**Supplemental materials and methods**

**Data description**

The data used in this study are from CML patients diagnosed and treated at the Brazilian National Cancer Institute (INCA). The data set was de-identified prior to analysis. The experimental protocols and data collection were approved by the INCA Institutional Review Board, which is subjected to the Brazilian National Ethics Committee (CONEP) under register number 129/10, and patients have been consented to contributing their data. This specific study is considered non-human subject research since it only accesses de-identified data. All methods were performed in accordance with the relevant guidelines and regulations, ensuring that they conformed to the ethical guidelines of the Declaration of Helsinki.

We only included patients who received IM treatment and we excluded patients if they were on DAS or NIL treatments for any of the time periods we considered. This resulted in a dataset of n=144 patients. We divided the available information into two types: (i) variables with value only at diagnosis and/or start of IM treatment, (ii) variables with values available at three-month intervals after the start of IM.

Table S1 provides our patient cohort distribution based on the so-called Sokal risk score – a hazard ratio of deaths computed using age, spleen size, platelet count, and percentage of myeloblasts in peripheral blood.I

For the patients in our dataset, we also have information related to (i) treatment, such as the use of cytoreduction agents like hydroxyurea or interferon (INF), normally used before starting TKI; imatinib (IM) starting date and dosage; and hematopoietic stem cell transplantation (HSCT); (ii) information related to responses such as Hematologic Response (HR), Cytogenetic Response (CyR), and Major Molecular Response (MMR); and (iii) the BCR-ABL1 ratios recorded at three-month intervals after the start of IM.

**Pre-processing and timeline strategy**

Molecular response was defined by the BCR-ABL1/ABL1 ratio measured by qPCR in International Scale at three-month intervals. Time 0 was defined at the start of IM, and then, despite BCR-ABL1 ratios being assessed every three months, we considered for this study the ELN monitoring milestones 3, 6, 12, 18, 24 and 60 months, because of the existence of too many missing values in other timepoints.

We defined binary variables to indicate use (1) or no use (0) of treatments with hydroxyurea, INF, and HSCT in each three-month time interval. To account for changes in IM dosage during treatment, we defined a continuous variable equal to the average of imatinib dosages in each three-month interval.

We imputed the missing values with the median for all variables, except BCR-ABL1 and cytogenetic response (CyR) ratios. While it would have been ideal to estimate the distribution of recorded BCR-ABL1 ratios in each 3-month time interval, missing values precluded accurate estimation. Instead, we computed summary statistics such as the median, mean, min, max, and standard deviation (std) of the recorded values of each patient, and we used these as features instead of the original recorded values. We call these new features *aggregated features*. In our final models, some of the *aggregated features* are not used because they have high correlation with each other, and we retain only one variable among each highly correlated pair, which is the median of BCR-ABL1 ratio measurements. This median shows up among the most important predictive variables in all our models.

We removed variables with zero variability (std<0.0001). To reduce the less informative features and simplify the models, we applied statistical feature selection. Specifically, we tested the null hypothesis of each variable having the same distribution in the two cohorts (achieving DMR and not) using the chi-squared test for binary variables and the Kolmogorov-Smirnov statistic for continuous variables. Variables for which we could not reject the null hypothesis (p-value>0.05) were removed. We also removed one from each pair of highly correlated variables (absolute value of correlation coefficient > 0.7).

**Models derived**

When predicting DMR at later months using patient characteristics at time zero (diagnosis) and early treatment information, missing BCR-ABL1 ratios limit our ability to make predictions for each 3-month interval. We considered the following models: (a) prediction of DMR at 18 months using information up to 6 months (model\_6\_18), (b) prediction of DMR at 18 months using information up to 3 months (model\_3\_18), and (c) prediction of DMR at 12 months using information up to 6 months (model\_6\_12). We also sought to inform drug discontinuation decisions by predicting long-term DMR. We defined the achievement of long-term DMR as follows: the patient reaches DMR at 24 months after the start of the treatment and maintains this status until 60 months after treatment initiation. This leads to another model that predicts the long-term DMR using information up to 12 months after the start of IM (model\_long\_12).

