

Title: **Clustering in the Presence of Missing Data: The Design of Preference-Assessment Surveys in Health Care**

Author(s): **Alfred Lin and Richard A. Olshen**

Technical Report number (Dept. of Statistics, Stanford Univ.): **196**

Date: **February 1998**

Abstract:

**Background.** The value that patients and the public place on health outcomes is patently important. Cluster analysis and associated imputation can be useful tools to define prototypical health-states empirically. One purpose of these descriptions is to facilitate the assessment of preferences.

**Methods.** Two hundred twenty-four patients with ventricular arrhythmias who were treated at Kaiser-Permanente of Northern California were surveyed about physical functioning using the Duke Activity Status Index (DASI), mental status and vitality using the Medical Outcomes Study Short-Form 36-item (SF-36), and symptoms using items from the Cardiac Arrhythmia Suppression Trial (CAST). Various nonparametric approaches to imputations were explored as preprocessors because there were missing data. Then a “*k*-means” clustering algorithm was used to identify prototypical health-states, in which patients in the same cluster shared similar responses to items in the survey.

**Results.** The imputation and clustering algorithms yielded four prototypical health-states. Cluster 1 is characterized by high scores on physical functioning, vitality and mental health. Cluster 2 had low physical function but high scores on vitality and mental health. Cluster 3 had low physical function, low vitality, but preserved mental health. Cluster 4 had low scores on all scales. These clusters serve as the basis of written descriptions of the health states.

**Conclusions.** Employing imputation and clustering algorithms to analyze health status survey data enables a data-driven, concise summary of the experiences of patients. These experiences can be used to generate written health-state descriptions for purpose of assessing their relative values to patients.