CONF-9708110--

LA-UR- 97-3296

Approved for public release; distribution is unlimited.

Title: COMPARING CANDIDATE HOSPITAL REPORT CARDS

Author(s): Tom L. Burr, Reid D. Rivenburgh, James C. Scovel, & James M. White

Submitted to: Joint Statistical Meetings, Sponsored by the American Statistical Association, Anaheim, California, August 10-14, 1997

19980406 097

Los Alamos
NATIONAL LABORATORY

DISCLAIMER

COMPARING CANDIDATE HOSPITAL REPORT CARDS

Tom L. Burr, Reid R. Rivenburgh, James C. Scovel, and James M. White
Los Alamos National Laboratory
Tom L. Burr, MS E541, Los Alamos NM, 87545, tburr@lanl.gov

1. Statement of the Problem

Our data is from 1082 acute care hospitals in 10 states representing 30 patient subgroups. Each subgroup is at risk for an undesired outcome like surgical complications, Cesarean section, or death. For each subgroup, for each hospital, there are five variables: number of cases (n), percent of cases with the undesired outcome (p), average length of stay (los), average cost (cost), and average charge (charge). For a complete description, see the 1997 ASA Data Exposition information on the hospital report cards data.

We present graphical and analytical methods that focus on multivariate outlier detection applied to the hospital report cards data. No two methods agree which hospitals are unusually good or bad, so we also present ways to compare the agreement between two methods. We identify factors that have a significant impact on the scoring such as: (1) whether hospitals are only compared to their peer group, and how the peer group is defined, and (2) multivariate outlier detection issues such as outlier masking and the effect of using robust estimation of the covariance matrices or of using ranks rather than raw data. Several report cards are proposed and each one leads to a different hospital ranking. We give qualitative conclusions concerning (1) which hospitals are unusually good or bad according to all report cards, and (2) which report card is recommended.

2. Description of the Candidate Report Cards

All of our report cards (methods) give a score to each of the 30 patient subgroups, and we use several weighting options to combine scores over the 30 subgroups to arrive at a single hospital score (note: patients in a given subgroup might care only about how well a hospital performs for that subgroup, but our focus here is to give a report card to the hospital overall so we combine scores over subgroups).

Patients want small p scores. Payers want acceptably small p with low los, cost, and charge. All of our candidate report cards attempt to judge how pleased both the patient and the payer would be with a given hospital, so they use both the p scores and the "cost"- related scores. Because one of our methods relies on having a full-rank covariance matrix among the variables, we report results here for the case of using three of the four quality measure variables: p, los, and charge.

The following scores are for each patient subgroup for each hospital.

Score 1: The Mahalanobis distance from 0. One common outlier detection method uses the Mahalanobis distance MD_i of each point $\mathbf{x}_i$ from an estimate of the mean $T(\mathbf{X})$. The usual definition is $MD_i^2 = (\mathbf{x_i} - \mathbf{T(X)})\mathbf{C(X)}^{-\mathbf{1}}(\mathbf{x_i} - \mathbf{T(X)})^{\mathbf{t}}$, where $C(\mathbf{X})$ is an **estimate** of the covariance matrix of the data matrix $\mathbf{X}$. We use both robust and nonrobust $C(\mathbf{X})$, and we set $T(\mathbf{X}) = \mathbf{0}$ **instead of the sample mean** because all variables are nonnegative and we want to assign good scores to small distances from 0. We use the minimum volume ellipsoid method (cov.mve in S-Plus) as our robust $C(\mathbf{X})$ [1].

Score 2: Sum of the scaled variables. Scale each variable to have zero mean and unit variance, then sum them. Alternatively, we also divide each variable by a robust (using the mean absolute deviation to estimate the population standard deviation, σ) estimate of its σ.

