A longitudinal cost-benefit analysis using health state models

Catherine Sugar*, Gareth James; Leslie Lenert† and Robert Rosenheck‡

February 26, 2002

Abstract

Typically the effectiveness of medical treatments is assessed by comparing single number summaries of patients’ levels of health. However, this approach is often overly simplistic for analyzing complex diseases. In this paper we show how health state models can be used to perform longitudinal cost benefit analysis for such diseases. Such models are formed by clustering patients according to the dimensions that summarize the health characteristics of the population of interest. We illustrate the basic procedure by studying the effectiveness of a new drug, clozapine, on patients with schizophrenia. After constructing the model we examine the distributions across health states of the patients on each medication. Over time, subjects on clozapine are generally more likely to shift to “better” health states than those on standard medications. One of the major benefits of this approach is the ability to calculate long run stationary distributions which can be used to estimate utilities and financial costs for an average patient on each treatment. The resulting cost/benefit analyses can form the basis for long term health policy decisions.

1 Introduction

Clinical trials generally assess outcomes along a variety of dimensions. Different aspects of physical and physiological health are measured using health status instruments consisting of dozens of item responses. To make the analysis more manageable, the data are often summarized using a few composite scores. The results are thus presented as simple comparisons of the mean values of a handful of scales at various time points. In this study we explore use of health state modeling, an alternative approach to the analysis of complex clinical data, which seeks to harness more of the inherent richness of the available structural information. Patients are grouped into clusters or health states which take into account not only quantitative differences in disease severity, but also qualitative factors such as the relative predominance of physical or mental symptoms. Health state models allow an accurate but parsimonious representation of the clinical status of a patient population. Clinical change is not measured in terms of a simple increase or decrease on preset continuous scales, but rather by a set of transition probabilities between a variety of discreet medically distinct health states. Thus, the benefit of a medication is assessed by the probability of an individual achieving a particular state of health given the specified treatment.

Health state models have several key advantages over more traditional approaches. One is the ability to estimate long term effectiveness of a medication using data from clinical trials which are

*Marshall School of Business, University of Southern California
†Staff Researcher and Physician, Veterans Health Administration San Diego Healthcare System
‡Departments of Psychiatry and Epidemiology and Public Health, Yale Medical School
necessarily of finite duration. Such calculations are extremely difficult using traditional approaches because of the inherent dangers of extrapolating well beyond the range of data. However, once a health state model has been fit, calculation of transition probabilities between clusters over time makes it possible to estimate long run stationary distributions for patients on each type of medication. For example, one might find that, in the long run, one medication moves individuals to a state associated with better overall health but serious side effects while another medication has fewer side effects but does not lead to as significant an overall improvement.

Another advantage of health state models is that they can be used in utility analysis. The landmark report of the Public Health Service’s expert panel on cost-effectiveness of medical treatments (Gold et al., 1996) determined that the standard measure of outcome in cost-effectiveness analysis should be a specific measure of utility, the Quality Adjusted Life Year (QALY). The QALY is a cardinal measure of the value of a health state to the society and is scaled from 0 (a state equal to or worse than death) to 1 (the state of perfect health). Thus a person living one year at a health level of 0.5 QALYs has the same utility as a perfectly healthy person living for one-half year. By measuring outcomes in QALYs it is possible to directly compare the magnitude of health benefits achieved across different diseases. This approach also facilitates the application of cost-utility analysis, a method for evaluating the efficiency of investments in health at a societal level based on the ratio of the marginal cost of a therapy to the gain in QALYs. Some attempts have been made in the past to calculate these quantities but QALYs are unfortunately difficult to estimate using traditional methods of analysis. For example, Rosenheck et al. (1998) developed a measure of utility using a “Worst Health” to “Good Health” scale but while this scale had some of the characteristics of QALYs it did not directly provide a measure of them. However, after fitting a health state model it is straightforward to generate descriptions of a typical patient in each state and survey the population to estimate utilities for these states. The utilities are then combined with the stationary distributions to produce the long run average QALY for a patient receiving a given treatment. This value can then be compared to long run cost estimates.

In this paper we apply a health state model approach to one of the most comprehensive studies of a new type of antipsychotic medication for schizophrenia (Rosenheck et al., 1997). Schizophrenia is a severely disabling psychiatric illness that affects as much as one percent of the US population (Bromet et al., 1996). Common manifestations include positive symptoms such as hallucinations, delusions and paranoid ideation, and negative symptoms such as apathy, anhedonia and withdrawal. People with schizophrenia often require intensive treatment for extended amounts of time, including periodic hospitalization, at considerable expense. They also have difficulty maintaining employment and participating in social activities, and many receive public support payments. Costs associated with this illness run into tens of billions of dollars each year (Rice and Miller, 1986). Since the 1950’s the mainstays of treatment for schizophrenia have been anti-psychotic medications such as thorazine, haloperidol, and trifluoperazine which have vastly reduced the costs of care (Rosenheck and Leslie, 2000). Unfortunately, nearly one third of patients are refractory or only partially responsive to these treatments. Clozapine, introduced in the United States in the late 1980’s, was the first of a class of new, more effective, medications referred to as “atypical antipsychotics” because of their distinctive lack of extra-pyramidal side effects. Clozapine, in particular, has shown special promise in the treatment of patients with refractory schizophrenia (Kane et al., 1988). The Rosenheck et al. (1997) study was a 12-month, double-blind trial that compared clozapine and the frequently used conventional antipsychotic treatment haloperidol, at fifteen Veterans Affairs (VA) medical centers. This study was the first to complete a long-term assessment of the impact of clozapine on social, vocational and community functioning, side effects and societal costs, as well as clinical factors including positive and negative symptoms and general
symptoms of psychological distress such as anxiety and depression. The data indicated that for patients with refractory schizophrenia and high rates of hospitalization, clozapine was somewhat more effective at treating symptoms, had a lower incidence of side effects, and was comparable in cost to the conventional alternative. However, these results were based on comparing mean scores on a handful of standardized scales. While they provided an easily interpretable overall assessment of the data they did not identify discreet psychotic states or transition probabilities, and therefore could not be used to undertake a standard utility analysis in QALYs. Furthermore it was not possible to make an accurate assessment of the long term effects of each medication. Using a health state model we are able to achieve both these objectives.

