View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Strathclyde Institutional Repository

Papachristou, Nikolaos and Miaskowski, Christine and Barnaghi, Payam and Maguire, Roma and Farajidavar, Nazli and Cooper, Bruce and Hu, Xiao (2016) Comparing machine learning clustering with latent class analysis on cancer symptoms' data. In: Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT), 2016 IEEE. IEEE, Piscataway, NJ., pp. 162-166. ISBN 9781509011674 , http://dx.doi.org/10.1109/HIC.2016.7797722

This version is available at https://strathprints.strath.ac.uk/62861/

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Unless otherwise explicitly stated on the manuscript, Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Please check the manuscript for details of any other licences that may have been applied. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (https://strathprints.strath.ac.uk/) and the content of this paper for research or private study, educational, or not-for-profit purposes without prior permission or charge.

Any correspondence concerning this service should be sent to the Strathprints administrator: strathprints@strath.ac.uk

The Strathprints institutional repository (https://strathprints.strath.ac.uk) is a digital archive of University of Strathclyde research outputs. It has been developed to disseminate open access research outputs, expose data about those outputs, and enable the management and persistent access to Strathclyde's intellectual output.

Comparing Machine Learning Clustering with Latent Class Analysis on Cancer Symptoms' Data

Nikolaos Papachristou[1], Christine Miaskowski[2], Payam Barnaghi[1], Roma Maguire[1], Nazli Farajidavar[1],
Bruce Cooper[2] and Xiao Hu[2]

Abstract— **Symptom Cluster Research is a major topic in Cancer Symptom Science. In spite of the several statistical and clinical approaches in this domain, there is not a consensus on which method performs better. Identifying a generally accepted analytical method is important in order to be able to utilize and process all the available data. In this paper we report a secondary analysis on cancer symptom data, comparing the performance of five Machine Learning (ML) clustering algorithms in doing so. Based on how well they separate specific subsets of symptom measurements we select the best of them and proceed to compare its performance with the Latent Class Analysis (LCA) method. This analysis is a part of an ongoing study for identifying suitable Machine Learning algorithms to analyse and predict cancer symptoms in cancer treatment.**

I. INTRODUCTION

One of the major advancements in the diagnosis, symptom management and prognostication for cancer care has been Symptom Cluster Research [1]–[3]. Under the NIH Panel recommendations [4] clinicians use several assessment tools for various symptoms throughout a cancer patient's treatment. On average, these patients report 10 unrelieved symptoms that have a negative impact on their functional status and quality of life (QOL) [5]. For the patient populations that specific symptom clusters are identified, clinicians can be in a better position to apply interventions that are targeted both at separate symptoms within a cluster (e.g. pain) as well at the entire cluster (e.g. pain, fatigue, depression) [1]. If these symptoms correspond to similar effects, for example the symptoms of the aforementioned cluster, the individual treatments for any of these symptoms separately can have relieving effects on the symptom cluster as a whole [6]. So, the pharmacologic agents used to alleviate symptoms can affect independent symptoms as well as the collective effects of a symptom cluster [1]. Nevertheless, cancer patients experience such symptoms with a significant amount of interindividual variability, with some of them experiencing a few symptoms whereas others experience every symptom associated with a given chemotherapy regimen [7]. Being able to analyse the demographic and clinical characteristics together with the molecular and behavioral profile of such patients could help us identify distinct subgroups of patients who will require more targeted symptom management interventions while undergoing cancer treatment.

There are many different approaches, both clinical and statistical, that are used on the field [3], [8], [9]. Clusters may vary depending on the assessment tools, the disease stage and type of cancer, the symptom domain used to cluster, the statistical methodology and the numbers of symptoms assessed. Furthermore, clusters may vary depending on the conceptual approach selected, more specifically whether we identify symptom clusters "de novo" or subgroups of patients based on their experiences within a specific symptom cluster [9]. As there are no standardized accepted conceptual approaches, statistical methods, assessment tools and symptom domains yet, Symptom Cluster Research requires further exploration before becoming part of the routine clinical care.

