American Joint Committee on Cancer Acceptance Criteria for Inclusion of Risk Models for Individualized Prognosis in the Practice of Precision Medicine

Michael W. Kattan, PhD; Kenneth R. Hess, PhD; Mahul B. Amin, MD; Ying Lu, PhD; Karl G.M. Moons, PhD; Jeffrey E. Gershenwald, MD; Phyllis A. Gimotty, PhD; Justin H. Guinney, PhD; Susan Halabi, PhD; Alexander J. Lazar, MD, PhD; Alyson L. Mahar, MSc; Tushar Patel, MD; Daniel J. Sargent, PhD; Martin R. Weiser, MD; Carolyn Compton, MD, PhD; members of the AJCC Precision Medicine Core

The American Joint Committee on Cancer (AJCC) has increasingly recognized the need for more personalized probabilistic predictions than those delivered by ordinal staging systems, particularly through the use of accurate risk models or calculators. However, judging the quality and acceptability of a risk model is complex. The AJCC Precision Medicine Core conducted a 2-day meeting to discuss characteristics necessary for a quality risk model in cancer patients. More specifically, the committee established inclusion and exclusion criteria necessary for a risk model to potentially be endorsed by the AJCC. This committee reviewed and discussed relevant literature before creating a checklist unique to this need of AJCC risk model endorsement. The committee identified 13 inclusion and 3 exclusion criteria for AJCC risk model endorsement in cancer. The emphasis centered on performance metrics, implementation clarity, and clinical relevance. The facilitation of personalized probabilistic predictions for cancer patients holds tremendous promise, and these criteria will hopefully greatly accelerate this process. Moreover, these criteria might be useful for a general audience when trying to judge the potential applicability of a published risk model in any clinical domain. CA Cancer J Clin 2016;66:370-374. © 2016 American Cancer Society.

Keywords: decision making, evidence-based medicine, patient preferences, personalized medicine

American Joint Committee on Cancer Staging Classification: Background From Inception to the Current 7th Edition

Since its inception in 1953, the American Joint Committee on Cancer (AJCC) has had the mission of developing and maintaining state-of-the-science anatomic staging systems for cancers and is the global leader in this endeavor. The staging system codifies principally the anatomic extent of disease at diagnosis, yet, overall, it remains the most accurate prognostic factor for solid malignancies. The system is based on three criteria: the local extent of the cancer within the site of origin (T), the degree of metastatic involvement of the regional lymph nodes (N), and the presence or absence of distant metastatic disease (M).
The stage of a cancer, otherwise known as the TNM stage grouping, is defined by one element from each staging criterion. For decades, the TNM staging system has been successfully deployed worldwide by the AJCC and its partner, the International Union for Cancer Control (UICC), and is the premier classifier for patients with solid tumors.

Over the years and through seven editions, this widely accepted system is appreciated because it is simple, easy to use due to its categorical character, and clinically useful to apply in patient management because it is associated with overall survival. As a general rule, stage groupings are based on a bin model, each bin containing one of the four subcategories of T, one of the three subcategories of N and one of the two subcategories of M. The bin model limits the number of elements that can be easily integrated, and adding even one additional element (either anatomic or nonanatomic prognostic element) creates unwieldy numbers of possible combinations.

The AJCC has increasingly recognized the growing need for more accurate and probabilistic individualized outcome prediction for precision medicine that would incorporate additional anatomic and nonanatomic prognostic factors beyond TNM. Toward this, since 2002 (6th edition), it has judiciously included, albeit very infrequently, nonanatomic factors that modified stage groupings. In the current 7th edition, the trend toward expansion of relevant markers to make clear treatment decisions included, for example, adding mitotic rate in gastrointestinal (GI) stromal tumors and melanomas, and prostate-specific antigen (PSA) level and Gleason score in stage grouping classification of prostate cancer.

The AJCC Staging System, 8th Edition: Building a Bridge From a Risk-Group–Based to a More Personalized Paradigm

To maintain the primary goal of the AJCC staging as the definitive, comprehensive, evidence-based, and clinically relevant classification, the editorial board of the 8th edition reaffirmed the anatomic stage and extent of disease basis of classification but with a deliberate effort to increasingly incorporate molecular biomarkers for accurate risk stratification that will help retain its fundamental role in defining a patient’s prognosis and appropriate disease management but will be enhanced by this new vision to guide precision therapy. The 8th edition, to be published in 2016, was expanded to include the Personalized Medicine Core (PMC) and the Evidence-Based Medicine and Statistics Core (EMBS).

The charter for the PMC is to provide Web-based outcome probability models for the major cancer sites that would include prognostic factors in addition to anatomic stage. This would allow more individualized prognosis calculation for patients whose disease fell within any given stage grouping. For the 8th edition, the following cancers will be first evaluated for existing prediction models: breast, colon, prostate, lung, melanoma, and head and neck cancer. We believe that this foundational effort will continue to be built upon beyond the 8th edition to comprehensively include all cancers where clinically relevant prediction models are applicable.

