Supplementary file

DECIDING ON THE NUMBER OF CLUSTERS - METHODOLOGICAL EXPLORATION

To segment the high-risk population, a k-means method was used. This method is efficient even for large sample sizes and produces roughly similar sized segments.[1] However, this method also require the number of clusters (k) to be specified before the analysis, rather than deducing it from the results afterwards. Therefore, a number of steps were taking to identify the optimal number of clusters for this population.

PSEUDO-F STATISTIC

The main method for determining the number of clusters was the Pseudo-F statistic.[2] This statistic is commonly used in healthcare clustering studies,[3-7] and has been identified as one of the best criteria to determine the number of clusters.[8] It compares the between-cluster to the within-cluster sum-of-squares, and a large Pseudo-F statistic indicates distinct clusters.[9]

The k-means analysis was run for 2 to 8 clusters, and the Pseudo-F statistic was calculated for each solution (see table 1). A peak could be observed around the 3- and 4-cluster solutions.

Table 1: Pseudo-F statistics for 2- to 8-cluster solutions

2 clusters	2249
3 clusters	2745
4 clusters	2662
5 clusters	2374
6 clusters	2267
7 clusters	2131
8 clusters	2041

WARD'S LINKAGE

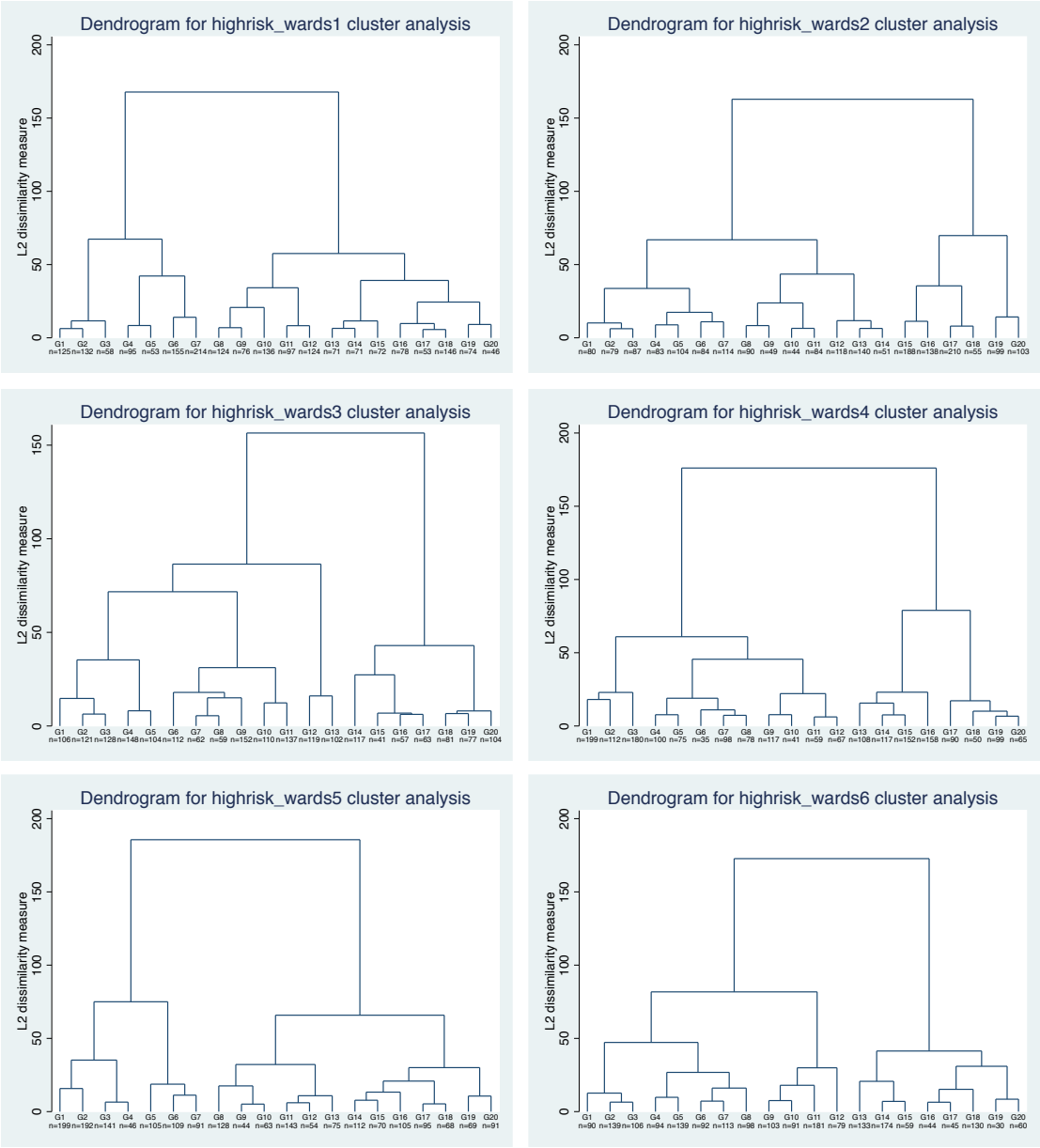
K-means is a non-hierarchical clustering method. Hierarchical methods, including the popular Ward's method, do not require k to be specified before the analysis.[8] Hierarchical clustering can be used to gain more insight into the data's structure. By displaying the results as a dendogram (a tree-like plot detailing each hierarchical step in the model) different clustering solutions can be visually explored.[10,11] Indeed, many studies combine hierarchical clustering with k-means in a two-stepped approach.[12-15]

