Supplement to "Predictive models for splenic response to JAK-inhibitor therapy in patients with myelofibrosis." by Menghrajani, et al.

S1 Supplemental Tables and Figures

Table S1: Summary of response and categorical candidate prognostic factors (at time of JAK2 inhibitor therapy initiation) among patients with at least one measured endpoint

Name	MA($n = 53$) n(%)	MI($n = 58$) n(%)	ST($n = 78$) n(%)	PST1($n = 194$) n(%)	PST2($n = 165$) n(%)
Early splenic response					
=NA	11	3	3	4	3
=Yes	27(64%)	19(35%)	23(31%)	81(43%)	50(31%)
=No	15(36)	36(65)	52(69)	109(57%)	112(69%)
Late splenic response					
=NA	10	4	15	43	56
=Yes	26(60)	27(50)	27(43)	71(47)	39(36)
=No	17(40)	27(50)	36(57)	80(53)	70(64)
Sex					
=NA	0	0	0	0	0
=Female	18(34)	22(62)	38(51)	85(44)	74(45)
=Male	35(66)	36(38)	40(49)	109(56)	91(55)
Diagnosis					
=PMF	29(55)	27(47)	40(51)	128(66)	115(70)
=PETM	6(11)	14(24)	13(17)	24(12)	17(10)
=PPV	18(34)	17(29)	25(32)	42(22)	33(20)
WBC at initiation, 10^9/L					
=NA	0	3	0	0	0
≤ 5	11(21)	5(9)	3(4)	36(19)	59(36)
$\in (5, 10]$	14(26)	7(13)	26(33)	68(35)	35(21)
> 10	28(53)	44(79)	49(63)	90(46)	71(43)
Platelet at initiation, 10^9/L					
=NA	0	3	0	0	3
≤ 50	3(6)	2(4)	2(3)	32(16)	79(49)
$\in (50, 100]$	6(11)	13(24)	10(13)	33(17)	55(34)
$\in (100, 150]$	8(15)	9(16)	13(17)	20(10)	18(11)
$\in (150, 450]$	23(43)	24(44)	41(53)	83(43)	9(6)
> 450	13(25)	7(13)	12(15)	26(13)	1(1)
Bone Marrow Cellularity					
=NA	2	1	17	15	23
=Hypocellular	8(16)	9(16)	4(7)	36(20)	42(30)
=Normocellular	4(8)	0	5(8)	21(12)	19(13)
=Hypercellular	39(77)	48(84)	52(85)	122(68)	81(57)
Fibrosis					
=NA	2	4	25	7	13
$\leq$MF-1	12(24)	8(15)	1(2)	28(15)	16(11)
=MF-2	18(35)	20(37)	14(26)	68(36)	45(30)
=MF-3	21(41)	26(48)	38(72)	91(49)	91(60)
% Blasts $\geq 1\%$					
=NA	1	10	24	7	11
=Yes	19(37)	31(65)	38(70)	97(52)	69(45)
=No	33(63)	17(35)	16(30)	90(48)	85(55)
Transfusion Status					
=NA	0	2	0	24	42
=Yes	12(23)	17(30)	9(12)	30(18)	30(24)
=No	41(77)	39(70)	69(88)	140(82)	93(76)

Table S2: Summary of continuous candidate prognostic factors (at time of JAK2 inhibitor therapy initiation) among patients with at least one measured endpoint