The discrete state approach relies upon a variety of multivariate statistical techniques including principal components analysis, cluster analysis, and Markov chain theory. There are four stages in the analysis of this data. First, in Section 3, we identify the set of variables that best describes the health of the population of patients using a principal components analysis. Second, in Section 4, we apply k-means clustering to these variables to partition the data space into health states. This section also discusses several graphical tools that can be used to aid a “medical expert” in assessing the best number of health states. In Section 5, we estimate transition probabilities between states in the model and use them to find the steady state distributions of patients on each medication. Finally, we estimate QALYs and financial costs for patients in each health state and combine these with the stationary distributions to perform a long run cost benefit analysis for each medication.

2 Background and Data

The majority of the data collected in the Rosenheck et al. (1997) study consisted of scores on several standard health status instruments or questionnaires that measure a broad spectrum of emotional, interpersonal, and physical functioning. The first instrument, the Positive and Negative Syndrome Scale (PANSS), consists of thirty questions each measured on a scale from 1 to 7 where 1 corresponds to the absence of a symptom, and 7 to extreme impairment (Kay et al., 1987). Seven of the questions deal with so called positive symptoms such as hallucinations, delusions, suspiciousness and hostility. Another seven questions focus on negative symptoms such as blunted affect, withdrawal, passivity, and difficulty in abstract thinking or spontaneity. The remaining sixteen questions address general emotional problems including somatic concern, anxiety, depression, guilt, social avoidance, preoccupation, and the like.

The second health status instrument, the Heinrichs-Carpenter Quality of Life Scale (Heinrichs), consists of three groups of questions rated on a seven point scale (Heinrichs et al., 1984). The first seven items deal with interpersonal relationships, with scores of 0 corresponding to a complete absence (or even active avoidance) of interaction and scores of 6 representing adequate involvement. The next four questions concern the patient’s ability to operate in a work environment, with 0 corresponding to a complete lack of role functioning or occupational satisfaction and 6 corresponding to a high level of performance. Because so many of the patients in this study were unemployed the 11th question pertaining to the degree of underemployment was frequently left unanswered and was dropped from further consideration. The remaining nine questions focus on the patient’s ability to handle everyday tasks and situations with 0 corresponding to a complete lack of functionality and 6 corresponding to little or no deficit.

Finally three separate questionnaires were used to measure movement disorders, known as extrapyramidal medication side effects in the medical literature. These are among the most common physical symptoms of schizophrenia. The Abnormal Involuntary Movement Scale (AIMS) is designed to measure tardive dyskinesia (Guy, 1976). It consists of ten questions, rated on a scale
from 0-4 with 0 corresponding to no symptoms, and 4 corresponding to severe impairment. The Simpson-Angus Scale (Simpson) measures extra-pyramidal syndromes such as gait, rigidity, tremor and salivation (Simpson and Angus, 1970). It also consists of ten questions measured from 0 (no symptoms) to 4 (high severity). Finally, the Barnes Akathisia Scale (akathisia) focuses on the patients’ restlessness, evaluated both objectively by the health-care practitioner, and subjectively by the patient (Barnes, 1989). The first three questions are measured on a scale from 0-3 (normal behavior to high severity or distress) and the final question is a global clinical assessment ranging from 0 (absence of disturbance) to 5 (severe akathisia). Since all three scales were measuring different aspects of extra-pyramidal medication side effects we treated the twenty-four questions as one composite scale (side effects).

For this study we were interested in forming separate health state models for each of the three scales, PANSS, Heinrichs and side effect. However, the procedure was similar for all three scales so in this paper we have only provided the PANSS and side effects models and used the Heinrichs scale as an independent validation variable. Various demographic and validation measures were also collected. These included patient age, gender, ethnicity, treatment status (haloperidol vs. clozapine), and global quality of life scores such as the Global Assessment of Functioning (GAF), Clinical Global Impressions (CGI) and the Lehman Quality of Life Scale (QOL). Of the 423 patients in this study 297 were Caucasian, 125 were African American and 1 was Hispanic. The average age was 43 years. The patients were treated at fifteen different centers around the country, and within each center patients were randomized to receive clozapine or haloperidol.

Some initial cleaning of the data was necessary. For each patient we potentially had questionnaires at six time-points (baseline, 6 weeks, 3 months, 6 months, 9 months, and 12 months.) In fact data were available for 87% of planned follow-up observations. Rather than imputing missing values, any patient-time combination with missing data was eliminated from further study since they tended to be missing complete questionnaires rather than answers to single questions. During the study some patients responded poorly to a medication and changed to an alternative treatment. Patients who crossed over from haloperidol to clozapine were treated as members of the control group before they switched medications and members of the treatment group afterwards, and crossovers in the other direction were handled analogously. Subjects who went off all medications or switched to a third form of treatment were analyzed on an intent to treat basis, meaning that they remained in the group to which they were originally assigned.