At the same time, advanced analytical techniques such as Machine Learning (ML) and Data Mining (DM) are constantly gaining popularity in providing insights into the complex nature and mechanisms of clinical problems and are used for building solutions for the critical decision making inside the clinical setting [10]–[17]. Although similar, these scientific methods have several significant distinctions from statistics [16]. For example, Machine Learning and Data Mining are more flexible about which methods to use, adopting both mathematical and heuristic approaches, on a specific dataset. Based on their superior computational efficiency nonstatistical Machine Learning methods are able to handle entire population of data samples, when statistics typically uses only a fraction sample of such datasets. Furthermore, while statistics usually handles quantified data, Machine Learning and Data Mining can handle various kinds of data (e.g. CT/MRI images, genomic data, sounds, text, etc.). Lastly, although a hypothesis is needed in order to run a statistical model, Machine Learning and Data Mining can explore the data and discover hidden patterns from it. In general, they seem to be more suitable and effective methods to produce a general, underlying understanding out of a specific, primary, complex set of data [16].

Drawing insights from the secondary analysis of past collected cancer symptoms data [7], [18] this study provides an evaluation of different ML algorithms in contrast to LCA, which is one of the common analytical methods [8] for identifying latent classes of patients based on their cancer symptoms' experience.

II. DATASET

Our dataset consists of oncology outpatients who were under chemotherapy (CTX) [7], [19]. We used two similar datasets, named as N_1 and N_2, consisting of n_1=582

[1] N. Papachristou, P. Barnaghi, R. Maguire, N. Farajidavar are with with the University of Surrey, Guildford, GU27XH UK (e-mail: n.papachristou@surrey.ac.uk).

[2] C. Miaskowski, B. Cooper and X. Hu are with the University of California, San Francisco, California, CA 94143 USA.

and n_2=1329 full patient registry records, respectively. The datasets contained demographic information such as age, sex, ethnicity, marital status, living arrangements, education, employment status, and income. In order to assess the patients' functional status the Karnofsky performance status (KPS) [20] scale was used. The Self-administered Comorbidity Questionnaires [21] evaluated the occurrence, treatment, and functional impact of comorbid conditions (e.g., diabetes, arthritis). The Memorial Symptom Assessment Scale (MSAS) [22] was used to evaluate the occurrence, severity, frequency, and distress of 32 cancer-related symptoms. Patients were asked to specify whether they had experienced each symptom within the past week (i.e., symptom occurrence) and rate its frequency of occurrence, severity, and distress.

Although there is such a plethora of data, previous analyses with LCA focused on a subset of the existing dataset, which are reported in [18] and [7]. Our experiment entailed the exploratory use of five different clustering algorithms on this cancer symptom dataset. Their respective performances were compared against each other and the algorithm with the best overall performance was selected for further analyses and comparison with the LCA method applied to the dataset N_2.

III. METHODS

A. *Comparing ML clustering algorithms*

Data was analysed using algorithms developed in PyCharm Professional Edition 4.5 and the Scikit-Learn library [23]. We explored dataset N_1 (n_1 samples x M features) utilizing different subsets from it (n_1 samples x m features; $m \leq M$).

Following the criterion used for LCA on similar datasets in previous analyses [7], [18], we divided our dataset in 2 different subsets. We used the MSAS symptoms that occurred at least in 30% (subset A = 25 Symptoms) and 40% (subset B = 15 Symptoms) of the patients. We ran our analysis only on the occurrence measurements of these symptoms, which take two discrete values "Yes" or "No".

During this initial analysis, we utilised both subsets and ran analyses with five different clustering algorithms (k-Means [24], Birch [25], Spectral-Clustering [26], Hierarchical Agglomerative Clustering and k-Modes [27], [28]). Following the LCA analysis on similar datasets, we preset the number of clusters to three and four in two distinct clusterings (3 for subset A, 4 for subset B) [7], [18].

