BACHELOR

Prediction of long-term symptom burden among cancer survivors
A machine learning analysis

Offringa, Sjoerd

Award date:
2023

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computer Science

Prediction of long-term symptom burden among cancer survivors: A machine learning analysis

Bachelor End Project Report

Sjoerd Offringa

Supervisors:
dr. Katrijn Van Deun
dr. Simone Oerlemans
dr. Belle de Rooij

21-01-2023

Abstract

Due to an increase in cancer survival rates and cancer incidence, the amount of cancer survivors steeply increases. However, a research gap has been identified regarding long-term symptom burden among cancer survivors. Multiple machine learning techniques were used in this research to predict and cluster symptom burden. This had the purpose of better understanding cancer survivorship and exploring the possibility for machine learning models to predict symptom burden in practice. A neural network was created and compared against less complex models. A significant improvement was observed, but predictions were largely biased towards the mean. K-means clustering on fatigue trajectories did not indicate that trajectories were clustered.

Contents

Introduction

Due to advancements in cancer diagnosis and treatment, a global increase in cancer survival rates can be observed. These advancements include chemotherapy, combination treatment and screening for certain cancers.[10] In the Netherlands for example, the cancer survival rate increases by 1% each year on average, with a current survival rate of 66%.[8] Data also shows a global positive trend in cancer incidence.[5] This can be explained by population ageing; people are getting older on average and older people have a much higher risk of getting cancer.[4] Another explanation of higher cancer incidence is natural selection; as more people survive cancer, they also pass on their genetic material to future generations more often.[11] These phenomena can be observed through higher cancer incidence rates in countries with better access to healthcare.

As a consequence of higher cancer incidence and higher cancer survival rates, the number of cancer survivors steeply increases. The term 'cancer survivor' is used for cancer patients from the moment of diagnosis.[7] Thus, the term does not indicate whether the patient has completed treatment or not. The majority of cancer survivors experience long-term effects from cancer and/or its treatment.[3] Common effects include fatigue, pain, insomnia and feelings of depression/anxiety and can last more than 10 years.[6] Existing research on this topic is mostly limited to finding factors that correlate with specific long-term symptoms. A research gap has been identified regarding analysis of longitudinal and multiple symptom burden of cancer survivors.[1] Research into these aspects is vital to understanding the course and extent of cancer survivorship.

This research will address this research gap by analysing longitudinal data of Health-Related Quality of Life (HRQoL) variables, with a focus on predicting symptom burden, using machine learning techniques. A conclusion will be drawn on the potential of symptom burden prediction using machine learning. Complementary explorative analysis will be done to gain insights into the variables used. Longitudinal clustering of patients will also be carried out using k-means clustering. A conclusion will be drawn on the potential to cluster a patient's symptom trajectory using machine learning.

Methods

2.1 Data description

The dataset used in this paper is a collection of longitudinal Patient-Reported Outcome (PRO) data. The data consists of 1931 cancer survivors who reported on their health related quality of life (HRQoL) using the EORTC QLQ-C30 questionnaire (Appendix A). Each person answered the questionnaire between 1 and 8 times after cancer diagnosis for a total of 6242 completed questionnaires. The data is combined from three different cohorts from the profiles registry: 'Procore', 'Lymphoma' and 'Rogy'.[9] Each cohort collected data from specific cancer types, for a total of 8 cancer types in the data.

The collected data can be divided into 7 categories shown in figure 2.1. Sociodemographics include the age(group) and sex of a patient, as well as their education level and relationship status. Medical data includes whether the patient currently, previously or never indulged in smoking and/or alcohol use. It also includes a calculated BMI and whether the patient suffers from zero, one, or more comorbidities. Cancer factors describe which type of cancer and stage the patient was diagnosed with and which treatment(s) he/she received. The dataset includes 8 types of cancer: Non-Hodgkin Lymphoma (NHL), Hodgkin Lymphoma (HL), Chronic Lymphocytic Leukemia (CLL), Multiple Myeloma (MM), ovarian cancer, endometrial cancer, colon cancer and rectal cancer. The stage ranges from 1 to 4. Possible treatments include chemotherapy, radiation therapy, hormone therapy, targeted therapy, surgery, systemic therapy, stem cell transplant, and watchful waiting, which means that no treatment is given, but the patient's condition is watched closely. Note that chemotherapy is a form of systemic therapy. Thus, if the patient has received chemotherapy, they will also have received systemic therapy. Data related to the questionnaire includes the patient's answers to 30 questions about his or her health condition in the past week. 28 questions are on a 4-point Likert scale and 2 questions are on a 7-point Likert scale. It also includes the time passed since diagnosis and the number of questionnaires taken including the current one, termed 'wave'. The functioning and symptoms categories are computed results from the questionnaire answers on a scale of 0 to 100. Which questions are used and how the score is computed can be found in figure A.3. A higher score for functioning factors indicates the patient is doing well, except for financial problems. A higher score for symptom factors indicate a higher symptom burden.

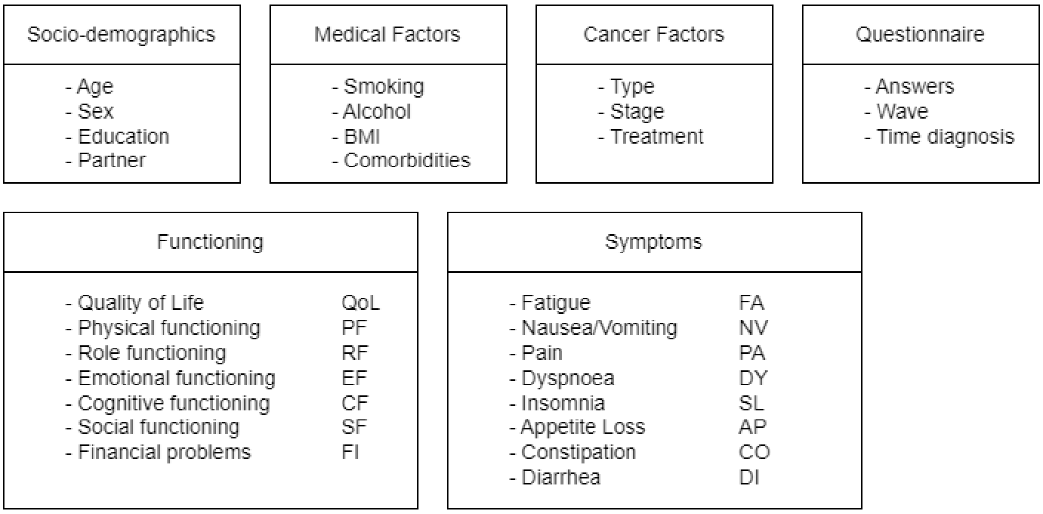


Figure 2.1: Variables within their specific categories and with their abbreviation if applicable.

2.2 Explorative analysis

2.2.1 Missing values

The dataset contains both structurally missing values and random missing values. Structurally missing values are caused by the three cohorts having discrepancies in the variables they collected and the method of collection. Through analysis of missing variables per cohort, it was found which structurally missing values exist and how many random missing values occur within each cohort. Structurally missing values were corrected based on well-grounded assumptions. From the amount of remaining random values, it was determined which values were suitable for further analysis.

2.2.2 Distributions

Distributions of variables were analysed in order to find possible outliers and under-representations. Outliers can result from incorrectly collected or computed data. Incorrect values can negatively affect further analysis, thus they

should be removed. Under-representations exist when there is little to no data of a category/group within an independent variable. Since there is less data on under-represented groups, there is more uncertainty regarding their effect on dependent variables.
The age at which patients were diagnosed was collected in 2 different ways. The Lymphoma and Rogy cohorts provided the age in years, whereas the Procore cohort provided age categories. The categories could be used for analysis, but this would be less informative than using the age in years. Therefore, it was chosen to convert the Procore age categories to age in years based on the average age for a given age category. The distribution of the diagnosis age variable in years was then visualised, using a histogram with a kernel density estimate. Other socio-demographics and medical factors were analysed and reported on if outliers or notable distributions were found.
It is possible that cancer stage can have a different effect on dependent variables per cancer type. Therefore, the number of patients for each cancer type and stage were visualised in order to find possible under-represented cancer forms. The same was done for treatment types per cancer type, because a treatment's effect on dependent variables may also differ per cancer type. Treatment types were computed as the percentage of patients that received the treatment per cancer type, because treatments are not mutually exclusive and patients can also receive no treatment at all. Chemotherapy was left out since it coincides with systemic therapy. Watchful waiting was included despite not being a physical treatment. It is however informative of which cancer types watchful waiting is applicable for.
In order to perform a longitudinal analysis, it is relevant to know how temporal data is distributed. The number of respondents per wave was plotted, as well as the distribution of how many questionnaires patients have taken. This can be used to determine how many waves have sufficient data to be included in a longitudinal analysis. The time elapsed between diagnosis and taking the questionnaire was also plotted for each wave. This is informative of the range and outliers within these distributions.

2.2.3 Dependent variables

This section provides distributions and summary statistics of the dependent variables (symptom variables, functioning variables and quality of life). This provides insights on the predictability of these variables and their relation with important independent variables. Distributions of all dependent variables were plotted and their mean and skewness, according to the Fisher-Pearson skewness coefficient, were computed. means of dependent variables were then computed for each cancer group (lymphoma, gynecologic cancers and colorectal cancers). These cancer groups correspond with the three cohorts. Cancer groups were used rather than cancer types to avoid low sample sizes such as in figure 3.2. One-way ANOVA was performed to find which symptoms differ significantly between cancer groups. Lastly, mean dependent variables were visualised per cancer group and stage. A single scale was used for symptom variables and functioning variables to allow for easy comparison within and between dependent variables.

