Individualized risk trajectories for iron-related adverse outcomes in repeat blood donors

W. Alton Russell1,2, David Schienker1,3,4,5, Brian Custer2,6

1Department of Management Science and Engineering, Stanford University Stanford, CA

2Epidemiology and Health Policy Science, Vitalant Research Institute, San Francisco, CA

3Lucile Packard Children’s Hospital Stanford, Palo Alto, CA, USA

4Pediatric Endocrinology, Stanford School of Medicine, Palo Alto, CA, USA

5Clinical Excellence Research Center, Stanford School of Medicine, Palo Alto, CA, USA

6Department of Laboratory Medicine, University of California, San Francisco, CA

**Corresponding author:** W. Alton Russell, Management Science and Engineering, Stanford University, Stanford CA 94305. email: [altonr@stanford.edu](mailto:altonr@stanford.edu).

**Key words:**

**Running title:**

# Abstract

**Background:**

**Methods:**

**Results:**

**Conclusions:**

# Introduction

Observational studies show that frequent blood donation can cause or exacerbate iron deficiency, with higher incidence among teen donors and premenopausal women [1–6]. Potential donors are screened using fingerstick hemoglobin or hematocrit tests and deferred if levels are below a minimum cutoff. Such hemoglobin deferrals prevent some collections from iron deficient donors but also consume time and resources from both donor and blood center, decreasing donor satisfaction and the likelihood of future donations [7]. Because fingerstick hemoglobin is an unreliable indicator of true iron stores, many donors qualify to donate despite having low or absent underlying iron stores [4]. More reliable measures of iron status include ferritin, zinc protoporphyrin, soluble transferrin receptor, and hepcidin, but these are more costly to measure and not available as point of care tests [8]. Past studies have identified several factors that increase risk of iron deficiency among blood donors. The Danish Blood Donor Study found that sex, menopause status, and donation history were the strongest predictors of iron deficiency among donors. Weight, age, vitamin use, and diet were also significant [5]. Similar results have been found for donors in the United States, Australia, and the Netherlands [1–4,6]. Other studies have analyzed predictors for a low hemoglobin deferral for repeat blood donors, identifying age, time since last donation, and donation history as strong predictors [9,10]. To our knowledge, no prediction model has been developed that simultaneously considers the competing risks of hemoglobin deferral and of collecting blood from a donor whose iron stores are low or absent at the time of collection.

In this study, we developed machine learning models to estimate how risk of hemoglobin deferral and completed donations from donors with low or absent iron stores develop as a function of time from an index donation to a subsequent donation attempt in a cohort of donors from the REDS-II Iron Status Evaluation (RISE) study. We also compared predictive performance with and without ferritin and soluble transferritin receptor (STfR), two biomarkers that were available for our study population but are not routinely collected by most US blood centers.

# Methods

Using data from the RISE study, we trained multiclass prediction models to predict the risk of three iron-related adverse outcomes at a subsequent donation attempt based on the time until the donor returned. The adverse outcomes were hemoglobin deferral and completing a donation with either low or absent iron stores. We assessed the models’ predictive performance, compared performance with and without the inclusion of two non-routine biomarkers (ferritin and sSTfR) as features for prediction, and generated and analyzed individual risk profiles for each donor’s likelihood of iron-related adverse donation outcomes at their next visit as a function of how long until they return.

## Data preprocessing and formatting

The RISE dataset contains data on all visits to a blood center for 2,425 donors over a 2-year period. Data elements include past donation history, biometrics for each visit, and questionnaire responses regarding demographics, diet, supplemental iron consumption, female reproductive health, and demographics from a baseline and final visit. We used 42 features from an index donation together with the time until the donor returns to predict the outcome of a follow-up donation attempt. We assumed that donor characteristics measured at the baseline visit such as diet, vitamin use, smoking, and female reproductive health would not change significantly over the study period, and we used them to predict outcomes following subsequent donations by the same donor. We re-coded or imputed missing values for some fields; ` contains these details for all features used for prediction. We also included a composite dietary iron consumption score that was generated for each donor in the RISE dataset as part of another secondary analysis [11].

To generate the model development dataset, we considered donations with at least 150 mL of red blood cell loss as potential index donations, which included whole blood donations, apheresis red blood cell donations, and some donations that were classified as ‘quantity not sufficient.’ We excluded potential index donations that were double red cell donations or that were missing a measurement of ferritin and either fingerstick hemoglobin or hematocrit. If follow-up visits were recorded after potential index donations, we generated labels with the time until the follow-up visit (in days) and its outcome. For all index donations followed by a visit with significant iron loss, defined as a loss of at least 55 mL of red blood cells, we generated a label for the index donation based on the first such follow-up visit. We also generated labels for each index donation based on any follow-up visits that did not result in significant iron loss (i.e., visits resulting in a deferral or apheresis donations of platelets or plasma with <55 mL of red blood cell loss) if they occurred before any follow-up visits with significant iron loss. For each index donation , the outcome of its follow-up visits () was classified as hemoglobin deferral () if one were recorded; as a low iron donation () if pre-donation ferritin was mg/dl and mg/dl for women or mg/dl and mg/dl for men; as an absent iron donation () if pre-donation ferritin was <12 mg/dl; and as a ‘no adverse outcome’ donation otherwise (). We excluded follow-up donations without ferritin measurements from the model development dataset.

## Prediction model development

### Model selection

We evaluated several candidate model types: gradient boosted machines, random forest, regression trees, and generalized linear models with elastic net regularization (with and without all second order interaction terms). For each model type we evaluated multiple hyperparameter settings via grid search, with specific hyperparameter values described in . We implemented a nested cross validation procedure with resampling to minimize bias in model selection and assessment [12]. In this process, we generated 15 *model assessment partitions* which consisted of 3 resamples of 5 equal-sized partitions of the entire dataset that were generated with stratified sampling to ensure the distribution of outcomes was balanced across partitions. For each model assessment partition, we defined all data not included in the partition as the corresponding *model tuning set*. Within the 15 tuning sets, we assessed all candidate model configurations (model type and hyperparameter setting) using 5-fold validation, assessing the multiclass area under the reliever operator characteristic curve (multiclass AUC) using the Hand and Till method [13]. We compared model configurations based on the average multiclass AUC across 5 cross validation folds averaged over all 15 tuning sets (assessing a total of 75 realizations of each candidate model configuration).

We also considered ensemble models, which combine the risk scores from multiple base models. We assessed two methods of combining risk scores from base models: a simple average and a weighted average, for which we weighted each model’s score proportionally to its accuracy raised to a power of four as suggested by Large et. al. [14] (**Q: the simple average outperformed the wieghted average. Should I just take this out and pretend we never did it?**). We assessed AUC for each candidate ensemble configuration across the same 5 cross validation folds within each of the 15 tuning sets.

we selected the top model configuration based on the highest multiclass AUC. To produce an unbiased assessment of the selected model configuration, we then assessed multiclass AUC on each of the 15 model assessment partitions. For each assessment partition, we trained the model configuration on all data not in the partition and used this model to generate risk scores on the assessment partition. We then calculated multiclass AUC for the partition.

To assess impact of measuring ferritin and STfR on ability to predict iron-related adverse outcomes at follow-up donation attempts, we repeated this model development process twice. In the “extra biomarkers” model, we used ferritin and soluble transferrin receptor and derived measures as variables for prediciton. Because these biomarkers are not currently measured for most blood donations in the United States, we also developed a “standard biomarkers” version that excluded ferritin and soluble transferritin receptor.

### Feature importance

For the two selected model configurations, we also assessed the importance of features for prediction using a random permutation method [15]. In this method, we randomly sample one column of the dataset in each model assessment partition and generate risk scores using a model developed using the top configuration trained on all data not in that partition. We then calculated the percent decrease in multiclass AUC. We assessed feature importance on each of the 15 tuning datasets and ranked features acording to the median decrease in multiclass AUC acros each tuning set.

### Calibration

To generate the final model, we retrained the selected model configuration on the entire model development dataset and then calibrated the predicted probabilities to the first return dataset. For this, we estimated the distribution of outcomes for follow-up visits from the first return dataset by assuming that the distribution of absent, low, and ‘no-adverse outcome’ donations in follow-up donations at which ferritin was not measured would be the same as for those with ferritin measured. Mathematically, we first totaled each follow-up outcome as , where correspond to the outcomes described above. We then calculated , an estimation of what the totals would have been if ferritin were measured for all follow-up donations, as (hemoglobin deferral) and for (completed donations). We then used our top model to generate the unnormalized probability vector for each index donation . We computed weights for the unnormalized probability of each outcome by solving the system of equations for . The final calibrated model used parameters , , and together with the uncalibrated scores from the model to produce the estimated likelihood of each outcome at a follow-up donation as where . This ensured that the expectation of the distribution of the predicted outcome for the first return dataset would correspond to our estimated totals .

# Risk trajectory analysis

For each index donation, we generated a risk trajectory by predicting the likelihood of each outcome at the donor’s next donation attempt using the calibrated model for every followup donation interval between 56 and 256 days, and we generated graphical representations of individual donors’ risk trajectories. We analyzed how risk of adverse events evolved from a post-donation interval of 56 days to one of 256 days. We performed multiple subgroup analyses, where-in we segmented the donors by one of the variables. For each subgroup, we calculated the mean and 95% confidence interval for the estimated risk of each adverse event for each donation interval from 56 to 250 days, which we plotted.

# Results

## Data processing

In the RISE dataset, a total of 7817 donations from 1922 donors were followed by at least one follow-up visit. We removed 520 index donations because hemoglobin was not recorded, and we removed a further 18 index donations from the first return dataset because the follow-up visit was less than 56 days later. The first return dataset contained 7279 index donations labeled with the outcome of the first follow-up donation. That outcome was a hemoglobin deferral for 636 index donations; a low-iron donation for 754; an absent iron donation for 568; no adverse outcome for 1340; and a completed donation with unknown iron status for 3981. The model development dataset included 3529 unique index donations from 1543 donors. 3149 index donations were labeled with one follow-up donation, 289 were labeled twice, and 91 were labeled with 3 or more follow-up visit outcomes (maximum of 8).