	MA($n = 53$)		MI($n = 58$)		ST($n = 78$)		PST1($n = 194$)		PST2($n = 206$)	
Name	# NA	Median (Range)	# NA	Median (Range)	# NA	Median (Range)	# NA	Median (Range)	# NA	Median (Range)
Age at initiation of therapy, years	0	69(42,87)	0	68(43,86)	0	70(43,85)	0	67(23,87)	0	68(39,91)
BMI, kg/m^2	0	25(19,35)	20	26(18,37)	7	25(18,42)	3	24(16,37)	6	24(15,46)
Relative init. JAK2 dose, % of RP2D†	2	100(33,166)	0	66(17,160)	0	100(6,136)	0	100(100,100)	0	100(100,100)
Spleen size at initiation, cm	0	12(0,26)	0	15(0,32)	0	16(0,34)	0	12(4,33)	0	12(3,33)
Time from DX, weeks	0	89.4(1.1,942)	0	103.1(0.4,774)	0	43.4(0.3,1266)	0	57.5(1,1466)	0	114.9(1,1442)
WBC at initiation, 10^9/L	0	12.4(2.1,159.0)	3	21.0(3.2,125.3)	0	15.0(1.5,89.6)	0	9.7(1.2,169.6)	0	8.0(1.2,221)
Platelet at initiation, 10^9/L	0	239(16,799)	3	209(25,854)	0	269(20,1520)	0	172(7,1065)	3	52.5(6,947)
Hgb at initiation, g/dL	0	10.5(7.6,15.2)	1	10.8(6.7,14.6)	0	10.5(7.5,16.6)	0	10.5(4.3,17.3)	0	9.6(5.1,17.1)

†Defined as the ratio of each patient's prescribed dose over the recommended phase II dose of JAK2 inihibitor

Table S3: Summary of cross-validated and validated predictive performance of multivariable logistic models for splenic response after initiation of JAK-inhibitor therapy, described in Table 1 of the manuscript. AUC measures the probability of correctly ordering the probabilities of a random responder and non-responder. Larger is better, with 0.5 being the null value and 1.0 being the ideal value.

AUC	Trained	Cross-Validated	Validated
Early	0.789	0.679	0.616
Late (Predicted at Baseline)	0.725	0.641	0.523
Late (Using Early Response)	0.797	0.765	0.792