2.3 Predictive analysis

Different methods of predicting symptom burden of different complexity have been constructed and compared. Symptoms were predicted separately from each other. If longitudinal data was used, it was used from the previous wave. For prediction of the first wave it was assumed that prior symptom burden scores were all 0.
As shown in figure 3.7, the symptom variables are all skewed towards 0. It is essential that cases where the symptom burden is not 0 are also accurately predicted, since it is more valuable to know which survivors will experience symptoms than who will not. Therefore, the performance of these models will be measured in Mean Squared Error (MSE), which punishes outliers more so than Mean Absolute Error (MAE) for example.

2.3.1 Baseline

Three baseline models were constructed with the following prediction rules:

- Baseline 1: Predict mean symptom burden.
- Baseline 2: Predict mean symptom burden per tumor type.
- Baseline 3: Predict symptom burden from previous wave. Predict 0 if it is the first wave.

The baselines were used as reference scores for more complex prediction models. The predictions from baseline models are very explainable due to their simplicity. Thus, if more complex models do not show improvement, the baseline models are preferable. Baseline 1 is the most simplistic. Baseline 2 and 3 were compared against baseline 1 to find which baseline provides the best reference.

2.3.2 Regression

Multiple linear regression models were constructed for each symptom, using variables that have few missing values. The same independent variables were used for all symptoms. Backwards selection was used to further reduce the amount of independent variables and variables were kept that had a significant effect for at least one symptom. *surgery* was excluded due to a high VIF of *surgery* and *tumortype.* This multicollinearity comes from the fact that surgery is never applicable for lymphomas and almost always for solid cancers (See figure 3.3).

2.3.3 Neural Network

A neural network was constructed for each symptom. The independent variables *sex, diagnosis age* and *wave* were used, as well as all tumor types as dummies and all treatment types. The symptom scores of the prior wave were also constructed and used as independent variables. Other variables were excluded due to a high amount of missing values. It was chosen to predict for all cancer types in one model, rather than training separate models on specific cancer types. This was supported by table 3.4, which shows that 4 out of 8 symptoms do not differ significantly between cancer groups. The Keras and Tensorflow libraries were used to construct the neural networks. The models all contained 3 hidden layers. The width of those layers was initially set to 15, but reduced if overfitting was detected. Symptoms differed in how sensitive they were to overfitting, thus different models used different widths. Early stopping was also used to prevent overfitting. Using a dropout layer, L1 regularization or L2 regularization did not seem to reduce overfitting. The hidden layers use ReLU as activation. The model used 'adam' as optimizer. K-fold cross validation (with k=10) was used to obtain reliable estimates.

2.3.4 Validation

After creating and testing different prediction models, the best performing model, the neural network, was validated. The purpose is to investigate whether the model has flaws or if it has potential to be used in practice. This was done by comparing the distribution of predicted values against the distribution of true values. Similar distribution shapes would provide evidence for an effective prediction model. The last fold of the cross-validation was used for the distributions, so the distributions show 10% of the data.
While MSE was used to compare model performance, this is not an easily

interpretable measurement of accuracy. In order to validate the accuracy of the neural network, a categorical accuracy measure was used for prediction of fatigue scores. The categories were 0 (no fatigue), between 0 and 40 (mild fatigue) and between 40 and 100 (clinically relevant fatigue). No changes were made to the neural network to obtain these results. Instead, the predictions and true values were categorised. The result was again obtained using 10-fold cross validation.

From the distribution of the predicted values and true values for fatigue by a neural network (figure 3.13), it was found that the network favours predictions near the mean. A random forest classifier was constructed to compare against the neural network on performance of predicting the aforementioned fatigue categories. If the random forest out-performs the network, it provides evidence that MSE is not a suitable loss function. The decision tree was created using the XGBoost random forest classifier with a max depth of 12. Results were obtained using 10-fold cross validation. A confusion matrix of the results was computed for the neural network and the random forest on the same 10% subset of test data.

2.4 Longitudinal clustering

K-means clustering was applied to longitudinal fatigue data. The purpose of this analysis is to find what trajectories patients can possibly be grouped in and what the optimal number of clusters is. Clustering was performed using the kml package in R, which provides k-means clustering, specifically for longitudinal data. Fatigue data was used from the first wave up to and including the fifth wave. Trajectories that had 3 or more missing values were not included. Between 2 and 8 clusters were compared for optimal clustering and evaluated by their Calinsky-Harabasz score. The algorithm was re-run 50 times with different starting conditions for optimal results.

The trajectories with their cluster means were plotted for 2 cluster means and 5 cluster means. It was chosen to plot 2 cluster means because it had the highest Calinski Harabasz score. 5 cluster means were plotted because this was the lowest number of clusters with cluster mean trajectories that are not constant over time.

Results

3.1 Explorative analysis

3.1.1 Missing values

Missing values of socio-demographics, medical factors, cancer factors and treatment types are given as a percentage in table 3.1. *Age* and *sex* can be seen to have little to no missing values, which makes them suitable for further analysis. *education* and *partner* have a higher number of missing values. These factors are also not expected to have a high predictive power of symptom burden, making them less suitable for further analysis. Medical factors are interesting to investigate, but all have a high number of missing values.

Cancer factors are the main independent variables of interest. It can be seen that there are no missing values for *tumortype* and most missing values in *stage* are found in the Lymphoma cohort. More specifically, most missing *stage* values occur within the tumor types Non-Hodgekin Lymphoma, Chronic Lymphocytic Leukemia and Multiple Myeloma.

Many treatment type variables have structurally missing values, which can be distinguished by a missing percentage of 100, indicated in bold. These treatment types are not collected by that cohort, because they are not relevant for said cohort. Therefore, missing treatment data that was not collected by the cohort is assumed to be 0 (i.e. treatment not underwent).

The remaining questionnaire, functioning and symptom variables all have a low number of missing values ($< 5\%$).

		procore	lymphoma	rogy
Socio-demographics	Age	0.0	0.2	0.3
	sex	0.0	0.0	0.0
	education	0.9	9.8	2.0
	partner	25.9	0.9	1.4
Medical factors	smoking	26.0	63.1	2.7
	alcohol use	26.3	65.6	10.6
	bmi	23.9	10.0	2.8
	comorbidities	26.2	3.8	34.5
Cancer factors	tumortype	0.0	0.0	0.0
	stage	0.3	49.0	8.0
Treatment types	chemotherapy	0.0	0.0	**100.0**
	radiotherapy	0.0	0.0	0.6
	hormonetherapy	0.0	**100.0**	0.6
	targetedtherapy	0.0	0.0	**100.0**
	surgery	4.6	**100.0**	0.6
	systemic	0.0	0.0	0.6
	watchfulwaiting	**100.0**	8.3	**100.0**
	stemcell	**100.0**	0.0	**100.0**

Table 3.1: Percentage of missing values of socio-demographic and medical factors and cohort.

3.1.2 Distributions

The distribution of age is plotted in figure 3.1. It can be seen that the dataset consists largely of older patients. Sex is equally distributed in the data with 52% being male. Of the patients that reported on their smoking behaviour, 11% currently smokes and 51% had previously smoked. Of those that reported on their alcohol use, 65% currently consumes alcohol and 8% previously consumed alcohol. BMI is found to have at least 25 impossible values, with a minimum of 1 and a maximum of 418. This indicates the BMI measurement may not be reliable. It was also found that many patients in the dataset suffer from comorbidities, with 42% having more than 1 comorbidity and 29% having 1 comorbidity.

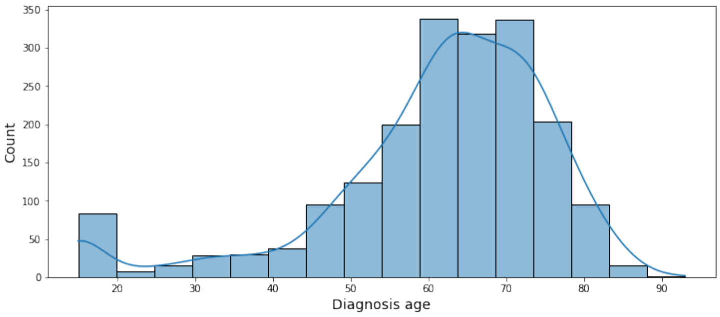
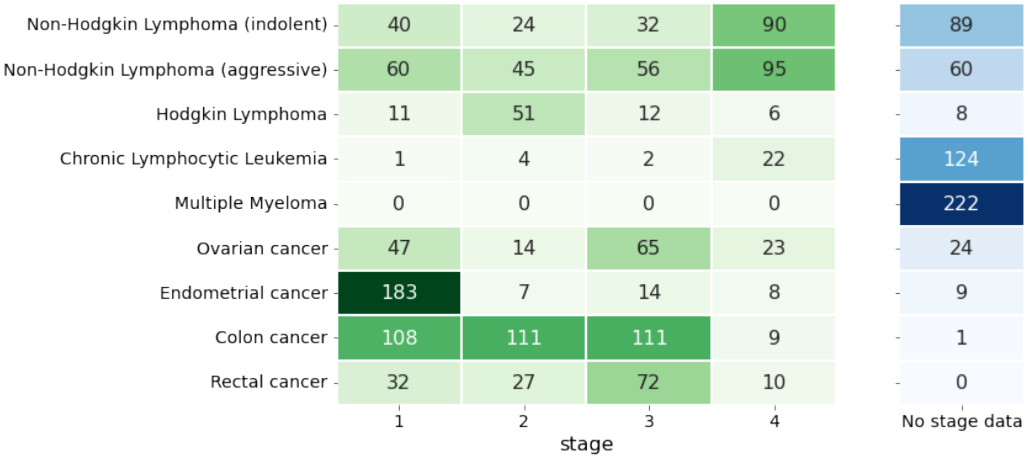


Figure 3.1: Age distribution of patients.

Figure 3.2 shows the number of participants for each cancer type and stage, including the patients where stage data is missing. It can be seen that endometrial cancer has an over-representation of first stage patients and solid cancers have an under-representation of stage 4 patients. Distribution of stages is less clear for the lymphomas due to the high amount of missing stage data. Notable is that stage 4 appears to be the largest group for Non-Hodgkin Lymphoma.



