Methods Section

Empirically Defined Health States for Depression from the SF-12

Catherine A. Sugar, Roland Sturm, Tina T. Lee, Cathy D. Sherbourne, Richard A. Olshen, Kenneth B. Wells, and Leslie A. Lenert

Objective. 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. Our model was designed to facilitate the elicitation of patients' and the public's values (utilities) for outcomes of depression.
Data Sources. From the depression panel of the Medical Outcomes Study. Data 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, was available for 166 of these patients.
Methodology. We use *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.
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 of 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.
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 preference and analyses of costs and utilities.

Key Words. Cluster analysis, health status measures, health states, utilities, quality of life

The cost-effectiveness of different strategies for the identification and treatment of depressive illness in primary care practice is the focus of one of the Patient Outcomes Research Team (PORT) projects of the Agency for Health Care Policy and Research. When medical services or treatments extend life, their values can be easily summarized. However, when it is primarily quality

of life that is improved, it is often difficult to assess the medical importance and cost-effectiveness of any particular intervention. In all but the most extreme cases of depression, treatments will primarily improve quality of life as opposed to its length. Strategies for measuring the significance of gains in quality of life in a patient population fall into two general categories: health inventories and assessments of utility (Guyatt, Feeny, and Patrick 1993). This article describes a strategy for merging these two approaches and its application in the context of a depression PORT project. By combining these strategies we make possible the interpretation of health status data from the PORT project from a cost-effectiveness standpoint.

The research plan of the PORT project calls for the use of a standard health inventory to measure repeatedly the effects of interventions on depressed patients' quality of life over a two-year period. A health inventory is a survey designed to measure a patient's level of functioning in different aspects of life. Examples include the Sickness Impact Profile (Berger et al. 1981), the Short Form (SF)-36 instrument (Stewart and Ware 1992; Ware and Sherbourne 1992; Hays, Sherbourne, and Manzel 1993), and most recently the SF-12 (Ware, Kosinski, and Keller 1996). The dimensions represented in health inventories vary, but typically include some distinction between mental and physical health. Many health inventories also involve the evaluation of attributes such as role functioning, anxiety, and expectations of future health. The goal of a health inventory is to enable a comparison of patients' health with that of the population as a whole and with that of groups defined by a specific illness. Health inventories are employed widely in clinical trials. They are often particularly valuable for the subtle and difficult task of quantifying the effects of a treatment on aspects of health that are not solely aspects of the disease under study.

This work was supported in parts by Grants HSR8349 and HSO0028-9 from the Agency for Health Care Policy and Research and Grant CA 59039 from the National Institutes of Health.

Address correspondence and requests for reprints to Leslie A. Lenert, M.D., Assistant Professor, Division of Clinical Pharmacology, Stanford University School of Medicine, Stanford, CA 94305-5113. E-mail: *lenert@SMI.stanford.edu.* Catherine A. Sugar, M.S. is a Doctoral Student in the Department of Statistics, Stanford University; Roland Sturm, Ph.D. is an Economist at RAND; Tina T. Lee, M.D. is a Fellow in the Division of Health Care Research and Policy, and General Internal Medicine, Stanford University School of Medicine; Cathy D. Sherbourne, Ph.D. is a Senior Health Policy Analyst, RAND; Richard A. Olshen, Ph.D. is a Professor in the Departments of Health Research and Policy, Electrical Engineering, and Statistics, Stanford University; and Kenneth B. Wells, M.D. is a Professor in the Department of Psychiatry, University of California at Los Angeles. This article, submitted to *Health Services Research* on February 13, 1997, was revised and accepted for publication on October 16, 1997.

Health inventories typically are self-administered questionnaires and thus provide a relatively low-cost way to measure changes in quality of life, especially in longitudinal studies. However, their use as outcome measures for trials may restrict the types of economic analyses that can be performed. One important limitation of most health inventories is that they do not provide a single number that summarizes the overall quality of life of a patient. When treatment results in gains in health status in one dimension and losses in another, the net impact of the therapy is unclear. A second problem is that scores are generally based on the degree of abnormality of functioning (relative to that of the population mean) rather than on the significance to the patient of that level of abnormality. For example, a change from the 70th to the 80th percentile of population functioning may or may not represent a clinically important improvement in health. As a consequence, measurements of health status may provide little insight into the appropriateness or relative value of a therapy. The elicitation and measurement of utilities is an alternative method for quantifying changes in global quality of life (Torrance 1986). Utility-based approaches summarize the overall desirability of a health state by a single number called the utility value, usually scaled to lie between 0 and 1. The lower end of the scale is equivalent to death (or a state valued even less—think here of something that amounts to torture), while the upper end of the scale represents perfect health. Various frameworks, such as the standard reference gamble, the time trade-off, and visual analog scales, are used to determine the distance of the health state from the worst and best states that anchor the utility scale. Direct measurement of the utility values of individuals for current health has been used occasionally in clinical trials (Bennett, Torrance, and Tugwel 1991). Models based on economic criteria often use utilities of hypothetical health states or patients' current health for purposes of assessing the overall benefit and cost-effectiveness of a therapy. However, in practice, utilities are difficult to measure. An interview involving either a trained research assistant or an automated computer program is necessary. Interviews, in general, require the physical presence of the patient; and as a result, the collection of serial observations of utilities is both time-consuming and expensive.