Score 3: The FRED score, $F = \sum_{i=1}^{d} I_i$, where the indicator variables, I_i, are defined as follows. Let T_i be quantile p for the i-th variable x_i, so that $100p$ percent of hospitals have $x_i > T_i$. Let $I_i = 1$ if $x_i > T_i$ and 0 otherwise. This score ranges from 0 to 3 for each subgroup. We report peer-group results for $p = .01, .05$, and $.1$. If the features were independent then $F \sim$ Binomial$(3, p)$. We can use the departure of the FRED scores from the binomial distribution as one measure of dependence of our variables.

Score 4: Average variable ranking. For each variable, rank the 1082 cases. The final rank is according to the average rank over all variables for each hospital.

Scores 1 and 2 have both robust and non-robust (with respect to outliers) versions.

Scores 3 and 4 are inherently robust to outliers.

We also considered the effect of only comparing a

hospital to its peer group. There are many possible ways to define peer groups. We present results for the following way. We searched for which of the nine peer group categories (number of beds, geographic region, etc.) exhibited the largest group-to-group variation (compared to within-group variation) in p. The two peer group categories that have significantly largest between-group to within-group variation are the percent Medicare and geographical area, which each had five categories. To reduce the total number of categories, we reduced the percent Medicare to low (A or B) and high (C, D, or E), and we retained the 5 geographical regions so our single peer group variable had 10 groups with the following number of hospitals in groups 1 to 10: 81, 238, 45, 211, 54, 104, 29, 28, 143, 149 (low Medicare: northeast, southeast, central, northwest, southwest, and then high Medicare with same regions). We report peer group results here only for Score 2. Score 1 was difficult to apply because the 3-by-3 covariance matrix was singular within peer groups for most of the 30 patient subgroups.

For each scoring method we must somehow combine scores over subgroups. Not all hospitals had $n > 0$ for all patient subgroups. The number of $n > 0$ subgroups ranged from 18 to 30 with an average of about 28. It is reasonable then to simply average the scores over the subgroups with $n > 0$, but we have choices for how to average. We use three weighted averaging methods: (1) w1: weights are the average p scores for that subgroup (subgroups having high p are weighted more heavily), (2) w2: weights are the average charge scores for that subgroup, (3) w3: weights are n for that subgroup, and (4) w4: weights = 1 (unweighted). Weights 1 and 2 vary with subgroup but are the same for each hospital. Weight 3 depends on both the subgroup and the hospital.

3. Multivariate Outlier Detection

Our main goal is to identify unusually good or bad hospitals and our second goal is to compare candidate report cards for hospitals. Our approaches can all be viewed as multivariate outlier detection methods [2]. There are too many notions of what it means to be an outlier to review here. Also, "the complexity of the multivariate case suggests it would be futile to search for a truly omnibus outlier detection procedure" [3]. Practical suggestions for detecting multivariate outliers therefore include

(a) try several methods and compare them [3],

(b) reduce dimensionality somehow, such as using principle components [2], and

(c) define a region where the outliers of interest should lie [2].

In our analyses we apply several methods, all of which reduce dimensionality because a scalar-valued score results. Also, all of our methods look for either "large" or "small" values of some of the features to take advantage of an inherent preferred direction for all three variables.

We cannot review the subject of multivariate outlier detection here, but we adopt the above practical suggestions and consider three issues:

(1) outlier masking (example: the presence of one outlier can make it difficult to detect a second outlier),

(2) the impact of having an *a priori* direction to search for outliers, and

(3) how to compare our outlier detection methods.

Our comparison methods to compare our ranking methods focus on the top n_{top} and bottom n_{bottom} hospitals. We define two distance measures to compare two ranking methods. Both distances lie between 0 and 1 with 0 being perfect agreement between two methods.