	1	2	3	4	No stage data
Non-Hodgkin Lymphoma (indolent)	40	24	32	90	89
Non-Hodgkin Lymphoma (aggressive)	60	45	56	95	60
Hodgkin Lymphoma	11	51	12	6	8
Chronic Lymphocytic Leukemia	1	4	2	22	124
Multiple Myeloma	0	0	0	0	222
Ovarian cancer	47	14	65	23	24
Endometrial cancer	183	7	14	8	9
Colon cancer	108	111	111	9	1
Rectal cancer	32	27	72	10	0

Figure 3.2: Number of patients for each cancer type and stage, including missing stage data.

Figure 3.3 shows the percentage of patients who received a treatment per cancer type. Although patients most commonly receive 1 type of treatment ($n = 3281$), they can also receive no treatment ($n = 360$), 2 types of treatment ($n = 2297$) or 3 to 4 types of treatment ($n = 304$).

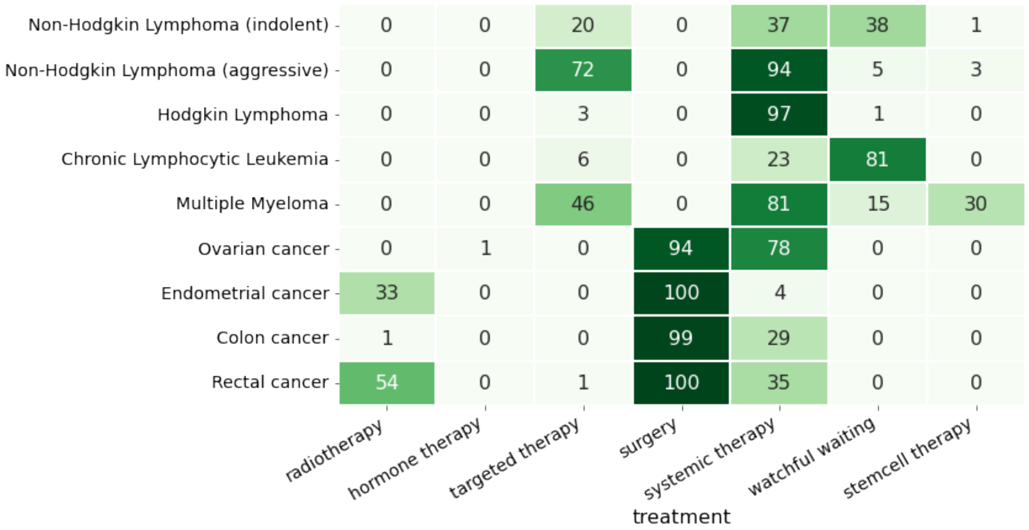


Figure 3.3: Percentage of treatment received per cancer type.

Figure 3.4 shows the amount of collected data for each wave. It can be seen that the amount gradually decreases with each wave. Figure 3.5 shows the number of waves patients have completed. It follows that patients have completed 4 questionnaires most commonly.

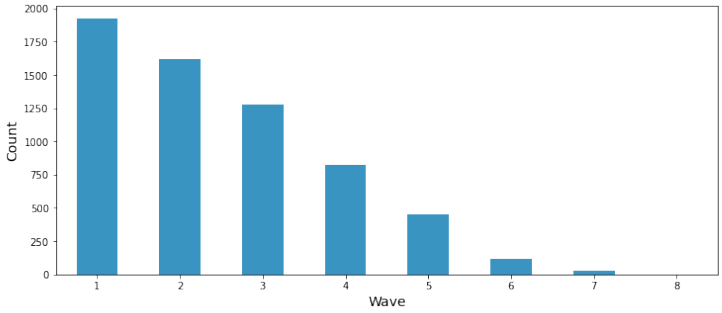


Figure 3.4: Amount of data per wave.

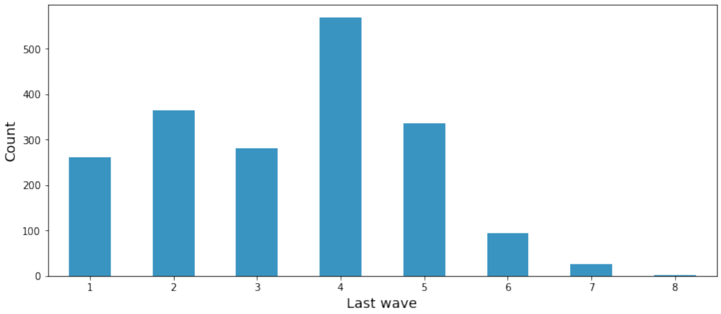


Figure 3.5: Distribution of number of questionnaires taken my patients.

Figure 3.6 shows the elapsed time between diagnosis and the moment of taking the questionnaire per wave. Of course, higher waves generally have a higher time since diagnosis. It can also be seen that each wave has a cutoff point at 5 years after which much fewer questionnaires are taken. Lastly, the first waves start from 0 years, the third wave from 1 year and the fourth and fifth wave from 2 years.

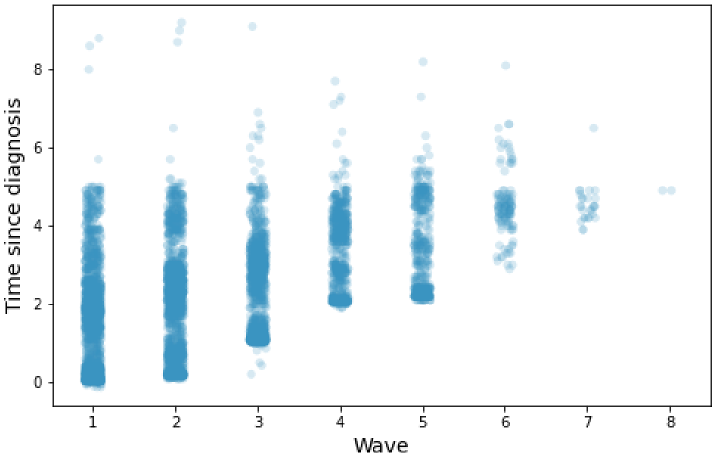


Figure 3.6: Time elapsed since diagnosis at the moment of taking the questionnaire for each wave.

3.1.3 Dependent variables

Figure 3.7 shows the distributions of the symptoms. All symptoms have 0 as mode value, which means that no symptom burden is most commonly observed. It can be seen that all symptom distributions are right-skewed, but differ in their degree of skewness. 3.2 shows the mean and skewness according to the Fisher-Pearson skewness coefficient for each symptom. The symptoms are computed from 1, 2 or 3 questionnaire items, with more items leading to more possible symptom scores. It appears from the *fatigue* distribution that the computation from three items leads to irregularities in the distribution.

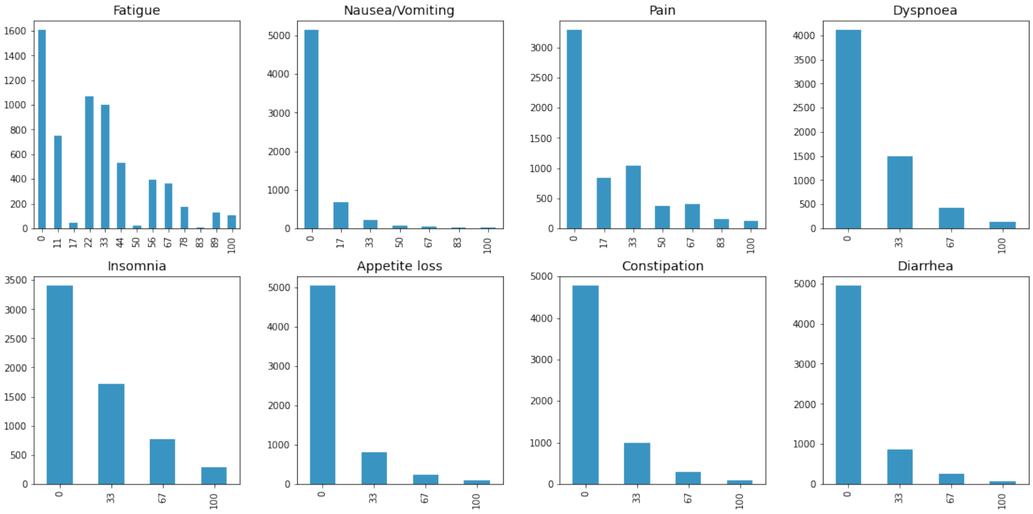


Figure 3.7: Distribution of symptom burden variables.

	mean	skew
Fatigue	27.82	0.84
Nausea/Vomiting	4.70	3.67
Pain	19.10	1.32
Dyspnoea	14.78	1.65
Insomnia	22.16	1.14
Appetite loss	8.62	2.62
Constipation	10.02	2.27
Diarrhea	8.83	2.47

Table 3.2: Distribution statistics of symptom burden variables.

Figure 3.8 shows the distributions of functioning variables and quality of life. Functioning variables have a mode value of 100, indicating that full functioning is most commonly observed. QoL has a mode value of 83. All distributions are left-skewed. Their mean and skewness according to the Fisher-Pearson skewness coefficient can be seen in 3.3. Emotional and physical functioning are computed from 4 and 5 questionnaire items respectively. The other variables are computed from 2 items. Many irregularities can be found in the distributions, which have likely resulted from the computation from multiple items.