Several investigators have developed models designed to link health inventories with the utilities of patients or of the general population. Examples include the Quality of Well Being Scale (Kaplan 1989), the Health Utilities Index (Feeny, Torrance, and Furlong 1996), and the EuroQol model (EuroQol Group 1990). To develop these models, investigators describe sets of health states to patients or normal subjects who, in turn, provide utilities using one or

more of the scaling methods. To use this approach it is necessary to develop descriptions of the effects of disease on quality of life for people with different patterns of health. However, no technology has been widely accepted for building these "health state descriptions." Previous approaches have relied primarily on the subjective judgment of putative experts, both for defining and describing the health states (Bennett and Torrance 1996). Most researchers begin with a factorial model, in which several important dimensions of health are selected and divided into equally spaced levels. However, the number of health states can rapidly become unwieldy with this approach. If a given index has d dimensions of health, and there are L levels in each dimension, then the total number of combinations of levels of health is L^d. Reducing the number of levels by erasing boundaries reduces the total number of states in the model and thus reduces the variability in assigning any individual to any particular health state, but it may increase the bias.

Here we describe an alternative approach to developing models that facilitates mapping between measures of health status and ratings of overall quality of life and that is also efficient in terms of the number of health states required. We begin with a health status model, the SF-12, which was developed by Ware, Kosinski, and Keller (1996). We apply this model specifically in the context of depressive disorders. The SF-12 was designed to measure two general aspects of the health of a population: mental health and physical health. In this study, rather than defining the health states using arbitrary levels of functioning on the mental and physical health scales, we use cluster analysis to identify the set of states that best summarizes the "clumping" of the data observed in a large population. The clumps are taken to correspond to health states. If our method has merit, as we believe it does, then the clumps (or clusters) found should permit easy interpretation. Thus, we illustrate how the method can be used to develop an empirical model of and objective descriptions of the effects of depression on quality of life in a manner suitable for use in experimental elicitation of preferences. When the utilities for these states are measured, in an appropriate population, it will be possible to estimate changes in quality-adjusted life years from a time series of SF-12 scores.

METHODS

Health Status Measures

The SF-12 is a short health status instrument, derived from the widely used 36-Item Short Form Health Survey (SF-36) (Stewart and Ware 1992; Ware

and Sherbourne 1992). Current approaches to summarizing outcomes of the SF-12 involve composite scores for mental and physical health (Hays, Sherbourne, and Manzel 1993). The scores summarize the degree of functioning for patients in each of these two aspects of quality of life but do not include preference weights. We use composite scores for the SF-12 for mental and physical health based on the scoring rules of Ware (Ware, Kosinski, and Keller 1996) to form our model of health status in depression.

The Medical Outcomes Study (MOS) was an observational study of care in many different systems. It included groups of patients with depression, recent myocardial infarction, congestive heart failure, hypertension, and diabetes mellitis. We use data from the depression panel of the MOS. Details about the design of this component of the study can be found in Wells and Sturm 1996). The MOS independently identified depressed patients in a general medical practice setting using a two-stage screening survey. All patients exceeding the depression symptoms screening survey cut-point for high probability of having depressive disorder were identified as having *depressive symptoms.* Those with current major depression or dysthymia, according to the criteria of the American Psychiatric Association's *Diagnostic and Statistical Manual, Third Edition* (DSM-III), were defined as having *current depressive disorder.* Those with both types of disorder were referred to as having *double depression.* Persons with depressive symptoms but no current disorder were considered to have *sub-threshold depression.* A distinction was made between those persons with lifetime (i.e., past) depressive disorder and those with no history of depressive disorder. For the longitudinal aspects of the study, all patients with a current disorder and half of those with sub-threshold symptoms were enrolled. Therefore, the data span a wide range of health states and represent many of the configurations found in actual psychiatric practices. However, due to the sampling procedure, the patients in our data set are somewhat more depressed than would be typical of patients in a general practice.

Clinical diagnoses were obtained by a structured, computer-assisted telephone interview at baseline and at one year. The SF-12 items were self-administered at baseline and at one year. Unfortunately, one item that elicits information about energy levels was omitted in the year one follow-up questionnaire, so that we essentially use SF-11 in the longitudinal section of our analysis.[1] This omission is unlikely to affect our conclusions since the mental and physical scores derived from the SF-11 and SF-12 at baseline are not just highly correlated, but in fact are virtually identical in magnitude. Complete SF-12 data are available at baseline for 716 of the 974 patients

initially enrolled in the study, and all subsequent analyses involve only this subset. The basic characteristics of the patient population at baseline can be found in Table 1.

