are all methods that cluster variables. However, one method to derive symptom clusters from patient clusters is to assign survivors to their respective cluster based on the clustering generated by the algorithm and subsequently calculate the mean scores of the survivors in a particular cluster for all symptoms. A bar plot can then be created with the different symptoms on the x-axis, the mean symptom scores on the y-axis, and the colours of the bars corresponding to the different patient clusters. A visual inspection of this plot could then lead to some insights regarding symptom clusters in the different patient clusters. However, this is an indirect way of finding symptom clusters and thus might not give the best results.

An alternative approach would be to use the `ClustOfVar` package (Chavent et al. 2012). This package was specifically developed for the clustering of variables, as opposed to the clustering of observations, which might be more suitable for the current study. A downside of this package is that only two algorithms are included: an algorithm based on hierarchical cluster analysis and an algorithm based on k-Means. However, because these algorithms cluster variables, the results can be more readily compared to the results produced by PCA, FA, and HCA, and symptom cluster are more directly investigated.

A more advanced approach would be to use biclustering algorithms. Regular clustering algorithms partition either the rows or the columns of a data set. Biclustering algorithms, on the other hand, work by simultaneously partitioning both the rows and the columns. To give an example with regards to symptom clusters, biclustering algorithms could, in theory, discover different symptom clusters for different subgroups of survivors. This method could yield fine-grained information on symptom clusters, and being able to link different symptom clusters to different subgroups of cancer survivors could be highly clinically relevant. R packages that can be used to perform biclustering are the `biclust` package, which includes many different biclustering algorithms, and the `BiclustGUI` package, which implements the algorithms from the `biclust` package and provides a plug-in graphical user interface (GUI) for 'R Commander', which is itself a statistics GUI package for R (De Troyer and Otava 2017; Fox 2005; Kaiser et al. 2018).

The use of biclustering algorithms might provide an answer to one of the most pressing questions in symptom science: Do different symptom clusters appear for differ-

ent subgroups of patients and can we use quantitative methods to finds these subgroups and symptom clusters? Therefore, one goal of this thesis project is to investigate the potential of biclustering algorithms for symptom cluster and patient subgroup identification. Given that biclustering algorithms will be used for the current study, biclustering will be discussed in more detail in the next section.

1.8 Biclustering

Biclustering is a technique that allows for discovery of subgroups of rows and columns in which the members belonging to a certain bicluster are as similar as possible to each other (on a subset of the variables) and are as different as possible to the members of a different bicluster and the members not included in any bicluster. This idea was first proposed by Hartigan (1972) but interest in biclustering grew enormously after the paper by Cheng and Church (2000) on biclustering of gene expression data. After this paper, many more biclustering algorithms were proposed for biclustering of gene expression data (see Padilha and Campello (2017) for a comprehensive overview of many different biclustering algorithms).

The main difference between biclustering methods and regular clustering methods is that for regular clustering methods each row is expressed over all the columns (global pattern) whereas for biclustering each row is expressed over only a subset of the columns (local pattern) (Busygin, Prokopyev, and Pardalos 2008; Chia and Karuturi 2010; De Troyer and Otava 2017). Thus, for traditional clustering, similarity between observations in a cluster is determined by considering all variables, resulting in clusters of observations that are most dissimilar to each other based on all variables. Biclustering, on the other hand, results in clusters of observations that differ from each other on only a subset of the variables, thus leading to potentially more fine-grained results (examples of biclusters can be seen in figure 1 below) (Kasim et al. 2016). Furthermore, methods such as PCA and FA create components or factors that are linear combinations of all the variables, often leading to issues with interpretation, as mentioned before. Biclustering algorithms, on the other hand, are capable finding discrete symptom clusters. That is to say, only those symptoms that are present in a certain subgroup are included in the cluster, all other symptoms are excluded. This leads to symptom clusters that are much easier to interpret compared to symptom clusters that include all symptoms to some extent.

An important reason as to why biclustering algorithms might be especially suited for discovering symptom clusters is that the data used for the current project come from NHL survivors, as opposed to NHL patients (even though some types of NHL are considered incurable). Many previous studies have only looked at symptom data collected from NHL patients that were either still in treatment or shortly after treatment. In contrast, the data collected for the current study come from NHL patients (survivors) whose time since the original diagnosis (in years) ranged from < 2 to > 10, with most survivors showing a time since the original diagnosis between 2 and 5 years. This means that it is highly likely that a large group of these survivors do not longer experience symptoms associated with the disease. Biclustering algorithms, as opposed to traditional clustering algorithms, are able to distinguish between subgroups of survivors that do not experience symptoms (or that experience less symptoms) and subgroups of survivors that do experience symptoms, without *"a priori"* selection of subgroups, because biclustering algorithms look for subgroups of survivors that are similar based on only a subset of the variables (local pattern). For traditional clustering algorithms, on the other hand, clusters are determined by considering all variables, which, in the

case of long-term survivors, is inappropriate because it is likely that a large group of survivors does not longer experience symptoms. Based on this, we would expect traditional clustering algorithms, particularly HCA, to perform poorly (i.e., they would find uninformative clusters), whereas biclustering algorithms should be able to distinguish between subgroups showing no symptoms and smaller subgroups experiencing different symptom clusters. PCA and FA should still be able to find informative symptom clusters but one should keep in mind that these symptoms clusters are then expressed over all survivors, including those who do not longer experience any symptoms.

Let X be an $n \times m$ matrix. Basically, biclustering involves finding smaller matrices in the matrix X where the observations in the smaller matrix are highly similar (De Troyer and Otava 2017). In the context of symptom clusters one can think of n as the survivors (rows) and m as the symptoms (columns). Thus, a_{nm} is the symptom experience of the nth survivor on the mth symptom. The main goal of biclustering in this context is to find subgroups of survivors that have similar symptom experiences on a subgroup of symptoms (i.e., symptom clusters). Thus, biclustering allows for simultaneously discovering subgroups of survivors as well as symptom clusters that are expressed differently for different subgroups of survivors. Naturally, this approach could provide more clinically relevant information regarding symptom clusters and survivor subgroups than other approaches considered so far, such as PCA, FA, and HCA, because these approaches only allow for either subgroup discovery (HCA) or identification of symptom clusters (PCA and FA). Furthermore, symptom clusters found by traditional clustering methods are expressed over all variables and potential differences between patient subgroups are not taken into account.

1.8.1 Bicluster Types. Similarity between observations in a bicluster can be calculated in many different ways. This has led to many different kinds of algorithms that can find different kinds of biclusters. Madeira and Oliveira (2004) distinguish between four types of biclustering algorithms and these algorithms differ with respect to how similarity is defined and thus, which types of biclusters they can find (examples of each type are shown in table 1):

1. **Biclusters with constant values**: Biclusters that have identical values for all rows and columns (top-left matrix in table 1).
2. **Biclusters with constant values on rows or columns**: Biclusters that have either identical values for all rows or identical values for all columns (second and third matrix from top-left in table 1).
3. **Biclusters with coherent values**: Every row or column can be calculated by adding a constant or multiplying by a constant. Finding this type of bicluster cannot be done by using the methods described for the other types above. Most importantly, algorithms that are capable of finding biclusters with coherent values, either in an additive or multiplicative manner, are capable of discovering more complex biclusters than algorithms that are only capable of finding biclusters with constant values (matrices on the second row in table 1).
4. **Biclusters with coherent evolutions**: The exact value of the data points in the matrix is not important because these types of algorithms look for either rows or columns (or both) that show coherent behavior. Coherent behavior can be defined in multiple ways and many different algorithms

have been proposed that all have a different approach to identifying coherent behavior (for an overview see Madeira and Oliveira (2004)). Perhaps the best example of this approach is the xMOTIF algorithm (Murali and Kasif 2003). The goal of this algorithm is to discover conserved gene expression motifs (xMOTIFs). An xMOTIF can be defined as a subset of genes that are in the same state for a subset of conditions. Genes are considered to be in the same state if their expression levels are within a given range and there are assumed to be a fixed number of states (i.e., the data are discretized). Thus, the exact value of a data point (e.g., the expression level of a gene) is only used to determine the state of the gene. This state is then used for further analysis instead of the exact values of the data points (matrices on the third row in table 1) (Murali and Kasif 2003).

This list is in order of complexity, with the least sophisticated biclustering algorithm only being able to find biclusters with constant values and more advanced algorithms being able to find biclusters with coherent values or coherent evolutions as well as biclusters with constant values (Kaiser 2011). The bicluster algorithms that look for biclusters with coherent evolutions are the most suitable for the current project because questionnaire data can be viewed as consisting of discretized items. That is to say, when answers are provided to a questionnaire item on a scale ranging from 1 to 4, with 1 referring to "not at all" and 4 referring to "very much", these answers can be interpreted as reflecting a certain state. Because the numerical difference between 1 and 2, and 2 and 3 is difficult to interpret (i.e., one cannot say with certainty that the actual difference between 1 and 2 is the same as the difference between 2 and 3, even though the numerical difference is 1 in both cases), it might be more useful to interpret these numerical values as representing certain states, where the interest lies in distinguishing between subgroups of patients that experience symptoms in different states.

1.8.2 Bicluster Structures. Biclustering algorithms can either assume that only one bicluster exists, or alternatively, that K biclusters exist (although most algorithms that look for only one bicluster can also be adapted to find more than one bicluster). Madeira and Oliveira (2004) distinguish between different types of bicluster structures that can be found using biclustering algorithms (see figure 1 for examples of each structure):

1. Single bicluster (a).
2. Exclusive row and column biclusters (b).
3. Checkerboard structure (c).
4. Exclusive row biclusters (d).
5. Exclusive column biclusters (e).
6. Non-overlapping biclusters with tree structure (f).
7. Non-overlapping non-exclusive biclusters (g).
8. Overlapping biclusters with hierarchical structure (h).
9. Arbitrarily positioned overlapping biclusters (i).

To provide a concrete example with respect to biclustering of symptom data, consider matrix c in figure 1. In this matrix, three distinct subgroups of patients (rows)

Table 1
Bicluster types (based on Madeira and Oliveira (2004))

constant values - overall

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

constant values - rows

1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

constant values - columns

1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0

coherent values - additive

1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

coherent values - multiplicative

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

coherent evolution - overall

S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1

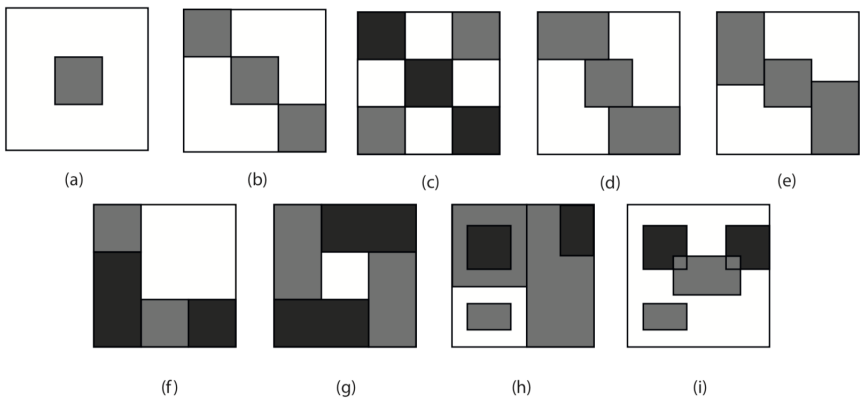
coherent evolution - rows

S1	S1	S1	S1
S2	S2	S2	S2
S3	S3	S3	S3
S4	S4	S4	S4

coherent evolution - columns

S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4

Figure 1
Bicluster structures. Discovered biclusters are colored as grey or black submatrices within the white data matrix, with a darker color indicating higher values (based on Madeira and Oliveira (2004)).



can be seen and three distinct symptom clusters (columns). For every patient subgroup, their expression for each symptom cluster may differ. For example, one patient subgroup may score high on symptom cluster 1 but low on the other symptom clusters and another patient subgroup may score high on symptom clusters 2 and 3 and low

on the other symptom clusters. This could yield valuable clinical information for each subgroup of patients.

However, keep in mind that this checkerboard example is for illustrative purposes only because, in reality, not every patient will be included in a bicluster and some biclusters might overlap, especially given that the patients included in the dataset that will be used for the current project are cancer survivors. This means that a large group of these survivors might no longer suffer from any symptoms related to their disease. Because of the large variety of bicluster types and bicluster structures, it is vital that the proper algorithms are chosen that best fit the structure of the data (Padilha and Campello 2017). This will be discussed in the next section.

1.8.3 Biclustering of Questionnaire Data. Because many biclustering algorithms were developed in the domain of biological data analysis, specifically for use with gene expression data, not all biclustering algorithms are suitable for questionnaire data (ordinal data) (Kaiser 2011). Given that the data for the current project were collected through a questionnaire (EORTC QLQ-C30, see section 2.3), in this section, biclustering of questionnaire data will be discussed.

In the field of symptom science, symptom data are mostly gathered through questionnaires where different items represent different symptoms. Patients are often asked to indicate on a scale from, for example, 1 to 4, to what extent they experience a certain symptom. This results in ordinal data. One issue with traditional clustering methods when applied to questionnaire data is that they would typically require a large sample size because of the ordinal nature of the data and the number of variables that are used (Kaiser 2011). For example, in the field of market segmentation (which has a similar goal to symptom cluster research, namely identification of subgroups that show similar patterns on certain variables), a sample size of 70 times the number of variables is advised (Dolnicar et al. 2014). This large sample size is necessary because, as the dimensionality of the data increases (i.e., the number of variables increases), the sample size needs to increase as well because otherwise the discovered groupings are random and no real patterns can be found. In data mining, this problem has been referred to as the 'curse of dimensionality' (Domingos 2012). Simply put, in high dimensional data (i.e., a dataset with a large number of variables) all data points are highly dissimilar to each other, which prevents meaningful groupings to be discovered. This would mean that, for the EORTC QLQ-C30 consisting of 30 items, a sample size of $70 \times 30 = 2100$ would be required. These large sample sizes are not typically found in the field of symptom science.

Furthermore, traditional clustering algorithms take into account all variables simultaneously. This means that all questionnaire items are weighted equally and this might not accurately reflect reality because it might be the case that some items are important only for a certain subgroup of patients whereas a completely different set of items might be important for another subgroup. Thus, whereas biclustering algorithms are able to discover local patterns, traditional clustering algorithms are only capable of discovering global patterns, leading to a loss of information when traditional clustering algorithms are used. As discussed above, biclustering algorithms can be used to deal with these issues. Below, two biclustering algorithms suitable for questionnaire data will be discussed.

1.8.4 Bimax Algorithm. One biclustering algorithm that is suitable for questionnaire data is the Bimax (binary inclusion-maximal biclustering) algorithm (Prelić et al. 2016). One downside is that the algorithm only works on binary data. Of course, information

on the severity or intensity of the symptom experience is then lost. However, given that many survivors included in the dataset for the current project might not experience many symptoms, this algorithm might still be useful because it provides knowledge on whether or not specific subgroups of survivors experience certain symptoms. The results of the algorithm could be highly clinically relevant because they would allow clinicians to categorize patients in two or more subgroups: either the patient does not experience any (or few) symptoms or the patient experiences multiple symptoms and thus belongs to one of the subgroups of patients that experience different symptom clusters.

The Bimax algorithm works by looking for submatrices (biclusters) that consist of only elements that have a value of 1, no zeroes are allowed. Biclusters that consist of only one element and smaller biclusters that can also be included in a larger bicluster are excluded (Kasim et al. 2016). With respect to the bicluster structures that the Bimax algorithm can find, the original Bimax algorithm is able to discover row overlapping biclusters whereas the repeated Bimax algorithm leads to each survivor belonging to only one symptom cluster. The repeated Bimax algorithm also solves an issue of the regular Bimax algorithm, namely that the regular Bimax algorithm returns too many small and uninteresting biclusters and that large biclusters are often missed. In the repeated Bimax algorithm the regular Bimax algorithm is run first and after that, all biclusters that can also be contained in other biclusters (i.e., a submatrix of ones that can be contained in a larger submatrix of ones) are thrown away. An advantage of the repeated Bimax algorithm, in the context of symptom science, is that each row (survivor) can only belong to one bicluster (subgroup) because the rows (survivors) that already belong to a previously found bicluster are removed. This means that the repeated Bimax algorithm leads to non-overlapping biclusters with respect to the rows (an example of this bicluster structure can be seen in figure 1, image d). Also, variables (symptoms) can belong to multiple biclusters. This is appropriate for symptom data because it might be that two subgroups both experience symptoms related to fatigue, but one subgroup also experiences symptoms related to depression in addition to the fatigue symptoms. This kind of information on differences between subgroup has high clinical relevance. For more details on the repeated Bimax algorithm see Dolnicar et al. (2011) and Kaiser (2011).

Dolnicar et al. (2011) used the repeated Bimax algorithm in a market segmentation study of tourism data to find subgroups of tourists that participated in specific activities on their holidays, based on binary questionnaire data. They found 11 different subgroups of tourists that participated in different activities which could then be used for more specific targeting of marketing approaches. Additionally, these subgroups also differed with respect to demographic variables, such as number of domestic holidays and number of days spent on the last domestic vacation, thus providing more evidence for the distinctness of the discovered subgroups.

Extending this approach to symptom data, one can imagine multiple distinct subgroups of patients, with each subgroup showing different symptoms. This could potentially lead to a more personalized treatment approach because specific symptoms that occur in certain subgroups can be targeted specifically. Another advantage of the Bimax algorithm is that it also shows for each subgroup which variables were not included in the bicluster. This could provide information on which specific symptoms certain subgroups of patients do not experience and thus do not need to be addressed[1]. Another

biclustering algorithm specifically developed for questionnaire data, Questmotif, will be discussed below.

.

1.8.5 Questmotif Algorithm. The Questmotif algorithm, developed by Kaiser (2011), is based on the xMOTIF algorithm that was also discussed in section 1.8.1 above, and where an xMOTIF can be defined as a subset of genes that are in the same state for a subset of conditions. (Murali and Kasif 2003). First, a detailed overview of the xMOTIF algorithm will be given. After that, the Questmotif algorithm will be discussed.

The xMOTIF algorithm is an iterative algorithm that tries to find the largest xMOTIF (bicluster) in every iteration. Murali and Kasif (2003) provide the following definition of an Xmotif:

> "Given a set of genes whose expression levels are measured across a set of samples and user defined parameters $0 < \alpha, \beta < 1$, a conserved gene expression motif or xMOTIF is a pair (C, G), where C is a subset of the samples and G is a subset of the genes, that satisfies the following condition:
>
> - **Size**: the number of samples in C is a least an α-fraction of all the samples.
> - **Conservation**: every gene in G is conserved across all the samples in C, i.e., the gene is in the same state in all the samples in C, and
> - **Maximality**: for every gene not in G, the gene is conserved in at most a β-fraction of the samples in C" (p. 79).

For every iteration, the samples in the data that belong to a bicluster are removed, then the algorithm looks for the biggest bicluster in the remaining data. This continues until every data point belongs to a bicluster. The algorithm can also be adapted so that the samples belonging to a particular bicluster are not removed, thereby making it possible for samples to belong to multiple biclusters. Another advantage of the xMOTIF algorithm is that genes (symptoms) can belong to multiple biclusters.