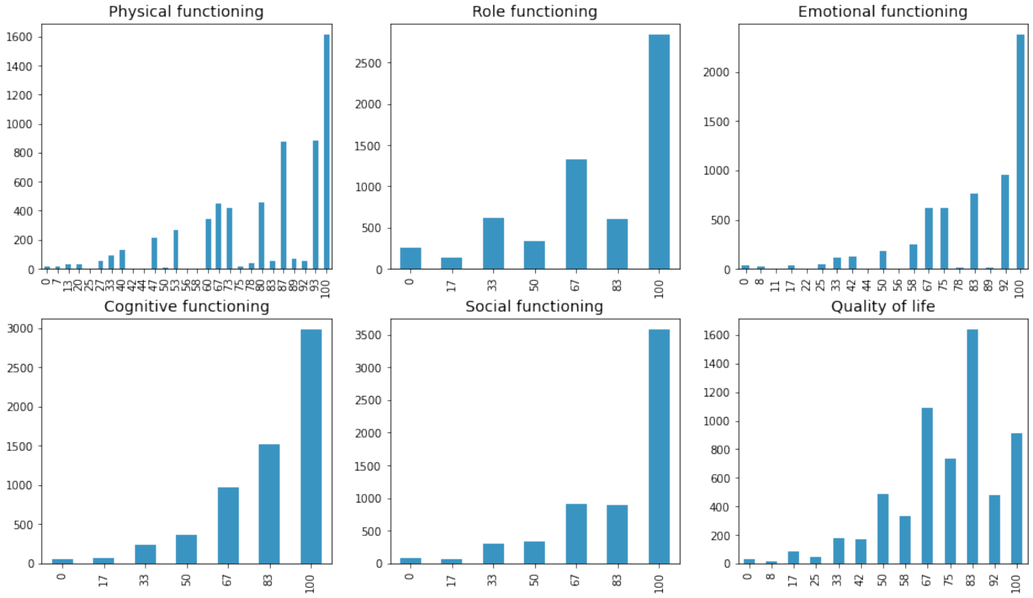


Figure 3.8: Distribution of functioning variables and quality of life.

	mean	skew
Physical functioning	80.40	-1.15
Role functioning	75.76	-1.03
Emotional functioning	83.18	-1.49
Cognitive functioning	83.39	-1.46
Social functioning	84.30	-1.58
Quality of life	74.08	-0.91

Table 3.3: Distribution statistics of functioning variables and quality of life.

Table 3.4 shows the mean symptom burden for each symptom per cancer group and the significance of a One-way ANOVA. The same analysis was performed for functioning variables and quality of life. Their mean values and ANOVA significance can be found in table 3.5.

	Lymphoma	Gynecologic	Colorectal	*p*
Fatigue	29.5	31.7	21.4	***
Nausea/Vomiting	4.4	6.8	3.9	
Pain	21.0	22.2	12.7	***
Dyspnoea	16.8	14.2	10.8	***
Insomnia	21.7	26.7	19.9	
Appetite loss	8.0	10.8	8.3	
Constipation	9.1	14.5	8.9	
Diarrhea	7.1	8.6	12.9	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.4: Mean symptom burden for each symptom per cancer group and significance of one-way ANOVA.

	Lymphoma	Gynecologic	Colorectal	*p*
Physical functioning	77.6	81.4	85.9	***
Role functioning	73.3	76.5	80.7	***
Emotional functioning	83.1	81.1	84.8	*
Cognitive functioning	81.5	83.5	87.4	***
Social functioning	83.4	82.2	87.7	***
Quality of life	73.2	73.5	76.4	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.5: Mean functioning/QoL per cancer group and significance of one-way ANOVA.

Figure 3.9 shows the average symptom burden for each symptom by cancer group and stage. Functioning variables and quality of life are plotted in the same manner in figure 3.10. Note that the colours are inverted for functioning variables and QoL, since higher functioning scores are positive and higher symptom scores are negative.

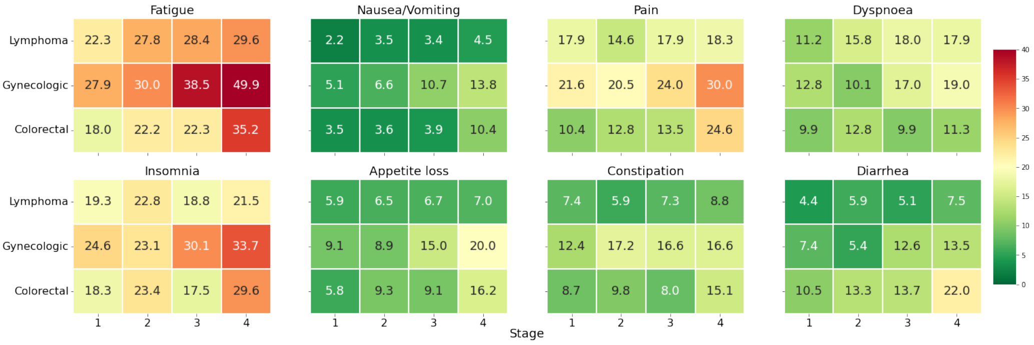


Figure 3.9: Average symptom burden per cancer group and stage.

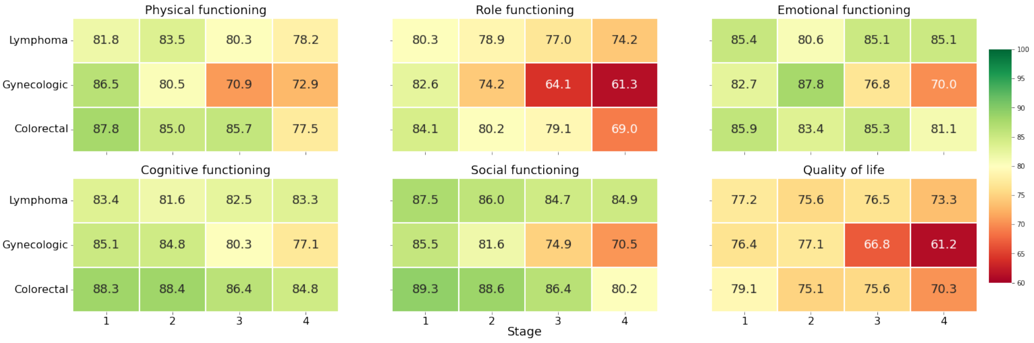


Figure 3.10: Average functioning score and quality of life per cancer group and stage.

3.2 Predictive analysis

3.2.1 Baseline

The results of the baselines described in 2.3.1 are visualised in 3.11. Baseline 2 has the best metric score and baseline 3 the worst for each symptom except dyspnoea. In the case of dyspnoea, baseline 3 has the best metric score and baseline 1 the worst. Baseline 1 is slightly worse than baseline 2 for each symptom. The MSE can also be seen to vary widely between symptoms. In conclusion, baseline 2 performs the best and can be used as a reference for more complex models.

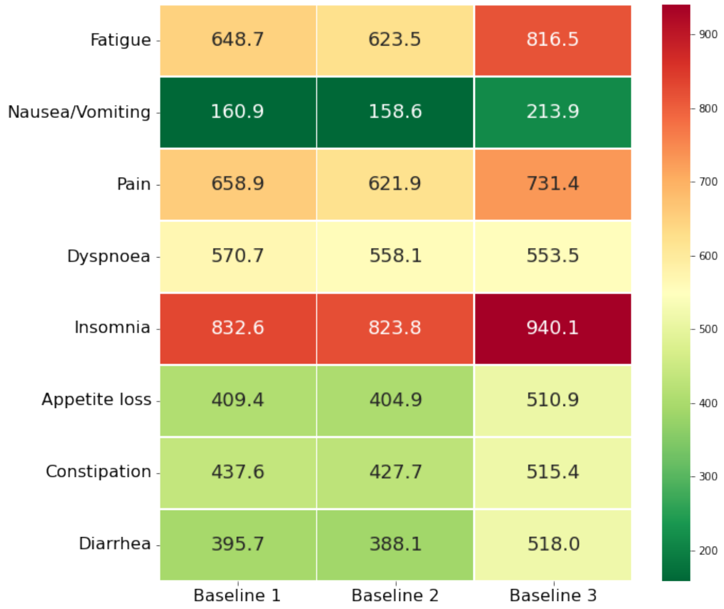


Figure 3.11: MSE of baseline predictions of each symptom.

3.2.2 Regression

Table 3.6 and 3.7 show the results of a multiple linear regression model of each symptom. Coefficient estimates are shown including their significance. The value between brackets is the standard error. *Non-Hodgekin Lymphoma (indolent)* is the reference variable of the cancer types.

	Fatigue	Nausea/ Vomiting	Pain	Dyspnoea
(Intercept)	20.19***	0.46	13.99***	13.73***
	(1.74)	(0.88)	(1.72)	(1.64)
sex	5.39***	2.30***	5.70***	-0.04
	(0.75)	(0.38)	(0.74)	(0.71)
Non-Hodgekin	-0.73	-0.76	-4.17^{**}	0.71
Lymphoma (aggresive)	(1.39)	(0.70)	(1.38)	(1.32)
Hodgekin Lymphoma	-4.57^{*}	-1.82^{*}	-10.44^{***}	-4.35^{*}
	(1.84)	(0.93)	(1.82)	(1.74)
Chronic Lymphocytic	2.13	0.29	0.05	1.11
Leukemia	(1.55)	(0.78)	(1.53)	(1.47)
Multiple Myeloma	8.08***	2.50***	12.81***	7.40***
	(1.39)	(0.70)	(1.37)	(1.32)
Ovarian cancer	3.41*	3.07***	-4.14^{*}	0.69
	(1.63)	(0.82)	(1.61)	(1.54)
Endometrial cancer	-0.46	0.91	-2.72	-1.03
	(1.59)	(0.80)	(1.57)	(1.50)
Colon cancer	-5.77^{***}	0.47	-10.38^{***}	-2.44
	(1.32)	(0.67)	(1.30)	(1.25)
Rectal cancer	-5.72^{***}	-0.04	-6.90^{***}	-6.87^{***}
	(1.59)	(0.80)	(1.58)	(1.50)
wave	-0.96^{***}	-0.27^{*}	-0.14	0.12
	(0.24)	(0.12)	(0.24)	(0.23)
targetedtherapy	-2.67^{*}	-0.12	-2.58^{*}	-0.76
	(1.17)	(0.59)	(1.16)	(1.11)
systemic	5.67***	1.75***	1.75	1.82
	(1.00)	(0.51)	(0.99)	(0.95)
watchfulwaiting	0.37	0.81	-5.36^{***}	0.73
	(1.49)	(0.75)	(1.48)	(1.41)
stemcell	-0.37	-1.32	-3.45^{*}	-4.27^{*}
	(1.77)	(0.89)	(1.75)	(1.67)
R^2	0.06	0.02	0.08	0.02
Adj. R^2	0.06	0.02	0.07	0.02
Num. obs.	5886	5888	5892	5854
MSE	613.5	156.8	602.6	547.3

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.6: Multiple linear regression model for each symptom (table 1/2).