3 Identifying dimensions of health

A critical first step in constructing a health state model for patients with refractory schizophrenia is to identify the dimensions of health that potentially differentiate among the members of this population. Like other controlled trials of treatment for schizophrenia, Rosenheck et al. (1997) concentrated on a set of univariate analyses based on summary statistics from the PANSS, Heinrichs and side effects health status instruments. Totals were taken over each of the five instruments plus the positive and negative symptoms sub-scales of the PANSS. Although this approach provides an easily interpretable overview of the data it fails to capture the inherent complexity of the problem. For example suppose we have two patients who score 50 on the PANSS, one because he has many positive symptoms and one because she is very depressed. It is not clear that these people’s problems are necessarily of equal severity nor is it clear that treatments will affect them in the same way. While the total is an obvious choice, it is not necessarily the best single number summary of an instrument. Furthermore, even if the total represents an important dimension of health it need not be the only one. In fact previous studies have shown that most of these health
status questionnaires contain information about multiple dimensions. The choice of appropriate composite scores was further complicated in our study by the fact that the refractory patients differ substantially from the general population of schizophrenics. For the above reasons it was necessary to perform a preliminary analysis of the data to determine the relevant dimensions of health. In Section 3.1 we outline the steps that were required to preprocess the data to remove systematic differences between the sites where patients were studied. In Section 3.2 we present a principal components analysis and provide interpretations of the resulting components.

3.1 Site effects

The Rosenheck et al. (1997) study was conducted at fifteen different Veterans Affairs medical centers. Every attempt was made to produce uniform assessments at each site. Unfortunately, the data still contained evidence of large site effects. Figure 1 provides an example. In Figure 1(a) we have plotted the average score at each time and site for one of the PANSS questions. There are significant differences between sites and also some evidence of interactions between site and time. This issue was not addressed in Rosenheck et al. (1997). The site effect problem is magnified when performing principal components analysis because PCA specifically targets dimensions with high variability and hence tends to create dimensions corresponding to differences between sites. If the effect is ignored and clustering is performed then site indicators becomes a highly significant predictor of cluster membership. Fortunately, since most sites contain numerous patients, it is possible to obtain accurate estimates of the site effects and standardize the data accordingly. Let \( Y_{ijkl} \) be the response to a particular question, for the \( j \)th individual from site \( i \) at time \( k \) on drug \( l \). Then we can express the observed values as a standard mixed effects model

\[
Y_{ijkl} = S_i + T_k + D_l + (ST)_{ik} + (DT)_{kl} + P_{ijl} + \epsilon_{ijkl}
\]

where \( S_i, T_k \) and \( D_l \) are, respectively, the main site, time, and drug effects, \( (ST)_{ik} \) and \( (DT)_{kl} \) are interaction terms and \( P_{ijl} \) is the individual person effect. We modeled site, time and drug as fixed
effects and person as a random effect. This choice was made primarily on the basis that people were nested within site as well as the fact that there were comparatively few measurements for each individual (between one and six) while a large number of observations were made on all other variables. Furthermore, the site and site-time interaction effects needed to be fixed since we wished to adjust for them. The model was fit to each question and the data were then standardized by subtracting off the site main effects and interaction terms, $S_i$ and $(ST)_{ij}$. Figure 1(b) illustrates the estimated site effects and Figure 1(c) the resulting transformed data. The site effects have been dramatically reduced and appear no longer to be a problem.

3.2 The principal components

After preprocessing the data, the side effects and PANSS scales were reduced to a small number of composite scores. There are numerous approaches to dimension reduction. We chose principal components (Seber, 1984) for several reasons. First, it seeks to minimize a clearly defined criteria. Second, it is an accepted and well understood procedure in both the statistics and medical communities. Finally, it allowed easy comparison with results from previous studies on this type of data, many of which have used principal components. We performed PCA separately on each of the instruments. However, the results were similar to those achieved by combining the scales. There were several reasons for doing a dimension reduction. First, by considering the weightings placed on the questions by each principal component it was possible to give a simple interpretation of each dimension. A weighting close to zero for a particular item indicates that question has little influence on the component while a large weight implies the opposite. For example, Figure 2 provides plots of the weightings for the first two side effects principal components. The first is basically an average over all questions with somewhat less weight on the Simpson-Angus scale. The second component is a contrast between the akathisia questions (positive weights) and the AIM scale (negative weights). Interpretations for the remaining side effects and PANSS components are provided in Appendix A. The second reason for doing PCA was to reduce variability in the health state formation. Clustering in four or five dimensions generally produces far more stable results than attempting to cluster twenty or thirty dimensional data.

The final choice of dimensions was made on both quantitative and qualitative grounds. We opted to include all dimensions for which the proportion of variance explained was higher than the average variance per dimension. While this is only one of numerous approaches to such problems, it produced a small number of easily interpretable dimensions. We use four dimensions for the side effects scale and five for the PANSS.
4 Clustering and Health States

In this section we produce a final health state model using the variables resulting from the principal components analysis. In the medical literature, such models are generally constructed using a factorial design. Each dimension or variable is divided into evenly spaced levels forming a grid. The hyper-rectangles correspond to health states and their Euclidean centers represent prototypical patients in those states. There are several problems with the standard grid design. First, for even a small number of variables it produces a large number of health states. More importantly, there is no a priori reason why the natural groupings of patients should follow a symmetric grid. This results in health states that are either empty or are poorly centered around their “typical” patient.

