Development of Patient Groupings for Effective Multi-disciplinary Treatment in Health Care Management

Pritam Mishra
ID 35390350

***Abstract*— At Present, healthcare management witnesses a growing demand for structured healthcare planning and resource management. Coordination is required especially in those cases in which hospitals have structured healthcare into specialty-oriented units, while a substantial portion of patient care is not limited to single units.From a logistic point of view, this multi-disciplinary patient care creates a tension between controlling the hospital's units, and the need for a control of the patient flow between units. A possible solution is the creation of new units in which different specialties work together for specific groups of patients[1]. Therefore, the concept of groupings of patients and treatment episodes are fundamental to better analysis of data for health services management. In this project, distance based and model based clustering techniques have been developed to identify the salient patient groups in need of multi-disciplinary care.**

I. INTRODUCTION

The most widely known patient classification developed for healthcare management purposes is diagnosis related groups (DRG)[2] which, on the basis of diagnosis and/or procedure, classifies acute hospital episodes into groups of similar expected resource use. Similarly, iso-resource groupings have been developed for other care settings, for example resource utilization groups (RUGs)[4] for long-stay care, ambulatory visit groups (AVGs)[5] and ambulatory patient groups (APGs)[6] for out-patient/ambulatory settings and so on.The driving force for these groupings has been a managerial approach to healthcare systems for logical development of program planning and budgeting to provide services (healthcare episodes) to distinct patient groups efficiently.

The scope of this project is to determine optimal subgroups out of a group of patients with several health conditions. This project could be applied to several applications including but not limited to healthcare planning, resource management and reduced healthcare cost of patients in terms of organised health insurance.

II. METHODOLOGY

This project takes advantage of the patient dataset, which contains 25 features (including 22 quality of life features and sex, age, relationship) for 377 patients from the hospital. Based on these salient features of the patients, they were clustered into several distinct sub-groups. Initially, the data has been pre-processed to remove any missing values and further treated as continuous variables (excluding age,sex,relationship for distance based clustering and GMM) for distance based clustering approaches(K-Means clustering, Partitioning around mediods clustering) and model based approaches (GMM). The next approach includes taking all 25 features into consideration as categorical variables with any missing values removed in the similar way and apply model based clustering technique (Latent Class Analysis). In these experiments, optimal clustering methodology has been selected based on statistical significance; for example elbow method,average silhouette width,Gap Statistic as a selection criteria for distance based clustering and BIC for model based clustering.

III. DATA DESCRIPTION AND PRE-PROCESSING

The data-set consists of 25 features out of which 22 are quality of life measures of 377 patients such as, short of breath, work limitations, any pain, trouble sleeping etc and three other salient features like sex, age, relationship. There are four different categories that describes 22 quality of life variables,

- $1 =$ do not have this quality at all
- $2 =$ have this quality a little
- $3 =$ have this quality quite a bit
- $4 =$ have this quality very much

The other three variables include age, sex (1=male, 2=female) and relationship (1=single, 2=married, 3=divorced, 4=widowed). The dataset also includes missing values coded as -9 which were converted into NA values initially further removed from the dataset.

IV. DISTANCE BASED CLUSTERING

In this project, K-means clustering and Partitioning around mediods clustering techniques have been implemented as part of distance based clustering approaches. In this approach, initially K-Means clustering PAM clustering have been applied on 22 quality of life variables for 4 different cluster centers. The procedure have been repeated for at least 200 times with different starting values as their means or random starting point. The results of these clustering algorithms have been discussed in the results section of this article. Further the cluster means as trajectories over each feature for K-Means PAM have been compared for more insights. In this section, we will focus on the selection criteria for optimal number of clusters and validate our initial assumption of using 4 cluster centers for these distance based clustering approaches. We will compare and discuss several methodologies for choosing optimal number of clusters for our dataset.

A. The Elbow Method

The Elbow method looks at the total within-cluster sum of square (WSS) or intra-cluster variation as a function of the number of clusters. The optimal number of clusters are selected so that adding another cluster doesn't improve the total WSS. The location of a bend knee in the plot is generally considered as an indicator of the appropriate number of clusters.