We used k-Means [24] which is a centroid based algorithm, clustering the samples into k groups of equal variances. These clusters are described by the mean of the samples inside them, which is called the cluster "centroid". k-Means aims at choosing centroids that minimise the within-cluster sum-of-squares distance. Spectral-Clustering [26] aims to cluster data that is connected but not necessarily compact or clustered within convex boundaries. It is used when the clusters are not linearly separated in the original space, providing better results than algorithms such as k-Means. In such occasions hierarchical clustering or density based methods may also provide better results.

Birch [25] is a hierarchical clustering algorithm which can provide an advantage in datasets which are non-uniformly distributed and every data point is not equally important. It concentrates on densely occupied partitions and following a hierarchical order of analysis focuses on calculating and updating measurements that capture the natural closeness of data. Therefore, it is more robust to "noise" (data points that are not part of the underlying pattern). Agglomerative hierarchical clustering is a bottom-up clustering method starting with every single data point in a single cluster. In each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster. A matrix tree plot visually demonstrates the hierarchy within the final cluster, where each merger is represented by a binary tree. Agglomerative hierarchical clustering can produce an ordering of the objects, which may be informative for data display. Furthermore, smaller clusters can be generated, which may be helpful for discovery.

Finally, k-Modes [27], [28], is used for clustering observations based on categorical variables. In comparison to k-Means it defines clusters based on the number of matching categories between data points and not on the Euclidean distance. It is considered a more appropriate solution for clustering categorical variables than k-Means. Its differences with k-Means lie on using a matching dissimilarity measure instead of the Euclidean distance between data points, replacing means of clusters by modes and using a frequency-based method to find the underlying modes for the final clusters of data points [27].

The identified clusters were evaluated by their Silhouette coefficient index [29], where the Jaccard similarity coefficient is used [30]. The Jaccard similarity coefficient is used to compare both the similarity and diversity of sample sets. It calculates similarity between fixed sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

On the other hand, the Silhouette Coefficient is calculated using the mean intra-cluster distance "a" and the mean nearest-cluster distance "b" for each sample. The Silhouette coefficient for a sample is:

$$\frac{a-b}{max\,(a,b)} \tag{2}$$

Where "b" is the distance between a sample and the nearest cluster that the sample is not a member of. Silhouette coefficient is defined only if the number of labels constrained is:

$$2 \leq n_{labels} \leq n_{samples} - 1 \tag{3}$$

The score is bounded between -1 for inappropriate clustering and +1 for highly compact clustering. Scores around zero

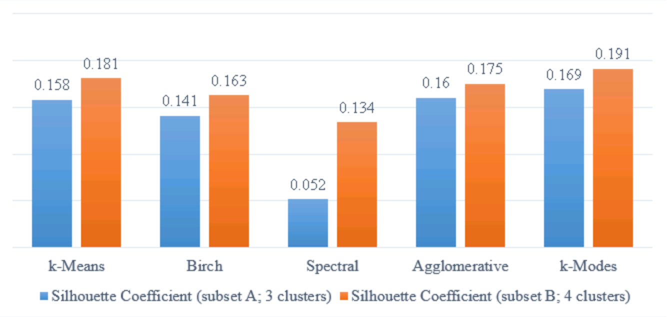

Fig. 1. Clustering performance during initial analysis.

indicate overlapping clusters. In general the score is higher when clusters are dense and well separated.

B. Comparing k-Modes with Latent Class Analysis

After evaluating the ML clustering methods to identify the one(s) with the best performance, we proceeded our experiment comparing symptom clusters made with k-Modes and LCA. For our purpose, we used dataset N_2. LCA and k-Modes were used to evaluate whether they could produce comparable findings. As our "gold" standard we used a recent study by Miaskowski et al. [18] where 4 distinct groups of patients were identified among cancer patients. In order to have a sufficient number of patients with each symptom we ran clustering on a total of 25 of 32 symptoms from the MSAS that occurred in $\geq 30\%$ of the patients.