Statistical Methods

The goal of our analysis was to produce a set of well-defined health states that summarize mental and physical health in depression. These dimensions were selected because they explain most of the variability captured by the SF-12. Thus, our data consist of 716 points in two-dimensional space, with one axis representing physical health and the other mental health. Typically, the next step would be to divide the range of the data into a uniform grid by evenly dividing each dimension into several levels. This approach is inefficient and restrictive. Many of the health states defined in this manner are empty or nearly so because the combinations of physical and mental health simply do not exist in the relevant patient population. Most importantly, there is no a priori reason why health states should be spaced symmetrically in the given scales. We use k-means cluster analysis to "allow the data to speak for themselves" in defining the states. The purpose is to produce a parsimonious set of discrete, possibly asymmetrically spaced, clusters that best summarize the data. Clusters are completely determined by points called cluster centers, each of which represents its respective prototypical cluster member.

We motivate the algorithm thus:

> Begin with k points in the range of the data; call them $c_1, c_2, \ldots, c_k$. Each point in the data set will be closest to one of those k points, provided that ties are broken by choosing the point that is both closest and has the lowest index. Define the ith cluster to be all points in the data set that are closest to the point c_i. Calculate the sum of squared distances from the data points to their respective cluster centers c_i. This quantity is the *distortion*. The k-means algorithm defines the optimal

Table 1: Descriptive Statistics for the Patient Population (s.d. in Parentheses)

Clinical Diagnosis at Baseline	*n*	*SF-12 Mental Health Score*	*SF-12 Physical Health Score*	*New Episode in Year 1*	*Remission in Year 1*
Sub-threshold	237	44.4 (19.2)	43.7 (11.5)	13%	NA
Lifetime	103	41.0 (10.6)	45.8 (10.6)	30%	NA
Major depression	163	37.4 (12.1)	47.6 (11.5)	NA	45%
Dysthymia	88	38.2 (11.7)	42.7 (11.9)	NA	35%
Double depression	125	32.0 (9.8)	45.6 (10.9)	NA	38%
Total	716	39.4 (11.7)	45.1 (11.4)	NA	NA

set of cluster centers to be the set of points $c_1, c_2, \ldots, c_k$ for which distortion is minimized. Data points are then assigned to clusters as before. It is not obvious how to find the cluster centers specified by the k-means algorithm. An iterative search procedure called the Lloyd algorithm is usually quite successful at finding "good" cluster centers, but there is, in general, no closed-form solution to the problem. (See Gersho and Gray 1992; Kaufman and Rousseeuw 1990, for more on cluster analysis and the Lloyd algorithm.)

Several issues are of technical concern regarding use of the Lloyd algorithm for k-means clustering, including the choice of dimensions on which to cluster, scaling of the dimensions, and initialization of the algorithm. Interested readers can find details on our Website at *Http://preferences.stanford.edu/cluster_details.html.* Note also that the k-means algorithm specifies cluster membership once the number of clusters is fixed, but does not specify the number of clusters that should be used. We will discuss this issue since it is a central part of our method for formulating models. Since the cluster centers are chosen based on distortion, the root mean-squared distance to cluster centers (hereafter called RMSE for root mean-squared error) seems a natural criterion. There is a clear trade-off between too few clusters and too many. With a single cluster, RMSE is large. With as many clusters as data points, RMSE is 0. In neither instance are the data summarized in a meaningful manner. For any given data set, a plot of RMSE versus number of clusters will produce a steadily decreasing curve. Typically, at some point, the rate of decrease will drop sharply because the data are genuinely clumped into a fixed number of clusters. The "kink" in the curve where the slope changes most abruptly determines the optimal number of clusters. It is often possible to identify the range of reasonable values for k by visual inspection. However, one can pinpoint the kink more precisely by fitting a broken-line regression to the RMSE curve. (See Sugar, Sturm, Lee, et al. 1997 for a discussion of the statistical properties of this approach. Details may also be obtained on our Website.) There is a technical point that has been neglected in this discussion. The RMSE of future interest is the predicted RMSE for a new patient "out of study." However, if the model is fit to a given data set, and these data are all that are available, the RMSE obtained will underestimate RMSE for a new patient. This is because the model was designed to minimize RMSE for points in the data set but not for points outside the data set. This underestimate of prediction error can be compensated for by using m-fold cross-validation. (See Breiman et al. 1984; Efron and Tibshirani 1993 for details.) We used tenfold cross-validation on our data set, and tried numbers of clusters from $k = 2$ to $k = 16$. For the best models that we identified by means of our RMSE plot, we estimated the cluster centers as accurately as possible by using all the data.

