University of Wisconsin
SCHOOL OF MEDICINE
AND PUBLIC HEALTH

Department of Biostatistics
& Medical Informatics

Technical Report 230

November 2012

A Systematic Selection Method for the Development of Cancer Staging Systems

Yunzhi Lin[1], Richard Chappell[1,2], and Mithat Gönen[3]

[1] Department of Statistics, University of Wisconsin, Madison, WI
[2] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison,WI
[3] Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, U.S.A.

Box 4675, K6/446 Clinical Science Center
600 Highland Avenue
Madison, WI 53792-4675

www.biostat.wisc.edu

Ph:608-263-1706
Fax: 608-263-5579

A Systematic Selection Method for the Development of Cancer Staging Systems

Yunzhi Lin[1], Richard Chappell[1,2], and Mithat Gönen[3]

[1]Department of Statistics, University of Wisconsin - Madison, Madison, WI 53706, U.S.A.

[2]Department of Biostatistics & Medical Informatics, University of Wisconsin - Madison, Madison, WI 53706, U.S.A.

[3]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, U.S.A.

Abstract

The tumor-node-metastasis (TNM) staging system has been the anchor of cancer diagnosis, treatment, and prognosis for many years. For meaningful clinical use, an orderly, progressive condensation of the T and N categories into an overall staging system needs to be defined, usually with respect to a time-to-event outcome. This can be considered as a cutpoint selection problem for a censored response partitioned with respect to two ordered categorical covariates and their interaction. The aim is to select the best grouping of the TN categories. A novel bootstrap cutpoint/model selection method is proposed for this task by maximizing bootstrap estimates of the chosen statistical criteria. The criteria are based on prognostic ability including a landmark measure of the explained variation, the area under the ROC curve, and a concordance probability generalized from Harrell's c-index. We illustrate the utility of our method by applying it to the staging of colorectal cancer.

Keywords: Cancer staging; TNM System; Bootstrap; Model selection; Survival Analysis.

1 Introduction

The development of accurate prognostic classification schemes is of great interest and concern in many areas of clinical research. In oncology, much effort has been made to define a cancer classification scheme that can facilitate diagnosis and prognosis, provide a basis for making treatment or other clinical decisions, and identify homogeneous groups of patients for clinical trials. Among various classification schemes, the tumor-node-metastasis (TNM) staging system is widely used because of its simplicity and prognostic ability.

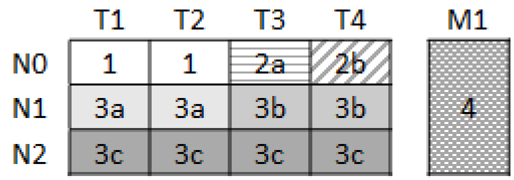
The basis of TNM staging is the anatomic extent of disease. It has three components: T for primary tumor, N for lymph nodes, and M for distant metastasis. TNM staging is periodically updated. Using its 6th edition, in the case of colorectal cancer which we will use as an example in this paper, there are 4 categories of T, 3 of N, and 2 of M [1]. Details of the categories are provided in Table 1. These TNM categories jointly define 24 distinct groups, which are unwieldy for meaningful clinical use [2]. Therefore, the American Joint Committee on Cancer (AJCC) and International Union against Cancer (UICC) defined an orderly, progressive grouping of the TNM categories which reduces the system to fewer stages (4 main stages and 7 sub stages under the 6th edition [1], see Figure 1). Alternative grouping schemes have also been proposed by other authors.

Table 1: The TNM staging system for colorectal cancer

T : Primary Tumor	
T1	Tumor invades submucosa
T2	Tumor invades muscularis propria
T3	Tumor invades into pericolorectal tissues
T4	Tumor directly invades or is adherent to other organs
N : Lymph Nodes	
N0	No regional lymph node metastasis
N1	Metastasis in 1-3 regional lymph nodes
N2	Metastasis in 4 or more regional lymph nodes
M : Distant Metastasis	
M0	No distant metastasis
M1	Distant metastasis

The value and usefulness of these TNM staging systems are, however, very much debated [3]. The main concerns are that the AJCC system is defined without systematic empirical investigation (by systematic, we mean the extensive division of the table in Figure 1 into all possible staging systems) and that it has too many (6) stages [2]. There is also a lack of commonly accepted statistical methods for developing staging systems. These critiques

Figure 1: Schematic showing the AJCC 6th edition staging system for colorectal cancer.



apply to staging all types of cancers.

Here and below, M1 patients will be omitted and relegated, as they usually are, to separate consideration. The reasons for this are two-fold: (1) M1 cancers are considered systemic diseases, as opposed to M0, which is considered localized; and (2) M1 has historically been the strong indicator of poor prognosis for almost all cancers. Currently, most studies on cancer staging systems are focused on evaluating and comparing existing proposals of TNM groupings [4, 5, 6]. However, we note that the AJCC system, as well as other proposed systems based on TNM, represent only a few of the numerous combinations of the T and N categories. We believe a thorough evaluation of all possible T and N combinations with respect to a possibly censored time-to-event outcome would be a more sensible way to develop good staging systems. Our goal is to answer the question: does the AJCC staging scheme outperform other possible T and N combinations in prognosis? If not, what is the best system out of all possible T and N combinations?