In this paper we present the symptom occurrences of only the 25 MSAS symptoms, based on the Physical and Psychological MSAS subscales [22]. In our full study, which is going to be presented in future work, we are exploring all patient characteristics and compare the clinical relevance of both analyses with descriptive statistics, means, and standard deviations for quantitative variables and frequencies and percentages for categorical variables. This specific analysis was generated using SPSS version 22.

IV. Results

A. Comparing ML clustering algorithms

We ran ten different analyses testing k-Means, Birch, Spectral-Clustering, Hierarchical Agglomerative Clustering and k-Modes on subset A (15 symptoms, prevalence $\geq 40\%$) and subset B (25 symptoms, prevalence $\geq 30\%$). Taking into account previous analyses on similar datasets of cancer patients' symptoms [7], [18], we ran our experiments under the assumption of 3 or 4 pre-existing clusters among our samples (3 for subset A, 4 for subset B). In overall, k-Modes identified the most separable clusters (see Figure 1, Table I).

B. Comparing k-Modes with Latent Class Analysis

The 4 identified groups of patients were categorised in accordance to the probability of their symptoms' occurrence within their cluster. The "ALL High" and "ALL Low" groups included patients that reported high and low occurrence rates of the 25 selected MSAS symptoms, respectively. The "Moderate Physical and High Psych" and "Moderate Physical and Low Psych" groups included patients that reported moderate occurrence rates for the majority of the physical symptoms (i.e., lack of energy, difficult concentrating, feeling drowsy, nausea, pain, difficulty sleeping, dry mouth, lack of appetite, change in the way food tastes, numbness/tingling, hair loss, constipations, feeling bloated, changes in skin, sweats, dizziness, hot flashes,, sexual problems, cough, diarrhea) and high or low occurrence rates, respectively, for the psychological symptoms (i.e., worrying, feeling irritable, feeling sad, feeling nervous, don't look like myself).

TABLE I

Performance of the clustering algorithms

Clustering algorithms	Silhouette Coefficient (clusters = 3)	Silhouette Coefficient (clusters = 4)
K-Modes	0.169	0.191
Birch	0.141	0.163
Spectral	0.052	0.134
Agglomerative	0.160	0.175
k-Means	0.158	0.181

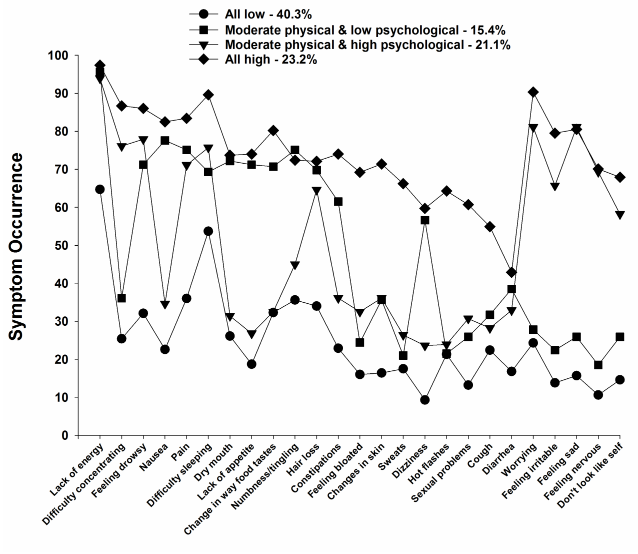

Fig. 2. Probability of symptom occurrence for K-Modes analysis

The two methods performed similarly for the "ALL High" and "ALL Low" groups, identifying two almost distinct clusters (see Figures 2, 3). The two middle groups, "Moderate Physical and Lower Psych" and "Moderate Physical and High Psych " had several overlaps among the probabilities of their symptoms' occurrence.

