Resampling with PROC MULTTEST: Providing Tools For Cardiac Surgeons To Identify Clinical Practice Improvement Opportunities

Gregory L. Pearce, Mission + St. Joseph's, Asheville, North Carolina
Peter H. Westfall, Texas Tech University, Lubbock, Texas

ABSTRACT

Mortality and major morbidity rates following coronary artery bypass graft (CABG) surgery are commonly used as measures to assess the effectiveness of clinical practice strategies. Of particular interest are surgeon specific mortality and morbidity (adverse event) rates. At this institution, all cardiac surgeons receive quarterly reports delineating their individual adverse event rates compared to the composite of all other surgeons. These comparisons involve six surgeons who perform approximately 1000 CABG surgeries each year. Mortality plus six major morbid events are included in the analyses. The primary reason for the reports is the identification of 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. Without adjustment, the probability of declaring a significant difference spuriously approaches 88%. Adjustment techniques to address the multiple comparison problem are available (e.g., Bonferroni) 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 multiplicity adjustments, thus protecting against Type I errors, while improving the power by incorporating discrete characteristics of the data. The Cochran-Armitage linear trend test and linear contrasts are used to make 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" of raw mortality data in New York and Pennsylvania newspapers (Green and Wintfeld, 1995). Such rudimentary reporting mechanisms are viewed by many health care professionals as lacking the scientific basis to be useful as practice improvement tools. However, appropriate evidence-based analytical tools are requisite to providing superior patient care (Evidence Based Medicine Working Group, 1992).

In recognition of the need for good decision-making tools, more than 700 institutions now participate voluntarily in the Society of Thoracic Surgeons National Cardiac Surgery Database (STSNDB). This institution has participated in the STSNDB since 1992. Six surgeons perform over 1,200 cardiac surgical procedures each year at MMH the majority (1000) of which are primary (first incidence) CABG surgeries.

The principal reason for our participation in the STSNDB 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. Specifically, where π_{ij} is the probability that an adverse event of type i will occur for surgeon j,

$$H_{ij}: \pi_{ij} = \overline{\pi_{i(-j)}}, \text{ where } \overline{\pi_{i(-j)}}$$

we test the hypotheses
is a weighted average of the proportions for the remaining physicians. The weights are implied by the Cochran-Armitage testing method; larger sample sizes for particular surgeons imply higher weights.

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 (j=1,...,6) being made on each of the seven adverse events (i=1,...7), we make 42 comparisons. 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 generally range 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 non-orthogonality 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. The resultant adjusted p-values can be defined as:

$$p'_{ij} = \Pr(\min P_{kl} \le p_{ij}), \qquad (1)$$

The adjusted p-value is easily interpreted as the probability that a p value as small as p_{ij} will be observed in the entire study when all null hypotheses are true.

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 tests, the second-smallest p is adjusted according to the distribution of min P over all 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). A thorough account of the history of efforts to use data to drive improvement in health delivery systems can be found in Moskowitz (1994). 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 6.12) 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. One-tailed (upper) testing was selected because we are interested in identifying situations where adverse event rates drift to an unacceptably high level. These methods could also be used to identify best practices, but we are not doing that currently. The Cochran-Armitage test was selected for this setting because it allows for a stratification variable while the Fisher's exact option does not. The number of resamples used was 200,000, which took approximately 3.5 minutes on a 2100 Alpha VAX EV-5 computer. As a general rule, one should use as many resamples as possible in order to minimize the Monte Carlo standard error, which is $\{p(1-p)/nresample\}^{1/2}$, where nresample is the number of resampled data sets. For example, if the adjusted p-value is 0.10, the Monte Carlo standard error is 0.00067 with 200,000 resampled data sets.

In this setting, 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 incorporating the discrete characteristics of the data in the multiplicity adjustments.