The Questmotif algorithm is an adaptation of the xMOTIF algorithm and works similarly (Kaiser 2011). The Questmotif algorithm can handle three different types of data that are commonly obtained through questionnaires: 1) nominal data, 2) ordinal data, and 3) continuous data. In a symptom science context, the Questmotif algorithm can be used to look for subgroups of survivors that differ with respect to symptom experience. This means that the Questmotif algorithm is able to simultaneously discover subgroups of survivors as well as symptom clusters. With respect to nominal data, the Questmotif algorithm is equal to the xMOTIF algorithm. With respect to ordinal data, the Questmotif algorithm has an important user-defined parameter, d, which represents an interval that determines to what extent answers to an item are considered as similar and which differentiates it from the xMOTIF algorithm. For example, imagine data collected on a 1 to 6 scale and the value of a specific data point (e.g., respondent 1, item 1) being 4. With $d = 1$, all other respondents that have the same value or that have a value of $4 - 1 = 3$ or $4 + 1 = 5$, for the specific item, are considered as being similar to each other (Kaiser 2011). This concept is somewhat comparable to the concept of the state of a gene for the xMOTIF algorithm. It should be noted that, because rows

1 Although it could be the case that a small group of patients in a certain subgroup experiences symptoms that are not included in the biclusters. This is due to how the Bimax algorithm constructs its biclusters, which will be discussed in section 4.2

are removed after they are included in a bicluster, this Questmotif algorithm cannot be used to find row overlapping biclusters. More details on the other parameters of the Questmotif algorithm can be found in section 2.5 below.

Given that the Questmotif algorithm was specifically developed for questionnaire data, it would seem to be particularly suitable for symptom data, which is often gathered through questionnaires. No other papers have looked at the use of biclustering algorithms, such as Bimax and Questmotif, on symptom data. This paper will try to add to the existing literature by investigating the potential of using biclustering algorithms on symptom data gathered from NHL survivors.

1.9 Research Question

To investigate the use of biclustering algorithms for finding symptom clusters and subgroups in NHL survivors, the following research question will be addressed in this paper:

> **Research Question:** Can biclustering algorithms, such as the repeated Bimax algorithm and Questmotif algorithm, be used for symptom cluster and subgroup identification based on symptom data collected from long-term NHL survivors, and, how do these biclusters compare to the clusters generated with PCA and HCA?

2. Materials and Method

2.1 PROFILES Registry

The data used for this project comes from the PROFILES registry (van de Poll-Franse et al. 2011). PROFILES stands for Patient Reported Outcomes Following Initial Treatment and Long term Evaluation of Survivorship and the registry was created in order to facilitate collection of patient reported outcomes and to be able to link this with clinical data. This makes it possible to relate PROs, such as the results of the EORTC QLQ-C30, to medical and demographic data. The medical and demographic data come from the Eindhoven Cancer Registry (ECR), which collects data on all persons newly diagnosed with cancer in the southern part of the Netherlands (van de Poll-Franse et al. 2011). Data in the PROFILES registry is collected both by electronic means through the internet as well as by traditional pen-and-paper surveys. Pen-and-paper collection methods are still included because a purely internet-based collection method could potentially result in an age and socioeconomic status (SES) bias (van de Poll-Franse et al. 2011). The data from the PROFILES registry can be freely accessed by researchers through the PROFILES website after permission has been granted. A more detailed description of the PROFILES registry can be found in van de Poll-Franse et al. (2011).

2.2 Patients

For the current project, two data sets from the PROFILES registry concerning Dutch patients diagnosed with any form of NHL were used. These data sets will be referred to as NHL_1 and NHL_2, respectively. The data in NHL_1 come from patients diagnosed with NHL between 1999 and 2008 in the southern region of the Netherlands and included 1062 patients. The data in NHL_2 come from 326 patients diagnosed with NHL between 2007 and 2009 as well as patients that were already included in NHL_1 and who were asked to fill in the survey again. Thus, NHL_2 includes new patients as well as patients

already present in NHL_1. For patients in NHL_2 that had already filled in a questionnaire once, only their first entry was kept. The time since diagnosis (in years) varied between between 0 and more than 10 (see also table 3). After filtering out non-responders, the number of responders and the response rate (between brackets) for the different data sets were as follows: NHL_1: 715 (67.33%), NHL_2: 120 (36.81%). Also, three patients were actually diagnosed with HL and were also filtered out. Subsequently, 24 respondents with more than 30% missing values on the variables were also filtered out. Combining both NHL data sets resulted in one NHL data set which will be used for further analyses (NHL, $n = 808$).

2.3 Instruments

The instruments used in this study included a demographic questionnaire, the EORTC QLQ-C30, and the EORTC QLQ CLL-17 (Aaronson et al. 1993; van de Poll-Franse et al. 2017). All questionnaires were administered in Dutch. Both the EORTC QLQ-C30 and the EORTC QLQ CLL-17 can be found in appendix A. The specific items used for the analyses are shown in table 2.

The demographic questionnaire included questions on gender, marital status, education level, and employment status. Other information available on patients was retrieved from patients' medical records, facilitated through the PROFILES registry. This included information on body mass index (BMI), age at the time the questionnaire was administered, SES, treatment type, years since diagnosis, and disease stage.

To evaluate quality of life and investigate symptom experience, the EORTC QLQ-C30 (version 3.0) was administered to all patients. The EORTC QLQ-C30 has a 1-week time frame and is a cancer-specific 30-item questionnaire consisting of five functional scales (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, and nausea and vomiting), and a global health and quality of life scale. Furthermore, there are additional single items that address other symptoms that are commonly experienced by cancer patients (dyspnea, appetite loss, sleep disturbance, constipation, diarrhea, and financial impact of the disease) (Aaronson et al. 1993). The fact that the EORTC QLQ-C30 has multiple scales that measure different (functional) aspects of HRQOL ties in well with the literature showing that HRQOL is necessarily a multi-faceted construct (Arden-Close, Pacey, and Eiser 2010). The mean of the items is interpreted as representing the intensity of the symptom experience and functioning.

All items in the questionnaire, with the exception of the global health and quality of life scale, have four response options: 1) not at all; 2) a little bit; 3) quite a bit; 4) very much. Patients are asked to provide a response on each item based on how they experienced the issue related to a particular item in the past week, with higher scores indicating worse functioning and higher symptom intensity. The global health and quality of life scale items have seven response options ranging from 1 (very poor) to 7 (excellent), with higher scores indicating a higher HRQOL. Symptoms did not need to be present in a certain percentage of patients to be included in the analyses.

Because the EORTC QLQ-C30 is a general measure (i.e., it can be used for patients with different types of cancer), it might not be as sensitive to symptoms and functioning in specific types of cancer, such as NHL. To increase this sensitivity to aspects of symptom experience and functioning specific to patients with lymphoma, the EORTC QLQ-CLL17 was developed (European Organisation for Research and Treatment of Cancer). The EORTC QLQ-CLL17 is a 17-item questionnaire, that can be used in conjunction with the EORTC QLQ-C30, consisting of 4 multi-item scales (fatigue, treatment side effects, disease effects, infection), and 3 single items (social problems, future health,

tingling hands or feet). The response options and scoring procedure are similar to those described above for the EORTC QLQ-C30.

For the current study, not all items included in the EORTC QLQ-C30 and QLQ-CLL17 were used. The decisions below were made on the advice of an expert in the field of symptom science and who has carried out multiple studies using the EORTC QLQ-C30. The item numbers below refer to the numbers of the questionnaire included in appendix A. Even though some scales in the QLQ-C30 are called functional scales, the items belonging to some of these scales can still be considered to represent symptoms. Therefore, the items belonging to the cognitive functioning scale (items 20 and 25) and the items belonging to the emotional functioning scale (items 21, 22, 23, and 24) were regarded as individual symptoms. Furthermore, the physical functioning scale (items 1, 2, 3, 4, and 5), the role functioning scale (items 6 and 7), the social functioning scale (items 26 and 27), the global health and quality of life scale (items 29 and 30), and the financial impact of the disease item (item 28) were excluded from the analyses. The physical, role, and social functioning scales were excluded because the items belonging to these scales are either not considered to represent individual symptoms or they represent symptoms in a wholly different domain than the other items in the questionnaire. The financial impact of the disease item was excluded because this item is considered as too different from the other symptoms included in the questionnaire and many other papers that used the EORT QLQ-C30 also excluded this item (Giesinger et al. 2016).

The global health and quality of life scale was excluded because previous research has shown that this scale is not sensitive enough to accurately capture HRQOL in cancer patients (Giesinger et al. 2016). Instead, Giesinger et al. (2016) argue that a summary score based on all items and scales of the EORTC QLQ-C30, excluding the financial impact item and the global health and quality of life scale, provides a more robust measure of HRQOL. This multi-faceted approach to HRQOL is also in line with findings from previous studies that show that HRQOL is necessarily a multi-faceted construct (Arden-Close, Pacey, and Eiser 2010). Thus, for the current project, the summary score proposed by Giesinger et al. (2016) will be used as a measure of HRQOL. With respect to the EORTC QLQ-CLL17, item 42 was excluded because this item is very similar to the role functioning scale in the EORTC QLQ-C30. Furthermore, items 44, 45, 46, and 47 were combined in an infection scale. In accordance with the scoring manual, raw scores were calculated by adding up the scores for all the items belonging to a particular scale and dividing by the total number of items of the scale. Eventually, 28 symptoms were considered for the analyses.

2.4 Pre-Processing & Missing Data Imputation

Initially, the full dataset consisted of more than 250 variables. These included the aforementioned demographic variables, the EORTC QLQ-C30, the EORTC QLQ-CLL17, as well as many more questionnaires on various topics (such as personality) that were not used for the current project. The first step in the pre-processing process consisted of filtering out the variables that were not used for further analyses. After that, non-responders and patients with more than 30% missing values on the variables were filtered out.

Missing data imputation was carried out using the `missMDA` package in R (Josse and Husson 2016). `missMDA` is a package that is designed for handling missing values in multivariate data analysis. Among other things, it allows for single imputation of an incomplete dataset consisting of mixed variables (i.e., both categorical and continuous

Table 2
EORTC QLQ-C30 and QLQ-CLL17 items used in current study. The item numbers refer to the questionnaire included in appendix A.

QLQ-C30		
Variable name	**Item Nr.**	**Question**
DY	8	Were you short of breath?
PA	9, 19	Have you had pain? Did pain interfere with your daily activities?
FA	10, 12, 18	Did you need to rest? Have you felt weak? Were you tired?
SL	11	Have you had trouble sleeping?
AP	13	Have you lacked appetite?
N	14	Have you felt nauseated?
V	15	Have you vomited?
CO	16	Have you been constipated?
DI	17	Have you had diarrhea?
C30_q20	20	Have you had difficulty in concentrating on things, like reading a newspaper or watching television?
C30_q21	21	Did you feel tense?
C30_q22	22	Did you worry?
C30_q23	23	Did you feel irritable?
C30_q24	24	Did you feel depressed?
C30_q25	25	Have you had difficulty remembering things?
QLQ-CLL17		
CLL_q1	31	Did you lose weight?
CLL_q2	32	Did you have a dry mouth?
CLL_q3	33	Did you experience bruising?
CLL_q4	34	Did you have an unpleasant feeling in your stomach?
CLL_q5	35	Did your temperature rise and fall?
CLL_q6	36	Did you sweat at night?
CLL_q7	37	Have you had skin issues?
CLL_q8	38	Did you feel sick or unwell?
CLL_q9	39	Did you feel apathetic?
CLL_q10	40	Did you feel washed out?
CLL_q11	41	Have you had tingling hands or feet?
CLL_q13	43	Did you worry about your future health?
CLL_INF	44, 45, 46, 47	Did you experience respiratory infections? Did you experience any other infections? Did you repeatedly need antibiotic treatment? Did you worry that you would incur an infection?

variables) based on principal component methods. Because the data for the current project consisted of both continuous (symptom variables) and categorical variables (demographic variables), factorial analysis for mixed data (FAMD), available in the `missMDA` package, was used because this method is capable of dealing with both data types simultaneously. While there has been some debate whether survey data based on variants of a Likert scale should be considered continuous or categorical, most of the symptom cluster literature has considered data gathered using questionnaires such as the EORTC QLQ-C30 and the EORTC QLQ-CLL17 as continuous. Furthermore, when survey data are considered as continuous, the distance between the values is preserved with respect to imputation, which is not the case when the survey data would be considered categorical. Therefore, the approach taken by other studies in the symptom science field was also used for the current project.

With respect to categorical variables, FAMD works by first transforming every categorical variable into dummy variables. Then, every dummy variable is divided by the square root of the proportion of observations that belong to a particular category of the associated variable. Continuous variables, similar to PCA, are standardized (Josse and Husson 2016). After this, PCA methods are carried out on this newly created matrix. Basically, FAMD is a mix of PCA and multiple correspondence analysis (MCA): PCA is used for the continuous variables and MCA is used for the categorical variables. In fact, FAMD is equivalent to PCA when only continuous variables are considered and it is equivalent to MCA when only categorical variables are considered (Josse and Husson 2016).

The `imputeFAMD()` function in the `missMDA` package takes the incomplete dataset as an argument, as well as the number of dimensions (components), `S`, that will be used for imputation and it returns the completed dataset with the missing values imputed. The optimal number of components to use can be found using the `estim_ncpFAMD()` function, which takes as arguments the incomplete dataset, the minimum number of components to consider, and the maximum number of components to consider. It returns the number of components, `S`, that minimizes the mean squared error of prediction (MSEP), based on one of two cross-validation methods: leave-one-out, and *k*-fold cross-validation. For example, with leave-one-out cross-validation every value of the dataset is removed once and for every number `S` the value is predicted using the FAMD method that is based on the dataset that excludes the particular value. Then, the prediction error is computed and this process is repeated for every value in the dataset and for all number of components. Finally, the `S` that results in the smallest MSEP is returned and should be used as an argument to the `imputeFAMD()` function. The *k*-fold cross-validation method functions similarly, except that it involves removing multiple values in the dataset simultaneously, as opposed to just one at a time for leave-one-out cross-validation (Josse and Husson 2016).

For the current project, the number of components was determined using the *k*-fold cross-validation method with the minimum number of components set to 0 and the maximum number of components set to 10. This method indicated 10 as the optimal number of components to use for imputation. Thus, the missing values in the incomplete dataset were imputed using the `imputeFAMD()` function with the number of components set to 10. Eventually, the complete dataset that was used for further analyses consisted of 28 symptom variables, 10 demographic variables and 808 observations.

2.5 Questmotif Biclustering

The Questmotif biclustering algorithm used for the current study is implemented in the R package `biclust` (Kaiser et al. 2018). As mentioned in section 1.8.5 above, the Questmotif algorithm was developed specifically for questionnaire data. The dataset used for the biclustering analyses consisted of all symptom variables, demographic variables were excluded. The `BCQuest()` function in the `biclust` package takes several arguments including:

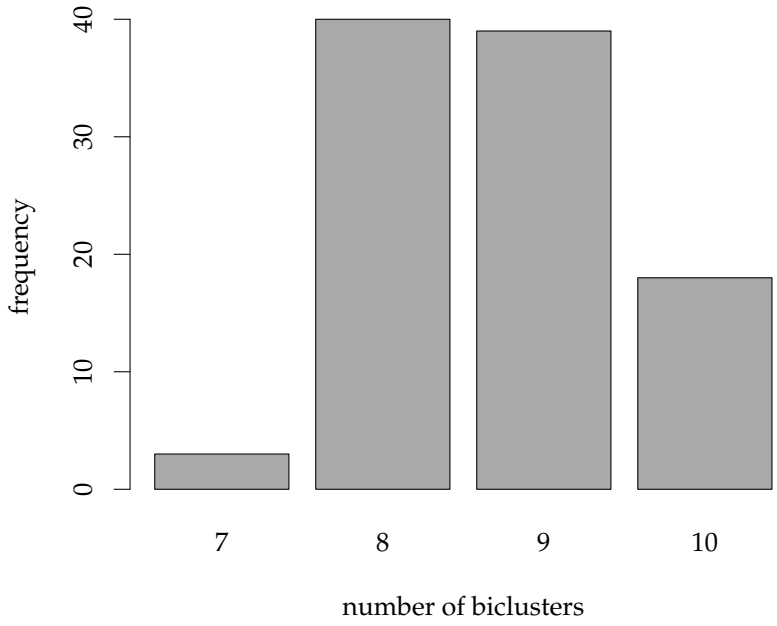
- **ns**: Number of questions chosen
- **nd**: Number of repetitions
- **sd**: Sample size in repetitions
- **d**: Half margin of interval that the question values should be in (interval is mean - d, mean + d)
- **number**: The maximum number of biclusters to be found

Because biclustering algorithms in general are highly sensitive to small changes in parameter settings, different values of these arguments often lead to a different number of biclusters or biclusters of different sizes. Furthermore, the Questmotif algorithm also depends on the starting value, meaning that different random seeds lead to different biclusters. Therefore, one run of a biclustering algorithm generally does not lead to a stable result (Kaiser 2011). This is shown in figure 2, which shows the distribution of the number of biclusters for 100 different runs of the Questmotif algorithm, with the same parameters used for every run. In order to address this issue, an ensemble method was proposed to construct so called superbiclusters (Tatsiana 2013; Pfundstein 2010; Kaiser 2011).

2.5.1 Superbiclusters. The ensemble method for constructing superbiclusters from a set of biclusters is implemented in the R package `superbiclust` and in the R commander plug-in `BiclustGUI` (Tatsiana 2014; De Troyer and Otava 2017). The basic idea behind the ensemble method as it was used for the current project is that the Questmotif algorithm is run multiple times on the data, using different parameter settings. The bicluster results from the different runs are then combined and a similarity matrix based on the Jaccard index is constructed. The Jaccard index, in this context, can be interpreted as a measure of similarity between two sets of biclusters. Subsequently, the hierarchy of the biclusters based on the similarity is calculated and a dendogram is drawn showing the hierarchy. Then, based on the dendogram, the robust biclusters can be obtained by cutting the tree at a certain height or by specifying a certain number of clusters. Combining the corresponding biclusters then results in the formation of superbiclusters (i.e., robust biclusters).

To give an example with respect to the Questmotif algorithm, one might use the ensemble method to obtain robust biclusters when the value of the `d` parameter is varied between 0 and 2 in steps of 0.2. Each algorithm with the various parameter values is then run multiple times on the data, eventually returning the robust biclusters. For a more detailed description of the `superbiclust` package and the implementation of the ensemble method in R, see Tatsiana (2013) and De Troyer and Otava (2017).

Figure 2
Distribution of the number of biclusters



For the current project, only a limited number of different parameter values was investigated because the main objective of this paper is to investigate the use of biclustering algorithms on symptom data and not to conduct an in-depth investigation into the effects of different parameter settings on the bicluster results. Kasim et al. (2016) advise to carry out parameter tuning on the data in order to find meaningful parameter value combinations. Initial parameter tuning showed that values of `d` between 1 and 1.5 and values of `nd` between 15 and 20 seemed to generate meaningful biclusters with clear differences between the biclusters. Therefore, six values of `d` between 1 and 1.5 in steps of 0.1 and six values of `nd` between 15 and 20 in steps of 1 were used in the ensemble method. In order to decrease the number of biclusters, not all parameter combinations were considered. To give an example, `d` = 1 was run in combination with `nd` = 15, `d` = 1.1 was run in combination with `nd` = 16, and so on, resulting in 6 parameter combinations. Moreover, in order to account for the seed dependency of the algorithm, each parameter combination was run 100 times, resulting in $6 \times 100 = 600$ total runs. The `number` parameter was set to 10 in order to decrease the number of biclusters, all other parameters were set to their default value.