We compared the relative efficiency of cluster analysis with the traditional grid design for forming health states. For the grid design, we divided each dimension of health into one, two, three, or four equally spaced levels. This produced models with one, four, nine, or sixteen health states. The center of each grid square was taken as the cluster center, and patients were classified to the cluster with the closest center, just as in the k-means framework.

It is also important to check the stability of the clusters produced by any clustering algorithm. A dramatic change in cluster centers based on which points are included in the sample provides evidence that what is being seen is an artifact of the data set rather than a genuine grouping. Further, if cluster analytic methods identify clinically distinct and meaningful groups, both the cluster centers and cluster membership should stay relatively stable as the number of clusters increases. We tested this by visual inspection and, more formally, by measuring the predictability of cluster assignments for consecutive values of k using a statistic that we denote by G_{ij}. This statistic is designed to find systematic patterns of covariation in ordinal data, much as the squared correlation coefficient does for ordered data; it is useful especially when these patterns are not visually obvious (Goodman and Kruskal 1979). Like R^2, G_{ij} lies between 0 and 1, but what constitutes a "large" value must be determined from the parameters of the data set. This is usually done by way of a simulation study. Details of the implementation may be found on our Website.

To validate health states defined by our methodology, we used several techniques. We checked whether cluster membership was significantly correlated with clinical diagnosis. One would not expect the match to be perfect since the SF-12 scores reflect patients' perceptions of their mental and physical health, which can vary widely within a diagnostic category. Nonetheless, the different diagnoses in our study, from sub-threshold to double depression, can be ordered according to severity; and one would expect the more severe diagnoses to be associated with the poorer health states. We tested this hypothesis using a chi-squared test for independence of diagnosis and cluster membership. The tests of cluster stability described earlier were also validatory in nature. Finally, we examined whether improvements (deteriorations) in clinical diagnosis were associated with movements to states with better (worse) mental health ratings. Follow-up data, one year after baseline, were available for 266 patients. These patients fell into two categories: those with depression and double depression on initial evaluation (active disease; $n = 137$) and those with dysthymia or sub-clinical symptoms (symptoms only; $n = 129$). Of the patients with initial active disease, 54 experienced

a remission within the first year. Of the patients with symptoms, only 28 developed depression or double depression within one year. We compared the joint distributions of cluster membership at baseline and one year for each of these two groups. A traditional chi-squared test of independence was found to be inappropriate due both to an insufficient amount of data for the number of cells and to the presence of many empty cells. Thus, we used an exact permutation test, subsampling the permutations (Efron and Tibshirani 1993). This method is analogous to a chi-squared test and uses the same test statistic. However, instead of assuming that the statistic has a chi-square distribution, one simulates the actual distribution. The p-values are then determined empirically based on the simulated distribution.

Health State Descriptions

After identifying the best model using k-means analysis tempered by clinical judgment, we studied the distribution of answers to the original SF-12 items within each cluster. For each of these items, we found, in each cluster, either a single response that accounted for at least 50 percent of patients in that cluster or, if there was no such response, the two responses with the most repetitions. (In every case, this accounted for more than 50 percent of patients.) We then developed health state descriptions for each cluster by working backwards from the item responses and casting them in the clinical context of depression.

RESULTS

The result of this stage of our research is an objectively based and clinically useful description of six typical patterns of physical and mental functioning observed in patients with depression. Each description is derived directly from the individual items that comprise the SF-12. As a basis for each description we take the one or two responses to an item that account for at least 50 percent of the patients in the cluster. We then construct a clinical description of the state based on those values. A full set of descriptions can be viewed on our Website. Consider as an example the description of State 3. It corresponds to a good physical health score on the SF-12 but a poor mental health score. The prose description of the state and the proportion of patients in each diagnostic category are shown in Table 2. Patients in this state do not experience difficulties performing moderate physical activities. However, they do exhibit standard features of depression, including sadness, anxiety, loss of concentration, and withdrawal from others. Descriptions for

Table 2: Distribution of Health States in Each Diagnostic Category for the Six-Cluster Model, with a Sample Description for Health State 3; More Severe Diagnoses (such as Double Depression) Have a Greater Proportion of Patients in the Worse Health States (states 3, 5, and 6)

Percent in Each State	*Sub-Threshold*	*Major Depression*	*Dysthymia*	*Double Depression*
State 1	19.1	13.5	17.0	5.6
State 2	28.8	19.0	18.2	8.0
State 3	7.1	24.5	15.9	24.8
State 4	17.1	17.8	8.0	16.8
State 5	17.9	14.7	18.2	20.8
State 6	10.0	10.4	22.7	24.0
	100%	100%	100%	100%

chi-square = 87.7.
$p < .0001$.

Sample Health State Description, State 3:

"You think your general health is good to very good. It is clear to you that it is not excellent, but it is better than fair.