The code used to generate the multiplicity adjustments follows:

```
proc multtest pvals stepperm n=200000;
        class mdcat;
        test ca(hdeath mi_ekg rfb infect neuro
                pulm renal/upper
permutation=50);
        strata risk;
        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 can be as large as 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

During the third quarter of 1996, 213 primary isolated CABG surgeries were performed by the six surgeons operating at our institution. Adverse event rates for each surgeon are presented in Table 1.

Example 1 - Reopen for Bleeding

Surgeon 3 had more patients returning to the operating room for bleeding complications than his peers (Table 1). The unadjusted p-value for this comparison suggests a statistically significant different (p = 0.0146)

Table 1. Adverse event rates for each surgeons.

		Surgeon					
Variable	Statistic	1	2	3	4	5	6
DTH	Count	3.00	0.00	0.00	2.00	0.00	0.00
	N	37.00	24.00	35.00	50.00	39.00	28.00
	Percent	8.11	0.00	0.00	4.00	0.00	0.00
CKMB150	Count	2.00	1.00	1.00	3.00	1.00	0.00
	N	37.00	24.00	35.00	50.00	39.00	28.00
	Percent	5.41	4.17	2.86	6.00	2.56	0.00
REOPEN	Count	0.00	0.00	3.00	1.00	0.00	0.00
	N	37.00	24.00	35.00	50.00	39.00	28.00
	Percent	0.00	0.00	8.57	2.00	0.00	0.00
WOUND	Count	0.00	0.00	0.00	0.00	0.00	0.00
	N	37.00	24.00	35.00	50.00	39.00	28.00
	Percent	0.00	0.00	0.00	0.00	0.00	0.00
NEURO	Count	1.00	1.00	0.00	2.00	3.00	0.00
	N	7.00	24.00	35.00	50.00	39.00	28.00
	Percent	2.70	4.17	0.00	4.00	7.69	0.00
PULM	Count	2.00	4.00	0.00	0.00	0.00	2.00
	N	37.00	24.00	35.00	50.00	39.00	28.00
	Percent	5.41	16.67	0.00	0.00	0.00	7.14
RENFAIL	Count	1.00	0.00	0.00	1.00	1.00	0.00
	N	37.00	24.00	35.00	50.00	39.00	28.00
	Percent	2.70	0.00	0.00	2.00	2.56	0.00

. However, the multiplicity adjusted p-value indicates that there is a 17% chance (p = 0.1713) that a p-value as low as the unadjusted 0.0146 value would be returned if nothing more than random variation were occurring. Therefore, it was not felt that any action should be taken with regard to surgeon 3. Figure 1 shows the reopen for bleeding rate over time for surgeon 3 versus his peers over time. The reference line indicates the national average rate (2.2%)

the upper-tail MULTTEST adjusted p-value.

RESULTS

Table 2. Results for Surgeon 2 versus other surgeons.

Var	Raw_p	Bon_p	StepBon_p	StepSid_p	StepPerm_p
DTH	1.0000	1.0000	1.0000	1.0000	1.0000
CKMB150	0.7684	1.0000	1.0000	1.0000	1.0000
REOPEN	0.0146	0.1900	0.1835	0.1682	0.1713
WOUND	1.0000	1.0000	1.0000	1.0000	1.0000
NEURO	1.0000	1.0000	1.0000	1.0000	1.0000
PULM	1.0000	1.0000	1.0000	1.0000	1.0000
RENFAIL	1.0000	1.0000	1.0000	1.0000	1.0000

Example 2 - Pulmonary Complications

The raw p-values resulting from the Cochran-Armitage exact contrast test showed surgeon 2 to have a higher pulmonary complication rate than the rest of the surgeons (p=0.0065). Multiplicity adjustments resulted in an adjusted p-value of 0.1024. This borderline result indicates only a 10% chance of observing an unadjusted p-value as low as 0.0065 under the condition of no surgeon effect. Table 3 presents the SAS output showing the adverse event rates for each surgeon, the raw p-value and the p-value resulting from the permutation multiplicity adjustment. These results indicated that further investigation was appropriate.

