Identification and potential use of colorectal and prostate patient clusters in clinical practice

Beuken, Maik Jozef Maria; Kanera, Iris Maria; Ezendam, Nicole Paulina Maria; Braun, Susy Michelle; Zoet, Martijn

DOI:
10.2196/preprints.42908

Publication date:
2022

Document Version
Early version, also known as pre-print

Identification and Potential Use of Colorectal and Prostate Patient Clusters in Clinical Practice: An Explorative mixed methods Study

Maik Jozef Maria Beuken, Iris Maria Kanera, Nicole Paulina Maria Ezendam, Susy Michelle Braun, Martijn Zoet

Submitted to: JMIR Cancer
on: September 23, 2022

Table of Contents

Identification and Potential Use of Colorectal and Prostate Patient Clusters in Clinical Practice: An Explorative mixed methods Study

Maik Jozef Maria Beuken[1] MEd; Iris Maria Kanera[2] PhD; Nicole Paulina Maria Ezendam[3] PhD; Susy Michelle Braun[2] PhD; Martijn Zoet[1] PhD

[1]Faculty of Financial Management, Research Centre for Future Proof Financials Zuyd University of Applied Sciences Sittard NL
[2]Faculty of Health, School of Physiotherapy, Research Centre for Nutrition, Lifestyle and Exercise Zuyd University of Applied Sciences Heerlen NL
[3]Department of Research Netherlands Comprehensive Cancer Organization (IKNL) Utrecht NL

Corresponding Author:
Maik Jozef Maria Beuken MEd
Faculty of Financial Management, Research Centre for Future Proof Financials
Zuyd University of Applied Sciences
Ligne 1
Sittard
NL

Abstract

Background: A steady increase in colorectal and prostate cancer patients and survivors is expected in the upcoming years. Due to primary cancer treatments, patients suffer from numerous additional complaints, which also increases the need for cancer aftercare. However, referrals to appropriate cancer aftercare remain inadequate, despite a wide range of aftercare options. Caregivers and patients often do not know which aftercare is the most appropriate for the individual patient. Since characteristics and complaints of patients within a diagnosis group can be different, predefined patient clusters could provide substantive and efficient support for professionals in the conversation about aftercare. By using advanced data analysis methods, clusters of patients who are different from one another within one diagnosis group can be identified.

Objective: The objective of this study was twofold: first, to identify, visualize, and describe potential patient clusters within colorectal and prostate cancer populations and, second, to explore the potential usability of these clusters in clinical practice.

Methods: First, we used cross-sectional data from colorectal and prostate cancer patients provided by the population-based Patient Reported Outcomes Following Initial Treatment and Long Term Evaluation of Survivorship registry, which was originally collected between 2008 and 2012. To identify and visualize different clusters among the two patient populations, we conducted cluster analyses by applying the K-means algorithm and multiple-factor analyses. Second, in a qualitative study, we presented the patient clusters to prostate and colorectal cancer patients and oncology professionals. To assess the usability of these clusters, we held expert panel group interviews. The interviews were videorecorded and transcribed. Three researchers independently performed content-directed data analysis to understand and describe the qualitative data. Quotes illustrate the most important results.

Results: We identified 3 patient clusters among colorectal cancer cases (N=3989) and 5 patient clusters among the prostate cancer cases (N=696), which were described in tabular form. Patient-experts (N=6) and professional-experts (N=17) recognized the patient clustering based on distinguishing variables. However, the tabular form was evaluated as less applicable in clinical practice. Instead, the experts suggested the development of a conversation tool (eg, decision tree) to guide professionals through the hierarchy of variables. In addition, participants suggested that information about possible aftercare initiatives should be offered and integrated. This would also ensure a good overview and seemed to be a precondition for finding suitable aftercare.

Conclusions: This study demonstrates that a fully data-driven approach can be used to identify distinguishable and in-routine care recognizable patient clusters in large datasets within cancer populations. Challenges for the future include the identification of more distinguishing key variables, the development of a smart digital conversation and referral tool, and the further development of new data analysis techniques to detect normal and abnormal recovery patterns among cancer patients. Clinical Trial: Trial ID NL9226 (Trial Register, The Netherlands)

(JMIR Preprints 23/09/2022:42908)

DOI: https://doi.org/10.2196/preprints.42908

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Authors

Maik Jozef Maria Beuken MEd[1], Iris Maria Kanera Phd[2], Nicole Paulina Maria Ezendam PhD[3,4], Susy Michelle Braun PhD[2], Martijn Zoet PhD[1],
[1] Faculty of Financial Management, Research Centre for Future Proof Financials, Zuyd University of Applied Sciences, Ligne 1, 6231 MT Sittard, PO Box 69, 6130 AB Sittard, The Netherlands
[2] Faculty of Health, School of Physiotherapy, Research Centre for Nutrition, Lifestyle and Exercise, Zuyd University of Applied Sciences, Nieuw Eyckholt 300, 6419 DJ Heerlen, PO Box 550, 6400 AN Heerlen, The Netherlands

[3] Department of Research, Netherlands Comprehensive Cancer Organization (IKNL), Utrecht, The Netherlands

[4] Department of Medical and Clinical Psychology, CoRPS - Center of Research on Psychology in Somatic Diseases, Tilburg University, Tilburg, The Netherlands

Corresponding author

Maik Beuken, maik.beuken@zuyd.nl

Identification and Potential Use of Colorectal and Prostate Patient Clusters in Clinical Practice: An Explorative mixed methods Study

Abstract

Background: A steady increase in colorectal and prostate cancer patients and survivors is expected in the upcoming years. Due to primary cancer treatments, patients suffer from numerous additional complaints, which also increases the need for cancer aftercare. However, referrals to appropriate

cancer aftercare remain inadequate, despite a wide range of aftercare options. Caregivers and patients often do not know which aftercare is the most appropriate for the individual patient. Since characteristics and complaints of patients within a diagnosis group can be different, predefined patient clusters could provide substantive and efficient support for professionals in the conversation about aftercare. By using advanced data analysis methods, clusters of patients who are different from one another within one diagnosis group can be identified.
Objective: The objective of this study was twofold: first, to identify, visualize, and describe potential patient clusters within colorectal and prostate cancer populations and, second, to explore the potential usability of these clusters in clinical practice.
Methods: First, we used cross-sectional data from colorectal and prostate cancer patients provided by the population-based Patient Reported Outcomes Following Initial Treatment and Long Term Evaluation of Survivorship registry, which was originally collected between 2008 and 2012. To identify and visualize different clusters among the two patient populations, we conducted cluster analyses by applying the K-means algorithm and multiple-factor analyses. Second, in a qualitative study, we presented the patient clusters to prostate and colorectal cancer patients and oncology professionals. To assess the usability of these clusters, we held expert panel group interviews. The interviews were videorecorded and transcribed. Three researchers independently performed content-directed data analysis to understand and describe the qualitative data. Quotes illustrate the most important results.
Results: We identified 3 patient clusters among colorectal cancer cases (N=3989) and 5 patient clusters among the prostate cancer cases (N=696), which were described in tabular form. Patient-experts (N=6) and professional-experts (N=17) recognized the patient clustering based on distinguishing variables. However, the tabular form was evaluated as less applicable in clinical practice. Instead, the experts suggested the development of a conversation tool (eg, decision tree) to guide professionals through the hierarchy of variables. In addition, participants suggested that information about possible aftercare initiatives should be offered and integrated. This would also ensure a good overview and seemed to be a precondition for finding suitable aftercare.
Conclusions: This study demonstrates that a fully data-driven approach can be used to identify distinguishable and in-routine care recognizable patient clusters in large datasets within cancer populations. Challenges for the future include the identification of more distinguishing key variables, the development of a smart digital conversation and referral tool, and the further development of new data analysis techniques to detect normal and abnormal recovery patterns among cancer patients.

Trial registration number: Trial ID NL9226 (Trial Register, The Netherlands)

Keywords:
Colorectal cancer; prostate cancer; referral to aftercare; patient clusters; cluster analysis; K-means cluster algorithm; multiple-factor analysis; expert panel group interviews

Introduction

Currently, cancer represents one of the major healthcare problems. Worldwide, in 2020, the incidence of all forms of cancer was higher than 18 million cases. Colorectal and prostate cancer are 2 of the top 4 most diagnosed cancers [1]. In 2020, in the Netherlands alone, approximately 11,500 new cases of colorectal cancer and over 12,000 new cases of prostate cancer were reported [2]. Within the next two decades, these annual numbers in the Netherlands are expected to increase by 35% for colorectal cancer cases and 25% for prostate cancer cases. Fortunately, due to improved diagnostics and treatments, 10-year survival of prostate cancer has risen to above 70% and of colorectal cancer up to almost 60% [2].