In this section we produce a superior, data driven, model using k-means clustering(Hartigan, 1975). The k-means algorithm finds the set of centroids which minimizes the sum of squared distances between each observation and its closest cluster center. This approach generally results in a far more efficient clustering with a much smaller number of health states being necessary to adequately differentiate the members the population. The cluster centroids are also much more representative of typical patients because they have been defined as the means of the data within their states.

One of the most important decisions that must be made when forming the side effects and PANSS health state models is the number of clusters to fit. This decision can be made on either statistical or medical grounds. Numerous statistical procedures have been proposed for selecting the number of clusters, but most are somewhat ad hoc. James and Sugar (2001) suggest a more rigorous approach based on $d_k$, the average distortion or squared Euclidean distance between each data point and its closest center when fitting $k$ clusters. Under fairly general conditions they prove that, for data from a $q$ dimensional multivariate normal, $d_k^{-q/2}$ will have a linear relationship to $k$.

Additionally for data from a mixture of $G$ multivariate normals they show that $d_k^{-q/2}$ will exhibit a distinct jump at $k = G$. This provides a simple statistical procedure for deciding whether distinct clusters exist and if so how many. Using this criteria there was no statistically detectable clustering of either the side effects or PANSS data. In this situation a “medical expert” should assess the optimal number of health states to use. Below we suggest three graphical tools which we found useful in aiding such a decision.

The first approach was to examine the cluster centers that were formed for varying values of $k$. Figure 3 illustrates the centers obtained when fitting a four cluster model (the first row) and a six cluster model (the second row) to the side effects data. The clusters are labeled according to their score on the first principal component. Since this component is effectively an average it can be thought of as providing an indicator of global side effects health. Low scores correspond to few problems while high scores indicate severe side effects disorders. Notice that cluster four in the six cluster model separates out people with low values in the third principal component. From Appendix A we note that this component corresponds largely to the Simpson-Angus scale which measures extra-pyramidal side effects. However, all the centers in the four cluster model have approximately the same value in the third component and fail to capture this grouping. This suggests that four clusters is too small to provide an adequate representation of the population in terms of extra-pyramidal medication side effects.

Cluster profile plots are another useful tool. An example, for the six cluster model, is given in Figure 4. For each cluster we have plotted the average score for each question among all patients falling in that health state. The questions have been centered by subtracting off their global means. Notice, for example, that cluster one has very low averages over all questions indicating patients are relatively well off in terms of side effects. Cluster six consists of patients with severe disorders.
Figure 3: Cluster centers plotted in the first three dimensions for the four and six cluster models on the side effects data. The first principal component is on the x-axis, the second component on the y-axis for the first column and the third component on the y-axis for the second column.

Note that cluster four, which had a low mean for the third principal component, has the highest average of all clusters for the Simpson-Angus scale, questions 11-20. The profile plots show clear distinctions among all the clusters which suggests that there are at least six medically interpretable states of side effects health for this population. A lack of differentiation between the clusters would indicate that the number of health states should be reduced.

A third useful approach to choosing the correct number of health states is to validate the clusters using independent variables. An example is shown in Figure 5 where we have plotted the average score on each of four independent validation variables for all patients in a given cluster. The variables are clinician’s rating of behavioral change since the beginning of the trial (CH), the Lehman quality of life scale (LHQ), clinical global impressions (CGI) and global assessment of functioning (GAF). Consider the top left plot which is divided into five rectangular regions corresponding to models fit with from four to eight clusters. Within each region the average score, for the variable CH is calculated for each cluster. The vertical lines though each mean represent least significant difference confidence intervals. If the lines for two clusters do not cross there is a statistically significant difference between the patients in those states. If, for a particular number of health states, two clusters are statistically indistinguishable over all four validation variables this suggests fewer health states may be required. For example, in the eight cluster model, the least significant difference intervals for clusters six and seven overlap for all the validation variables which suggests that eight health states is too many.

Based on these graphical tools, we decided to use a six cluster model for the side effects data and a seven cluster model for the PANSS data. Each cluster corresponds to a unique state of health. A complete set of descriptions is given in Appendix B.
5 Longitudinal Analyses

Next we use the two health state models developed in Section 4 to analyze the movement of patients between health states over time. Section 5.1 examines the effects of the different medications at each of the six time periods by calculating distributions over health states and also explores race effects. In Section 5.2 we test the data for Markovian structure. The existence of such a structure has many benefits, in particular allowing one to estimate long-run stationary distribution of patients across health states. These final stationary distributions are obtained for different subgroups of the population. Finally, in Section 5.3 we estimate financial costs and QALYs for a typical patient in each health state. We then use the stationary distributions to assess the long run effects of each medication from a financial and utility perspective.

5.1 Longitudinal analysis of health states

Using the models developed in Section 4 one can examine differences between the distributions among the health states of patients on haloperidol and those on clozapine. The data consist of measurements at six time periods: baseline and after 6 weeks, 3 months, 6 months, 9 months and 12 months of treatment. An important question is whether patients on different medications have different distributions across the health states and if so at what times these differences occur. Table 1 gives p-values for chi-squared tests of independence between medication and health state for the PANSS and side effects scales at each time. Two sets of p-values are given corresponding to chi-squared tests for all health states and tests comparing just the best and worst health states, relative to the first principal component. As one would hope, there are no significant differences between the medications at baseline. However, both scales exhibit evidence of separation at other times. The side effects scale shows differences that are both highly significant and increasing for
all time periods beyond baseline. Indeed as early as 6 weeks there are significantly more clozapine patients in state one which corresponds to the best side effects health levels. However, it appears that there is a delayed effect for the PANSS scale. It shows significant differences at 6 and 9 months when considering all health states and significant differences at all times beyond baseline when considering only best and worst states. Patients on clozapine, in the later time periods generally have a higher probability of moving to States one, two and four on the PANSS scale. These correspond to the better health states. The estimated distributions at each time point for both PANSS and side effects are shown in Tables 2 and 3.