Moreover, we excluded patients who have no BCR-ABL1 ratios recorded during the period used to obtain predictive variables according to each model’s definition. Specifically, in model\_6\_18, we retain 87 patients and among them 25.3% achieve DMR at 18 months. In model\_3\_18, we retain 81 patients and among them 25.9% achieve DMR at 18 months. In model\_6\_12, we retain 104 patients and among them 12.5% achieve DMR at 12 months. In model\_long\_12, we retain 95 patients and among them 15.9% achieve long-term DMR.

**Classification methods**

We explored a variety of supervised classification methods both linear and tree-based algorithms. Linear classifiers included Logistic Regression (LR) and linear Support Vector Machines (SVM).II These lead to interpretable predictions; for instance, the LR coefficient of a feature represents the sensitivity of the predicted likelihood to that feature and the absolute value of this coefficient can be interpreted as feature importance. Linear classifiers were fitted with an additional regularization term seeking to prevent the influence of outliers in training or test data.III In this study, we considered both an L1-norm regularizer (models L1LR and L1SVM) and an L2-norm regularizer (models L2LR and L2SVM). Tree-based learning algorithms included Gradient Boosted Trees (LightGBM)IV and Random Forest (RF),V which are more complex and generally yield better classification performance, as well as the feed forward Multilayer Perceptron Neural Network (MLP).

**Model training evaluation**

All variables were standardized by subtracting the mean and dividing by the standard deviation. We randomly split the dataset into three equal parts, where two parts were used as the training set, and the third part as the test set. We used the training set to tune the model hyperparameters via 9-fold cross-validation. We evaluated the performance metrics on the test set. We repeated training and testing twenty times, each time with a different random split between the training and test sets. The mean and standard deviation of all metrics on the test sets over the twenty repetitions are reported.

We evaluated model performance using two metrics: Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) and Weighted-F1 score. The ROC is created by plotting the true positive rate (*i.e.*, sensitivity) against the false positive rate (equal to one minus specificity) at various thresholds. The c-statistic, or the Area Under the ROC Curve (AUC), is used to evaluate prediction performance. A perfect predictor has an AUC of 1 and a predictor which makes random guesses has an AUC of 0.5. The F1 score is the harmonic mean of recall and precision. The weighted F1-score is the average of the F1-scores of each class weighted by the number of participants in each class. The weighted-F1 score is between 0 to 1, and a higher value represents a better model. The AUC is more easily interpretable (it is the probability that the model assigns a higher score to a random positive sample than to a random negative sample), while the weighted F1-score is more robust to class imbalance.

We also use Recursive Feature Elimination (RFE) with L1-penalized logistic regression (L1-regularized RFE) to extract the most informative features and develop parsimonious models. Specifically, after running L1LR we obtain weights associated with the variables (i.e., the coefficients of the model); we then eliminate the variable with the smallest absolute weight and perform L1LR to obtain a new model. We keep iterating in this fashion, eliminating one variable at each iteration, to select a model that maximizes a metric equal to the mean AUC minus the standard deviation of the AUC in a validation dataset (using 9-fold cross-validation). We will be referring to the resulting model as the *parsimonious* model.

**Tables**

*Table S1. Patient distribution according to the Sokal risk score.*

|  |  |
| --- | --- |
| Sokal Risk Score | Number of Patients (%) |
| Low (< 0.8) | 74 (51.4%) |
| Intermediate (0.8–1.2) | 49 (34.0%) |
| High (> 1.2) | 21 (14.6%) |

Supplemental Material References

I. Hasford, J. *et al.* Predicting complete cytogenetic response and subsequent progression-free survival in 2060 patients with CML on imatinib treatment: the EUTOS score. *Blood* **118**, 686–692 (2011).

II. Cortes, C. & Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273–297 (1995).

III. Chen, R. & Paschalidis, I. C. Distributionally Robust Learning. *OPT* **4**, 1–243 (2020).

IV. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *NIPS* (2017).

V. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).