## Prediction model

We evaluated over 2,000 model configurations (model type and hyperparameter settings) across the five candidate model types. shows the average overall AUC within the 15 tuning datasets for each model configuration, and shows the top hyperparameter setting for each model type. The overall top model was a gradient boosted decision tree. It had an AUC of 75.0% – 75.8% across the 15 tuning set and a mean AUC of 76.1% as estimated using the outer cross-validation folds. As shown in , discriminative performance was highest for predicting no adverse outcome donations and lowest for predicting low iron donations. In a secondary analysis, we found that use of ferritin, soluble transferrin receptor, and body iron increased the overall AUC from 77% to 82% among the subset of donations for which those values were recorded. Inclusion of these biomarkers increased discriminative performance most substantially for identifying absent iron donations and had little effect on ability to discriminate hemoglobin deferrals.

Variable importance for the top model in the primary analysis is shown in and for the model using extra biomarkers as predicters in . In the primary analysis, hemoglobin and return time were most important for predicting the outcome of a follow-up donation. When additional biomarkers were used as predictors, ferritin became the most important.

We calculated normalization weights to calibrate the model scores to the expected distribution of outcomes in the first return dataset. They were 1.4 for the probability of no adverse outcome; 0.47 for the probability of a hemoglobin deferral; 1.1 for the probability of a low iron donation; and 1.2 for the probability of an absent iron donation.

## Individual risk profiles

Figure shows the individual risk trajectories from two donations: one for a donor whose risk of an adverse outcome was high at day 56 but declined over time and another for a donor who had a low risk of adverse outcomes even at day 56. shows the same plots for 60 randomly selected index donations from the first return dataset. Notably, estimated risk did not monotonically decrease for all adverse events for all donors. For example, the risk of a low iron donation increased as risk of hemoglobin deferral or an absent iron donation fell for some donors.

Figure shows the probability of any adverse outcome at post-donation day 56 and post-donation day 250. The median risk of any adverse outcome at day 56 was 71% (IQR 43% – 86%), but this dropped to 23% (IQR 12% – 41%) at post-donation day 250. While risk of an adverse outcome fell for most donors, some continued to have a high risk even at post-donation day 250. For 787 donors (11%), estimated risk of any adverse outcome was above 60% at post-donation day 250, which may indicate an underlying iron-related condition unrelated to repeat blood donation.

shows different trajectory types.

is by iron status.

is by tertile of venous hemoglobin.

shows different gender.

shows tertiles of red blood cells lost.

shows whether people take daily iron supplementation.

shows tertiles of the composite iron scores.

# Discussion

Risk of iron-related adverse outcomes at follow-up donations can be estimated as a function of the return time using data available at an index donation. Estimated risk decreased precipitously for most donors if they waited longer to return, suggesting that tailoring donors’ IDIs to individual donors’ risk profiles may be an effective strategy for managing risk of iron-related adverse donation outcomes without unduly restricting return donations from low-risk donors. Risk for some donors remained high even 250 days later, suggesting that these methods could also be used to identify individuals who may be poor candidates for repeat blood donation due to underlying iron deficiency. Including ferritin as a predictor improved risk estimation, particularly with respect to estimating risk of absent iron donations.

Our analysis has several limitations. The RISE study asked participants to commit to frequent blood donation and targeted recruitment to achieve proportional representation based on gender and donation history [16]. Further study is needed to assess the generalizability of our prediction model’s performance to a general blood donor population. Additionally, the outcomes we estimated at the population level are specific to the RISE cohort. In particular, the baseline rate of adverse outcomes and the reduction in supply introduced by tailored IDIs may be lower in populations with lower return rates. Another limitation of the RISE data is that ferritin and other biomarkers were not measured for all follow-up visits. We factored this into our analysis by assuming these biomarkers were missing at random, but this may not be the case.

Two other limitations must be overcome before tailored IDIs can be implemented in practice. First, a decision-maker must identify risk thresholds for each adverse outcome in our method. Experts may not agree on the level of risk that is acceptable and how the sufficiency of the blood supply should be weighed against risks to donors. Further work is needed to understand these trade-offs and identify reasonable risk thresholds. Second, this method may face significant barriers to implementation. Sophisticated machine learning techniques for decision-making require technical expertise to develop and maintain and are opaque in the sense that humans cannot readily understand how the system arrived at a decision. In a growing literature on interpretable machine learning, methods have been developed for constructing simpler decision rules that sometimes perform on par with advanced machine learning techniques [17–21]. Further work is needed to assess barriers to the adoption of the tailored IDI method developed here and to determine whether simpler decision rules might be easier to implement and perform similarly. Despite these limitations, our analysis suggests that individual risk prediction could be a useful tool for ensuring a sufficient blood supply while managing iron-related risks to repeat blood donors.

# Declarations

**Funding:** A

**Conflicts:** A

**Ethics/Consent:** A

**Data and materials:** A

**Code availability:** A

**Authors’ contributions:**

# References

1. Cable RG, Glynn SA, Kiss JE, et al. Iron deficiency in blood donors: the REDS-II donor iron status evaluation (RISE) study. *Transfusion*. 2012;52(4):702-711. doi:[10.1111/j.1537-2995.2011.03401.x](https://doi.org/10.1111/j.1537-2995.2011.03401.x)

2. Salvin HE, Pasricha SR, Marks DC, Speedy J. Iron deficiency in blood donors: A national cross-sectional study. *Transfusion*. 2014;54(10):2434-2444. doi:[10.1111/trf.12647](https://doi.org/10.1111/trf.12647)

3. Spencer BR, Bialkowski W, Creel DV, et al. Elevated risk for iron depletion in high-school age blood donors. *Transfusion*. 2019;59(5):1706-1716. doi:[10.1111/trf.15133](https://doi.org/10.1111/trf.15133)

4. Baart AM, Van Noord PAH, Vergouwe Y, et al. High prevalence of subclinical iron deficiency in whole blood donors not deferred for low hemoglobin. *Transfusion*. 2013;53(8):1670-1677. doi:[10.1111/j.1537-2995.2012.03956.x](https://doi.org/10.1111/j.1537-2995.2012.03956.x)

5. Rigas AS, Sørensen CJ, Pedersen OB, et al. Predictors of iron levels in 14,737 Danish blood donors: results from the Danish Blood Donor Study. *Transfusion*. 2014;54(3 Pt 2):789-796. doi:[10.1111/trf.12518](https://doi.org/10.1111/trf.12518)

6. Patel EU, White JL, Bloch EM, et al. Association of blood donation with iron deficiency among adolescent and adult females in the United States: a nationally representative study. *Transfusion*. 2019;59(5):1723-1733. doi:[10.1111/trf.15179](https://doi.org/10.1111/trf.15179)

7. Custer B, Chinn A, Hirschler NV, Busch MP, Murphy EL. The consequences of temporary deferral on future whole blood donation. *Transfusion*. 2007;47(8):1514-1523. doi:[10.1111/j.1537-2995.2007.01292.x](https://doi.org/10.1111/j.1537-2995.2007.01292.x)

8. Kiss JE, Vassallo RR. How do we manage iron deficiency after blood donation? 2018;181:590-603. doi:[10.1111/bjh.15136](https://doi.org/10.1111/bjh.15136)

9. Baart AM, De Kort WLAM, Moons KGM, Vergouwe Y. Prediction of low haemoglobin levels in whole blood donors. *Vox Sanguinis*. 2011;100(2):204-211. doi:[10.1111/j.1423-0410.2010.01382.x](https://doi.org/10.1111/j.1423-0410.2010.01382.x)

10. Baart AM, De Kort WLAM, Atsma F, Moons KGM, Vergouwe Y. Development and validation of a prediction model for low hemoglobin deferral in a large cohort of whole blood donors. *Transfusion*. 2012;52(12):2559-2569. doi:[10.1111/j.1537-2995.2012.03655.x](https://doi.org/10.1111/j.1537-2995.2012.03655.x)

11. Spencer BR, Fox M, Wise L, Cable R. A composite measure of heme iron consumption predicts incident iron depletion in repeat blood donors. In: *Abstract of 29th Regional Congress of the ISBT*. Vol 114. Basel: Vox Sanguinis; 2019:5-240. doi:[10.1111/vox.12792](https://doi.org/10.1111/vox.12792)

12. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7(1):91. doi:[10.1186/1471-2105-7-91](https://doi.org/10.1186/1471-2105-7-91)

13. Hand DJ, Till RJ. A Simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*. 2001;45(2):171-186. doi:[10.1023/A:1010920819831](https://doi.org/10.1023/A:1010920819831)

14. Large J, Lines J, Bagnall A. A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. *Data Mining and Knowledge Discovery*. 2019;33(6):1674-1709. doi:[10.1007/s10618-019-00638-y](https://doi.org/10.1007/s10618-019-00638-y)

15. Breiman L. Random forests. *Machine Learning 2001 45:1*. 2001;45(1):5-32. doi:[10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)

16. Cable RG, Glynn SA, Kiss JE, et al. Iron deficiency in blood donors: Analysis of enrollment data from the REDS-II Donor Iron Status Evaluation (RISE) study. *Transfusion*. 2011;51(3):511-522. doi:[10.1111/j.1537-2995.2010.02865.x](https://doi.org/10.1111/j.1537-2995.2010.02865.x)

17. Jung J, Concannon C, Shroff R, Goel S, Goldstein DG. Simple rules for complex decisions. *arXiv*. 2017. doi:[10.1145/nnnnnnn.nnnnnnn](https://doi.org/10.1145/nnnnnnn.nnnnnnn)