We also checked for differences in the distribution over health states between the two main ethnic groups, Caucasians and African Americans. Surprisingly, while there was no detectable difference on the PANSS scale at baseline (p-value of 0.830), there was a highly significant difference between ethnic groups for the side effects scale (p-value of 0.003). Table 4 shows the actual distributions by ethnicity for the first and last time period on the side effects scale. Notice that Caucasians are far more likely to start in States three and six which correspond to moderate to severe akathisia problems. African Americans are far more likely to begin in State one which corresponds to the best level of health. Interestingly, these differences are no longer statistically significant beyond baseline. Race effects have not been observed in previous studies of refractory schizophrenia so this point requires further investigation. Examination of the original side effects questions reveals a weak race effect. However, the differences only become statistically significant after adjusting for site effects and constructing the health state model.

5.2 Transition probabilities and stationary distributions

In this section we estimate the long run stationary distributions of patients over health states. In order for a stationary distribution to exist we must verify that the transitions between health
### Table 1: P-values for tests of independence between medication and health states for the PANSS and side effects scales and each of the six time periods.

<table>
<thead>
<tr>
<th>Time</th>
<th>Drug</th>
<th>Health State</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>clozapine</td>
<td>0.02</td>
<td>0.224</td>
<td>0.144</td>
<td>0.06</td>
<td>0.199</td>
<td>0.194</td>
<td>0.159</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.00</td>
<td>0.227</td>
<td>0.150</td>
<td>0.05</td>
<td>0.164</td>
<td>0.264</td>
<td>0.145</td>
<td></td>
</tr>
<tr>
<td>6 Weeks</td>
<td>clozapine</td>
<td>0.162</td>
<td>0.168</td>
<td>0.139</td>
<td>0.156</td>
<td>0.162</td>
<td>0.098</td>
<td>0.116</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.092</td>
<td>0.184</td>
<td>0.154</td>
<td>0.114</td>
<td>0.154</td>
<td>0.154</td>
<td>0.149</td>
<td></td>
</tr>
<tr>
<td>3 Months</td>
<td>clozapine</td>
<td>0.220</td>
<td>0.164</td>
<td>0.158</td>
<td>0.175</td>
<td>0.102</td>
<td>0.090</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.164</td>
<td>0.155</td>
<td>0.123</td>
<td>0.155</td>
<td>0.118</td>
<td>0.186</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>6 Months</td>
<td>clozapine</td>
<td>0.232</td>
<td>0.190</td>
<td>0.214</td>
<td>0.167</td>
<td>0.060</td>
<td>0.071</td>
<td>0.065</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.201</td>
<td>0.114</td>
<td>0.157</td>
<td>0.148</td>
<td>0.122</td>
<td>0.153</td>
<td>0.105</td>
<td></td>
</tr>
<tr>
<td>9 Months</td>
<td>clozapine</td>
<td>0.262</td>
<td>0.157</td>
<td>0.134</td>
<td>0.169</td>
<td>0.105</td>
<td>0.099</td>
<td>0.076</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.187</td>
<td>0.139</td>
<td>0.187</td>
<td>0.096</td>
<td>0.110</td>
<td>0.201</td>
<td>0.081</td>
<td></td>
</tr>
<tr>
<td>12 Months</td>
<td>clozapine</td>
<td>0.235</td>
<td>0.188</td>
<td>0.153</td>
<td>0.182</td>
<td>0.094</td>
<td>0.094</td>
<td>0.053</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.203</td>
<td>0.170</td>
<td>0.208</td>
<td>0.118</td>
<td>0.080</td>
<td>0.142</td>
<td>0.080</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Distributions of patients taking clozapine or haloperidol over the seven PANSS health states at each of the six time periods.

<table>
<thead>
<tr>
<th>Time</th>
<th>Drug</th>
<th>Health State</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>clozapine</td>
<td>0.264</td>
<td>0.159</td>
<td>0.259</td>
<td>0.100</td>
<td>0.035</td>
<td>0.184</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.244</td>
<td>0.175</td>
<td>0.249</td>
<td>0.092</td>
<td>0.060</td>
<td>0.180</td>
<td></td>
</tr>
<tr>
<td>6 Weeks</td>
<td>clozapine</td>
<td>0.529</td>
<td>0.188</td>
<td>0.165</td>
<td>0.029</td>
<td>0.053</td>
<td>0.035</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.327</td>
<td>0.212</td>
<td>0.202</td>
<td>0.062</td>
<td>0.072</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>3 Months</td>
<td>clozapine</td>
<td>0.552</td>
<td>0.218</td>
<td>0.138</td>
<td>0.017</td>
<td>0.034</td>
<td>0.040</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.267</td>
<td>0.187</td>
<td>0.289</td>
<td>0.037</td>
<td>0.096</td>
<td>0.123</td>
<td></td>
</tr>
<tr>
<td>6 Months</td>
<td>clozapine</td>
<td>0.577</td>
<td>0.184</td>
<td>0.141</td>
<td>0.012</td>
<td>0.061</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.351</td>
<td>0.193</td>
<td>0.234</td>
<td>0.041</td>
<td>0.058</td>
<td>0.123</td>
<td></td>
</tr>
<tr>
<td>9 Months</td>
<td>clozapine</td>
<td>0.548</td>
<td>0.283</td>
<td>0.096</td>
<td>0.006</td>
<td>0.030</td>
<td>0.036</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.347</td>
<td>0.231</td>
<td>0.177</td>
<td>0.054</td>
<td>0.102</td>
<td>0.088</td>
<td></td>
</tr>
<tr>
<td>12 Months</td>
<td>clozapine</td>
<td>0.602</td>
<td>0.217</td>
<td>0.087</td>
<td>0.006</td>
<td>0.056</td>
<td>0.031</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.376</td>
<td>0.227</td>
<td>0.199</td>
<td>0.064</td>
<td>0.099</td>
<td>0.035</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Distributions of patients taking clozapine or haloperidol over the six side effects health states at each of the six time periods.
<table>
<thead>
<tr>
<th>Time</th>
<th>Race</th>
<th>Health State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Baseline</td>
<td>African American</td>
<td>0.339</td>
</tr>
<tr>
<td></td>
<td>Caucasian</td>
<td>0.218</td>
</tr>
<tr>
<td>12 Months</td>
<td>African American</td>
<td>0.458</td>
</tr>
<tr>
<td></td>
<td>Caucasian</td>
<td>0.511</td>
</tr>
</tbody>
</table>