Searching for the best TNM grouping posed a challenging statistical problem. In this paper, a bootstrap model/cutpoint selection method is proposed for this task based on the following considerations. First, not all TNM combinations are eligible staging systems. As both categories are ordinal, only those combinations are eligible which are ordered in T given N and vice versa. A search algorithm that satisfies this partial ordering rule is proposed for generating all eligible staging systems. Second, the best staging systems can be simply defined as the ones that optimize the selection criterion chosen. Ideally, an external validation with a new population is desirable before determining the best system. In the absence of independently collected data, bootstrapping could be used as an alternative to provide replicate data sets for validating the selection [7]. Hence a bootstrap resampling strategy is proposed to estimate the optimal staging system, and to provide inference procedures (e.g. confidence intervals).

Selection criteria need to be identified that quantify the prognostic ability of candidate staging systems. A common approach for model development based on censored survival data is through the use of Cox proportional hazards model. Whereas the partial likelihood function as a statistical criterion is informative for looking at magnitude of effect, in certain clinical situations it might not be the most desirable option. It might be difficult to interpret for a non-statistician. Furthermore, since our problem is centered on evaluating prognostic classification schemes, which are inherently fully categorical and hence model-free, measures that check goodness-of-fit or that address model selection are less suitable for the task at hand. In view of these considerations, we elect to use measures that directly assess the prognostic ability of the staging systems. Several measures and ad hoc methods have been proposed for assessing prognostic ability; detailed reviews of these measures have been given by Schemper and Stare [8] and by Graf et al. [9], among others. In this paper, we elect to use the three criteria proposed by Begg et al. [6] and adapt them for comparison with our search algorithm: the explained variation for a specified "landmark" time, the area under the ROC curve for a landmark, and a concordance probability generalized from Harrell's c-index.

The structure of the paper is as follows. In Section 2 we describe the motivating data example of colorectal cancer patients. The bootstrap selection method is described in Section 3 and the criteria for finding the optimal staging system are explained in Section 4. The method is then illustrated on the colorectal cancer example in Section 5. Discussions and conclusions are given in Section 6.

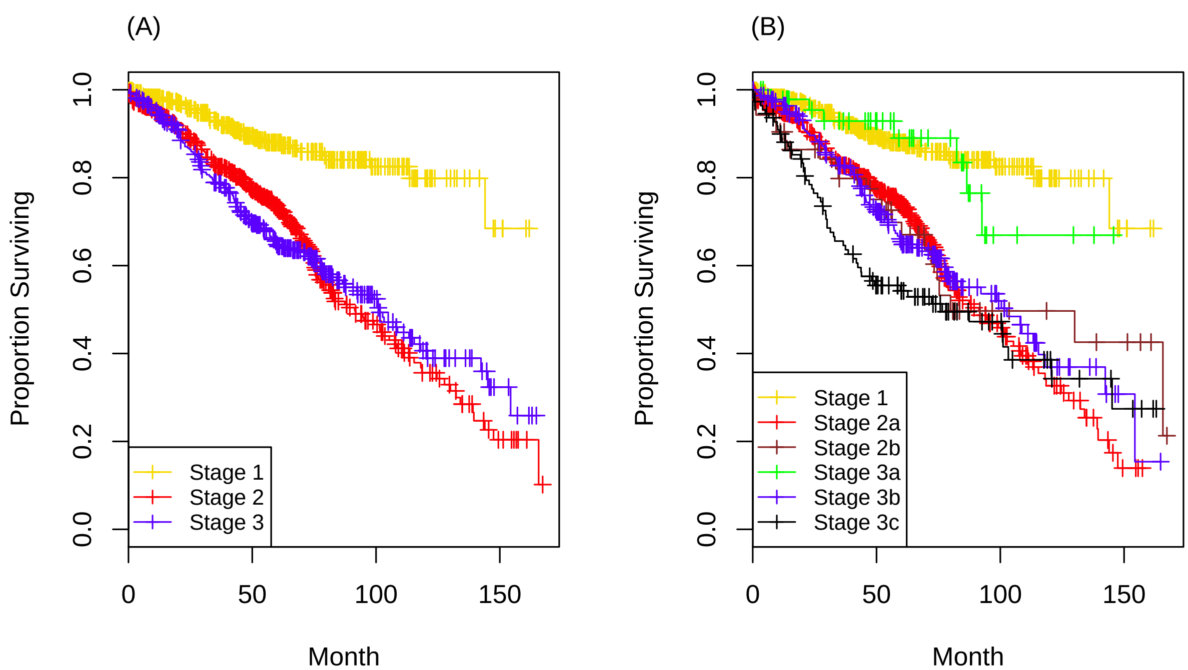
2 Motivating Example: Colorectal Cancer

We based our analysis on the de-identified database of 1,326 patients with non-metastatic colon cancer treated at Memorial Sloan-Kettering Cancer Center between January 1, 1990 and December 27, 2000 [10]. All patients are diagnosed with AJCC stage 1 to 3c disease (6th edition). The primary outcome used in the analysis is cancer-specific survival (only deaths attributable to recurrent cancer were counted as events). Of the 1,326 patients, 379 died by end of follow-up and the median survival was 115 months. Median follow-up time was 61.4 months. Table 2 presents the sample size, hazard ratio, and 10-year survival for each cell in the T$\times$N table. With a couple of exceptions, apparently due to small sample sizes, there is a strong upward trend in risk with increasing T and N involvement. However, we observe a relatively poor separation of the Kaplan-Meier survival curves under the AJCC 6th edition staging system (Figure 2).

Table 2: Estimated cancer-specific 10-year survivals/hazard ratios by TNM classifications (sample size).