"Your physical health does not limit your ability to perform moderate activities, work, or do simple things like climbing several flights of stairs. Still, for half the people in this state, their physical health prevents them from accomplishing things that they would like to do.

"Mostly, it is emotional problems that prevent you from accomplishing all that you would like to do. You have difficulty concentrating at times. Problems with concentration also keep you from working as carefully as you normally would. It is not pain that causes you to have difficulty, but other problems. Anxiety is your constant companion. You almost never feel peaceful or calm. You do not have your usual level of energy. Some people in this state have a lot of energy a good bit of the time; others only have it a little of the time. You could fall anywhere in between. A feeling of sadness is commonplace in your life. A good bit of the time to most of the time you feel downhearted and blue. You spend more of your time alone than you typically would. People in this state find that their emotional health limits their social activities at least some of the time. For some people, their emotional health limits their social activities most of the time. You could fall anywhere in between."

the remaining clusters have similar clinical coherency, as readers can discern easily for themselves. In addition to having clinical coherency, these six states accurately reflect the distribution of functioning in a population of fairly typical patients. Further, the test of the descriptions is relatively objective, as it is tied closely to the language of the original questionnaire.

The-six-state model is the product of a series of statistical analyses designed to identify the optimal number of states in combination with clinical judgment. Figure 1 shows how the RMSE varies with numbers of clusters for factorial models and cluster-analytic models. For any given k, the model

defined by cluster analysis has a lower RMSE, and a specified level of RMSE is achieved using fewer than half as many states as are entailed by a competitive factorial design. The optimal number of clusters appears to be between four and six, since this is where the plot of RMSE versus number of clusters exhibits a marked change in slope. This result is confirmed by fitting a broken-line regression to the RMSE data points. The best obtained sets of cluster centers for models with four, five, and six states are shown in Figure 2. By looking at any of the cluster models, we see why the factorial design is inefficient; the states are indeed asymmetric.

A critical issue for our methodology is whether the cluster analysis identifies distinct and clinically meaningful groups. If it does, then the location of cluster centers and cluster membership both should be conserved as the number of clusters increases. When a new cluster is introduced, we would hope that some states in the model would be split and that others would remain largely unaffected. This is evidenced in the models for our data set.

Figure 1: Root Mean Distortion Versus Number of Clusters for the Standard Grid and *k*-Means Methods; the *k*-Means Methodology Requires Many Fewer Clusters to Achieve a Given Level of Distortion

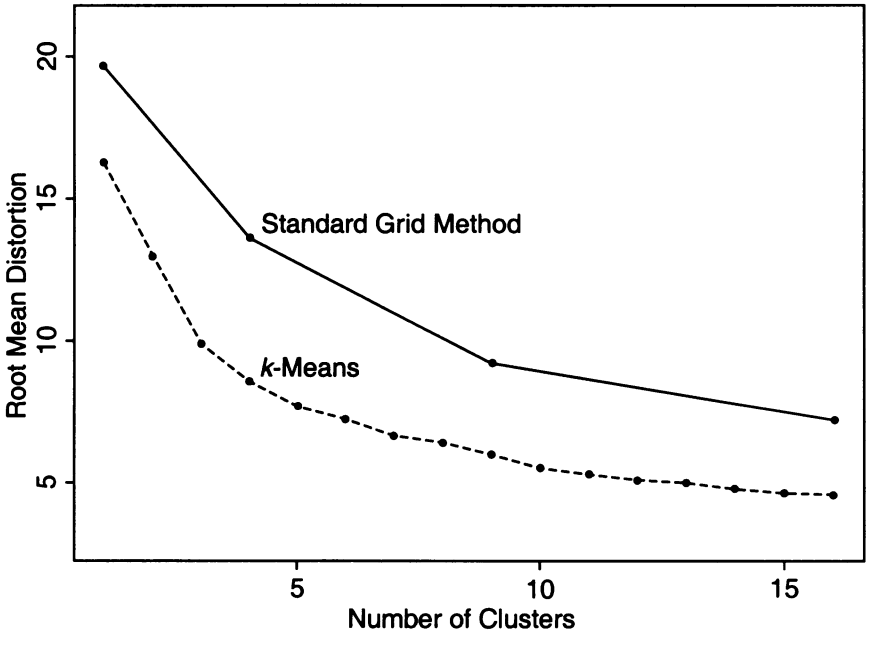
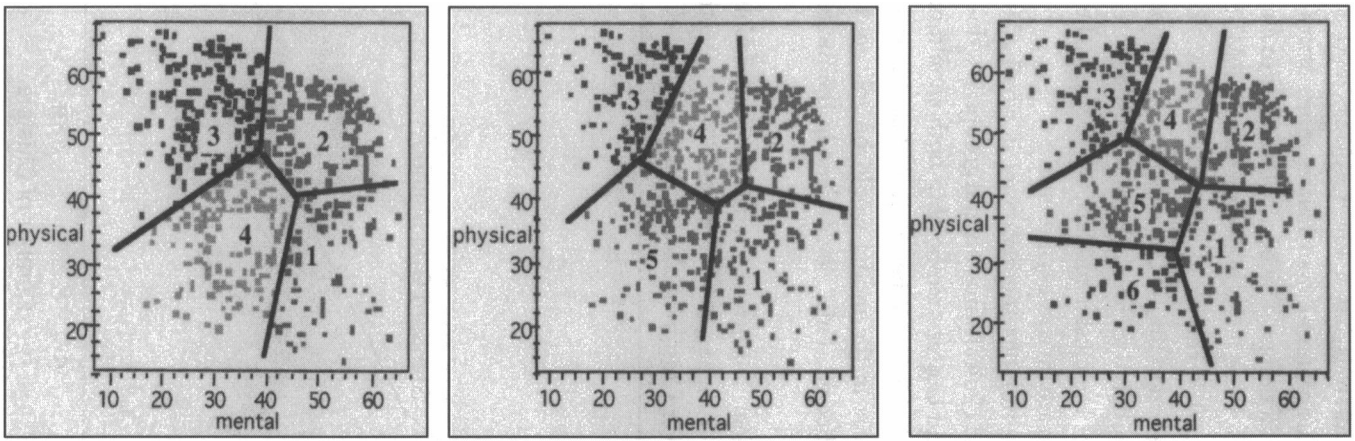