On January 23 and 24, 2015, the AJCC PMC committee met in Phoenix, Arizona to discuss this new initiative and agree on an approach to evaluating a statistical model of high quality that would be both useful and usable and would include tumor-related and patient-related prognostic factors in addition to anatomic stage. The committee postulated that a tool, such as a comprehensive Web-based prognostic model for each cancer site that covered all stages of disease, could possibly be built in modular fashion from existing and/or newly developed, high-quality models created for stage-specific use. A key issue for the PMC was to decide upon criteria for endorsement by the AJCC of any probability or risk model, either one to be built or an existing one, which reflected the commitment to quality and reliability. The PMC established criteria that it would use to judge existing risk models or require of newly built models, and the final result was a checklist of 16 items necessary for possible approval of an AJCC risk model.

The philosophy of the PMC was that validated predictive accuracy of a risk calculator or model was paramount. We believed it was important to underscore the complexity of validation and generalization. Each of these has various levels/types, such that true validation will require multiple data sets over various levels of time and place.

The Checklist

For purposes of illustration, we assume that a published prediction or risk model is under consideration for AJCC endorsement. Before applying the checklist, we felt the first step was to complete a brief form describing the prediction model.

Model Description Form

A. Describe the cancer that the patients being modeled have (eg, clinically localized prostate cancer) and whatever inclusion/exclusion criteria should be applied (eg, no prior cancer, untreated for prostate cancer).
B. Describe the diagnosis or treatment that defines the baseline or prognostic time zero, ie, at which time the outcome prediction will be calculated in future patients (eg, preoperatively, prior to potential treatment, at start of treatment).
C. List the predictors measured at baseline and how they were measured (eg, PSA by the Hybritech assay [ng/mL], stage by UICC 1992).
D. Choose the endpoint being predicted: overall survival, disease-specific survival (DSS), or disease-specific mortality (DSM).

E. Define the horizon time point being predicted (how far in time from baseline; eg, 10-year survival probability).

F. Describe how having this prediction might change clinical practice (eg, patients always ask about this outcome prior to choosing a specific treatment or, better, something actionable, such as, “if the prediction is <X, I would not recommend this treatment”).

For the checklist that follows, it is obvious that some items are possibly fixable by the developers/authors, some are not, and some are simply issues that were not reported (and, in that sense, are possibly fixable). So we anticipate some degree of back and forth as an author/developer is provided with his checklist result—perhaps the only failure item was something that was simply not reported and can easily be remedied, which would potentially result in endorsement.

Inclusion Criteria (Model Must Have All of These Characteristics)

1. The probability of overall survival, DSS, or DSM must be the outcome predicted. Overall survival is the endpoint consistent with the prior work of the AJCC and has the least methodological issues. However, clinicians and patients are generally more interested in DSS despite the potential difficulty in reliably assigning cause of death. Another drawback of the DSS endpoint lies in communicating the risk to the patient. DSS considers the only event to be death due to the cancer; patients who remain alive or who died of another cause are censored. This creates a hypothetical probability for the patient: his chances of surviving his particular form of cancer assuming he does not die of another cause first. Alternatively, a DSM probability is attractive as an outcome being predicted, because DSM provides the (nonhypothetical) probability of death due to a particular form of cancer and can appropriately handle competing risks. For this reason, DSM does not necessarily equal 1-DSS.

We realize that not allowing progression-free, disease-free, or recurrence-free survival-type endpoints is a narrow perspective that needs to be considered in the future. However, this will be complex due to the definitions of progression, frequency of assessment, etc.

2. The model should address a clinically relevant question—a prediction someone cares about. Is the treatment assumed by the prediction model relevant today? Do the inclusion and exclusion criteria in place for this model define a patient population of interest to the clinician? Is the endpoint interesting to predict? This is a somewhat subjective criterion best assessed by clinicians with disease management expertise.

3. At face value, the model should include the relevant predictors or explain why something relevant was not included. We do not want to endorse a prediction model that lacks a predictor that most clinicians would expect to see in this context. If the missing predictor was evaluated by the modeling team and removed because of a lack of incremental predictive ability, this is acceptable and does not constitute failure for endorsement. This subtle checklist item is likely best judged by clinicians with disease management expertise, potentially with a vote.

4. The model validation study should specify precisely which patients were used to evaluate the model and the validation data set’s inclusion/exclusion criteria. The end user of an endorsed prediction model needs to know whether the model is applicable to his or her patient. Therefore, we need a clear understanding of how patients were chosen for inclusion in the validation data set(s). This is the only way to know whether the end user’s patient would have qualified for the validation study.

5. The model should be assessed for generalizability and external validation. The key assumption being made with endorsement is that a risk model is valid in future patients. Because validation first requires patient follow-up and data analysis, validation cannot be known immediately. We must rely upon presently available studies to be comfortable with validation. Those studies should separate reproducibility (ie, simply new patients) from transportability (ie, patients who differ in one or more respect from the development data set). In general, true validity assessment will likely require multiple studies over time. In the interim, state-of-the-art internal and internal-external validation procedures are recommended.