	T1	T2	T3	T4
N0	0.87/1.00 (213)	0.73/2.44 (209)	0.33/5.63 (468)	0.50/4.99 (53)
N1	0.83/0.93 (14)	0.57/2.54 (34)	0.36/5.56 (197)	0.58/6.00 (27)
N2	0.50/4.37 (3)	1.00/0.00 (5)	0.33/8.00 (81)	0.43/10.27 (22)

Figure 2: Cancer-survival of colorectal cancer patients by the 6th edition AJCC staging system: (A) three main stages; (B) including sub stages.



3 Method

To identify the best staging system, we propose a search algorithm that scans through all eligible possibilities. In general, suppose the T descriptor has p categories, the N descriptor has q categories, and a k-stage system is desirable. The problem can be described by borrowing the framework of an outcome-oriented cutpoint selection problem for a censored response partitioned with respect to two ordered categorical covariates and their interaction. That is, we aim to estimate the best $k-1$ partition lines (cutpoints) that classify a partially ordered $p \times q$ table into k ordinal groups.

Calculating the number of all eligible partitions falls into the general mathematical problem of compositions of a grid graph [11], yet an analytical solution is not available for the general case. Numerical solutions can be obtained through computerized enumeration for small k, p, and q values, and they are given in Table 3 for small k's with $p = 4$ and $q = 3$, relevant to our colorectal cancer example.

Table 3: Number of eligible staging systems given k.

number of stage k	2	3	4	5	6
number of eligible systems	33	388	2,362	8,671	20,707

A value $n_{\min}$ is pre-specified for the minimum size of a stage, for example, $n_{\min}$ equals 5% of the sample size. Any system violating the $n_{\min}$ criterion will be dropped, and the remaining are the candidate systems. Let S denote the set of candidate systems, and let T_s denote the selection criterion value (see technical discussions in Section 4) for candidate system $s \in S$. The maximally selected criterion is

$$T_{\max} = \max_s T_s. \tag{1}$$

The maximally selected TN combination s^* is defined to be the one for which the maximum is attained, that is, for which the value of statistical criterion T_{s^*} equals $T_{\max}$ given k.

Our task differs from the usual cutpoint estimation problem which utilizes the maximally selected statistics. Under the maximally selected tests, the null hypothesis of interest is the independence between the covariate (to be dichotomized) and the response, and the estimation of a cutpoint comes after the rejection of the null hypothesis. This null hypothesis is irrelevant in our case as the prognostic ability of the T and N categories is well established and assumed to hold. Our inquiry takes one step further to ask, is the maximally selected TN combination s^* truly the optimal staging system for the population?

A bootstrap model selection strategy is therefore applied to estimate the optimal staging system. B bootstrap samples of size n (where n is the original sample size) are drawn with replacement from the original data. Denoting the bootstrap replication of $\hat{T}_s$ by $\hat{T}_s^b$, the bootstrap estimated criterion for candidate system s is given by

$$\tilde{T}_s = \frac{1}{B} \sum_{b=1}^{B} \hat{T}_s^b. \tag{2}$$

The bootstrap estimate of the best staging system $\tilde{s}^*$ is defined as the system that maximizes $\tilde{T}_s$.

There are two reasons that lead us to choose the bootstrap procedure:

1. The bootstrap method provides inference procedures (e.g. confidence intervals) for not only the optimal selected but all candidate systems, which enables us to examine the relative performance of any staging systems of interest and allows flexibility in the decision making process for clinical researchers and practitioners. In the analysis in Section 5, the standard bootstrap variance estimate was employed to construct the variance estimates of the measures for each candidate system, and the confidence intervals are also produced by bootstrapping.

2. The bootstrap selection procedure can easily adopt any measure of prognostic ability. In the next part of this section, three such measures will be introduced.

Note that a complete search through all eligible partitions for all bootstrap samples is, however, thwarted by combinational explosion. To overcome this problem, we can first compute the criterion for each eligible system using the complete data, and then only include the top m systems, say $m = 200$, (and the currently used staging system proposed by AJCC) as the "finalists" for the bootstrap selection procedure.

4 Criteria for Assessing Staging Systems

In this section, we discuss the three measures / criteria we choose for assessing the prognostic power of candidate staging systems.

4.1 Landmark Measures

An appealing way to simplify the analysis of survival data is to use a "landmark" time-point, such as 5-year or 10-year survival, and deal with only the censored binary outcome. This is frequently used in medical investigations. Here we elect to use the two landmark measures described by Begg et al. [6], the explained variation and the area under the ROC curve. Let θ_i, $i = 1, \ldots, c$, denote the probabilities of survival at the chosen landmark time or each of the c categories in the staging system, and p_i denote the prevalence of the stage categories. Let $\mu = \sum p_i \theta_i$ represent the unconditional mean outcome, and ν_i be the variance of θ_i. Then the estimated proportion of explained variation $\hat{\pi}$ is given by

$$\hat{\pi} = \frac{\sum \hat{p}_i \hat{\theta}_i^2 - (\sum \hat{p}_i \hat{\theta}_i)^2 - \sum \hat{p}_i \hat{\nu}_i}{(\sum \hat{p}_i \hat{\theta}_i)(1 - \sum \hat{p}_i \hat{\theta}_i)} \,, \tag{3}$$

and the area under the ROC curve $\hat{A}$ is estimated as

$$\hat{A} = \sum_{i=1}^{c} \frac{\hat{p}_i(1-\hat{\theta}_i)}{2\hat{\mu}(1-\hat{\mu})} \left\{ 2\sum_{j=1}^{i-1} \hat{p}_j\hat{\theta}_j + \hat{p}_i\hat{\theta}_i \right\} \tag{4}$$

where $\{\hat{\theta}_i\}$ are the Kaplan-Meier [12] estimates of the survival probabilities at the landmark time, $\{\hat{p}_i\}$ are the observed relative frequencies of the staging categories, and $\{\hat{\nu}_i\}$ are the variances of the observed values of $\{\hat{\theta}_i\}$ obtained from the Greenwood formula.