The "All Low" group consisted of 40.3% (n=536) of the patients for the k-Modes analysis and 31.5% (n=419) for the LCA. Probability of occurrence for the MSAS symptoms for this group ranged from 0.647 to 0.093 for k-Modes and from 0.549 to 0.064 for LCA. The "ALL High" group was the smallest for LCA (13.4%, n=178) and the second biggest for k-Modes (23.2%, n=308). Probability of occurrence for the MSAS symptoms for this group ranged from 0.974 to 0.429 for k-Modes and from 0.994 to 0.562 for LCA. The second

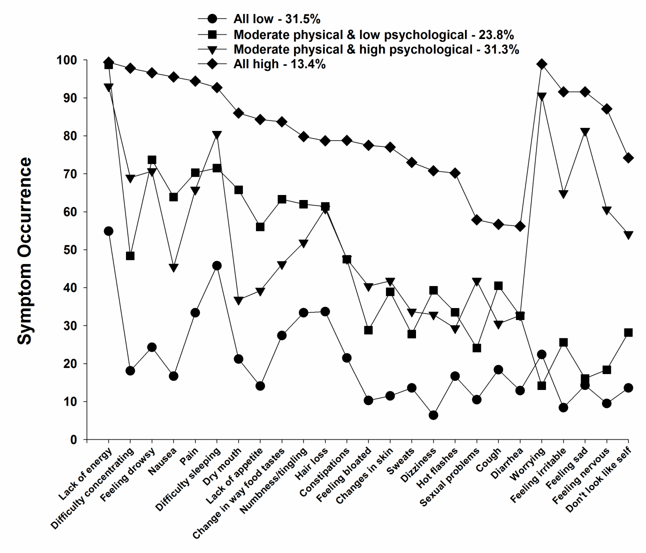

Fig. 3. Probability of symptom occurrence for Latent Class Analysis

largest separated group for LCA was the "Moderate Physical and High Psych" one, consisting of 31.3% (n=416) of the patients. In contrast the same group was the third largest identified for k-Modes with 21.1% (n=279) of the patients. The probability of occurrence for the MSAS symptom ranged from 0.93 to 0.293 for LCA and from 0.939 to 0.236 for k-Modes. For both analyses, there were moderate occurrence rates for the majority of the physical symptoms, ranging from 0.939 to 0.236 for k-Modes and from 0.93 to 0.293 for LCA, and high occurrence rates for the psychological symptoms ranging from 0.811 to 0.582 for k-Modes and from 0.906 to 0.541 for LCA. The third largest group (23.8%, n=316) for LCA was the "Moderate Physical and Lower Psych" class while the same group was the smallest one for k-Modes (15.4%, n=205). Probability of occurrence for the MSAS symptoms for this class ranged from 0.956 to 0.185 for k-Modes and from 0.987 to 0.142 for LCA. The probability of occurrence for the physical symptoms ranged from 0.956 to 0.21 for k-Modes and from 0.987 to 0.241 for LCA. For the psychological symptoms the range was from 0.278 to 0.185 for k-Modes and from 0.282 to 0.142 for LCA.

V. CONCLUSIONS

We investigated the performance of readily Machine Learning algorithms in relation to common practice statistical methods, such as LCA. The results of our study were in congruence with the hypothesis and the findings of four distinct subgroups of cancer patients that were made with the LCA method and the clinical observations by Miaskowski et al. [18]. Our proposed model provides highly compatible results with LCA, being capable of automated grouping of patients also. We have initiated a comparison among different clustering algorithms on cancer symptom datasets based on the hypothesis that Symptom Cluster Research will benefit from exploring all the multivariate versions of these datasets, taking the domain a step forward from utilising only categorical variables such as symptoms prevalence, distress and discomfort levels. Although this experiment is only based on categorical variables, the aforementioned principal is the base of ongoing experiments, taking advantage the multivariate nature of different ML methods.

In the era of precision medicine [31] and big data [32], coupled with the use of electronic medical records [33] and smart phone technology [34], [35], it can be assumed that symptom data will be collected in real time from oncology patients receiving CTX. The application of more sophisticated algorithms with analytical techniques such as Machine Learning will allow clinicians to assess patients' phenotypic and molecular data on an ongoing basis. The integration of these types of information across multiple patients will assist clinicians to identify patients at highest risk for the most severe symptom profiles and to treat their most common and severe symptoms on a more timely and effective manner. This type of risk profiling and aggressive symptom management could reduce oncology patients' symptom burden and improve their QOL.

