iScience, Volume *25*

Supplemental information

A machine learning model of response to hypomethylating agents in myelodysplastic syndromes

Nathan Radakovich, David A. Sallman, Rena Buckstein, Andrew Brunner, Amy Dezern, Sudipto Mukerjee, Rami Komrokji, Najla Al-Ali, Jacob Shreve, Yazan Rouphail, Anne Parmentier, Alexandre Mamedov, Mohammed Siddiqui, Yihong Guan, Teodora Kuzmanovic, Metis Hasipek, Babal Jha, Jaroslaw P. Maciejewski, Mikkael A. Sekeres, and Aziz Nazha

Supplemental Figures:

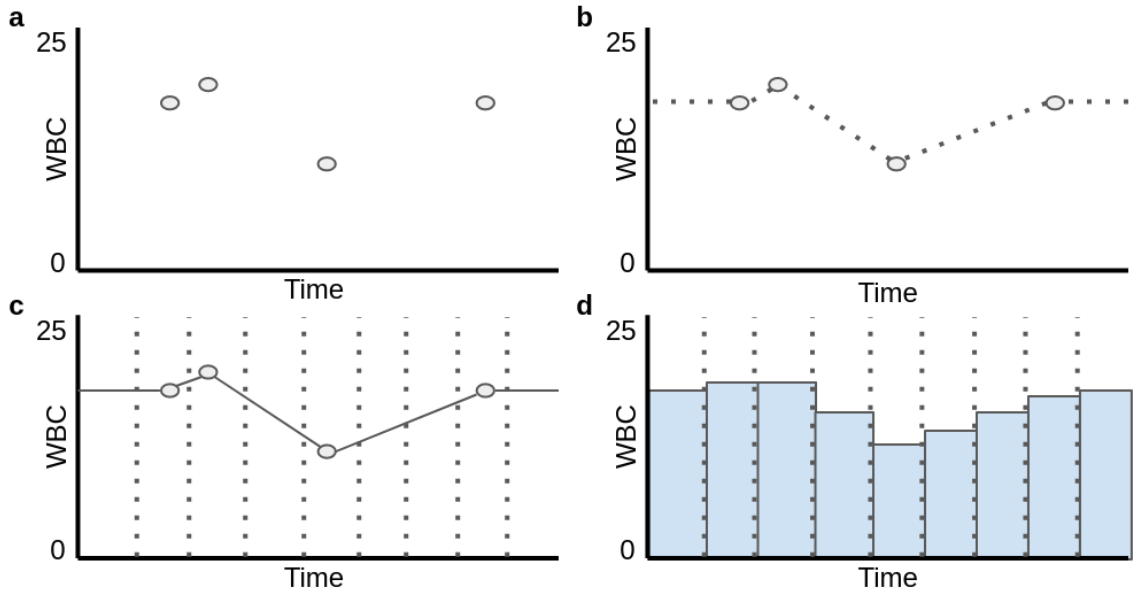

Figure S1, related to STAR methods: Visual depiction of strategy used to handle irregular time series data. Starting data occur at irregular intervals (a). These data are interpolated in between time points where data are available, forward- and back-filled (b), binned into discrete, regular intervals (c), and averaged within those intervals (d).

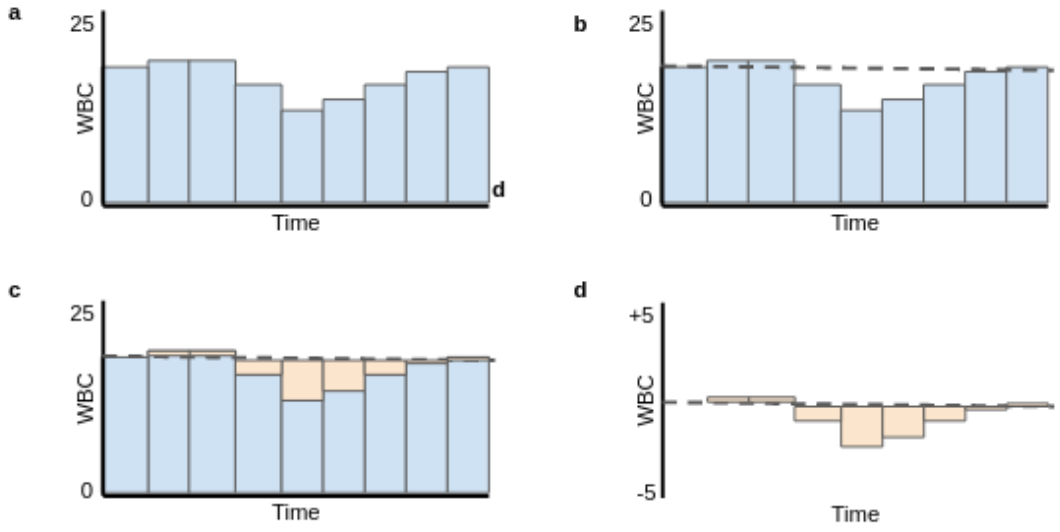

Figure S2, related to STAR methods: Visual depiction of calculating change in CBC values from baseline. Downsampled labs are created (a), with the first time period serving as a baseline against which the subsequent time periods are compared (b). The baseline value is subtracted from subsequent time periods (c) in order to contextualize lab values for a given time period (d).