AUC = Area Under Receiver Operating Characteristic

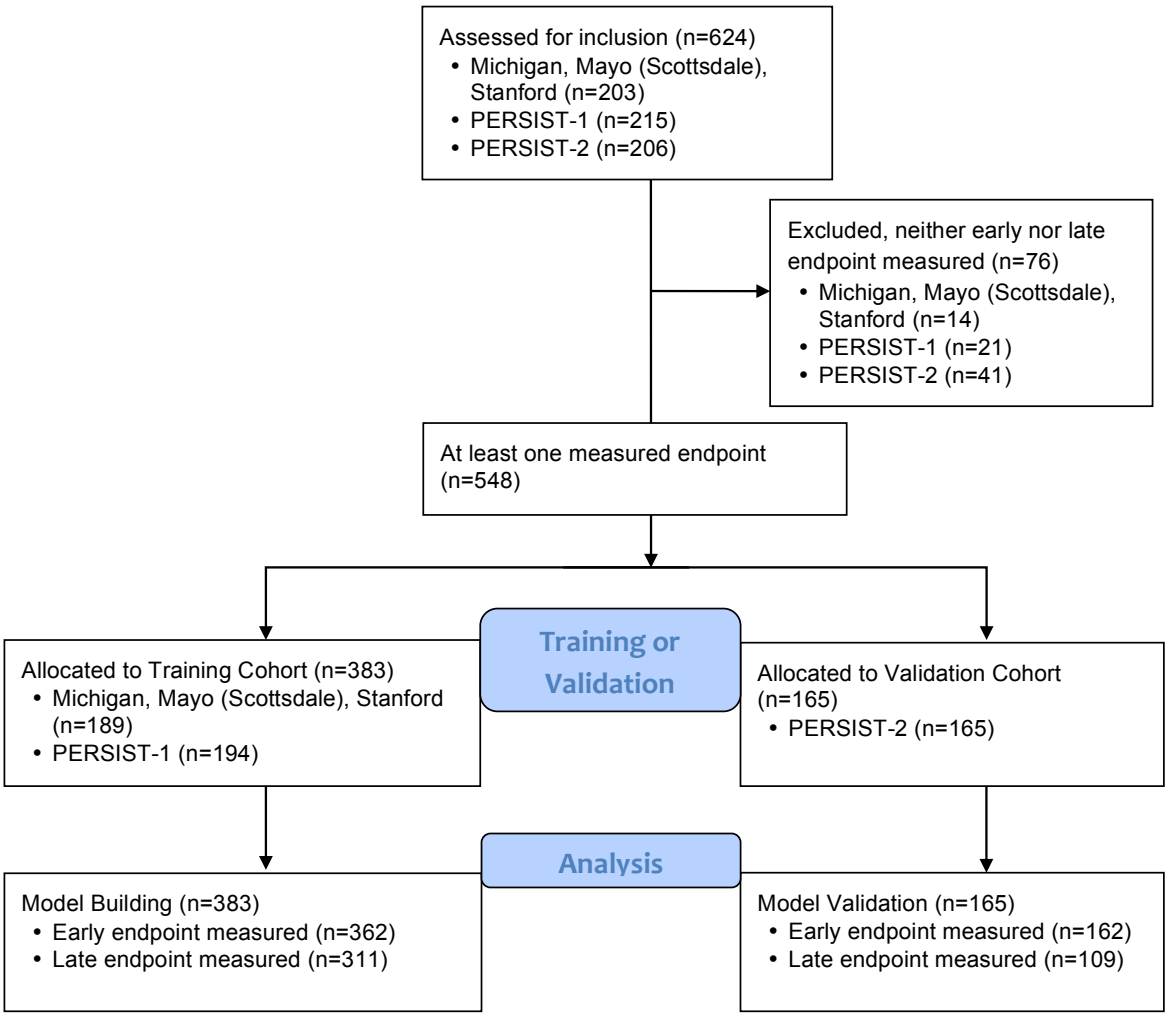

Figure S1: Flow chart delineating construction of model training and model building cohorts

