

**Title: An Application of Cluster Analysis to Health Services Research: Empirically Defined Health States for Depression from the SF-12**

**Author(s): Catherine A. Sugar, Leslie A. Lenert, and Richard A. Olshen**

**Technical Report number (Dept. of Statistics, Stanford Univ.): 203**

**Date: June 1999**

**Abstract:**

**Objective** The primary purpose of this research is to define objectively and describe a set of clinically relevant health states that encompass the typical effects of depression on quality of life in an actual patient population. The model we present is designed so as to facilitate the elicitation of patients' and the public's values (utilities) for outcomes of depression. During the development of the requisite analytical techniques, many interesting statistical issues arose which we also attempt to address. These include preprocessing of survey data for clustering, choosing the optimal number of clusters to represent a data set, measuring the stability of groups found via cluster analysis, and studying longitudinal trends for clustered data.

**Data Sources** Our data come from the depression panel of the Medical Outcomes Study. They include scores on the 12-Item Short Form Health Survey (SF-12) as well as independently obtained diagnoses of depression for 716 patients. Follow up information, one year after baseline, is available for 266 of these patients.

**Methodology** We use a new methodology based on k-means cluster analysis to group the patients according to appropriate dimensions of health derived from the SF-12 scores. Chi-squared and exact permutation tests are used to validate the health states thus obtained, by checking for baseline and longitudinal correlation of cluster membership and clinical diagnosis. Techniques based on cross-validated distortion, broken-line regression and rate-distortion theory are developed for identifying the underlying structure and optimal number of clusters present in a data set.

**Principal Findings** We find, on the basis of a combination of statistical and clinical criteria, that six states are optimal for summarizing the range of health experienced by depressed patients. Each state is described in terms of a subject who is typical in a sense that is articulated with our cluster analytic approach. In all our models, the relationship between health state membership and clinical diagnosis is highly statistically significant. The models are also sensitive to changes in patients' clinical status over time. Furthermore, the techniques developed for identifying structure and optimal number of clusters are shown to be extremely accurate in cases where the underlying distribution

of the data is known and a theoretical basis for the success of the methodology is provided using ideas and results from rate-distortion theory.

**Conclusions** Cluster analysis is demonstrably a powerful methodology for forming clinically valid health states from health status data. The states produced are suitable for the experimental elicitation of preferences and analyses of costs and utilities. The techniques developed here for forming health states can be applied more generally to study the underlying structure of clustered data.