Figure 2: The Four-, Five-, and Six-Cluster Models for SF-12 Scores in Depression



Note: Note stability in the pattern of clustering over increasing numbers of cluster centers in this range; these results suggest that cluster centers represent biological types identified by the statistical algorithm.

The five-cluster model is similar to the four-cluster model, except that in the five-cluster model, three clusters rather than two are used to describe the range of mental health in the context of normal physical functioning. The clusters involving physical impairment are largely unchanged. The cluster assignments for the four- and five-state models were highly related, with a value of $G_{ij} = .768$. The six-cluster model extends the five-cluster model by splitting patients with severe mental and physical impairments (State 5 in Figure 2(b)) into two separate groups according to their physical functioning. Cluster assignment between the five- and six-state models was even more stable, with $G_{ij} = .794$. Permutation distributions were calculated for the G_{ij} statistic in each instance. For the $k = 4$ to $k = 5$ comparison, the maximum value out of 10,000 trials was $G_{ij} = .0514$ and for the $k = 5$ to $k = 6$ comparison, the corresponding maximum was $G_{ij} = .0586$. The p-values for a test of no correlation were therefore less than .0001 in both cases.

In a biologically valid model, cluster membership should be associated with clinical diagnosis. To test this hypothesis, we combined all patients with sub-clinical disease into a single category, producing four groups: sub-clinical disease, depression (by DSM-III criteria), dysthymia (by DSM-III criteria), and both depression and dysthymia (so-called double depression). Using these groupings, we examined the association between clinical diagnosis and health state membership for the four-, five-, and six-cluster models. The results for the six-state model are shown in Table 2. Associations between clinical diagnosis and health state membership are highly statistically significant for all the models ($p < .0001$). Similar findings obtain for the four- and five-state models.

A clinically useful model should also be responsive to changes in clinical diagnoses over time. Over the one-year follow-up period of the MOS study, the clinical diagnoses and health state assignments changed for some patients. (We can determine a patient's health state at the end of the study by using follow-up SF-11 physical and mental scores, properly adjusted, to assign them to a cluster from the original model.) Just as health states were associated with diagnoses at baseline, we expect changes in health states to be associated with changes in diagnoses. We separated the patients with follow-up data, as described earlier, and for each group tested whether movement patterns were significantly different for patients who did or did not experience a change in diagnosis. For the patients who started with an active illness, the test for a difference in distribution for patients who did and did not experience remission was significant at $p < .0001$. For the patients who had no active illness at baseline, the test for a difference in distribution for patients who did and did not experience the onset of active depression had $p < .024$.

A good model should lead to improved understanding of the way in which clinical illnesses affect quality of life. Focusing in on the types of clinical associations seen with the six-cluster model (Table 1), we observe an interesting pattern. States with loss of physical functioning are more strongly associated with the chronicity of patients' depressive symptoms than with the severity of those symptoms. States 5 and 6, the states with loss of both mental and physical functioning have a high proportion of patients with dysthymia, a clinical diagnosis characterized by a lower severity and a longer duration of symptoms than acute depression. In contrast, State 3, which represents patients with poor mental health but little impairment of physical functioning has a high proportion of patients with acute depression. These findings clarify previous observations by Wells et al. (1992) with regard to the effects of dysthymia on patients' quality of life based on the same set of data. Wells et al. attributed the lower physical health of dysthymic patients to the "disease entity" of dysthymia. Our results suggest that there is a subgroup of dysthymic patients, identifiable by cluster-analytic methods, who have severe impairments in both mental and physical health. These patients may have mental health problems that are secondary to the physical effects of their primary illness rather than physical illnesses caused by their mental disorder. Specific evaluation of this hypothesis is possible but beyond the scope of this article.