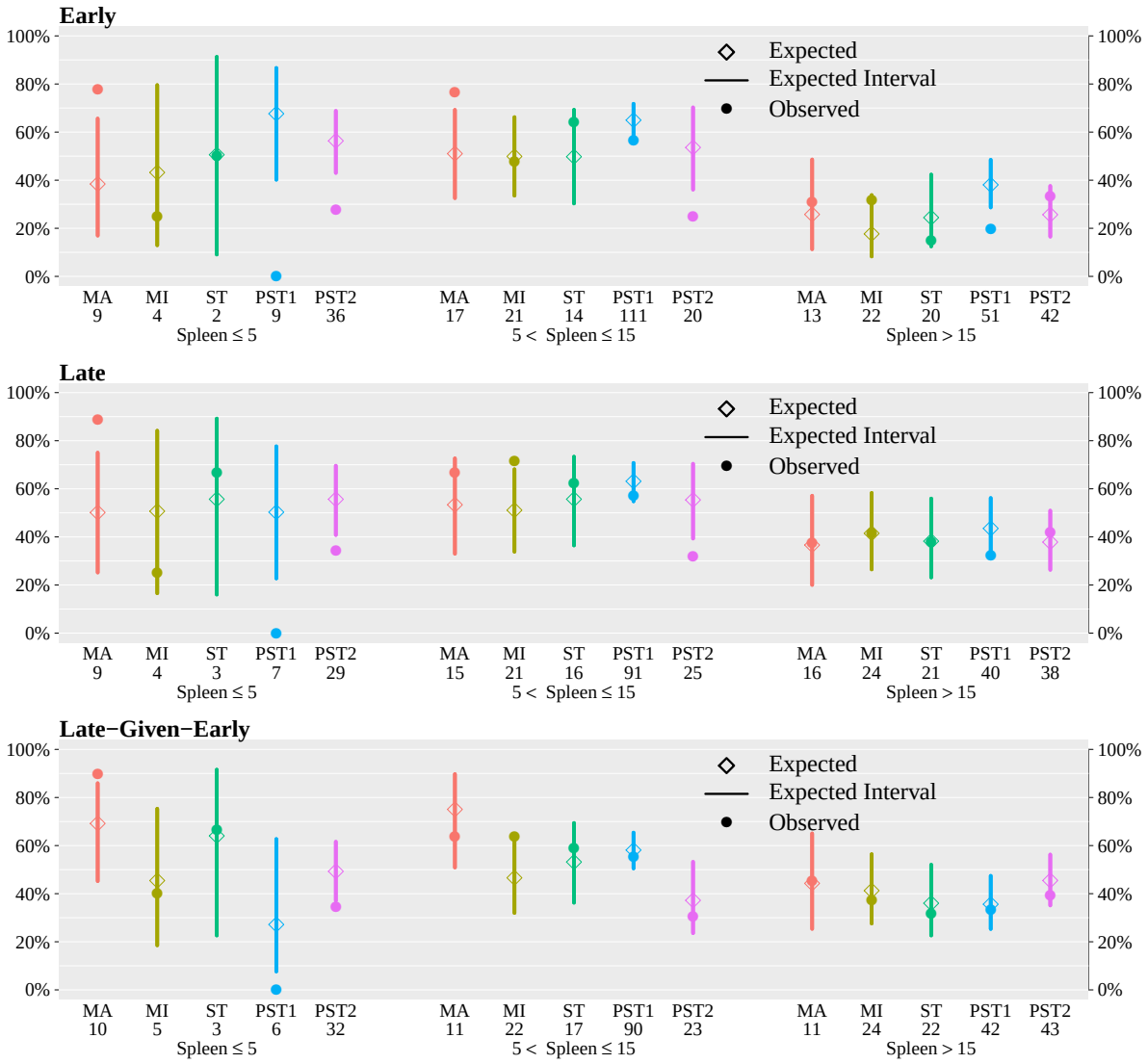

Figure S2: Calibration across the five centers (left-to-right colored bars; MA=red, MI=yellow, ST=green, PST1=blue, PST2=lavender), grouped by spleen size at initiation (Spleen < 5; $5 <$ Spleen ≤ 15; Spleen > 15), for the three models (top panel = Early, middle panel = Late, bottom panel = Late-Given-Early). The number of patients in each group is given below each center/splenic group combination. The open square is the expected proportion of responders for each category. The filled circle is the observed proportion of responders. The vertical line spans where the observed proportion should fall with probability 95%, if the model was perfectly calibrated.