Cancer survivors are at a higher risk of developing new forms of cancer and comorbidities, as well as long-term physical, lifestyle, and psychosocial problems and difficulties with work. Consequently, an increasing number of survivors require information and support [3,4]. Earlier research has indicated that adequate cancer aftercare can support survivors to increase and maintain health, wellbeing, and quality of life [5-7].{Balhareth, 2019 #105;Kanera, 2017 #106;Willems, 2017 #107}{Maechler, , cluster citation info}

The European Academy of Cancer Sciences and several European organizations and cancer centers have emphasized the urgency of tailored aftercare in their published research agenda to reduce the major cancer burden and improve the health-related quality of life by promoting cost-effective, evidence-based best practices in cancer prevention, treatment, care and aftercare [8]. One of their recommendations for psychosocial oncology, rehabilitation, and survivorship research is to develop tools to enhance communication with patients and shared decision-making, such as the development and testing of decision aids for selecting aftercare. These are also key points in the recently published Dutch National Cancer & Life Action Plan [9].

In this paper, we explore the potential benefits and barriers of patient clusters within the referral process. Referral to an aftercare option might be more appropriate and faster if distinguishing characteristics are taken into account. Clustering patient groups with similar characteristics may provide substantive and efficient support for professionals in the conversation about aftercare. For clustering, we consider variables, which are related to long-term problems after cancer, such as sociodemographic, health-related, psychosocial, lifestyle factors, and quality of life. To verify this fully data-driven approach in daily practice we combine it with a qualitative evaluation among professionals and former and current cancer patients.

The aim of this study was twofold: first, to identify, visualize, and describe potential patient clusters within colorectal and prostate cancer populations and, second, to explore the potential usability of these patient clusters in clinical practice.

Methods

In part one, we address the first aim of identifying, visualizing, and describing patient clusters. The clinical usability of the identified patient clusters is reported in the second part. This study was carried out in accordance with the Ethics committee METC- Z, ID number METCZ20200203.

PART ONE: Patient Clusters

Design

To identify patient clusters, we used cross-sectional data from the population-based Patient-Reported Outcomes Following Initial treatment and Long-term Evaluation of Survivorship (PROFILES) [10]. The PROFILES registry collects patient-reported outcomes in a large cohort to study the psychosocial and physical impact of cancer and its treatment. PROFILES data are available for non-commercial scientific research, subject to study question, privacy and confidentially restrictions, and registration [11].

Study Population

From the PROFILES registry, we included 2 patient samples with colorectal cancer collected between 2008 and 2011 and one patient sample with prostate cancer collected between 2011 and 2012. A detailed description of the data collection method within the PROFILES registry has been reported elsewhere [10]. A population-based sampling frame was used, where patients were selected from the Netherlands Cancer Registry from a selected set of participating hospital. Patients needed to be able to complete a Dutch questionnaire and be 18 years or older. Patients were invited by their treating physician.

Measurements

For the cluster analysis, we used all available variables from the PPROFILES dataset provided, including the following self-reported measures: sociodemographic information (regarding marital status, educational level, and employment), socioeconomic status [12], and emotional and cognitive functioning. We included all available patient-related outcome measurements (Appendix 1).

Statistical Analysis

We conducted the data analyses on colorectal and prostate cancer samples separately. We merged both colorectal cancer samples and assessed all data for aberrant measurement data, missing data, and outliers.

Missing data were imputed by using the K nearest-neighbor method (KNN, VIM package) [13]. All variables were used to impute missing values. In the KNN function, the distance computation was based on an extension of the Gower distance [14]. For continuous variables, we used the median to give a central measurement for the 5 nearest neighbors that were used to impute a missing value, and, for categorical variables, we used the mode to impute [13].

We used RStudio (4.0.3 (2020-10-10), R Foundation for Statistical Computing) as a programming language.

Further handling of missing data, including data imputation and the handling of outliers, as well as other used software packages, are described in Appendix 2.

Identification of Patient Clusters

To assign patients to clusters, we performed a K-means cluster algorithm. By using the K-means algorithm after data-cleaning, individual cases were clustered into a k number of clusters using the squared Euclidean distance variable [15]. We minimized the distance between so-called centroids (one centroid for each cluster) and the objects of each cluster. To evaluate the result of the K-means algorithm (number of clusters), we used the silhouette coefficient (SC), which gives a measure for the cohesion and segregation of each data point [16]. The closer the SC value gets to the value of 1, the stronger the cohesion of data points within one cluster and the segregation between data points within one cluster relative to data points in another cluster. We determined the optimal number of patient clusters by the highest SC value for each diagnosis group.

Visualization and Description of Patient Clusters

To enable visualization and to describe characteristics of the identified patient clusters, we employed a multiple-factor analysis (MFA) [17]. Since the patient clusters consisted of quantitative and qualitative variables, we applied a factorial method to visualize the mutual relationships of the variables. We mapped quantitative variables by using the correlation circle based on principal components analysis (PCA). Qualitative variables, as well as the cluster numbers, were visualized by using the individual factor map [18]. We grouped positively correlated variables in a correlation circle, which was visualized by arrows that lie together in the same direction in the correlation circle. Negatively correlated variables were presented opposed to each other. The further away the variables lied from the center of the correlation circle—visualized by longer arrows—the better these variables were represented within the concept (ie, a particular topic is assessed by a number of questions; those questions together illuminate a concept (eg, perception is a concept that is elucidated by 8 items of the BIPQ questionnaire)). For each concept, we performed this MFA analysis based on the prostate and colorectal cancer data (Appendix 3).

To standardize, we used a cut-off point of 0.5 for the quality of the projection of a variable on one of the dimensions in the correlation circle. The same threshold was applied for the individual factor map when describing the characteristics of the clusters. We accounted for the variables drawn above these thresholds.

The variables that clustered together based on these procedures were described in different patient clusters for colorectal cancer and prostate cancer separately.

PART TWO: Usability Study

Design

To assess the clinical usability of the identified patient clusters, we applied a qualitative approach by conducting expert panel group interviews. Due to the COVID-19 pandemic, the group interviews were held online.

Study population

Both professionals and cancer patients formed the panel of experts. Eligible healthcare professionals were professionals from various care disciplines with expertise in the field of oncology, including prostate or colorectal cancer. Eligible participants for the patient-expert panel were adult former and current patients with colorectal or prostate cancer who completed primary cancer treatment and may still receive adjuvant therapy. Other inclusion criteria included having basic computer skills, internet access, and a digital device with a camera and speakers.

Procedure and Data Collection

Through an information letter, we recruited potential participating healthcare professionals from two regional hospitals, a general practitioner society and an oncology physiotherapy network. These professionals approached other eligible health professionals and patients (snowball sampling). The researchers assessed the eligibility criteria, and detailed information was offered by phone. All participants provided informed consent before enrolment in the study.

We interviewed the professional-expert panel, the colorectal cancer patient–expert panel, and the prostate cancer patient–expert panel separately. We held semi-structured group interviews based on a topic list (Appendix 4) with a maximum duration of 120 minutes to gain insight into the potential clinical usability of the identified patient clusters as assessed by the healthcare professionals and cancer patients. The group interviews followed a fixed structure. After a short introduction of the project in which the purpose of the meeting was explained again, the patient clusters were presented to the panel, and the following topics were discussed: (1) the number of the patient clusters and recognizability of the content, (2) the forms of cancer aftercare that best fit each cluster, (3) the usefulness, meaningfulness, and opportunities of patient clusters concerning tailor-made aftercare referral, and (4) the preconditions for implementing patient clusters in clinical practice. Prior to the group interviews, the participants received information about the patient clusters and regional cancer aftercare possibilities. Moreover, they received a brief online questionnaire in order to gather information about personal characteristics. The participating healthcare professionals additionally received some preparation questions.

