Guidance for Statistical Analysis of Biomarker Data Generated From the NCI Colon Tissue Microarray

Last revised 2/24/08 by L.M.M.

The information provided in this document is for guidance only and does not eliminate the need for input from an individual with statistical expertise.

The NCI Colon TMA was designed to allow investigators to examine associations of the markers they are studying with tumor stage, clinical outcome, and other clinico-pathologic variables. The cases represented on the TMA were selected to ensure that there would be adequate representation of cases that eventually recurred as well as cases that did not recur within a substantial length of follow-up. The number of cases on the TMA breaks down as displayed in Table 1. This case selection method ensures that statistical power will be high to compare the distribution of the marker (e.g., prevalence of positive for binary markers) between recurrent and non-recurrent cases *within* each stage group separately. A consequence of this is that the proportion of recurrent cases on the TMA for stage II is higher than would be expected had II cases been selected completely at random (i.e., as a simple random sample from all stage II cases without stratification by recurrence status). Stage III recurrent cases were also slightly oversampled compared to stage III cases not recurring within 5 years.

If the goal of the analysis is to compare marker values between recurrent and non-recurrent cancer cases (within 5 years of diagnosis) within each stage group *individually*, this stratified sample design for case selection presents no particular statistical challenges. The usual two-sample test of binomial proportions (binary markers) or two sample t-tests (continuous markers) are appropriate for comparing the recurrers and non-recurrers within a given stage group.

Additional cases with less than five years of follow-up were included on the TMA to make available additional cases for investigators who wish to consider other time windows for recurrence. For example, if an investigator wishes to compare marker values between cases who recurred within two years of original diagnosis and cases who remained recurrence free for at least two years after diagnosis, then the cases with less than five years follow-up but with at least two years of follow-up could be used in the analyses.

If it is desired to estimate the mean marker level (continuous marker) or prevalence of positivity of the marker (binary marker), either within stage group II or III or combining across stage groups, then it is necessary to statistically adjust for the fact that cancer cases were selected differentially depending on recurrence status and stage. This adjustment is accomplished through the use of stratified estimators. The cancer cases were selected from eight strata: 1) stage I, 2) stage II-recurred within 5 years, 3) stage II-not recurred within 5 years, 4) stage II-followed less than 5 years, 5) stage III-recurred within 5 years, 6) stage III-not recurred within 5 years, 7) stage III-followed less than 5 years, and 8) stage IV. Formulas for means, prevalence estimates, and their estimated variances are given below.

For continuous marker measurements, the following estimators are used.

Stratified Weighted Mean Estimator:

$$\overline{Y} = \frac{1}{N} \sum_{k \in K} N_k \overline{y}_k \tag{1}$$

Stratified Weighted Variance Estimator for Mean:

$$\hat{var}(\overline{Y}) = \frac{1}{N^2} \sum_{k \in K} N_k^2 s_k^2 / n_k \tag{2}$$

Notation in the above formulas is defined as follows.

K is the set of indices designating the strata over which one is combining (see column 6 of Table 1).

n_k is the total number of cases selected from stratum k for inclusion on the TMAs (see column 7 in Table 1).

N_k is the potential number of samples in stratum k (see column 8 in Table 1).

$N = \sum_{k \in K} N_k$.

$\overline{y}_k = \sum_{i-1}^{n_k} y_{ki} / n_k$ is the sample mean of all marker values measured on samples from stratum k.

$s_k^2 = \sum_{i=1}^{n_k} (y_{ki} - \overline{y}_k)^2 / (n_k - 1)$ is the sample variance of all marker values measured on samples from stratum k.

For example, if one were interested in estimating the mean marker value for all stage II cases, formulas (1) and (2) would be applied across strata 2, 3, and 4 as defined in column 6 of Table 1, i.e., $K = \{2,3,4\}$.

If the marker value is binary (0/1) rather than a continuous measurement, then formulas (1) and (2) are replaced by formulas (3) and (4) below.

Stratified Weighted Proportion Estimator:

$$\hat{p} = \frac{1}{N} \sum_{k \in K} N_k \hat{p}_k \tag{3}$$

where $\hat{p}_k$ is the proportion of marker-positive results observed on samples from stratum k.

Stratified Weighted Variance Estimator for Proportion:

$$\hat{var}(\hat{p}) = \frac{1}{N^2} \sum_{k \in K} N_k^2 \hat{p}_k (1 - \hat{p}_k) / n_k \qquad (4)$$

Statistical comparisons between mean or proportion estimates produced using formula (1) or formula (3) can be carried out using approximate z- tests. For example, suppose that one wishes to compare the proportion of marker positive results for stage II invasive cancers versus stage III invasive cancers. Let $\hat{p}_{II}$ and $\hat{p}_{III}$ denote the estimates obtained using formula (3) with $K = \{2,3,4\}$ and $K = \{5,6,7\}$, respectively, and $\hat{\text{var}}(\hat{p}_{II})$ and $\hat{\text{var}}(\hat{p}_{III})$ be their respective variance estimates obtained using formula (4). A test of marker positivity rates between stage II and stage III cases could be based on the approximate z-statistic given by $z = (\hat{p}_{II} - \hat{p}_{III}) / \sqrt{\hat{\text{var}}(\hat{p}_{II}) + \hat{\text{var}}(\hat{p}_{III})}$. Tests for means follow in an analogous manner, using formulas (1) and (2). If one of the two proportions or means being compared is that obtained from the normal colon tissue from diverticulitis patients, then the proportion or mean for the normal colon tissue and their variance estimates are obtained using standard non-stratified estimators.