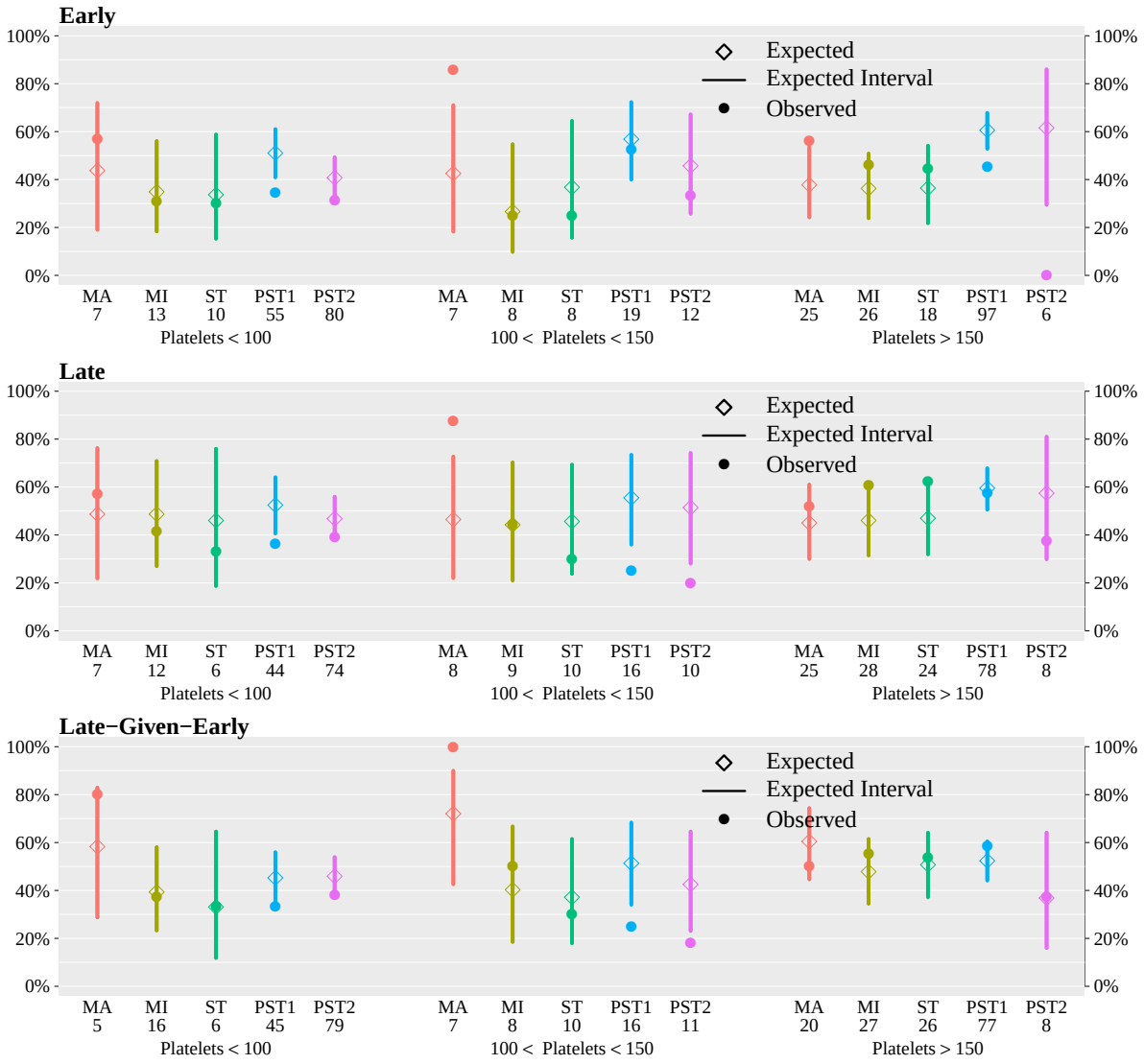

Figure S3: Calibration across the five centers (left-to-right colored bars; MA=red, MI=yellow, ST=green, PST1=blue, PST2=lavender), grouped by platelet counts at initiation (Platelets < 100; $100 <$ Platelets < 150; Platelets > 150), for the three models (top panel = Early, middle panel = Late, bottom panel = Late-Given-Early). The number of patients in each group is given below each center/platelet group combination. The open square is the expected proportion of responders for each category. The filled circle is the observed proportion of responders. The vertical line spans where the observed proportion should fall with probability 95%, if the model was perfectly calibrated.

S2 Statistical Methodology

Multiple Imputation Step

To make use of the partial information available in the observations with missing data, we used multiple imputation (Rubin, 2004), specifically, iterative chained equations (Van Buuren et al., 2006, White et al., 2011) using the R package MI (R Core Team, 2017, Su et al., 2011). Age and BMI are constrained to be positive during imputations. A separate Bayesian logistic regression model was fit to each imputed dataset (as described below), with the results combined across imputations as described in Gelman et al. (2004, p. 520) and Zhou and Reiter (2010). Zhou and Reiter found that this approach of combining Bayesian inferences required many more imputations than the standard recommendation of 5–10 completed datasets necessary for maximum likelihood inference; specifically, we imputed 150 datasets.