Data Analysis Expert Panels

We analyzed personal characteristics descriptively. Video recordings and additional notes of the online group interviews were analyzed based on an abridged transcript. We employed content-directed analysis [19] to describe and understand the collected qualitative data systematically [20]. We coded and categorized the data based on the structure of the topics and questions in line with the topic list. Three researchers (IK, health scientist; PE, health scientist; AK, student research assistant) independently performed the coding and categorizing. To increase trustworthiness, four researchers (WE, research assistant; RJ, student research assistant; IK, health scientist; PE, health scientist), reviewed the codes and categories and reached an agreement on the results [21]. Subsequently, the participants received a summary of the key points for verification of the content (member check).

Results

PART ONE: Patient Clusters

In total, 3989 colorectal-cancer cases (1371 participants in the colorectal 2009 wave and 2618 participants in the colorectal 2010 wave) and 696 prostate cancer cases were included in the cluster analysis (Table 1). Participants varied in age between 29 and 85 years. Description of all characteristics appears in Appendix 5.

Table 1. Basic Characteristics of Participants with Colorectal Cancer (N = 3989) and Prostate Cancer (N = 696)

Variable	**Category**	Colorectal cancer	Prostate cancer
Gender, N (%)			
	Male Female	2,220 (55.6) 1,769 (44.4)	696 (100.0) 0 (0.0)
Age in years, M[a] (SD[b])			
	At the time of diagnosis	64.7 (9.8)	67.4 (7.3)
	At time of questionnaire	69 (9.6)	70.8 (7.2)
Marital status, N (%)			
	Married Divorced Widowed Never married	3011 (75.5) 204 (5.1) 640 (16.0) 134 (3.4)	586 (84.2) 27 (3.9) 65 (9.3) 18 (2.6)
Educational level, N (%)			

	Lower education Secondary education Secondary vocational education University	777 (19.5) 1247 (31.3) 1179 (29.6) 786 (9.7)	117 (16.8) 162 (23.3) 249 (35.8) 168 (24.1)
Employment status, N (%)			
	Yes No	604 (15.1) 3385 (84.9)	89 (12.8) 607 (87.2)
Socioeconomic status, N (%)			
	1. Low 2. Medium 3. High 4. Living in a nursing home	833 (20.9) 1631 (40.9) 1454 (36.4) 71 (1.8)	118 (17.0) 270 (38.8) 292 (41.9) 16 (2.3)
Body Mass Index, M (SD)			
		26.7 (4.2)	26.5 (3.3)
Assigned numbering cluster, N (%)			
	Cluster no. 1 Cluster no. 2 Cluster no. 3 Cluster no. 4 Cluster no. 5	1788 (44.8) 1144 (28.7) 1057 (26.5) N/A[c] N/A	197 (28.3) 85 (12.2) 144 (20.7) 159 (22.8) 111 (16.0)

[a]M = Mean
[b]SD = Standard Deviation
[c]N/A = Not Applicable

Identification of Patient Clusters

We calculated the highest SC value was calculated within the prostate-cancer sample for 5 patient clusters and the highest SC value within the colorectal-cancer sample for 3 patient clusters (Table 2).

Table 2. Silhouette Coefficients Per Diagnosis Group and Number of Clusters

	Colorectal cancer, N = 3989	Prostate cancer, N = 696
	SC[a]	SC
K[b]=3		
	0.15127883	0.06516783
K=4		
	0.04583991	0.09671970
K=5		
	0.04350592	0.13308123
K=6		
	0.06356493	0.11919456
K=7		
	0.04058240	0.08339187
K=8		
	0.02672504	0.09207357
K=9		
	0.01644865	0.06816820
K=10		
	0.01136183	0.04821369

[a]SC = Silhouette coefficient
[b]K = Number of patient clusters

The main distinguishing characteristics of the patient clusters are described in Table 3.

Table 3. Main Characteristics of the Patient Clusters for Colorectal Cancer and Prostate Cancer

	Colorectal cancer, N = 3989	Prostate cancer, N = 696
	Interpretation clusters	Interpretation clusters
Patient cluster 1		

	• Higher socioeconomic status • Have a lower BMI • More patients who have been diagnosed with their disease some time ago • Drink alcohol more often, mainly wine • More patients who exercise or do sports • Lower stage of disease • Do not frequently have an appointment with the specialist and have no need for one • Have fewest comorbidities • Sense a small effect on their lives because of their illness • More likely to think that their illness will not last long, have a sense of control, and are confident that the treatment will work • Have a high understanding of their disease • Recognize fewer symptoms and worry less about their illness • Experience a small emotional effect Score high on the functioning scales, including the highest on emotional functioning and quality of life.	• Younger • Relatively higher education but not highest education • More often have a paid job • More smokers • Tend to drink alcohol more often • Do not feel well informed, are less satisfied with the information they receive, and find that information less helpful • Use the internet more often to find information about their disease.
Patient cluster 2		
	• Lower socioeconomic status • Have a higher BMI • More often elderly patients who are widows or widowers • More often have lower education • More patients who have been diagnosed with their disease a shorter time ago • More often deceased • Tend to represent less alcohol users and smokers • Least active in terms of exercise • Have most often a higher stage of the disease • Visit more often the general practitioner and specialist about cancer • Discussed to come back more often • Have a higher number of comorbidities • Problems with personality and fatigue on a physical and mental level and more characterized by anxiety and depression • More likely to report a high degree of impact on their lives; think the illness will last longer • Indicate a lower level of control • Experience many symptoms • Have a high degree of concern about their illness • Feel an extreme effect on an emotional level • Have reasonable confidence in the success of their treatment • Score lower on the functioning scales • Score high on fatigue, breath shortness, insomnia, pain, loss of appetite, nausea, and vomiting.	• Younger • More often have higher education Higher socio-economic status • Lower stage of disease • Tend to drink alcohol more often, even more than cluster 1 • More liver problems • Understand their illness better and have more confidence in their treatment • Higher score on the physical, emotional, and social scales and lower score on fatigue and pain • Feel better informed and have less need for more information about their disease • Use the internet more often to find information about their disease.
Patient cluster 3		
	• Younger • More often divorced • Higher representation of middle	• Lower education • Lower socio-economic status • Do household tasks more often

	socioeconomic status and people who live in an institution • More often patients who have a job • Drink alcohol more often • More patients who exercise or do sports • Have a higher stage of disease compared to cluster 1 • More often have an appointment with the specialist regarding cancer and have also discussed returning to the specialist more often compared to cluster 1 • Have fewer comorbidities, but depression is more common • Relatively fewer problems with personality, fatigue, and depression compared to cluster 2 • Relatively more fears and more negative affectation compared to cluster 1 • Have a more neutral perception of their disease • Not very distinctive on quality of life	• More often stopped drinking alcohol • More comorbidities • Have a more negative self-image, feel a greater impact on their lives and emotions, and are more concerned • Lower score on the physical, emotional, and social scales and higher score on fatigue and pain • Do not feel well informed, are less satisfied with the information they receive, and find that information less helpful.
Patient cluster 4		
		• Higher education but not the highest • More often have an advanced stage of disease • More often deceased • More often disabled due to their disease • More often stopped drinking alcohol • More comorbidities • Have a more negative self-image, illness has a greater impact on their lives and emotions, and are more concerned • Lower score on the physical, emotional, and social scales and higher score on fatigue and pain.
Patient cluster 5		
		• Lower education • Lower socio-economic status • More often stay in a nursing home • More often without a partner • More often stopped drinking alcohol • Understand their illness better and have more confidence in their treatment • Use the internet less often to find information about their disease.

Visualization and Description of Patient Clusters

We described participant characteristics of five prostate cancer patient clusters and the three colorectal cancer clusters in Table 3 based on the MFA analysis. Not all the same concepts were measured in the different data sets available (ie, colorectal data and prostate data), as displayed in Table 1. As a result, certain concepts could not be reflected in the clusters.

PART TWO: Usability

Expert Panel Participants

Twenty-three people participated in this part of the study (Table 4). Of the 8 patient experts approached, 6 filled in the brief online questionnaire (prostate cancer N = 3; colorectal cancer N = 3), and 5 took part in the group interviews. Reasons for not participating included not wanting to participate digitally (n=1) and an emergency medical appointment (n=1). One person did not state a reason (n=1). Of the 20 professional-experts, 17 participated. Reasons for non-participation were maternity leave (n=1), no time (n=1), and unknown (no response, n=1).