Using landmark times are less efficient statistically than using the entire survival distribution but provides for easier communication of results. In fact certain landmark times have become standards of reporting in various cancers such as 5 and 10 years in localized colorectal cancer. We include these measures also because in some situations landmark survival analysis can be more desirable than using the full survival. These include comparisons in which proportionality is obviously violated (e.g., when one stage is usually treated with a therapy which has a substantial immediate failure rate and another stage's failures tend to occur later) or those in which a landmark analysis is preferred for scientific reasons. An example of the latter might be a childhood cancer in which life extension is less relevant than the cure rate, and so a landmark measure such as 5-year survival could be used to stage these patients as a surrogate for cure.

4.2 Concordance Probability

Harrell et al. [13, 14] proposed the c-index as a way of estimating the concordance probability for survival data. It is defined as the probability that, for a randomly selected pair of participants, the person who fails first has the worse prognosis as predicted by the model. A limitation of Harrell's c-index is that it only takes into account usable pairs of subjects, at least one of whom has suffered the event. Begg et al. proposed an improved estimator of concordance which is adapted to account for all pairs of observations, including those for which the ordering of the survival times cannot be determined with certainty [6]. It requires the estimation of the probability of concordance for each pair of subjects and thus is computationally intensive for large sample sizes, particularly when bootstrapping. It also assumes that if the patient with the shorter censored value lives as long as the observed censored survival time in the paired patient, the remaining conditional probability of concordance is 1/2. As a result there is likely to be a conservative bias in the concordance estimator in the presence of high censoring rates [6].

Here we develop an estimator of the concordance probability under a classification scheme. Similar to Begg's approach, the new method utilizes the Kaplan-Meier estimates to evaluate the probabilities. Let K be the probability of concordance. For two patients randomly selected with stage (class) and survival time denoted by (S_1, T_1) and (S_2, T_2),

$$K = P\{(S_1 > S_2,\ T_1 < T_2) \text{ or } (S_1 < S_2,\ T_1 > T_2)\}. \tag{5}$$

Here we assume the survival time is inherently continuous although there could be ties in observed survival times. If $S1 = S2$, then the most common approach is to consider it equivalent to $S1 > S2$ with probability 1/2 and to $S1 < S2$ with probability 1/2. Thus (5) can be written as

$$K = 2P(S_1 > S_2,\ T_1 < T_2) + P(S_1 = S_2,\ T_1 < T_2). \tag{6}$$

Letting $S_1 = j$ and $S_2 = i$, $1 \le i < j \le k$, the first part of (6) can be estimated as

$$\begin{aligned} \hat{P}(S_1 > S_2,\ T_1 < T_2) &= \hat{P}(T_1 < T_2 | S_1 > S_2)\hat{P}(S_1 < S_2) \\ &= \sum\sum_{j>i} \hat{P}(T_1 < T_2 | j, i)\hat{P}(j, i) \\ &= \sum\sum_{j>i} \hat{P}(T_1 < T_2 | j, i)\frac{N_j N_i}{N(N-1)} \end{aligned} \tag{7}$$

where N_i, N_j are the sample sizes of stages i and j, respectively, and N is the total sample size.

Given i and j, and the last event time in all groups denoted by $t_{\max}$, we have

$$P(T_1 < T_2) = P(T_1 < T_2,\ T_1 \le t_{\max}) + P(T_1 < T_2,\ T_1 > t_{\max}). \tag{8}$$

When at least one event occurred,

$$\begin{aligned} P(T_1 < T_2,\ T_1 \le t_{\max}) &= \int_0^\infty dt_2 \int_0^{t_2} f_1(t_1) f_2(t_2) dt_1 \\ &= \sum_{t \in \{t_j\}} [S_j(t^-) - S_j(t)] S_i(t) \end{aligned} \tag{9}$$

where S_i and S_j can be estimated by the Kaplan-Meier survival estimators in stage i and j, and $\{t_j\}$ are the observed event times in stage j.

In the case when both observations are censored,

$$P(T_1 < T_2,\ T_1 > t_{\max}) = S_1(t_{\max}) S_2(t_{\max}) P(T_1 < T_2 | T_1, T_2 > t_{\max}). \tag{10}$$

The conditional probability $P(T_1 < T_2 | T_1, T_2 > t_{\max})$ is not estimable, but can be conservatively assumed to be 1/2 as in Begg et al., or assumed to be equal to the overall concordance

$P(T_1 < T_2)$. The latter is adopted in our method. That is,

$$\hat{P}(T_1 < T_2) = \frac{\sum_{t \in \{t\}_j} [\hat{S}_j(t^-) - \hat{S}_j(t)]\hat{S}_i(t)}{1 - \hat{S}_1(t_{\max})\hat{S}_2(t_{\max})} . \tag{11}$$

