A GENETIC SIGNATURE PREDICTING SURVIVAL AND METASTASIS FOR MELANOMA PATIENTS

J. B. NATION

This short note presents a 2-gene signature that divides later stage melanoma (SKCM) patients into low-risk and high-risk groups for survival and recurrence/metastasis.

Method. The TCGA database has mRNA expression and clinical data for 104 melanoma patients. Of these, 102 were diagnosed at stage 2, 3, or 4, with the remaining 2 at stage 1. The gene expression matrix was preprocessed by log transform, row-centering, and quantile-normalization. The LUST algorithm, which can be found at the website https://github.com/tristanh314/lust-cancer-2019, was used to obtain the signature.

LUST is a two-step algorithm. The first step looks for signals in the expression data using a version of association rules; it is unsupervised by clinical outcomes. The second step restricts the expression matrix to a set of genes identified by the first step, and looks for small subsets that are associated with survival. These candidate signatures are then tested to see if they have predictive value.

For a given signature, every patient is assigned a score that is a linear combination of his/her expression of those genes, by projecting the patient's expression onto the first singular value of the expression matrix for those genes only.

Next, a time T is chosen as the target survival date. Patients who die before T days from diagnosis are considered short survivors, while those living more than T days are long survivors. The choice of T depends very much on the disease and patient population (e.g., stage).

A threshold θ is determined to divide the patients into high-risk and low-risk groups. Patients with a score $s < \theta$ are in the high-risk group, while those with a score $s \geq \theta$ are in the low-risk group. (Depending on the nature of the genes in the signature, this could be reversed.) For this note, the threshold was chosen to maximize the fraction of patients assigned to the correct risk pool. This measure is the *accuracy* of the predictor, and is one of several options. See [1] for a discussion of various measures of predictors.

Date: September 30, 2019.

Results: survival. For SKCM, the signature A5 consists of 2 genes relating to immune response: CD48, SLAMF6. The threshold $\theta = -0.55$ gave the maximum accuracy. Patients with a score lower than the threshold, representing low expression of the genes in the signature, were assigned to the high-risk group.

For survival, we used 550 days (roughly 1.5 years) as the criterion for long survival. The results are shown in Figure 1 and Table 1. Of the 104 patients, 61 either died before 550 days or survived at least that long. The remaining 43 in the study, who were still alive but less than 550 days after diagnosis at the time of their last follow-up, were censored. (The censored group included both stage 1 patients.)

In Figure 1, the blue curve indicates the patients' scores, which are arranged in ascending order, as indicated on the left vertical axis. The vertical green line is at the threshold $\theta = -0.55$: patients to the left of the green line had lower scores, patients to the right had higher scores. A red "+" sign indicates a patient who died, at a time indicated on the right vertical axis. Blue circles correspond to surviving patients (longer than 550 days) at the time of their last follow-up.

These results are to be interpreted as follows. Patients with a low score constitute the high-risk group. Of those 19 patients, 12 died before 550 days (63%). Patients with a high score constitute the lower-risk group. Of those 42 patients, 8 died before 550 days (19%). The overall accuracy (patients placed in the correct group) is then 46 ($= 12 + 34$) out of 61 (75%).

The Kaplan-Meier curves for the high-risk and low-risk groups are given in Figure 2, with the p-values for the logrank and Cox tests.

	low score	high score
long survival	7	34
short survival	12	8

TABLE 1. Signature A5: survival. A low score is $s < -0.55$ based on the expression of the genes CD48 and SLAMF6; a high score is $s \geq -0.55$. Long survival is ≥ 550 days, while short survival is < 550 days.

Results: disease-free survival. We can do the same analysis, but this time using a return of the disease (either distant metastasis or local recurrence) or death as the termination event. For simplicity, we use the same signature (denoted A5), time (550 days), and threshold. The results are given in Figure 3 and Table 2. Note that fewer patients