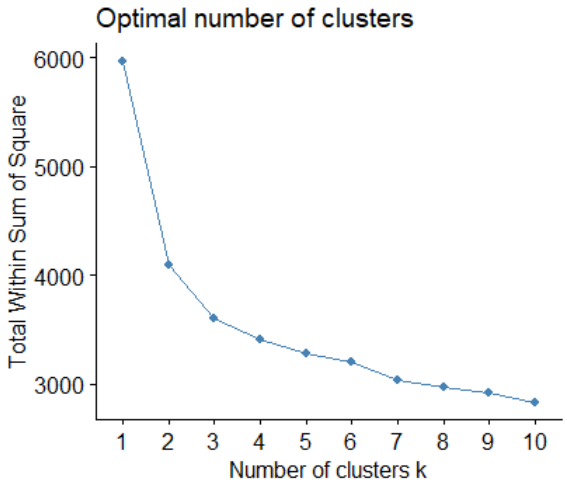

Figure 1: Elbow Method

We could analyse from the figure 1 there is a sharp decrease in WSS till k=3 and beyond the third cluster no significant change in WSS can be observed. However, we could also observe a relatively sharp decrease in WSS from cluster 6 to cluster 7 and beyond cluster 7 there's minimal change in WSS. Therefore we can not conclusively determine the optimal number of clusters with this method for our dataset.

B. The Silhouette Method

Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k.

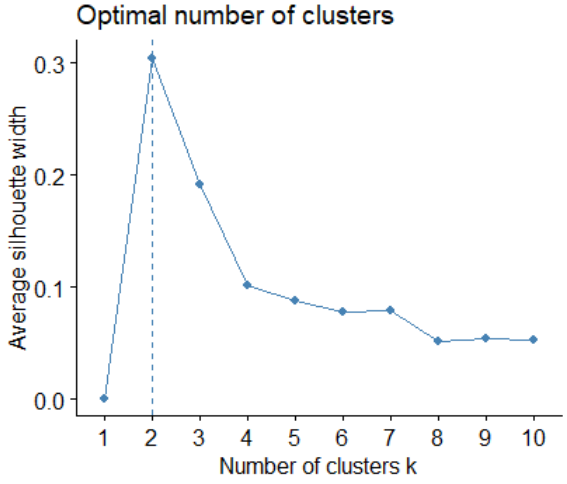

Figure 2: The Silhouette Method

We can observe from Figure 2, for k=2 it maximises the average silhouette for all possible values of k. Therefore, the optimal number of clusters in this case could be predicted as two with this method.

C. Gap Statistic

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The plot for gap statistic shows the statistics by number of clusters (k) with standard errors drawn with vertical segments. The optimal value of k is usually highlighted as the vertical dashed blue line in the plot.

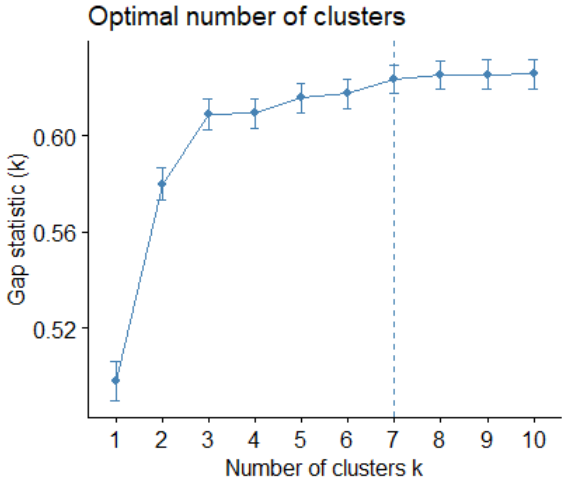

Figure 3: Gap Statistic

We can observe the optimal number of clusters obtained in this method is 7 which was also observed at the time of analysis the elbow method in figure 1. We could investigate further through a dendrogram to conclude our experiments.