However, hierarchical methods present some limitations. The approach is computational intensive and struggles to handle large datasets with more than a thousand observations.[10,11] In addition, hierarchical clustering based on Ward's method can be sensitive to outliers.[8]

The high-risk population in the test sample, consisting of 7,433 people, was too large to include in its entirety in a hierarchical cluster analysis. Therefore, three unique, random samples of 2,000 people we used. After reshuffling the data, another three 2,000 people samples were taken and clustered. These results were then analysed through dendograms

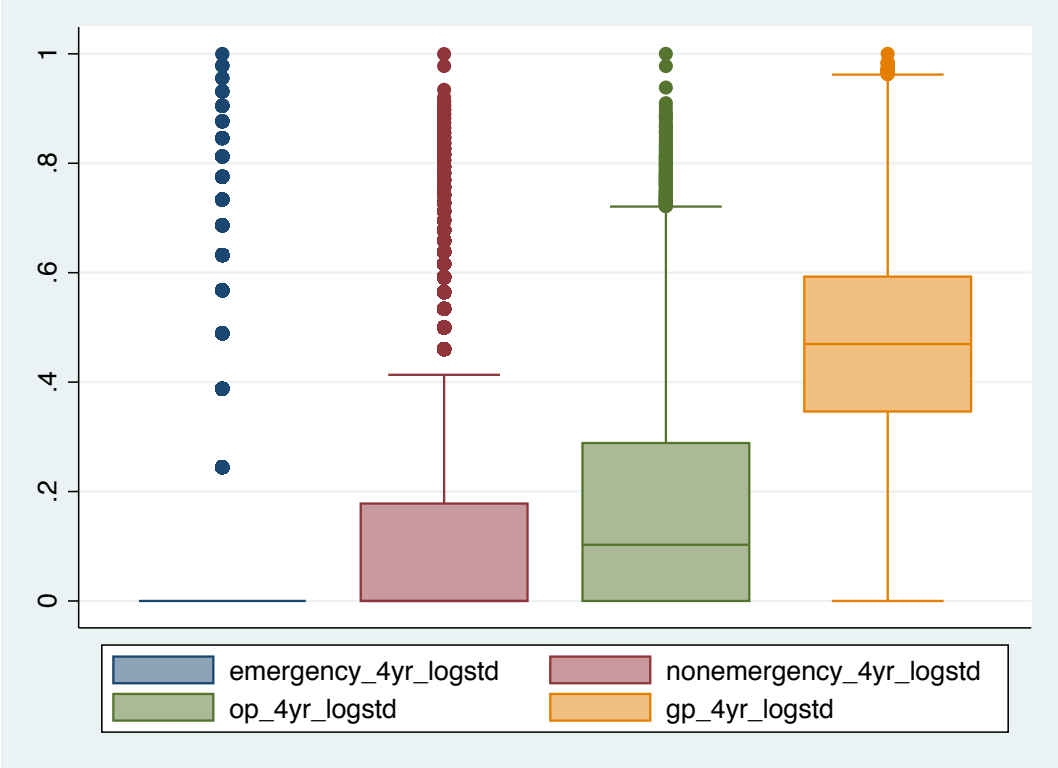
(see figure 1). All samples favoured a two-cluster solution, reflecting high- and low-utilisation groups, with the next split being further down the graph. The samples showed different results regarding the next best split. Sample 1,2, 3 and 5 can be interpreted as indicating the existence of four distinct clusters. Sample 4 favoured five clusters, and sample 6 could be interpreted as three or five clusters. Overall, the differences at this level are small.

