SUPPLEMENTARY MATERIAL

Supplemental Methods

Classification and Regression Algorithms

In order to develop PRiMeUM we evaluated the following metastasis prediction approaches and models. First, we considered univariate and multivariate cox models as described in the main text. Next, we considered the prediction task as both a regression problem and a classification as described below.

For regression, we first tested a Cox semiparametric model,[1] and Accelerated Failure Time (AFT),[1, 2] both standard methods in the field. The Cox semiparametric model resulted with a 83% concordance (COX specific accuracy result, comparable to AUC in ROC plots), showing slightly poorer results in prediction of metastasis before 48 months than the final PRiMeUM model described in the main text. However, it appeared unstable when bootstrapping training sets, with concordance varying between 77% - 83%. The observed variation is likely due to the limited amount of positive labels for some of the features. Compared to COX, the AFT model gave inferior and even less stable results, again likely due to the sparsity of the features. This survival analysis was done using two different implementations lifelines (in python) and MATLAB Cox and AFT functions.

To frame the prediction task as a classification problem, we used the clinical criteria of metastasis within 48 month. A similar performance of the AUC metric of 84% was observed when we used a 24 month cutoff (data not shown). There were 46 cases where metastasis occurred after 48 months (range=49 to133 months) which were not included in the training data as their label was not well defined under this modeling assumption. The option of including them as metastasis-negative at 48 months was considered misleading due to the possibility that metastasis had occurred but was as yet undetected. The alternative approach of not including these 48 tumors with metastasis in the model resulted in a more accurate result relative to the remainder of the samples,

For classification, we evaluated the performance of decision trees, random forests, a mixture of decision trees trained using boosting[3], and logistic regression with and without sparseness control (lasso). All algorithms were evaluated using the Sci-kit llearn package,[4] and the mixture of decision trees was also tested using an algorithm previously developed in Barash et al [5] as it allowed partial (weighted) samples as described below. For each of these algorithms we tuned the matching hyper-parameters using standard train/validation/test sets procedures.

We evaluated two model training setups: a 1 step and a 2 step procedure. The first step procedure uses only labeled data to train the classifier. The 2 step procedure then adds a second stage where the initial model is used to predict the label of the unlabeled data and adds these samples with the associated predicted value as weighted samples.[6] We found the 2 step procedure to slightly improve the accuracy of the models, with the logistic classifier giving overall best prediction accuracy and model stability under subsampling.

Finally, we note that given the sparsity of some of the model features and the instability of some of the models under subsampling, we expect that non-linear and regression based models may prove more accurate as more data is accumulated.

REFERENCES

1. Cox DR. Regression models and life tables (with discussion). *J Royal Stat Soc.* 1972;34:187-220.
2. Wei L-J. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat Med.* 1992;11:1871-1879.