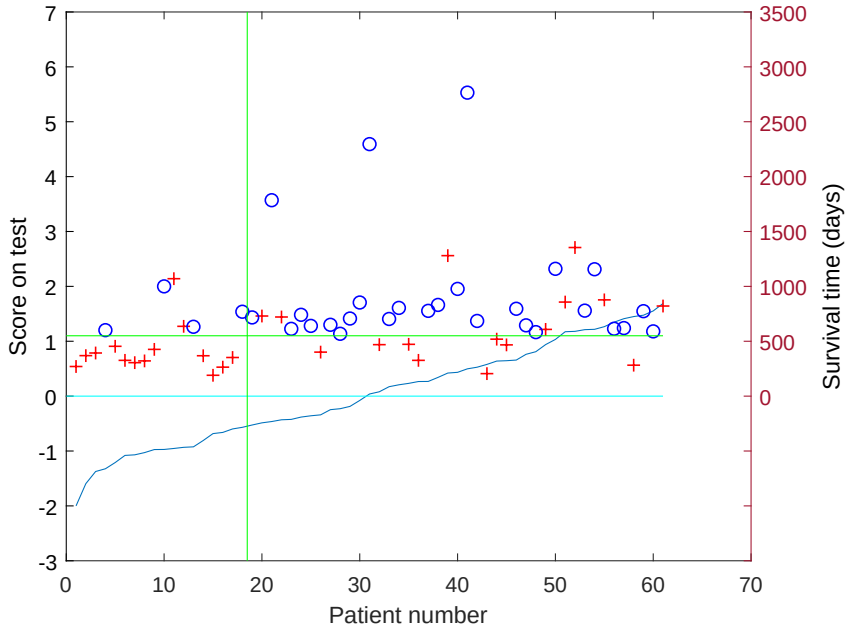

FIGURE 1. Signature A5 survival vs. score (blue curve). The vertical green line is at the threshold score -0.55, the horizontal green line at 550 days. A red + indicates death, while a blue $\circ$ is the survival time at last follow-up. Censored patients (those alive at last follow-up $<$ 550 days) are not shown.

are censored, being only those who did not have recurrence or death before 550 days (69 patients).

In the high-risk group, 19 of 22 patients had recurrence or death before 550 days (86%). In the lower-risk group, it was 18 out of 47 (38%). The overall accuracy was 70% (48 out of 69).

The Kaplan-Meier curves for disease-free survival of the high-risk and low-risk groups are given in Figure 4.

	low score	high score
long disease-free survival	3	29
short disease-free survival	19	18

TABLE 2. Signature A5: disease-free survival, that is, time until death or local recurrence or distant metastasis.

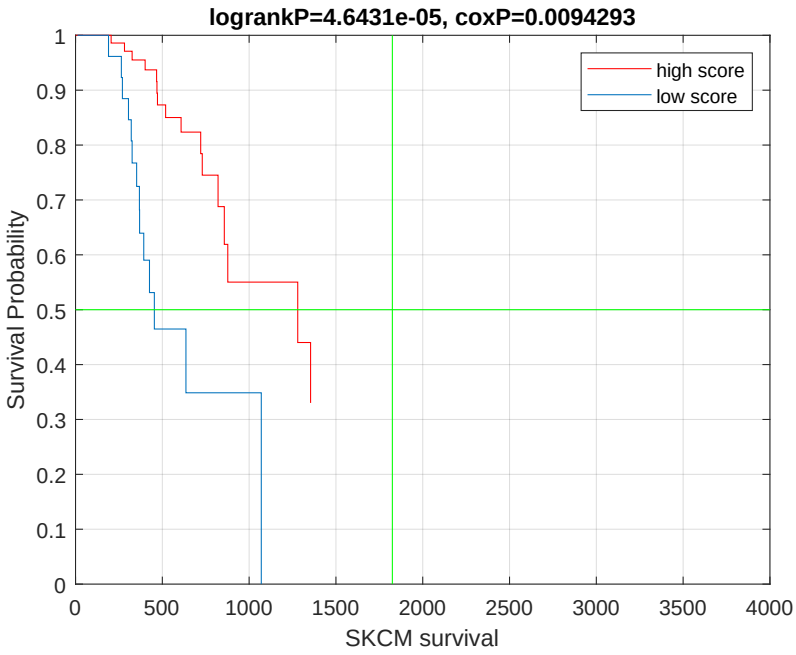

FIGURE 2. Kaplan-Meier curves for the high-risk and low-risk groups with respect to survival

Discussion. This signature is very good at identifying high-risk patients. Presumably, these patients might warrant more aggressive treatment. It would be particularly interesting to see if the signature is valid for earlier stage melanomas.

The threshold was chosen to optimize overall accuracy. The cost is that for survival, only 60% of the short-term survivors were identified as high-risk, and for disease-free survival, just over 50%. In traditional terms, the test is specific but not very sensitive. Raising the threshold would put more patients into the high-risk group, including more short-term survivors, at the cost of more false positives. The choice of the threshold, which depends on the weighting of the weighting of the accuracy function, is a medical question. Or perhaps a better signature can be found, or a combination with other factors to assess patient risk. Attempts to incorporate lymph node involvement (the n-score) into the predictor yielded no improvement.