Table 4: Distributions of African American and Caucasian patients over the six side effects health states at the first and last time periods.

<table>
<thead>
<tr>
<th>Scale and Drug</th>
<th>First vs Last</th>
<th>Second vs Third</th>
<th>Second vs Last</th>
<th>Third vs Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANSS clozapine</td>
<td>0.205</td>
<td>0.190</td>
<td>0.765</td>
<td>0.745</td>
</tr>
<tr>
<td>PANSS haloperidol</td>
<td>0.505</td>
<td>0.120</td>
<td>0.170</td>
<td>0.405</td>
</tr>
<tr>
<td>Side effects clozapine</td>
<td>0.000</td>
<td>0.140</td>
<td>0.520</td>
<td>0.630</td>
</tr>
<tr>
<td>Side effects haloperidol</td>
<td>0.065</td>
<td>0.450</td>
<td>0.215</td>
<td>0.175</td>
</tr>
</tbody>
</table>

Table 5: P-values when testing for a Markovian structure to the transition matrices. The large p-values indicate that the data are consistent with Markovian transitions.

states are Markovian. There are many possible approaches to testing for such a structure. We opted to compare estimated transition matrices over different time periods. The $ij$th entry of a transition matrix gives of the probability of moving to state $j$ given a patient is already in state $i$. For example, one can compute the matrix of transitions from the start to the end of the first quarter of the study and compare this to the transitions during the fourth quarter. If the structure is Markovian these two transition matrices should be equal, up to errors in the estimates. To test for differences in the transition matrices we calculated a pooled estimate of the transition probabilities, generated random data for the first quarter and the last quarter according to these probabilities, calculated the two estimated transition matrices from the new data and recorded the sum of squared component-wise deviations between these two matrices. This procedure was repeated 200 times to produce an empirical distribution of the difference between first and forth quarter under the assumption of Markovianess. The p-value for our test was then simply the fraction of these deviations greater than the observed difference between the transition matrices for the original data. The resulting p-values are shown in Table 5. We tested for a Markovian structure in each scale for both medications using several different pairs of time periods. Notice that the only significant p-value is for the side effects scale comparing the first and last quarters for patients on clozapine. The p-value for the same period for patients on haloperidol is also marginally significant. This is likely a result of the rapid improvement of side effects scores at the beginning of the study coupled with the fact that patients had just been released from hospital. However, all other p-values are insignificant, indicating that after the first quarter the data are consistent with a Markovian structure.

Since the transitions after the first quarter appear to be Markovian one can estimate final stationary distributions for each set of patients. Estimated stationary distributions for patients on clozapine and haloperidol are shown in Table 6. Since there was some evidence of a non-Markovian structure in the early time periods, the stationary distributions were calculated using estimated transition matrices based on combining all movements from one state to another between
3 months and 6 months, 6 months and 9 months and 9 months and 12 months. Patients that changed drugs during one of these time intervals were not used for that period. The stationary distributions predict clear long run differences between the medications for the side effects scales with 61% of those on clozapine ending up in the best state, one, compared to 38% of those on haloperidol. We tested the significance of these differences by randomly permuting the treatment variable, recalculating the transition matrices and stationary distributions, and calculating the sum of squared differences between the probabilities for the two stationary distributions. This procedure was repeated 1000 times. All the simulated differences were much smaller than that for the original stationary distributions, yielding a p-value of 0 for the side effects scale. There is also some weak evidence of a long run improvement for clozapine patients over haloperidol patients on the PANSS scale. More of the clozapine patients end up in the best health state, one, and fewer in the worst states, six and seven. When the above permutation procedure was applied it gave a p-value of 0.121.

5.3 Long run effects

The long run stationary distributions calculated in the previous section can be used to estimate the limiting average differences between patients on haloperidol and on clozapine. Next we measure the long run effects on the PANSS scale using three different criteria.