An individual run of the algorithm was executed as follows:

```
biclust(as.matrix(NHL), method = 'BCQuestord', quant = 0.25,
vari = 1, d = 1, ns = 10, nd = 10, sd = 5, alpha = 0.05,
number = 10)
```

2.6 Repeated Bimax Biclustering

The repeated Bimax biclustering algorithm used for the current study is implemented in the R package `biclust` (Kaiser et al. 2018). The dataset used is the same as the dataset used with the Questmotif algorithm except that the data were binarized with values > 1 set to 1 and values ≤ 1 set to 0.

The repeated Bimax algorithm is not seed-dependent and thus the same biclusters are found every time the algorithm is run with the same parameters. As is also the case with the Questmotif algorithm, different parameter settings lead to different bicluster results. However, contrary to the Questmotif algorithm, the parameters for the repeated Bimax algorithm are more straightforward and proper values can be determined by considering one's research question and research goal. Therefore, it makes less sense to use the ensemble method with the repeated Bimax algorithm because interest is not necessarily in the robust biclusters that result from the ensemble method. Rather, interest is in the individual results generated by the repeated Bimax algorithm with a limited number of different parameter settings. That is to say, it could be clinically relevant to know what the difference is between biclusters that contain at least 50 survivors as opposed to biclusters that contain at least 25 survivors. Forming robust biclusters from these results is not as interesting because potentially clinically relevant information might then be lost. The `BCrepBimax()` function in the `biclust` package takes several arguments including:

- **minr**: Minimum row size of resulting bicluster
- **minc**: Minimum column size of resulting bicluster
- **number**: Number of biclusters to be found
- **maxc**: Maximum column size of resulting bicluster

The `minr` and `minc` parameters are important because these have a large effect on the number and size of biclusters that the algorithm discovers (Dolnicar et al. 2011; Kasim et al. 2016). Decreasing the minimum row size (`minr`) and minimum column size (`minc`) often leads to a higher number of biclusters of a smaller size whereas increasing `minr` and `minc` often leads to a smaller number of biclusters of a larger size. That is to say, when the algorithm is run with small values for `minr` and `minc`, smaller subgroups are more likely to be found and when the algorithm is run with larger values for `minr` and `minc`, larger subgroups are more likely to be found. As of yet, no method has been developed that could be used to decide optimal values for `minr` and `minc`.

Because there are many survivors in the current dataset that experience few to no symptoms, the value for `minr` was kept relatively small in order to be able to find small subgroups of survivors that share a similar symptom experience. Three different values for `minr` were used: 25, 35, and 50. The value of 25 was chosen because subgroups smaller than 25 are likely not clinically relevant. The other values were chosen to investigate the effect of this parameter on the bicluster solution and to see whether clinically meaningful symptom clusters can still be discovered that apply to a larger percentage of the sample. The value for `minc` was set to 2 in every run because, as mentioned in section 1.5 above, the minimum number of symptoms to occur simultaneously in order to speak of a symptom cluster is two. The `maxc` parameter was set to 12 because

symptom clusters that include a large number of symptoms can be difficult to interpret. The analyses were executed using the following command (with `minr` set to either 25, 35, or 50):

```
biclust(NHL.b, method = 'BCrepBimax', minr = 25, minc = 2,
number = 30, maxc = 12)
```

In order to test for differences with respect to demographic variables and HRQOL between survivors included in the biclusters and survivors not included in the biclusters, Fisher's exact test, Welch t-test, and the one-way analysis of variance (ANOVA) were used.

2.7 PCA and HCA

To be able to compare the findings of the bicluster algorithms with the findings of more traditional clustering methods, PCA and HCA were applied to the dataset.

PCA with varimax rotation was performed on the symptom data using the `principal()` function from the R package `psych` (Revelle 2018). The number of components to extract was determined with the `fa.parallel()` function from the psych package, which runs a parallel analysis to compare the scree of the components of the actual data with the scree of simulated data with the same size as the actual dataset and returns an optimal number of components to extract. The components to retain are those components (based on the actual data) whose eigenvalues are larger than those of the components extracted from the simulated data. The resulting plot is shown in figure 3 and the suggested number of components to extract was three. The internal consistency of the resulting symptom clusters was estimated with Cronbach's α. The varimax rotation was used in order to increase interpretability of the resulting symptom clusters since a varimax rotation usually has the effect of decreasing the value of the loadings of some variables so that only a few variables with large loadings remain.

HCA was performed on the symptom data using the `hclust()` function from the `stats` package which is included in base R, with mean linkage clustering as the agglomeration method and the Euclidean distance as the distance measure (R Core Team 2017). The mean linkage agglomeration method was used because this is the most commonly used agglomeration method in general and because is it also often used in research on symptom science.

The analyses were executed using the following commands:

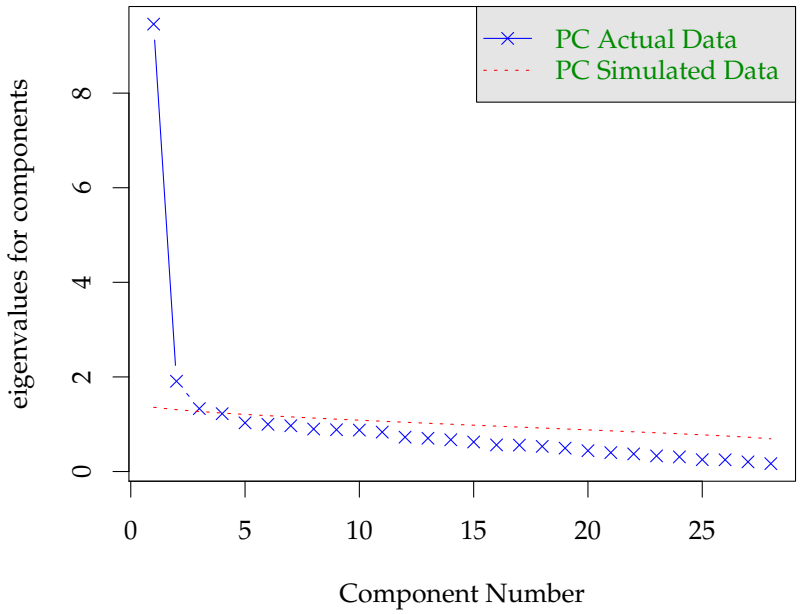
```
principal(NHL.m, nfactors = 6, rotate = 'varimax')
```

```
hclust(dist(NHL.t), method = 'average')
```

3. Results

Demographic information about the 808 survivors included in the dataset is shown in table 3. As can be seen in table 3, the number of survivors with a time since diagnosis of over 10 years is small (12 survivors, 1.5% of the total sample). This is likely due to the recency of the data. The largest group of survivors has a time since diagnosis between 2 and 5 years.

Figure 3
Parallel analysis scree plot

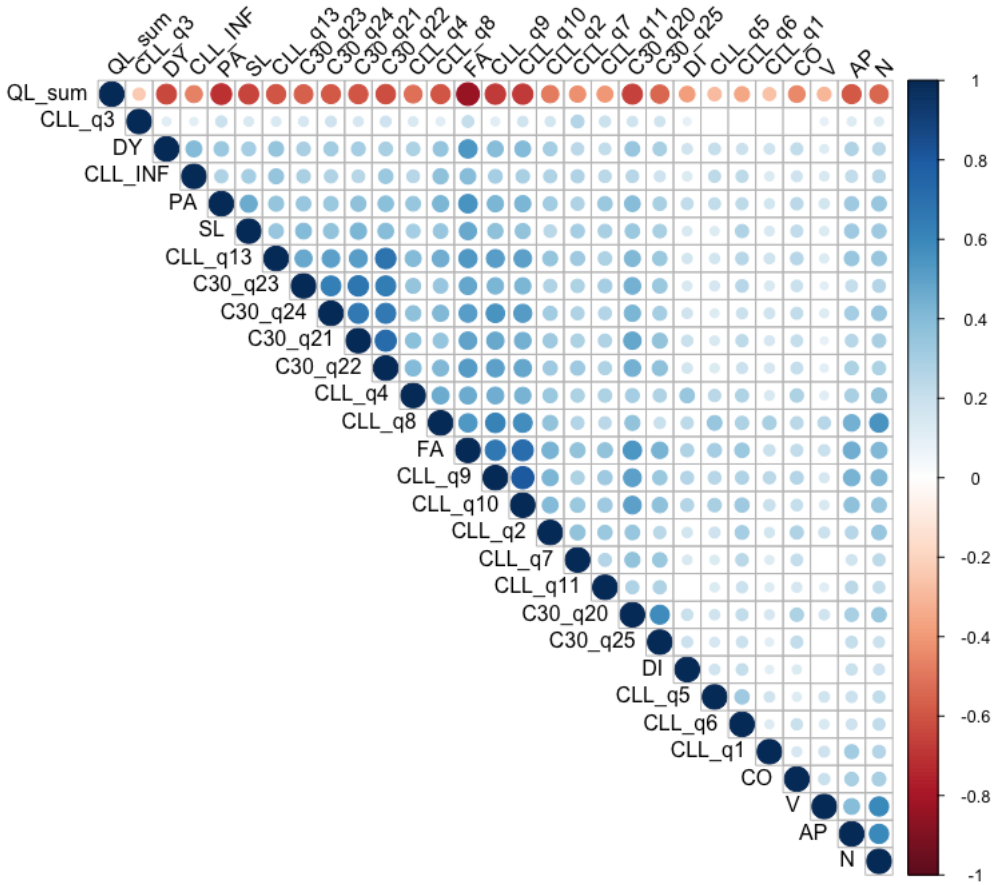


A plot of the correlations between the symptoms and HRQOL can be seen in figure 4. A red color indicates a significant negative correlation and a blue color indicates a positive correlation (at the 0.01 level of significance). Blank squares in the figure indicate non-significant correlations. As can be seen in figure 4, all symptoms are significantly negatively correlated with HRQOL and most variables are positively correlated with each other. A strong negative correlation (> 0.8) was found between HRQOL and fatigue (FA). Furthermore, moderately strong correlations (> 0.5) were found between HRQOL and symptoms related to shortness of breath (DY), pain (PA), trouble sleeping (SL), appetite loss (AP), nausea (N), vomiting (V), emotional health (C30_q21, C30_q22, C30_q23, C30_q24, and CLL_q13), cognitive functioning (C30_q20, C30_q25), and sickness (CLL_q8).

Moderately strong positive correlations (> 0.5) were found between symptoms related to emotional health (C30_q21, C30_q22, C30_q23, C30_q24, and CLL_q13). This is unsurprising given that these symptoms formed an emotional functioning scale in the QLQ-C30 (with the exception of CLL_q13). Strong positive correlations (> 0.65) can also be seen between symptoms related to fatigue (FA, CLL_q9, and CLL_q10). Interestingly, moderately strong positive correlations (> 0.42) were also also found between symptoms pertaining to emotional health and fatigue, potentially suggesting that these symptoms are related. Moderately strong positive correlations (> 0.5) were also observed between symptoms related to nausea (N) and vomiting (V), and between symptoms related to nausea (N) and appetite loss (AP).

In section 3.1, the interpretation of the bicluster result plots will be discussed. In section 3.1.1, the discovered biclusters will be discussed in more detail. Section 3.1.2 focuses on one particular Bimax model and the subgroups found by the model are presented in more detail. In section 3.1.3, the differences between subgroups discovered

Figure 4
Correlation plot



by the Bimax model with relation to HRQOL and other demographic variables will be discussed. In section 3.2 and 3.3, the results from the PCA and HCA are presented. Finally, in section 3.4, the results from the Questmotif algorithm are discussed.

3.1 Repeated Bimax

Biclusters generated by the repeated Bimax algorithm with different parameter settings are shown in figures 5, 6, and 7. The different subgroups are shown on the x axis and the different symptoms are shown on the y axis. Fields that contain colored rectangles indicate symptoms that all survivors belonging to this subgroup have in common. Thus, the colored fields show symptom clusters for the different subgroups. The color in the middle of a rectangle represents the mean value of this variable across all survivors, ranging from 0 (red) to 1 (green). This means that a variable that has a rectangle with a red color in the middle indicates that the mean value of this variable is low in the overall sample but that the mean value of this variable is high for a particular subgroup. Thus,

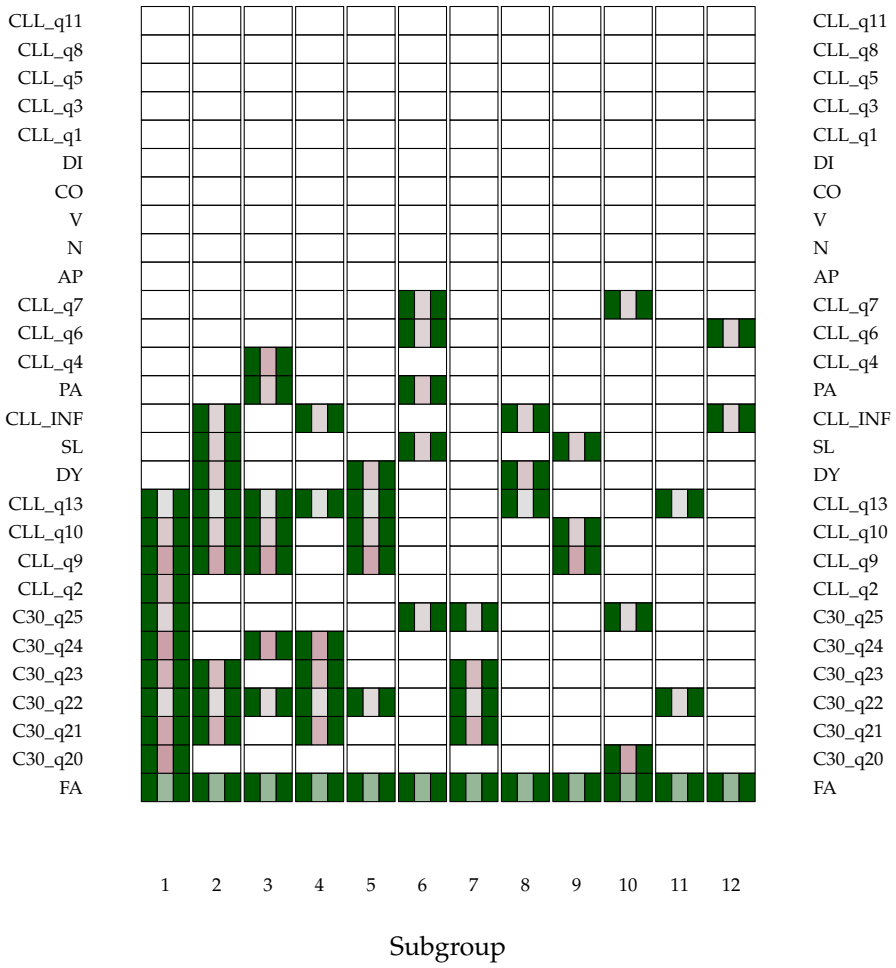
Table 3
Survivor demographics ($N = 808$)

Variable		Variable	
Age (in years); mean (range)	64 (21-85)	**BMI; mean (range)**	26.4 (17.2-43.6)
Gender	N (%)	**Type**	N (%)
Male	505 (62.5)	CLL	164 (20.3)
Female	303 (37.5)	Agressive NHL	360 (44.6)
Years since diagnosis	N (%)	Indolent NHL	251 (31.1)
< 2	210 (26)	Other NHL	33 (4.1)
\geq 2 and < 5	315 (39)	**Paid job**	N (%)
\geq 5 and < 10	271 (33.5)	Yes	219 (27.1)
\geq 10	12 (1.5)	No	589 (72.9)
SES	N (%)	**Marital Status**	N (%)
Low	160 (19.8)	Married / Cohabiting	650 (80.4)
Medium	321 (39.7)	Divorced / Separated	42 (5.2)
High	314 (38.9)	Widowed	75 (9.3)
Living in care institution	13 (1.6)	Never married / Never cohabited	41 (5.1)
Treatment	N (%)	**Education**	N (%)
Wait and see	215 (26.6)	Higher education	177 (21.9)
Chemotherapy	394 (48.8)	Medium education	502 (62.1)
Radiotherapy	74 (9.2)	Lower education	129 (16)
Transplantation	3 (0.4)		
Radio and Chemo	109 (13.5)		
Other therapies	11 (1.4)		
Surgery	2 (0.2)		

the variables that have rectangles with a red color are the most interesting because these variables distinguish the subgroup from the entire sample (Dolnicar et al. 2011).

Variables that have rectangles with a green color in the middle, on the other hand, are less interesting with respect to subgroup identification because these represent symptoms that are experienced by the majority of the sample. For example, in all figures, we can see that the FA (fatigue) variable has a green color in the middle of the rectangle for all subgroups. This indicates that fatigue is a symptom that is experienced by survivors in every subgroup and that it has a high mean value in the entire sample. Even though this is a highly clinically relevant finding, because the mean value for this symptom is high for the entire sample it is not helpful in distinguishing between the different subgroups. Interestingly, the color of the center of the rectangles of most of the other symptoms is closer to red, indicating that these symptoms have a low mean value in the overall sample and that the symptoms are unique for the discovered subgroups. That is to say, survivors in the discovered subgroups experience these particular symptoms to a higher extent than the other survivors in the overall sample. Table 5 shows the subgroup sizes and the symptoms experienced by each subgroup for the different parameter settings (detailed information on variable names can be found in table 2 above).

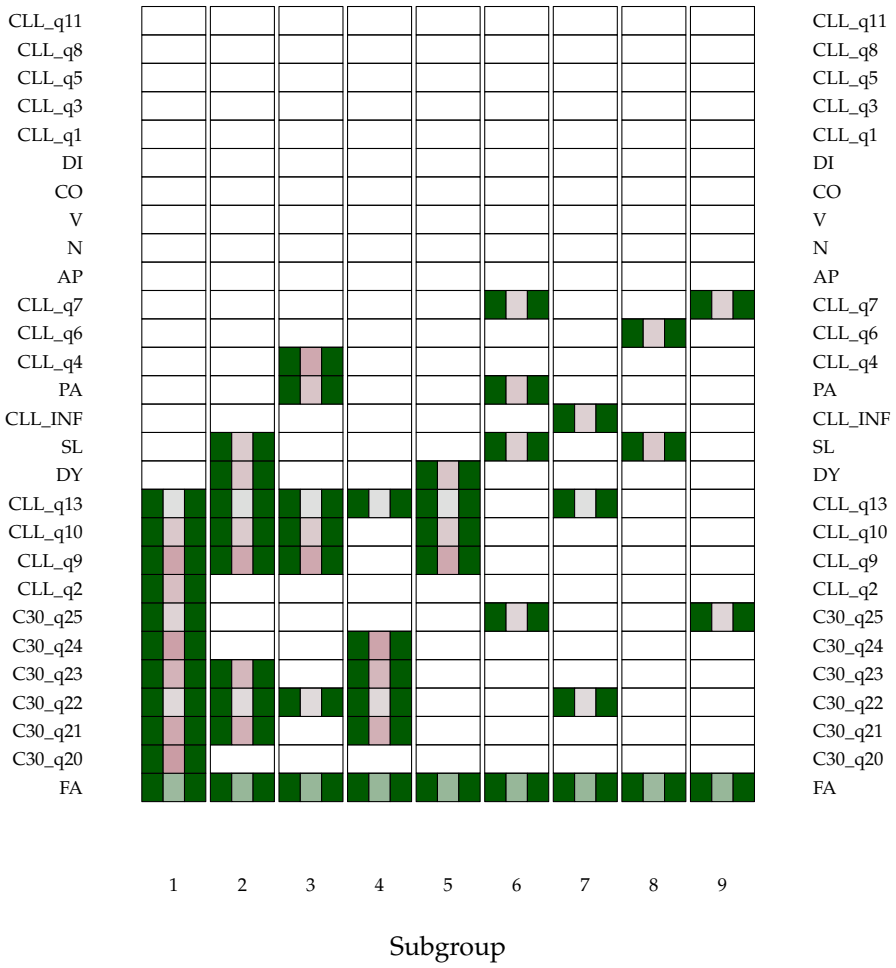
Figure 5
Model 1: Repeated Bimax Biclustering with minr = 25



3.1.1 Repeated Bimax Biclusters. The biclusters generated by the repeated Bimax algorithm with the `minr` parameter set to 25 (**model 1**) are shown in figure 5. In total, 429 survivors were clustered, which amounts to roughly 53% of the total sample. The results show 12 different subgroups of survivors that differ with respect to the symptoms they experience. The biclusters generated by the repeated Bimax algorithm with the `minr` parameter set to 35 (**model 2**) are shown in figure 6. In total, 398 survivors were clustered, which amounts to roughly 49% of the total sample. The results show 9 different subgroups of survivors that differ with respect to the symptoms they experience. The biclusters generated by the repeated Bimax algorithm with the `minr` parameter set to 50 (**model 3**) are shown in figure 7. In total, 378 survivors were clustered, which amounts to roughly 47% of the total sample. The results show 6 different subgroups of survivors that differ with respect to the symptoms they experience.