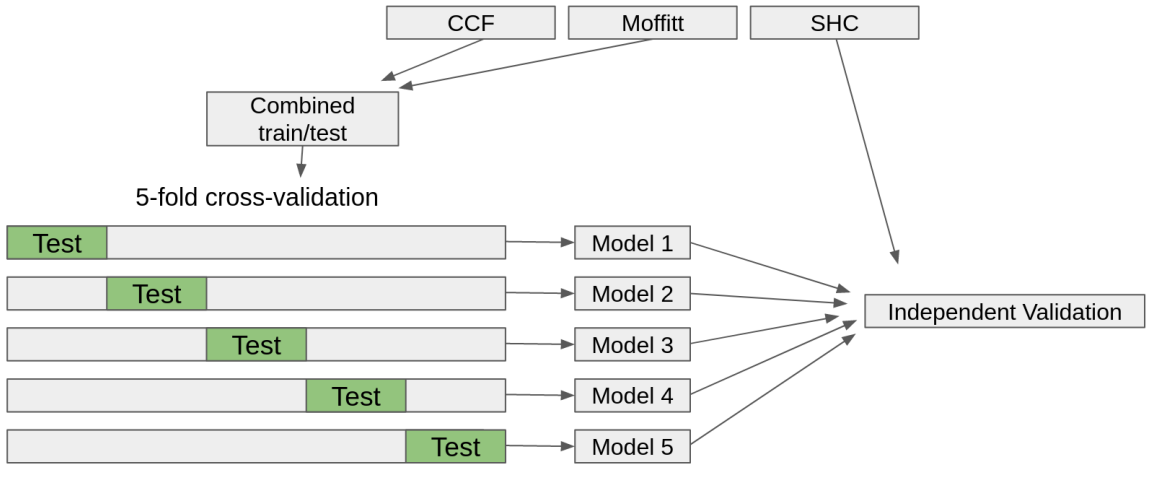

Figure S3, related to STAR methods: Schematic of train/test/validation schema. CCF and Moffitt data are pooled; five non-overlapping test subsets are created from the pooled data and used to generate five separate models; an independent cohort (SHC) is then used to test models' external validity.

In order to estimate the variability of models' performance, bootstrapping was used to

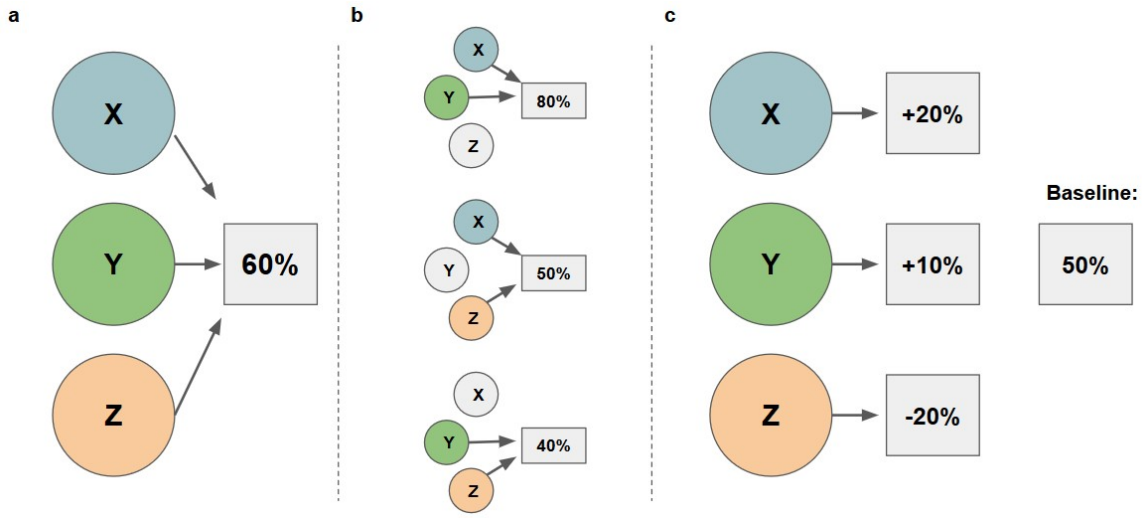

Figure S4, related to STAR methods: visual representation of Shapley value approach. Three variables of unknown significance, x, y, and z, contribute to a predicted likelihood (a). In order to determine how each variable affects the outcome, predictions are generated by iteratively removing each variable and observing how its absence affects predictions (b). A baseline (i.e., default) probability and individual contributions are then inferred (c).