To conduct regression analyses, e.g. Cox proportional hazards regression, using the marker values obtained for the invasive cancer cases, the stratified case selection method can be accounted for by performing stratified weighted analyses. The sample weights to be applied to each case within each stratum are provided in the last column of Table 1. For example, for stratum 3 (stage II cancer that was followed and did not recur within 5 years of diagnosis), the sample weight that should be applied to each observation is 1.629. Some statistical packages such as SUDAAN and R ("survey" library) allow for such stratified weighted analyses. SUDAAN is a commercially available statistical software package that can be called from within SAS or used as a stand-alone application. R is freely available statistical software. The main R software and numerous libraries such as the "survey" library are available for download from the website http://www.r-project.org. Example SUDAAN (SAS-callable version) code (MarkerX_analysis_example_in_SUDAAN_colon_TMA_users.sas) and R code (MarkerX_analysis_example_in_R_colon_TMA_users.R) for analyzing a simulated marker dataset generated using the TMA are provided with this document. Tab-delimited text file and Excel file versions of the example data are contained in the files Colon_TMA_markerX_full_data_colon_TMA_users.txt and Colon_TMA_markerX_full_data_colon_TMA_users.xls, respectively. Output files generated using the example dataset with the SUDAAN and R code are provided in the files MarkerX_analysis_example_in_SUDAAN_output_colon_TMA_users.doc and MarkerX_analysis_example_in_R_output_colon_TMA_users.txt. The example code is provided to demonstrate the proper specification of the sampling design. Specific regression models to be fit would vary depending on the aims of the study and nature of the marker data collected.

Table 1. Distribution of Cases on the NCI Colon TMA

Type of Core	Case Set 1	Case Set 2	Case Set 3	Case Set 4	Stratum number for cancer cases (k)	Total # of cases selected from stratum (n_k)	Potential # of samples in stratum (N_k)	Sample weight for each observation in stratum ($w_k=N_k/n_k$)
Stage I colon cancer	12	13	12	12	1	49	105	2.143
Stage II colon cancer								
a: Recurred ≤ 5 years after diagnosis	8	6	6	6	2	26	28	1.077
b: Followed and not recurred within 5 years of diagnosis	18	81	17	17	3	70	114	1.629
c: Not recurred but not followed a minimum of five years	5	7	7	7	4	26	56	2.154
Stage III colon cancer								
a: Recurred ≤ 5 years after diagnosis	15	19	15	16	5	65	69	1.062
b: Followed and not recurred within 5 years of diagnosis	15	14	16	15	6	60	69	1.150
c: Not recurred but not followed a minimum of five years	5	3	6	5	7	19	25	1.316
Stage IV colon cancer	13	13	12	14	8	52	110	2.115
Diverticulitis *(normal colon tissue from patient without cancer)*	10	10	10	10		40		
Normal colon tissue matched to colon cancer patient								
Stage I patient	2	1	2	1		6		
Stage IIa patient	1	1	1	1		4		
Stage IIb patient	1	1	1	1		4		
Stage IIc patient	1	1	1	1		4		
Stage IIIa patiet	1	1	1	1		4		
Stage IIIb patient	1	1	1	1		4		
Stage IIIc patient	1	1	1	1		4		
Stage IV patient	1	1	1	1		4		
Adenomatous colon polyp *(from patient without cancer)*	9	9	10	9		37		
Total colon tissue cores	**119**	**120**	**120**	**119**		**478**		

Type of Core	Case Set 1	Case Set 2	Case Set 3	Case Set 4	Stratum number for cancer cases (k)	Total # of cases selected from stratum (n_k)	Potential # of samples in stratum (N_k)	Sample weight for each observation in stratum ($w_k=N_k/n_k$)
Cell line controls								
CLC-a: SW403	2	2	2	2		8		
CLC-b: HT-29	2	2	2	2		8		
CLC-c: Caco-2	2	2	2	2		8		
CLC-d: SW48	2	2	2	2		8		
CLC-e: Colo205	2	2	2	2		8		
CLC-f: HCT-15	2	2	2	2		8		
Normal non-colon tissue controls								
NNC-a: Kidney	2	2	2	2		8		
NNC-b: Endometrium	2	2	2	2		8		
NNC-c: Prostate	2	2	2	2		8		
NNC-d: Breast	2	2	2	2		8		
Total control cores	**20**	**20**	**20**	**20**		**80**		
Grand Total Cores	**139**	**140**	**140**	**139**		**558**		

The four case sets indicated in Table 1 have been arrayed in quadruplicate. Therefore, a total of 16 TMA blocks were constructed. The number of cores represented on any single TMA section is 139-140 (119-120 colon tissue cores + 20 control cores). All cores are 0.6 mm in diameter. Each TMA user receives a minimum of two sections (from different replicate array blocks) from each case set. Thus, the minimum number of TMA sections that any investigator would receive is eight (4 case sets × 2 replicate TMA blocks per case set × 1 section per TMA block).