Published in final edited form as:

Ann Surg Oncol. 2018 August ; 25(8): 2172–2177. doi:10.1245/s10434-018-6370-4.

Variability in Predictions from Online Tools: A Demonstration Using Internet-Based Melanoma Predictors

Emily C. Zabor, MS^{1,*}, Daniel Coit, MD², Jeffrey E. Gershenwald, MD³, Kelly M. McMasters, MD, PhD⁴, James S. Michaelson, PhD⁵, Arnold J. Stromberg, PhD⁶, and Katherine S. Panageas, DrPH¹

¹Memorial Sloan Kettering Cancer Center, Department of Epidemiology and Biostatistics

²Memorial Sloan Kettering Cancer Center, Department of Surgery

³MD Anderson Cancer Center, Department of Surgical Oncology

⁴University of Louisville, Department of Surgical Oncology

⁵Massachusetts General Hospital, Laboratory for Quantitative Medicine

⁶University of Kentucky, Department of Statistics

Abstract

Background—Prognostic models are increasingly being made available online, where they can be publicly accessed by both patients and clinicians. These online tools are an important resource for patients to better understand their prognosis and for clinicians to make informed decisions about treatment and follow-up. The goal of this analysis was to highlight the possible variability in multiple online prognostic tools in a single disease.

Methods—To demonstrate the variability in survival predictions across online prognostic tools, we applied a single validation dataset to three online melanoma prognostic tools. Data on melanoma patients treated at Memorial Sloan Kettering Cancer Center between 2000 and 2014 were retrospectively collected. Calibration was assessed using calibration plots and discrimination was assessed using the *C*-index.

Results—In this demonstration project, we found important differences across the three models that led to variability in individual patients' predicted survival across the tools, especially in the lower range of predictions. In a validation test using a single institution data set, calibration and discrimination varied across the three models.

Conclusions—This study underscores the potential variability both within and across online tools, and highlights the importance of using methodological rigor when developing a prognostic model that will be made publicly available online. The results also reinforce that careful development and thoughtful interpretation, including understanding a given tool's limitations, are required in order for online prognostic tools that provide survival predictions to be a useful resource for both patients and clinicians.

*Corresponding author: Emily C. Zabor, Memorial Sloan Kettering Cancer Center, Department of Epidemiology and Biostatistics, 485 Lexington Ave, 2nd floor, New York, NY 10017, zabore@mskcc.org, 646-888-8299.

Introduction

Cancer staging systems have historically been used to stratify patients into risk groups with fixed ranges of predicted survival outcomes. For example, the American Joint Committee on Cancer (AJCC) melanoma staging system initially classified patients based on tumor, nodal, and metastatic stage (TNM staging). But the AJCC staging system, and staging systems more generally, has evolved over time to incorporate additional patient characteristics in an effort to address the fact that anatomic TNM staging is not very useful for predicting prognosis of individual patients. The 7th edition of the AJCC melanoma staging system utilizes numerous patient characteristics in addition to anatomic information to divide patients into four main stage groups, or nine substage groups.¹ As expected, survival still differed by stage, but there was also significant heterogeneity in survival described by substage, especially among patients with stage III disease. Furthermore, there was substantial overlap in survival estimates across stages I, II, and III disease even after inclusion of these additional patient characteristics.

In an effort to overcome some of the problems with survival estimates derived from staging systems, nomograms and other prognostic models have been gaining in popularity as a way to predict an outcome for an individual patient based on a collection of variables. Often, corresponding internet-based tools have been made available for easy access by patients and providers. These online prediction tools are designed to help providers and patients assess individual risk and make informed decisions about treatment and follow-up. Online prediction tools are particularly relevant in settings where standard of care is rapidly evolving, such as in melanoma, where effective adjuvant therapies are now treatment options for patients with stage III and IV resected melanoma.^{2,3}

Rabin and colleagues recently conducted a systematic review of prognostic tools across all cancers and identified 107 calculators for use in 89 different cancers.⁴ With one exception, all tools were available online. A recent systematic review that focused specifically on melanoma prognostic models identified 17 tools, five of which were available as internet-based calculators.⁵ Three of these melanoma prognostic models are currently available online. One group analyzed the same AJCC data that were used in the creation of the 2009 AJCC staging system (AJCC predictor).⁶ Another group analyzed data from the Sunbelt Melanoma Trial (Sunbelt predictor).⁷ And a third group used data from the Massachusetts General Hospital (MGH) (MGH predictor).^{8–10} Despite improvements in prediction accuracy over staging systems, use of different statistical methodology and inclusion of varied patient characteristics can lead to variation in predictions from online tools.