	Insomnia	Appetite loss	Constipation	Diarrhea
(Intercept)	17.33^{***}	1.80	7.35^{***}	3.15^{*}
	(2.01)	(1.40)	(1.45)	(1.39)
sex	6.39^{***}	4.21^{***}	2.26^{***}	2.19^{***}
	(0.86)	(0.60)	(0.63)	(0.60)
Non-Hodgekin	-2.18	-1.29	-0.19	-0.82
Lymphoma (aggresive)	(1.61)	(1.12)	(1.17)	(1.12)
Hodgekin Lymphoma	-4.87^{*}	-2.25	-4.73^{**}	-3.16^{*}
	(2.12)	(1.48)	(1.54)	(1.47)
Chronic Lymphocytic	0.85	3.59^{**}	-1.87	-0.60
Leukemia	(1.79)	(1.25)	(1.29)	(1.24)
Multiple Myeloma	1.02	6.15^{***}	6.58^{***}	2.81^{*}
	(1.60)	(1.12)	(1.16)	(1.11)
Ovarian cancer	3.47	1.51	6.98^{***}	1.29
	(1.88)	(1.31)	(1.36)	(1.30)
Endometrial cancer	-4.73^{*}	1.76	1.49	1.63
	(1.84)	(1.28)	(1.32)	(1.27)
Colon cancer	-4.74^{**}	1.90	-0.48	6.87^{***}
	(1.53)	(1.06)	(1.10)	(1.06)
Rectal cancer	-3.38	1.51	-0.50	7.76^{***}
	(1.84)	(1.28)	(1.33)	(1.27)
wave	-0.48	-0.94^{***}	-0.34	-0.42^{*}
	(0.28)	(0.19)	(0.20)	(0.19)
targetedtherapy	0.04	0.15	0.47	0.37
	(1.36)	(0.95)	(0.98)	(0.94)
systemic	-2.55^{*}	3.66^{***}	-0.98	1.75^{*}
	(1.16)	(0.81)	(0.84)	(0.80)
watchfulwaiting	-2.18	-2.16	-1.58	2.95^{*}
	(1.73)	(1.20)	(1.25)	(1.19)
stemcell	1.42	-4.92^{***}	-5.06^{***}	-0.75
	(2.04)	(1.42)	(1.48)	(1.42)
R^2	0.02	0.03	0.03	0.02
Adj. R^2	0.02	0.03	0.03	0.02
Num. obs.	5874	5883	5851	5853
MSE	820.5	399.6	427.2	392.2

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.7: Multiple linear regression model for each symptom (table 2/2).

3.2.3 Neural Network

Table 3.8 shows the results of the neural network prediction for each symptom. MSE is given with its standard error, obtained from 10-fold cross-validation. The table also shows how many neurons the three layers of the neural network contained for each symptom.

			neurons
	MSE	SE	per layer
Fatigue	444.7	9.3	12
Nausea/vomiting	144.1	7.3	6
Pain	471.5	13.2	12
Dyspnoea	417.9	8.5	6
Insomnia	631.0	11.7	15
Appetite loss	339.5	14.7	12
Constipation	360.4	11.9	12
Diarrhea	353.7	14.8	6

Table 3.8: Neural network performance for each symptom.

Figure 3.12 Shows the MSE of the best baseline model (baseline 2), the regression model and the neural network for each symptom. While the regression model only shows slight improvement from the baseline, the neural network shows substantial improvement.

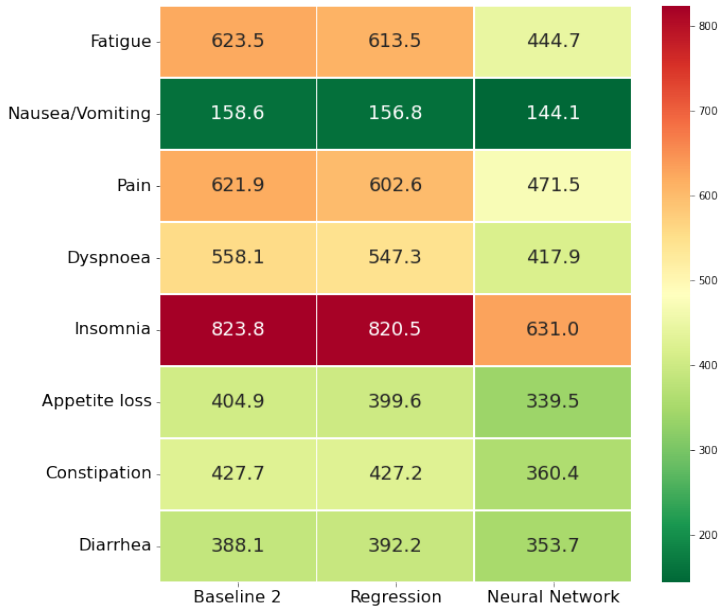


Figure 3.12: MSE of prediction models of each symptom.

3.2.4 Validation

Figure 3.13 shows the distribution of predicted values for fatigue by a neural network and the corresponding distribution of the true values. It is visible that the distributions do not have the same shape. The network predicts very few near 0 values and many values between 10 and 40, whereas true values have mostly values of 0. This indicates that the neural network favours predictions around the mean.

The accuracy of the neural network on categorical prediction of fatigue scores is 52.34% (SE = 0.70). The categories were 0 (no fatigue), between 0 and 40 (mild fatigue) and between 40 and 100 (clinically relevant fatigue). The accuracy of the random forest classifier on categorical prediction of fatigue scores is 55.23% (SE = 0.66). Figure 3.14 shows the results of the neural network and the random forest classifier in a confusion matrix. The accuracy of the random forest classifier is slightly higher than the neural network. Despite this slight difference in overall accuracy, it can clearly be seen in the confusion matrices how the random forest classifier shows less bias towards predicting mild fatigue.

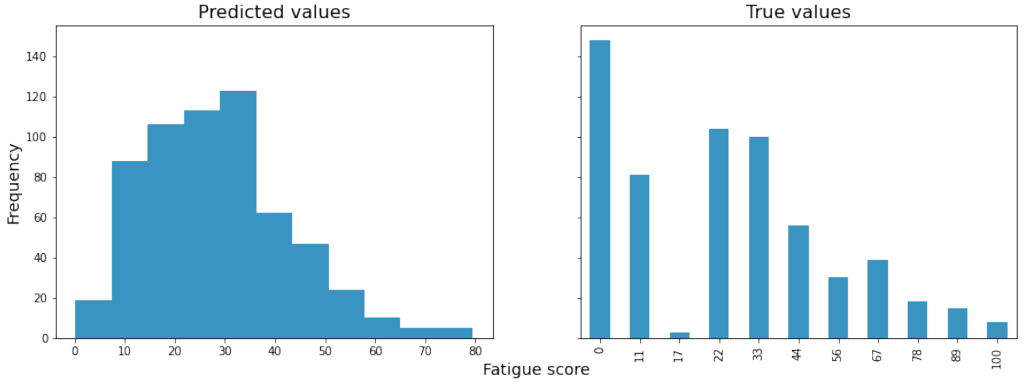


Figure 3.13: Distribution of predicted values and true values for fatigue by a neural network.

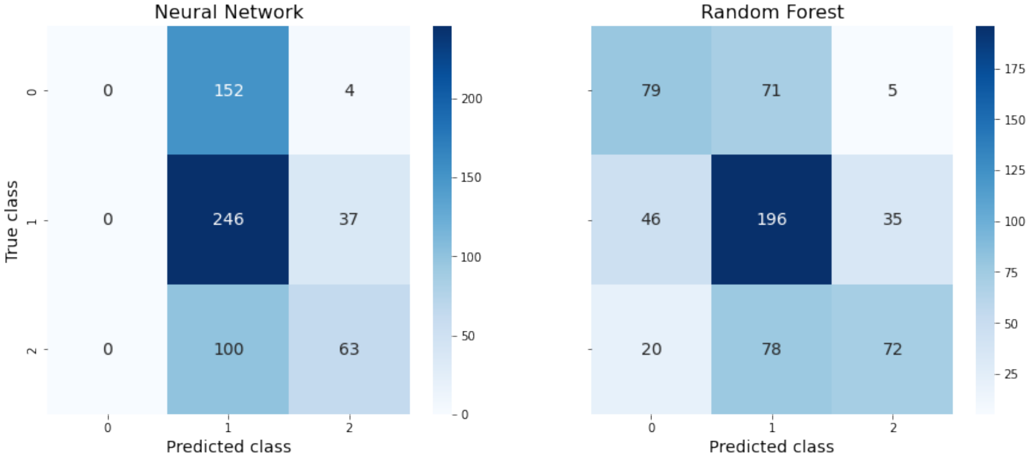


Figure 3.14: Confusion matrices of fatigue class prediction by the neural network and the random forest classifier.

3.3 Longitudinal clustering

Figure 3.15 shows the Calinski-Harabasz score for each number of clusters. It can be seen that a lower number of clusters has a higher Calinski-Harabasz score, meaning that less clusters work better on the data than more clusters.