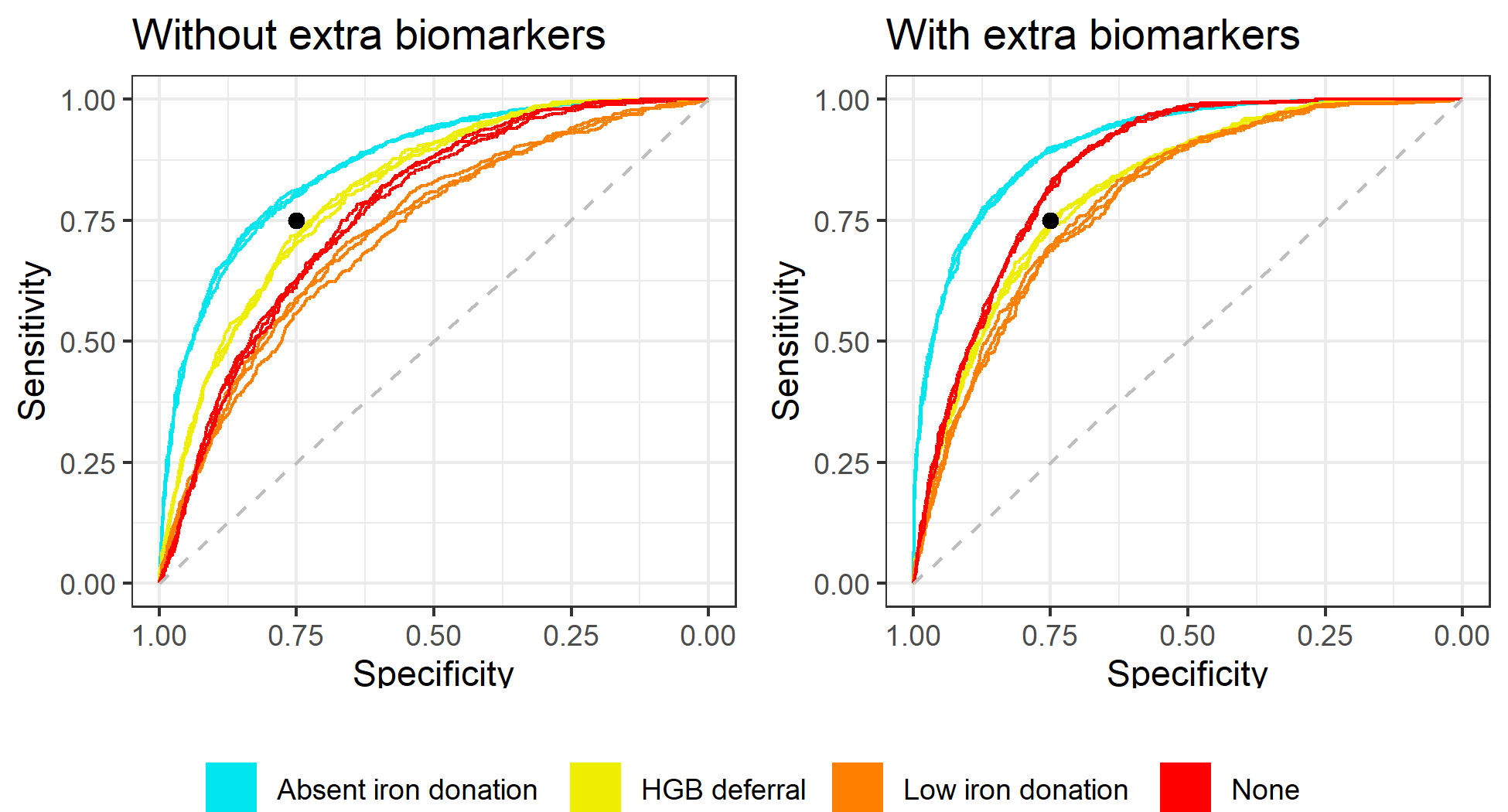
18. Ustun B, Rudin C. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*. 2016;102(3):349-391. doi:[10.1007/s10994-015-5528-6](https://doi.org/10.1007/s10994-015-5528-6)

19. Letham B, Rudin C, McCormick TH, Madigan D. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*. 2015;9(3):1350-1371. doi:[10.1214/15-AOAS848](https://doi.org/10.1214/15-AOAS848)

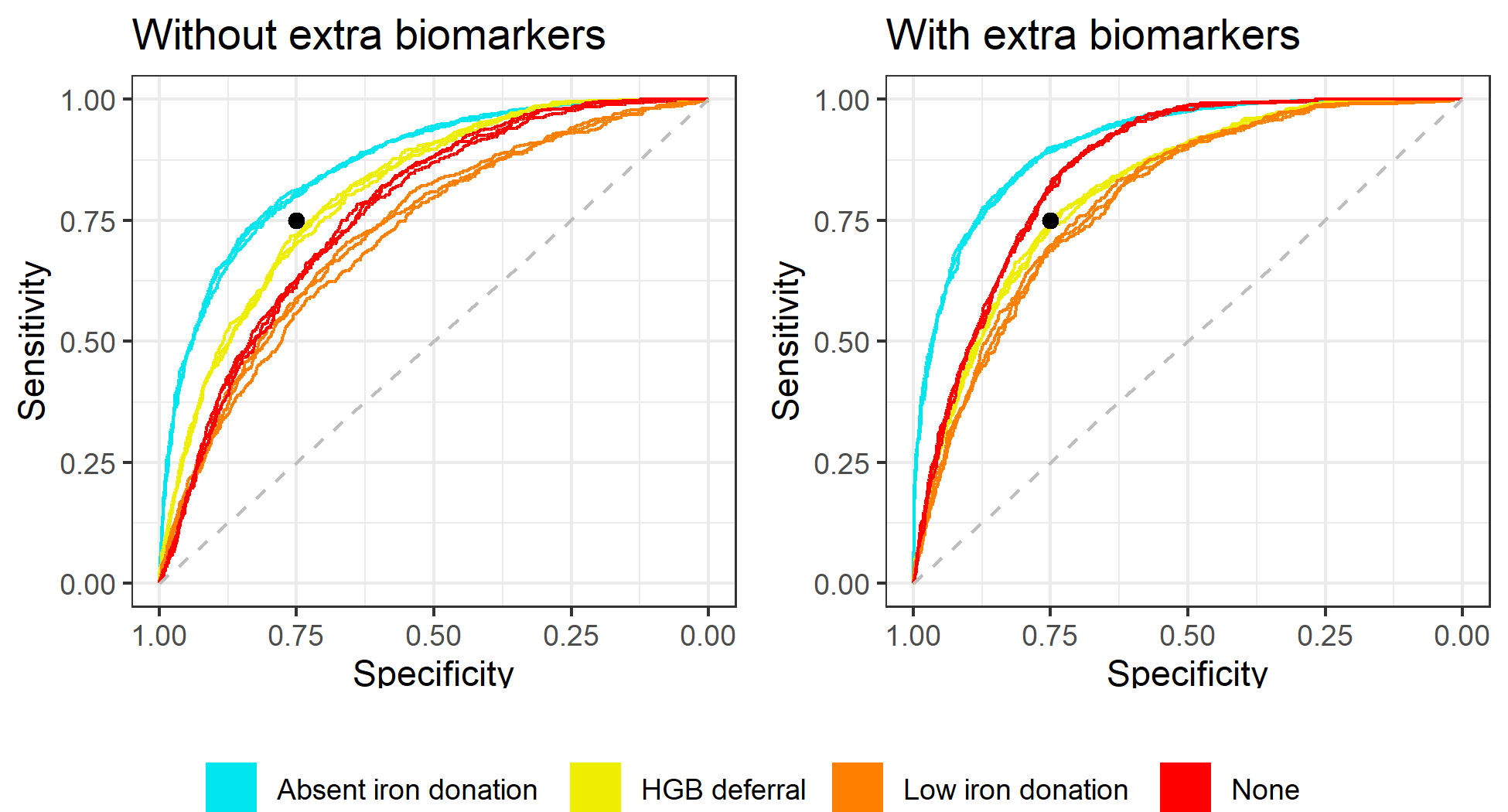
20. Lakkaraju H, Bach SH, Leskovec J. Interpretable decision sets: A joint framework for description and prediction. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol 13-17-Augu. New York, NY, USA: Association for Computing Machinery; 2016:1675-1684. doi:[10.1145/2939672.2939874](https://doi.org/10.1145/2939672.2939874)

21. Ustun B, Rudin C. Learning optimized risk scores. *Journal of Machine Learning Research*. 2019;20(150):75. <http://arxiv.org/abs/1610.00168>.

