Published in final edited form as:

Clin Trials. 2017 December ; 14(6): 659–660. doi:10.1177/1740774517722141.

Comparisons of cancer staging systems should be based on overall performance in the population

Kathleen F Kerr¹, Michael LeBlanc², and Holly Janes²

¹Department of Biostatistics, University of Washington, Seattle, WA, USA

²Fred Hutchinson Cancer Research Center, Seattle WA, USA

Keywords

Cancer staging; prognosis; reclassification table; c-index

In a recent article in the journal, Kattan and Gerds propose a method for evaluating whether a new cancer staging system is more prognostic than the existing staging system.¹ The proposed method first checks whether both the “old” or established staging system and the “new” candidate staging system meet a minimum level of prognostic performance in the patient population. If they do, the proposed method then evaluates the prognostic performance of one staging system within levels defined by the other. Drawing on our background in evaluating the prognostic capacity of biomarkers, we caution that this approach does not reliably assess which staging system is more prognostic.

The Kattan-Gerds method relies on a “reclassification table” that cross-tabulates patients using the old and new staging systems (see Table 1 for an example). The method examines the prognostic performance of the new staging system within levels of the old staging system. (Or, if the new system has fewer levels than the old system, the old system is instead evaluated within levels of the new system.) If the new system meets a minimum level of overall population performance (its c-index exceeds 0.5) and is prognostic within every level of the old system, the Kattan-Gerds algorithm prescribes that the new system should be adopted as the more prognostic staging system.

Table 1 gives a simple, hypothetical example showing that a new system can be prognostic in every level of the old system but have inferior prognostic capacity in the population overall. In the example, the new System B is prognostic in each level of old System A (c-index = 0.71), yet the population prognostic capacity is lower for System B (c-index = 0.62) than for System A (c-index = 0.74). According to the Kattan-Gerds method, we should abandon the old system A and adopt the new system B, despite the fact that B is less prognostic overall than A. (In fact, the approach could have one going in circles, with each system deemed better than the other according to the proposed procedure.)

Our view is that a head-to-head comparison of the competing systems should be based on their performances in the patient population. One can choose one's preferred measure and compare that measure between the systems. For example, using the c-index favored by Kattan and Gerds for the example in Table 1, the old system A (c-index = 0.74) is superior to the new system B (c-index = 0.62) in the head-to-head comparison of population performance.

Table 1 gives the survival times of patients in each group so that readers can explore the example on their own. In addition, we provide an R function (<http://faculty.washington.edu/katiek/software.html>) that simulates data with a similar structure with exponential distributions of survival times. One can specify sample sizes to be arbitrarily large and get the same qualitative result. The results from applying the Kattan-Gerds method to the example in Table 1 are not attributable to small sample sizes or the particular choice of survival distribution. Nor do results depend on the choice of the c-index as the measure of prognostic capacity. For example, if prognostic capacity is measured by contrasting mean survival time across levels we arrive at the same conclusion: one system can be prognostic within every level of another system and yet be less prognostic overall.

We wonder if the concept of comparing cancer staging systems using a reclassification table was adopted from the field of risk prediction, where there is often an “old” or established risk prediction model and a candidate “new” or “updated” risk model (which often adds a new biomarker to the old model). In this context it has become common to cross-classify subjects with respect to “old” and “new” risk models using a reclassification table.^{2, 3} While the reclassification table is an interesting exploratory tool, we and others have argued against the notion that a new risk model should be evaluated by the degree to which it reclassifies individuals compared to the old risk model.^{4–6} In the end, the decision is which model or system to adopt for the relevant population, so we recommend that each candidate be evaluated in terms of its performance in that population.⁷

Acknowledgments

Funding: This work was supported by funding from the National Institutes of Health R01HL085757 to KFK; U10CA180819 and P01CA53996 to ML; and R01CA152089 to HJ

References

1. Kattan MW, Gerds TA. Comparison of two ordinal prediction models: A cancer staging system example. *Clin Trials*. 2015; 12:342–347. [PubMed: 25701155]
2. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115:928–935. [PubMed: 17309939]
3. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008; 27:157–172. discussion 207–212. [PubMed: 17569110]
4. Kerr KF, Wang Z, Janes H, et al. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology*. 2014; 25:114–121. [PubMed: 24240655]
5. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *Am J Epidemiol*. 2011; 173:1327–1335. [PubMed: 21555714]
6. Hilden J. Commentary: On NRI, IDI, and “good-looking” statistics with nothing underneath. *Epidemiology*. 2014; 25:265–267. [PubMed: 24487208]

7. Pepe, M., Janes, H. Methods for evaluating prediction performance of biomarkers and tests. In: Lee, M-LT, Gail, M., Pfeiffer, R., editors. Risk assessment and evaluation of predictions. Silver Spring: Springer Science and Business Media, LLC; 2013. p. 107-142.

Table 1
Two Hypothetical Staging Systems

The reclassification table cross-tabulates patients according to the two staging systems. The interior cells of the table give the survival times for patients in that cell. C-indices for each staging system within levels of the other staging system all exceed 0.5. However, we cannot conclude that either staging system is more prognostic than the other on this basis; this rationale would lead us to conclude that each staging system is superior to the other. A head-to-head comparison of the performance of the two systems is a better way to compare their prognostic capacity. From the head-to-head comparison we conclude that System A (c-index 0.74) is more prognostic than System B (c-index 0.62).

	SYSTEM B LEVEL 1	SYSTEM B LEVEL 2	C-index of B within Levels of A	C-index of B overall
SYSTEM A LEVEL 1	163, 173, 183, 193, 203, 213, 223, 233, 243, 253	112, 122, 132, 142, 152, 162, 172, 182, 192, 202	0.71	0.62
SYSTEM A LEVEL 2	61, 71, 81, 91, 101, 111, 121, 131, 141, 151	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	0.71	
C-index of A within Levels of B	0.76	0.76		
C-index of A overall	0.74			