Supplementary material

Telomere length

Samples were made up of either separated peripheral blood mononuclear cells (PBMC) or whole blood collected in ACD-A blood collection vacutainers and kept frozen at -80°C. All PBMC samples were separated and placed in storage within 3 days of draw date.

DNA was extracted from PBMC and whole blood samples using the Qiagen QIAamp DNA Mini Blood Extraction kits (Catalog # 51306) according to manufacturers' specifications.

Average telomere length (ATL) was measured by quantitative PCR (qPCR) by comparing telomere sequence copy number in a subject sample (T), to single-copy number (S) from the same sample, as described by Dr. Richard Cawthon[22, 23] and analysis by Dr. Jue Lin[24]. The resulting T/S ratio is proportional to ATL[25].

Average Telomere Length results or "T/S ratios" were determined for each sample by calculating the average telomere signal normalized to a single copy gene, based off of the standard curve.

Variable Pre-processing

Age, height, weight, and telomere length are continuous variables, whereas the remaining variables are discrete. Full descriptions of the variables are given in Supplementary file 1. All variables were imputed for missing values and (for discrete variables) encoded. Furthermore, to account for variable interactions, we formed all possible products of input variables and added them to the model. This turned out to significantly improve cross-validation performance of the algorithm as measured by the survival difference between predicted-positive and predicted-negative groups. The introduction of interactions (products) in the model substantially increased computational complexity of learning, and reduced interpretability. Nevertheless, the gain in estimated performance justified the penalty.

Following the formation of products, all resulting variables were centered to zero mean and scaled to unit standard deviation. Therefore, the final set of variables that was input to the machine learning workflow consisted of the pre-processed normalized values of the variables listed in Table 2, and their products.

Variable Selection and Classifiers

Most practical classification devices utilize two stages: feature selector or extractor, and classifier. The feature selector chooses which of the input variables shall be input to the classifier algorithm, whereas feature extractor transforms all input variables into a (usually) smaller number of derived variables, which are then input to the classification algorithm. In our context, the features are the patient/donor variables listed in Table 2, and their products. Some algorithms, like ElasticNet[26], perform feature selection internally, whereas others require external feature selector. In this work, we used external feature ranking based on Pearson correlation coefficient between feature value and class label[27] (Preferred or NotPreferred); the features with highest absolute value of the correlation were fed to the classifier stage.

We experimented with several state-of-the-art classifiers, including Support Vector Machines[28], Random Forest[29], Elastic Net and Normal Mixture Modeling[29] as well as survival models as explained above. In the search for the clinically most useful model, we employed meaningful combinations of feature selectors, feature extractors and classifiers. The best combination that we found entailed combining the features selected by Pearson correlation ranking in combination with the SVM classifier. The final set of features used in the classifier is listed in the Supplementary file 2.

Model Selection

Each predictive model (i.e., classifier) is defined by a set of parameters (weights), similar to coefficients of linear regression. Most of the weights, but not all, are optimized using a core, off-the-shelf classification algorithm. The process of deciding the values of the remaining parameters is called model selection. For example, a non-linear Support Vector Machine model is defined by the weights and four additional parameters: 1) the list of input features 2) cost 3) gamma 4) decision threshold. Model selection is the process of determining these values. The approach that we used is called "grid-search"[30], and it involves trying out a relatively large number of possible combinations of these four parameters, and choosing the best one. Other methods of model selection exist, but none have been shown to consistently match or surpass performance of the grid search when the number of parameters is low.

The decision threshold plays a special and crucial role in controlling whether the classification system will have clinical utility. Most core classifiers output a score (numerical value). The classifiers that we considered produce positive scores which increase with the probability that the donor is conferring better survival on the recipient (i.e., higher score means better donor for the given recipient). The decision threshold is used to convert this score to the Preferred or

NotPreferred label by comparing the classifier score with the decision threshold value. High values of the decision threshold correspond to classification systems which consider fewer donors to be Preferred, but each has a higher likelihood of being a good match for the recipient (as a result, more patients would be unlikely to find a Preferred donor when high decision threshold values are used). Conversely, lower values of the decision threshold correspond to models that identify more donors as Preferred, albeit with lesser chance of achieving the five-year survival outcome. Therefore, there is a trade-off between clinical utility (increasing survival chances) and practicality (identifying at least one Preferred donor for a reasonable proportion of acute leukemia HCT recipients). Decision thresholds correspond to the "stringency" of a test in calling a donor Preferred.

Given these considerations, the model selection process proceeded as follows. We evaluated each grid point (i.e., each donor classifier) using repeated ten-fold cross-validation[31] analysis of the classifier performance, with each cross-validation run was repeated 30 times. This is, to our knowledge, the most reliable method available to assess the quality (i.e., accuracy) of the classifier predictions, short of fully independent validation. The result was a table of (proportion, survival) values, one for each classifier, where "proportion" measures how frequently the corresponding test labeled donors as Preferred, and "survival" measures the probability that a patient would survive at least five years if receiving HCT from a donor labeled Preferred by the model. The best model was the one which labeled at least 10% of the donors Preferred, and maximized the survival benefit experienced by the recipient.

To facilitate the selection process and make it more intuitive, this table was visualized as a two-dimensional dot plot of (proportion, survival) values. An example of such a graph is given in Fig. 2. It should be noted that statistics of each model (i.e., combination of weights) was evaluated for 100 different values of decision threshold, resulting in about a million different classifiers, although the number of substantially distinct models was about 10,000.