The genes CD48 and SLAMF6 are part of a larger family regulating the differentiation and activation of a wide variety of immune cells (GeneCards). Thus our signature measures the strength of some aspect of the patients immune response to the cancer; those with a poor response, as reflected by low expression of these immune system genes,

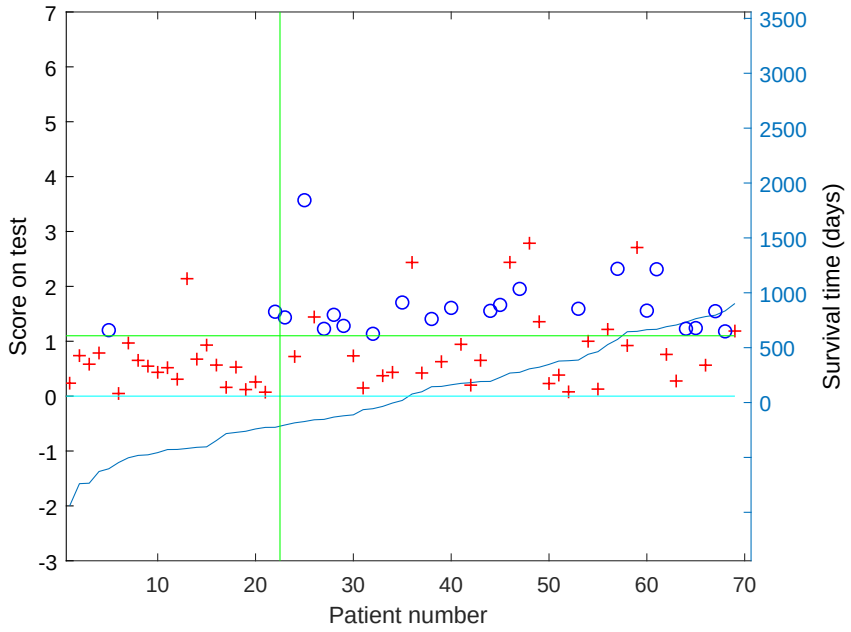

FIGURE 3. Signature A5 disease-free survival. For this graph, a red + indicates death or local recurrence or distant metastasis.

are at higher risk. Adding the related gene SLAMF1 to the signature gives nearly identical results, and might make the signature more robust.

Other signatures obtained by LUST from the immune system genes give results that are close to these. The signatures that we have tested that measure gene expression for other factors (not immune response) are only very weakly predictive of survival or recurrence/metastasis.

REFERENCES

[1] J. M. Hughes-Oliver, Assessment of prediction algorithms for ranking objects. Notices Amer. Math. Soc., Feb. 2019, 182–190.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF HAWAI'I, HONOLULU, HI 96822, USA

Email address: jb@math.hawaii.edu

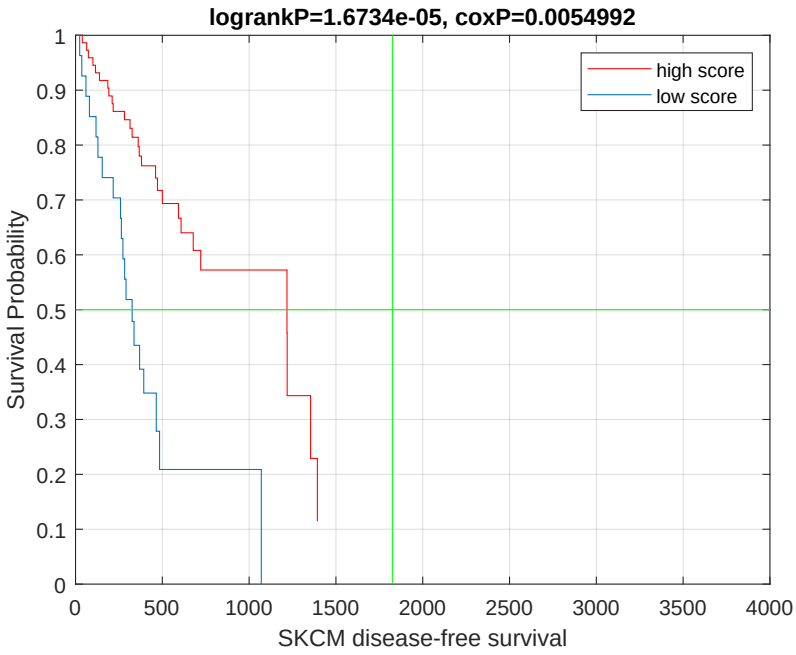

FIGURE 4. Kaplan-Meier curves for the high-risk and low-risk groups with respect to disease-free survival.