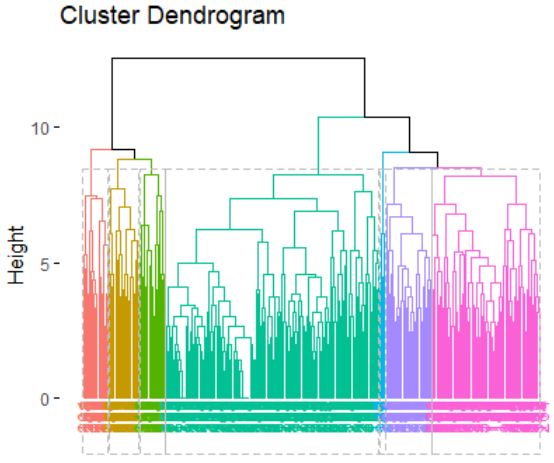

Figure 4: Dendrogram

From Figure 4, we can analyze if we choose 2 to be optimal number of clusters in this case, that generalizes the subgroups a lot and might be considered as a naive approach. Among all of the methodologies implemented in this project

to determine optimal number of clusters, Gap Statistic was the most sophisticated statistical approach. Therefore, we can conclude our initial impression of choosing 4 clusters for distance based clustering might not be an optimal solution, rather k=7 or k=2 is much more statistically significant for clustering the given dataset.

V. Model Based Clustering

Gaussian Mixture Model and Latent Class Analysis have been implemented in this project as part of the model based clustering approach. While, GMM has been applied to 22 life variables treated as continuous data, all 25 features have been used with LCA as categorical variables. Bayesian Information Criterion or BIC has been identified as the model selection criteria for both of these clustering techniques.

A. Bayesian Information Criterion

This criterion provides an estimation on how good a Gaussian Mixture Model (GMM) or Latent Class Analysis (LCA) model is in terms of predicting on a given data. The lower value of the BIC is, the better is the model to accurately predict the data, and by extension, the true, unknown, distribution. In order to avoid over-fitting, this technique penalizes models with big number of clusters.

VI. Result

Based on our initial assumption, K-means clustering technique has been implemented on the given dataset for 4 different cluster centers and at least 200 different random starts with maximum iteration parameter being 100. The result for this distance based clustering technique can be shown below with four distinct clusters as different subgroups,

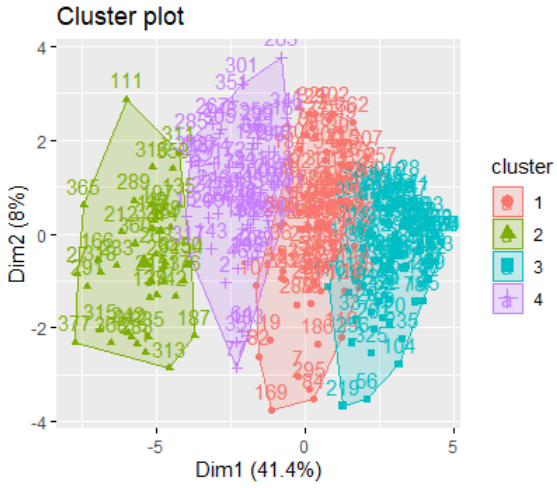

Figure 5: K-Means Clustering with k=4

A. cluster characteristics

From Figure 5, we can observe that dimension 1 and dimension 2 has total explained variance of almost 50 percent. Therefore, the data points are not accurately explained in a two dimensional figure for maximum explained variance of the dataset. Based on the above clustering experiment, the four cluster sizes in which the patients have been divided are 96, 42, 100 and 61 respectively. The within cluster sum of squares for these clusters have been observed to be 1098.0625, 735.4048, 608.4200 and 962.3934 respectively. So the percentage of data within the four clusters can be determined as 32.1, 14, 33.4 and 20.4 percent respectively. Therefore the average within sum of squares for each cluster can be computed as 11.43, 17.5, 6.08 and 9.97 respectively. We can conclude, the average variability of the observations in cluster 3 is lowest while for cluster 2 highest average variance of data can be observed.