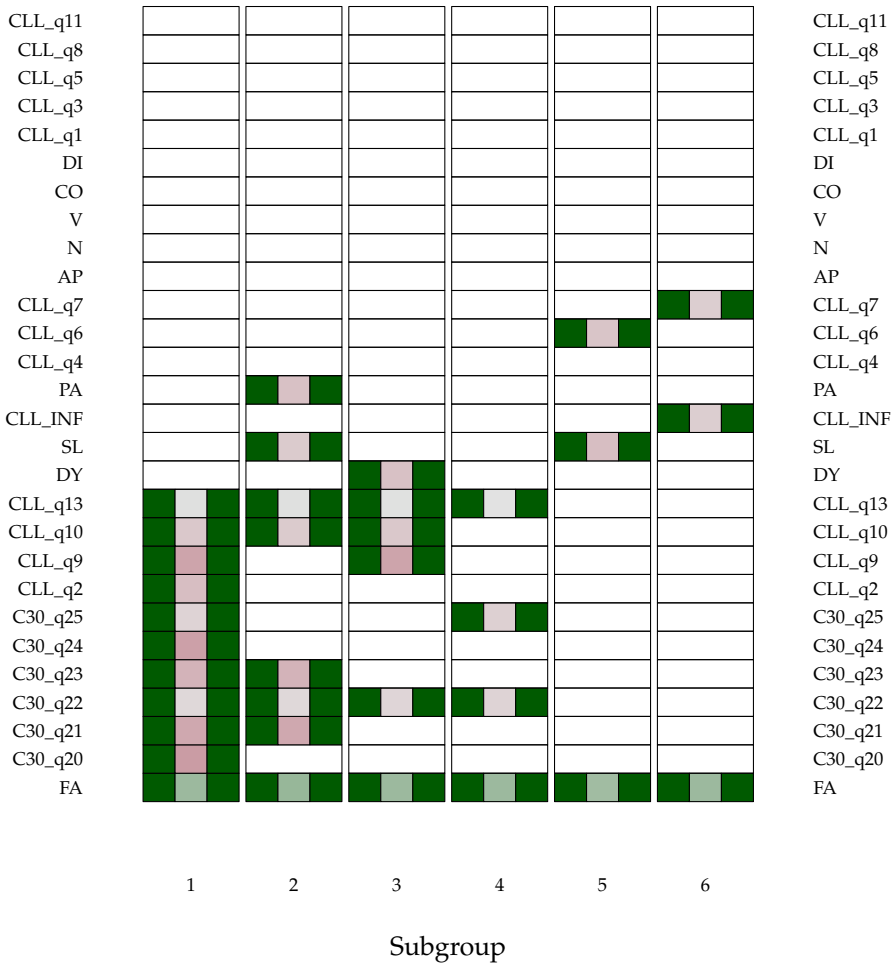
Crucially, the most important finding is that different symptom clusters are found for different subgroups of patients, providing promising evidence that biclustering

Figure 6
Model 2: Repeated Bimax Biclustering with minr = 35



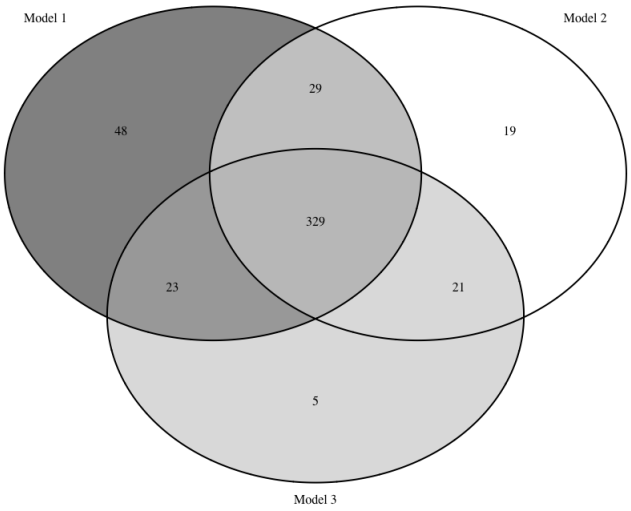
algorithms are highly suitable for discovering subgroups and symptom clusters from symptom data and that different subgroups of survivors experience different symptom clusters. As expected, a greater value for `minr` leads to a smaller number of biclusters of a larger size and a smaller value for `minr` leads to a larger number of biclusters of a smaller size. As can be seen in the Venn diagram in figure 8, the three different models show a high degree of overlap in terms of the survivors that are clustered. More precisely, 329 survivors were included in all three different models. 48 survivors are unique to model 1, 19 survivors are unique to model 2, and 5 survivors are unique to model 3. In terms of overlap in the symptoms that were clustered, 17 symptoms (see table 5) were included in each model. One symptom (CLL_q4: Did you have an unpleasant feeling in your stomach) was included only in models 1 and 2. The following 10 symptoms were not included in any of the biclusters: AP, N, V, CO, DI, CLL_q1, CLL_q3, CLL_q5, CLL_q8, CLL_q11.

Figure 7
Model 3: Repeated Bimax Biclustering with minr = 50



Because of the large overlap in survivors and symptoms between the different models (suggesting highly similar models), in the following section, the results of model 2 will be discussed in more detail. Model 2 was chosen for further analyses because this model shows a good balance between the number of survivors included per subgroup (at least 5% of the total sample is included in each subgroup), and the number of symptoms included per subgroup. Model 1 found more subgroups than model 2 but since some of these subgroups are relatively small, the clinical relevance of this model is limited. Model 3, on the other hand, found a smaller number of subgroups, but of a larger size. However, because of the small number of subgroups, more fine-grained information on differences between survivors might be lost. Thus, model 2 provides a compromise between subgroups that are large enough to be clinically relevant and, at the same time, small enough to capture subgroups with unique symptom patterns.

Figure 8
Overlap in the survivors clustered by the different models



3.1.2 Repeated Bimax Model 2. In this section, the bicluster results from the repeated Bimax model 2 are discussed in more detail. In the next section, differences between subgroups discovered by this model, with relation to HRQOL and other demographic variables, will be discussed.

As mentioned before, the results of model 2 can be seen in figure 6 and table 5. In total, 398 survivors were clustered, which amounts to roughly 49% of the total sample. Importantly, this also means that 410 survivors were not included in any of the subgroups. This finding can be interpreted in three different ways: 1) a large number of survivors do not longer significantly experience any symptoms several years after the original diagnosis, or 2) these survivors experience symptoms that lead to a less severe symptom experience, or 3) these survivors experience less symptoms overall.

Further analyses of the survivors not included in any of the subgroups of model 2 (410 survivors) showed that 50 survivors (12.2%) did not experience any symptoms. Furthermore, 274 survivors (66.8%) experienced 5 symptoms or less. On average, the survivors not included in any of the subgroups experienced 4 symptoms. This is in stark contrast to the survivors included in the biclusters, who experienced, on average, 14 symptoms.

Below, the most important findings of the biclustering results of model 2 are discussed in more detail:

- **Subgroup 1:** Survivors in subgroup 1 (54 survivors) experience symptoms related to fatigue (FA, CLL_q9, CLL_q10), concentration (C30_q20), emotional health (C30_q21 through C30_q24, CLL_q13), memory (C30_q25), and having a dry mouth (CLL_q2).
- **Subgroup 2:** Subgroup 2 (40 survivors) is somewhat similar to subgroup 1 with respect to some of the symptoms relating to emotional health and fatigue. However, one symptom measuring a particular aspect of emotional health is not present (C30_q24, which asks about depression),

indicating that the emotional health of this subgroup might be slightly better than that of subgroup 1. Furthermore, subgroup 2 is distinguished by additional symptoms related to shortness of breath (DY) and trouble sleeping (SL).

- **Subgroup 3:** Survivors in subgroup 3 (41 survivors) also seem to experience symptoms related to fatigue (FA, CLL_q9, CLL_q10), similar to subgroup 1, 2, and 5. However, only two symptoms relating to emotional health are present for this subgroup (C30_q22, CLL_q13). Furthermore, two additional symptoms are present in this subgroup: one relating to pain (PA) and one relating to unpleasant feelings in the stomach (CLL_q4).
- **Subgroup 4:** Subgroup 4 (43 survivors) is similar to subgroup 1 with respect to symptoms related to emotional health (C30_q21 through C30_q24, CLL_q13). In fact, apart from one symptom relating to fatigue (FA), these are the only symptoms present in this subgroup, indicating that the symptoms experienced by this subgroup are not of a physical nature but rather a psychological one.
- **Subgroup 5:** The survivors in this subgroup (45 survivors) seem to present mostly with symptoms related to fatigue (FA, CLL_q9, CLL_q10) with two additional symptoms relating to shortness of breath (DY) and worrying (CLL_q13). In fact, this subgroup shares some similarities with subgroup 2 to the extent that both subgroups experience symptoms related to fatigue in combination with shortness of breath. However, survivors in subgroup 2 present with more symptoms pertaining to emotional health, which are mostly absent in subgroup 5.
- **Subgroup 6:** Subgroup 6 (37 survivors) is a somewhat unique subgroup of survivors that experience symptoms related to fatigue (FA), memory (C30_q25), trouble sleeping (SL), pain (PA), and skin issues (CLL_q7).
- **Subgroup 7:** The survivors in this subgroup (41 survivors) present with symptoms relating to worrying (C30_q22, CLL_q13), infections (CLL_INF), and fatigue (FA).
- **Subgroup 8:** Survivors in subgroup 8 (55 survivors) experience symptoms related to trouble sleeping (SL), fatigue (FA), and sweating at night (CLL_q6). It is interesting to note that two symptoms (SL, CLL_q6), both relating to sleep in some way, cluster together for this subgroup.
- **Subgroup 9:** Survivors in subgroup 9 (42 survivors) experience symptoms related to fatigue (FA), memory (C30_q25), and skin issues (CLL_q7).

One symptom that all subgroups have in common is FA (fatigue). Often, other symptoms relating to fatigue are also present (CLL_q9, CLL_q10), indicating that fatigue is still a major issue for survivors, even long after the original diagnosis. Furthermore, other symptoms that often appear together in the different subgroups are C30_q21, C30_q22, C30_q23, C30_q24, and CLL_q13. This is perhaps not surprising since these symptoms originally formed the emotional functioning scale in the EORTC QLQ-C30, with the exception of CLL_q13. Nevertheless, the fact that these items are included in the bicluster results indicates that many survivors experience symptoms

relating to emotional health (e.g., depression) and worrying. Subgroups 1, 2, and 4 all show activation of these symptoms.

Interestingly, further investigation of the biclusters showed that the first subgroup (bicluster) of model 2, consisting of 54 survivors and 11 symptoms, was similar for all models. That is to say, this exact bicluster, with the same survivors and symptoms, was consistently found by the repeated Bimax algorithm, even with different parameter settings. This finding, in addition to the large degree of overlap between the survivors in the bicluster results of the different models, seems to indicate that the findings are stable.

3.1.3 Repeated Bimax Model 2 Subgroup Differences. The group of survivors not included in any of the biclusters of model 2 will be referred to as subgroup 0. Unless noted otherwise, subgroup 0 was not included in any of the following analyses. Table 4 shows the mean and standard deviation of the HRQOL score for the different subgroups discovered by model 2. Figure 9 shows the mean HRQOL score plotted by subgroup. Demographics for the various subgroups can be found in appendix B.

Significant differences were found between subgroups with respect to some demographic variables, including gender (Fisher's exact test p-value < 0.01), years since diagnosis (Fisher's exact test p-value = 0.027), NHL type (Fisher's exact test p-value = 0.044), and age (ANOVA p-value = 0.033) (see also appendix B). These results provide some external validity and provide additional support for the claim that the discovered subgroups are meaningfully different.

Crucially, a significant difference was also found with respect to HRQOL between survivors included in the biclusters and survivors not included in any of the biclusters (subgroup 0). Results of the Welch two sample t-test showed a significant difference in mean HRQOL score between the two groups: t(584.81) = -22.712, p < 0.001, 95% CI [-19.913, -16.743]. The mean HRQOL score of survivors included in the biclusters (M = 75.44, SD = 14.40) was significantly lower than the HRQOL score of survivors not included in any of the biclusters (M = 93.77, SD = 7.31). This suggests that survivors which are not clustered by the repeated Bimax algorithm seem to suffer from less symptoms, or symptoms that have less impact on HRQOL, as evidenced by their higher HRQOL. In contrast, survivors included in the biclusters show an impaired HRQOL, likely due to their more severe symptom experience.

In order to test for differences in HRQOL score between survivors included in the different models, multiple t-tests were conducted. No significant differences were found in HRQOL score between survivors included in model 1 (M = 76.13, SD = 14.38) and model 2 (t(820.22) = 0.696, p = 0.487, 95% CI [-1.269, 2.663]), model 1 and model 3 (M = 75.17, SD = 14.54, t(789.97) = 0.943, p = 0.346, 95% CI [-1.041, 2.966]), and model 2 and model 3 (t(771.08) = 0.256, p = 0.798, 95% CI [-1.775, 2.307]). This indicates that HRQOL is similar for all survivors included in the different models (without taking into account differences between subgroups).

The results of a one-way ANOVA showed a significant difference with respect to mean HRQOL score between the different model 2 subgroups (survivors from subgroup 0 were excluded): F(8, 389) = 33.52, p < 0.001. A Tukey post hoc test revealed the following results:

- The mean HRQOL score was lower for subgroup 1 compared with every other subgroup (p < 0.01).

- The mean HRQOL score was lower for subgroup 2 compared with subgroup 4, subgroup 5, subgroup 7, subgroup 8, and subgroup 9 ($p < 0.001$).
- The mean HRQOL score was lower for subgroup 3 compared with subgroup 7, subgroup 8, and subgroup 9 ($p < 0.001$).
- The mean HRQOL score was lower for subgroup 4 compared with subgroup 9 ($p < 0.001$).
- The mean HRQOL score was lower for subgroup 5 compared with subgroup 9 ($p < 0.001$).
- The mean HRQOL score was lower for subgroup 6 compared with subgroup 7 ($p < 0.01$), subgroup 8 ($p = 0.014$), and subgroup 9 ($p < 0.001$).
- No significant difference in mean HRQOL score was found between subgroup 7 and subgroup 8, subgroup 7 and subgroup 9, and subgroup 8 and subgroup 9.

These results provide additional support for the claim that the discovered subgroups are meaningfully different. Moreover, they indicate that the mean HRQOL score varies from as low as 57.75 (subgroup 1) to as high as 88.61 (subgroup 9). This shows that even within survivors that experience symptoms, large differences exist between their reported HRQOL. Furthermore, these results also suggest that some symptoms could have less impact on HRQOL than other symptoms. For example, the symptoms experienced by survivors in subgroup 9 include symptoms related to infections and skin issues. One can imagine that these symptoms impact HRQOL differently than symptoms related to depression and fatigue, which are experienced by survivors in subgroups with a lower mean HRQOL score.

3.2 PCA

The results of the PCA with varimax rotation are shown in figure 10. Cronbach's α was 0.88 for cluster 1, 0.75 for cluster 2, and 0.84 for cluster three. Variance explained was 19% for cluster 1, 11% for cluster 2, and 16% for cluster three, resulting in a total variance explained of 45%. Only symptoms with an absolute loading value ≥ 0.5 were considered to be part of a cluster.

As can be seen in figure 10, the first symptom cluster consists of symptoms relating to emotional health (C30_q21, C30_q22, C30_q23, C30_q24, CLL_13), concentration (C30_q20) and fatigue (FA, CLL_q9). The second cluster consists of symptoms relating to lack of appetite (AP), nausea (N), vomiting (V), and feeling sick or unwell (CLL_q8). This is interesting since further investigation of these symptoms showed that 682 survivors (84%) did not experience issues with nausea, 763 survivors (94%) did not experience issues related to vomiting, 681 survivors (84%) did not experience issues relating to a lack of appetite, and 650 survivors (80%) did not experience issues related to feeling sick or unwell. The third symptom cluster is somewhat unclear since the loadings of many variables are close to the threshold of 0.5 and no variables have a loading larger than 0.6. Based on the threshold of 0.5, symptom cluster 3 consists of symptoms related to shortness of breath (DY), fatigue (FA), dry mouth (CLL_q2), body temperature fluctuations (CLL_q5), sweating at night (CLL_q6), and feeling washed out (CLL_q10).

Table 4
Mean HRQOL score per subgroup

Subgroup (N)	Mean (SD)
Subgroup 0 (410)	93.77 (7.31)
Subgroup 1 (54)	57.75 (16)
Subgroup 2 (40)	67.02 (10.84)
Subgroup 3 (41)	72.27 (10.92)
Subgroup 4 (43)	77.56 (9.97)
Subgroup 5 (45)	77.7 (10.89)
Subgroup 6 (37)	73.93 (10.4)
Subgroup 7 (41)	84.09 (8.8)
Subgroup 8 (55)	82.28 (11.97)
Subgroup 9 (42)	88.61 (6.43)

Overall, the symptoms in symptom cluster 1 are somewhat similar to symptoms experienced by subgroup 1 and subgroup 4 of the repeated Bimax results from model 2 and form a fatigue and emotional health symptom cluster. The symptoms in symptom cluster 2 form a sickness symptom cluster because the symptoms are related to nausea, vomiting, and feeling sick. Furthermore, this symptom cluster is surprising since these symptoms do not appear for any of the subgroups of Bimax model 2. There does not seem to be a pattern unifying the symptoms in symptom cluster 3. That is to say, while the symptoms in cluster 1 and 2 can all be seen as belonging to the same underlying construct, this is not the case for the symptoms in symptom cluster 3. Also note that, because PCA can only be used to find symptom clusters, no subgroups can be derived from the PCA results [2].