Figure 1: *Dendograms for the six 2,000 people samples clustered using Ward's linkage*



One of the reasons the results are different across the samples is the impact of outliers, which Ward's method is sensitive to.[8] Despite the log-normalisation of the clustering variables, there still exist a large number of outliers (see figure 2). Especially in the smaller samples used for the clustering, these outliers could have changed the resulting clusters.

Figure 2: Box plots of the standardised, log-normalised clustering variables

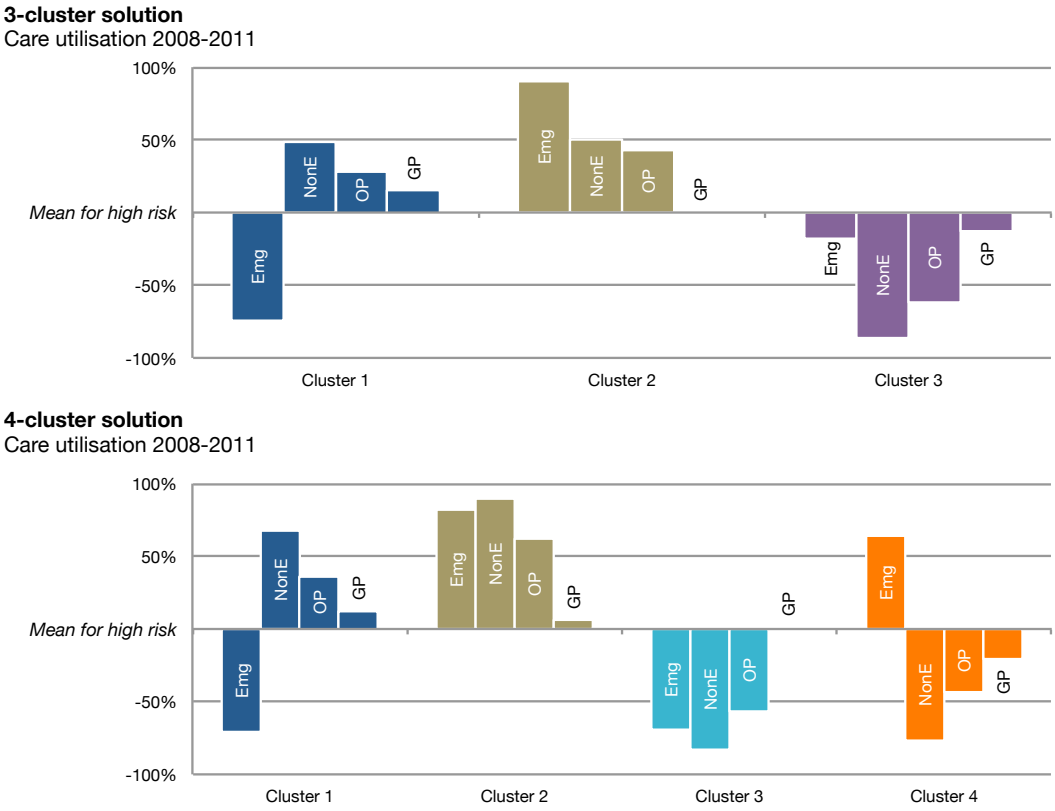


POST-HOC ANALYSIS

It is important to keep in mind that in cluster analysis, there is no absolute 'right' answer[16] - it all depends on the purpose of the clustering. Some other aspects to consider in evaluating the number of segments are, for example, interpretability, actionability and ease of use.[11]