Prognostic models made available online for public use warrant careful examination and require methodological rigor to ensure that clinicians obtain clinically useful predictions when they access these tools to make treatment decisions and inform conversations about prognosis with their patients, and that patients get accurate information when they access these tools to directly obtain survival estimates. To demonstrate the potential differences in prognostic models in a single disease site, in this study we compare the input variables and modeling procedures of three internet-based melanoma prognostic models, and evaluate model performance by applying a single external data set to all three tools. In doing so we

sought to critically assess the development and deployment of online prognostic tools more generally.

Internet-Based Predictors in Melanoma

Table 1 presents details of the underlying models and input parameters for the three internet-based melanoma predictors studied here. The AJCC predictor can be found at melanomaprognosis.net.⁶ The final predictions were generated from a series of multivariable Cox regression models, stratified by tumor thickness. The reported concordance correlation coefficients (CCCs) for the 5- and 10-year MSS rates were 0.90 and 0.93, respectively.

The Sunbelt predictor can be found at MelanomaCalculator.com.⁷ The model is only generalizable to patients aged 18 to 70 years with primary tumor size ≤ 1 mm who were staged by SLN biopsy. A final multivariable model was constructed using Cox regression for overall survival (OS) with stepwise variable selection. The reported CCC for 5-year OS was 0.98. The online Sunbelt predictor produces OS, disease-free survival (DFS) and locoregional recurrence-free survival (LR-RFS) predictions. In addition, after communication with the study statisticians, it was revealed that the model used in the online predictor differs somewhat from the model presented in the manuscript, in that age and Breslow thickness were included as continuous covariates, and a quadratic term for Breslow thickness was also included, in the online predictor. Table 1 reflects the variables utilized in the internet-based predictor rather than those reported in the original publication.⁷

The MGH predictor can be found at lifemath.net/cancer/melanoma.^{8–10} The modeling technique employed is called the “binary-biologic model of cancer metastasis” and is based on a theory of the underlying spread of cancer cells.¹⁰ Simply put, this model assumes that the probability of cancer death increases with each additional cancer cell present and that prognostic factors have a multiplicative impact on risk of cancer death. The reported R^2 value was 0.96 for predicting 15-year MSS. Although the online predictor generates only MSS predictions, predicted OS is also available from this model, as it is used to determine life expectancy by the online version.

Data and Methods

Data on all melanoma patients who presented to Memorial Sloan Kettering Cancer Center (MSKCC) from January 2000 through December 2014 were extracted from a retrospectively maintained database (n=13,853) after Institutional Review Board approval. Patients were excluded if they presented to MSKCC > 120 days after initial diagnosis (n=4,221), had stage 0 (n=2,680) or unknown stage (n=34) disease at presentation, or had Clark level I invasion tumors (n=7), leaving a total of 6,911 patients eligible for analysis.

Figure 1 shows a Venn diagram of the sample sizes available to validate each predictor. The Sunbelt predictor applies only to patients between 18 and 70 years of age with tumors ≤ 1 mm who were staged using SLN biopsy. After limiting the MSKCC data based on these conditions, and to patients with complete data on the included predictors, 1500 patients remained for analysis. Median follow-up time among survivors was 3.4 years (IQR: 1.1–6.7), and during that time there were 293 deaths from any cause and 374 recurrences or

deaths from melanoma. Baseline survival and parameter estimates for the Sunbelt predictor were obtained from the study statisticians⁷ and were used to generate predicted 5-year OS and DFS, a composite of recurrence or death from melanoma. Patients were grouped into quartiles based on these predictions, and the Kaplan-Meier method was used to estimate 5-year observed OS and DFS within each quartile.