Table 4. Characteristics of Expert Panel Participants (N = 23)

	Patient experts	Professional experts
	N = 6	N = 17
Gender, female, N (%)		
	1 (16.7)	13 (76.5)
Age, median (min-max)		
	60 (48-79)	48 (33-64)
Prostate cancer diagnosis, N (%)		
	3 (50)	
Colorectal cancer diagnosis, N (%)		
	3 (50)	
Time since diagnosis, median (min-max)		
	2.8 (1-8)	
Still cancer detected during control visit, N (%)		
	2 (33.3)	
Nurse specialist hospital, N (%)		
		2 (11.8)
Nurse specialist general practice, N (%)		
		2 (11.8)
General practitioner, N (%)		
		2 (11.8)
Internist oncologist, N (%)		
		2 (11.8)
Psychologist, N (%)		
		2 (11.8)
Oncology physiotherapist, N (%)		
		2 (11.8)
Oncology surgeon, N (%)		
		1 (5.9)
Rehabilitation physician, N (%)		
		1 (5.9)
Complementary health therapist/lifestyle coach, N (%)		
		1 (5.9)
Acupuncturist, herbalist, N (%)		
		1 (5.9)
Staff advisor oncology, N (%)		
		1 (5.9)
Years of work experience (oncology), median (min-max)		
		15 (0.5-40)
Cancer aftercare provider, N (%)		
		14 (82.4)

Expert Panel Interviews

In total, 7 group interviews took place. We conducted one group interview with prostate cancer patients (N = 3) and one with colorectal cancer patients (N = 2). Five professional expert panel group interviews took place in varying compositions regarding the profession and with a group size between 3 and 5 participants. One individual interview was conducted.

Clinical Usability of the Patient Clusters

Most of the participants recognized the clustering as distinctive 'profiles,' and all variables described were assessed as important factors regarding tailored referral to aftercare. They indicated that the variables follow a certain hierarchy that should be taken into account when considering referral to appropriate aftercare. The expert panel stated that providing the description of the clusters in tabular form with many variables outlined in text was too difficult to oversee. Moreover, participants were concerned that patients would be placed into fixed categories by using this tabular format. Furthermore, a conversation with patients would be necessary to clarify the support needs. The clusters could also serve as a valuable starting point and guidance for this conversation because they provide meaningful content and structure.

> "*Care providers often don't look beyond their specialism. A broad view is missing. Other fields should also be considered in the conversation about aftercare.*" [Prostate cancer patient]

Therefore, participants suggested the development of a conversation tool that could provide insight into the content and structure of these clusters. To guide professionals through the hierarchy of variables, a decision tree could be integrated into this tool. In addition, participants suggested access to information about available aftercare initiatives be made available. This would also ensure a good overview and seemed to be a precondition for finding suitable aftercare.

> "*As a patient, you don't know what the disease entails and what you can expect, so you don't know what aftercare you need. You need to be well-informed; only then do you know what you need!*" [Colorectal cancer patient]

> "*You are very much searching and constantly re-telling your whole story. It would be nice to have a choice of pre-sorted relevant options of aftercare. The disease already costs you a lot of energy. Searching also takes a lot of energy!*" [Colorectal cancer patient]

The tool content should be comprehensive, clearly structured, and easy to use. The patient, not the professional or the application, should always make the final decision on aftercare. The professional experts also wished to link existing data from the electronic patient files to the decision tool.

> "*Using a decision aid based on the patient clusters would be a good tool for care providers to gain a better understanding and to get an overview when it comes to referral to the right aftercare.*" [Prostate cancer patient]

> "*This kind of tool could take the administrative burden off the nurses' shoulders.*" [Oncology specialist]

Discussion

This study aimed to 1) identify, visualize, and describe patient clusters within colorectal and prostate cancer populations and to 2) explore the potential usability of the patient clusters in clinical practice to improve referral to cancer aftercare.

We identified, described, and presented 5 patient clusters among a prostate cancer population and three patient clusters among a colorectal cancer population to an expert panel for evaluation.

Most notably, by performing the cross-sectional data-analysis, we included all available variables in the datasets without any human pre-selection and the number of patient clusters was solely determined by the SC. Our approach to cluster the data of individuals based on their characteristics is consistent with the situation in clinical practice, in which an oncology professional encounters a patient with individual characteristics. In our results, easily detectable characteristics, such as age, employment status, and socioeconomic status clustered with less easily recognizable characteristics, such as illness perception. This interrelationship between different characteristics can support healthcare providers in the conversation with patients to ultimately refer to appropriate follow-up care.

Contrary to our method, de Rooij et al. [22] explored the relation of symptoms among a selection of PROFILES registry variables in their network analysis (ie, EORTC QLQ- C30 symptom scales and the emotional and cognitive functioning scales). Noticeably, however, our results among colorectal cancer data are in line with the findings of de Rooij et al. regarding the corresponding variables (eg, fatigue, pain, dyspnea, sleeping problems, appetite loss, and nausea and vomiting), which might strengthen our findings.

Professional and patient experts considered the insight that different subgroups can be distinguished within one diagnosis group to have been valuable for ultimately referring patients to the appropriate aftercare. Participants largely recognized the classification into the clusters. However, the expert panel deemed the way of presenting the clusters in textual tabular form as standalones to be unpractical for routine care. In order to have a meaningful conversation about referral to appropriate aftercare, professionals and patients would like to have guidance to help them discuss relevant topics, which then can lead to the most suitable choices for cancer aftercare. Therefore, a complete overview of current aftercare initiatives is also needed. The experts suggested developing a digital decision and referral aid based on the patient clusters to detect the patient's support needs and risks and link them to the available aftercare options.

Overall, this study succeeded in identifying patient clusters that are also seen in routine care and recognized by healthcare professionals. Results show that the presented holistic, explorative machine-learning approach can provide a foundation to identify clinically meaningful patient clusters. Consequently, our results can serve as a first step to improve referrals to cancer aftercare in daily practice, which is in line with the goals of the Taskforce Cancer Survivorship [8,9].

Limitations

Like all research, this study has its limitations. Data of participants were not highly distinguishable for all variables because not all answer options were distinguishable (ie, the distinguishing variables had a lot of overlap and were therefore not good indicators for distinguishing between clusters). This problem could technically be solved by using a larger number of patient clusters. However, this would be less appropriate for clinical use, because a larger number of clusters makes it difficult for professionals to get an overview of the clusters.

The data from the PROFILES registry was generated about 10 years ago, while we retrieved the data from the qualitative study in 2020. However, we do not expect a negative impact from this time difference, as we assume that cancer patients are not significantly different now than they were 10 years ago.

Finally, we interviewed mainly professional experts, patient experts and their opinions were relatively underrepresented. Consequently, we may not have achieved data saturation.

Future Directions

Since the identification and use of patient clusters among colorectal and prostate cancer populations is still in its infancy, future research should further focus on identifying distinguishing key variables in order to optimize the number and content of patient clusters. Building upon a data-driven approach, as shown in this study, an additional expert-driven approach could provide a qualitative improvement of the selection of variables. Both patient and professional experts should be equally involved in this process. Researchers should explore in what form a digital referral aid could be of added value in clinical practice. Our results might provide valuable insights as a basis for the development of smart referral technology.

Furthermore, identifying longitudinal patient patterns, based on data gathered over time, might be a next step to generate insights into the course of the patients' situation and about deviations from 'expected recovery.' The process of identifying patient patterns could be automated by creating a data tunnel linked to electronic patient records and by automatically generating trend analyses that can provide insights into the development of the individuals' disease and recovery over time.

Conclusions

This study demonstrates that a fully data-driven approach can be used to identify distinguishable and recognizable patient clusters in large datasets within colorectal and prostate cancer populations. Presenting the clusters in tabular form does not provide the support needed for professionals and patients to arrive at a balanced decision about appropriate cancer aftercare. Challenges for the future involve the development of a smart digital conversation and referral tool based on relevant key topics and the further development of new data analysis techniques to detect normal and abnormal recovery patterns among cancer patients.

Acknowledgments

The authors thank Dr. Kees van Berkel for advice on the writing of the scripts and for advice regarding the quantitative data analyses. Dr. Laura Hochstenbach took responsibility for the acquisition of the data from the PROFILES study and provided input for the first analysis of the prostate data. We also thank Dr. Pieter Eijgenraam, Willem Emons, Roy Jorissen, and Alina Kramme for their contributions to the development, conducting, and analysis of the interviews. Moreover, we thank all involved patients and professionals for their time, cooperation, and valuable input to this study.