Data Processing, Additional Details

To allow for differently changing risk profiles among patients with leukopenia and leukocytosis, we treated WBC count as a mixed discrete-continuous variable. We first mapped each WBC count into a three-category vector, based upon whether it was below 5, between 5 and 10, or greater than 10. The numeric value of the middle category was always zero and used as a reference. The numeric value of the extreme categories was only non-zero if the count fell in that category. In that case, we calculated the absolute value of the log2-ratio of the count and the boundary value (either 5 or 10). So, for example, a WBC count of 4×10^9/L was mapped to the vector (0.322, 0, 0), because log2(5/4) = 0.322, whereas a WBC count of 12 was mapped to (0, 0, 0.263), because log2(12/10)=0.263. We transformed platelet counts similarly, using 150 and 450 as normal boundaries.

Two other continuous covariates, HGB and time from diagnosis, were log2-transformed.

Finally, all covariates were centered and scaled by twice the empirical standard deviation, to allow for relative comparisons of associations across different types of prognostic factors.

After transforming covariates as described in the main text, there were 33 possible prognostic factors considered. We location-scale transformed each imputed dataset to a common standard, using as location the observed mean (ignoring missing values) of each covariate. For scale, we used twice the standard deviation of the observed data. This is recommended by Gelman (2008) in the presence of a mixture of continuous and binary covariates, because a change in two standard deviations for a binary predictor with probability 0.5 is interpreted as the difference in association between 0 and 1.

Model Fitting

For each imputed dataset and for each time point, we fit a multiple Bayesian logistic regression model, allowing for separate, center-level intercept parameters (one for MI, MA, ST, or PST1). For the ith patient at the j center, $Y_{ij}^{(t)} = 0$ and $Y_{ij}^{(t)} = 1$ indicate, respectively, no splenic response or splenic response at time t months ($t \in \{3, 6\}$), and $\boldsymbol{X}_{ij}$ indicates the patient's vector of covariates. Then,

$$\operatorname{logit} \Pr(Y_{ij}^{(t)} = 1 | \boldsymbol{X}_i) = \mu_j^{(t)} + \boldsymbol{X}_i^\top \boldsymbol{\beta}^{(t)}, \tag{S1}$$

where $\operatorname{logit}(x) = \log(x/[1-x])$. The vector $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^{33}$ is the vector of log-odds ratios (ORs). The reported ORs in Table 1 of the manuscript are un-scaled, i.e. interpreted in natural units of measurements rather than on the standard deviation.

We used the hierarchical shrinkage "horseshoe" prior for $\boldsymbol{\beta}^{(t)}$ (Carvalho et al., 2009, 2010,

Piironen and Vehtari, 2016). Specifically, a priori,

$$\beta_k \overset{\text{ind}}{\sim} N(0, \lambda_k \tau); \lambda_k \overset{\text{iid}}{\sim} C^+(0,1); \tau \sim C^+(0, 2/\sqrt{n})$$

Here, $N(0,s)$ is the normal distribution with standard deviation s, $C^+(0,s)$ is the half-Cauchy distribution (with support constrained to the positive real line) with scale parameter s, and n is the number of observations in the model. This particular prior, specifically our choice of scale parameter on τ, corresponds to an approximate prior assumption that about half of the elements of $\boldsymbol{\beta}$ are non-zero, as derived in Piironen and Vehtari (2016).

For the center-level intercepts, μ_j, we used the following prior:

$$\mu_j \overset{\text{ind}}{\sim} \text{Logistic}(\theta, \sigma); \theta \sim \text{Logistic}(0,1); \sigma \sim \text{Logistic}^+(0, 2.5)$$

where $\text{Logistic}(m,s)$ is the Logistic distribution with location m and scale s, and $\text{Logistic}^+(m,s)$ is the half-Logistic distribution with support constrained to the positive real line.