DISCUSSION

There are many possible strategies for developing a globally complete set of descriptions of the effects of depression on quality of life for the purpose of studying patients' or society's preferences for health outcomes. Which are the most common problems? Which clinical problems best illustrate the pathophysiology of the disease? Which have the clearest implications for treatment? Examples of classical effects of depression on health can be found in the case studies book for the *Diagnostic and Statistical Manual* (DSM-IV), and in most textbooks on psychiatry. In some of these descriptions, emphasis is placed on symptoms that are clinically worrisome or that could suggest differences in treatment approaches. These include psychotic symptoms, recurrent depression, suicidal tendencies, or psychiatric comorbidities. Such descriptions do not necessarily separate all relevant health states in depression. In particular, they tend to exclude distinctions based on degree of physical functioning. Further, because some classic pathological symptoms are rarely seen in primary care populations, their relevance to general health policy

studies is limited. Therefore, it is difficult to determine the most salient health states based on an entirely clinical perspective. In the absence of an obvious clinical approach to defining a set of health states that adequately describe most depressed patients, we have turned to a purely empirical strategy.

The clusters, and hence the descriptions, of health states identified using this approach reflect the model of quality of life inherent in the health status instrument. The SF-12 model is oriented toward measuring functioning across a broad range of disorders, rather than symptomatology. Thus, our model may not capture the full emotional burden of sadness, hopelessness, and anxiety borne by patients with depressive illnesses. Rather, it captures the degree to which individuals are successful in coping with these burdens. The focus on functioning rather than symptoms allows the comparisons of levels of impairments across different diseases using the SF-12. However, there is no particular reason to limit analyses using this methodology to data from health inventories. Use of disease-specific scales to form clusters that include perceived levels of symptoms, such as the Hamilton Depressive Index, may be more useful to individuals attempting to make specific medical decisions as opposed to general policy decisions.

The standard method for forming health states for use in experiments in the elicitation of preferences is to form a factorial design, where each dimension represents a specific aspect of quality of life and is divided into multiple levels of functioning. It is unlikely that the types of health impairments seen in depression are unrelated (as is implicit in factorial designs for which cluster membership is defined only in terms of marginal rather than joint values of physical and mental health). It makes no sense, for example, to define states of depression with high psychological distress but no social limitations because it is likely that both impairments share the same primary cause. Our results suggest that cluster analysis is a useful alternative method that allows the investigator to identify the states that are prevalent in a typical patient population. In this application, for models with between 4 and 16 states, cluster analysis was a far more efficient approach in terms of the number of clusters needed to achieve a given level of "tightness" of the clusters. The RMSE was at least 50 percent lower for cluster analysis models than for their corresponding factorial models, no matter the number of states in the model (see Figure 1). This is due, at least in part, to the asymmetry of the states.

The models developed using cluster analysis appear to be clinically valid with strong statistical evidence of association between diagnosis and health state membership. This suggests that the model captures disease effects in a cross-sectional sense. However, we observe patients with both nearly normal and severely abnormal mental and physical health scores in every

diagnostic category. Therefore, our model captures biology in a way different from that of clinical diagnosis. The sub-types identified in the model may, in fact, represent biological types. As we introduce new states into the model, previously identified cluster centers are largely preserved (see Figure 2). If cluster analysis were merely summarizing the data rather than identifying unique types, we would not expect to see this remarkable stability. The cluster-analytic models also appear to capture the longitudinal variability associated with disease. One-year follow-up data show that asymptomatic patients who experienced the onset of major depression during the study have a different pattern of movement from state to state in the model than do healthy patients who did not develop depression. Patients who were depressed initially, but whose symptoms of depression went into remission, also have a pattern of movement different from that of ill patients whose symptoms did not go into remission. This suggests that when we apply this model with preference data, we are likely to have sufficient power to identify important changes in health due to medical interventions. However, further work is needed to establish whether the model is sensitive enough to detect clinically important changes. Additional work may also be needed to compare sensitivity of the model to that of general health status models such as the QWB scale, the Health Utilities Index, or EuroQol models.

Future Work

We are currently into the next phase of the research reported here: using the six-state model to measure utilities for primary care patients with sub-clinical or active depression. This model was selected to maximize the chances of detecting small improvements in the main dimensions of health. We are pursuing two approaches to assessment of preference for health states in the model: (1) measurement of preferences of patients who are in each state (as determined by their SF-12 scores), and (2) measurement of preferences of patients for the six states in our model, considered hypothetically. We will compare the utilities of subjects in each state to the utilities for the states when rated on a hypothetical basis. Once we have obtained reliable estimates for the mean utility for each state, it will be possible to estimate the utility for any SF-12 score in the context of depression.