We present results here for $n_{top} = 10$ and 20 and the same for n_{bottom}. Distance 1 is 1 minus the percent of cases that are in the n_{top} by both methods. The second distance finds the providers in the n_{top} ranks by method 1, and then records their ranks by method 2. The Spearman correlation, ρ_1, is computed between those two sets of ranks, and then the viewpoint is reversed; we compute ρ_2 and then symmetrize the comparison method by defining $\rho = \frac{\rho_1+\rho_2}{2}$. Distance 2 is $.5 \times (1 - \rho)$ which ranges from 0 to 1, where 0 (1) corresponds to an average Spearman correlation (of the two viewpoints) of 1 (-1).

4. Data Analysis Results

In Fig. 1d we plot the overall hospital scores versus hospital number for Score 2 using n as the weighting over subgroups method. We also plot in Figs. 1a-1c the overall hospital scores using only (respectively) p, charge, and los. This allows us to see that, for example, Hospital 805 is an outlier according to Score 2 because it has large p and large los scores (though its charge score is not large).

To show the effect of our four weighting methods, in Fig. 2 we show Score 1 (MD) with each of the four weighting methods. In Fig. 3 we show results from several scoring methods, each combined across

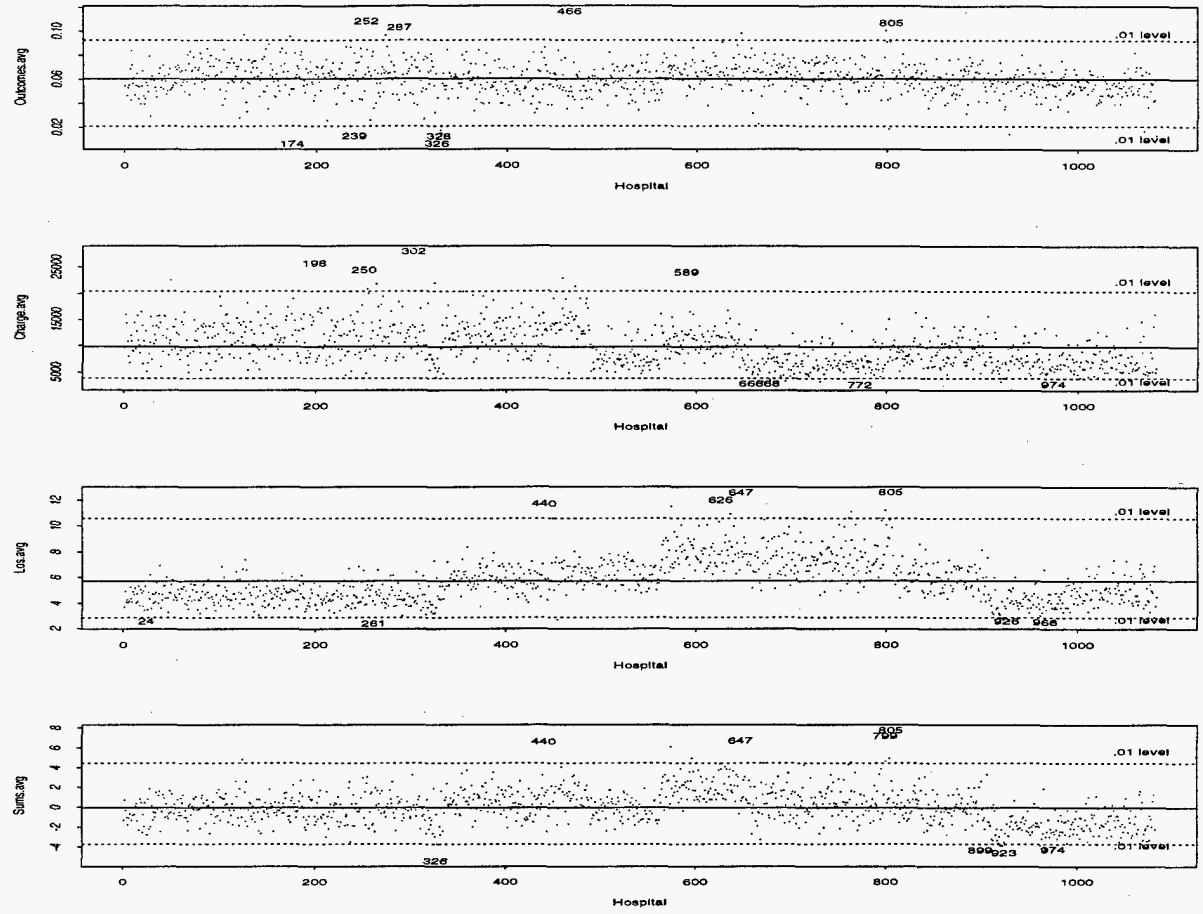


Figure 1: Outcomes, charges, los, and combination (Score 2).

patient subgroups using n as weights.

