Using Resampling Techniques in PROC MULTTEST to Evaluate Surgeon Specific Results Following Coronary Artery Bypass Graft (CABG) Surgery

Gregory L. Pearce, Memorial Mission Hospital, Asheville, North Carolina
Peter H. Westfall, Texas Tech University, Lubbock, Texas

ABSTRACT

Six surgeons perform over 800 CABG surgeries at Memorial Mission Hospital each year. Surgeon specific results are evaluated quarterly. Each surgeon is compared to the remainder of the group for seven adverse events with the intent of identifying continuous quality improvement (CQI) opportunities for clinical practice. In order to drive out fear in the CQI process, the probability of declaring a false significance must be controlled. Adjustment techniques to address the multiple comparisons problem are available (e.g. Bonferroni, Sidak) but may prove too conservative to identify CQI opportunities for the surgeons. Therefore, a method that balances the risk of falsely declaring a significant result with the ability to detect clinically important differences is desirable. We have employed PROC MULTTEST to resample the data to make permutation adjustments. This method approximates the distribution of the minimum p-value of all tests and this distribution is then used to adjust individual raw p-values. The Cochran-Armitage linear trend test is used to make the surgeon specific comparisons.

INTRODUCTION

Health care has entered into the evidence based decision making era. In no field is that more evident than cardiac surgery as evidenced by the publication of surgeon "report cards" in New York and Pennsylvania newspapers. (Green and Wintfeld, 1995). More than 750 institutions now participate voluntarily in the Society of Thoracic Surgeons National Cardiac Surgery Database (STS Database). Memorial Mission Hospital (MMH), an affiliate of the Mission - St. Joseph's Health System, and Asheville Cardiovascular and Thoracic Surgeons, P.A. have participated in the STS Database since 1992. Six surgeons perform over 1,100 cardiac surgical procedures each year at MMH the majority (> 800) of which are primary (first incidence) CABG surgeries.

The purpose for our participation in the STS Database is to use the database as a tool in the continuous quality improvement (CQI) of clinical practice. Hospital death (HDEATH), perioperative myocardial infarction (MI_EKG), reoperation for bleeding (RFB), surgical wound infection (INFECT), cerebrovascular accident (NEURO), pulmonary complications (PULM) and renal failure (RENAL) are examined on a quarterly basis. Each of these adverse events is measured as a percentage of the total surgical procedures performed (individually and in total). Quarterly evaluations are made at the institutional level and the individual level. At the institution level, national reports are used in bench marking and internal comparisons are made longitudinally. At the individual level, each of the preceding adverse events is examined on a surgeon specific basis. These examinations consist of testing the multiple hypotheses that each individual surgeon's outcomes for each adverse event do not differ significantly from the remainder of the group.

A critical question to this CQI process is how to preserve the Type I error protection rate in light of the multiple comparisons that are being made. With six tests (one for each surgeon) being made on each of the seven adverse events, 42 comparisons are being made. The multiple comparisons issue results in an increased likelihood of declaring a false significance. Under independent and uniformly distributed p-values, the probability that at least one of the 42 comparisons is significant is 88.4%. This high probability of spuriously identifying a surgeon with a significantly higher adverse event rate will lead to fear and mistrust of the CQI process. Therefore, it is imperative that the multiplicity problem be addressed. Conventional methods for preserving the family-wise Type I error rate are available (e.g. Bonferroni and Sidak), but may be too conservative to identify areas for improvement in clinical practice. The problems with Bonferroni-type methods in this application are (I) they do not account for the correlations among the tests, and (ii) they do not account for the extreme discreteness of the data (adverse event rates from 1-5% following CABG). Correlations result from the dependence among binary indicators of adverse events (if one adverse outcome appears, then another is also more likely to appear) and from the nonorthogonality of the contrasts used to compare on physician against all others. Incorporating correlations and discrete characteristics usually makes the multiplicity adjustments less conservative. Use of discrete characteristics can dramatically reduce amount of required multiplicity adjustment, as discussed in Westfall and Young (1993, pp. 156-169).

We have employed PROC MULTTEST to make multiplicity adjustments. The procedure incorporates distributional and correlational characteristics in obtaining the distribution of the minimum p-value for all tests, and each p-value is adjusted according to the distribution of the min P statistic. To achieve improved power, tests are performed in step-down fashion (Westfall and Young, pp. 66-67), so that the minimum p-value is adjusted according to the distribution of min P over all k tests, the second-smallest p is adjusted according to the distribution of min P over the (k-1) hypotheses excluding the most significant, and so on. The method controls the probability of declaring a false significance, and we have preserved the confidence of the surgeons that the CQI process will identify clinically relevant opportunities to enhance patient outcomes following CABG surgery. The confidence of physicians in the CQI process is critical because, traditionally, there exists an uneasy alliance between physicians and hospitals in matters of quality. This negative perception is due largely to a history of programs that focused primarily on "...finding errors in medical practice and imposing punitive, sometimes humiliating sanctions..." rather than on improving processes viewed by physicians as important for patient care (Chassin, 1996). This history is not trivial, and effort must be spent on getting clinicians to accept the value of CQI tools.