We fit the models using Hamiltonian Monte Carlo methods (Neal, 2011), using the STAN software interface in R (R Core Team, 2017, Stan Development Team, 2016). For each of the 150 imputed datasets, we ran the sampler for a warm-up period of 400 iterations, keeping the subsequent 400 draws of $\boldsymbol{\beta}$, and combining these across all 150 imputed datasets, so as to propagate imputation uncertainty (p. 520, Gelman et al., 2004, Zhou and Reiter, 2010). This yielded a combined 60,000 draws of $\boldsymbol{\beta}$ from the posterior distribution, accounting for the imputation uncertainty.

Model Selection

The kth predictor was selected if the posterior probability that the association was strictly positive or strictly negative exceed 2/3. Mathematically, this is written as:

$$\max\{\Pr(\beta_k < 0), \Pr(\beta_k > 0)\} > 2/3$$

We approximated these probabilities by counting the proportion of the 60,000 posterior draws of each β_k falling above or below zero. The maximum of these estimated probabilities is what we called the relevance percentage in the manuscript. We then re-fit the models after excluding those predictors that did not satisfy the above criterion to obtain the final models. Any "main effects" were automatically included if a corresponding interaction was included, and all baseline spleen size groups were included in a model if any spleen size group was selected.

Model Assessment

We first cross-validated each model by sequentially leaving out each of the four sites in the model-building data, repeating the entire imputation-plus-model-building process on the remaining three sites, and testing against the held-out site. We also validated the performance of our model on the PST2 data, which was held apart until the predictive model was finalized.

The final models at each time point were summarized via ORs and relevance percentages (defined above). We assessed discrimination with the 'Area Under the receiver-operating characteristic Curve' (AUC), a number between 0.5 and 1, with larger values indicating improved ability to distinguish between high- and low-risk individuals. These are reported in Table S3. We assessed model calibration, i.e. the ability to correctly estimate probabilities, by reporting the difference between model-based expected response rate and the observed

response rate, separately for each endpoint. This is sometimes called the 'goodness-of-fit' of the model. All analyses were conducted in the R statistical environment (R Core Team, 2017, Stan Development Team, 2016)

Model Application

Table 1 provides the necessary information to calculate the model-estimated probability of response for any patient. For a given patient, start with the baseline odds of response (which is the row labeled 'Intercept'), multiplicatively adjust these odds up or down based upon the relevant prognostic factors, then transform the result into a probability using the transformation Odds/(Odds + 1). To exemplify this, we provide here the details of the calculation for the 'Early response probability' for the hypothetical patient in Figure 1 of the manuscript, using a baseline spleen size of 8 cm and a baseline WBC count of 35×10^9/L, which is plotted in the left-most panel of Figure 1 in the manuscript. The x-axis corresponds to baseline spleen size, thus the fitted probability for this patient is approximately 0.50.

1. Start with the baseline odds, the Intercept in Table 1, corresponding to all prognostic factors set to zero: $x = 0.03$

2. Determine the multiplicative adjustment from the four relevant categorical variables in the 'Early' column (spleen size, transfusion requirement, cellularity, and fibrosis grade)

   (a) Patient's spleen length at baseline is in (5cm,10cm] $\Rightarrow$ multiply odds by 1.94. New odds are $x = 1.94 \times 0.03 = 0.0582$

   (b) Patient is transfusion dependent $\Rightarrow$ multiply odds by 1.13. New odds are $x = 1.13 \times 0.0582 = 0.0658$

   (c) Patient is not hypocellular $\Rightarrow$ odds are unchanged

   (d) Patient's MF has fibrosis grade MF-3 $\Rightarrow$ multiply odds by 0.80. New odds are

$x = 0.80 \times 0.0658 = 0.0526$

3. Determine the multiplicative adjustment from the three relevant continuous variables in the 'Early' column (HGB, Age, Relative Dose)

   (a) HGB is 10.7 g/dL, which is $\log_2(10.7) = 3.42$ doublings $\Rightarrow$ multiply odds by $2.07^{3.42} = 12.04$. New odds are $x = 12.04 \times 0.0526 = 0.6334$