M.B., I.K., and S.B. designed the study. M.B. conducted the quantitative sub-study. I.K. and S.B. conducted the qualitative sub-study. M.B. wrote the script for the data cleaning, the K-means algorithm, and the MFA. I.K. and M.B. performed the analyses and denoted the clusters in conversation with M.Z. I.K. and M.B. drafted the manuscript, and S.B. and M.Z. reviewed all versions of the manuscript. All authors read and contributed to editing the manuscript for final submission.

Conflicts of Interest

None declared.

Abbreviations

BIPQ: Brief illness perception questionnaire
BMI: Body mass index
FAS: Fatigue assessment scale
HADS: Hospital anxiety and depression scale
KNN: K nearest neighbor
MFA: Multiple-factor analysis

NA: Not available
NNI: Nearest neighbor imputation
PCA: Principal Components analysis
RCT: Randomized controlled trial
SC: Silhouette coefficient

Appendix

Appendix 1. Variables included in the cluster analysis available from PROFILES.

Lifestyle was assessed by questioning, tobacco and alcohol use, and comorbidity were assessed by using the Comorbidity questionnaire [23]. To assess the cognitive and emotional representations of illness, eight items of the Brief Illness Perception Questionnaire (BIPQ) were included [24]. Clinical and cancer-related data were used which originated from the Netherlands Cancer Registry [10], comprising the time since diagnosis, age at the time of diagnosis, age at the time of filling out the questionnaire, body mass index (BMI), TNM tumor classification, vital status, cancer treatment (eg, surgery, systemic treatment, radiotherapy, hormonal therapy, no treatment, treatment unknown). Various items were used to assess the utilization of cancer care [25]. Data on health-related quality of life were used by including all subscales from the EORTC-QLQ-C30 [26]. This data was not available in the 2009 dataset of colorectal cancer patients. The following data was only present in both colorectal cancer samples: data on physical activity, type D personality (DS-14 subscales negative affectivity and social inhibition) [27,28], fatigue including the subscales physical fatigue and mental fatigue from the Fatigue Assessment Scale (FAS) [29], anxiety and depression including the two subscales for anxiety and depression of the Hospital Anxiety and Depression Scale (HADS) [30]. From only the prostate cancer dataset, data from the EORTC QLQ-INFO26 concerning the perception of received information could be used [31].

Appendix 2. Handling missing data including imputation and handling outliers and used software packages.

Handling missing data

The data analyses on colorectal and prostate cancer samples were conducted separately. The two colorectal cancer samples were merged and all data were assessed for outliers and aberrant measurement data. Complete data sets are an important precondition for performing the cluster analysis and therefore non-responders and variables with more than 50% missing data (non-available's; NA's) were removed.

NAs were imputed conform recommendations of Kalton & Kasprzyk [32] and Rubin [33] by using the nearest neighbor imputation (NNI) technique which is appropriate to apply for survey data with a high number of respondents [34-36]. In this study, the fife nearest neighbors-imputation technique was applied which is derived from the NNI by using donor observations of the actual data [13,37], with nearest defined by a distance function of the auxiliary variables [38]. This imputation method is applicable for samples with multiple missing values and suitable for both discrete and continuous variables [39]. This method leads to a consistent imputation that is based on all included variables.

Handling outliers

To downsize the effects of large size variables (or having a great variability) on cluster analysis, several standardization methods were conducted [40]. To accommodate extreme outliers in continuous variables (except for BMI) Winsorized Trimming was conducted which replaces outliers on the high side (and low side) by the next value to the highest (respectively lowest) value within the boundary of the outer fence [41]. To highlight values that are considered to be extreme outliers the outer fences are set to three times the interquartile range [41]. Continuous variables were standardized with a z-score-standardization method for normally distributed variables, whereas skewed data is standardized with a Min-Max-standardization method. For nominal or ordinal variables with categories that had a lower number of objects, lower than the square root of the total number of objects, the categories were aggregated with the nearest category to the category with low numbers. Subsequently, these variables were standardized using dummy variables.

Removing near-zero variance variables

Variables with a near-zero variance were removed because they do not contribute information and therefore the minority of the values that are represented in a near-zero variable could have an undue influence on the model [42].

Used Software packages

- FactoMineR, used to exploratory analyze the data with respect to identifying hidden patterns in the dataset. In particular to use the MFA for variables structured in groups [43].
- Factoextra, used to create and visualize the output of multivariate data analyses with Multiple Factor Analysis (MFA) [44].
- Provides ggplot2 Cluster, used to cluster the data with the K-means algorithm.
- MASS, used to support Venables and Ripley [45].
- VIM package, used to impute missing data with the use of the KNN method [13].

Appendix 3. The interpretation of the MFA.

In the correlation circle on the right-hand side in figure 2, derived from the prostate cancer data, the relationship between variables considering the concept: lifestyle, in terms of how many glasses of wine do you drink a day and how many cigarettes do you smoke a day (amongst other questions, see table 1), the quality of the representation and the correlation between these variables and the dimensions are shown. The first dimension mostly correlates positively with wine consumption as does the second dimension with cigarette consumption, positioned opposed to this the time since a participant quit smoking is shown and is the variable that correlates negatively with the second dimension.
In the plot on the left-hand side (figure 1) the qualitative variables considering the concept: of lifestyle are shown. Participants who smoke (ROOK_3) have positive coordinates on the second axis along with participants that are clustered in clusters 1 and 4 (assignment5_1 and assignment5_4), thus cluster 1 can be looked at as the group where the number of participants who smoke is more represented. In cluster 2 (assignment5_2), people who drink more wine are mainly represented, at the same time, these are the people who have more often stopped smoking for a longer time ago, both variables score low on the second dimension. Cluster 5 (assignment5_5) often includes participants who have stopped drinking alcohol (ALCOHOL_2) or who indicate that they do not drink alcohol at all (ALCOHOL_1). For clusters 3 and 4 this picture is not so clear.
Most of the other qualitative or quantitative variable categories are close to the origin. This indicates that these categories are not related to the first or second dimension.

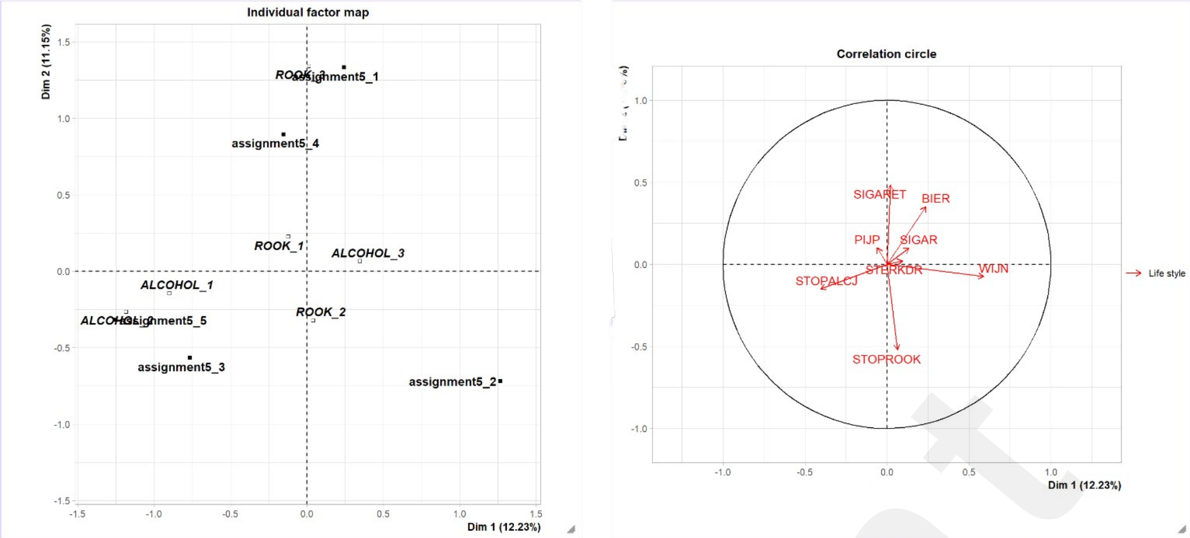


Figure 1. MFA-plot

Appendix 4. Topic list expert-panel interview.