The MGH predictor applies to all patients with data on tumor size, resulting in a sample size of 6873 patients for analysis. Median follow-up time among survivors was 2.1 years (IQR: 0.1–17.1), and during this time there were 1460 deaths from any cause, 636 of which were from melanoma. Source code for the MGH predictor was obtained from the online predictor site and converted into R code in order to generate predicted 5-year OS and MSS for each patient. Patients were grouped into quartiles based on these predictions, and the Kaplan-Meier method was used to estimate 5-year observed OS and MSS within each quartile.

For the AJCC predictor, data were first limited to patients presenting in 2009 or later (n=3034) because MSKCC contributed data to the development of the AJCC model, which included patients diagnosed through 2008. We next limited to those diagnosed with localized disease (n=2,452) and with complete information on the included variables (n=2,291). Median follow-up time among survivors was 1.4 years (interquartile range [IQR]: 0.1–3.9), and during this time, 57 patients died from melanoma. Because the AJCC predictor is based on combinations of binary variables, only 38 distinct predictions are possible. To obtain predictions from the online tool, we simply entered data for each of the 38 possible values. Patients were grouped into quartiles based on these predictions, and the Kaplan-Meier method was used to estimate 5-year observed MSS within each quartile.

Calibration, a measure of agreement between predictions and observed data, was assessed for each model by plotting predicted versus observed survival. Discrimination, a measure of how accurately patients were classified as alive versus dead at 5 years, was evaluated using an inverse probability censoring weighted concordance-index (C-index). All statistical analyses were conducted using R software version 3.4.1¹¹ including the “survival” and “pec” packages.

Results

In all plots perfect calibration is indicated by the dotted 45-degree line; a higher C-index indicates better discrimination. Calibration and discrimination results for the Sunbelt predictor are shown in Figure 2a for DFS and Figure 2b for OS. Calibration is quite good for OS, with most points falling almost directly on the 45-degree line, whereas for DFS, at the lower end of the range the 95% confidence intervals do not cross the 45-degree line, suggesting that calibration is poor in the lower range of predicted DFS. The Sunbelt predictor has discrimination of 0.754 for DFS and 0.767 for OS, suggesting that the model differentiates well between those who have and those who do not have the event for both endpoints.

Calibration and discrimination results for the MGH predictor are shown in Figure 3a for MSS and Figure 3b for OS. We see that the MGH predictor results in a higher range of MSS

rates than those of the Sunbelt predictor. For MSS, the MGH predictor does not appear to be well calibrated, and it results in a poor C -index 0.542. On the other hand, whereas most of the predicted 5-year OS values are slightly lower than the estimated values (indicated by the points being above the 45-degree line), the confidence intervals are tight and the points follow a straight-line pattern. A C -index of 0.767 for OS using the MGH predictor indicates good discrimination and is comparable with the OS discrimination of the Sunbelt predictor.

Finally, calibration and discrimination results for the AJCC predictor are presented in Figure 4. Although we obtain a very high C -index of 0.836 for discrimination with the AJCC predictor, we see that the model results in large confidence intervals, suggesting a great deal of uncertainty in the model calibration.

Discussion

Generating accurate predictions should be the most important goal when building a prediction model. Whereas most prediction models are disseminated through academic publications and therefore primarily reach an audience with technical and subject-area expertise, prediction models made available online are broadly available to the public, and therefore require additional scrutiny. Using three online melanoma predictors as an example, we found that some models demonstrated acceptable calibration and discrimination in a validation test using a single institution data set, others were not as accurate or introduced a large amount of uncertainty, either overall within a specific range of predictions.