Limitations

The approach illustrated in this article requires a large clinical database and is difficult to apply when no such database exists. The validity of the states defined depends on the extent to which the patients in the study accurately

represent the population of interest. Our approach yields health states that are specific to depression as measured by the SF-12. Other illnesses may result in different patterns of mental and physical impairment, and other health status instruments might detect different patterns of loss. Further work is needed to explore the biological relevance of the states defined by this approach in depression as well as their relevance to other diseases.

CONCLUSIONS

The accurate definition of health states is a necessary first step to an exploration of the relationship between preferences and health outcomes. This article describes preliminary work on a new approach to defining health states based on statistical techniques. Cluster analysis appears to be a useful tool for identifying genuine health states, many fewer in number than with the traditional factorial designs. It can produce a clinically relevant summary of health, at least for depression. The method as we applied it produced "objective" descriptions of health states suitable for use in the experimental elicitation of utilities from depressed patients.

NOTES

1. The omitted question refers to the patient's experiences in the past four weeks. It reads: "Did you have a lot of energy: (a) all of the time, (b) most of the time, (c) some of the time, (d) a little of the time, (e) none of the time."

REFERENCES

Bennett, K. J., G. Torrance, and P. Tugwel. 1991. "Methodologic Challenges in the Development of Utility Measures of Health-related Quality of Life in Rheumatoid Arthritis." *Controlled Clinical Trials* (Supplement): 118S–28S.

Bennett, K. J., and G. W. Torrance. 1996. "Measuring Health Preferences and Utilities Rating Scale, Time Trade-Off and Standard Gamble Methods." In *Health Utilities Index: Quality of Life and Pharmacoeconomics in Clinical Trials*, edited by B. Spliker, pp. 235–65. Philadelphia: Lippincott-Raven Publishers.

Berger, M., R. A. Babbit, W. B. Carter, and JB. S. Gilson. 1981. "The Sickness Impact Profile: Development and Final Revision of a Health Status Measure." *Medical Care* 19: (6): 787–805.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees,* pp. 12, 75–77, 223, 234, 306–309. Belmont, CA: Wadsworth; since 1993, New York: Chapman & Hall.

Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap,* pp. 202–19, 239–41. New York: Chapman & Hall.

The EuroQol Group. 1990. "EuroQol: A New Facility for the Measurement of Health-Related Quality of Life." *Health Policy* 16 (3): 199–208.

Feeny, D. H., G. W. Torrance, and W. J. Furlong. 1996. *Health Utilities Index: Quality of Life and Pharmacoeconomics in Clinical Trails,* edited by B. Spliker, pp. 239–52. Philadelphia: Lippincott-Raven Publishers.

Gersho, A., and R. M. Gray. 1992. *Vector Quantization and Signal Compression,* pp. 323–31, 362–69. Boston: Kluwer Academic Publishers.

Goodman, L. A., and W. H. Kruskal. 1979. *Measures of Association for Cross Classification.* New York: Springer-Verlag.

Guyatt, G. H., D. H. Feeny, and D. L. Patrick. 1993. "Measuring Health-related Quality of Life." *Annals of Internal Medicine* 118 (8): 622–29.

Hays, R., C. Sherbourne, and R. Manzel. 1993. "The RAND 36-Item Health Survey 1.0." *Health Economics* 2 (3): 217–27.

Kaplan, R. M. 1989. "Health Outcome Models for Policy Analysis." *Health Psychology* 8 (6): 723–35.

Kaufman, L., and P. J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis,* pp. 111–16. New York: Wiley.

Stewart, A., and J. Ware, eds. 1992. *Measuring Functioning and Well Being: The Medical Outcomes Study Approach.* Durham, NC: Duke University Press.

Sugar, C. A., R. Sturm, T. T. Lee, C. D. Sherbourne, R. A. Olshen, K. B. Wells, and L. A. Lenert. 1997. "An Application of Cluster Analysis to Health Services Research: Empirically Defined Health States for Depression from the SF-12." Division of Biostatistics, Stanford University.

Torrance, G. W. 1986. "Measurement of Health State Utilities for Economic Appraisal: A Review." *Journal of Health Economics* 5 (1): 1–30.

Ware, J., and C. Sherbourne. 1992. "The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection." *Medical Care* 30 (6): 473–83.

Ware, J. J., M. Kosinski, and S. W. Keller. 1996. "A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity." *Medical Care* 34 (3): 220–33.

Wells, K. B., M. A. Burman, W. Rogers, and R. Hays. 1992. "The Course of Depression in Adult Outpatients: Results from the Medical Outcomes Study." *Archives of General Psychiatry* 49 (10): 788–94.

Wells, K. B., and R. Sturm. 1996. *Caring for Depression.* Cambridge, MA: Harvard University Press.