

Identifying Clinical Practice Improvement Opportunities Using Resampling Techniques in PROC MULTTEST

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

ABSTRACT

Six surgeons perform more than 800 coronary artery bypass graft (CABG) surgeries annually at the Owen Heart Center, a program of the Mission + St. Joseph's Health System (MSJHS). 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. 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 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 and linear contrasts are 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" of

raw mortality data in New York and Pennsylvania newspapers. (Green and Wintfeld, 1995). Such rudimentary reporting mechanisms are perceived to lack 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 750 institutions now participate voluntarily in the Society of Thoracic Surgeons National Cardiac Surgery Database (STS Database). MSJHS and Asheville Cardiovascular and Thoracic Surgeons, P.A. have participated in the STS Database since 1992. Six surgeons perform over 1,200 cardiac surgical procedures each year at MMH the majority (> 800) of which are primary (first incidence) CABG surgeries.

The principal reason 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 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 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 $\{p(1-p)/nresample\}^{1/2}$, 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 set, 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 mdcatt;
  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 third quarter of 1996, 225 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 2 to have a higher pulmonary complication rate than the rest of the surgeons ($p=.0065$). However, when making multiplicity adjustments, there was insufficient evidence to reject the null hypothesis of no surgeon effect ($p = 0.1010$). 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 a highly significant result. However, when all tests are considered jointly, there is a 10% chance of seeing a p-value as small as 0.0065, 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 .0065 would be performed as $1-(1-0.0065)^{42} = 0.24$. The corresponding MULTTEST adjustment 0.1010, while still not statistically significant, shows the potential improvement in power that can be obtained.

Another adjustment that can be made to the MULTTEST model is the addition of a stratification variable. For example, an

evaluation delineating high risk surgery could be made (e.g. STS Risk Stratification Score, Edwards et al., 1997). The analyses could then be stratified according to risk level to present a more detailed look at the morbidity issue.

DISCUSSION

The issue of multiple comparisons in this paper is dealt with by assuming all $6 \times 7 = 42$ 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 $7 \times 0.05 = 0.35$ (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 $6 \times 6 = 36$ 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 $2 \times 0.05 = 0.10$. 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.

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.

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.

SAS and SAS/STAT are registered trademarks of SAS Institute Inc. In the USA and other countries. ®Indicates USA registration.

Other brand and product names are registered

statistics

trademarks or trademarks of their respective
companies.

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 trend
 Exact permutation distribution used: Everywhere
 Tails for discrete tests: Upper-tailed
 Strata adjustment? No
 P-value adjustments: Bonferroni Permutation
 Number of resamples: 20000
 Seed: 48905

MULTTEST COEFFICIENTS

Test	1	2	3	Class	4	5	6
MD1 v rest	5	-1	-1	-1	-1	-1	-1
MD2 v rest	-1	5	-1	-1	-1	-1	-1
MD3 v rest	-1	-1	5	-1	-1	-1	-1
MD4 v rest	-1	-1	-1	5	-1	-1	-1
MD5 v rest	-1	-1	-1	-1	5	-1	-1
MD6 v rest	-1	-1	-1	-1	-1	5	-1

MULTTEST TABLES

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

MULTTEST P-VALUES

Variable	MD1 v rest			MD2 v rest			MD3 v rest		
	Raw_p	Adj_p	Bon_p	Raw_p	Adj_p	Bon_p	Raw_p	Adj_p	Bon_p
DTH	0.0378	0.4385	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CKMB150	0.4188	1.0000	1.0000	0.6222	1.0000	1.0000	0.7684	1.0000	1.0000
REOPEN	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0146	0.1764	0.6133
WOUND	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
NEURO	0.7425	1.0000	1.0000	0.5724	1.0000	1.0000	1.0000	1.0000	1.0000
PULM	0.4188	1.0000	1.0000	0.0065	0.1010	0.2724	1.0000	1.0000	1.0000
RENFAIL	0.4375	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Variable	MD4 v rest			MD5 v rest			MD6 v rest		
	Raw_p	Adj_p	Bon_p	Raw_p	Adj_p	Bon_p	Raw_p	Adj_p	Bon_p
DTH	0.3347	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
CKMB150	0.2818	0.9999	1.0000	0.8076	1.0000	1.0000	1.0000	1.0000	1.0000
REOPEN	0.6600	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
WOUND	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
NEURO	0.5188	1.0000	1.0000	0.1168	0.9022	1.0000	1.0000	1.0000	1.0000
PULM	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.2834	1.0000	1.0000
RENFAIL	0.5538	1.0000	1.0000	0.4566	1.0000	1.0000	1.0000	1.0000	1.0000