To visually compare our methods agreement on the top 10 and 20 (and bottom 10 and 20) hospitals, we use a minimal spanning tree (MST) [4] and hierarchical clustering. The MST is a more complete comparison of any two methods because it compares their similarity not just to each other but also to all other methods. For more details on this approach with another data set, see [5].

We present results (Fig. 4) here for $n_{top} = 10$ and 20 (we also consider $n_{bottom} = 10$ and 20 when we compare methods) for 30 methods: mds.w1 to mds.w4, rmds.w1 to rmds.w4, sums.w1 to sums.w4, rsums.w1 to rsums.w4, sums.pg.w1 to sums.pg.w4, fred.01.w4, fred.05.w4, fred.10.w4, spearman.w1, spearman.w2, spearman.w3, spearman.w4, outcomes.scores, charge.scores, and los.scores (the last three, methods 28, 29, 30 in Fig. 4, are weighted by n).

5. Summary

We selected p, los, and charge as the three primary variables and presented several hospital report cards. There is no consistent agreement on which hospitals are best, but there is nearly consistent agreement that two hospitals (449 and 805 in FL and NY, respectively) are noticeably worse than the other 1080 hospitals. Hospital 805 has high average p and high average los, while hospital 449 is modestly high on all three variables. We can affect our conclusions by comparing hospitals only to their subjectively defined peer groups. We presented one way to define peer group that used geographic region and percent Medicare to define the group which resulted in 10 groups (high and low Medicare usage and five geographic regions). A unique feature of our presentation is the way we graphically compared candidate report cards using a two-dimensional minimal spanning tree for a custom-defined measure of agreement