VI. ACKNOWLEDGEMENTS

Part of this study was funded by the National Cancer Institute (CA134900). In addition, this project received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number 602289.

REFERENCES

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

[2] C. Miaskowski, "Symptom clusters: establishing the link between clinical practice and symptom management research," *Supportive Care in Cancer*, vol. 14, no. 8, pp. 792–794, 2006.

[3] J. Kirkova, A. Aktas, D. Walsh, and M. P. Davis, "Cancer symptom clusters: clinical and research methodology," *Journal of palliative medicine*, vol. 14, no. 10, pp. 1149–1166, 2011.

[4] D. Patrick, S. Ferketich, P. Frame, J. Harris, C. Hendricks, B. Levin, M. Link, C. Lustig, J. McLaughlin, L. Ried *et al.*, "National institutes of health state-ofthe-science panel. national institutes of health state-of-the-science conference statement: symptom management in cancer: pain, depression, and fatigue; july 15-17, 2002," *J Natl Cancer Inst*, vol. 95, no. 15, pp. 1110–1117, 2003.

[5] J.-E. E. Kim, M. J. Dodd, B. E. Aouizerat, T. Jahan, and C. Miaskowski, "A review of the prevalence and impact of multiple symptoms in oncology patients," *Journal of pain and symptom management*, vol. 37, no. 4, pp. 715–736, 2009.

[6] S. B. Fleishman, "Treatment of symptom clusters: Pain, depression, and fatigue," *JNCI Monographs*, vol. 2004, no. 32, pp. 119–123, 2004.

[7] C. Miaskowski, B. A. Cooper, M. Melisko, L.-M. Chen, J. Mastick, C. West, S. M. Paul, L. B. Dunn, B. L. Schmidt, M. Hammer *et al.*, "Disease and treatment characteristics do not predict symptom occurrence profiles in oncology outpatients receiving chemotherapy," *Cancer*, vol. 120, no. 15, pp. 2371–2378, 2014.

[8] H.-J. Kim, I. Abraham, and P. S. Malone, "Analytical methods and issues for symptom cluster research in oncology," *Current opinion in supportive and palliative care*, vol. 7, no. 1, pp. 45–53, 2013.

[9] C. Miaskowski, B. E. Aouizerat, M. Dodd, and B. Cooper, "Conceptual issues in symptom clusters research and their implications for quality-of-life assessment in patients with cancer." *JNCI: Journal of the National Cancer Institute*, no. 37, 2007.

[10] N. Emanet, H. R. Öz, N. Bayram, and D. Delen, "A comparative analysis of machine learning methods for classification type decision problems in healthcare," *Decision Analytics*, vol. 1, no. 1, p. 1, 2014.

[11] D. Tomar and S. Agarwal, “A survey on data mining approaches for healthcare,” *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[12] I. Kononenko, “Machine learning for medical diagnosis: history, state of the art and perspective,” *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001.

[13] I. Parvathi and S. Rautaray, “Survey on data mining techniques for the diagnosis of diseases in medical domain,” *International Journal of Computer Science and Information Technologies*, vol. 5, no. 1, pp. 838–846, 2014.

[14] P. Mahindrakar and D. M. Hanumanthappa, “Data mining in healthcare: A survey of techniques and algorithms with its limitations and challenges,” *Int. Journal of Engineering Research and Applications, ISSN*, pp. 2248–9622, 2013.

[15] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, “Data-mining technologies for diabetes: a systematic review,” *Journal of diabetes science and technology*, vol. 5, no. 6, pp. 1549–1556, 2011.

[16] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, “Data mining in healthcare and biomedicine: a survey of the literature,” *Journal of medical systems*, vol. 36, no. 4, pp. 2431–2448, 2012.

[17] S. A. A. Balamurugan, S. Sasikala, and S. Geetha, “A survey on predictive data mining approaches for medical informatics,” *International Journal of Scientific Engineering Research*, vol. 3, no. 9, 2012.