Since symptoms related to nausea, vomiting, and appetite loss were not included in the bicluster results, further analyses of the data were carried out. The results showed that, overall, 127 survivors experienced symptoms related to appetite loss, 126 survivors experienced symptoms related to nausea, and 45 survivors experienced symptoms related to vomiting. Of the 127 survivors that experienced symptoms related to appetite loss, 109 survivors (85.8%) were included in the biclusters and 18 survivors (14.2%) were not included in the biclusters. With respect to survivors that experienced symptoms related to nausea, 110 survivors (87.3%) were included in the biclusters and 16 survivors (12.7%) were not included in the biclusters. Finally, of the 45 survivors that experienced symptoms related to vomiting, 32 survivors (71.1%) were included in the biclusters

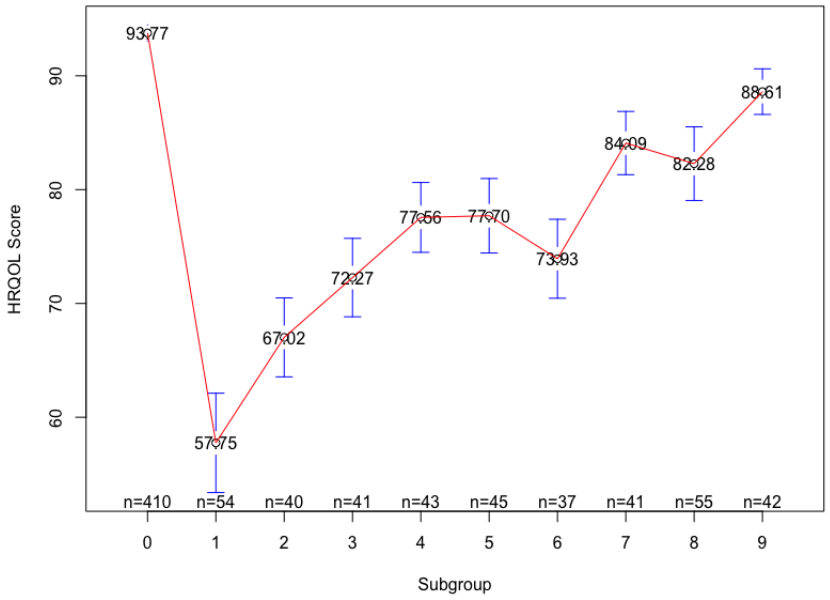
Table 5
Subgroup sizes and symptoms experienced by the different subgroups

minr = 25	**Size (% of total sample)**	**Symptoms**
Subgroup 1	54 (6.7)	FA, C30_q20, C30_q21, C30_q22, C30_q23, C30_q24, C30_q25, CLL_q2, CLL_q9, CLL_q10, CLL_q13
Subgroup 2	32 (4)	FA, C30_q21, C30_q22, C30_q23, CLL_q9, CLL_q10, CLL_q13, DY, SL, CLL_INF
Subgroup 3	30 (3.7)	FA, C30_q22, C30_q24, CLL_q9, CLL_q10, CLL_q13, PA, CLL_q4
Subgroup 4	28 (3.5)	FA, C30_q21, C30_q22, C30_q23, C30_q24, CLL_q13, CLL_INF
Subgroup 5	36 (4.5)	FA, C30_q22, CLL_q9, CLL_q10, CLL_q13, DY
Subgroup 6	29 (3.6)	FA, C30_q25, SL, PA, CLL_q6, CLL_q7
Subgroup 7	31 (3.8)	FA, C30_q21, C30_q22, C30_q23, C30_q25
Subgroup 8	43 (5.3)	FA, CLL_q13, DY, CLL_INF
Subgroup 9	35 (4.3)	FA, CLL_q9, CLL_q10, SL
Subgroup 10	26 (3.2)	FA, C30_q20, C30_q25, CLL_q7
Subgroup 11	50 (6.2)	FA, C30_q22, CLL_q13
Subgroup 12	35 (4.3)	FA, CLL_INF, CLL_q6

minr = 35	**Size (% of total sample)**	**Symptoms**
Subgroup 1	54 (6.7)	FA, C30_q20, C30_q21, C30_q22, C30_q23, C30_q24, C30_q25, CLL_q2, CLL_q9, CLL_q10, CLL_q13
Subgroup 2	40 (5)	FA, C30_q21, C30_q22, C30_q23, CLL_q9, CLL_q10, CLL_q13, DY, SL
Subgroup 3	41 (5.1)	FA, C30_q22, CLL_q9, CLL_q10, CLL_q13, PA, CLL_q4
Subgroup 4	43 (5.3)	FA, C30_q21, C30_q22, C30_q23, C30_q24, CLL_q13
Subgroup 5	45 (5.6)	FA, CLL_q9, CLL_q10, CLL_q13, DY
Subgroup 6	37 (4.6)	FA, C30_q25, SL, PA, CLL_q7
Subgroup 7	41 (5.1)	FA, C30_q22, CLL_q13, CLL_INF
Subgroup 8	55 (6.8)	FA, SL, CLL_q6
Subgroup 9	42 (5.2)	FA, C30_q25, C30_q7

minr = 50	**Size (% of total sample)**	**Symptoms**
Subgroup 1	54 (6.7)	FA, C30_q20, C30_q21, C30_q22, C30_q23, C30_q24, C30_q25, CLL_q2, CLL_q9, CLL_q10, CLL_q13
Subgroup 2	50 (6.2)	FA, C30_q21, C30_q22, C30_q23, CLL_q10, CLL_q13, PA, SL
Subgroup 3	57 (7.1)	FA, C30_q22, CLL_q9, CLL_q10, CLL_q13, DY
Subgroup 4	72 (8.9)	FA, C30_q22, C30_q25, CLL_q13
Subgroup 5	86 (10.6)	FA, SL, CLL_q6
Subgroup 6	59 (7.3)	FA, CLL_q7, CLL_INF

Figure 9
Plot of HRQOL means by subgroup (Repeated Bimax Model 2)



and 13 survivors (28.9%) were not included in the biclusters. The implications of these findings will discussed in section 4.2.

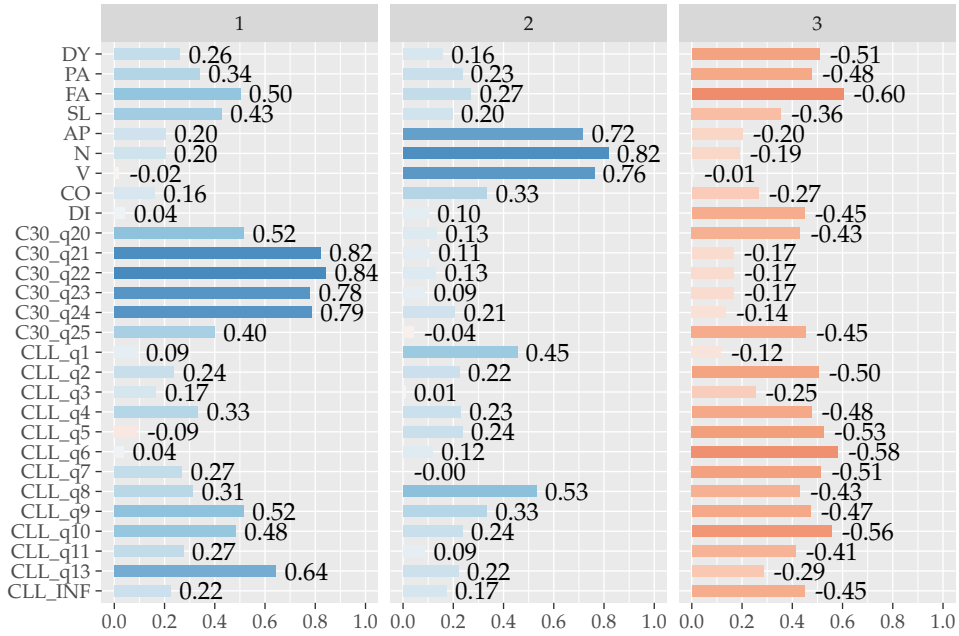
3.3 HCA

The results of the HCA are shown in the form of a dendogram in figure 11. Depending on where we cut the tree and how we interpret the results, two to four symptom clusters can be identified. In the right hand side of the dendogram we can see, somewhat similar to symptom cluster 1 in the PCA results and subgroup 1 and 4 of Bimax model 2, that the symptoms related to emotional health (C30_q21, C30_q22, C30_q23, C30_q24) are clustered together before they are joined with symptoms relating to fatigue (FA, CLL_q9, CLL_q10), forming symptom cluster 1. Even more higher up in the dendogram they are eventually joined with symptoms relating to pain (PA), memory (C30_q25), and concentration (C30_q20). This cluster of PA, C30_q25, and C30_q20 could be seen as a separate symptom cluster, or it could be combined with the symptoms in symptom cluster 1 to form a larger cluster.

To the left of this cluster, we see another potential symptom cluster. Interestingly, similar to symptom cluster 2 of the PCA results, nausea (N) and vomiting (V) are

2 It should be noted here that it is possible to derive subgroups from PCA results in an ad hoc fashion. This is a two-step procedure which, in this context, would first involve performing a PCA on the symptom data and subsequently performing a clustering of the component scores of the survivors on the different components.

Figure 10
Rotated component loadings

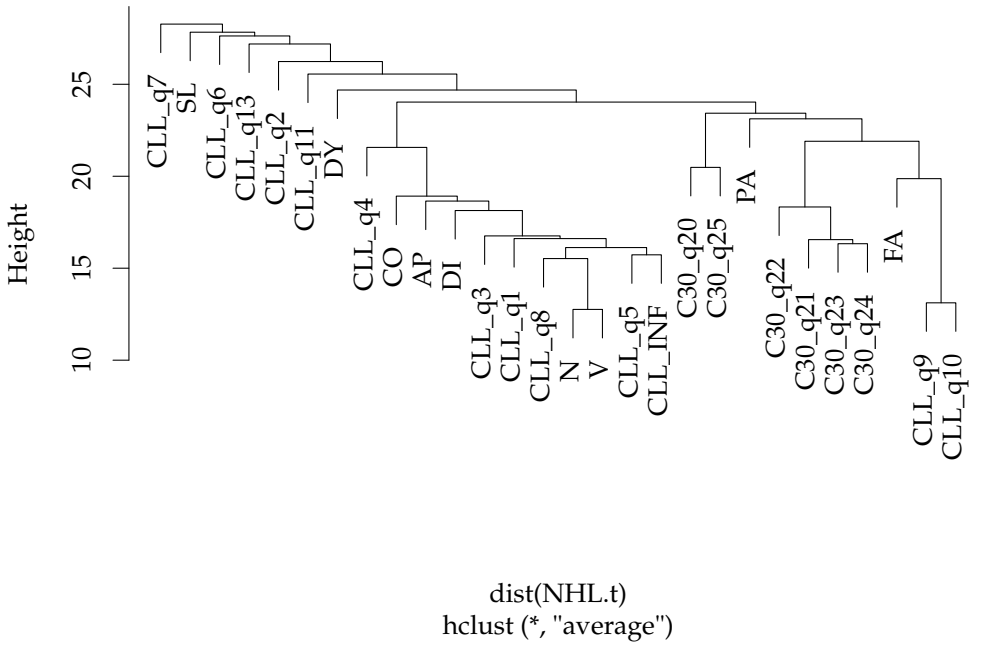


clustered together and they are joined with a symptom related to feeling sick or unwell (CLL_q8). A symptom related to fluctuations in body temperature (CLL_q5) and a symptom related to infections (CLL_inf) are clustered together before being joined with the symptoms related to nausea, vomiting, and feeling sick. Slightly higher up in the tree, these symptoms are joined with symptoms relating to weight loss (CLL_q1) and bruising (CLL_q3), forming another potential symptom cluster. After that, this cluster is joined with symptoms related to diarrhea (DI), lack of appetite (AP), constipation (CO), and stomach aches (CLL_q4). The other symptoms in the top left of the dendogram do not seem to belong to any particular symptom cluster. These include symptoms related to shortness of breath (DY), tingling hands or feet (CLL_q11), dry mouth (CLL_q2), worrying (CLL_q13), sweating at night (CLL_q6), trouble sleeping (SL), and skin issues (CLL_q7). Interestingly, all these symptoms are present for some of the subgroups in the repeated Bimax results, with the exception of tingling hands or feet (CLL_q11).

3.4 Questmotif

Unfortunately, the results from the ensemble run of the Questmotif algorithm were very unstable. In total, 5095 biclusters were discovered and a similarity matrix based on the Jaccard index was computed. The resulting dendogram based on this similarity matrix is shown in figure 12. Ideally, one would like to see many biclusters joined together at the bottom of the dendogram, which would indicate highly similar biclusters. However, as can be seen in figure 12, this is not the case here. It is virtually impossible to extract robust superbiclusters from this dendogram because there are no clear clusters visible.

Figure 11
Symptom cluster dendogram



This indicates that the discovered biclusters are all highly dissimilar to each other. Reasons as to why this might be the case are discussed in section 4.

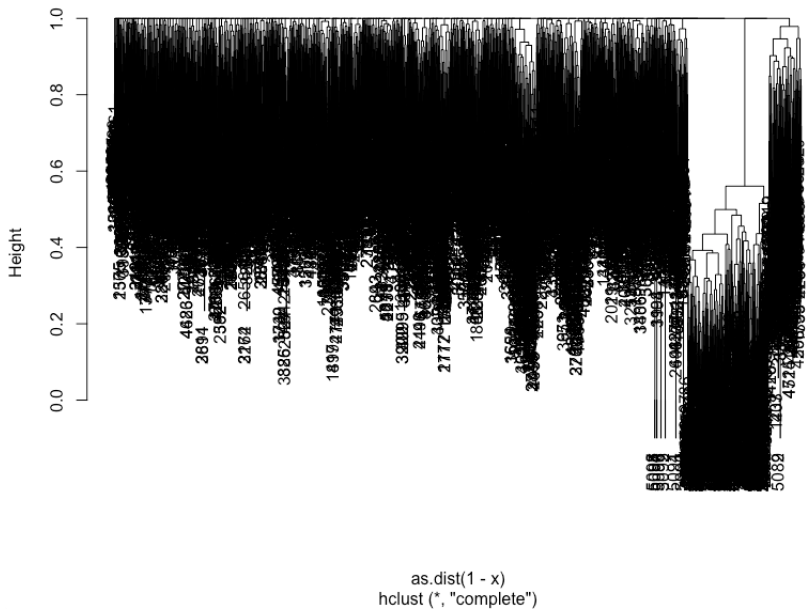
4. Discussion

In section 4.1, the contribution of this study to the current body of research on symptom science is discussed. More specifically, the clinical relevance of the findings and the advantage of biclustering algorithms over traditional clustering methods is discussed. In section 4.2, the implications of the findings from the repeated Bimax algorithms and the resulting biclusters are discussed in more detail. In particular, the discovered subgroups and symptoms clusters are discussed and compared with symptom clusters previously found in the literature. In section 4.2.1 the differences between subgroups with respect to HRQOL are discussed. Finally, in section 4.3 the limitations of the current study are discussed and recommendations with respect to future research in biclustering and symptom data are given.

4.1 General Discussion

As mentioned above, the number of long-term cancer survivors is increasing (Verdecchia et al. 2007; Engert et al. 2012, 2017; Oerlemans et al. 2013). Among these long-term survivors there are many who do not longer experience any symptoms. However, a smaller but substantial group of survivors experiences symptoms even years after their original diagnosis (Stein, Syrjala, and Andrykowski 2008; Harrington et al. 2010). Findings from the current study support this claim; roughly half of the survivors in-

Figure 12
Questmotif dendogram showing similarity between the discovered biclusters



cluded in the total sample were included in a bicluster (indicating that they suffer from symptoms), meaning that the other half of the survivors did either not experience any symptoms or did not experience symptoms to such a degree that they were included in a bicluster. Most of the research on symptom clusters has focused on patients during treatment, or shortly thereafter. Consequently, less is known about the symptom experience of long-term cancer survivors (Harrington et al. 2010). This study aimed to fill this gap by looking at symptom data gathered from long-term NHL survivors. Because many survivors do not longer experience any symptoms, or experience only a small number of symptoms, biclustering algorithms are particularly suitable for discovering subgroups of survivors in the data that do experience symptoms.

Overall, the most important finding from this study is that biclustering algorithms, such as the repeated Bimax algorithm, can be used to find subgroups of survivors that differ from each other with respect to the symptoms they experience. More specifically, nine different subgroups of survivors were found that differ with respect to the symptoms they experience. This finding is of clinical importance because based on the biclustering results, clinicians are able to target specific subgroups of survivors with treatments that take into account the different symptoms that the survivors in the subgroups experience.

Furthermore, the biclustering results provide more clinically relevant information on both patient subgroups as well as their specific symptom experience than the results obtained from PCA and HCA since both PCA and HCA can only be used to find either symptom clusters or patient subgroups, but never both simultaneously [2]. Next to that, the symptoms clusters identified by PCA and HCA are expressed over all survivors. Even though PCA and HCA might provide relatively good results with symptom data gathered from patients that are still in treatment or shortly after treatment, since it is

likely that most of these patients experience symptoms, they are probably less suitable for symptom data gathered from long-term survivors, many of whom do no longer experience any symptoms or experience only a small number of symptoms. Crucially, the main difference between the biclustering results and the results from PCA and HCA is that the biclustering results can be used to identify unique subgroups that differ with respect to their symptom experience, whereas PCA and HCA only allow for identification of symptom clusters, without subgroup identification. Thus, with PCA and HCA, potentially crucial information on differences in symptom experience between subgroups is lost.

To my knowledge, this study is the first to use biclustering algorithms for discovering subgroups and symptom clusters from symptom data. The bicluster results from the repeated Bimax algorithm show that this particular biclustering algorithm is highly suitable for application to symptom data. Furthermore, the results also show that the repeated Bimax algorithm is capable of distinguishing between survivors that experience symptoms and survivors that do not experience symptoms (or that experience only a few symptoms), without having to define these groups *"a priori"*. This result is backed up by the the finding that there is a significant difference in HRQOL score between these two groups, indicating that survivors included in the biclusters had a lower HRQOL score than survivors not included in the biclusters. While the relation between HRQOL and symptom experience has been found in the existing literature for cancer patients, the findings from the current study add to this by showing that a relation between symptom experience and HRQOL might exist for long-term cancer survivors as well (Stein, Syrjala, and Andrykowski 2008; Harrington et al. 2010; Leach et al. 2014; Barsevick 2016; Miaskowski et al. 2017).

As mentioned before, traditional methods such as PCA and HCA can only be used to find symptom clusters, they cannot be used to simultaneously discover subgroups that differ with respect to symptom experience. Furthermore, the symptom clusters discovered with PCA and HCA are often difficult to interpret. Because the components (clusters) of PCA are a linear combination of all variables, loadings for each individual symptom will be non-zero. User-defined thresholds are then required to determine whether or not a certain symptom belongs to a cluster (Fan, Hadi, and Chow 2007; Barsevick 2016). This was also the case with the PCA results in the current study: because of the non-zero loadings on the individual variables, a choice had to be made with respect to the minimal loading value. Furthermore, the amount of variance explained by each of the components is based on all variables, selecting only a few of these to belong to a symptom cluster means that the variance explained cannot be attributed to these variables alone. This makes it difficult to gauge the actual contribution of the extracted symptom cluster to the variation present in the data. In contrast, symptom clusters found with the repeated Bimax algorithm can be interpreted without having to resort to any user-defined thresholds because a symptom is either present for a particular subgroup or it is not, thus leading to discrete symptom clusters.

It should be noted here, however, that even though some symptoms are not included in the biclusters, they may still be experienced by a small number of survivors included in the biclusters. The fact that these symptoms are not included in any of the biclusters simply means that these particular symptoms are not experienced by *all* survivors in a particular bicluster. This is due to how the repeated Bimax algorithm constructs the biclusters. The repeated Bimax algorithm looks for perfect biclusters. That is to say, only rows that have ones for all the columns belonging to a particular bicluster are included, no noise (zeroes) is allowed. This could mean that, for example, of the survivors included in subgroup 1, a small number also experienced symptoms

related to nausea and vomiting, but because of the requirement of perfect biclusters these symptoms would only be included in the bicluster if all survivors in this particular bicluster experienced these symptoms.