   (b) Patient is 70 yo, which is 14 sets of 5-year increments $\Rightarrow$ multiply odds by $1.08^{14} = 2.94$. New odds are $x = 2.94 \times 0.6334 = 1.861$

   (c) Patient starts at 100% of RP2D of JAK inhibitor therapy, which is a relative initial dose of 0 $\Rightarrow$ odds are unchanged ($1.90^0 = 1$)

4. Determine the multiplicative adjustment from the two relevant mixed discrete-continuous variables in the 'Early' column (WBC and Platelets)

   (a) Patient's baseline platelet count is 210×10^9/L, which does not exceed 450×10^9/L, $\Rightarrow$ odds are unchanged

   (b) Patient's baseline WBC is 35×10^9/L, which is $\log_2(35/10) = 1.81$ doublings beyond the high threshold of 10×10^9/L $\Rightarrow$ multiply odds by $0.72^{1.81} = 0.552$. New odds are $x = 0.552 \times 1.861 = 1.027$

5. Determine the multiplicative adjustment from the statistical interactions

   (a) Patient starts at 100% of RP2D of JAK inhibitor therapy, which is a relative initial dose of 0. Patient's spleen length at baseline is in (5cm,10cm] $\Rightarrow$ odds are unchanged ($0.99^0 = 1$)

   (b) Patient's baseline WBC is 35×10^9/L, which is $\log_2(35/10) = 1.81$ doublings beyond the high threshold of 10×10^9/L. Patient's spleen length at baseline is in (5cm,10cm] $\Rightarrow$ multiply odds by $1.00^{1.81} = 1.00$. Odds are unchanged.

6. The patient's model-based estimate of probability of Early splenic response is $x/(x +$

$1) = 1.027/(1 + 1.027) = 0.507.$

Due to the presence of statistical interactions, the associations between spleen size at baseline, dose of JAK inhibitor, and WBC must be interpreted with caution. Specifically, the so-called 'main effects' corresponding to these predictors represent the associations of the reference group only. For example, in the Early model, the OR of 1.90 for dose of JAK inhibitor is the dose-response association among patients with a spleen size of 5cm or less at initiation. To calculate the dose-response association for patients with a different spleen size, multiply the main effect for dose by the corresponding interaction term. Thus, from Table 1 in the manuscript, the dose-response association for patients with a spleen size of 16cm at initiation is 1.90*1.04=1.98; however, the relevance percentage corresponding to the second term (the interaction parameter) is less than 66%, suggesting that this is very close to the dose-response association in the reference spleen size group, i.e. 1.90.

Calibration

Figures S2 and S3 more precisely characterize calibration of the response models. The difference between model-based expected response rate of each site (an open square) and the observed response rate (a circle) in that site is plotted for the Early model (top panel), Late model (middle panel), and Late-given-Early model (lower panel). Within a panel, patients are stratified based upon their baseline spleen size (Figure S2) or baseline platelet count (Figure S3). The line segment represents the interval in which the response rate would be expected to fall in a perfectly calibrated model. From Figure S2, the calibration of our models improves with larger baseline spleen size. In other words, patients with small (< 5 cm) spleens had better-than-expected response rates (in the case of MA) or worse-than-expected response rates (PST1 and PST2). In contrast, response rates in patients with large (> 20 cm) spleens generally agreed with the models' predictions.

References

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *International Conference on Artificial Intelligence and Statistics*, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97,** 465–480.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine* **27,** 2865–2873.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Taylor & Francis, 2 edition.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton, FL.

Piironen, J. and Vehtari, A. (2016). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *arXiv preprint arXiv:1610.05559* .

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons.

Stan Development Team (2016). RStan: the R interface to Stan. R package version 2.14.1.

Su, Y.-S., Gelman, A., Hill, J., Yajima, M., et al. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software* **45,** 1–31.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76,** 1049–1064.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine* **30,** 377–399.

Zhou, X. and Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician* **64,** 159–163.