1. The recognizability of content and number of the patient clusters
    a. In your opinion, to what extent are the characteristics described recognizable as important characteristics for referral to aftercare?
        i. Would you consider adding or removing certain characteristics?
    b. To what extent do you recognize the composition of characteristics of the clusters from your own experience?
        i. Would you consider merging certain profiles? If so, which ones?
        ii. Would you consider adding a completely different profile? If so, what are the characteristics that this new profile should comprise?
    c. In your opinion, which characteristics within a profile are crucial in the choice for suitable aftercare?

2. Linking cancer aftercare possibilities to patient clusters (provided with an overview of all available options in the region)
    a. Looking at cluster x and imagining a patient who fits this cluster, what form(s) of cancer aftercare could be appropriate for this patient, in your opinion?

3. Usability and opportunities of patient clusters in health care practice
    a. To what extent are patient clusters useful when it comes to improving referral to cancer aftercare?
    b. In what way could the patient clusters actually be used in clinical practice?

4. Preconditions for implementing patient clusters in clinical practice
    a. What are the essential preconditions for implementing patient clusters in healthcare practice?
    b. What kind of barriers do you expect?
    c. In which particular ways do you think patient clusters can be successfully brought into clinical practice?

Appendix 5. Remaining characteristics of colorectal cancer (N = 3989) and prostate cancer (N = 696) participants.

Variable	**Category**	Colorectal cancer	Prostate cancer
Stage disease, N (%)			
	1	1060 (26.6)	N/A[a]
	2	295 (7.4)	N/A
	2A	1097 (27.5)	N/A
	2B	123 (3.1)	N/A
	3	215 (5.4)	N/A
	3A	136 (3.4)	N/A
	3B	585 (14.7)	N/A
	3C	240 (6.0)	N/A
	4	187 (4.7)	N/A
	X	39 (1.0)	N/A
	[2 others]	12 (0.3)	N/A
	1	N/A	2 (0.3)
	2	N/A	497 (71.4)
	3	N/A	132 (19.0)
	4	N/A	65 (9.3)
Vital status, N (%)			
	Alive	2937 (73.6)	560 (80.5)
	Deceased	1052 (26.4)	136 (19.5)
Time since diagnosis in years, M[b] (SD[c])			
		4.8 (2.7)	4 (1.2)
Treatment (KR data), N (%)			
	Surgery	3946 (98.9)	197 (28.3)
	Radiotherapy	1094 (27.4)	273 (39.2)
	Systemic therapy	1193 (29.9)	0 (0.0)
	Hormonal therapy	3 (0.1)	209 (30.0)
	No therapy or active surveillance	4 (0.1)	130 (18.7)
Number of consults in the past 12 months, M (SD)			
	General practitioner	4 (6.1)	3.4 (2.9)
	General practitioner, due to cancer	1.2 (3.5)	N/A
	Specialist	1.1 (4.8)	1.2 (2.1)
	Specialist due to cancer	3 (4.3)	2.2 (1.6)
Still follow up Appointments, N (%)			
	Yes	3149 (78.9)	655 (94.1)
	No	840 (21.1)	41 (5.9)
Discussed with specialist how often to come back from this moment on, N (%)			
	Yes, every 3 months	473 (11.9)	74 (10.6)
	Yes, every 4 months	142 (3.6)	26 (3.7)
	Yes, every 6 months	1574 (39.5)	360 (51.7)
	Yes, once a year	950 (23.8)	200 (28.7)
	Yes, every 2 years	327 (8.2)	4 (0.6)
	No	523 (13.1)	32 (4.6)
Comfortable with follow up scheme, N (%)			
	Yes	3567 (89.4)	637 (91.5)
	No, want more follow up	218 (5.5)	29 (4.2)

	No, want less follow up No, want no follow up	73 (1.8) 131 (3.3)	14 (2.0) 16 (2.3)
Received cancer aftercare, N (%)			
	Received aftercare overall	N/A	349 (50.1)
	Psychologist	N/A	20 (2.9)
	Sexologist	N/A	6 (0.9)
	Social worker	N/A	3 (0.4)
	Pastoral worker	N/A	1 (0.1)
	General practitioner	N/A	42 (6.0)
	Dietitian	N/A	12 (1.7)
	Physiotherapist	N/A	108 (15.5)
	Recovery group program	N/A	10 (1.4)
	Creative therapy	N/A	2 (0.3)
	Oncological nurse	N/A	16 (2.3)
	Contact with fellow patients/ survivors	N/A	7 (1.0)
	Others	N/A	43 (6.2)
Comorbidities, N (%)			
	Heart condition	727 (18.2)	141 (20.3)
	Stroke	96 (2.4)	18 (2.6)
	High Blood pressure	1288 (32.3)	223 (32.0)
	Long disease	399 (10.0)	80 (11.5)
	Diabetes	536 (13.4)	94 (13.5)
	Ulcer	52 (1.3)	10 (1.4)
	Kidney disease	134 (3.4)	19 (2.7)
	Liver disease	124 (3.1)	2 (0.3)
	Anemia	127 (3.2)	29 (4.2)
	Thyroid disease	189 (4.7)	16 (2.3)
	Depression	250 (6.3)	43 (6.2)
	Arthritis	988 (24.8)	156 (22.4)
	Backache	994 (24.9)	171 (24.6)
	Rheumatism	238 (6.0)	45 (6.5)
Number of hours paid job, M (SD)			
		4.5 (11.7)	4.1 (11.9)
Unable to work, due to cancer, N (%)			
	Not applicable I was always able to work I wasn't able to work	N/A N/A N/A	624 (89.7) 18 (2.6) 54 (7.8)
Number of hours unable to work per week, M (SD)			
		N/A	1 (4.1)
Employment status, N (%)			
	Having a job Pension/early retirement Scholar/student Unemployed	604 (15.1) 2819 (70.7) 1 (0.0) 40 (1.0)	89 (12.8) 558 (80.2) 0 (0.0) 6 (0.9)

	Disabled Managing the household Other	226 (5.7) 220 (5.5) 79 (2.0)	29 (4.2) 3 (0.4) 11 (1.6)
Disability percentage, M (SD)			
		3.5 (18)	2.9 (16.3)
Disability due to the disease, N (%)			
	NA Yes No	3789 (95.0) 130 (3.3) 70 (1.8)	674 (96.8) 5 (0.7) 17 (2.4)
Smoking, N(%)			
	No No, but I used to Yes	1284 (32.2) 2267 (56.8) 438 (11.0)	154 (22.1) 459 (66.0) 83 (11.9)
Time since stopped smoking in years, M (SD)			
		N/A	16.1 (16.4)
Number of cigarettes per day, M (SD)			
		1.3 (4.8)	1.1 (4.3)
Number of cigars per week, M (SD)			
		0.5 (4.7)	0.5 (4.1)
Number of packages of pipe tobacco per week, M (SD)			
		0 (0.1)	0 (0.3)
Alcohol consumption, N (%)			
	No No, but I used to Yes	1055 (26.5) 361 (9.0) 2573 (64.5)	90 (12.9) 84 (12.1) 522 (75.0)
Time since stopped drinking in years, M (SD)			
		N/A	1.3 (5.4)
Glasses/consumption per week, M (SD)			
	Beer Wine Liquor	1.7 (4.7) 2.7 (5.1) 0.9 (3.1)	2.8 (5.2) 3.2 (5.4) 1.4 (3.5)
Physical Activity, hours per week, M (SD)			
	Walking summer Walking winter Biking summer Biking winter Gardening summer Gardening winter Household summer Household winter	5.2 (5.4) 3.9 (4.6) 4.9 (7) 2 (3.5) 3 (4.7) 0.7 (1.6) 7.9 (10.1) 7.7 (10.1)	N/A
Weekly sporting activities in the past year, N (%)			
	No Yes	2674 (67.0) 1315 (33.0)	N/A
Type –D personality (DS-14)[d], M (SD)			
	Negative affectivity	7.3 (6.2)	N/A
	Social Inhibition	7.9 (6.2)	N/A
Illness Perception (BIPQ)[e], M (SD)			
	Affect on life	3.9 (2.6)	3.7 (2.5)
	Time illness continues	4.4 (3.4)	5.7 (3.6)