Of course, the parameter settings of the algorithm still have to be determined by the user and these depend largely on the research goals of the researcher / clinician. As expected, a greater value for `minr` leads to a smaller number of biclusters of a larger size and a smaller value for `minr` leads to a larger number of biclusters of a smaller size. This means that the researcher should decide, based on her domain knowledge and in consultation with a clinician, whether interest is in finding a larger number of smaller subgroups or in finding a smaller number of larger subgroups. Of course, when subgroups of larger sizes are found, some observations that would have potentially formed their own subgroup (if smaller subgroup sizes would have been acceptable), are now included in the larger subgroups. On the other hand, subgroups that include only 10 observations are potentially not very clinically relevant. This is a consideration that has to be made by the researcher and clinician. Another option would be to use a cross-validation approach or other data driven methods to determine the optimal parameter settings. One issue with this approach is that a suitable measure has to be found to determine bicluster quality in order to allow for quantitative comparison between bicluster results. Since this is the first study that has looked at biclustering algorithms applied to symptom data, more research has to be carried out in this regard.

So far, much of the research in the field of symptom science has focused on symptom cluster consistency across statistical methods, across different cancer sites, and over time, with mixed results (Dong et al. 2016; Chen et al. 2012b; Nguyen et al. 2011; Chen et al. 2011; Kim et al. 2009). However, as mentioned in section 1.5.7, perhaps symptom cluster consistency can never be fully expected because of differences in statistical methods, patient populations, cancer type, and disease stage. Even with relatively homogeneous patient populations (i.e., patient populations with the same cancer type, time since diagnosis, etc.) consistent symptom clusters are not always found (Nguyen et al. 2011). This seems to indicate that the currently used methods are perhaps not the most suitable for symptom cluster identification, because differences between subgroups are not taken into account. The results from the current study show that biclustering algorithms can be used to identify meaningful subgroups in a relatively homogeneous population of NHL survivors. Future research should also look into the use of biclustering algorithms for identification of subgroups that share a similar symptom experience in a more heterogeneous patient population.

For example, consider a population consisting of patients suffering from two different types of cancer. It might be the case that the patients suffering from one type of cancer experience different symptoms than the patients suffering from another type of cancer. Biclustering algorithms could be used on the entire population, comprised of patients with both types of cancer, and if differences between these two groups exist the algorithm should be able to detect them. On the other hand, it could also be that the type of cancer has no effect on which kinds of symptoms the patients experience, but rather that other (demographic) factors determine the patients' symptom experience. In this case, the biclustering algorithm could still be used because it would be able to find those subgroups with a similar symptom experience, regardless of cancer type. This then also allows for further investigating whether different cancer types lead to different symptoms. Of course, up to this point this is all theoretical but the potential of biclustering algorithms in the field of symptom science should be acknowledged and more research should be carried out in this regard.

This study also adds to the existing body of research on biclustering algorithms. Most of the research on biclustering algorithms has been done in the field of computational biology, more specifically the biclustering of gene expression data (Pontes, Giráldez, and Aguilar-Ruiz 2015; Padilha and Campello 2017). Outside of the field of computational biology, biclustering algorithms have been applied to tourism data in a market segmentation study and some interest has been shown in applying biclustering in data mining (Dolnicar et al. 2011; Busygin, Prokopyev, and Pardalos 2008). The results of the current study show that biclustering algorithms can also be applied to symptom data.

4.2 Biclusters

The biclustering results from the repeated Bimax (**model 2**) showed that symptoms related to fatigue are very commonly experienced by long-term NHL survivors, indicating that issues with fatigue are present even longer after the original diagnosis. This finding is in line with previous research on cancer patients, where a cluster of symptoms relating to fatigue is often found (Dong et al. 2016; Oerlemans et al. 2013; Ahlberg, Ekman, and Gaston-Johansson 2005; Cheng and Lee 2011; Hjermstad et al. 2006; Jenkins and McCoy 2015). Symptoms related to fatigue are present in each of the nine different subgroups. This suggests that clinicians should emphasize the treatment of symptoms related to fatigue in their treatment approaches. Furthermore, symptoms related to fatigue also loaded onto component 1 (symptom cluster 1) of the PCA results and were found to cluster together in the HCA dendogram (as can be seen in figure 10 and 11).

Other symptoms commonly experienced by long-term NHL survivors include symptoms related to emotional health. This finding suggests that many long-term survivors struggle with issues related to depression, anxiety, and worrying. This finding is also in line with previous research on cancer patients, where a symptom cluster related to depression and anxiety is often found (Dong et al. 2016; Stein, Syrjala, and Andrykowski 2008; Fobair et al. 1986; Loge et al. 1999; Fosså, Dahl, and Loge 2003; Breland et al. 2015; Jenkins and McCoy 2015). It should also be noted that symptoms related to fatigue have often been associated with emotional health (psychological distress) (Brown and Kroenke 2009; Jenkins and McCoy 2015). Given that many subgroups experience both symptoms related to fatigue and emotional health and that significant positive correlations were found between symptoms related to fatigue and emotional health, it could be interesting for future research to further investigate the relationship between these two constructs.

Symptoms related to emotional health are not present in every subgroup. This is a good example of why biclustering algorithms should be preferred over traditional clustering methods; because the symptoms are not present in every subgroup, treatment can be adapted for each individual survivor, depending on which subgroup she belongs to. Even though the results from the PCA and HCA also indicated a symptom cluster related to emotional health (and fatigue), it is impossible to know from those results which particular subgroups of survivors experience these symptoms since these methods do not allow for simultaneous subgroup discovery. Thus, the repeated Bimax algorithm has a significant advantage over the traditional clustering methods.

Interestingly, the PCA results showed a symptom cluster related to nausea, vomiting, and lack of appetite. This cluster has consistently been found in previous research (Chen et al. 2011; Walsh and Rybicki 2006; Miaskowski et al. 2017). It is somewhat surprising to see a symptom cluster related to nausea, vomiting and appetite loss

since these symptoms are usually associated with the treatment of cancer (e.g., as a result of chemotherapy) and the current study included many survivors which have not received any cancer treatment for years. However, it could be that a small group of survivors included in the data received chemotherapy or other types of treatment, shortly before the questionnaire was administered, because of the recurrence of their cancer. Unfortunately, no data was available to verify this.

Symptoms related to vomiting, nausea, and appetite loss were not included in any of the biclusters from the repeated Bimax algorithm. This could be due to how the repeated Bimax algorithm constructs its biclusters. Biclusters that can be contained within in a larger bicluster are thrown away, so it could be that the bicluster related to nausea and vomiting is contained within any of the other biclusters. However, it could also be that the symptoms do not occur in the bicluster results because of the dichotomization of the data before the Bimax algorithm is applied. It could be that a small number of survivors reported high values (i.e., values of 3 or 4) with respect to the nausea and vomiting symptoms while a larger group of survivors did not experience these particular symptoms. Because the intensity of the symptom experience is lost with the dichotomization of the data, the nausea and vomiting symptoms may not appear in the repeated Bimax results.

Finally, it could also be due to the fact that the Bimax algorithm looks for perfect biclusters. That is to say, only rows that have ones for all the columns belonging to this bicluster are included, no noise is allowed. After a row (survivor) is included in a bicluster it is removed from the data and thus cannot be included in another bicluster. Because of this, it could be that some survivors that experienced nausea and vomiting were already included in another bicluster, which did not contain these particular symptoms, and thus these symptoms may not be discovered by the repeated Bimax algorithm. Further analyses of the results seem to support this notion. Most of the survivors that experienced symptoms related to nausea, vomiting, and appetite loss were included in the biclusters, even though these symptoms were not found by the algorithm.

The Bibit algorithm, which is quite similar to the Bimax algorithm, has been proposed to deal with this issue of perfect biclusters. The Bibit algorithm allows some noise in its biclusters, which means that the rows included in a bicluster do not need to consist of only ones, some zeros are also allowed (Rodriguez-Baena, Perez-Pulido, and Aguilar-Ruiz 2011). The downside of this algorithm is that a large number of biclusters is typically found, because of the noise allowance, and an ensemble method is required to extract robust biclusters from this large set of biclusters. Nevertheless, future research should look at the application of the Bibit algorithm to symptom data and compare the resulting biclusters with the results of the Bimax algorithm.

Many symptoms that occur in the biclusters are not included in the symptom cluster results of the PCA. This could indicate that these symptoms are only experienced by a small subset of the data, leading to PCA being unable to detect this because it does not take differences between subgroups into account.

Other symptoms that were included in the biclusters include symptoms related to pain, stomach issues, shortness of breath, memory, trouble sleeping, skin issues, infections, and sweating at night. An interesting symptom cluster emerged for subgroup 8, where symptoms related to trouble sleeping and sweating at night are clustered together, suggesting a potential relation between these two symptoms.

Significant differences were found between subgroups with respect to some demographic variables, including gender, years since diagnosis, NHL type, and age. Since significant differences between subgroups with respect to NHL type were found, this

could indicate that different types of NHL lead to a different symptom experience for long-term NHL survivors. This is perhaps not surprising since indolent NHL, for example, is generally considered to be incurable (Leukemia & Lymphoma Society 2013a). The fact that significant differences between subgroups with respect to years since diagnosis were found could indicate that the number of years since the original diagnosis has an effect on the symptom experience of long-term NHL survivors. However, more research needs to be carried out in this regard.

With respect to the group of survivors not included in the biclusters, the results showed that 12.2% of these survivors did not experience any symptoms. Furthermore, 66.8% of the survivors from this group experienced 5 symptoms or less. The fact that some of these survivors still suffer from symptoms but are not included in a bicluster likely stems from the nature of how the repeated Bimax algorithm works. As mentioned above, the repeated Bimax algorithm looks for perfect biclusters. This means that the group of survivors not included in the biclusters likely does not fit within any of the discovered biclusters without the allowance of some noise. This is a downside of the repeated Bimax algorithm and future research should look into the use of biclustering algorithms that allow imperfect biclusters, such as the Bibit algorithm. Nevertheless, the finding that the HRQOL of survivors not included in any of the biclusters was significantly better than the HRQOL of survivors included in the biclusters provides support to the claim that both groups of survivors are different and that the repeated Bimax algorithm is able to distinguish between these groups.

4.2.1 HRQOL. Previous research has shown that symptoms affect a patient's HRQOL (Leach et al. 2014; Amro et al. 2014; Arden-Close, Pacey, and Eiser 2010; Dong et al. 2016; Hjermstad et al. 2006; Jenkins and McCoy 2015; Lee and Jeon 2015). This finding is supported by the results from the current study. The HRQOL score of survivors not included in any of the biclusters was higher than the HRQOL score of survivors that were included in a bicluster. This indicates that survivors that experience more symptoms, or experience symptoms with a higher intensity, have a lower HRQOL than survivors that experience symptoms to a lesser degree. Thus, there exists great variability in the number of symptoms long-term NHL survivors experience and the extent to which these symptoms affect their HRQOL.

Differences in HRQOL were also found between the survivor subgroups from model 2. Among other things, the results showed a large range of HRQOL scores, with mean scores as low as 57.75 for subgroup 1 and mean scores as high as 88.61 for subgroup 9. These results again highlight the clinical usefulness of the biclustering results. Clinicians are able to distinguish between groups of survivors that have a low HRQOL score, who might benefit from immediate treatment of their symptoms, on the one hand, and groups of survivors that have a relatively high HRQOL score and for whom immediate treatment of their symptoms might not be as pressing, on the other hand. Future research could also look at using the subgroup assignments as a predictor in a model to predict HRQOL. This could provide further evidence of a relation between symptoms and HRQOL and it could also emphasize the usefulness of the discovered subgroups.

The difference in HRQOL scores might also shed some light on which symptoms have the largest impact on HRQOL. For example, symptoms experienced by subgroup 1 (the group with the lowest mean HRQOL score) include symptoms related to fatigue and emotional health, whereas symptoms experienced by subgroup 9 (the group with highest mean HRQOL score) include symptoms related to fatigue, memory, and skin issues. This can be interpreted in three ways: 1) the symptoms experienced by subgroup

1 are more severe and thus have a greater impact on HRQOL than the symptoms experienced by subgroup 9, 2) there is no difference in the severity of the symptoms experienced by both subgroups but the intensity of the symptom experience is different for both groups, or 3) in the calculation of the HRQOL score more symptoms related to emotional health than to fatigue are used, meaning that the HRQOL score could be biased. More research is needed in this regard to be able to distinguish between these three explanations.

4.3 Limitations and Future Research

Several limitations of the current study have to be taken into consideration. Firstly, since the repeated Bimax algorithm was designed for binary data, information on symptom intensity is unfortunately lost. Secondly, the results of the Questmotif algorithm were unstable and thus were not discussed further. These unstable results might be due to the number of different parameters that was tested, leading to a high number of unique biclusters which have only a limited similarity to each other. Future research could look into this issue and perhaps find a way to apply the Questmotif to symptom data so that it gives interpretable and stable results. Thirdly, due to time constraints, only single imputation of the missing data could be carried out. There are currently no methods available that enables one to easily implement multiple imputation in combination with biclustering. Future research could be aimed at developing these methods. Furthermore, future research should also investigate the use of biclustering algorithms with different patient populations to see whether clinically relevant biclusters can be obtained. Another shortcoming of this study is that bicluster quality was not objectively quantified (Chia and Karuturi 2010). Effort should be put towards the development of methods to objectively quantify the quality of a bicluster found from symptom data. Next to that, being able to tune algorithm parameters in a data driven way would be a nice addition to the repeated Bimax algorithm. Finally, future research should look into the difference in impact of different symptoms/symptom clusters on HRQOL. It could be that different symptoms or different symptom clusters have a different impact on HRQOL.

4.4 Conclusion

This study is the first study to use biclustering algorithms to discover subgroups and symptom clusters from symptom data. Overall, the findings of this study show that biclustering algorithms, such as the repeated Bimax algorithm, can be used to find clinically meaningful subgroups of NHL survivors that differ with respect to the symptoms they experience. More specifically, nine subgroups of survivors were found, with each group showing a different symptom experience. These subgroups also differed with respect to their HRQOL score. This provides additional support to the claim that the subgroups discovered by the biclustering algorithm are distinct. Furthermore, biclustering algorithms solve some important issues related to more traditional clustering methods, such as PCA and HCA. These traditional methods are only capable of either clustering variables or observations, but never both simultaneously, and the discovered symptom clusters are expressed over all survivors. Biclustering algorithms solve this issue by clustering variables and observations simultaneously, which allows for discovery of subgroups with distinct symptom profiles. This, in turn, leads to a higher clinical relevance of the findings because clinicians can use this information to target specific subgroups of survivors and develop a treatment plan that is tailored to the symptom

profiles of the specific subgroups. To conclude, the results from this study show the potential of using biclustering algorithms to discover clinically relevant subgroups and symptom clusters from symptom data and more research should be carried out in this regard in order to explore this promising direction further.

References

Aaronson, N. K., S. Ahmedzai, B. Bergman, M. Bullinger, A. Cull, N. J. Duez, A. Filiberti, H. Flechtner, S. B. Fleishman, J. C. J. M. de Haes, S. Kaasa, M. Klee, D. Osoba, D. Razavi, P. B. Rofe, S. Schraub, K. Sneeuw, M. Sullivan, and F. Takeda. 1993. The European organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, 85(5):365–376.

Ahlberg, K., T. Ekman, and F. Gaston-Johansson. 2005. Fatigue, psychological distress, coping resources, and functional status during radiotherapy for uterine cancer. *Oncology Nursing Forum*, 32(3):633–640.

Amro, A., B. Waldum, T. Dammen, C. Miaskowski, and I. Os. 2014. Symptom clusters in patients on dialysis and their association with quality-of-life outcomes. *Journal of Renal Care*, 40(1):23–33.

Arden-Close, E., A. Pacey, and C. Eiser. 2010. Health-related quality of life in survivors of lymphoma: A systematic review and methodological ciritique. *Leukemia & Lymphoma*, 51(4):628–640.

Barsevick, A. 2016. Defining the symptom cluster: How far have we come? *Seminars in Oncology Nursing*, 32(4):334–350.

Barsevick, A. M. 2007. The elusive concept of the symptom cluster. *Oncology Nursing Forum*, 34(5):971–980.

Bender, C. M., S. J. Engberg, H. S. Donovan, S. M. Cohen, M. P. Houze, M. Q. Rosenzweig, G. A. Mallory, J. Dunbar-Jacob, and S. M. Sereika. 2008. Symptom clusters in adults with chronic health problems and cancer as a comorbidity. *Oncology Nursing Forum*, 35(1):1–11.

Bray, F., J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424.

Breland, J. Y., N. E. Hundt, T. L. Barrera, J. Mignogna, N. J. Petersen, M. A. Stanley, and J. A. Cully. 2015. Identification of anxiety symptom clusters in patients with COPD: Implications for assessment and treatment. *International Journal of Behavioral Medicine*, 22(5):590–596.

Brock, G., V. Pihur, S. Datta, and S. Datta. 2008. clValid: An R package for cluster validation. *Journal of Statistical Software*, 25(4).

Brown, L. F. and K. Kroenke. 2009. Cancer-related fatigue and its associations with depression and anxiety: A systematic review. *Psychosomatics*, 50(5):440–447.

Busygin, S., O. Prokopyev, and P. M. Pardalos. 2008. Biclustering in data mining. *Computers & Operations Research*, 35:2964–2987.

Chavent, M., V. Kuentz, B. Liquet, and J. Saracco. 2012. ClustOfVar: An R package for the clustering of variables. *Journal of Statistical Software*, 50(2).

Chen, E., L. Khan, L. Zhang, J. Nguyen, G. Cramarossa, M. Tsao, C. Danjoux, E. Barnes, A. Sahgal, L. Holden, F. Jon, K. Dennis, S. Culleton, and E. Chow. 2012a. Symptom clusters in patients with bone metastases - a reanalysis comparing different statistical methods. *Supportive Care in Cancer*, 20(11):2811–2820.

Chen, E., J. Nguyen, G. Cramarossa, L. Khan, A. Leung, S. Lutz, and E. Chow. 2011. Symptom clusters in patients with lung cancer: A literature review. *Expert Review of Pharmacoeconomics & Outcomes*, 11(4):433–439.

Chen, E., J. Nguyen, L. Khan, L. Zhang, G. Cramarossa, M. Tsao, C. Danjoux, E. Barnes, A. Sahgal, L. Holden, F. Jon, and E. Chow. 2012b. Symptom clusters in patients with advanced cancer: A reanalysis comparing different statistical methods. *Journal of Pain and Symptom Management*, 44(1):23–32.

Cheng, K. K. F. and D. T. F. Lee. 2011. Effects of pain, fatigue, insomnia, and mood disturbance on functional status and quality of life of elderly patients with cancer. *Critical Reviews in Oncology/Hematology*, 78:127–137.

Cheng, Y. and G. M. Church. 2000. Biclustering of expression data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 1:93–103.

Cheung, W. Y., L. W. Le, and C. Zimmerman. 2009. Symptom clusters in patients with advanced cancers. *Supportive Care in Cancer*, 17(9):1223–1230.

Chia, B. K. H. and R. K. M. Karuturi. 2010. Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. *Algorithms for Molecular Biology*, 5.