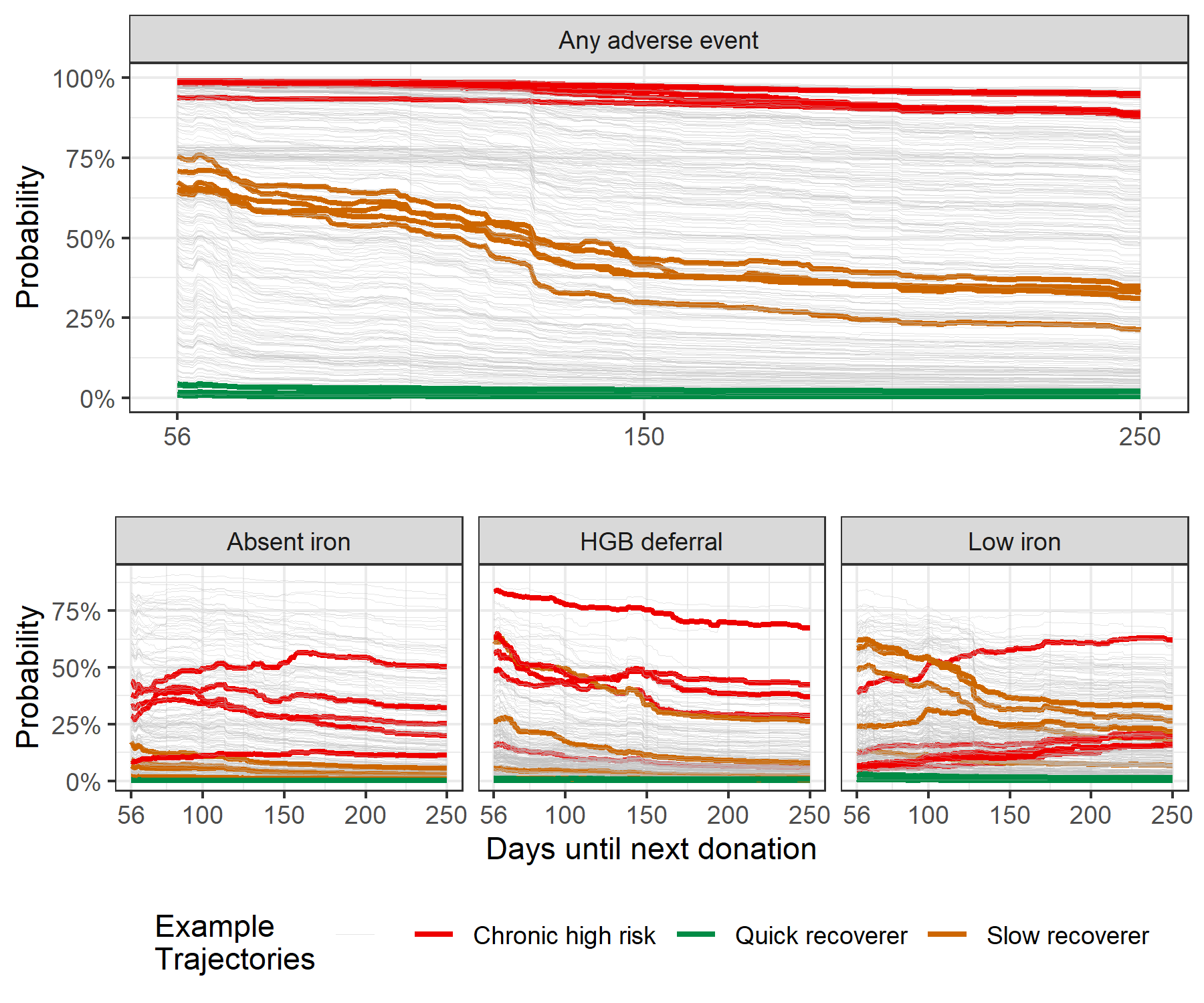
# Figures



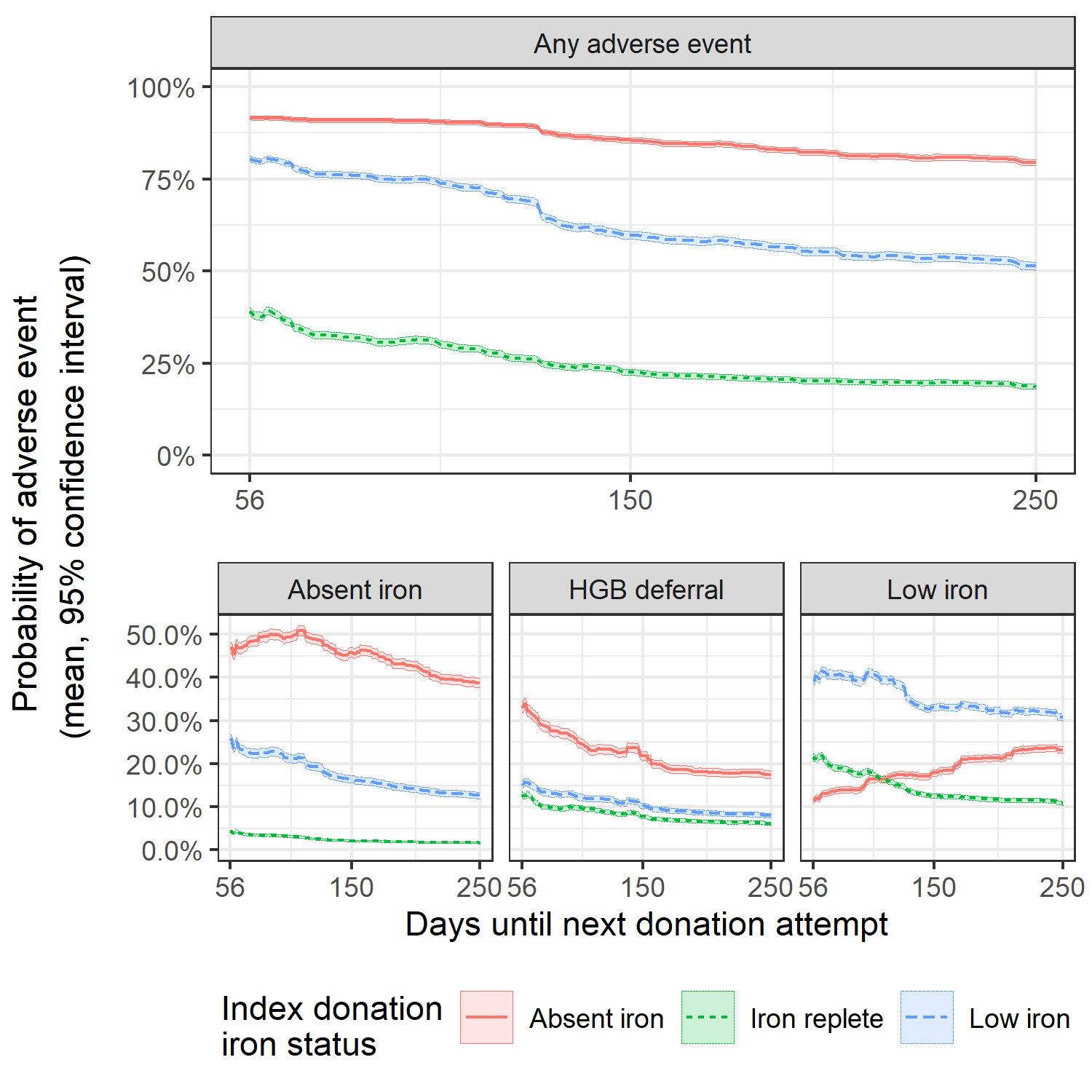
**Figure :** One vs. all ROC curves with and without ferritin, soluble transferrin receptor, and derived measures. Black dot at 75% sensitivity and 75% specificity for visual reference.



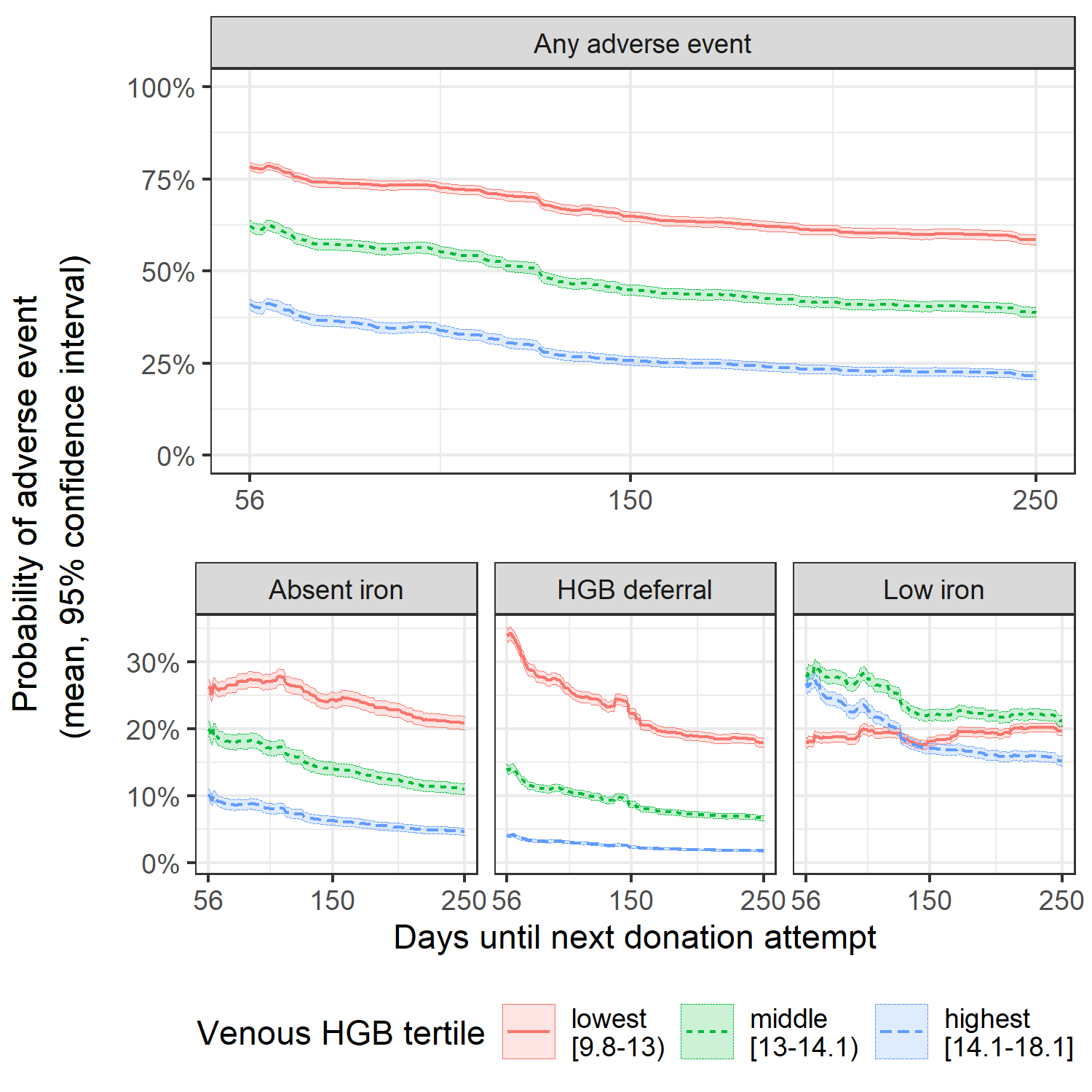
**Figure :** Relative importance of features for the top "extra biomarkers" and "Standard biomarkers" model. Features were included if among the top 15 most important features for either model.



**Figure :** Risk trajectory for any adverse event (top) or a specific adverse event (bottom) for 100 randomly-selected donors. Five 'chronic high risk' donors with more than 80% risk of an adverse event if they return on post-donation day 250 are shown in red; five 'quick recoverer' donors with less than 5% risk of an adverse event if they return on post-donation day 56 shown in green; and 5 'slow recoverer' donors whose risk of an adverse event was above 60% on post-donation day 56 but fell below 35% by postdonation day 250.



**Figure :** .



**Figure :** .

# Supplemental materials

# A. Calculations for estimating the outcomes of pathogen inactivation

Blank.