Similarly the second part of (6) can be estimated as

$$\hat{P}(S_1 = S_2,\ T_1 < T_2) = \frac{1}{2} \sum_i \frac{N_i(N_i - 1)}{N(N-1)} \tag{12}$$

and the overall concordance estimator is given by

$$\hat{K} = 2 \sum \sum_{j>i} \left\{ \frac{N_j N_i}{N(N-1)} \frac{\sum_{t \in \{t\}_j} [\hat{S}_j(t^-) - \hat{S}_j(t)]\hat{S}_i(t)}{1 - \hat{S}_1(t_{\max})\hat{S}_2(t_{\max})} \right\} + \frac{1}{2} \sum_i \frac{N_i(N_i - 1)}{N(N-1)} . \tag{13}$$

The new estimator improves upon Harrell's c-index, particularly in the presence of a large amount of censoring, by including comparisons between censored individuals. It is also much faster to implement than Begg's method. The statistic suffers from the usual criticism applied to concordance statistics; that is, they look only at the ranks of individuals and thus might be insensitive to small model improvements. Using survival times, however, often requires parametric modeling and alternative measures that are sensitive to small changes can also be sensitive to model choice. Using ranks can also be a benefit in that K is robust to outlying observations.

5 Analysis of Colorectal Cancer Data

We illustrated the utility of the proposed method by applying it to the staging of colorectal cancer. The number of stages is given as $k = 3$ and $k = 6$, corresponding to numbers of main and sub-stages in the 6th edition AJCC staging system. For the percent explained variation and the area under the curve measures, we tried both landmark times of 5 years and 10 years, based on the median follow-up time, and the results are very similar. We hence report here only the results from the 10-year landmark analysis.

5.1 Bootstrap Selection

The staging systems selected by maximizing the bootstrap estimates of each of the criteria described in Section 4, given $k = 3$ and $k = 6$, respectively, are presented in Figure 3, as well as the AJCC system for comparison. The systems selected by the three criteria are similar to each other and quite different from the AJCC system. Unlike the AJCC which separates stage 3 horizontally at N1, the bootstrap selected systems all classify groups primarily by the T categories (vertically). This is consistent with what we observe in Table 2, where the estimated 10-year survivals are much lower and the hazard ratios are much greater in categories T3 and T4.

Figure 3: Schematic showing staging systems selected by bootstrap and the AJCC 6th edition staing system. VAR: explained variation; AUC: area under the ROC curve; K: concordance probability.

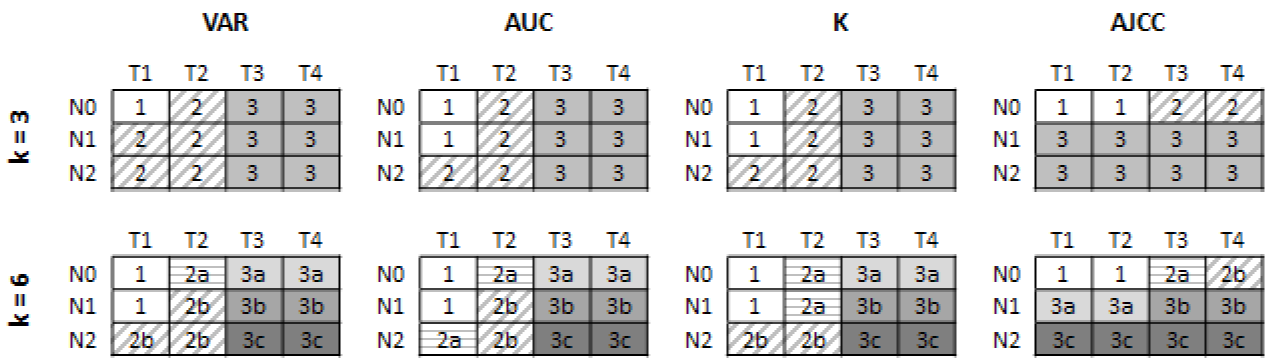


Table 4: Selected systems and the AJCC: the estimated criteria and their standard errors.

		Criteria (SE)		
	System	VAR	AUC	K
k = 3	A1	0.684 (0.010)	0.705 (0.011)	0.662 (0.008)
	A2	0.684 (0.010)	0.705 (0.012)	0.663 (0.007)
	AJCC	0.627 (0.008)	0.643 (0.011)	0.622 (0.008)
k = 6	B1	0.688 (0.009)	0.708 (0.011)	0.667 (0.008)
	B2	0.688 (0.009)	0.709 (0.011)	0.666 (0.008)
	B3	0.687 (0.009)	0.708 (0.011)	0.667 (0.007)
	AJCC	0.642 (0.012)	0.660 (0.013)	0.623 (0.008)

Let A1 and A2 denote the 3-stage systems selected by explained variation, and by area under the curve and concordance probability, respectively, and let B1, B2, and B3 denote the 6-stage systems selected by the three criteria, respectively. Table 4 shows the estimated value of the three criteria for these selected systems and the AJCC. The bootstrap selected systems are very similar with respect to all three criteria, which is not surprising given that the systems highly resemble each other. The prognostic power increases minimally as the number of stages increase from 3 to 6, indicating there is not much to gain by adding more sub-stages. In addition, of course, a 3-stage system is easier to use than a 6-stage one. The AJCC system is inferior to the selected ones in all cases.

Kaplan-Meier survival curves for the selected staging systems are displayed in Figure 4. All five systems show a substantial degree of prognostic separation and a clear advantage over the AJCC system in Figure 2. There is considerable overlap of survival curves in the right panel because of the larger number of stages, which again raise the question whether 6 distinct stages are too many.