The models compared here were all developed by respected research teams that included experienced statisticians and used appropriate statistical methodology, including model development in a training data set followed by model validation in a separate external data set. In all cases a measure of correlation between the predicted and observed survival estimates was presented to assess the calibration of each model. However, none of the original reports for these three predictors^{6–8} included a measure of discrimination, such as the C -index that we evaluated here. The three models also have several important distinctions. Both the AJCC predictor and the Sunbelt predictor are based on a Cox regression model. However, the AJCC predictions are generated from Cox regression models with only a few binary predictors and are stratified by tumor size (as a categorical variable), so only a finite number of predicted values is possible, thereby limiting the accuracy of predictions by design.⁶ In contrast, the Sunbelt predictor is based on a single multivariable Cox regression model that incorporates continuous predictors, allowing for a wider range of resulting predicted values.⁷ The MGH predictor is not based on a Cox regression model, but rather on a novel approach that seeks to predict survival based on the probability of the spread of individual tumor cells.^{9,10} The MGH predictor is unique in its ability to handle missing values, and results in predictions that are similar to the other two models in many cases despite the divergent methodology. It is important to note that this analysis is limited by somewhat small sample sizes and relatively short follow-up time, and results are intended for use as an illustrative example.

This example validation of three online melanoma predictors demonstrates the challenges faced when working to develop a prediction model that will be made publicly available on

the internet. In 2016 the 8th edition of the AJCC staging manual was expanded to include the Personalized Medicine Core, tasked with providing online probability models that would incorporate prognostic markers in addition to anatomic stage to provide more personalized predictions. In preparation for this new undertaking, a checklist of sixteen criteria that should be used to evaluate a prognostic model for endorsement by the AJCC was created.¹² The checklist highlights the importance for researchers to keep in mind that the ultimate goal of a prognostic model is to provide doctors and patients alike with accurate measures of predicted survival at a specific timepoint so that results can be used for personalized decisions about treatment and follow-up. These predictions should ideally come from a model that incorporates only clinically important variables that are widely available to patients with the disease of interest. If a variable is not widely available, or if a new prognostically relevant variable is discovered, then the model needs a way of handling missing data to ensure broad applicability. The model also needs to report a measure of uncertainty, such as a confidence interval or standard deviation, so that those accessing the predictions have an idea of the range of possible values. In addition, as emphasized in a recent review by Alba and colleagues¹³, the model needs to be critically assessed in an appropriate external data set for measures of both calibration and discrimination. The limitations of a prediction model should be reported alongside the online predictor. Interestingly, none of the three online tools assessed herein would have passed the checklist criteria simply for lack of a reported measure of discrimination.

A recommendation as to which of these 3 melanoma tools should be used would require additional analyses of clinical utility that are outside the scope of this paper. Instead, we used this example validation study to demonstrate potential differences in prognostic models for a single disease site. As clinicians it is important to be aware of the online tools that are available in one's field, and to know that patients may have already accessed survival predictions online at the time of initial evaluation. Given the predictive variability both within and across predictors, it is important to understand the limitations of online predictors, and to be able to synthesize information across tools to give patients a better idea of the possible range of survival times they are facing. Careful development and thoughtful interpretation are required in order to ensure that online prognostic tools that provide survival predictions will be a useful resource for both patients and clinicians.

Acknowledgments

This work was supported with funding from the Core Grant (P30 CA008748).

References

1. Balch CM, Gershenwald JE, Soong SJ, et al. Final version of 2009 AJCC melanoma staging and classification. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2009; 27(36):6199–6206. [PubMed: 19917835]
2. Long GV, Hauschild A, Santinami M, et al. Adjuvant Dabrafenib plus Trametinib in Stage III BRAF-Mutated Melanoma. *The New England journal of medicine*. 2017
3. Weber J, Mandala M, Del Vecchio M, et al. Adjuvant Nivolumab versus Ipilimumab in Resected Stage III or IV Melanoma. *The New England journal of medicine*. 2017
4. Rabin BA, Gaglio B, Sanders T, et al. Predicting cancer prognosis using interactive online tools: a systematic review and implications for cancer care providers. *Cancer epidemiology, biomarkers &*

prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2013; 22(10):1645–1656.