# Supplemental tables

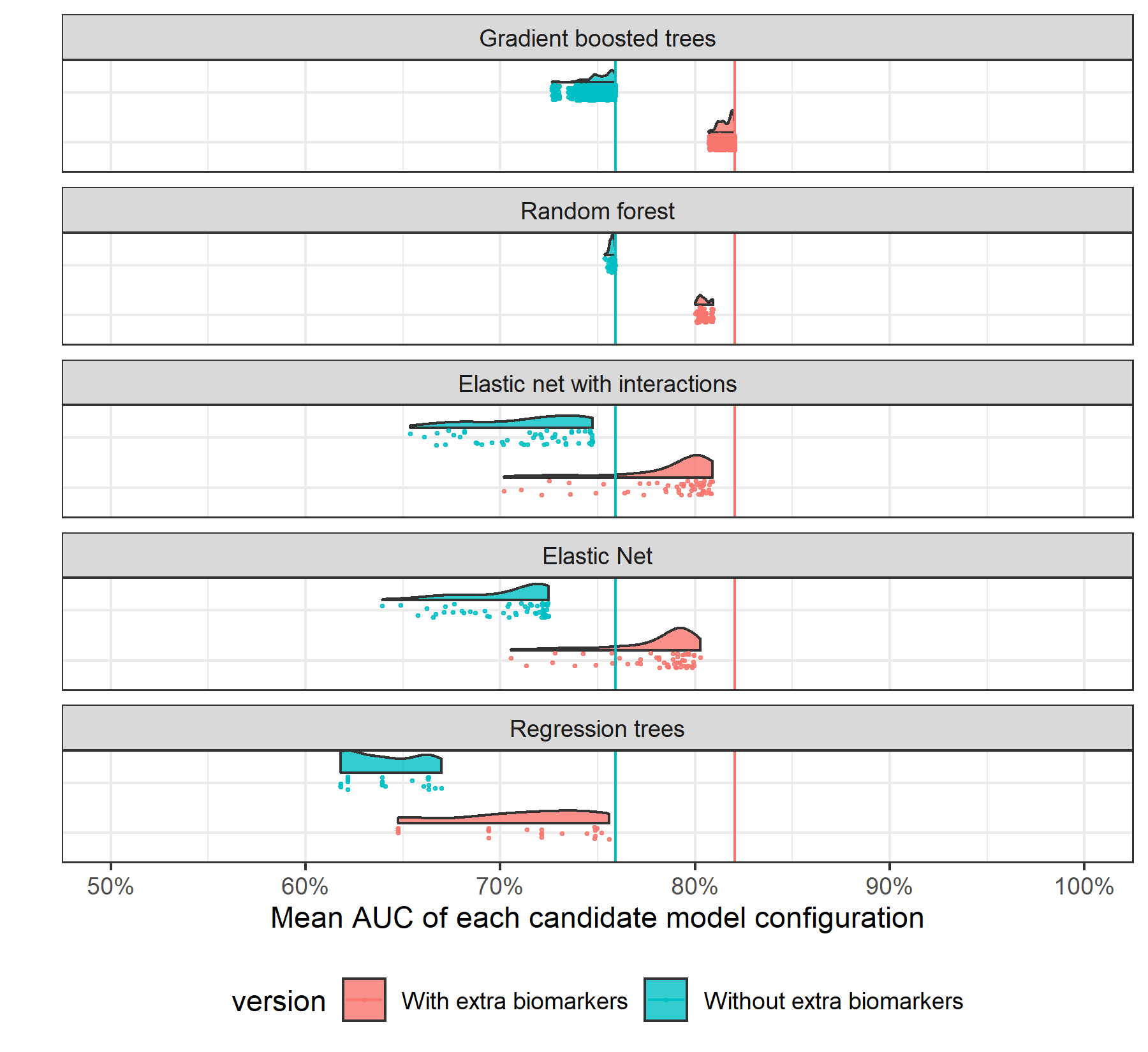
**Table S:** List of features for prediction model with description and notes from feature engineering.

| **Variable name** | **Description** | **Feature engineering** |
| --- | --- | --- |
| **Donation history** | | |
| DER\_RBC\_Last12months | Red blood cell loss in last 12 months | Used as-is. |
| DER\_RBC\_Last24months | Red blood cell loss in last 24 months | Used as-is. |
| DER\_RBCLoss\_Units | Units of red blood cell lost at this donation | Used as-is. |
| DER\_RBCLoss\_mL | Volume of red blood cell loss at this donation (mL) | Used as-is. |
| DER\_DaysRBCLoss | Days since last red blood cell loss | Set NAs (36%) to 10 years ago (10\*365 days) |
| DER\_DaysDRLoss | Days since last double red blood cell donation | Set NAs (36%) to 10 years ago (10\*365 days) |
| RQ1\_Ever\_Donated | Ever donated blood before | Recoded to 0 (no), 1 (yes); set as 1 for all followup donations since they gave an index donation |
| cumLifetimeDonations | Total lifetime donations | If lifetime donations missing from baseline questionaire, set to 0. For follow-up visits, increased by 1 for all donations. |
| **Biometric** | | |
| FingerstickHGB\_equiv | Fingerstick HGB or HCT | Used fingerstick HGB if measured. Otherwise, used fingerstick HCT/3.04. If also missing, used adjusted venous HGB. |
| DD\_ABO\_RH | ABO-Rh blood type | 1 donor has UNT. Setting to most common value in dataset, O+ (40% of donors at baseline) |
| DER\_AdjVenousHgb | Venous HGB (converted to pre-donation if measured from a post-donation sample) | Not collected at all follow-up visits. If missing, set to fingerstick value |
| DER\_Weight | Weight (pounds) | Missing for 5 donors with 22 follow-up donations. 4 M, 1 F, ages 27 - 69. Imputing mean weight by gender among those aged >25 |
| DER\_Height | Height (inches) | Missing for 6 donors (same 5 missing weight plus one additional) with 23 follow-up donations. Age 27 - 69. Imputing mean height by gender amon those age >25. |
| BMI | Body Mass Index | Missing for same 6 donors missing height. Calculating from (imputed) weight and height |
| DER\_EBV | Estimated Blood Volume (Nadler's equations) | None missing |
| DER\_RedCellVolume | Total body red blood cell volume | Calculated total body red cell volume; none missing |
| DER\_PercentRBCLoss | Estimated percent red blood cell volume loss at visit | Percent of red cell lost at donation; none missing |
| DER\_Age | Age | Available for all, 18 - 87 |
| **Survey responses from baseline visit** | | |
| DD\_Country | US born | 1 if US-born, 0 otherwise. Missing for <1%; if missing re-coding as US-born (>95%) |
| DD\_Gender | Gender | no missing. 52% female at baseline; 51% for followups |
| DD\_Raceth | Race/ethnicity (Asian, Black, Hispanic, Other, White) | 0.7% missing; Recoding to "O" (other) |
| RQ7\_Ever\_Smoked | Smoked at least 100 cigarettes in lifetime | Recoding to 0 (no) and 1(yes). For "don't know" (26) and "missing" (10), coding to "no" because majority (60%) |
| RQ8\_Smoked\_Past\_90Days | Smoked in last 90 days | Recoding to 0 no and 1 yes. If NA (60%) or don't know (3%), recoding to no (68% of respondents) |
| RQ11\_Liver | Liver consumption (times per week) | Recoding as times per week 0 (never), 0.5 (<1/wk), 1 (1/wk), 2 (2/wk), 3.5 (3-4x/wk), 5.5 (5-6x/wk), 7 (daily), 14 (2x or more/day); 59 are 99 (blank/don't know); setting them to most common which was never |
| RQ11\_Beef | Beef consumption (times per week) | Recoding as times per week 0 (never), 0.5 (<1/wk), 1 (1/wk), 2 (2/wk), 3.5 (3-4x/wk), 5.5 (5-6x/wk), 7 (daily), 14 (2x or more/day); 98 are 99 (blank/don't know); setting them to most common which was 2x/wk |
| RQ11\_LPCT | Lamb, pork, chicken, or turkey consumption (times per week) | Recoding as times per week 0 (never), 0.5 (<1/wk), 1 (1/wk), 2 (2/wk), 3.5 (3-4x/wk), 5.5 (5-6x/wk), 7 (daily), 14 (2x or more/day); 59 are 99 (blank/don't know); setting them to most common which was 3.5x/wk |
| RQ11\_Clams | Clams consumption (times per week) | Recoding as times per week 0 (never), 0.5 (<1/wk), 1 (1/wk), 2 (2/wk), 3.5 (3-4x/wk), 5.5 (5-6x/wk), 7 (daily), 14 (2x or more/day); 121 are 99 (blank/don't know); setting them to most common which was never |
| RQ11\_OMSS | Oysters, mussels, shrimp, or sardines consumption (times per week) | Oysters, mussels, shrimp, sardines. Recoding as times per week 0 (never), 0.5 (<1/wk), 1 (1/wk), 2 (2/wk), 3.5 (3-4x/wk), 5.5 (5-6x/wk), 7 (daily), 14 (2x or more/day); 100 are 99 (blank/don't know); setting them to most common which was less than 1/wk |
| RQ11\_OtrFish | Other fish consumption (times per week) | Recoding as times per week 0 (never), 0.5 (<1/wk), 1 (1/wk), 2 (2/wk), 3.5 (3-4x/wk), 5.5 (5-6x/wk), 7 (daily), 14 (2x or more/day); 38 are 99 (blank/don't know); setting them to most common which was <1x/wk |
| RQ11\_Eggs | Egg consumption (times per week) | Recoding as times per week 0 (never), 0.5 (<1/wk), 1 (1/wk), 2 (2/wk), 3.5 (3-4x/wk), 5.5 (5-6x/wk), 7 (daily), 14 (2x or more/day); 62 are 99 (blank/don't know); setting them to most common which was 1x/wk |
| RQ11\_Dairy | Dairy consumption (times per week) | Recoding as times per week 0 (never), 0.5 (<1/wk), 1 (1/wk), 2 (2/wk), 3.5 (3-4x/wk), 5.5 (5-6x/wk), 7 (daily), 14 (2x or more/day); 33 are 99 (blank/don't know); setting them to most common which was 1x/day |
| compositeIronScore | Composite dietary iron score | Missing for 13 at baseline. Imputing mean. |
| supp\_iron\_pct\_of\_daily | Supplemental iron | Value between 0 (no iron suppl) and 1 (daily iron suppl), where multivitamins or iron supplementation are both counted as supplemental iron. Computed as min(1, (ironsupp\_per\_week + RQ12C\_MV\_WithIron\_YN \* multivitamins\_per\_week) /7)) |
| multivitamins\_per\_week | Multivitimin consumption (times per week) | Based on RQ12A\_MultiVitamins\_YN and RQ12B\_MultiVitamins\_How\_Often. Recoding to times per week. "1: everyday" becomes 7; "2: 4-6x per week" becomes 5; "3: 1-3 days per week" becomes 2. Don't know/missing (51) set to daily (7/wk). If they answered yes to taking multivitamins but missing or don't know for how often, coding as 4-6x per week. |
| RQ17\_NumberOfPeriods | Periods in last year | Original is NA for all men and all women who say their period stopped, and for one F 19yo donor who answered NA to menstrual status question, likely because she is pre-menstrual and that wasn't an option. Coding all NAs as 0. Also, equal to 99 (don't know) for 13 donors who said yes to having periods. Coding to the most common value which is 12. |
| RQ18\_Menstrual\_Flow | Menstrual flow intensity | Recoding as 0 for no period, 1 for spotting, 2 for very light, 3 for light, 4 for moderate, 5 for heavy, 6 for very heavy/gushing. 3 women who have periods had a '9', presumably don't know or refused to answer. Recoding to most common answer, which was 4 (moderate), |
| menstrual\_flow\_times\_freq | Menstrual frequency and flow | RQ19\_MenstrualFlow/5 \* RQ17\_NumberOf\_Periods/12 |
| RQ19\_Ever\_Pregnant | Ever pregnant | Originally NA for all men and 10 women. Recoding as 0 (no) and 1 (yes), all NAs coded as 0. |
| RQ20\_NumberOfPregnancies | Number of pregnancies | Originally NA for all men and 356 women (same as never pregnant), 1-6 for other women. Coding NAs as 0. |
| RQ21\_NumberOfLiveBirths | Number of live births | Originally NA fo rall men and 356 women; Coding NAs as 0. |
| gender\_menstrating\_cohorts | Gender & menstration cohort | 3 categories: male, female menstrating (>0 periods in last year), female not menstrating (0 periods in last year) |
| **Additional biometrics available for some donations** | | |
| ARUP\_Ferritin | Ferritin (mg/dL) | Use as-is for secondary analysis. |
| ARUP\_STR | Soluble Transferrin Receptor (STfR) (nmol/L) | Use as-is for secondary analysis. |
| DER\_ARUP\_log\_Ferr | Log of ferritin | Use as-is for secondary analysis. |
| DER\_ARUP\_log\_STfR\_Ferr | STfR divided by the log of ferritin | Use as-is for secondary analysis. |
| DER\_BodyIron | Body iron in mg/kg calculated from STfR | Use as-is for secondary analysis. |