5.3.1 Long run differences in Heinrichs scores

First we compare the medications using Heinrichs scores. The first principal component, which essentially gives an average over all questions, is taken as a measure of a person’s overall quality of life. High values indicate better health. For each of the 423 patients, Heinrichs scores were obtained at between one and six time points, for a total of 2113 measurements. For each of these observations a corresponding health state assignment on the PANSS scale was determined and then the mean Heinrichs score was calculated for each of the seven health states. The results are shown in the first row of Table 7. There is a clear trend with State one having by far the best level of quality of life,

<table>
<thead>
<tr>
<th>Scale</th>
<th>Drug</th>
<th>State 1</th>
<th>State 2</th>
<th>State 3</th>
<th>State 4</th>
<th>State 5</th>
<th>State 6</th>
<th>State 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANSS</td>
<td>clozapine</td>
<td>0.275</td>
<td>0.201</td>
<td>0.152</td>
<td>0.168</td>
<td>0.093</td>
<td>0.074</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.239</td>
<td>0.165</td>
<td>0.181</td>
<td>0.109</td>
<td>0.080</td>
<td>0.155</td>
<td>0.071</td>
</tr>
<tr>
<td>Side effects</td>
<td>clozapine</td>
<td>0.611</td>
<td>0.227</td>
<td>0.089</td>
<td>0.014</td>
<td>0.041</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haloperidol</td>
<td>0.377</td>
<td>0.209</td>
<td>0.213</td>
<td>0.033</td>
<td>0.083</td>
<td>0.084</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Estimated stationary distributions for patients on clozapine or haloperidol.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Health State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heinrichs</td>
<td>12.27 9.95 9.57 7.72 7.58 7.38 5.39</td>
</tr>
<tr>
<td>Financial ($)</td>
<td>997 947 944 1176 1306 1516 1620</td>
</tr>
<tr>
<td>QALY</td>
<td>0.88 0.75 0.75 0.63 0.63 0.63 0.42</td>
</tr>
</tbody>
</table>

Table 7: Values for a typical patient for each of the 7 PANSS health states. The values measure Heinrichs, financial cost and QALYs.
followed by States two and three, States four, five and six have similar ratings and State seven by far the worst. By taking an average of these scores weighted according to the estimated stationary distributions from Table 6 we can calculate the long run average level of well being for a patient on each medication. Patients on clozapine will have an average Heinrichs score of 9.57 while patients on haloperidol will have an average of 9.28. Thus the improvement produced by clozapine is 0.29 greater than that produced by haloperidol, a small but clinically significant difference.

5.3.2 Long run differences in financial costs

Clozapine is a considerably more expensive drug than haloperidol. However, medication costs make up only a fraction of the financial burden to society of caring for a schizophrenic patient. Other costs include hospitalization and lost earning potential. If clozapine reduces these other costs it may in fact result in an overall reduction in expense. To provide a conclusive answer we need to calculate the cost for an average patient in each of the 7 PANSS health states and then compute long run weighted averages for the two medications. For each of the 423 patients the total societal cost, including medication, was recorded for the year of the study. However, since few patients remained in the same health state over that time period, we do not have direct measurements of health state costs. Instead we fit a multiple linear regression using the patients as observations, the number of weeks a patient spent in each health state as the seven predictor variables with no intercept term, and cost as the response variable. The fitted coefficients provide estimates of the weekly cost of maintaining a patient in a given health state. Patients with missing observations were removed. The estimated weekly cost for a patient in each health state is shown in the second row of Table 6. Standard errors of these estimates ranged from $100 to $150. Notice that, as we might expect, the better states have lower health costs. For example, State seven has an annual health cost of $84,240 compared to only $51,844 for State one. When the long run health state distributions are used to estimate the average weekly cost for patients on each medication we find that the typical expense for patients on clozapine is $1098 while that for patients on haloperidol is $1148. This result suggests that even though clozapine is more expensive, in the long run it may actually result in slightly lower overall health costs as well as higher levels of health. These results are consistent with those of previous studies (Rosenheck et al., 1997).

5.3.3 Long run differences in utility levels

Finally we consider the long run difference in utility levels or QALYs for patients on each medication. Recall that a QALY score of 0 corresponds to a state equal to or worse than death while 1 corresponds to perfect health. Once a health state model is produced one can calculate QALYs for a typical person in each state by interviewing patients and caregivers (Lenert et al., 2002). A long run goal of this study is to use the health state models developed in this article to calculate such values. To illustrate how these QALYs could be used in practice we have computed approximate values for each of the PANSS health states using the results from an earlier study (Lenert et al., 2002). The QALYs are shown in the third row of Table 6. Notice that the QALY score for State seven is less than half that of State one indicating that a patient would prefer to live half a year in State one than a full year in State seven. We can also calculate long run average QALY levels for patients on each medication using the previously obtained stationary distributions. For patients on clozapine the long run average QALY is 0.733 while for those on haloperidol it is 0.716, a small but clinically significant difference of 0.017. This effect is similar in magnitude to that found by Rosenheck et al. (1998) using a more ad hoc measure of utility.
6 Discussion and Conclusions

In this paper we have used k-means clustering to produce health state models for patients with refractory schizophrenia. In addition, we discuss three useful graphical tools to aid in the choice of health states. The k-means approach produces significantly more efficient models than a traditional factorial procedure.

Health state models have several distinct advantages over traditional approaches to analyzing clinical trials. First, they provide a parsimonious but detailed representation of the population. Second, by forming discrete states for the patients, one can study transitions over time and hence estimate the long run fraction of people in each health state. This provides information not only about the long run health of an “average” patient but also about the distribution of patients across health states. For example, it is possible for a medication to work well for the average patient but still leave a high percentage of people in “unacceptable” states. Finally, QALY values can be estimated for each health state and combined with the stationary distributions to calculate the average utility for each treatment. This allows one to make objective long term health policy decisions by balancing treatment effectiveness against societal costs on a quantitative basis.