5.2 Confidence Intervals

The bootstrap selection provides inference procedures for not only the optimal selected, but all candidate systems. Figure 5 shows the confidence intervals of each of the criteria for the top-ranked systems and the AJCC. The systems are ordered by their rankings with regard to the bootstrap estimated criteria. The top systems are in fact very close in terms of prognostic power, especially for 6-stage systems where the top 100 systems are virtually identical due to the fact that the systems are only slightly different from one another in their definition. It is hence difficult to select one best system, but it allows flexibility in the decision making process for clinical researchers, who might incorporate both statistical evidence and medical insight into their considerations. Among the 388 3-stage systems (only the top 150 shown), the AJCC ranks around 135 (35%), and it ranks around 12500 (60%) among the 20707 6-stage systems (the top 125 shown). Again, the AJCC demonstrates clearly lower prognostic power than the top systems.

5.3 Cross-Validation

Table 4 compares maximally selected values with the one given by fixed model, the AJCC, without adjusting for the maximization process. This could result in maximization bias that elevates the performance of the bootstrap selected systems, a problem well known in the context of model selection or cutpoint selection [15, 16, 17]. One approach for correcting the maximization bias is the use of cross-validation [15]. Here we use 10-fold cross-validation to reevaluate the bootstrap procedure. Basically, the data are randomly split into ten parts of similar size. Ten times we use 9/10 of the data for selection and each time apply the selected system to the omitted 1/10th of the data. Once the procedure is complete, all patients in the sample have been assigned to a stage. We then compute the estimated criteria under this staging assignment. For example, here we use the bootstrap selection procedure with the concordance criterion and set the desired number of stages to be 3. With the cross-validation adjustment, the estimated criteria are 0.669 (VAR), 0.688 (AUC), and 0.650 (K), which are still substantially superior to the estimates for AJCC in Table 4, indicating there is much room for improvement in the current system.

6 Discussion and Conclusions

An accurate staging system is crucial for predicting patient outcome and guiding treatment strategy. For decades investigators have developed and refined stage groupings using a combination of medical knowledge and observational studies, yet there appears to be no well established statistical method for objectively incorporating quantitative evidence into this process. In this paper, we have proposed a systematic selection method for the development of cancer staging systems, and illustrated the utility of this method by applying it to the staging of colorectal cancer. The staging systems selected by the three criteria are similar to each other while quite different from and superior to the current AJCC system, indicating there might be room for improvement in selecting it.

It is important to remember that staging, here and in the innumerable articles in the medical literature which discuss it, is undoubtedly confounded with treatment. The practical implication of this is that two or more groups of patients which are placed in the same stage may belong together either because their cancers' prognoses are intrinsically similar or because additional treatment to those with more advanced disease makes them so, or some combination of the two. Begg et al. concisely summarized one way to view the issue: "However, in thymoma, as in cancer in general, the relative impact of available treatments on cancer survival is much smaller than the impact of anatomical stage at diagnosis, and thus any confounding effect of the treatment is likely to be small. Furthermore, for thymoma, there is no widespread agreement on the ideal therapy. This, allied to the fact that the patients in our series were assembled from many different institutions for referral pathology, resulted in a wide variation of treatments administered by stage."

Our analysis of the colorectal cancer data has provided some insight into the prognostic power of the TNM staging system. The selected systems (A1, A2, B1, B2, and B3) are virtually identical in their prognostic accuracy regardless of which of the three evaluative measures is used. The selected 6-stage systems are a further division of the 3-stage systems with no apparent improvement in separating the survivals. Thus, it might be reasonable to favor a more parsimonious system as urged in Gönen and Weiser [2]. All final systems suggest that the most essential information is contained in the contrast between the tumor invading through the muscularis propria (T3 and T4) and otherwise (T1 and T2). This is in sharp contrast to AJCC where the primary distinction is between node-positive (N1 and N2) and node-negative (N0) cancers. Other near-top systems can be identified from the confidence interval plot or the bar plot showing the majority voted systems, and a compromise can be reach between statistical evaluation and medical judgment and common sense.

The choice between systems with three, six, or other numbers of stages involves a variety of considerations. The nature of the problem means that parsimony is important. But even if say three stages are preferred, such a system might be unsatisfactory if it glosses over obvious heterogeneity. The solution is an essentially medical one which combines issues of treatment regimen distinctions, diagnostic ease, and clinical practice. We recommend that an analyst give medical researchers several staging systems in a range of practical sizes along with their performance score as in Table 4 to allow them to compare the systems' prognostic capabilities.

We use bootstraps to provide bias-corrected estimates of performance for the staging systems. This addresses the internal validity which is a prerequisite for external validity yet does not guarantee it. External validity of a prognostic system can be established by being tested and found accurate across increasingly diverse settings. The selected systems should be tested across multiple independent investigators, geographic sites, and follow-up periods for accuracy and generalizability. The use of population-based datasets is important in establishing a staging system that is useful for the general patient population.

TNM staging is applicable to virtually any type of solid tumor hence, although we used colorectal cancer as illustration, our methodology has general appeal. In addition to cancers, many other diseases also use aggregate risk scores based on ordinal (or ordinalized) risk factors, such as the ATP III score for high-blood cholesterol that can benefit from optimal aggregation [18]. Our methodology is applicable in principle to binary outcomes as well since, in this case, $\hat{\theta}_i$'s in (3) can be directly estimated from the observed event rates in each risk category.