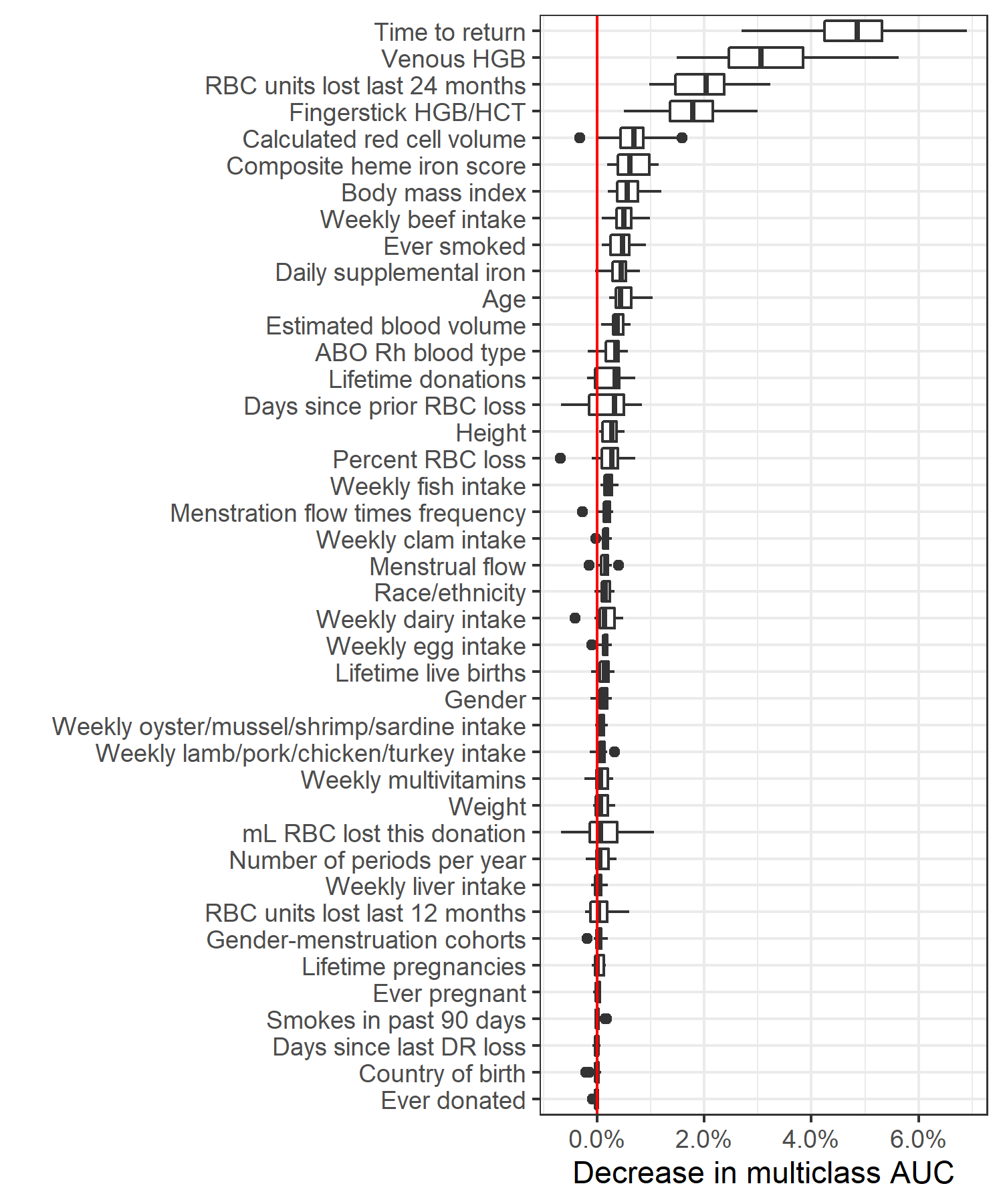
**Table S:** List of features for prediction model with description and notes from feature engineering.

| **Hyperparameters** | **Values assessed** | **Top configuration hyperparameter value** | **Top configuration AUC (mean and range)** |
| --- | --- | --- | --- |
| **Gradient boosted decision trees (R package xgboost)** | | | |
| Learning rate | 0.01, 0.05, 0.1, 0.3 | 0.01 | 75.4% (75.1 to 75.8) |
| Maximum tree depth | 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 | 18 |
| Minimum child weight | 0, 1, 2, 4, 8 | 1 |
| Row subsampling per tree | 0.65, 0.8, 1 | 0.65 |
| column subsampling per tree | 0.8, 0.9, 1 | 0.8 |
| **Random forest (R package randomForest)** | | | |
| Minimum size of terminal nodes | 1, 2, 4, 8 | 4 | 75.2% (74.7 to 75.7) |
| Maximum number of trees | 250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250, 2500 | 1500 |
| Sampling with replacement? | yes, no | yes |
| **Elasticnet penalized logistic regression (R package glmnet)** | | | |
| alpha (0 = ridge, 1 = lasso) | 0, .25, .5, .75, 1 | NA | 72.5% (72.1 to 72.8) |
| lambda (penalty weight) | 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1 | 0 |
| **Elasticnet with all second order interactions (R package glmnet)** | | | |
| alpha (0 = ridge, 1 = lasso) | 0, .25, .5, .75, 1 | 0 | 73.8% (73.3 to 74.1) |
| lambda (penalty weight) | 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1 | 0.05 |
| **Regression trees (R package rpart)** | | | |
| Complexity parameter | 0.001, 0.005, 0.01, 0.05, 0.1 | 0.005 | 66.4% (65.8 to 67.0) |
| Minimum observations per split | 10, 15, 20, 25, 30 | 20 |