The cluster means of the 3- and 4-cluster solutions were compared to review the practical usefulness of the resulting population groups (see figure 3). Both solutions found clusters of people with high utilisation but low emergency care use (clusters one), and people with overall high utilisations (clusters two). As the third group, the 3-cluster solution identified people with low overall utilisation but average emergency care use. However, the 4-cluster solution split this final cluster into two very distinct groups: people with overall low utilisation, and people with low utilisation but high emergency care use.

Figure 3: Practical comparison of the 3- and 4-cluster solutions



Considering the relevance of emergency care use for risk stratification, the difference between clusters three and four are important to the interpretability of the results. In terms of actionability, differentiating between these two groups allows tailored initiatives to be developed that target those people with low care utilisation but high emergency admissions. Taking this, and the previously described analyses into account, the 4-cluster solution was ultimately selected.

References

1. Han J, Kamber M. *Data Mining: Concepts and Techniques*. 1st ed. San Diego, CA: Academic Press, 2001.
2. Stata Statistical Software: Release 14. [program]. College Station, TX: StataCorp LP, 2015.
3. Armstrong JJ, Zhu M, Hirdes JP, et al. K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population. Arch Phys Med Rehabil 2012;**93**(12):2198-205.
4. Coste J, Bouyer J, Fernandez H, et al. A population-based analytical approach to assessing patterns, determinants, and outcomes of health care with application to ectopic pregnancy. Med Care 2000;**38**(7):739-49.
5. Cryer PC, Saunders J, Jenkins LM, et al. Clusters within a general adult population of alcohol abstainers. International Journal of Epidemiology 2001;**30**(4):756-65.
6. Kendig H, Mealing N, Carr R, et al. Assessing patterns of home and community care service use and client profiles in Australia: a cluster analysis approach using linked data. Health & Social Care in the Community 2012;**20**(4):375-87.
7. Pud D, Ben Ami S, Cooper BA, et al. The Symptom Experience of Oncology Outpatients Has a Different Impact on Quality-of-Life Outcomes. J Pain Symptom Manage 2008;**35**(2):162-70.
8. Everitt BS, Landau S, Leese M, et al. *Cluster analysis*. 5th ed. Chichester: John Wiley & Sons, 2011.
9. StataCorp. Stata 14 Cluster Stop reference manual. College Station, TX: Stata Press, 2015.
10. IBM Corp. IBM SPSS Statistics Base 22. Chicago, IL: IBM Software Group, 2013.
11. Tsiptsis K, Chorianopoulos A. *Data Mining Techniques in CRM: Inside Customer Segmentation*. Chichester: John Wiley & Sons, 2009.
12. Lega F, Mengoni A. Profiling the different needs and expectations of patients for population-based medicine: a case study using segmentation analysis. BMC Health Serv Res 2012;**12**(1):473.
13. Liu C-Y, Liu J-S. Mining the optimal clustering of people's characteristics of health care choices. Expert Systems with Applications 2011;**38**(3):1400-04.
14. Nagel GC, Schmidt S, Strauss BM, et al. Quality of life in breast cancer patients: a cluster analytic approach. Breast Cancer Res Treat 2001;**68**(1):75-87.
15. Leijon O, Härenstam A, Waldenström K, et al. Target groups for prevention of neck/shoulder and low back disorders: an exploratory cluster analysis of working and living conditions. Work 2006;**27**(2):189.
16. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. 3rd ed. Waltham, MA: Morgan Kaufmann, 2011.