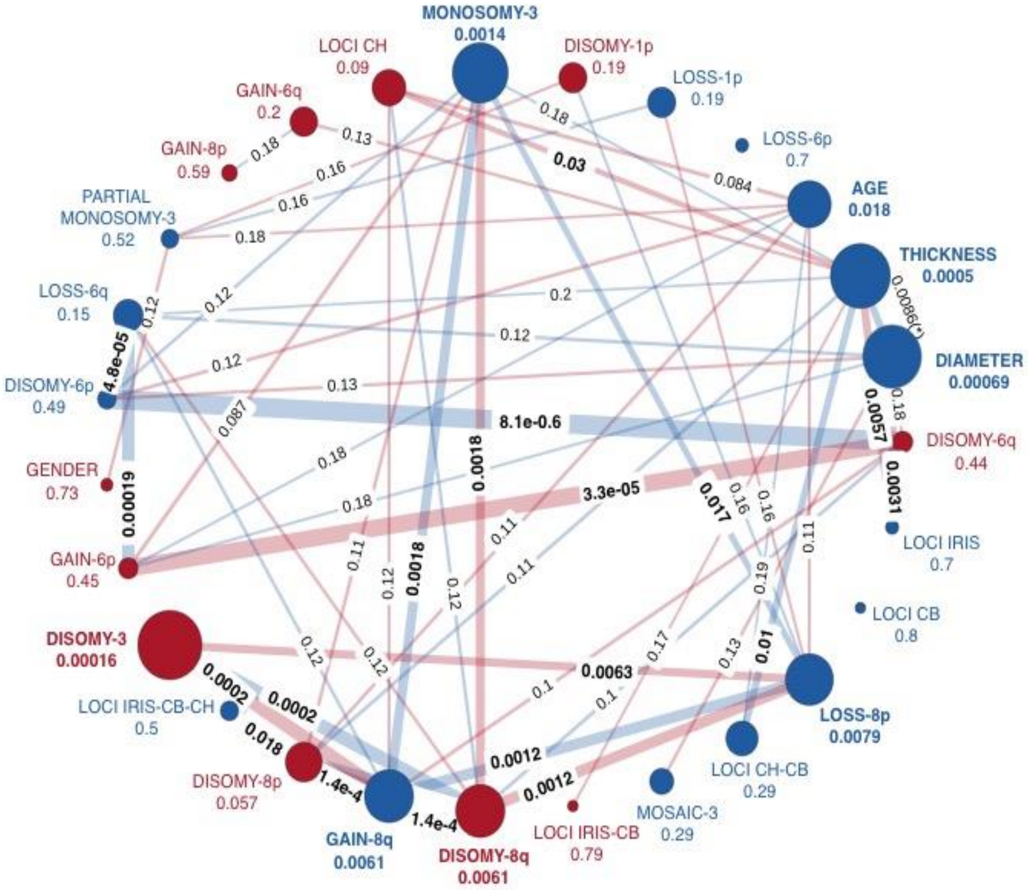
3. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat.* 2001;29:1189-1232.

4. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.

5. Barash Y, Calarco JA, Gao W, et al. Deciphering the splicing code. *Nature.* 2010;465:53-59.

6. lkan C, Noto K. (2008). Learning classifiers from only positive and unlabeled data. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (ACM)*. 2008: 213–220.

Supplementary Figure S1. Information content measure for feature SET 3, based on a two-sided Fisher exact test between the features and the true labels of the nodes (metastasis yes/no) and the information content between features (edges). The color indicates negative (blue) or positive (red) correlation. The size of the nodes and the width of the edges indicate the significance (larger nodes or thicker edges indicates lower p-value).


MONOSOMY-3 0.0014
LOCI CH 0.09
DISOMY-1p 0.19
LOSS-1p 0.19
GAIN-6q 0.2
LOSS-6p 0.7
GAIN-8p 0.59
AGE 0.018
PARTIAL MONOSOMY-3 0.52
THICKNESS 0.0005
LOSS-6q 0.15
DIAMETER 0.00069
DISOMY-6p 0.49
DISOMY-6q 0.44
GENDER 0.73
LOCI IRIS 0.7
GAIN-6p 0.45
LOCI CB 0.8
DISOMY-3 0.00016
LOSS-8p 0.0079
LOCI IRIS-CB-CH 0.5
LOCI CH-CB 0.29
DISOMY-8p 0.057
MOSAIC-3 0.29
LOCI IRIS-CB 0.79
GAIN-8q 0.0061
DISOMY-8q 0.0061
0.18
0.13
0.03
0.084
0.18
0.16
0.16
0.18
0.12
0.12
0.2
0.12
0.0086(*)
4.8e-05
0.12
0.13
0.087
8.1e-0.6
0.0057
0.18
0.0018
0.18
0.18
3.3e-05
0.017
0.16
0.16
0.0031
0.00019
0.11
0.11
0.11
0.11
0.12
0.0018
0.12
0.19
0.12
0.12
0.17
0.13
0.01
0.0063
0.0002
0.0002
0.1
0.1
0.018
0.0012
0.0012
1.4e-4
1.4e-4