De Troyer, E. and M. Otava. 2017. *RcmdrPlugin.BiclustGUI: 'Rcmdr' Plug-in GUI for Biclustering*. R package version 1.1.1.

Dodd, M. J., C. Miaskowski, and S. M. Paul. 2001. Symptom clusters and their effect on the functional status of patients with cancer. *Oncology Nursing Forum*, 28(3):465–470.

Dolnicar, S., B. Grün, F. Leisch, and K. Schmidt. 2014. Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research*, 53(3):296–306.

Dolnicar, S., S. Kaiser, K. Lazarevski, and F. Leisch. 2011. Biclustering: Overcoming data dimensionality problems in market segmentation. *Journal of Travel Research*, 51(1):41–49.

Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Dong, S. T., D. S. J. Costa, P. N. Butow, Me. R. Lovell, M. Agar, G. Velikova, P. Teckle, A. Tong, N. C. Tebbutt, S. J. Clarke, K. van der Hoek, M. T. King, and P. M. Fayers. 2016. Symptom clusters in advanced cancer patients: An empirical comparison of statistical methods and the impact on quality of life. *Journal of Pain and Symptom Management*, 51(1):88–98.

Engert, A., H. Goergen, J. Markova, T. Pabst, J. Meissner, J. M. Zijlstra, Z. Král, D. A. Eichenauer, M. Soekler, R. Greil, S. Kreissl, R. Scheuvens, R. Eich, H. Eich, C. Kobe, M. Dietlein, H. Stein, M. Fuchs, V. Diehl, and P. Borchmann. 2017. Reduced-intensity chemotherapy in patients with advanced-stage Hodgkin lymphoma. *Hemasphere*, 1(1).

Engert, A., H. Haverkamp, C. Kobe, J. Markova, C. Renner, A. Ho, J. Zijlstra, Z. Král, M. Fuchs, M. Hallek, L. Kanz, H. Döhner, H Dörken, B. Dörken, N. Engel, M Topp, S. Klutmann, H. Amthauer, A. Bockisch, R. Kluge, C. Kratochwill, O. Schober, R. Greil, R. Andreesen, M. Kneba, M. Pfreundschuh, H. Stein, H. T. Eich, R. P. Müller, M. Dietlein, P. Borchmann, and V. Diehl. 2012. Reduced-intensity chemotherapy and PET-guided radiotherapy in patients with advanced stage Hodgkin's lymphoma (HD15 trial): A randomised, open-label, phase 3 non-inferiority trial. *The Lancet*, 379:1791–1799.

Esther Kim, J. E., M. J. Dodd, B. E. Aoizerat, T. Jahan, and C. Miaskowski. 2009. A review of the prevalence and impact of multiple symptoms in oncology patients. *Journal of Pain and Symptom Management*, 37(4):715–736.

European Organisation for Research and Treatment of Cancer. CLL-17.

Fan, G., S. Hadi, and E. Chow. 2007. Symptom clusters in patients with advanced-stage cancer referred for palliative radiation therapy in an outpatient setting. *Supportive Cancer Therapy*, 4(3):157–162.

Fobair, P., R. T. Hoppe, J. Bloom, R. Cox, A. Varghese, and D. Spiegel. 1986. Psychosocial problems among survivors of Hodgkin's disease. *Journal of Clinical Oncology*, 4(5):805–814.

Fosså, S. D., A. A. Dahl, and J. H. Loge. 2003. Fatigue, anxiety, and depression in long-term survivors of testicular cancer. *Journal of Clinical Oncology*, 21(7):1249–1254.

Fox, J. 2005. The R Commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software*, 14(9).

Giesinger, J. M., J. M. Kieffer, P. M. Fayers, M. Groenvold, M. A. Petersen, N. W. Scott, M. A. G. Sprangers, G. Velikova, and N. K. Aaronson. 2016. Replication and validation of higher order models demonstrated that a summary score for the EORTC QLQ-C30 is robust. *Journal of Clinical Epidemiology*, 69:79–88.

Gift, A. G., M. Stommel, A. Jablonski, and W. Given. 2003. A cluster of symptoms over time in patients with lung cancer. *Nursing Research*, 52(6):393–400.

Hadi, S., G. Fan, A. E. Hird, A. Kirou-Mauro, L. A. Flipczak, and E. Chow. 2008a. Symptom clusters in patients with cancer with metastatic bone pain. *Journal of Palliative medicine*, 11(4):591–600.

Hadi, S., L. Zhang, A. Hird, E. de Sa, and E. Chow. 2008b. Validation of symptom clusters in patients with metastatic bone pain. *Current Oncology*, 15:211–218.

Harrington, C. B., J. A. Hansen, M. Moskowitz, B. L. Todd, and M. Feuerstein. 2010. It's not over when it's over: Long-term symptoms in cancer survivors - a systematic review. *International Journal of Psychiatry in Medicine*, 40(2):163–181.

Hartigan, J. A. 1972. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129.

Henoch, I., A. Ploner, and C. Tishelman. 2009. Increasing stringency in symptom cluster research: A methodological exploration of symptom clusters in patients with inoperable lung cancer. *Oncology Nursing Forum*, 36(6):283–292.

Hjermstad, M. J., L. Oldervoll, S. D. Fosså, H. Holte, A. B. Jacobsen, and J. H. Loge. 2006. Quality of life in long-term Hodgkin's disease survivors with chronic fatigue. *European Journal of Cancer*, 42(3):327–333.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *Introduction to Statistical Learning*. Springer.
Jenkins, J. B. and T. P. McCoy. 2015. Symptom clusters, functional status, and quality of life in older adults with osteoarthritis. *Orthopaedic Nursing*, 34(1):36–42.
Jiménez, A., R. Madero, A. Alonso, V. Martínez-Marín, Y. Vilches, B. Martínez, M. Feliu, L. Díaz, Espinosa E., and J. Feliu. 2011. Symptom clusters in advanced cancer. *Journal of Pain and Symptom Management*, 42(1):24–31.
Jolliffe, I. T. and J. Cadima. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society*, 374.
Josse, J. and F. Husson. 2016. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1).
Jurgens, C. Y., D. K. Moser, R. Armola, B. Carlson, K. Sethares, and B. Riegel. 2009. Symptom clusters of heart failure. *Research in Nursing & Health*, 32(5):551–560.
Kaiser, S. 2011. *Biclustering: Methods, software and application*. Ph.D. thesis, Ludwig Maximilian University of Munich.
Kaiser, S., R. Santamaria, T. Khamiakova, M. Sill, R. Theron, L. Quintales, F. Leisch, and E. De Troyer. 2018. *biclust: Bicluster Algorithms*. R package version 2.0.1.
Kankerbestrijding. 2011. Kanker in Nederland tot 2020, trends en prognoses.
Kasim, A., Z. Shkedy, S. Kaiser, S. Hochreiter, and W. Talloen, editors. 2016. *Applied Biclustering Methods for Big and High-Dimensional Data Using R*. CRC Biostatistics Series. Chapman & Hall.
Kim, E., T. Jahan, B. E. Aouizerat, M. J. Dodd, B. A. Cooper, S. M. Paul, C. West, K. Lee, P. S. Swift, W. Wara, and C. Miaskowski. 2009. Changes in symptom clusters in patients undergoing radiation therapy. *Supportive Care in Cancer*, 17(11):1383–1391.
Kim, H. J., A. M. Barsevick, S. L. Beck, and W. Dudley. 2012. Clinical subgroups of a psychoneurologic symptom cluster in women receiving treatment for breast cancer: A secondary analysis. *Oncology Nursing Forum*, 39(1):20–30.
Kim, H. J., A. M. Barsevick, L. Tulman, and P. A. McDermott. 2008. Treatment-related symptom clusters in breast cancers: A secondary analysis. *Journal of Pain and Symptom Management*, 36(5):468–479.
Kim, H. J., D. B. McGuire, L. Tulman, and A. M. Barsevick. 2005. Symptom clusters. Concept analysis and clinical implications for cancer nursing. *Cancer Nursing*, 28(4):270–282.
Leach, C. R., K. E. Weaver, N. M. Aziz, C. M. Alfano, K. M. Bellizzi, E. E. Kent, L. P. Forsythe, and J. H. Rowland. 2014. The complex health profile of long-term cancer survivors: Prevalence and predictors of comorbid conditions. *Journal of Cancer Survivorship*, 9(2):239–251.
Lee, S. J. and J. Jeon. 2015. Relationship between symptom clusters and quality of life in patients at stages 2 to 4 chronic kidney disease in Korea. *Applied Nursing Research*, 28(4):e13–e19.
Leukemia & Lymphoma Society. 2013a. The Lymphoma Guide.
Leukemia & Lymphoma Society. 2013b. Hodgkin Lymphoma.
Leukemia & Lymphoma Society. 2013c. Non-Hodgkin Lymphoma.
Loge, J. H., A. F. Abrahamsen, Ø. Ekeberg, and S. Kaasa. 1999. Reduced health-related quality of life among Hodgkin's disease survivors: A comparative study with general population norms. *Annals of Oncology*, 10:71–77.
Madeira, S. C. and A. L. Oliveira. 2004. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45.
McCorkle, R. and V. Young. 1978. Development of a symptom distress scale. *Cancer Nursing*, 1(5):373–378.
Miaskowski, C., A. Barsevick, A. Berger, R. Casagrande, P. A. Grady, P. Jacobsen, J. Kutner, D. Patrick, L. Zimmerman, C. Xiao, M. Matocha, and S. Marden. 2017. Advancing symptom science through symptom cluster research: Expert panel proceedings and recommendations. *Journal of the National Cancer Institute*, 109(4):1–9.
Moens, K., R. J. Siegert, S. Taylor, E. Namisango, and R. Harding. 2015. Symptom clusters in people living with HIV attending five palliative care facilities in two sub-saharan African countries: A hierarchical cluster analysis. *PLOS ONE*, 10(5).
Molassiotis, A., M. Lowe, F. Blackhall, and P. Lorigan. 2011. A qualitative exploration of a respiratory distress symptom cluster in lung cancer: Cough, breathlessness and fatigue. *Lung Cancer*, 71:94–102.
Murali, T. M. and S. Kasif. 2003. Extracting conserved gene expression motifs from gene expression data. *Pacific Symposium on Biocomputing*, 8:77–88.

Namisango, E., R. Harding, E. T. Katabira, R. J. Siegert, R. A. Powell, L. Atuhaire, K. Moens, and S. Taylor. 2015. A novel symptom cluster analysis among ambulatory HIV/AIDS patients in Uganda. *AIDS Care*, 27(8):954–963.

National Coalition for Cancer Survivorship. 2006. About cancer survivorship research: Survivorship definitions.

Nederlandse Kankerregistratie, IKNL. Incidentiecijfers agressief non-Hodgkinlymfoom.

Ness, K., M. Wall, J. Oakes, L. Robison, and J. Gurney. 2006. Physical performance limitations and participation restrictions among cancer survivors: A population-based study. *Annals of Epidemiology*, 16(3):197–205.

Newton, R., J. Ferlay, V. Beral, and S. Devesa. 1997. The epidemiology of non-Hodgkin's lymphoma: Comparison of nodal and extra-nodal sites. *International Journal of Cancer*, 72(6):923–930.

Nguyen, J., G. Cramarossa, D. Bruner, E. Chen, L. Khan, A. Leung, S. Lutz, and E. Chow. 2011. A literature review of symptom clusters in patients with breast cancer. *Expert Review of Pharmacoeconomics & Outcomes*, 11(5):533–539.

Oerlemans, S., F. Mols, D. E. Issa, J. H. F. M. Pruijt, W. G. Peters, M. Lybeert, W. Zijlstra, J. W. W. Coebergh, and L. V. van de Poll-Franse. 2013. A high level of fatigue among long-term survivors of non-Hodgkin's lymphoma: results from the longitudinal population-based PROFILES registry in the south of the Netherlands. *Haematologica*, 98(3):479–486.

Padilha, V. A. and R. J. G. B. Campello. 2017. A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics*, 18.

Papachristou, N., C. Miaskowski, P. Barnaghi, R. Maguire, N. Farajidavar, B. Cooper, and X. Hu. 2016. Comparing machine learning clustering with latent class analysis on cancer symptoms' data. *2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT)*.

Park, S. K. and J. L. Larson. 2014. Symptom cluster, healthcare use and mortality in patients with severe chronic obstructive pulmonary disease. *Journal of Clinical Nursing*, 23:2658–2671.

Pfundstein, G. 2010. Ensemble methods for plaid bicluster algorithm. Bachelor thesis, Institut fur Statistik, LMU Munchen.

Pontes, B., R. Giráldez, and J. S. Aguilar-Ruiz. 2015. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180.

Prelić, A., S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. 2016. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revelle, W. 2018. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.8.12.

Rodriguez-Baena, D. S., A. J. Perez-Pulido, and J. S. Aguilar-Ruiz. 2011. A biclustering algorithm for extracting bit-patters from binary datasets. *Bioinformatics*, 27(19):2738–2745.

van de Schans, S. A. M., D. E. Issa, O. Visser, P. Nooijen, P. C. Huijgens, H.E. Karim-Kos, M. L. G. Janssen-Heijnen, and J. W. W. Coebergh. 2011. Diverging trends in incidence and mortality, and improved survival of non-Hodgkin's lymphoma, in the Netherlands, 1989-2007. *Annals of Oncology*, 23(1):171–182.

Skerman, H. M., P. M. Yates, and D. Battistutta. 2009. Multivariate methods to identify cancer-related symptom clusters. *Research in Nursing & Health*, 32(3):345–360.

Stein, K. D., K. L. Syrjala, and M. A. Andrykowski. 2008. Physical and psychological long-term and late effects of cancer. *Cancer*, 112:2577–2592.

Sweeney, C., K. H. Schmitz, D. Lazovich, B. A. Virnig, R. B. Wallace, and A. R. Folson. 2006. Functional limitations in elderly female cancer survivors. *Journal of the National Cancer Institute*, 98(8):521–529.

Tatsiana, K. 2013. *Statistical methods for analysis of high throughput experiments in early drug development*. Ph.D. thesis, Universiteit Hasselt.

Tatsiana, K. 2014. *superbiclust: Generating robust biclusters from a bicluster set (ensemble biclustering)*. R package version 1.1.

Thavarajah, N., E. Chen, L. Zeng, G. Bedard, J. D. Giovanni, M. Lemke, N. Lauzon, M. Zhou, D. Chu, and E. Chow. 2012. Symptom clusters in patients with metastatic cancer: A literature review. *Expert Review of Pharmacoeconomics & Outcomes*, 12(5):597–604.

van de Poll-Franse, L., S. Oerlemans, A. Bredart, C. Kyriakou, M. Sztankay, S. Pallua, L. Daniëls, C. L. Creutzberg, K. Cocks, S. Malak, G. Caocci, S. Molica, W. Chie, and F. Efficace. 2017.

Iinternation development of four EORTC disease-specific quality of life questionnaires for patients with Hodgkin lymphoma, high- and low-grade non-Hodgkin lymphoma and chronic lymphocytic leukaemia. *Qualify of Life Research*, 27(2):333–345.

van de Poll-Franse, L. V., N. Horevoorts, M. van Eenbergen, J. Denollet, J. A. Roukeman, N. K. Aaronson, A. Vingerhoets, J. W. Coeberg, J. de Vries, M. L. Essink-Bot, and F. Mols. 2011. The patient reported outcomes following initial treatment and long term evaluation of survivorship registry: Scope, rationale and design of an infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts. *European Journal of Cancer*, 47(14):2188–2194.

Velicer, W. F. and D. N. Jackson. 1990. Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, 25(1):1–28.

Verdecchia, A., S. Francisci, H. Brenner, G. Gatta, A. Micheli, L. Mangone, and I. Kunkler. 2007. The EUROCARE-4 working group. Recent cancer survival in Europe: a 2000-02 period analysis of EUROCARE-4 data. *The Lancet Oncology*, 8(9):784–796.

Walsh, D. and L. Rybicki. 2006. Symptom clustering in advanced cancer. *Supportive Care in Cancer*, 14(8):831–836.

Zou, H., T. Hastie, and R. Tibshirani. 2006. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

Appendix A: EORTC QLQ-C30 & EORTC QLQ-CLL17

integraal kankercentrum Zuid

Universiteit van Tilburg

Leven na de diagnose Lymfeklierkanker

Dit is een vragenlijst voor mensen die zijn gediagnosticeerd met lymfeklierkanker zoals het Hodgkin lymfoom, non-Hodgkin lymfoom of chronische lymfatische leukemie. Vult u de vragenlijst zelf, op uw gemak, in. U kunt steeds antwoord geven door het hokje/cijfer aan te kruisen dat het beste op u van toepassing is. Als u twijfelt, geef dan toch het antwoord dat het dichtst in de buurt komt van uw situatie. Er zijn geen goede of foute antwoorden; het gaat alleen om uw persoonlijke mening. Hoewel sommige vragen op elkaar kunnen lijken, is toch iedere vraag weer anders. Het kan ook zijn dat sommige vragen voor u overbodig of eigenlijk niet op u van toepassing lijken. Wilt u toch proberen alle vragen te beantwoorden?

De antwoorden op deze vragenlijst worden vertrouwelijk behandeld en uitsluitend anoniem gebruikt voor dit onderzoek.

Datum:

☐☐ - ☐☐ - ☐☐☐☐

Studienummer:

☐☐☐☐☐☐☐☐☐☐☐

Invulinstructie

- Vul de vragenlijst in met een zwarte of blauwe pen (geen viltstift).
- Zet een duidelijk kruisje in het antwoordvakje.
- Als u een fout antwoord invult, laat dan het foutieve kruisje staan en maak het goede vakje **helemaal zwart.**
- Vul bij een getal één cijfer per vakje in. Het hele cijfer moet binnen het vakje komen. Geen streepjes zetten als u iets niet hoeft in te vullen.
- Kruis bij elke vraag één hokje aan. Als er bij een vraag meer antwoorden gegeven mogen worden staat dit aangegeven.

Uw gezondheid

Wij zijn geïnteresseerd in bepaalde dingen over u en uw gezondheid. Wilt u alle vragen zelf beantwoorden door een kruisje te zetten onder het antwoord dat het meest op u van toepassing is. Er zijn geen "juiste" of "onjuiste" antwoorden.