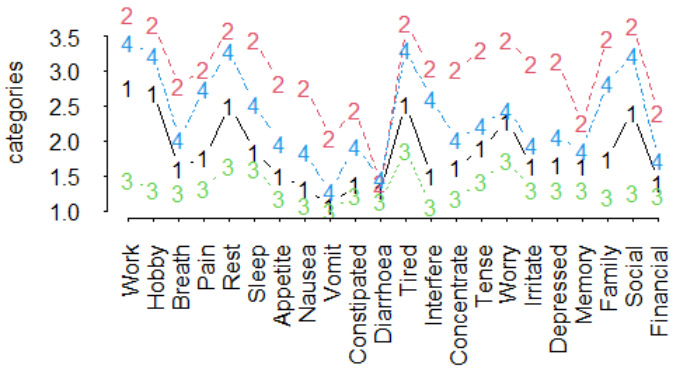

Figure 6: Cluster means as trajectories for K-Means

From figure 6, we can observe the cluster means as the dotted lines with it's cluster numbers while in the x axis all 22 features explain how likely those means are related to those health conditions with 1-4 as the severity of likelihood present in y axis. From the discussion above, we observed very less variance in cluster 3 and now we see the group of people belongs to this cluster are very less likely to have these health condition as shown in the figure. We can conclude this with confidence since the variance or average within sum of squares appears to comparatively low than other clusters. We could also observe it is quite unlikely for most of the patients to be diagnosed with diarrhoea. Also, we see patients belong to cluster 2 are more likely to have these health conditions followed by cluster 4 and cluster 1.

Figure 7: PAM Clustering with k=4

In the other distance based approach, PAM clustering technique has been implemented with 4 cluster centers.

B. Cluster Characteristics and Comparison With K-Means

We have observed 4 different clusters using PAM clustering with cluster sizes 49, 91, 83 and 76 respectively and average distances from the center being 16.510204, 12.923077, 9.730159 and 6.729167 respectively. This gives us insight that observations belong to cluster 4 are more centralized while variance is much higher for cluster 1. The discernible difference between PAM clusters K-Means is that PAM clusters appear to be overlapped with each other with cluster 2, 3 and 4 having isolation value greater than 2.

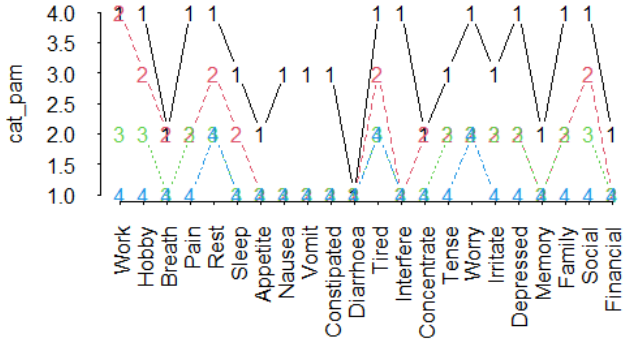

Figure 8: Cluster means as trajectories for PAM

From Figure 8, we observe some noticeable changes in mean trajectories between PAM and K-Means clustering. While, they both appear similar for feature diarrhoea, only cluster 1 seems more likely to have health conditions such as appetite,nausea, vomit. We can also observe some irregular patterns for cluster 1 and cluster 2 for PAM while for K-Means the pattern was very similar for all the clusters.

For model based clustering techniques, MclustBIC has been applied on the given data to select the best Gaussian Mixture model based on BIC as the selection criteria. Based on this experiment, the top 3 models with BIC has been observed as VEE,2 -14850.82; VEI,3 -14878.65 and VII,3 -14986.74. Out of these models, the model with minimum BIC value has been selected VEE (ellipsoidal, equal shape and orientation) model with 2 components.

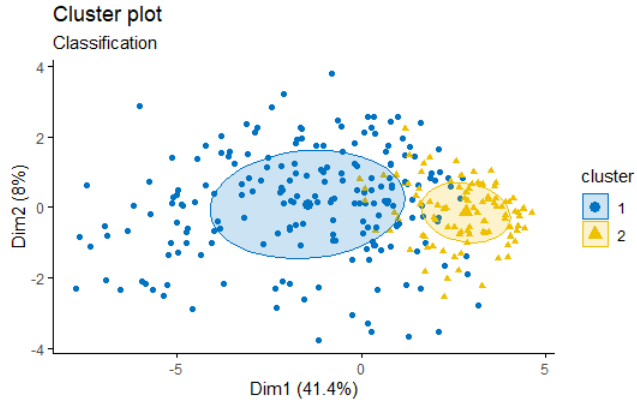

Figure 9: Gaussian Mixture Model with k=4

As we can observe from the figure 9 as well, the GMM model obtains two clusters containing 197 and 102 observations respectively with mixing probabilities 66.78 percent and 33.21 percent respectively. The likelihood and degrees of freedom for this model observed to be -6573.195 and 299 respectively.

