METHODS

Variables

Cost, inpatient days, hospital admissions, ICU days, ED visits, and major OR procedures were log-transformed. All variables were then range standardized[[1]](#endnote-1).

Statistical Analysis

A two-step process was used to cluster patients. First, a hierarchical clustering was created using Ward’s minimum-variance method. Ward’s method is an agglomerative clustering algorithm; at each agglomeration step the increase in within-cluster error sum of squares (ΔSSE) is minimized. The hierarchical clustering diagram was used to choose initial patient clusters by taking a level cut’, i.e. by taking the highest-level clusters with a ΔSSE below a cutoff. Expert judgment was used to decide the coarseness of the initial clusters. Factors taken into consideration for this decision included the characteristics of potential clusters, PCA plots, and the ΔSSE of different cluster numbers. In the second clustering step, the k-medoids algorithm was run to produce our final clusters. The cluster number and initial seeds for k-medoids were given by our initial clusters.

**Bootstrap:**

The stability the clustering solution was assessed by bootstrap. The entire clustering process from range standardization to k-medoids was re-run for 1000 bootstrap samples. The bootstrap stability of each patient was the proportion of samples wherein the bootstrap cluster matched the reference cluster. Bootstrap cluster labels were matched to reference label by best total stability score.

**CV:**

The stability the clustering solution was assessed using 50/50 cross-validation[[2]](#endnote-2)[[3]](#endnote-3). The data were repeatedly split into random halves and each half was independently clustered. The entire clustering process from range standardization to k-medoids was run on each half, with the exception that the cluster number k=5 was fixed. The clustering in each half of the split was then further used to cross-classify patients in the opposite half. Cross-classification was by assignment to nearest medoid, effectively applying the final k-medoids assignment step to across-split patients. The cross-validation stability of each patient was then defined as the proportion of repetitions wherein clustering labels matched cross-classification labels. Cluster labels for the two halves were matched by best total stability score.

Descriptive analysis of the resulting patient clusters was performed. Association of the utilization variables with cluster membership was assessed by the Kruskal-Wallis test. In addition, the relative incidence of 24 conditions in each cluster was assessed. Odds ratios and statistical significance for the conditions were calculated by logistic regression. Odds ratios were calculated in comparison to mean incidence. R version 3.0.2 was used for all analyses.

1. (Milligan, 1988) [↑](#endnote-ref-1)
2. (Bennetts, 2013) [↑](#endnote-ref-2)
3. (Jain, 2010)

   Google Scholar links for citations:

   <http://scholar.google.com/scholar?hl=en&q=A+Study+of+Standardization+of+Variables+in+Cluster+Analysis&btnG=&as_sdt=1%2C36&as_sdtp=#>

   <http://scholar.google.com/scholar?q=Clustering+and+classification+of+regional+peak+plantar+pressures+of+diabetic+feet&btnG=&hl=en&as_sdt=0%2C36#>

   <http://scholar.google.com/scholar?cluster=16205305846126879965&hl=en&as_sdt=0,36> [↑](#endnote-ref-3)