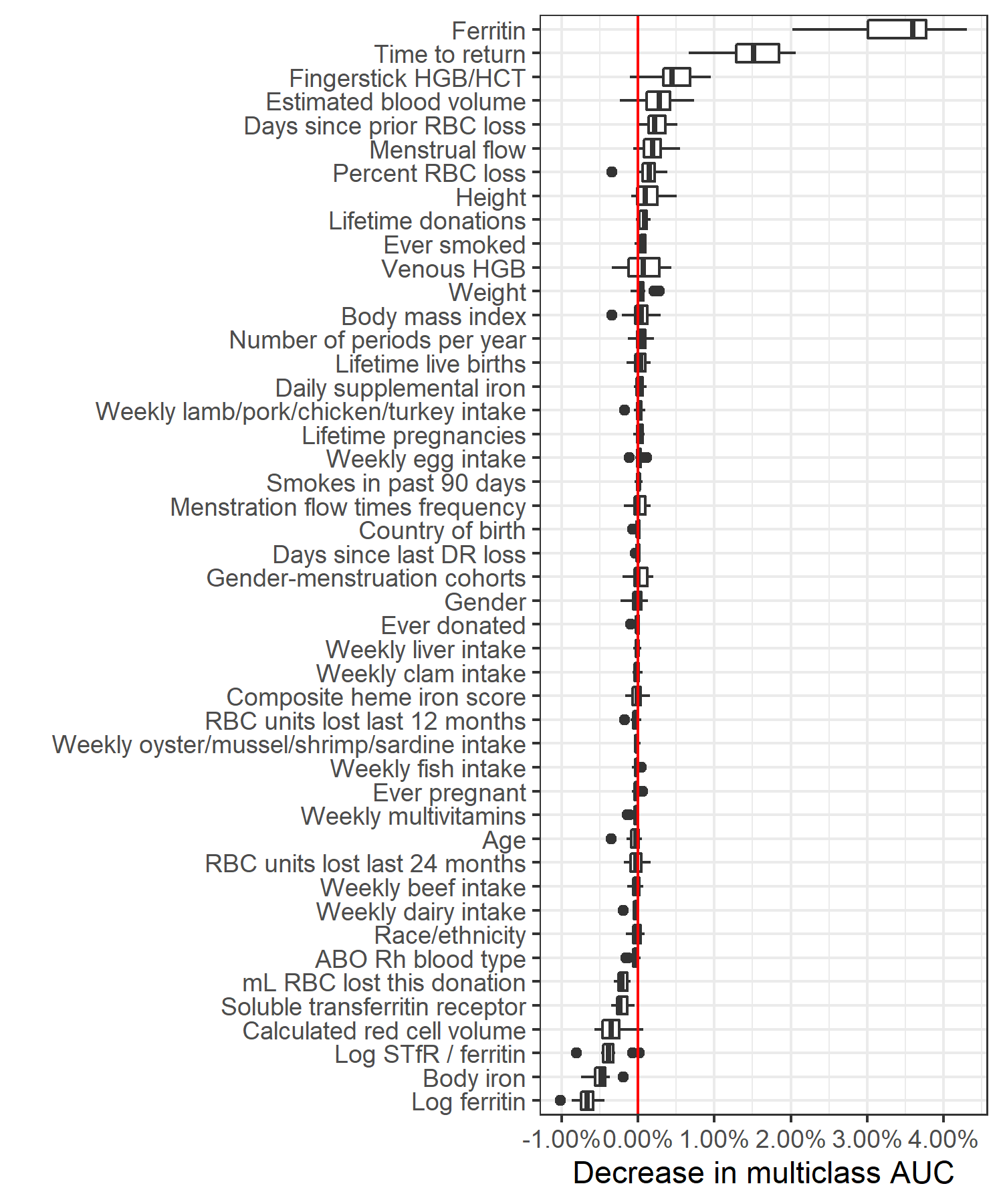
# Supplemental figures



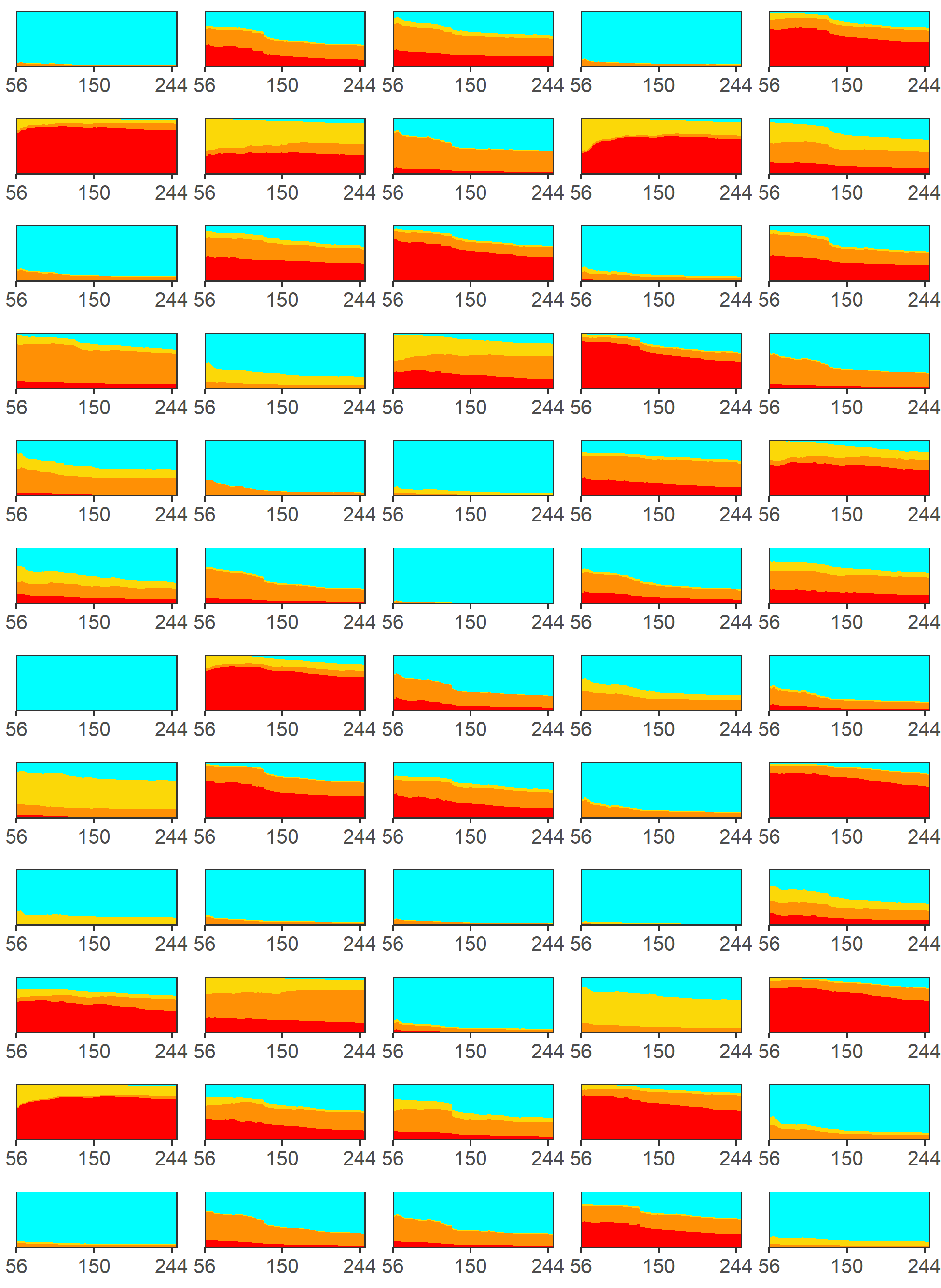
**Figure S:** Overall AUC for each evaluated model configuration assessed using 5-fold cross validation and averaged across 15 tuning sets plotted as a dot. Distributions for each version and model configuration plotted.



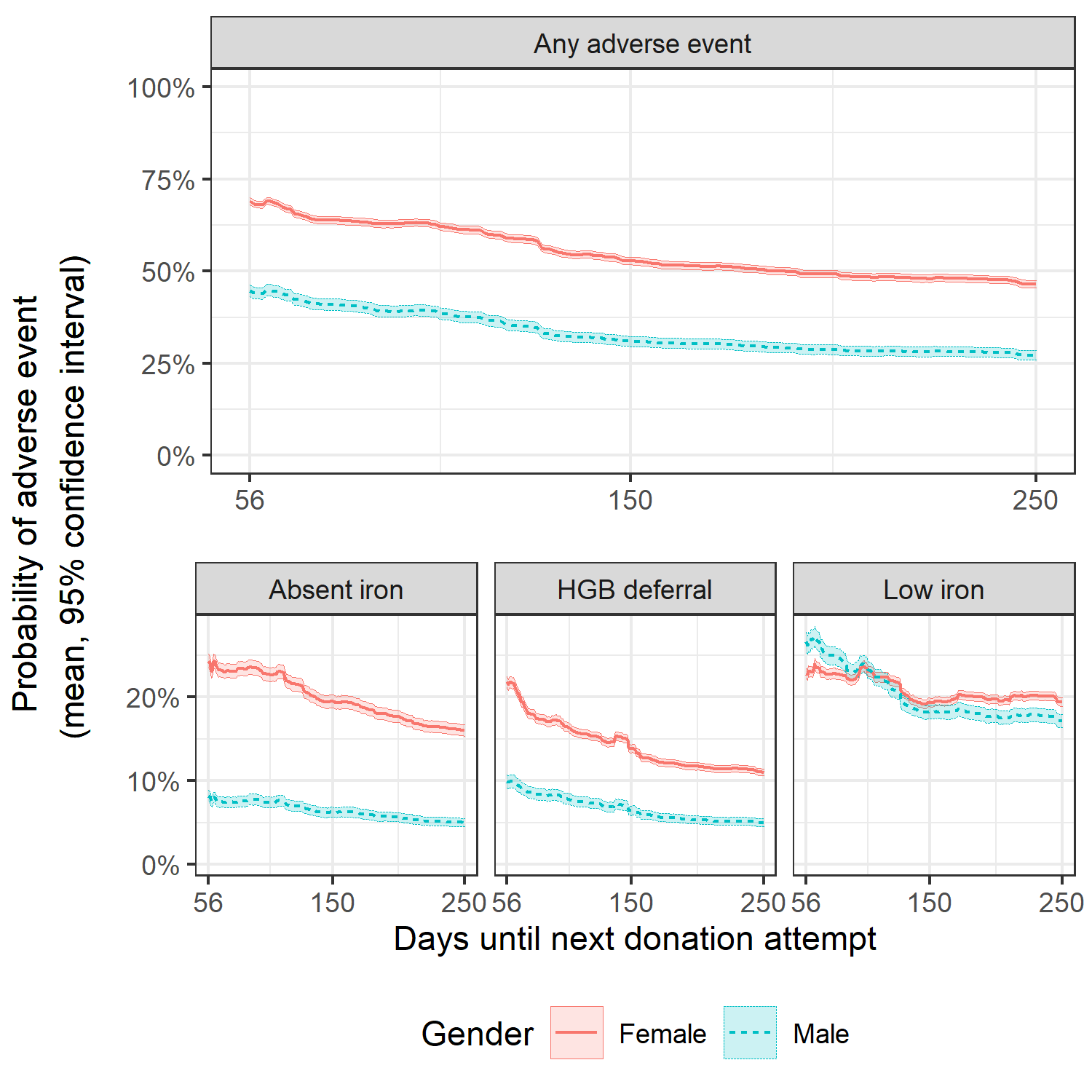
**Figure :** Relative variable importance for the top "standard biomarkers" model.