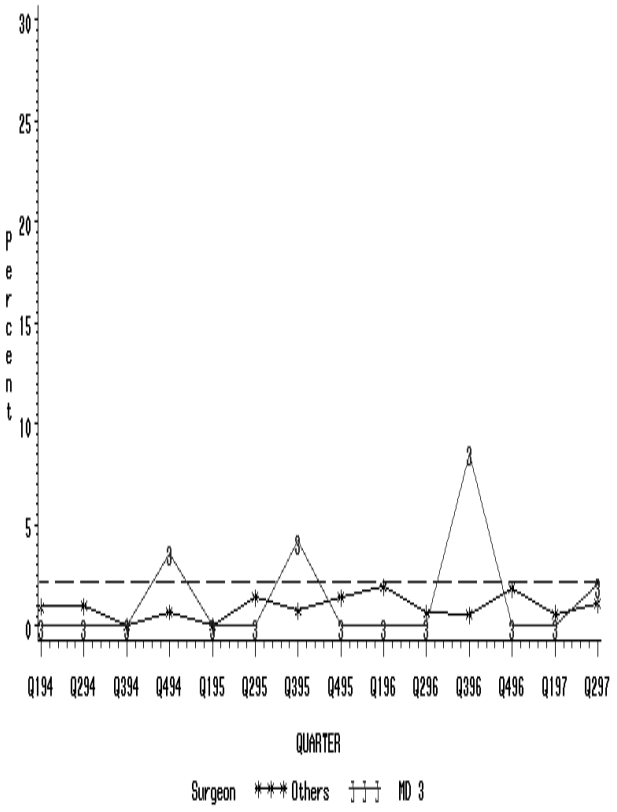
Table 3. Results for Surgeon 2 versus other surgeons.

Variable	Raw_p	Bon_p	StepBon_p	StepSid_p	StepPerm_p
DTH	1.0000	1.0000	1.0000	1.0000	1.0000
CKMB150	0.6222	1.0000	1.0000	1.0000	1.0000
REOPEN	1.0000	1.0000	1.0000	1.0000	1.0000
WOUND	1.0000	1.0000	1.0000	1.0000	1.0000
NEURO	0.5724	1.0000	1.0000	1.0000	1.0000
PULM	0.0065	0.1064	0.1064	0.1012	0.1024
RENFAIL	1.0000	1.0000	1.0000	1.0000	1.0000

Often, the first question raised with regard to surgical morbidity is case mix. That is, were the cases of the surgeon in question of similar risk to the remainder of the population. The STSNDB provides a risk stratification score was employed as a stratification variable to identify high-risk patients (Edwards et al., 1997).

Table 3 presents results of risk-stratified multiplicity adjustments. For this example, we conclude that differing case mix is not a reasonable explanation for the variability in the pulmonary complication rate between surgeon 3 and the remainder of the practice. The p-value of 0.11 still leaves us without a clear-cut conclusion as to whether the higher pulmonary complication rate could be a result of random variation.

Figure 1. Reopen for Bleeding Over Time



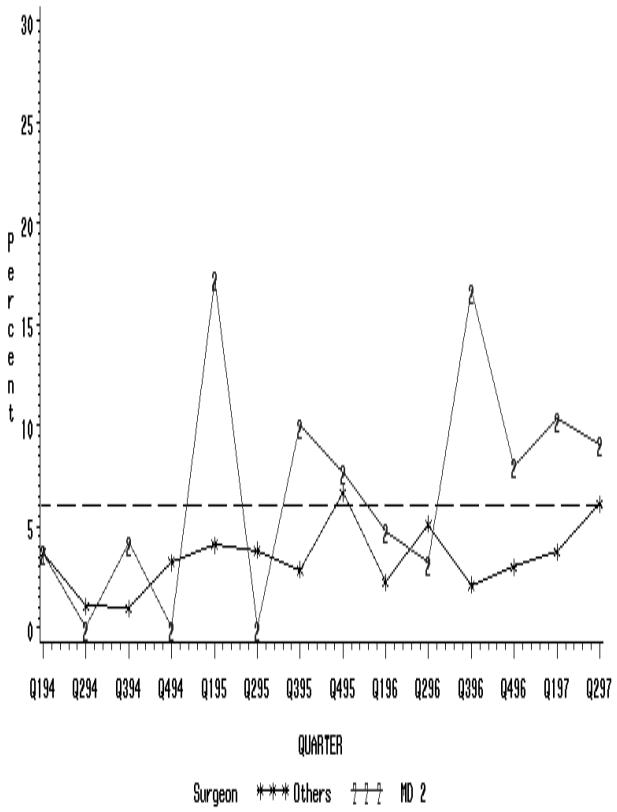
However, an examination of the pulmonary complication rates over time (Figure 2) indicates that the surgeon in question tended to have higher rates that his peers and the national average (6%). Therefore, the physician was informed of the likelihood that his patients were experiencing more pulmonary complications than those of his peers. At this point, any action to affect this rate was left to the surgeon. However, it can be seen that his pulmonary complication rate has been lower since this finding.

In these examples individual evaluation of surgeon-specific adverse outcome rates would lead to two statistically significant results. However, when all tests are considered jointly, there is a 17% chance of seeing a p-value as small as 0.0146 would be seen for reopen for bleeding and a 10% chance of seeing a p-value as small as 0.0065 for pulmonary complications. Examination of the graphs of the adverse events over time confirms the conclusions reached through the resampling based comparisons.

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 adjustments for the evaluation and comparison of surgeon-specific adverse outcomes. PROC MULTTEST performs such adjustments. It protects the familywise error rate (FWE) while achieving improved power through incorporation of correlational and distributional characteristics. In this example, an independence-assuming adjustment of the p-value .0065 would be performed as $1-(1-0.0065)^{42} = 0.24$. The corresponding MULTTEST adjustment 0.1024, while still not statistically significant, shows the potential improvement in power that can be obtained.

This improvement in power can be illustrated simply using the new "stepbon" option of PROC MULTTEST. These adjustments use the permutation distributions to calculate (1), but approximate the probability