Let us discuss whether GMM should be appropriate for our given dataset. We realize that, the given dataset consists of categorical variables which has been treated as continuous for both distance based clustering approaches and GMM. Further, GMM is ideal for data points that have maultiple mean and standard deviation following a mixture of Gaussian distribution. However, in this case categorical features could not take advantage of density based soft clustering in GMM. Therefore it is not appropriate for this dataset.

Latent Class Analysis is another very sophisticated clustering approaches where categorical variables can be used to divide the observations into subgroups. In our first experiment, number of latent classes considered for this approach has been assumed to be 5. After the results of the first experiment based on the AIC and BIC values along with log-likelihood, the optimal number of latent classes has been selected as 3.

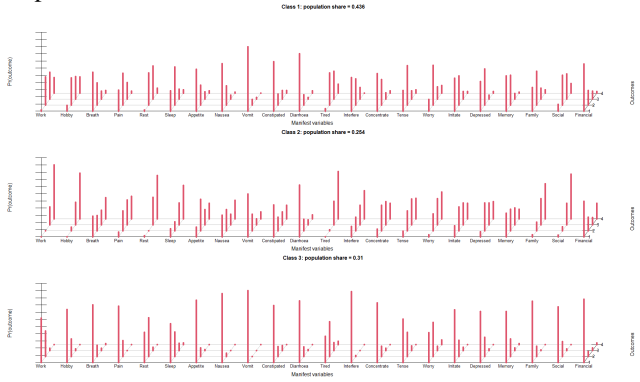

Figure 10: Gaussian Mixture Model with k=3

From Figure 10, we can observe 3 latent classes with population share 0.436, 0.254 and 0.31 respectively. Class 3 from the observation is least likely to have the specified health conditions on the dataset while class 1 is moderately likely to have health conditions for all the features. Class 2 is moderately likely to have these health conditions except some features where they are highly likely like hobby, tired, rest, family etc. If sex is included as a covariate it is likely for females to be included in class 3, more details can be found on appendix.

VII. CONCLUSIONS

In this project, several clustering techniques have been compared and contrasted with each other to obtain optimal solution from separate approaches. The features in the dataset provided in this project were mostly categorical and therefore distance based clustering techniques and also GMM model based techniques were not very useful in determining patterns in different clusters as we already discussed the primary 2 components does not have high explained variance

for these clustering techniques. Further details can be found in the appendix section. Thus LCA was very useful in determining subgroups in this project.

APPENDIX

Based on the model selection as discussed in the earlier section, we have taken 2 as the optimal number of clusters considering average silhouette width, the result of which can be demonstrated as follows,