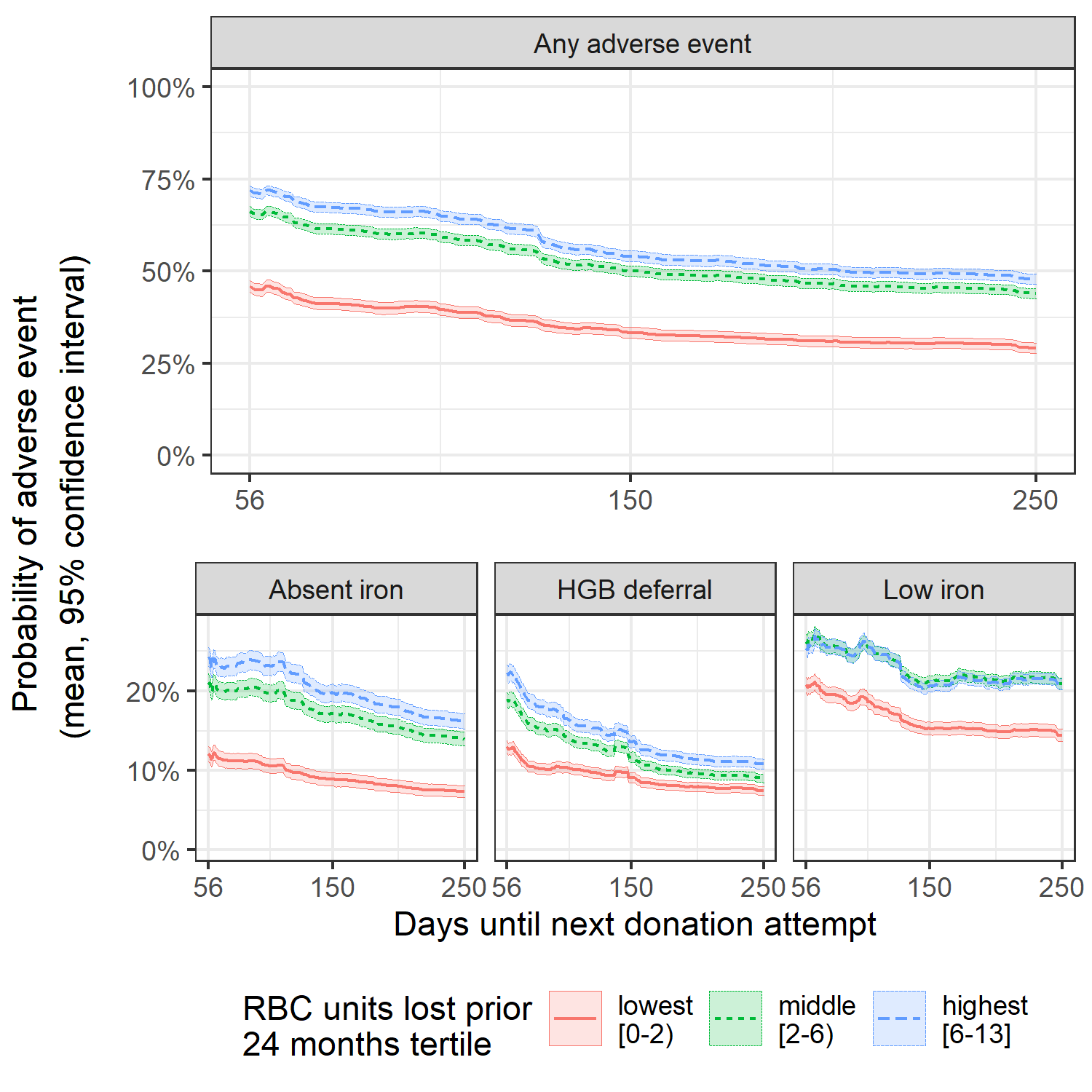
**Figure :** Relative variable importance for the top "extra biomarkers" model.



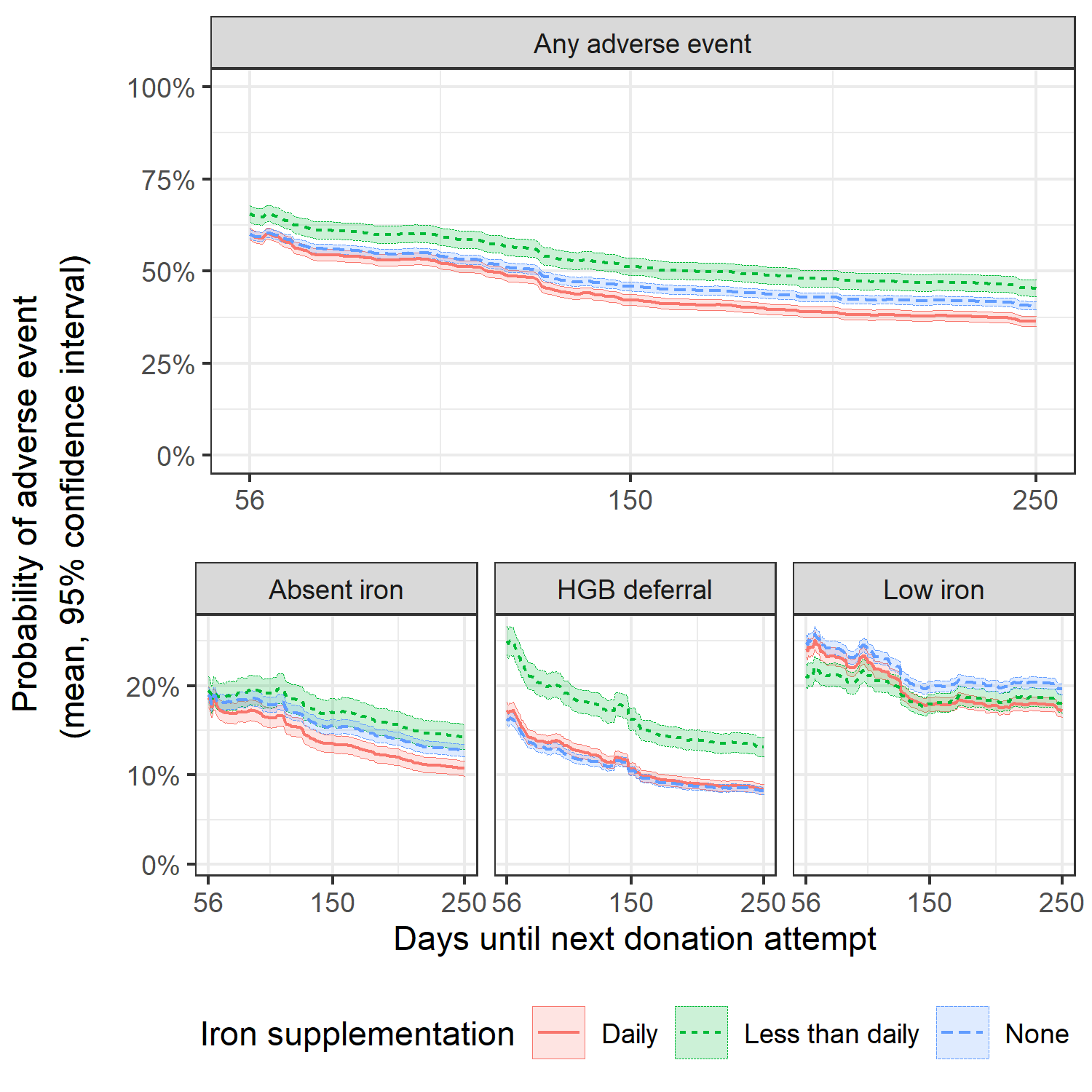
**Figure S:** Individual risk trajectory for sixty randomly-selected index donations.



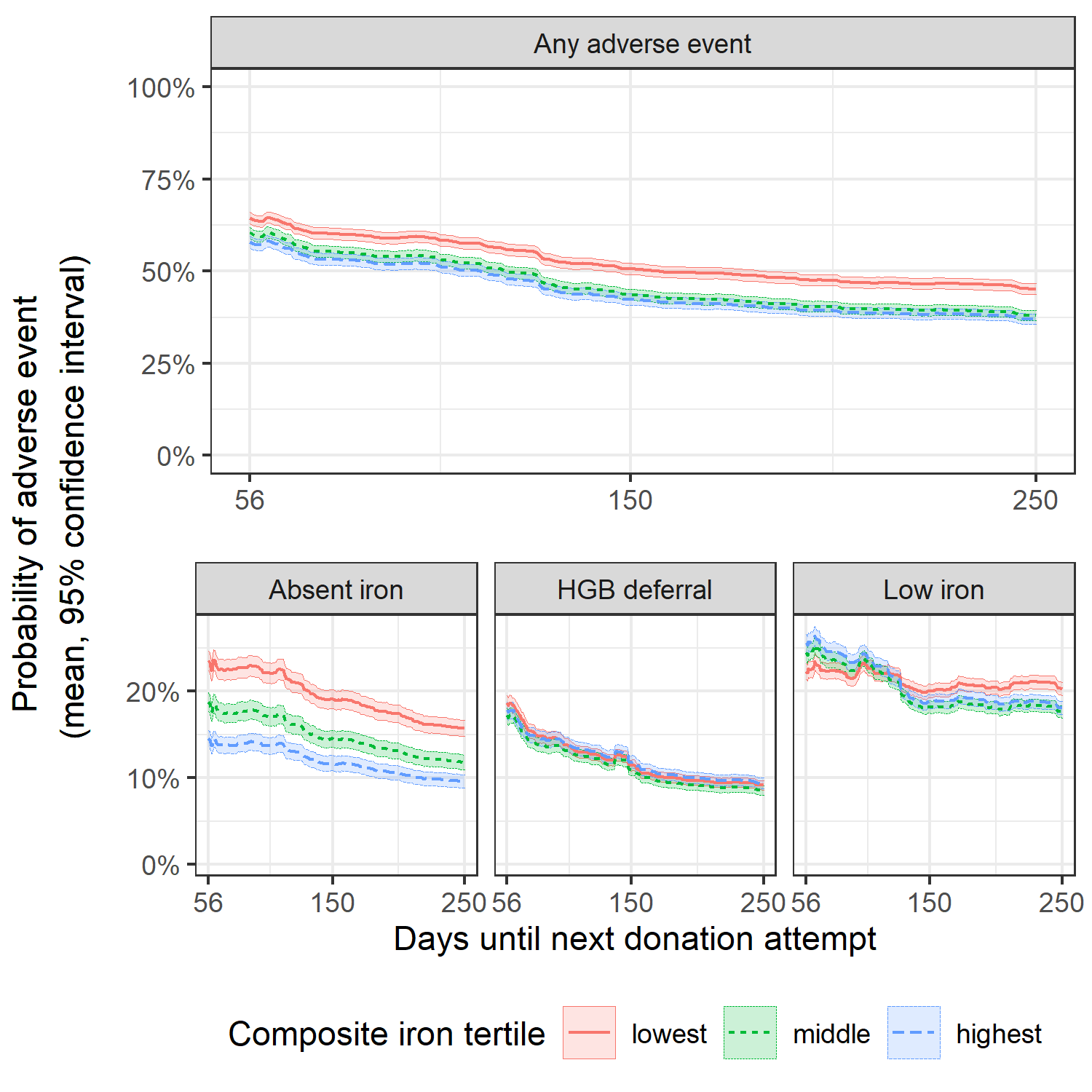
**Figure S:** .



**Figure S:** .



**Figure S:** .



**Figure S:** .