[18] C. Miaskowski, L. Dunn, C. Ritchie, S. M. Paul, B. Cooper, B. E. Aouizerat, K. Alexander, H. Skerman, and P. Yates, “Latent class analysis reveals distinct subgroups of patients based on symptom occurrence and demographic and clinical characteristics,” *Journal of pain and symptom management*, vol. 50, no. 1, pp. 28–37, 2015.

[19] C. Miaskowski, B. Cooper, B. Aouizerat, M. Melisko, L.-M. Chen, L. Dunn, X. Hu, K. Kober, J. Mastick, J. Levine *et al.*, “The symptom phenotype of oncology outpatients remains relatively stable from prior to through 1 week following chemotherapy,” *European journal of cancer care*, 2016.

[20] D. A. Karnofsky, W. H. Abelmann, L. F. Craver, and J. H. Burchenal, “The use of the nitrogen mustards in the palliative treatment of carcinoma. with particular reference to bronchogenic carcinoma,” *Cancer*, vol. 1, no. 4, pp. 634–656, 1948.

[21] O. Sangha, G. Stucki, M. H. Liang, A. H. Fossel, and J. N. Katz, “The self-administered comorbidity questionnaire: A new method to assess comorbidity for clinical and health services research,” *Arthritis Care & Research*, vol. 49, no. 2, pp. 156–163, 2003.

[22] R. K. Portenoy, H. T. Thaler, A. B. Kornblith, J. M. Lepore, H. Friedlander-Klar, E. Kiyasu, K. Sobel, N. Coyle, N. Kemeny, L. Norton *et al.*, “The memorial symptom assessment scale: an instrument for the evaluation of symptom prevalence, characteristics and distress,” *European Journal of Cancer*, vol. 30, no. 9, pp. 1326–1336, 1994.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[24] D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[25] T. Zhang, R. Ramakrishnan, and M. Livny, “Birch: an efficient data clustering method for very large databases,” in *ACM Sigmod Record*, vol. 25, no. 2. ACM, 1996, pp. 103–114.

[26] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, “On spectral clustering: Analysis and an algorithm,” *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[27] Z. Huang, “Extensions to the k-means algorithm for clustering large data sets with categorical values,” *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[28] F. Cao, J. Liang, and L. Bai, “A new initialization method for categorical data clustering,” *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 223–10 228, 2009.

[29] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[30] S.-S. Choi, S.-H. Cha, and C. C. Tappert, “A survey of binary similarity and distance measures,” *Journal of Systemics, Cybernetics and Informatics*, vol. 8, no. 1, pp. 43–48, 2010.

[31] F. S. Collins and H. Varmus, “A new initiative on precision medicine,” *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[32] C. Yoo, L. Ramirez, and J. Liuzzi, “Big data analysis using modern statistical and machine learning methods in medicine,” *International neurourology journal*, vol. 18, no. 2, pp. 50–57, 2014.

[33] D. Blumenthal and M. Tavenner, “The meaningful use regulation for electronic health records,” *New England Journal of Medicine*, vol. 363, no. 6, pp. 501–504, 2010.

[34] N. Kearney, L. McCann, J. Norrie, L. Taylor, P. Gray, M. McGee-Lennon, M. Sage, M. Miller, and R. Maguire, “Evaluation of a mobile phone-based, advanced symptom management system (asyms©) in the management of chemotherapy-related toxicity,” *Supportive Care in Cancer*, vol. 17, no. 4, pp. 437–444, 2009.

[35] R. Maguire, E. Ream, A. Richardson, J. Connaghan, B. Johnston, G. Kotronoulas, V. Pedersen, J. McPhelim, N. Pattison, A. Smith *et al.*, “Development of a novel remote patient monitoring system: the advanced symptom management system for radiotherapy to improve the symptom experience of patients with lung cancer receiving radiotherapy,” *Cancer nursing*, vol. 38, no. 2, pp. E37–E47, 2015.