		Helemaal niet	Een beetje	Nogal	Heel veel
1.	Heeft u moeite met het doen van inspannende activiteiten zoals het dragen van een zware boodschappentas of een koffer?	❒	❒	❒	❒
2.	Heeft u moeite met het maken van een lange wandeling?	❒	❒	❒	❒
3.	Heeft u moeite met het maken van een korte wandeling buitenshuis?	❒	❒	❒	❒
4.	Moet u overdag in bed of in een stoel blijven?	❒	❒	❒	❒
5.	Heeft u hulp nodig met eten, aankleden, u zelf wassen of naar het toilet gaan?	❒	❒	❒	❒

Gedurende de afgelopen week:		Helemaal niet	Een beetje	Nogal	Heel erg
6.	Was u beperkt bij het doen van uw werk of andere dagelijkse bezigheden?	❒	❒	❒	❒
7.	Was u beperkt in het uitoefenen van uw hobby's of bij andere bezigheden die u in uw vrije tijd doet?	❒	❒	❒	❒
8.	Was u kortademig?	❒	❒	❒	❒
9.	Heeft u pijn gehad?	❒	❒	❒	❒
10.	Had u behoefte te rusten?	❒	❒	❒	❒
11.	Heeft u moeite met slapen gehad?	❒	❒	❒	❒
12.	Heeft u zich slap gevoeld?	❒	❒	❒	❒
13.	Heeft u gebrek aan eetlust gehad?	❒	❒	❒	❒
14.	Heeft u zich misselijk gevoeld?	❒	❒	❒	❒
15.	Heeft u overgegeven?	❒	❒	❒	❒

Gedurende de afgelopen week:	**Helemaal niet**	**Een beetje**	**Nogal**	**Heel erg**
16. Had u last van obstipatie (was u verstopt)?	❒	❒	❒	❒
17. Had u diarree?	❒	❒	❒	❒
18. Was u moe?	❒	❒	❒	❒
19. Heeft pijn u gehinderd in uw dagelijkse bezigheden?	❒	❒	❒	❒
20. Heeft u moeite gehad met het concentreren op dingen, zoals een krant lezen of televisie kijken?	❒	❒	❒	❒
21. Voelde u zich gespannen?	❒	❒	❒	❒
22. Maakte u zich zorgen?	❒	❒	❒	❒
23. Voelde u zich prikkelbaar?	❒	❒	❒	❒
24. Voelde u zich neerslachtig?	❒	❒	❒	❒
25. Heeft u moeite gehad met het herinneren van dingen?	❒	❒	❒	❒
26. Heeft uw lichamelijke toestand of medische behandeling uw familieleven in de weg gestaan?	❒	❒	❒	❒
27. Heeft uw lichamelijke toestand of medische behandeling u belemmerd in uw sociale bezigheden?	❒	❒	❒	❒
28. Heeft uw lichamelijke toestand of medische behandeling financiële moeilijkheden met zich meegebracht?	❒	❒	❒	❒

Wilt u voor de volgende vragen het getal tussen de 1 en 7 aankruisen dat het meest op u van toepassing is.

29. Hoe zou u uw algehele gezondheid gedurende de afgelopen week beoordelen?

Erg slecht						Uitstekend
1	2	3	4	5	6	7
❒	❒	❒	❒	❒	❒	❒

30. Hoe zou u uw algehele "kwaliteit van het leven" gedurende de afgelopen week beoordelen?

Erg slecht						Uitstekend
1	2	3	4	5	6	7
❒	❒	❒	❒	❒	❒	❒

Klachten

Soms melden patiënten dat ze de volgende symptomen of problemen hebben. Gelieve aan te duiden in welke mate u deze symptomen of problemen gedurende de afgelopen week heeft ervaren. Kruis het hokje aan onder het antwoord dat het meest op u van toepassing is.

Gedurende de afgelopen week:	**Helemaal niet**	**Een beetje**	**Nogal**	**Heel erg**
31. Bent u afgevallen?	❒	❒	❒	❒
32. Had u een droge mond?	❒	❒	❒	❒
33. Liep u kneuzingen op?	❒	❒	❒	❒
34. Heeft u een onprettig gevoel in uw buik gehad?	❒	❒	❒	❒
35. Ging uw temperatuur op en neer?	❒	❒	❒	❒
36. Zweette u ‘s nachts?	❒	❒	❒	❒
37. Heeft u huidproblemen gehad (bijv. jeukerig, droog)?	❒	❒	❒	❒
38. Voelde u zich ziek of onwel?	❒	❒	❒	❒
39. Voelde u zich lusteloos?	❒	❒	❒	❒
40. Voelde u zich ‘futloos’?	❒	❒	❒	❒
41. Had u tintelende handen of voeten?	❒	❒	❒	❒
42. Was u beperkt in het plannen van activiteiten (bv. met vrienden afspreken)?	❒	❒	❒	❒
43. Maakte u zich zorgen over uw toekomstige gezondheidstoestand?	❒	❒	❒	❒

Tijdens de laatste vier weken:	**Helemaal niet**	**Een beetje**	**Vrij veel**	**Heel erg**
44. Had u last van luchtweginfecties?	❒	❒	❒	❒
45. Had u last van andere infecties?	❒	❒	❒	❒
46. Had u herhaaldelijk een antibioticabehandeling nodig?	❒	❒	❒	❒
47. Maakte u zich zorgen dat u een infectie zou oplopen?	❒	❒	❒	❒

Appendix B: Subgroup Demographics Bimax Model 2

Table 1
Subgroup 0 demographics (survivors not included in any subgroup) ($N = 410$)

Variable		Variable	
Age (in years); mean (range)	63.7 (22.3-83.8)	**BMI; mean (range)**	26.4 (18.1-43.6)
Gender	N (%)	**Type**	N (%)
Male	279 (68)	CLL	86 (21)
Female	131 (0.32)	Agressive NHL	180 (43.9)
Years since diagnosis	N (%)	Indolent NHL	126 (30.7)
< 2	106 (25.9)	Other NHL	18 (4.4)
\geq 2 and < 5	165 (40.2)	**Paid job**	N (%)
\geq 5 and < 10	135 (32.9)	Yes	120 (29.3)
\geq 10	4 (1.0)	No	290 (70.7)
SES	N (%)	**Marital Status**	N (%)
Low	79 (19.3)	Married / Cohabiting	333 (81.2)
Medium	147 (35.9)	Divorced / Separated	18 (4.4)
High	179 (43.7)	Widowed	41 (10)
Living in care institution	5 (1.2)	Never married / Never cohabited	18 (4.4)
Treatment	N (%)	**Education**	N (%)
Wait and see	106 (25.9)	Higher education	113 (27.6)
Chemotherapy	195 (47.6)	Medium education	236 (57.6)
Radiotherapy	44 (10.7)	Lower education	61 (14.9)
Transplantation	2 (0.5)		
Radio and Chemo	57 (13.9)		
Other therapies	4 (1.0)		
Surgery	2 (0.5)		

Table 2
All survivors included in a subgroup ($N = 398$)

Variable		Variable	
Age (in years); mean (range)	64.1 (21.7-85.1)	**BMI; mean (range)**	26.5 (17.2-43.5)
Gender	N (%)	**Type**	N (%)
Male	226 (56.8)	CLL	78 (19.6)
Female	172 (43.2)	Agressive NHL	180 (45.2)
Years since diagnosis	N (%)	Indolent NHL	125 (31.4)
< 2	104 (26.1)	Other NHL	15 (3.8)
\geq 2 and < 5	150 (37.7)	**Paid job**	N (%)
\geq 5 and < 10	136 (34.2)	Yes	99 (24.9)
\geq 10	8 (2)	No	299 (75.1)
SES	N (%)	**Marital Status**	N (%)
Low	81 (20.4)	Married / Cohabiting	317 (79.6)
Medium	174 (43.7)	Divorced / Separated	24 (6)
High	135 (33.9)	Widowed	34 (8.5)
Living in care institution	8 (2)	Never married / Never cohabited	23 (5.8)
Treatment	N (%)	**Education**	N (%)
Wait and see	109 (27.4)	Higher education	64 (16.1)
Chemotherapy	199 (50)	Medium education	266 (66.8)
Radiotherapy	30 (7.5)	Lower education	68 (17.1)
Transplantation	1 (0.3)		
Radio and Chemo	52 (13.1)		
Other therapies	7 (1.8)		
Surgery	0		

Table 3
Subgroup 1 demographics ($N = 54$)

Variable		Variable	
Age (in years); mean (range)	62.3 (27.7-85)	**BMI; mean (range)**	26.3 (18.3-43.5)
Gender	N (%)	**Type**	N (%)
Male	39 (72.2)	CLL	11 (20.4)
Female	15 (27.8)	Agressive NHL	27 (50)
Years since diagnosis	N (%)	Indolent NHL	15 (27.8)
< 2	13 (24.1)	Other NHL	1 (1.9)
\geq 2 and < 5	27 (50)	**Paid job**	N (%)
\geq 5 and < 10	13 (24.1)	Yes	11 (20.4)
\geq 10	1 (1.9)	No	43 (79.6)
SES	N (%)	**Marital Status**	N (%)
Low	17 (31.5)	Married / Cohabiting	42 (77.8)
Medium	24 (44.4)	Divorced / Separated	6 (11.1)
High	11 (20.4)	Widowed	2 (3.7)
Living in care institution	2 (3.7)	Never married / Never cohabited	4 (7.4)
Treatment	N (%)	**Education**	N (%)
Wait and see	17 (31.5)	Higher education	9 (16.7)
Chemotherapy	22 (40.7)	Medium education	32 (59.3)
Radiotherapy	6 (11.1)	Lower education	13 (24.1)
Transplantation	0		
Radio and Chemo	8 (14.8)		
Other therapies	1 (1.9)		
Surgery	0		

Table 4
Subgroup 2 demographics ($N = 40$)

Variable		Variable	
Age (in years); mean (range)	63.05 (21.7-84.3)	**BMI; mean (range)**	27.3 (19.7-40.1)
Gender	N (%)	**Type**	N (%)
Male	18 (45)	CLL	4 (10)
Female	22 (55)	Agressive NHL	22 (55)
Years since diagnosis	N (%)	Indolent NHL	13 (32.5)
< 2	6 (15)	Other NHL	1 (2.5)
\geq 2 and < 5	15 (37.5)	**Paid job**	N (%)
\geq 5 and < 10	19 (47.5)	Yes	9 (22.5)
\geq 10	0	No	31 (77.5)
SES	N (%)	**Marital Status**	N (%)
Low	10 (25)	Married / Cohabiting	31 (77.5)
Medium	17 (42.5)	Divorced / Separated	4 (10)
High	12 (30)	Widowed	3 (7.5)
Living in care institution	1 (2.5)	Never married / Never cohabited	2 (5)
Treatment	N (%)	**Education**	N (%)
Wait and see	8 (20)	Higher education	4 (10)
Chemotherapy	21 (52.5)	Medium education	26 (65)
Radiotherapy	2 (5)	Lower education	10 (25)
Transplantation	0		
Radio and Chemo	9 (22.5)		
Other therapies	0		
Surgery	0		

Table 5
Subgroup 3 demographics ($N = 41$)

Variable		Variable	
Age (in years); mean (range)	64.7 (23.3-83.7)	**BMI; mean (range)**	26 (18.6-38.8)
Gender	N (%)	**Type**	N (%)
Male	20 (48.8)	CLL	6 (14.6)
Female	21 (51.2)	Agressive NHL	19 (46.3)
Years since diagnosis	N (%)	Indolent NHL	12 (29.3)
< 2	17 (41.5)	Other NHL	4 (9.8)
\geq 2 and < 5	14 (34.1)	**Paid job**	N (%)
\geq 5 and < 10	9 (22)	Yes	9 (22)
\geq 10	1 (2.4)	No	32 (78)
SES	N (%)	**Marital Status**	N (%)
Low	7 (17.1)	Married / Cohabiting	29 (70.7)
Medium	19 (46.3)	Divorced / Separated	4 (9.8)
High	14 (34.1)	Widowed	5 (12.2)
Living in care institution	1 (2.4)	Never married / Never cohabited	3 (7.3)
Treatment	N (%)	**Education**	N (%)
Wait and see	13 (31.7)	Higher education	4 (9.8)
Chemotherapy	17 (41.5)	Medium education	35 (85.4)
Radiotherapy	3 (7.3)	Lower education	2 (4.9)
Transplantation	1 (2.4)		
Radio and Chemo	6 (14.6)		
Other therapies	1 (2.4)		
Surgery	0		

Table 6
Subgroup 4 demographics ($N = 43$)

Variable		Variable	
Age (in years); mean (range)	61.5 (29.6-82)	**BMI; mean (range)**	26.8 (18-38.2)
Gender	N (%)	**Type**	N (%)
Male	28 (65.1)	CLL	13 (30.2)
Female	15 (34.9)	Agressive NHL	13 (30.2)
Years since diagnosis	N (%)	Indolent NHL	17 (39.5)
< 2	14 (32.6)	Other NHL	0
\geq 2 and < 5	11 (25.6)	**Paid job**	N (%)
\geq 5 and < 10	18 (41.9)	Yes	16 (37.2)
\geq 10	0	No	27 (62.8)
SES	N (%)	**Marital Status**	N (%)
Low	12 (27.9)	Married / Cohabiting	37 (86)
Medium	18 (41.9)	Divorced / Separated	1 (2.3)
High	12 (27.9)	Widowed	3 (7)
Living in care institution	1 (2.3)	Never married / Never cohabited	2 (4.7)
Treatment	N (%)	**Education**	N (%)
Wait and see	11 (25.6)	Higher education	8 (18.6)
Chemotherapy	23 (53.5)	Medium education	23 (53.5)
Radiotherapy	5 (11.6)	Lower education	12 (27.9)
Transplantation	0		
Radio and Chemo	4 (9.3)		
Other therapies	0		
Surgery	0		

Table 7
Subgroup 5 demographics ($N = 45$)

Variable		Variable	
Age (in years); mean (range)	62.9 (27.5-84.4)	**BMI; mean (range)**	27.7 (17.2-37.7)
Gender	N (%)	**Type**	N (%)
Male	31 (68.9)	CLL	10 (22.2)
Female	14 (31.1)	Agressive NHL	21 (46.7)
Years since diagnosis	N (%)	**Indolent NHL**	14 (31.1)
< 2	15 (33.3)	Other NHL	0
\geq 2 and < 5	16 (35.6)	**Paid job**	N (%)
\geq 5 and < 10	13 (28.9)	Yes	16 (35.6)
\geq 10	1 (2.2)	No	29 (64.4)
SES	N (%)	**Marital Status**	N (%)
Low	7 (15.6)	Married / Cohabiting	36 (80)
Medium	23 (51.1)	Divorced / Separated	3 (6.7)
High	14 (31.1)	Widowed	3 (6.7)
Living in care institution	1 (2.2)	Never married / Never cohabited	3 (6.7)
Treatment	N (%)	**Education**	N (%)
Wait and see	14 (31.1)	Higher education	10 (22.2)
Chemotherapy	25 (55.6)	Medium education	30 (66.7)
Radiotherapy	2 (4.4)	Lower education	5 (11.1)
Transplantation	0		
Radio and Chemo	4 (8.9)		
Other therapies	0		
Surgery	0		

Table 8
Subgroup 6 demographics ($N = 37$)

Variable		Variable	
Age (in years); mean (range)	64.8 (36.7-84.7)	**BMI; mean (range)**	27.2 (20.2-40)
Gender	N (%)	**Type**	N (%)
Male	21 (56.8)	CLL	14 (37.8)
Female	16 (43.2)	Agressive NHL	15 (40.5)
Years since diagnosis	N (%)	Indolent NHL	7 (18.9)
< 2	8 (21.6)	Other NHL	1 (2.7)
\geq 2 and < 5	11 (29.7)	**Paid job**	N (%)
\geq 5 and < 10	14 (37.8)	Yes	4 (10.8)
\geq 10	4 (10.8	No	33 (89.2)
SES	N (%)	**Marital Status**	N (%)
Low	5 (13.5)	Married / Cohabiting	30 (81.1)
Medium	15 (40.5)	Divorced / Separated	2 (5.4)
High	16 (43.2)	Widowed	2 (5.4)
Living in care institution	1 (2.7)	Never married / Never cohabited	3 (8.1)
Treatment	N (%)	**Education**	N (%)
Wait and see	12 (32.4)	Higher education	8 (21.6)
Chemotherapy	17 (45.9)	Medium education	26 (70.3)
Radiotherapy	3 (8.1)	Lower education	3 (8.1)
Transplantation	0		
Radio and Chemo	3 (8.1)		
Other therapies	2 (5.4)		
Surgery	0		

Table 9
Subgroup 7 demographics ($N = 41$)

Variable		Variable	
Age (in years); mean (range)	67.8 (38.8-84.1)	**BMI; mean (range)**	25.2 (18-36.8)
Gender	N (%)	**Type**	N (%)
Male	19 (46.3)	CLL	8 (19.5)
Female	22 (53.7)	Agressive NHL	17 (41.5)
Years since diagnosis	N (%)	Indolent NHL	15 (36.6)
< 2	11 (26.8)	Other NHL	1 (2.4)
\geq 2 and < 5	21 (51.2)	**Paid job**	N (%)
\geq 5 and < 10	9 (22)	Yes	9 (22)
\geq 10	0	No	32 (78)
SES	N (%)	**Marital Status**	N (%)
Low	7 (17.1)	Married / Cohabiting	34 (82.9)
Medium	17 (41.5)	Divorced / Separated	1 (2.4)
High	17 (41.5)	Widowed	5 (12.2)
Living in care institution	0	Never married / Never cohabited	1 (2.4)
Treatment	N (%)	**Education**	N (%)
Wait and see	14 (34.1)	Higher education	8 (19.5)
Chemotherapy	18 (43.9)	Medium education	26 (63.4)
Radiotherapy	2 (4.9)	Lower education	7 (17.1)
Transplantation	0		
Radio and Chemo	5 (12.2)		
Other therapies	2 (4.9)		
Surgery	0		

Table 10
Subgroup 8 demographics ($N = 55$)

Variable		Variable	
Age (in years); mean (range)	63.9 (34.4-85.1)	**BMI; mean (range)**	25.8 (18.1-37.4)
Gender	N (%)	**Type**	N (%)
Male	22 (40)	CLL	10 (18.2)
Female	33 (60)	Agressive NHL	24 (43.6)
Years since diagnosis	N (%)	Indolent NHL	18 (32.7)
< 2	10 (18.2)	Other NHL	3 (5.5)
\geq 2 and < 5	18 (32.7)	**Paid job**	N (%)
\geq 5 and < 10	26 (47.3)	Yes	16 (29.1)
\geq 10	1 (1.8)	No	39 (70.9)
SES	N (%)	**Marital Status**	N (%)
Low	9 (16.4)	Married / Cohabiting	45 (81.8)
Medium	26 (47.3)	Divorced / Separated	1 (1.8)
High	19 (34.5)	Widowed	6 (10.9)
Living in care institution	1 (1.8)	Never married / Never cohabited	3 (5.5)
Treatment	N (%)	**Education**	N (%)
Wait and see	13 (23.6)	Higher education	6 (10.9)
Chemotherapy	32 (58.2)	Medium education	37 (67.3)
Radiotherapy	4 (7.3)	Lower education	12 (21.8)
Transplantation	0		
Radio and Chemo	6 (10.9)		
Other therapies	0		
Surgery	0		

Table 11
Subgroup 9 demographics ($N = 42$)

Variable		Variable	
Age (in years); mean (range)	67 (45.2-83.8)	**BMI; mean (range)**	26.2 (19.9-39.8)
Gender	N (%)	**Type**	N (%)
Male	28 (66.7)	CLL	2 (4.8)
Female	14 (33.3)	Agressive NHL	22 (52.4)
Years since diagnosis	N (%)	Indolent NHL	14 (33.3)
< 2	10 (23.8)	Other NHL	4 (9.5)
\geq 2 and < 5	17 (40.5)	**Paid job**	N (%)
\geq 5 and < 10	15 (35.7)	Yes	9 (21.4)
\geq 10	0	No	33 (78.6)
SES	N (%)	**Marital Status**	N (%)
Low	7 (16.7)	Married / Cohabiting	33 (78.6)
Medium	15 (35.7)	Divorced / Separated	2 (4.8)
High	20 (47.6)	Widowed	5 (11.9)
Living in care institution	0	Never married / Never cohabited	2 (4.8)
Treatment	N (%)	**Education**	N (%)
Wait and see	7 (16.7)	Higher education	7 (16.7)
Chemotherapy	24 (57.1)	Medium education	31 (73.8)
Radiotherapy	3 (7.1)	Lower education	4 (9.5)
Transplantation	0		
Radio and Chemo	7 (16.7)		
Other therapies	1 (2.4)		
Surgery	0		