Cancer staging has been as much about anatomic interpretation as it is about accurate prognosis. A staging system that is prognostically optimal is unlikely to be adopted if it does not respect the anatomic extent of disease. Our results for colorectal cancer suggest that prognostically optimal systems are also anatomically interpretable. The substantial difference between the prognostic ability of the optimal systems and the AJCC categories is concerning especially in light of the fact that optimal systems are comparable to AJCC in terms of simplicity and interpretability.

Acknowledgements

The authors thank the anonymous reviewers and the Associate Editor for their insightful comments and suggestions, which have led to a significantly improved paper. We also thank Ronald Gangnon for many helpful discussions and suggestions. This research was supported in part by NIH/NCI Grant Number P30 CA014520 to the UW Carbone Cancer Center, Madison, WI. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the the NIH.

Figure 4: Cancer-specific survival of colorectal cancer patients by the selected staging systems. Left panel: 3-stage systems; right panel: 6-stage systems.

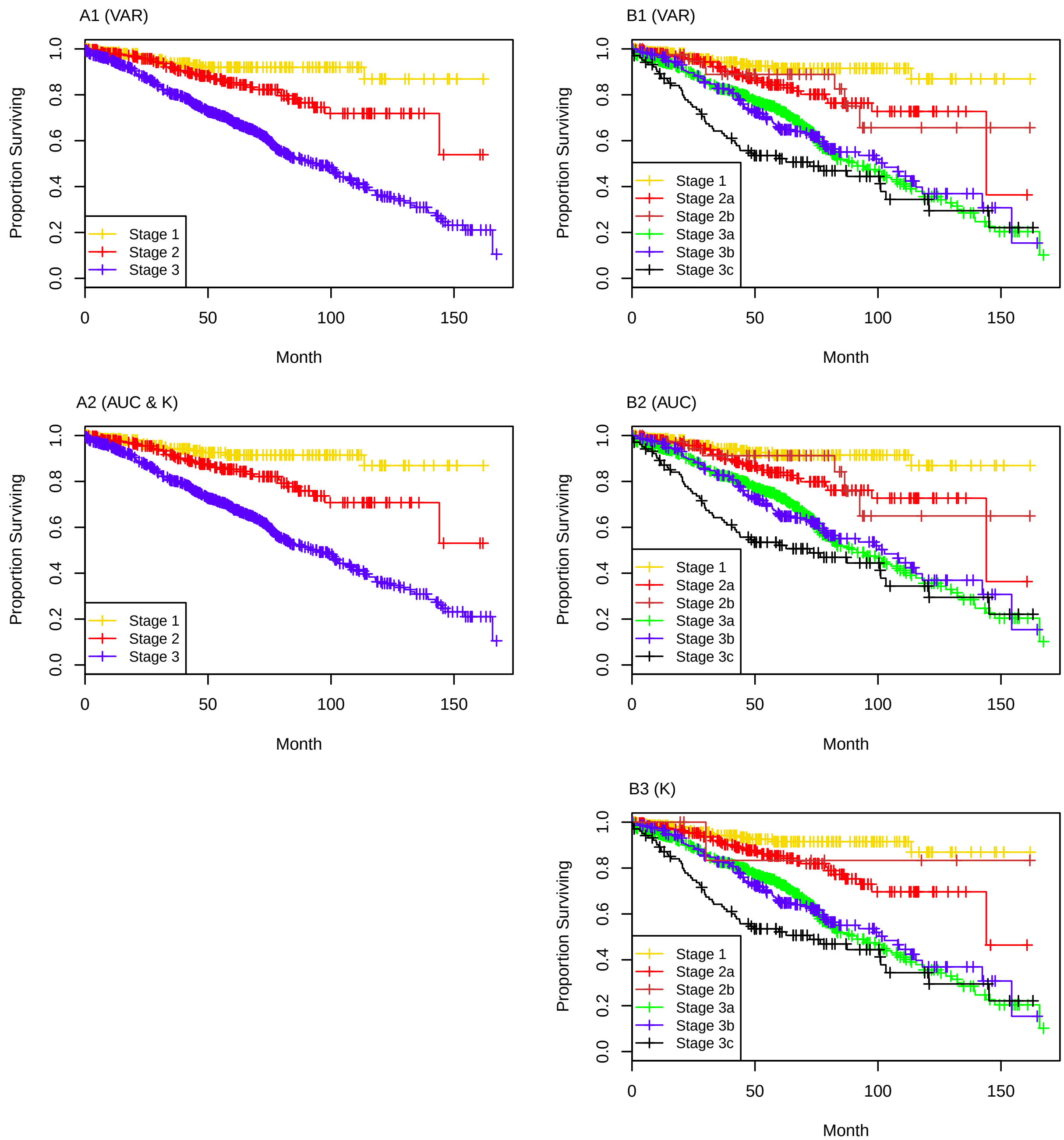
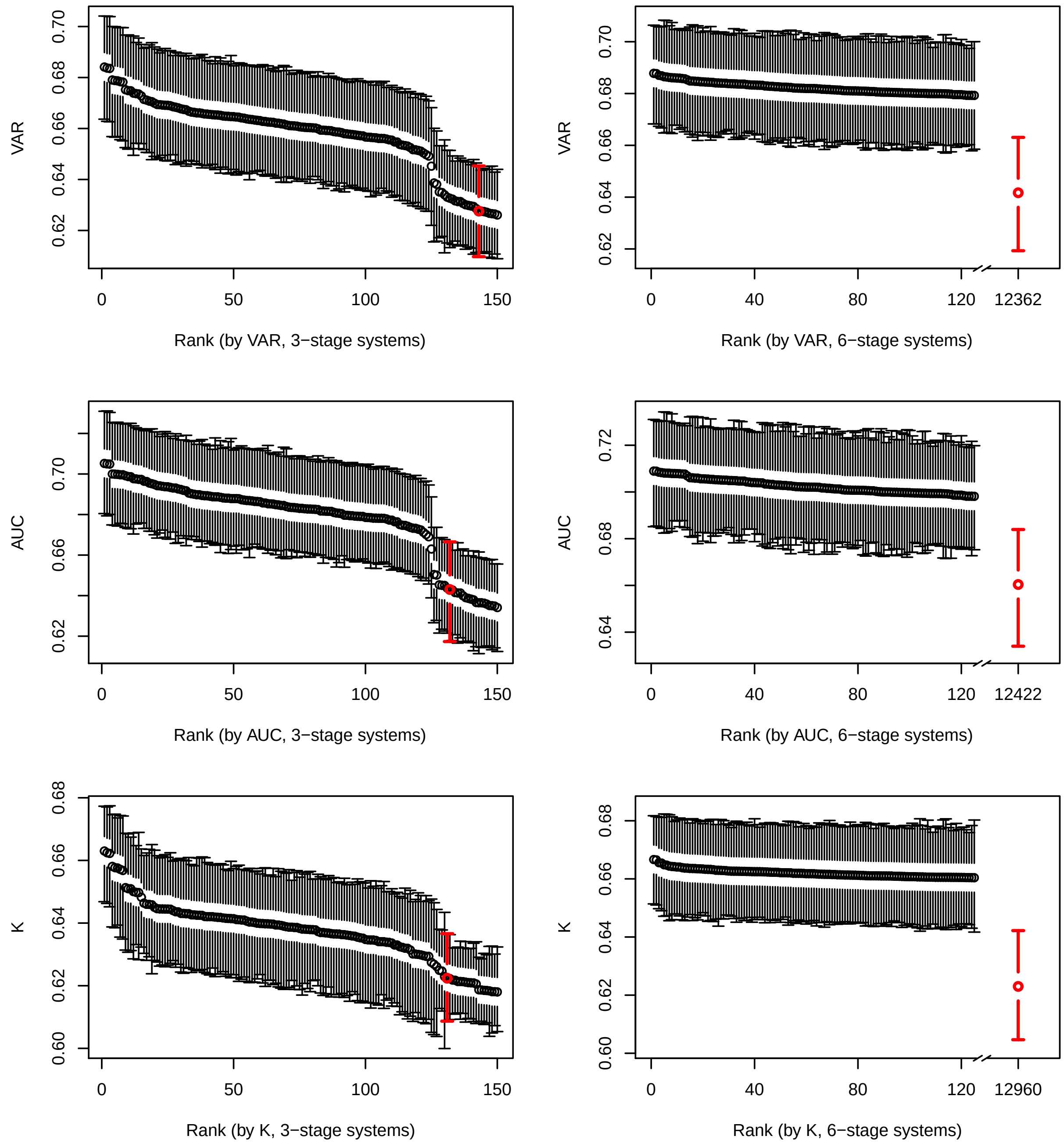