METHODS AND MATERIALS

PROC MULTTEST under SAS/STAT® (Version 6.11 for Windows®) was used to perform the permutation resampling procedures presented. Exact upper-tailed permutational Cochran-Armitage tests with step-down permutational resampling-based multiplicity adjustments balance the opportunity for identifying clinically meaningful differences among surgeons with protection against declaring false significance. The number of resamples used was 20,000 which required 49 seconds to execute using an Intel Pentium® 120 MHZ processor. As a general rule, one should use as many resamples as possible in order to minimize the Monte Carlo Error, which is $\{p(1-p)/\text{nresample}\}^{1/2}$, where nresample is the number of resampled data sets. In this example, the surgeons' vectors of binary outcomes are resampled to preserve the correlations among the binary adverse event outcomes. Since p-values are computed for all tests within each resampled data set, correlations among non-orthogonal contrasts also are incorporated. Finally, exact tests are computed for all tests for each resampled data sets, therefore the discrete characteristics of the data are also incorporated in the multiplicity adjustments.

The code used to generate the multiplicity adjustments follows:

```
proc multtest pvals stepperm n=20000;
        class mdcat;
        test ca(hdeath mi_ekg rfb infect neuro
        pulm renal
           /upper permutation=50);
        contrast "1 vs. rest" 5  -1 -1 -1 -1 -1 ;
        contrast "2 vs. rest" -1  5 -1 -1 -1 -1 ;
        contrast "3 vs. rest" -1 -1  5 -1 -1 -1 ;
        contrast "4 vs. rest" -1 -1 -1  5 -1 -1 ;
        contrast "5 vs. rest" -1 -1 -1 -1  5 -1 ;
        contrast "6 vs. rest" -1 -1 -1 -1 -1  5 ;
run;
```

As mentioned previously, the probability of declaring at least one of the 42 comparisons to be falsely significant is 0.884 (1-[1-0.05]^{42}). The assumptions of independence and uniform distributions made in this calculation are not valid in our example because of the discreteness of the measures, thus 88% is presented as an upper bound. Fortunately, PROC MULTTEST allows for and incorporates dependencies and non-uniformity of the distributions into the multiplicity adjustment. In contrast to traditional methods, when the complete null hypothesis is true, the probability of erroneously declaring a significant surgeon effect remains approximately 5% when using the upper-tail MULTTEST adjusted p-value.

RESULTS

During the first quarter of 1996, 197 primary CABG surgeries were performed by the six surgeons operating at MMH. The raw p-values resulting from the Cochran-Armitage exact contrast test showed surgeon 3 to have a higher MI_EKG rate than the rest (p=.0499), and surgeon 5 to have higher NEURO adverse outcome rate than the rest (p=.0446). However, when making multiplicity adjustments, there was insufficient evidence to reject the null hypothesis of no surgeon effect. Table 1 presents the SAS output showing the adverse event rates for each surgeon, the raw p-value, the p-value resulting from the permutation multiplicity adjustment and the p-value resulting from Bonferroni adjustment.

In this example, evaluation of surgeon-specific adverse outcome rates would lead to two borderline significant results. However, when all tests are considered jointly, there is a 0.4587 chance of seeing a p-value as small as 0.0446, when there is no difference among surgeons. In light of the potential negative consequences of the CQI process, it is paramount to protect against spurious results. Therefore, we recommend use of multiplicity adjustment for the evaluation and comparison of surgeon-specific adverse outcomes. PROC MULTTEST performs such adjustments. It protects the familywise error rate while achieving improved power through incorporation of correlational and distributional characteristics. In this example, an independence-assuming adjustment of the p-value .0446 would be performed as 1-(1-0.0446)^{42} = 0.853. The corresponding MULTTEST adjustment 0.4587, while still insignificant, shows the potential improvement in power that can be obtained.

DISCUSSION

The issue of multiple comparisons in this paper is dealt with by assuming all 6x7=42 tests constitute a single "family." There are other possible approaches. For example, one might use 7 separate families, one for each adverse outcome, each containing the 6 comparisons among surgeons. In this case, the familywise Type I error rate (FWE) is controlled for each family, but when all families are considered jointly, the overall Type I error rate can be as large as 7x0.05 = 0.35 (approximately, using Bonferroni). An alternative is to "weight" the families differently: because hospital death is much more important than the remaining adverse events, one might consider the six surgeon-specific comparisons within the HDEATH category as one family, and the remaining 6x6=36 comparisons as a second family. Use of PROC MULTTEST for each of these families will control the FWE at 0.05 for each family individually, and it will control the FWE for both together at a rate no larger than 2x0.05=0.10. As another alternative, a composite score could be used (e.g., weighted sum of all adverse events which counts hospital death more heavily). This approach might lead to confusion, however, when it comes time to identify improvement opportunities.