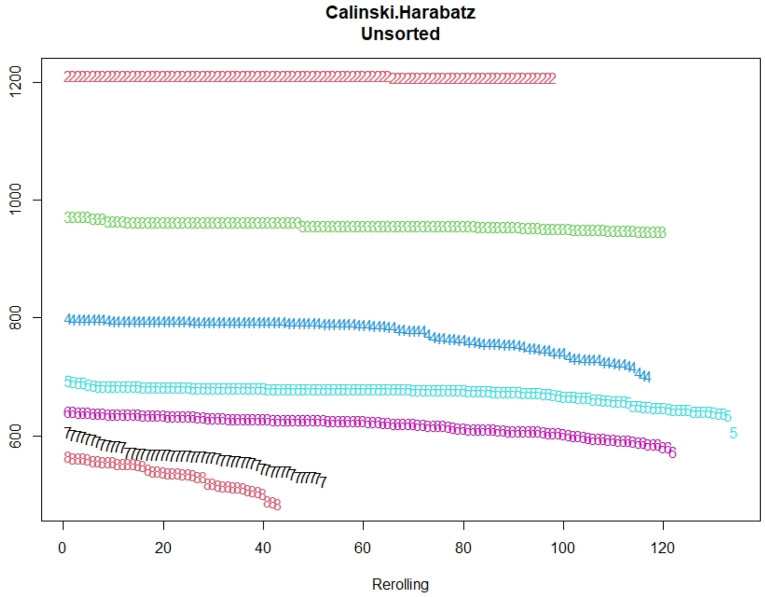


Figure 3.15: Calinski Harabatz score for 2 to 8 clusters.

Figure 3.16 shows the mean trajectories of fatigue data of wave 1 to 5. The trajectories are clustered in 2 groups. The cluster means are shown in different colours and trajectories within the same cluster are given the same colour. The percentage of trajectories belonging to each cluster is shown above the graph. The mean trajectories are relatively constant over time and divided in a high and low mean. The low mean trajectory includes the majority (64.1%) of patients.
The same plot for 5 cluster means can be found in figure 3.17. These trajectory means are not all constant over time. Clusters D increases over time and cluster E decreases and then stagnates.

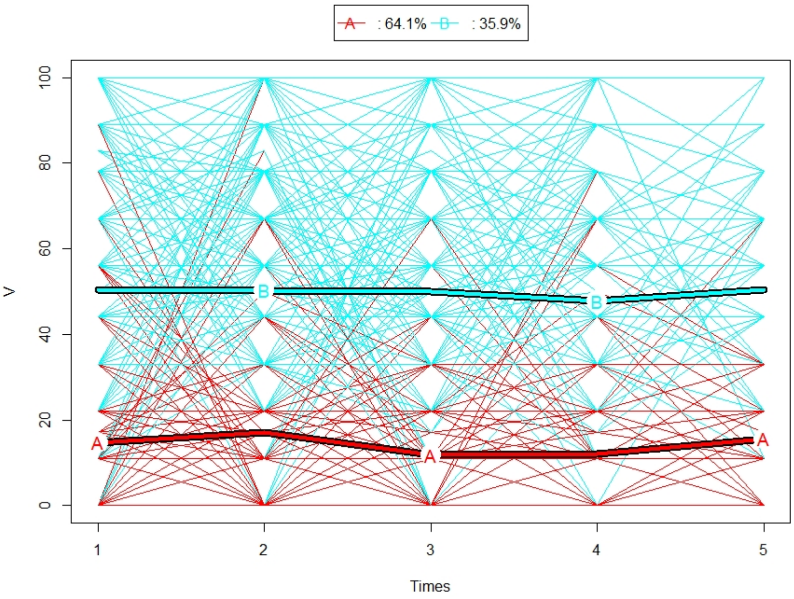


Figure 3.16: 2-means clustering of fatigue data from wave 1 to 5.

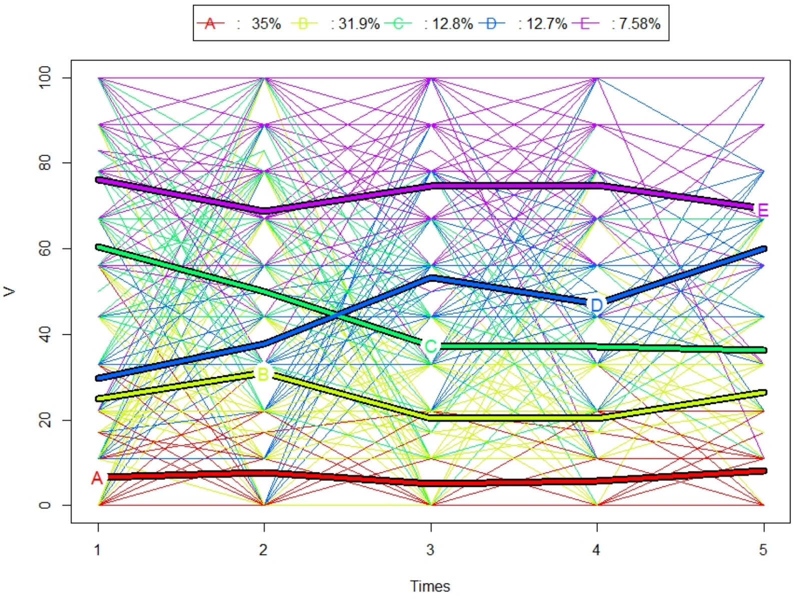


Figure 3.17: 5-means clustering of fatigue data from wave 1 to 5.

Discussion

4.1 Interpretation of results

The main focus of this research was to predict symptom burden of cancer survivors. Models of different levels of complexity were compared on this task. From figure 3.12, it can be seen that linear regression only shows slight improvement over a simple baseline (even predicting worse on diarrhea). The neural network, however, shows substantial improvement.

When comparing results between symptoms, nausea/vomiting has a much lower MSE in each model than other symptoms, and insomnia has a much higher MSE. The high accuracy on predicting nausea/vomiting can be attributed to its highly skewed distribution towards 0 (figure 3.7). The vast majority of patients do not experience nausea and/or vomiting, thus predicting a low score for all patients results in a low MSE. The high MSE for predicting insomnia can also be explained by its distribution; It has the highest mean and lowest skew after fatigue. This raises the question why fatigue does not have a higher MSE than insomnia. It is not the case that fatigue can be predicted more accurately from the independent variables than insomnia. Baseline 1, which is only based on the symptom mean, also has a higher MSE for insomnia than fatigue (figure 3.11). Therefore, the higher MSE of insomnia than other symptoms can solely be attributed to the distribution of the insomnia variable.

After establishing that the neural network shows substantial improvement in predicting symptom scores than less complex models, it was investigated how meaningful these predictions were and if the model shows potential for practical use. This was done by comparing the distribution of predictions with the distribution of true values (figure 3.13). It was found that predictions are much more centred around the mean. The implication this has is that it will often be predicted that a patient has mild fatigue when they have no or severe fatigue in reality. Especially no fatigue occurs often in reality, but is virtually never predicted by the neural network. In order to measure the severity of this issue, fatigue values and predictions by the neural network were transformed to 3 ordinal categories (no fatigue, mild fatigue and clinically relevant fatigue). The neural network predicted the correct category in 52.34% (SE = 0.70) of the cases. A random forest classifier was constructed to investigate whether the bias towards the mean could be reduced by training on the categorised data. It was found that the bias was reduced, but mild fatigue was still over-predicted (figure 3.14). The accuracy of the classifier was 55.23% (SE = 0.66). It can be concluded that the neural network has potential for predicting symptom burden, since it is substantially better than linear regression. However, it is too biased to have any practical use in its current state. A random forest classifier for 3 categories is able to decrease this bias. Its disadvantage is that it can only predict 3 values, whereas a neural network can predict any value between 0 and 100, which would be more informative if used in practice.

longitudinal k-means clustering was performed in order to find clusters within longitudinal fatigue data. The Calinski Harabasz score was lower for each additional amount of cluster, which was measured for 2 to 8 clusters (figure 3.15). This means having 2 clusters fits the data the best, although it is likely that this is the case because the data is not clustered at all. The mean cluster trajectories of 2 to 4 clusters were all constant over time such as in figure 3.16. When using 5 cluster means, a cluster with increasing fatigue appears, as well as a cluster that first drops and then stagnates. It can be concluded that the main characteristic separating fatigue trajectories is the severity of the fatigue, rather than the trajectory.

4.2 Limitations

The dataset contained many random missing values that could not be corrected. These missing values occurred for the variables *education, partner, smoking, alcohol use, bmi, comorbidities* and *stage*. These variables could not be used for much of the analysis since it would reduce the sample size significantly. Stage data is especially relevant for predicting symptom burden. This can be seen in 3.9, where many cancer groups show increased symptom burden for higher stages.

From the age distribution in figure 3.1, it can be seen that younger ages are under-represented. Although this distribution is representative of cancer incidence, it should be kept in mind that the lack of data on younger patients may influence the prediction accuracy for younger patients. Stage 4 solid cancers, or metastatic cancer, is under-represented in the dataset (figure 3.2). This is likely due to the fact that metastatic cancer is not curable in most cases.

Overall, the dataset could be improved by more complete data collection. It could also be improved by more accurate or informing data, such as the age in years instead of age groups for all cohorts and more elaborate data on the nature of possible comorbidities, for example.

4.3 Future research

The analyses presented in this research provide a starting point for more predictive analyses, among others. Since the dataset includes multiple questionnaires of the same respondent over time, a multilevel regression could be used to gain insights on the variance of dependent variables within and between patients. Insights on which independent variables to use could be gathered from the multiple regression in this research. More research can also be done on using random forest classifiers for prediction. The neural network could be improved on predicting ordinal data such as the 3 fatigue categories presented in this research using ordinal classification techniques for neural networks.(2) Moreover, prediction can be extended towards functioning variables and quality of life, since this research has focused on symptom burden. Longitudinal clustering research can be extended by applying 3d k-means clustering, which would cluster patients based on all symptoms rather than one.