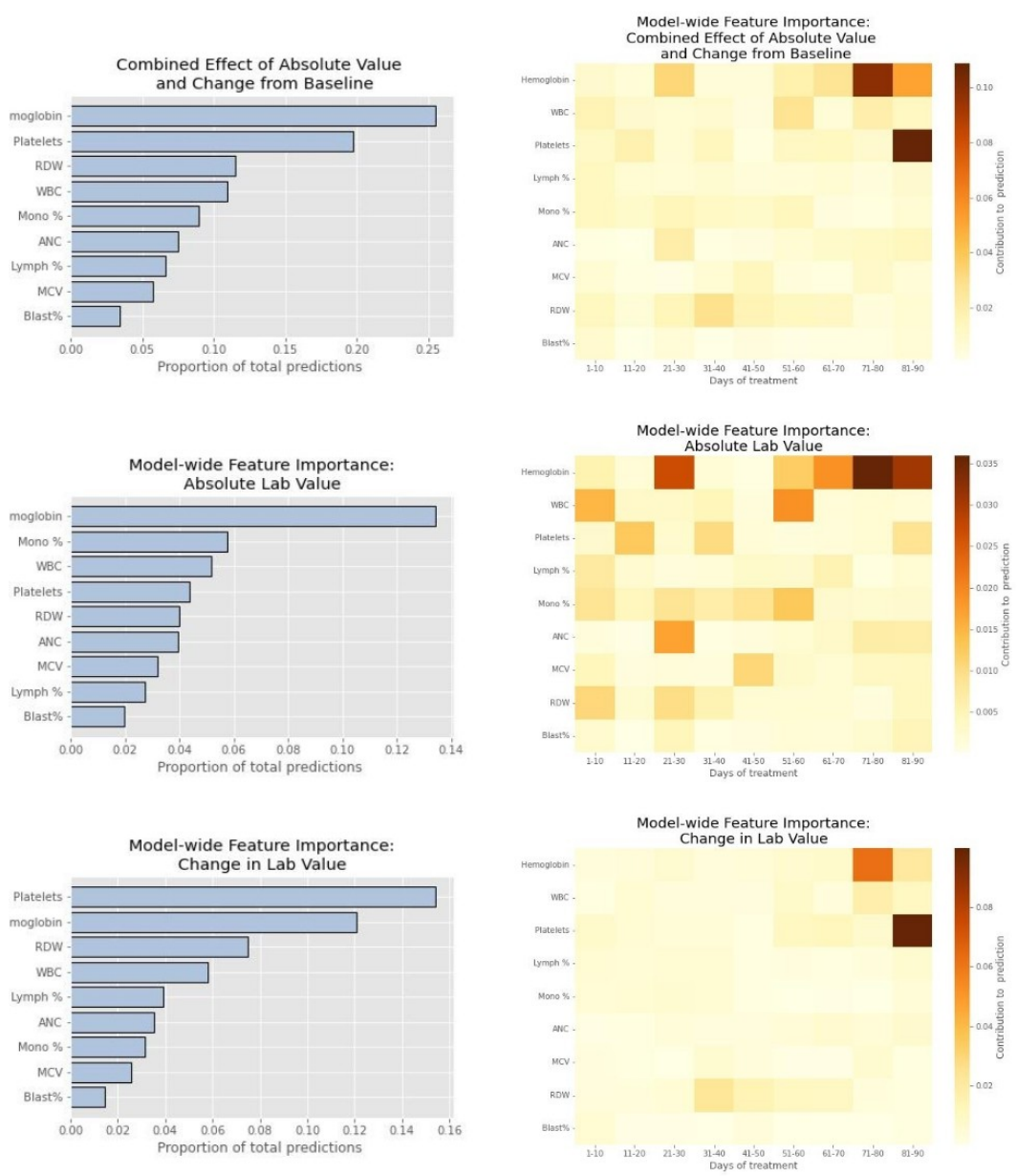

Figure S5, related to figure 2. Feature importance for absolute laboratory values and change in laboratory values. Bar graphs on the left column depict the overall contribution of different laboratory values to model predictions considering combined effects of absolute values and change in values from baseline (top), absolute laboratory value alone (middle), and change in laboratory value from baseline (bottom). The heatmaps on the right depict different laboratory tests in rows, and timepoints from treatment in columns. Feature importance as determined by SHAP values is depicted via color coding, with darker colors corresponding to higher feature importance.