6. The model should have a well-defined prognostic time zero. (What starts the clock/when is prediction calculated?) A prediction is calculated at a certain time in a patient’s course of illness. What time this is must be obvious for any model. Examples include immediately after diagnosis, before a certain treatment, and immediately after a particular treatment.

7. All predictors must be known at time zero and sufficiently defined for someone else to use. Clearly, the end user needs to know what to enter in the risk model. This should be clear before endorsement is granted.

8. Sufficient detail must be available to implement the model (need the equation, not a crippled version or simple, not yet validated score chart) OR the author
must allow free access to the model. If a prediction model is otherwise a candidate for endorsement, the AJCC would typically like access to the underlying equation/formula. However, this would not be required if the developer provides a free online risk model that still has the underlying equation/formula. Note that availability or access refers to the actual model that was validated, and the developer must notify AJCC if they modify the online model or calculator from the published validation.

9. A measure of discrimination must have been reported. This is often measured as the concordance index and needs to be assessed on the validation data set(s).

10. Calibration in the small must be assessed (from the external validation data set) and provided. Calibration in the small is a plot of predicted probability versus observed proportion. Here, we want to see that, across the spectrum of predictions, the observed proportion of deaths closely resembles the predicted probability in the external data set.

11. The model should be validated over a time frame and in a practice setting that are relevant to contemporary patients with disease. The validation data set should be reflective of the patient being evaluated today. The treatment(s) applied in the validation data set should be similar to those used today, and the clinician should be comfortable that the disease reflected in the validation data set resembles the disease he or she sees today. Furthermore, the setting in which the validation was performed should be similar to the setting of interest to the clinician today.

12. It should be clear which initial treatment(s), if any, were applied, and with what frequency. The initial treatment need not be a specific predictor, nor must the model be restricted to a single treatment. In other words, the prediction model does not need to include the treatment type as a predictor, but it can. We just need to know how patients in the validation data set were treated (i.e., the proportions who received each form of therapy, if multiple initial therapies were present). Adjuvant treatments should be described but should be ignored as predictors in the model.

13. The development and/or the validation of the prediction model must appear as a peer-reviewed journal article. The reference(s) needs to be provided.

Exclusion Criteria (Any of These Exclude a Model From Consideration)

1. A substantial proportion of patients had essentially no follow-up (either missing entirely or very short censored follow-up) in the validation data set. This item is intentionally on the subjective side. However, the concern here is that selection bias may be introduced essentially by excluding many patients because of missing outcome data, and they may be quite different from the rest of the patients who were included. Those with missing follow-up need to be compared with the rest.

2. No information is available on the number of missing values in the validation data set. Virtually any data set will have missing values. When a data set appears to have no missing values for any variables, it is often the case that those observations with missing values were excluded previously, before the analysis. This may create large bias. As such, we need to see the true time frame for patient accrual into the validation data set to understand the extent of missingness.

3. The number of events in the validation data set is small. This is a relatively unexplored literature to make firm suggestions on how small is small. However, 100 events may be the minimum needed.

Discussion

The goal of the AJCC PMC checklist was to establish criteria for the endorsement of an online risk model or calculator. Satisfying the checklist does not establish when or how the checklist should be used. Rather, the checklist sets minimum quality criteria necessary for the risk model to be considered further for possible application to prognostic tools developed by the AJCC. Similarly, the checklist is not useful for comparing one risk model versus another, because both must meet all criteria (i.e., endorsed risk models will be equal according to these criteria).

At some point, the AJCC may host a Website of all endorsed models. Alternatively, the AJCC may provide an updated listing of endorsed models. The exact approach has not been decided. Our criteria complement the recently published transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) document. The TRIPOD document sets forth the details deemed necessary for reporting on what was done in a prediction modeling study: development or validation. Our checklist mirrors these requirements, although is not intended as another reporting checklist but rather as a checklist with minimal requirements for prediction models in cancers to become endorsed by end users.

References

1. Gospodarowicz M, Benedet L, Hutter R, Fleming I, Henson DE, Sobin LH. History and international developments in cancer staging. Cancer Prev Control. 1998;2:262-268.

2. Asare EA, Washington MK, Gress DM, Gershenwald GE, Greene FL. Improving the quality of cancer staging. CA Cancer J Clin. 2015;65:261-263.

3. Gershenwald JE, Soong SJ, Balch CM; American Joint Committee on Cancer (AJCC) Melanoma Staging Committee. 2010 TNM staging
system for cutaneous melanoma...and beyond. Ann Surg Oncol. 2010;17:1475-1477.

4. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;44:201-209.

5. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130:515-524.

6. Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Stat Med. 1999;18:695-706.

7. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation [published online ahead of print April 18, 2015]. J Clin Epidemiol. doi: 10.1016/j.jclinepi.2015.04.005.

8. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247:2543-2546.

9. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study [published online ahead of print November 9, 2015]. Stat Med. doi: 10.1002/sim.6786.

10. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1-W73.