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. 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 taken 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. This statistic is commonly used in healthcare clustering studies, and has been identified as one of the best criteria to determine the number of clusters. It compares the between-cluster to the within-cluster sum-of-squares, and a large Pseudo-F statistic indicates distinct clusters.

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

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Pseudo-F Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 clusters</td>
<td>2249</td>
</tr>
<tr>
<td>3 clusters</td>
<td>2745</td>
</tr>
<tr>
<td>4 clusters</td>
<td>2662</td>
</tr>
<tr>
<td>5 clusters</td>
<td>2374</td>
</tr>
<tr>
<td>6 clusters</td>
<td>2267</td>
</tr>
<tr>
<td>7 clusters</td>
<td>2131</td>
</tr>
<tr>
<td>8 clusters</td>
<td>2041</td>
</tr>
</tbody>
</table>

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. 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. Indeed, many studies combine hierarchical clustering with k-means in a two-stepped approach.

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

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
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. 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\textsuperscript{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.\textsuperscript{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.
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
10. IBM Corp. IBM SPSS Statistics Base 22. Chicago, IL: IBM Software Group, 2013.