The example of using resampling techniques for CQI purposes in a hospital setting is very promising. Traditionally, a barrier to acceptance of CQI techniques in the health care setting has been the concern that an individual might be erroneously indicated to have unacceptable performance. Standard statistical techniques without adjustments for multiplicity may well make that fear justified. However, we have shown that with resampling adjustments, protection against false significance can be preserved. Moreover, this protection against Type I errors does not compromise the ability to detect clinically important differences. This balance has convinced cardiovascular surgeons that appropriate statistical tools have been identified to enhance patient outcomes through the CQI process.

REFERENCES

Chassin, M.R. (1996), "Quality of Health Care: Improving the Quality of Care," *The New England Journal of Medicine*, 335(14), 1060-1063.

Green, J.and Wintfeld, N. (1995), "Sounding Board: Report Cards on Cardiac Surgeons--Assessing New York State's Approach," *The New England Journal of Medicine*, 332(18), 1229-1232.

Westfall, P.H. and Young, S.S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, New York: John Wiley & Sons, Inc.

Gregory L. Pearce,
USWNCGLP@IBMMAIL.COM

Peter H. Westfall,
WESTFALL@TTU.EDU

Table 1. PROC MULTTEST Output

```
                                         MULTTEST PROCEDURE

Test for discrete variables:            Cochran-Armitage
Exact permutation distribution used:    Everywhere
Tails for discrete tests:               Upper-tailed
Strata adjustment?                      No
P-value adjustments:                    Stepdown Permutation
Number of resamples:                    20000
Seed:                                   66477

                          MULTTEST COEFFICIENTS
                                                 Class
  Contrast                     1        2        3        4        5        6

  1 vs. rest                   5       -1       -1       -1       -1       -1
  2 vs. rest                  -1        5       -1       -1       -1       -1
  3 vs. rest                  -1       -1        5       -1       -1       -1
  4 vs. rest                  -1       -1       -1        5       -1       -1
  5 vs. rest                  -1       -1       -1       -1        5       -1
  6 vs. rest                  -1       -1       -1       -1       -1        5


                             MULTTEST TABLES
                                                  Class
 Variable Statistic            1        2        3        4        5        6

 HDEATH   Count             1.00     0.00     0.00     1.00     0.00     0.00
          N                44.00    21.00    27.00    30.00    42.00    33.00
          Percent           2.27     0.00     0.00     3.33     0.00     0.00
 MI_EKG   Count             0.00     0.00     2.00     0.00     1.00     0.00
          N                44.00    21.00    27.00    30.00    42.00    33.00
          Percent           0.00     0.00     7.41     0.00     2.38     0.00
 RFB      Count             1.00     0.00     0.00     1.00     0.00     1.00
          N                44.00    21.00    27.00    30.00    42.00    33.00
          Percent           2.27     0.00     0.00     3.33     0.00     3.03
 INFECT   Count             0.00     0.00     1.00     0.00     0.00     0.00
          N                44.00    21.00    27.00    30.00    42.00    33.00
          Percent           0.00     0.00     3.70     0.00     0.00     0.00
 NEURO    Count             0.00     0.00     0.00     0.00     2.00     0.00
          N                44.00    21.00    27.00    30.00    42.00    33.00
          Percent           0.00     0.00     0.00     0.00     4.76     0.00
 PULM     Count             1.00     1.00     2.00     0.00     1.00     0.00
          N                44.00    21.00    27.00    30.00    42.00    33.00
          Percent           2.27     4.76     7.41     0.00     2.38     0.00
 RENAL    Count             1.00     0.00     0.00     1.00     0.00     0.00
          N                44.00    21.00    27.00    30.00    42.00    33.00
          Percent           2.27     0.00     0.00     3.33     0.00     0.00

                             MULTTEST P-VALUES

              1 vs. rest             2 vs. rest             3 vs. rest
Variable Raw_p   StepPerm_p     Raw_p   StepPerm_p     Raw_p   StepPerm_p

HDEATH   0.3977       1.0000    1.0000      1.0000     1.0000      1.0000
MI_EKG   1.0000       1.0000    1.0000      1.0000     0.0499      0.5723
RFB      0.5336       1.0000    1.0000      1.0000     1.0000      1.0000
INFECT   1.0000       1.0000    1.0000      1.0000     0.1371      0.8940
NEURO    1.0000       1.0000    1.0000      1.0000     1.0000      1.0000
PULM     0.7216       1.0000    0.4343      1.0000     0.1394      0.8981
RENAL    0.3977       1.0000    1.0000      1.0000     1.0000      1.0000

              4 vs. rest             5 vs. rest             6 vs. rest
Variable Raw_p   StepPerm_p     Raw_p   StepPerm_p     Raw_p   StepPerm_p

HDEATH   0.2820       0.9992    1.0000      1.0000     1.0000      1.0000
MI_EKG   1.0000       1.0000    0.5149      1.0000     1.0000      1.0000
RFB      0.3925       1.0000    1.0000      1.0000     0.4248      1.0000
INFECT   1.0000       1.0000    1.0000      1.0000     1.0000      1.0000
NEURO    1.0000       1.0000    0.0446      0.4587     1.0000      1.0000
PULM     1.0000       1.0000    0.7027      1.0000     1.0000      1.0000
RENAL    0.2820       0.9992    1.0000      1.0000     1.0000      1.0000
```