	Control over illness	5.1 (3.1)	5.3 (3.3)
	Treatment helps	7.3 (2.7)	7.5 (2.7)
	Experience symptoms	3.4 (2.6)	3.5 (2.6)
	Concerned about illness	4 (2.7)	3.7 (2.7)
	Understanding illness	6.9 (2.9)	7.4 (2.6)
	Illness affects emotionally	3.4 (2.5)	3.3 (2.6)
Fatigue (FAS)[f], M (SD)			
	Physical subscale	11.6 (4.1)	N/A
	Mental subscale	9.3 (3.6)	N/A
Anxiety and Depression (HADS)[g], M (SD)			
	Anxiety subscale	4.7 (3.8)	N/A
	Depression subscale	4.7 (3.6)	N/A
Health-Related Quality of life (EORTC QLQ-C30)[h] , M (SD)			
	Physical Functioning	71.1 (21.7)	83.1 (19.1)
	Role Functioning	74 (24.1)	81.2 (26.7)
	Emotional Functioning	79.8 (19.8)	87.4 (18.7)
	Cognitive Functioning	78.9 (18.7)	84.5 (20.1)
	Social Functioning	75.4 (26.6)	89.5 (19.4)
	Global health status	78.7 (16.3)	77.7 (18.1)
	Fatigue	28 (21.7)	19.9 (22.3)
	Nausea / Vomiting	10.5 (16.8)	2.2 (9)
	Pain	21.3 (26.1)	15.7 (24.3)
	Dyspnea	18.5 (24.9)	15.4 (25.5)
	Insomnia	29.3 (29)	18.4 (27.6)
	Appetite loss	8 (16.8)	3.3 (12.5)
	Constipation	7.6 (17.6)	6.7 (17.9)
	Diarrhea	9.3 (19.9)	5.3 (15.8)
	Financial Problems	5.3 (16.4)	4.5 (13.8)
Information (EORTC QLQ-INFO25), M (SD)			
	Treatment	N/A	2.9 (1)
	Disease	N/A	53.8 (21.5)
	Medical tests	N/A	62 (27.7)
	Other services	N/A	18.8 (23.5)
	Different places of care	N/A	17.5 (29)
	Things you can do to help yourself	N/A	22.4 (29.6)
	Written information	N/A	74.7 (43.5)
	On CD tape/video	N/A	5.3 (22,4)
	Satisfaction	N/A	60.1 (27.7)
	Wish for more	N/A	25.6 (43.6)
	Wish for less	N/A	3.6 (18.6)
	Helpful	N/A	64.7 (26.2)
Use of internet, N (%)			
	Daily Weekly Monthly No	N/A	336 (48.3) 103 (14.8) 31 (4.4) 226 (32.5)
Search information via the internet, N (%)			

	Yes No	N/A	352 (50.6) 344 (49.4)

Note:
[a]N/A = Not Applicable
[b]M = Mean
[c]SD = Standard Deviation
[d]Subscale used for Type-D personality; Negative affection (range: 0-28); Social inhibition (range: 0-28); type D if both N/A and SI score ≥ 10 [27]
[e]Brief Illness Perception Questionnaire (BIPQ); item score range: 0-10 [24].
[f]FAS; Subscale score range: 5-25 [46].
[g]HADS: subscale score range: 0-21 [30].
[h]EORTC QLQ-C30: item score range 0-100; Higher scores on functional scales represent higher levels of functioning and higher score for global health status represents a higher level of quality of life; high scores for the symptoms scales represent a higher level of problems [26].

1. Ferlay J EM, Lam F, Colombet M, et al. Cancer Today. Lyon: International Agency for Research on Cancer. https://gco.iarc.fr/today [accessed Oct 1, 2021].
2. Organisation NCC. IKNL and the NCR. https://iknl.nl/nkr [accessed Oct 1, 2021].
3. Kanera IM, Bolman CAW, Mesters I, et al. Prevalence and correlates of healthy lifestyle behaviors among early cancer survivors. BMC Cancer. 2016;16(1):4. doi:10.1186/s12885-015-2019-x
4. Willems RA, Bolman CA, Mesters I, et al. Cancer survivors in the first year after treatment: The prevalence and correlates of unmet needs in different domains. Psych-Oncol. 2016 Jan;25(1):51-7. doi:10.1002/pon.3870
5. Balhareth A, Aldossary MY, McNamara D. Impact of physical activity and diet on colorectal cancer survivors' quality of life: A systematic review. World J Surg Oncol. 2019;17(1):153-153. doi:10.1186/s12957-019-1697-2
6. Kanera IM, Willems RA, Bolman CA, et al. Long-term effects of a web-based cancer aftercare intervention on moderate physical activity and vegetable consumption among early cancer survivors: A randomized controlled trial. Int J Behav Nutr Phys Act. 2017 Feb;14(1):19. doi:10.1186/s12966-017-0474-2
7. Willems RA, Mesters I, Lechner L, Kanera IM, Bolman CAW. Long-term effectiveness and moderators of a web-based tailored intervention for cancer survivors on social and emotional functioning, depression, and fatigue: Randomized controlled trial. J Cancer Surviv. 2017 Dec;11(6):691-703. doi:10.1007/s11764-017-0625-0
8. Berns A, Ringborg U, Celis JE, et al. Towards a cancer mission in Horizon Europe: Recommendations. Mol Oncol. 2020 Aug;14(8):1589-1615. doi:10.1002/1878-0261.12763
9. Care TCS. Nationaal Actieplan Kanker & Leven. Taskforce Cancer Survivorship Care. https://taskforcecancersurvivorshipcare.nl/wp-content/uploads/2020/05/Visiedocument-TFCSC_2020_def.pdf [accessed Oct 1, 2021].
10. van de Poll-Franse LV, Horevoorts N, van Eenbergen M, et al. 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. Eur J Cancer. 2011 Sep;47(14):2188-94. doi:10.1016/j.ejca.2011.04.034
11. Profiles. Available from: https://www.profilesregistry.nl/ [accessed Feb 23, 2022].
12. van Duijn C, Keij I. Sociaal-economische status indicator op postcode niveau.Maandstatistiek van de bevolking.2002;50:32–35.
13. Kowarik A, Templ M. Imputation with the R package VIM. J Stat Sftwar. 2016;74(7):1-16. doi:10.18637/jss.v074.i07
14. Gower, J. A General Coefficient of Similarity and Some of Its Properties. Biomtrcs. 1971; 27(4):857–874. https://doi.org/10.2307/2528823. [access October 15, 2021]
15. Gan G, Ma C, Wu J. Data clustering: Theory, algorithms, and applications. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2007. ISBN:9780898718348
16. Rousseeuw PJ. Silhouettes: A graphical aid to interpretation and validation of