4.4 Conclusion

A potential for machine learning techniques to be used for symptom burden prediction for cancer survivors has been established in this research. This was shown through significant improvement in accuracy of a neural network model over a baseline model and multiple regression model. However, the model currently has a large bias towards the mean. It is expected that this can be reduced if more complete and elaborate data is used, and the combination of machine learning technique and performance metric is optimized.
A potential for machine learning to cluster symptom trajectories of cancer survivors was less evident. It is likely that there is no clustering between trajectories of fatigue. If they are clustered, they are mainly clustered by their severity, not their trajectory. Clustering by severity has little practical use, since it would not provide any additional insights into the trajectory a patient's symptoms will take.

References

Burkett, V. S., & Cleeland, C. S. (2007, 6). Symptom burden in cancer survivorship. *Journal of Cancer Survivorship*, *1*, 167-175. Retrieved from `https://link.springer.com/article/10.1007/s11764-007-0017-y` doi: 10.1007/S11764-007-0017-Y/TABLES/2

da Costa, J. P., & Cardoso, J. S. (2005). Classification of ordinal data using neural networks. In J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, & L. Torgo (Eds.), *Machine learning: Ecml 2005* (pp. 690–697). Berlin, Heidelberg: Springer Berlin Heidelberg.

de Rooij, B. H., Oerlemans, S., van Deun, K., Mols, F., de Ligt, K. M., Husson, O., ... Schoormans, D. (2021, 12). Symptom clusters in 1330 survivors of 7 cancer types from the profiles registry: A network analysis. *Cancer*, *127*, 4665-4674. Retrieved from `https://pubmed.ncbi.nlm.nih.gov/34387856/` doi: 10.1002/CNCR.33852

European cancer information system: 21% increase in new cancer cases by 2040. (2022, 3). Retrieved from `https://joint-research-centre.ec.europa.eu/jrc-news/european-cancer-information-system-21-increase-new-cancer-cases-2040-2022-03-16_en`

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., ... Bray, F. (2019, 4). Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods. *International Journal of Cancer*, *144*, 1941-1953. Retrieved from `https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.31937` doi: 10.1002/IJC.31937

Harrington, C. B., Hansen, J. A., Moskowitz, M., Todd, B. L., & Feuerstein, M. (2010, 1). It's not over when it's over: long-term symptoms in cancer survivors–a systematic review. *International journal of psychiatry in medicine*, *40*, 163-181. Retrieved from `https://pubmed.ncbi.nlm.nih.gov/20848873/` doi: 10.2190/PM.40.2.C

Oerlemans, S., Mols, F., Issa, D. E., Pruijt, J. H., Peters, W. G., Lybeert, M., ... van de Poll-Franse, L. V. (2013, 3). 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*, 479-486. Retrieved from `https://haematologica.org/article/view/6604` doi: 10.3324/HAEMATOL.2012.064907

Overleving kankerpatiënten stijgt, maar niet bij alle kankersoorten. (2022, 8). Retrieved from https://iknl.nl/nieuws/2022/overleving-kankerpatienten-stijgt,-maar-niet-bij-a

Poll-Franse, L. V. V. D., Horevoorts, N., Eenbergen, M. V., Denollet, J., Roukema, J. A., Aaronson, N. K., ... Mols, F. (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*. doi: 10.1016/j.ejca.2011.04.034

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022, 1). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*, 7-33. Retrieved from https://www.cancer.org/latest-news/facts-and-figures-2022.html doi: 10.3322/CAAC.21708

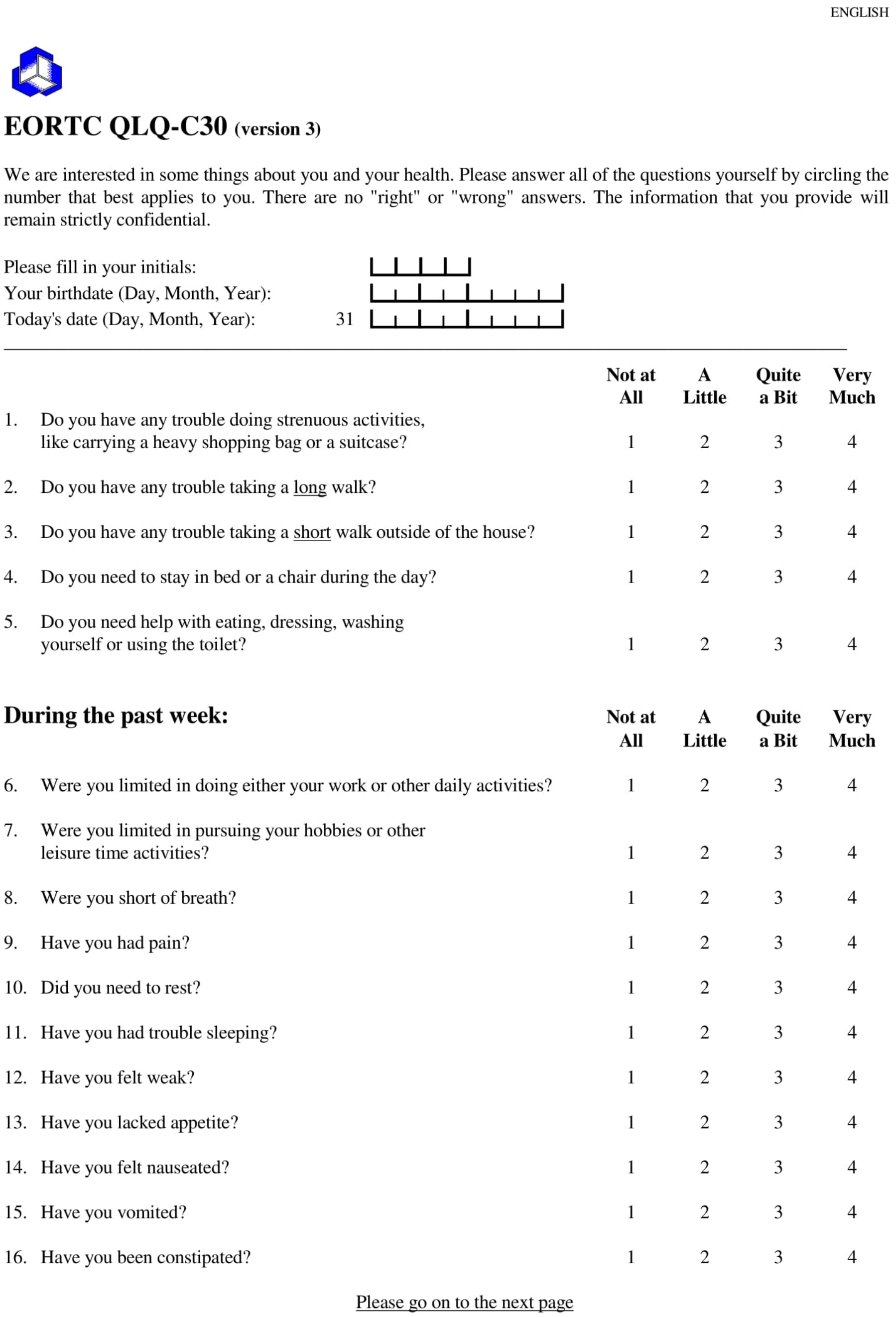
You, W., & Henneberg, M. (2018, 2). Cancer incidence increasing globally: The role of relaxed natural selection. *Evolutionary Applications*, *11*, 140-152. Retrieved from https://www.sciencedaily.com/releases/2017/10/171011100708.htm doi: 10.1111/EVA.12523

List of Figures

List of Tables

Appendix

A.1 EORTC QLQ-C30



ENGLISH

EORTC QLQ-C30 **(version 3)**

We are interested in some things about you and your health. Please answer all of the questions yourself by circling the number that best applies to you. There are no "right" or "wrong" answers. The information that you provide will remain strictly confidential.

Please fill in your initials:

Your birthdate (Day, Month, Year):