For the patients in our study there was clear evidence of dramatic, and immediate, improvement on the side effects scale for patients on clozapine. The stationary distributions also indicated that, over the long run, patients on clozapine will generally experience fewer side effects than those on haloperidol. The differences on the PANSS scale are less dramatic, and slower to develop, but are still apparent. In particular, in the long run, a higher proportion of clozapine patients will be in the best two states, one and two, and a lower proportion in the worst states, six and seven. Finally we demonstrated how QALYs and societal costs can be calculated for an average patient in each of the PANSS health states. By combining these with the long run distributions for patients on each medication we find that clozapine produces a small but clinically significant improvement in QALYs over that caused by haloperidol. Also, even though clozapine is a more expensive medication, over the long run clozapine patients have a slightly lower financial cost than those on haloperidol.

Acknowledgments

This work was partially funded by the New England Mental Illness, Research and Education Center of the Department of Veterans Affairs and the NIMH Clinical Antipsychotic Trials of Intervention Effectiveness (J. Lieberman, principle investigator).

A Interpretation of principal components

Side effects

- PC1: The first component is roughly an average of all the side effects questions with somewhat less emphasis on the Simpson-Angus instrument than the other two scales. It measures overall the degree to which a patient experiences side effects problems. High positive scores mean severe problems.

- PC2: The second component is a contrast between the akathesia and AIMS scales. It puts positive weights on the akathesia questions and negative weights on the AIMS. A high positive score means severe akathesia problems but low tardive dyskinesia and vice versa.
• PC3: The third component separates out the Simpson-Angus scale with the exception of the akathisia and salivation questions. High negative scores mean problems with extra-pyramidal syndromes such as gait, rigidity, tremor and salivation.

• PC4: The final component focuses on the AIMS scale. It seems to be largely a contrast between facial/oral movements (which get negative scores) and the other questions, especially those about the extremities, which get positive scores. The other two scales have little weight. High positive scores mean problems with extremity movements and high negative scores mean problems with facial movements.

PANSS

• PC1: The first component is fundamentally an average although lower weights are put on some of the general emotional concerns questions such as depression and anxiety. High positive scores indicate severe problems.

• PC2: The second component is a contrast between positive and negative symptoms. High positive scores indicate problems with positive but not negative symptoms. High negative scores mean the reverse.

• PC3: The third component is a mixture of positive and negative weights on several questions. However, the questions about depression, anxiety, guilt and somatic concern are significantly more negative. High negative scores on this component indicate the patient has problems with general negative feelings.

• PC4: The fourth component measures hostility. Excitement, hostility, tension, un-cooperativeness, and poor impulse control all get higher positive weights, so high positive scores correspond to greater hostility.

• PC5: The final component corresponds to thought disturbances. High negative weights are put on questions like conceptual disorganization, problems with abstract thinking, lack of judgment and so forth.

B Interpretation of health states

Side effects

Health State one is the best and State six the worst overall, although five is also fairly bad. States three and six correspond to akathisia problems while patients in State five have problems with abnormal involuntary movements. Finally, State four corresponds to problems on the Simpson-Angus scale.

• State one: No side effects problems.
  These people are below average on all the side effects questions so they are relatively speaking in good shape. Typical average scores per question are around 0.25 to 0.5.

• State two: Mild tardive dyskinesia.
  These people have worse scores than average on the AIMS, average scores on the Simpson-Angus, and better than average scores on the akathasia questions. Questions on the AIMS average close to 1.
• State three: Mild akathisia.
  These people are average or slightly better than average on all questions except the akathisia scale where they are markedly worse than average. Typical scores on the akathisia questions range from 1 to 1.5.

• State four: Extra-pyramidal syndromes.
  These people are right on average in every area except the first eight Simpson-Angus questions on which they are significantly worse than average. The typical scores on the Simpson-Angus questions range from 1.5 to 2.

• State five: Frank tardive dyskinesia.
  These people are worse than average on most questions but only really strongly so on the AIMS where their average scores range from 1.5 all the way to 3.

• State six: Serious side effects disorders, tardive dyskinesia and akathisia.
  These people fare poorly across the board on the side effects questions, with particularly severe akathisia problems and moderately severe AIMS, although the AIMS is not as bad as State five. Typical akathisia scores average around 2.

PANSS

Health State one is the best and State seven the worst overall. States four, five, and seven correspond to negative values on the second principal component, meaning more negative symptoms than positive. The other states have the reverse pattern. States two and five correspond to high negative scores in the third principal component meaning significant problems with depression etc.

• State one: Mild symptoms.
  These people have better than average scores on all PANSS questions. Typical question scores are around 2.

• State two: Moderate symptoms with high subjective distress.
  These people are better than average on most questions except that they have higher than average levels of anxiety, depression and other general emotional disturbances.

• State three: Moderate symptoms with grandiosity.
  These people are worse than average on positive symptoms and slightly better than average on other questions. Typical scores on most of the positive symptom questions are from 3 to 4.

• State four: Severe negative predominance of symptoms with low subjective distress.
  This state is a reverse of State three. The patients are better than average on positive symptoms and depression related issues and above average on negative symptoms and some of the general questions on similar topics.

• State five: Severe negative predominance of symptoms with high subjective distress.
  These people are about average on positive symptoms, and worse than average on negative symptoms and depression. They can be considered as similar to State four with depression added in.
- State six: Severe positive predominant symptoms.
  These people have severe positive symptoms, are average on negative symptoms, and have moderately bad problems across the board on the general symptoms.

- State seven: Highly symptomatic with low subjective distress.
  These people are bad on everything except the depression/anxiety scale on which they are roughly average.

References