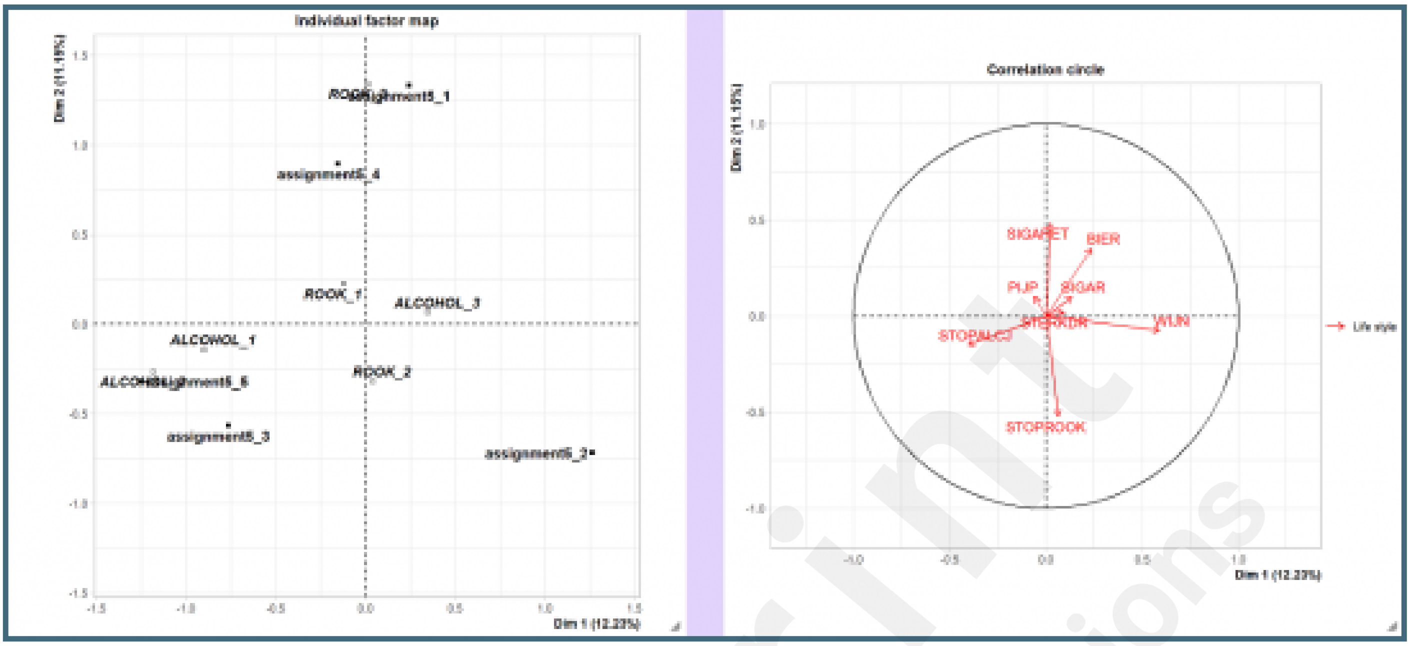
cluster analysis. J Comp & Appld Math. 1987 Nov;20:53-65. doi:10.1016/0377-0427(87)90125-7
17. Greenacre M, Blasius J. Multiple Correspondence Analysis and Related Methods. Boca Raton, FL: Chapman-Hall/CRC; 2006. ISBN:9781584886280
18. Abdi W, Valentin H, J L. Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. Wires Comp Stat. 2013. wires.wiley.com/compstats [accessed Feb 14, 2021].
19. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res*. 2005 Nov;15(9):1277-1288. doi:10.1177/1049732305276687
20. Verhoeven N. Thematische analyse. Patronen vinden bij kwalitatief onderzoek. Amsterdam, NL: Boom; 2020. ISBN:9789024427550
21. Korstjens I, Moser A. Series: Practical guidance to qualitative research. Part 4: Trustworthiness and publishing. Eur J Gen Pract. 2018 Dec;24(1):120-124. doi:10.1080/13814788.2017.1375092
22. De Rooij BH, Oerlemans S, van Deun K, et al. Symptom clusters in 1330 survivors of 7 cancer types from the PROFILES registry: A network analysis. Cancer. 2021 Dec 15;127(24):4665-4674. doi: 10.1002/cncr.33852
23. Sangha O, Stucki G, Liang MH, Fossel AH, Katz JN. The Self-Administered Comorbidity Questionnaire: a new method to assess comorbidity for clinical and health services research. Arthritis Rheum. 2003 Apr 15;49(2):156-163. doi:10.1002/art.10993
24. Broadbent E, Petrie KJ, Main J, Weinman J. The brief illness perception questionnaire. J Psychosom Res. 2006 Jun;60(6):631-637. doi:10.1016/j.jpsychores.2005.10.020.
25. Van de Poll-Franse LV, Mols F, Vingerhoets AJ, et al. Increased health care utilisation among 10-year breast cancer survivors. Sprt Care Cancer. 2006 May;14(5):436-443. doi: 10.1007/s00520-005-0007-4
26. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993 Mar 3;85(5):365-376. doi:10.1093/jnci/85.5.365
27. Denollet J. DS14: standard assessment of negative affectivity, social inhibition, and Type D personality. Psychosom Med. 2005 Jan-Feb;67(1):89-97. doi:10.1097/01.psy.0000149256.81953.49
28. Grande G, Romppel M, Glaesmer H, Petrowski K, Hermann-Lingen C. The type-D scale (DS14) – Norms and prevalence of type-D personality in a population-based representative sample in Germany. Prsnlty & Indvdl Diff. 2010 June;48(8):935-939. doi: 10.1016/j.paid.2010.02.026.
29. Michielsen HJ, de Vries J, van Heck GL, van de Vijver FJR, Sijtsma K. Examination of the dimensionality of fatigue: The construction of the Fatigue Assessment Scale (FAS). Eur J Psych Assmnt. 2004;20(1):39-48.
30. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand. 1983 June;67(6):361-370. doi:10.1111/j.1600-0447.1983.tb09716.x.
31. Arraras JI, Kuljanic-Vlasic K, Bjordal K, et al. EORTC Quality of Life Group. EORTC QLQ-INFO26: a questionnaire to assess information given to cancer patients

a preliminary analysis in eight countries. Psychooncology. 2007 Mar;16(3):249-254. doi:10.1002/pon.1047
32. Kalton G, Kasprzyk D. The treatment of missing survey data. Survey Methodology. 1986 Jun;12(1):1-16.
33. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ: John Wiley & Sons; 1987. ISBN: 9780470316696
34. Functional Description of the Generalized Edit and Imputation System. Statistics Canada; 1991.
35. Kovar JG, Whitridge P, MacMillan J. Generalized Edit and Imputation System for Economic Surveys at Statistics Canada. Proceedings of the Section on Survey Research Methods; American Statistical Association; 1988;690-695.
36. Rancourt E, Särndal CE, Lee H. Estimation of the Variance in the Presence of Nearest Neighbor Imputation. Proceedings of the Section on Survey Research Methods; Statistics Canada; 1994;888-893.
37. Chen S. Nearest neighbor imputation for survey data. J Official Stats. 2000 Jan;16(2):113–131.
38. Kalton G, Kasprzyk D. Imputing for missing survey responses. American Statistical Association. 1982:22-31.
39. Batista GEAPA, Monard MC. A study of K-nearest neighbour as a model-based method to treat missing data. 2001;1-9.
40. Mohamed IB, Usman D. Standardization and its effects on k-means clustering algorithm. Rsrch J App Sci, Eng & Tech. 2013 Sep;6(17):3299-3303. doi:10.19026/rjaset.6.3638
41. Tukey JW. Exploratory Data Analysis. Boston, MA: Addison-Wesley; 1977. ISBN:0201076160
42. Kuhn M, Johnson K. Applied Predictive Modeling. New York City, NY: Springer Link; 2013. ISBN: 9781461468493
43. Sébastien L, Josse J, Husson F. FactoMineR: An R package for multivariate analysis. J Stat Sftwar. 2008 Mar;25(1):1-18. doi:10.18637/jss.v025.i01
44. Factoextra: Extract and visualize the results of multivariate data analyses. 2020. https://CRAN.R-project.org/package=factoextra [accessed Jan 21, 2022].
45. Venables W, Ripley B. Modern Applied Statistics with S. 4th edition. New York City, NY: Springer Link; 2002. ISBN: 9780387217062
46. Hendriks C, Drent M, Elfferich M, De Vries J. The fatigue assessment scale: Quality and availability in sarcoidosis and other diseases. Curr Opin Pulm Med. 2018 Sep;24(5):495-503. doi:10.1097/mcp.0000000000000496

Supplementary Files

Figures

MFA-plot.


Individual factor map
Dim 2 (11.15%)
Dim 1 (12.23%)
assignment5_1
assignment5_4
ROOK_1
ALCOHOL_3
ALCOHOL_1
ROOK_2
assignment5_5
assignment5_3
assignment5_2
Correlation circle
Dim 2 (11.15%)
Dim 1 (12.23%)
SIGARET
BIER
PIJP
SIGAR
WIJN
STOPALC
STOPROOK
Life style

Multimedia Appendixes

Variables included in the cluster analysis available from PROFILES.
URL: http://asset.jmir.pub/assets/e8de82b431adc8c46326dd8804e4267d.docx

Handling missing data including imputation and handling outliers and used software packages.
URL: http://asset.jmir.pub/assets/5ae165e2f9f8fbd0b1b884bf79e7df3f.docx

The interpretation of the MFA.
URL: http://asset.jmir.pub/assets/5a9522746cf5776211f814c0b4797eb6.docx

Topic list expert-panel interview.
URL: http://asset.jmir.pub/assets/0e9c44165c1c87555b52fb8326ae57fc.docx

Remaining characteristics of colorectal cancer (N = 3989) and prostate cancer (N = 696) participants.
URL: http://asset.jmir.pub/assets/52ead623eee85da87e63f1777c86796e.docx