5. Mahar AL, Compton C, Halabi S, et al. Critical Assessment of Clinical Prognostic Tools in Melanoma. *Annals of surgical oncology*. 2016; 23(9):2753–2761. [PubMed: 27052645]
6. Soong SJ, Ding S, Coit D, et al. Predicting survival outcome of localized melanoma: an electronic prediction tool based on the AJCC Melanoma Database. *Annals of surgical oncology*. 2010; 17(8): 2006–2014. [PubMed: 20379784]
7. Callender GG, Gershenwald JE, Egger ME, et al. A novel and accurate computer model of melanoma prognosis for patients staged by sentinel lymph node biopsy: comparison with the American Joint Committee on Cancer model. *Journal of the American College of Surgeons*. 2012; 214(4):608–617. discussion 617–609. [PubMed: 22342785]
8. Chen LL, Nolan ME, Silverstein MJ, et al. The impact of primary tumor size, lymph node status, and other prognostic factors on the risk of cancer death. *Cancer*. 2009; 115(21):5071–5083. [PubMed: 19658184]
9. Michaelson JS, Chen LL, Silverstein MJ, et al. Why cancer at the primary site and in the lymph nodes contributes to the risk of cancer death. *Cancer*. 2009; 115(21):5084–5094. [PubMed: 19670457]
10. Michaelson JS, Chen LL, Silverstein MJ, et al. How cancer at the primary site and in the lymph nodes contributes to the risk of cancer death. *Cancer*. 2009; 115(21):5095–5107. [PubMed: 19670458]
11. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2017.
12. Kattan MW, Hess KR, Amin MB, et al. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. *CA: a cancer journal for clinicians*. 2016; 66(5):370–374. [PubMed: 26784705]
13. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *Jama*. 2017; 318(14):1377–1384. [PubMed: 29049590]

Synopsis

This analysis demonstrated the variability in individual patients' predicted survival from different online melanoma prediction tools, and highlights the need for more rigorous development of prognostic models that will be made available online and for careful interpretation of results, accounting for a given tool's limitations.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
Venn diagram of available validation sample sizes for each predictor

Figure 2.
Calibration Plots for a) DFS and b) OS for the Sunbelt predictor

Figure 3.
Calibration Plots for a) MSS and b) OS for the MGH predictor

Figure 4.
Calibration Plot for MSS for the AJCC predictor

Table 1

Comparison of Data, Modeling Strategy, and Input and Output variables for three Online Melanoma Predictors.

Predictor	AJCC predictor melanomaprognois.net	Sunbelt predictor melanomacalculator.com	MGH predictor Cancer.lifemath.net
Disease stage included	Stage I and II	Staged by SLN ^a biopsy	All stages
Training data	n=14,760	n=2,507	n=2,770
Validation data	n=10,974	n=1,001	n=20,215
Model	Cox regression	Cox regression	Binary-biological model of cancer metastases
Endpoint(s)	Melanoma-specific survival	Overall survival, Disease-free survival and Recurrence-free survival	Melanoma-specific survival, Death from other causes, Overall survival
Estimated survival times	1-, 2-, 5-, and 10-years	5-years	15-years
Confidence intervals	Yes	Yes	No
Included factors			
Ulceration	No, Yes	No, Yes	No, Yes
Lesion site	Extremity, Non-extremity	Extremity, Non-extremity	External ear, Face, Scalp or neck, Trunk, Upper limb/shoulder, Lower limb/hip
Tumor thickness(mm)	0.50, 0.51–1.00, 1.01–2.00, 2.01–3.00, 3.01–6.00, >6.00	Continuous and quadratic term	Continuous
Patient age (yr)	<60, 60	Continuous	----
Sex	----	Male, Female	Male, Female
Nodal tumor burden	----	Positive SLN, Negative SLN	----
Number of positive nodes	----	1, >1 (SLNs); 0, 1 (non-SLNs)	Continuous
Clark Level	----	----	I, II, III, IV, V
Histology	----	----	Desmoplastic, Nodular, Acral lentiginous, Superficial spreading, Lentiginous, Spindle, Epithelioid, Epithelioid-spindle, Regressing, Giant pigmented nevus

^aSLN = sentinel lymph node