Figure 5: Confidence intervals for the criteria: the top-ranked systems and the AJCC (red). Left panel: 3-stage systems; right panel: 6-stage systems.



References

[1] Greene FL, Page DL, Fleming ID, et al. AJCC cancer staging manual. 6th ed. New York: Springer-Verlag, 2002.

[2] Gönen M, Weiser MR. Whither TNM? Semin Oncol 2010; 37:27-30.

[3] Benson AB III, Schrag D, Somerfield MR, et al. American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. J Clin Oncol 2004; 22:3408-19.

[4] Groome PA, Schulze KM, Mackillop WJ, et al. A comparison of published head and neck stage groupings in carcinomas of the tonsillar region. Cancer 2001; 92:1484-1494.

[5] Lee AW, Foo W, Law SC, et al. Staging of nasopharyngeal carcinoma: From Ho's to the new UICC system. Int J Cancer 1999; 842:179-187.

[6] Begg CB, Cramer LD, Venkatraman ES, Rosai J. Comparing tumor staging and grading systems: a case study and a review of the issues, using thymoma as a model. Stat Med 2000; 19:1997-2014.

[7] Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: Application to the Cox regression model. Statistics in Medicine 1992; 11:20932109.

[8] Schemper M, Stare J. Explained variation in survival analysis. Statistics in Medicine 1996; 15:1999-2012.

[9] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine 1999; 18:2529-2545.

[10] Weiser MR, Gönen M, Chou JF, Kattan MW, Schrag D. Predicting survival after curative colectomy for cancer: individualizing colon cancer staging. J Clin Oncol 2011; 29:4796-802.

[11] Kasteleyn PW. The statistics of dimers on a lattice. Physica 1961; 27:1209-1225.

[12] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958; 53:457-481.

[13] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Journal of the American Medical Association 1982; 247:2543-2546.

[14] Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 1996; 15:361-387.

[15] Faraggi D, Simon R. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. Statistics in Medicine 1996; 15:2203-2213.

[16] Lausen B, Schumacher M. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. Computational Statistics and Data Analysis 1996; 21:307-326.

[17] Siegmund D. Confidence Sets in Change-Point Problems. International Statistical Review 1988; 56:31-48.

[18] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). JAMA 2001; 285:2486-97.