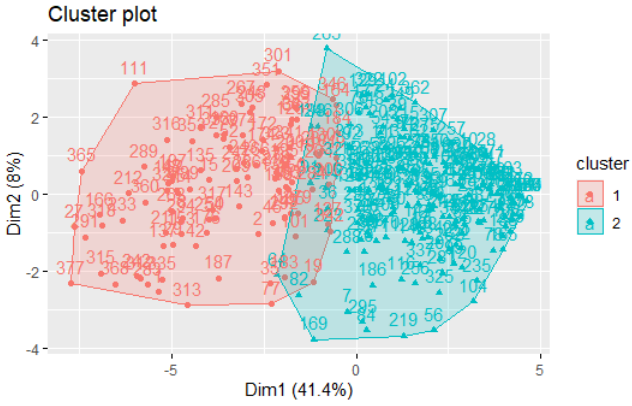

Figure 11:PAM Clustering with k=2

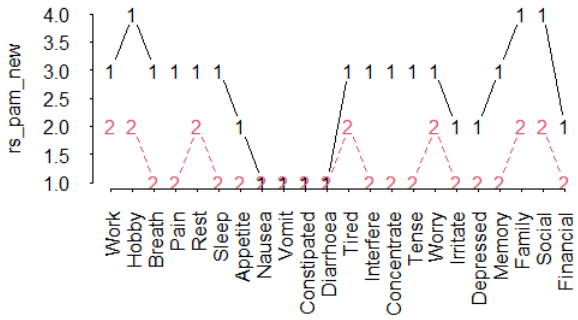

Figure 12: Cluster means as trajectories for PAM

```
Estimated class population shares
 0.3114 0.2455 0.4431

Predicted class memberships (by modal posterior prob.)
 0.3116 0.2466 0.4418

=========================================================
Fit for 3 latent classes:
=========================================================
2 / 1
             Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -0.16017     0.40757   -0.393     0.695
Sex2           -0.12353     0.59947   -0.206     0.837
=========================================================
3 / 1
             Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)     0.17110     0.32013    0.534     0.594
Sex2            0.27068     0.41800    0.648     0.519
=========================================================
number of observations: 292
number of estimated parameters: 202
residual degrees of freedom: 90
maximum log-likelihood: -6178.991
```

Figure 11: Gaussian Mixture Model with k=3 with sex covariates

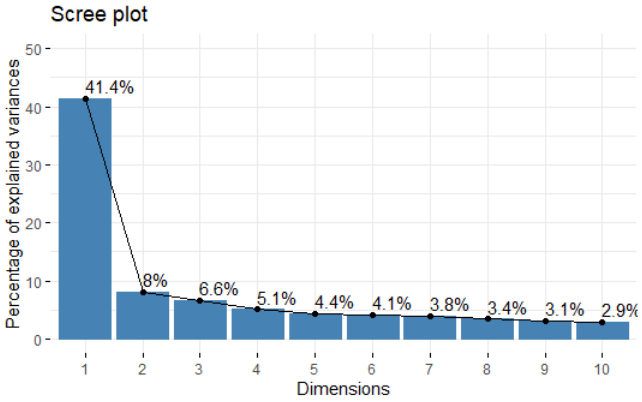

Figure 12: PCA Explained variance

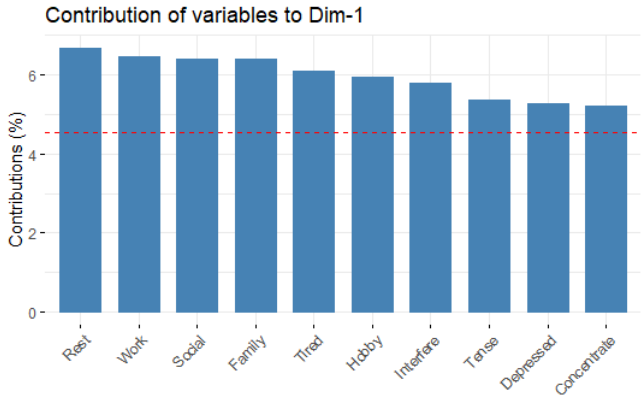

Figure 13: Feature Contribution for Component 1

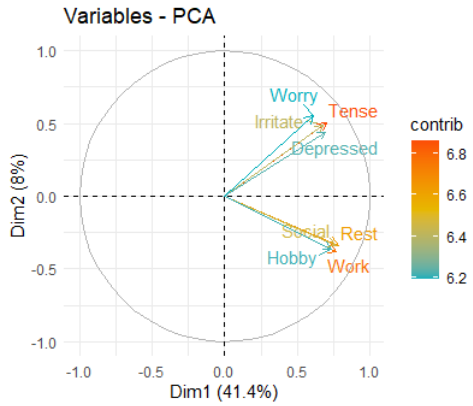

Figure 14: Feature Contribution for Components 1 and 2

REFERENCES

[1] "Logistic-based patient grouping for multi-disciplinary treatment." Maruşter L1, Weijters T, de Vries G, van den Bosch A, Daelemans W.

[2] "Case mix definition by diagnosis related groups. Med Care 1980" Fetter RB, Shin Y, Freeman JL, et al.

[3] "The development of patient groupings for more effective management of health care" HUGH SANDERSON, LEONIE MOUNTNEY .

[4] " Refining a casemix measure for nursing homes: resource utilization groups (RUG-lll).Med Care 1994,32:668-85." Fries BE, Schneider DP, Foley WJ, et al.

[5] " The new ICD-9-CM ambulatory visit groups classification scheme: definitions manual. New Haven, CT: Yale University, "Schneider KC, Lichtenstein JL, Fetter RB, et al.

[6] ". Ambulatory patient groups: definitions manual. Wallingford, CT: 3M Health Information Systems, 1991." Averill RF, Goldfield Nl, McGuire TE, et al.