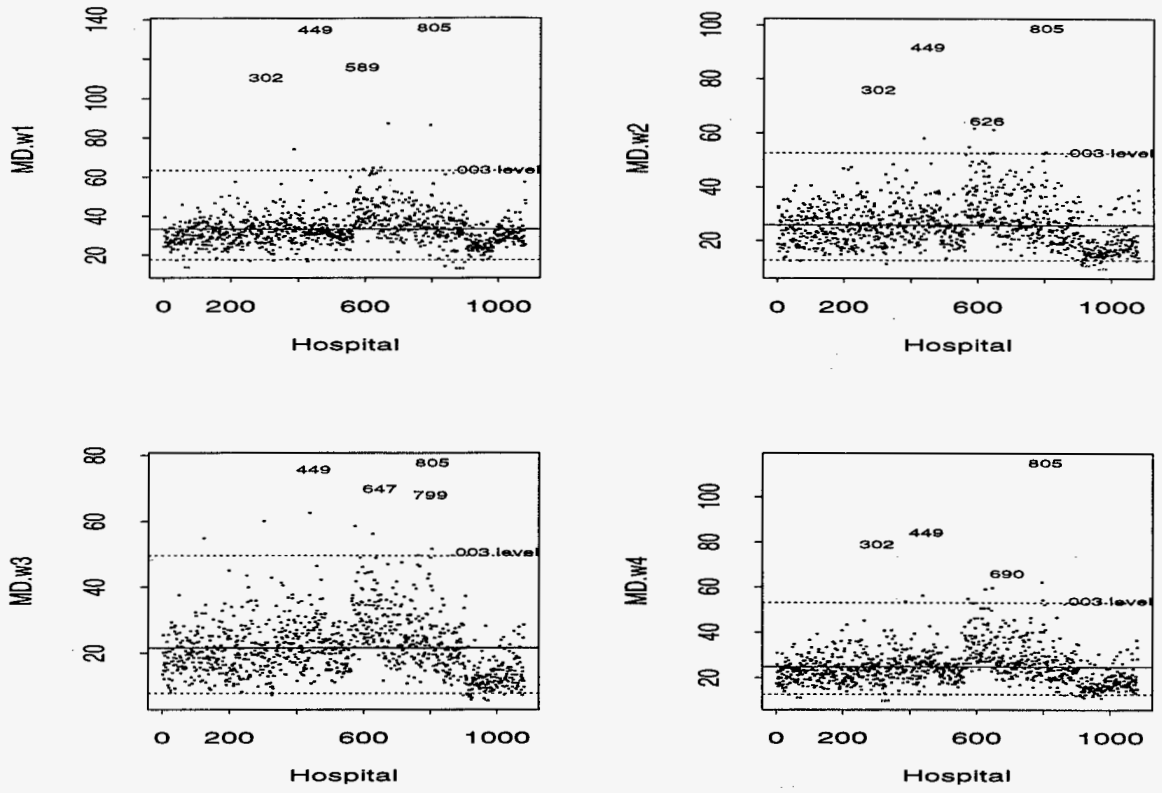


Figure 2: MD (Score 1) with four weighting methods.

between two scoring methods. If we must choose only one report card, we suggest averaging the nine presented in Fig. 3 (all weighted by n, the number of cases) plus the same nine weighted by the difficulty of the outcome as defined by the average p. An ordinary average of those 18 scores results in the following **top 11** hospitals (from best to worst):

899, 326, 68, 864, 840, 883, 328, 974, 918, 966, 922, and the following **bottom 11** hospitals from (best to worst):

804, 626, 302, 574, 799, 589, 440, 647, 449, 805.

Note that hospitals 449 and 805 are the two worst hospitals.

As we can see in Fig. 1, there is a significant gap between approximately the worst five hospitals and the "average hospitals," and similarly for the top five hospitals. In particular, hospitals 899 and 326 have significantly better average scores (over all methods considered) than the other hospitals, and hospitals 805, 449, 647, 440, 589, 799, 574, and 302 have significantly worse average scores than the other hospitals.

6. References

[1] Rousseeuw, Peter J. and Zomeren, Bert C. (1990), "Unmasking Multivariate Outliers and Leverage Points," Journal of the American Statistical Association, 85, No. 411, 633-651.

[2] Beckman, R. J, and Cook, R. D. (1983), "Outliers ...," Technometrics, V 25, No. 2, 119-149.

[3] Gnanadesikan, R. and Kettenring, J. R. (1972), "Robust Estimates, Residuals, and Outlier Detection With Multiresponse Data," Biometrics, 28, 81-124.

[4] Friedman, J. H. and Rafsky, L. C. (1981), "Graphics for the Multivariate Two-Sample Problem," Journal of the American Statistical Association 76, 277-287.

[5] Burr, T., Hale, C. Kantor, M., Rivenburgh, R., Siciliano, C., Scovel, C., Weiss, D., and White, J., "Fraud Detection in Medicare Claims: A Multivariate Outlier Detection Approach," LA-UR-97-1142, Los Alamos National Laboratory Report, May 1997.