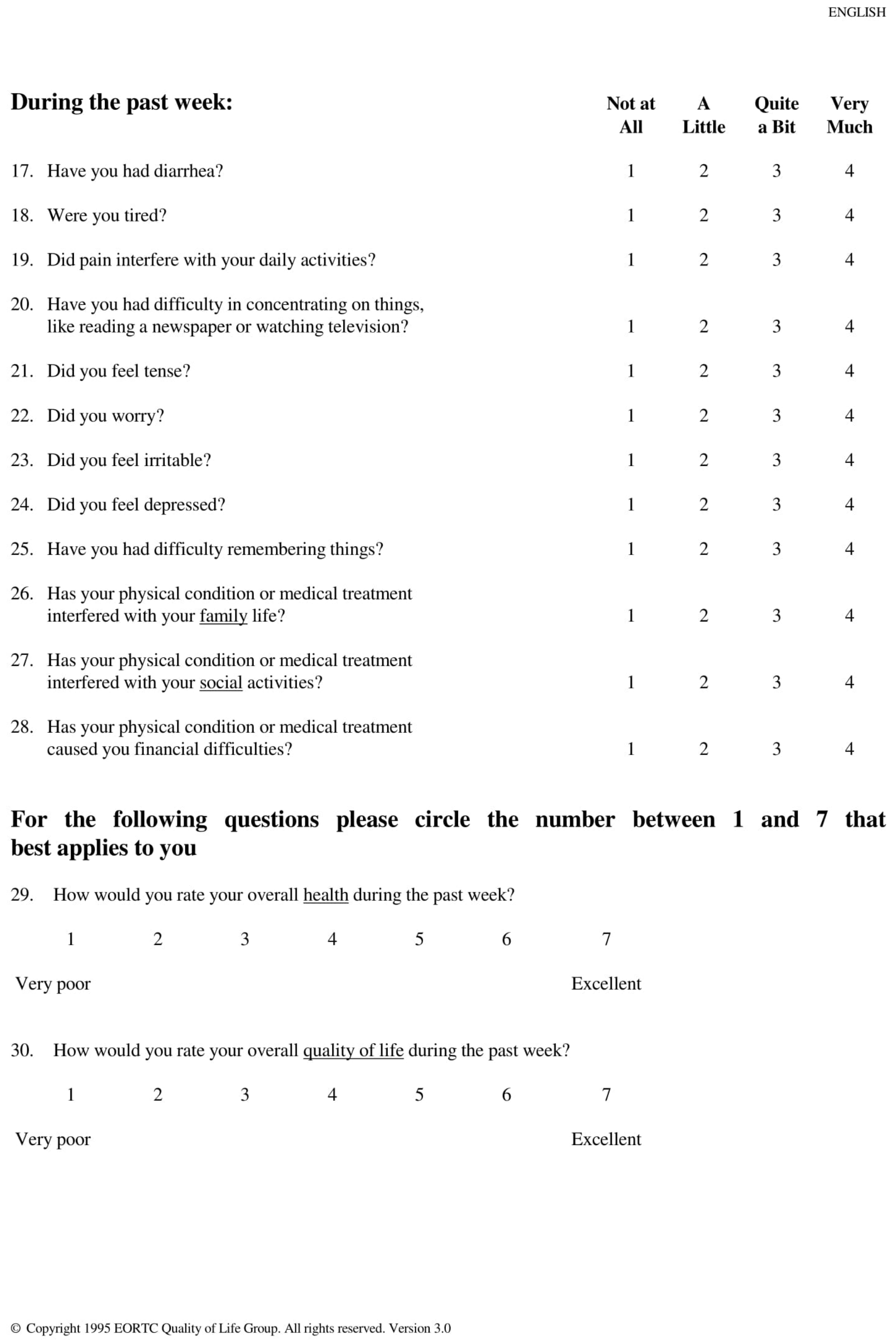
Today's date (Day, Month, Year): 31

		Not at All	A Little	Quite a Bit	Very Much
1.	Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?	1	2	3	4
2.	Do you have any trouble taking a long walk?	1	2	3	4
3.	Do you have any trouble taking a short walk outside of the house?	1	2	3	4
4.	Do you need to stay in bed or a chair during the day?	1	2	3	4
5.	Do you need help with eating, dressing, washing yourself or using the toilet?	1	2	3	4

During the past week:		Not at All	A Little	Quite a Bit	Very Much
6.	Were you limited in doing either your work or other daily activities?	1	2	3	4
7.	Were you limited in pursuing your hobbies or other leisure time activities?	1	2	3	4
8.	Were you short of breath?	1	2	3	4
9.	Have you had pain?	1	2	3	4
10.	Did you need to rest?	1	2	3	4
11.	Have you had trouble sleeping?	1	2	3	4
12.	Have you felt weak?	1	2	3	4
13.	Have you lacked appetite?	1	2	3	4
14.	Have you felt nauseated?	1	2	3	4
15.	Have you vomited?	1	2	3	4
16.	Have you been constipated?	1	2	3	4

Please go on to the next page

Figure A.1: EORTC QLQ-C30 Page 1



ENGLISH

During the past week:	Not at All	A Little	Quite a Bit	Very Much
17. Have you had diarrhea?	1	2	3	4
18. Were you tired?	1	2	3	4
19. Did pain interfere with your daily activities?	1	2	3	4
20. Have you had difficulty in concentrating on things, like reading a newspaper or watching television?	1	2	3	4
21. Did you feel tense?	1	2	3	4
22. Did you worry?	1	2	3	4
23. Did you feel irritable?	1	2	3	4
24. Did you feel depressed?	1	2	3	4
25. Have you had difficulty remembering things?	1	2	3	4
26. Has your physical condition or medical treatment interfered with your family life?	1	2	3	4
27. Has your physical condition or medical treatment interfered with your social activities?	1	2	3	4
28. Has your physical condition or medical treatment caused you financial difficulties?	1	2	3	4

For the following questions please circle the number between 1 and 7 that best applies to you

29. How would you rate your overall health during the past week?

1 2 3 4 5 6 7

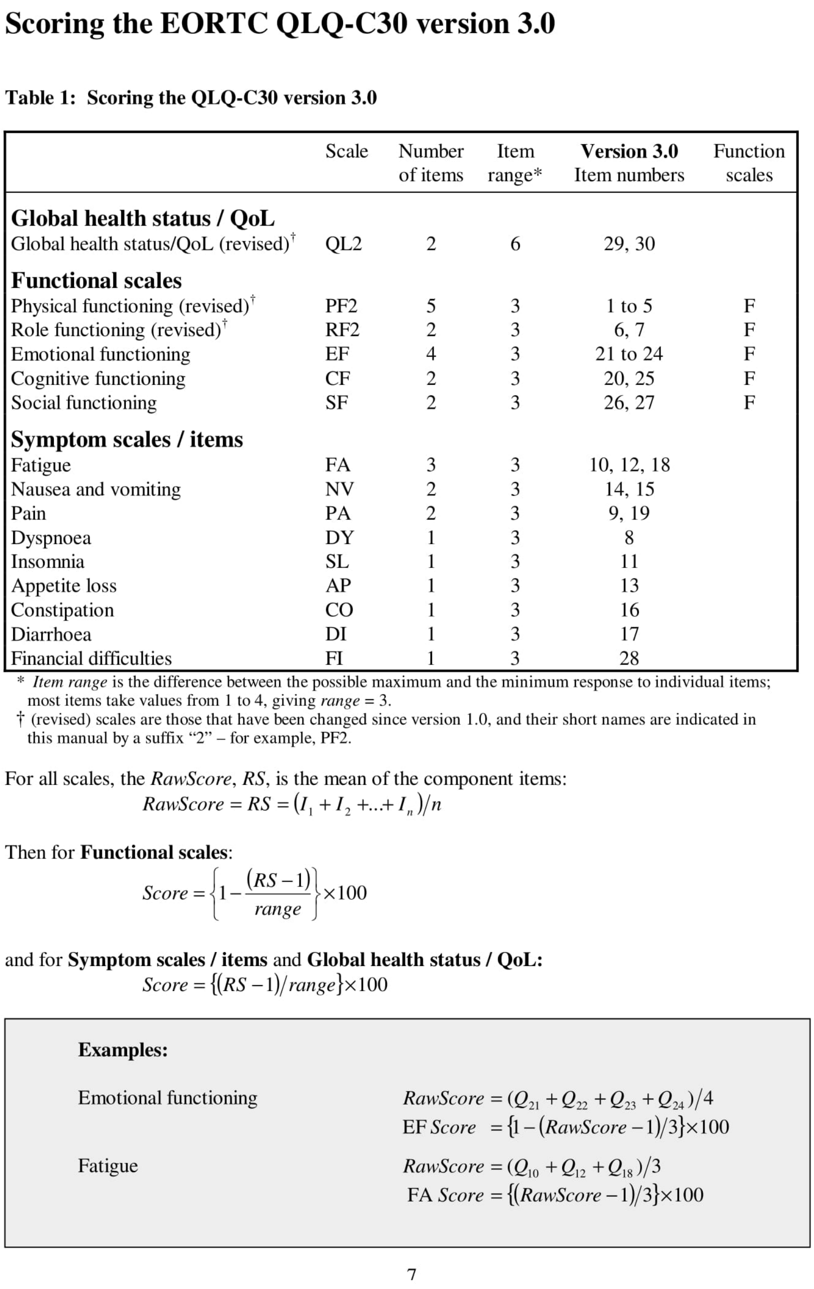
Very poor Excellent

30. How would you rate your overall quality of life during the past week?

1 2 3 4 5 6 7

Very poor Excellent

Figure A.2: EORTC QLQ-C30 Page 2



Scoring the EORTC QLQ-C30 version 3.0

Table 1: Scoring the QLQ-C30 version 3.0

	Scale	Number of items	Item range*	**Version 3.0** Item numbers	Function scales
Global health status / QoL					
Global health status/QoL (revised)†	QL2	2	6	29, 30	
Functional scales					
Physical functioning (revised)†	PF2	5	3	1 to 5	F
Role functioning (revised)†	RF2	2	3	6, 7	F
Emotional functioning	EF	4	3	21 to 24	F
Cognitive functioning	CF	2	3	20, 25	F
Social functioning	SF	2	3	26, 27	F
Symptom scales / items					
Fatigue	FA	3	3	10, 12, 18	
Nausea and vomiting	NV	2	3	14, 15	
Pain	PA	2	3	9, 19	
Dyspnoea	DY	1	3	8	
Insomnia	SL	1	3	11	
Appetite loss	AP	1	3	13	
Constipation	CO	1	3	16	
Diarrhoea	DI	1	3	17	
Financial difficulties	FI	1	3	28	

* *Item range* is the difference between the possible maximum and the minimum response to individual items; most items take values from 1 to 4, giving *range* = 3.

† (revised) scales are those that have been changed since version 1.0, and their short names are indicated in this manual by a suffix "2" – for example, PF2.

For all scales, the *RawScore*, *RS*, is the mean of the component items:

$$RawScore = RS = \left(I_1 + I_2 + ... + I_n\right)/n$$

Then for **Functional scales**:

$$Score = \left\{1 - \frac{(RS-1)}{range}\right\} \times 100$$

and for **Symptom scales / items** and **Global health status / QoL:**

$$Score = \{(RS-1)/range\} \times 100$$

Examples:

Emotional functioning	$RawScore = (Q_{21} + Q_{22} + Q_{23} + Q_{24})/4$ $\text{EF}\,Score \;= \{1 - (RawScore - 1)/3\} \times 100$
Fatigue	$RawScore = (Q_{10} + Q_{12} + Q_{18})/3$ $\text{FA}\,Score = \{(RawScore - 1)/3\} \times 100$

7

Figure A.3: EORTC QLQ-C30 Scoring system