Figure 2. Pulmonary Complications Over Time



using the Bonferroni inequality (see Westfall and Wolfinger, 1997, eq. 4). In Table 4, we see the stepbon adjustments, which incorporate discreteness, are substantially less than the simple Bonferroni adjustments (the "SIMPLE" column in Table 4 step-down Bonferroni adjustments (calculated as 42 x 0.0065 = 0.2726, 41 x 0.0146 = 0.5986, etc.) The effect of incorporating correlation is only slight compared to the effect of incorporating discreteness, as can be seen by comparing the stepbon with the stepperm columns, but it is seen that incorporating correlations improves power. Note also that the independence-assuming stepsid adjustments are anticonservative. We recommend using the steperm adjustments for these data.

Table 4. Test results sorted by raw p-value showing effects of incorporating discreteness

and correlations.

VAR	_LABEL_	RAW_P	SIMPLE	STPBON_P	STPSID_P	STPPERMP
PULM	2 v rest	0.0065	0.2724	0.1064	0.1012	0.1024
REOPEN	3 v rest	0.0146	0.6133	0.1835	0.1682	0.1713
DTH	1 v rest	0.0378	1.0000	0.5304	0.4152	0.4252
NEURO	5 v rest	0.1168	1.0000	1.0000	0.8528	0.8876
CKMB150	4 v rest	0.2818	1.0000	1.0000	0.9927	0.9988
PULM	6 v rest	0.2834	1.0000	1.0000	0.9939	0.9988
DTH	4 v rest	0.3347	1.0000	1.0000	0.9967	0.9995
CKMB150	1 v rest	0.4188	1.0000	1.0000	0.9998	1.0000

...35 more tests...

DISCUSSION

The use of resampling for surgeons specific comparisons should be viewed as a potential tool used to help physicians make better patient care decisions. It is certainly not the only tool. In cases where Type II error is more important (e.g. examination of practice-wide trends), multiplicity adjustments may not be necessary or even desirable.

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 FWE is controlled for each family. However, 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.

Some controversy surrounds the use of multiple testing methods. Criticism sometimes is based upon the idea that multiplicity adjustments propagate the "p-value culture." However, the reality that applied statisticians in health care face is that health care professionals have been highly exposed to the p-value and generally understand the meaning. The p-value terminology actually helps us communicate with the practitioner and helps bridge the gap from statistical theory to patient care.

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.

Edwards F.H., Grover F.L., Shroyer L.W.,

Schwartz M., and Bero J. (1997) "The Society of Thoracic Surgeons National Cardiac Surgery Database: Current Risk Assessment," *The Annals of Thoracic Surgery*, 63:903-908.

Evidence-Based Medicine Working Group (1992), "Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine," *JAMA*, 268:2420-2428.

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.

Moskowitz, D.B. (1994), *Ranking Hospitals and Physicians: The Use and Misuse of Performance Data*, New York: Faulkner and Gray, Inc.

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.

Westfall, P.H. and Wolfinger, R.D. (1997), "Multiple Tests with Discrete Distribution," *The American Statistician*, 51(1), 3-8.

Gregory L. Pearce,
MSJ.CHSGLP@MEMO.MSJ.ORG

Peter H. Westfall,
WESTFALL@TTU.EDU