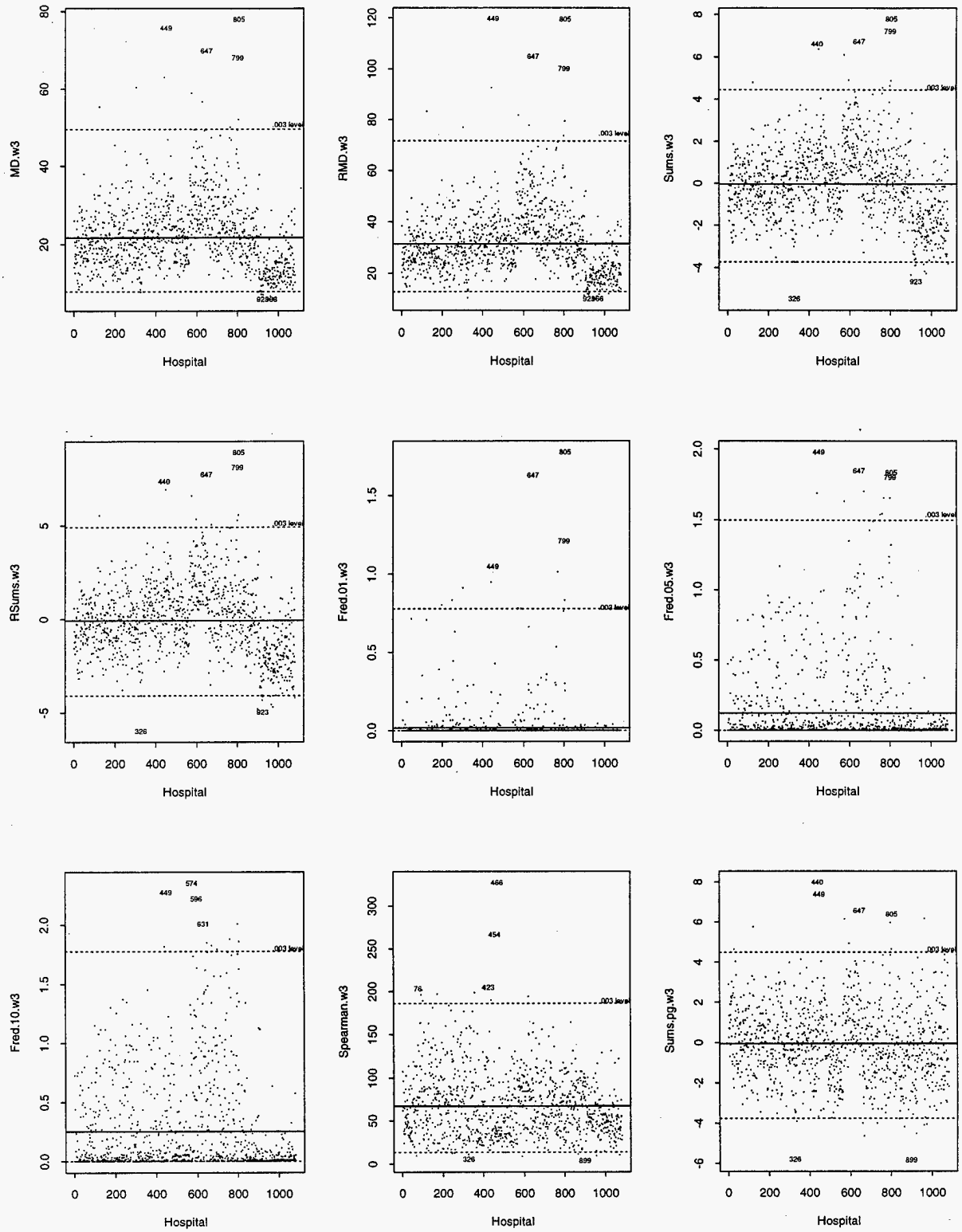
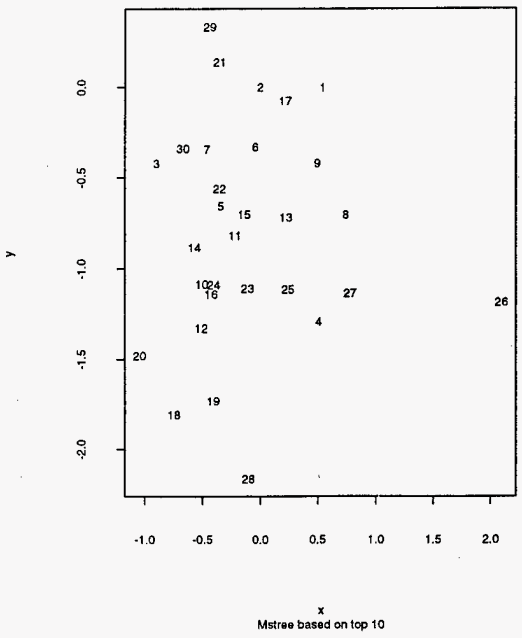
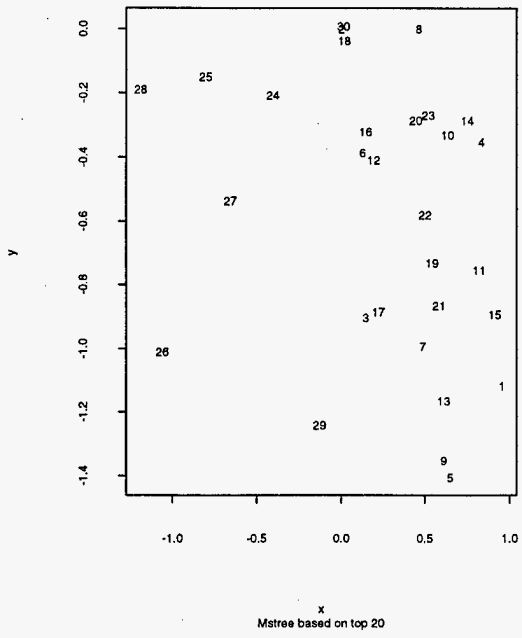
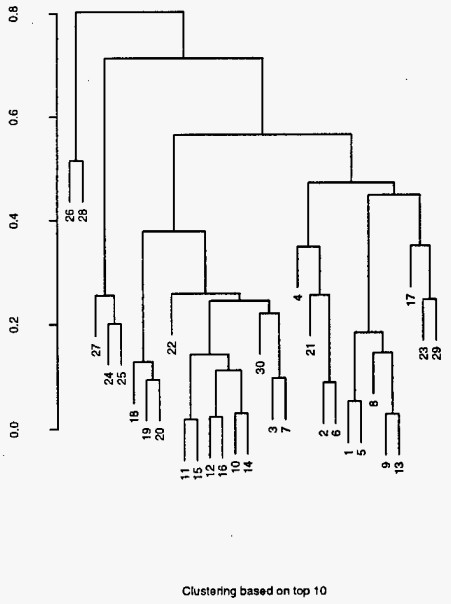


Figure 3: Nine scoring methods, each weighted by n.







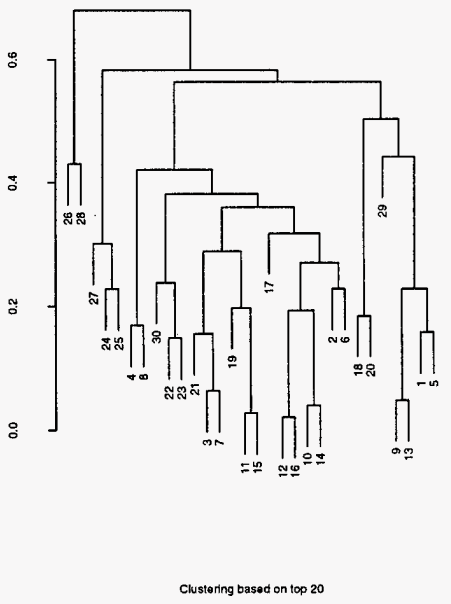


Figure 4: Comparing 30 methods on top 10 and 20 hospitals.

Report Number (14) LA-UR--97-3296
CONF-9708110--

Publ. Date (11) 199710
Sponsor Code (18) DOD, XF
UC Category (19) UC-000, DOE/ER

DOE