Analysis of variation between model runs in Hb-predictor Docker container

# Github

All analyses are available on github, although the data cannot be shared because they contain private information. There are more figures on github than are included in this document.

<https://github.com/SanguinStats/hb_prediction_international/>

# Background

After running the models with mode final for the first time, we noticed quite a discrepancy in model performance between modes final and initial. Further investigation showed that changing the seed in the user interface results in considerable differences between runs. To find out where this variation stems from, and if it is a problem, we investigated several possible causes. The results are in this document.

We identified three places where the seed is or could be of influence on the results:

1. **In the selection of the 10k subsample**This selection is completely dependent on the seed. Doing multiple 10k runs with different seeds will tell us how the 10k donor selection affects the model fit and performance.
2. **In the selection of 1000 donors from the test set to calculate SHAP values**

This selection is completely dependent on the seed. We take one 10k run and outside of the container calculate SHAP values multiple times with different seeds. This tells us how stable the SHAP values are between different subsamples in the test set.

1. **Elsewhere in the model fitting**

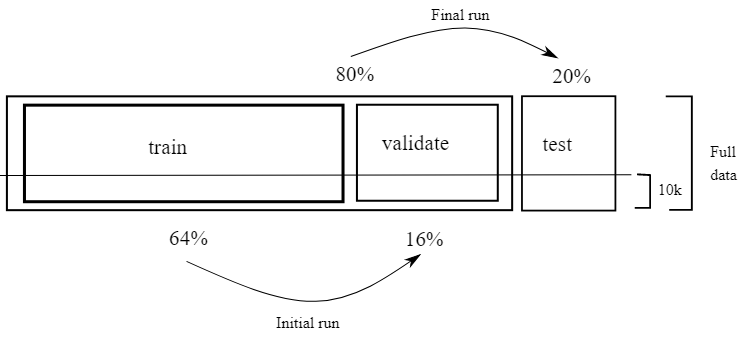
Random forest (RF) and support vector machine (SVM) models are trained using existing packages. These may use the random seed as well. We do multiple runs with the exact same 10k subsample, but with different seeds: this way there is no effect of the seed on the 10k selection. By calculating SHAP values with the same seed in these runs, we can see if the model seed affects the models and if so, how much.

# Modes and 10k selection

Division in train/validate/test is fixed and is not affected by seed or anything else. This is done once during preprocessing, and all runs are done on the same preprocessed file.

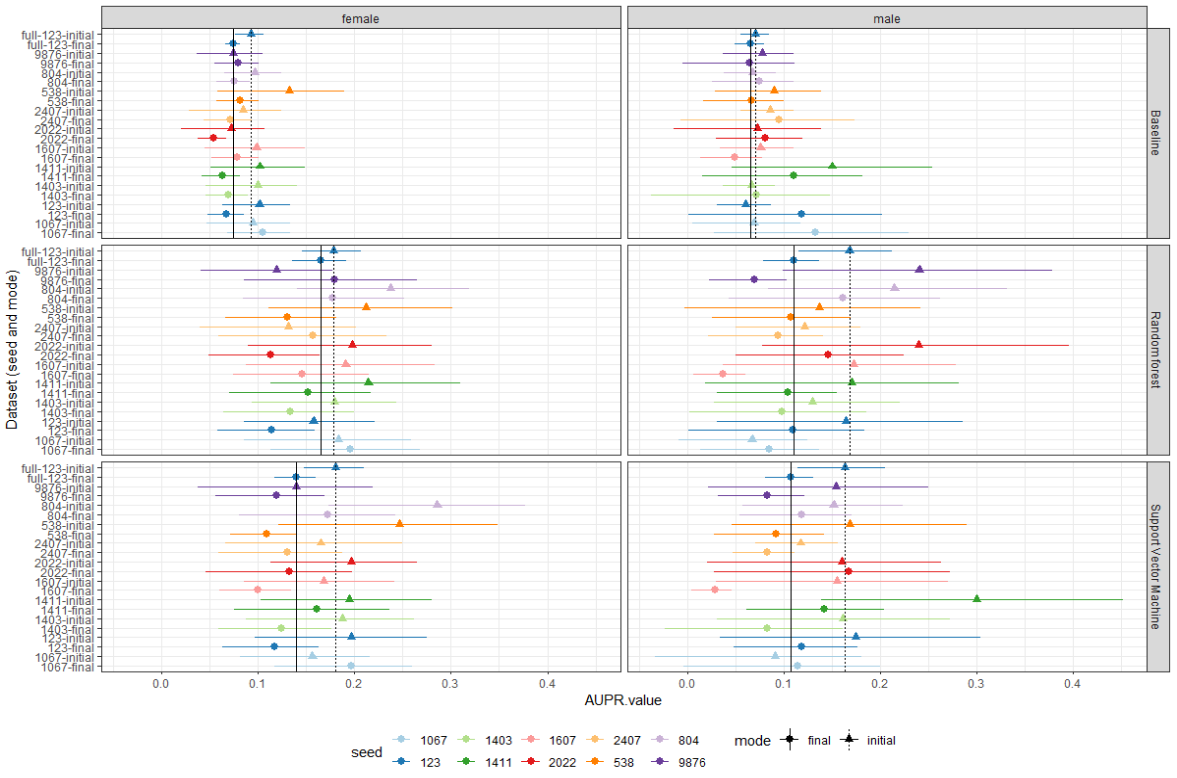
As seen in the image below, the sample sizes for mode final are larger than for mode initial. For a 10k subsample, the sample sizes (number of donors) are:

|  |  |  |
| --- | --- | --- |
|  | Mode initial | Mode final |
| Data for model training | 6400 | 8000 |
| Data for calculating performance and variable importance | 1600 | 2000 |



# Results

* 1. **Effect of seed on 10k subsample**



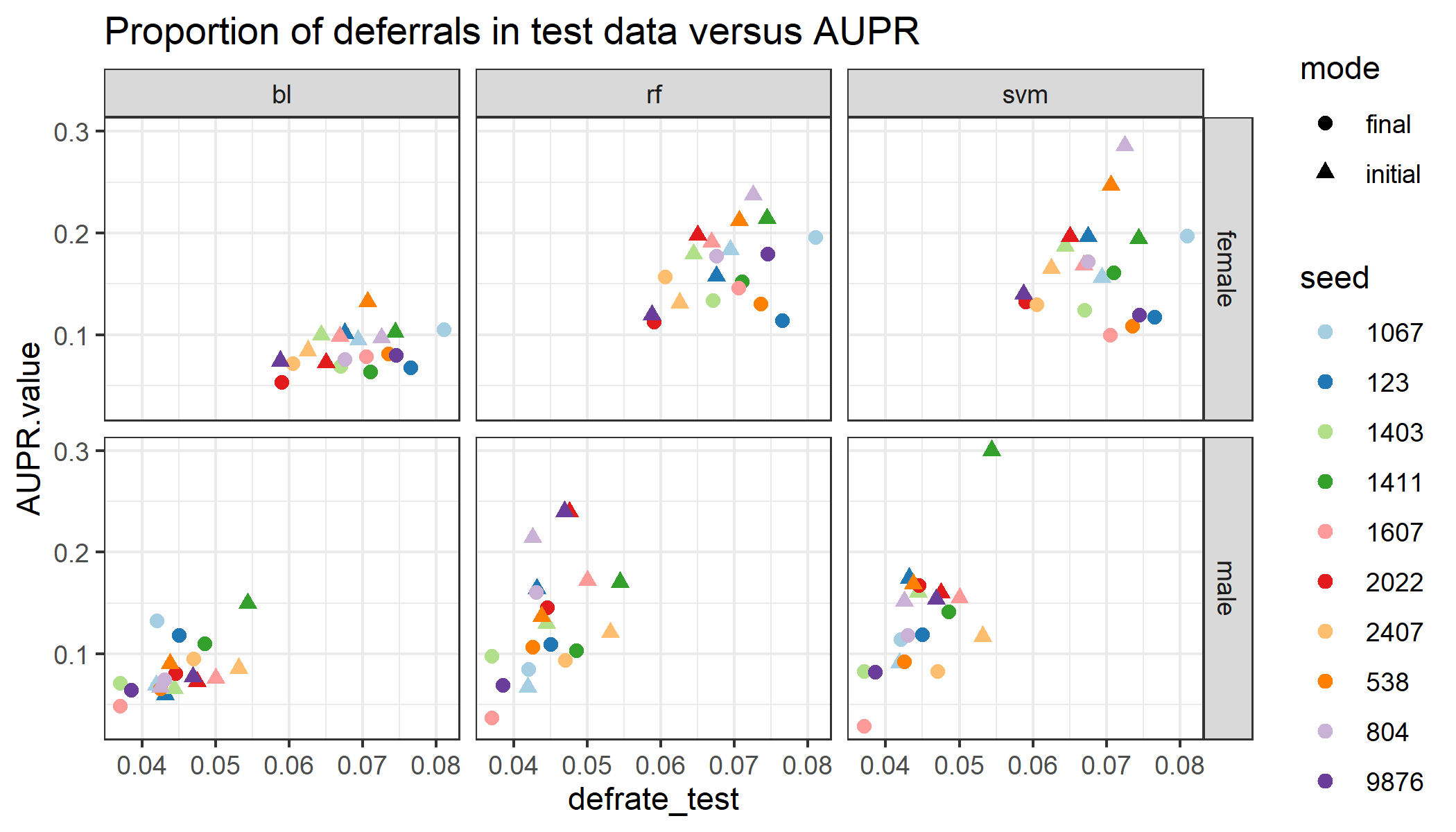
The above plot shows that performance is dependent on the selected 10k subsample. For most seeds, performance is better in the initial mode than in the final mode, although the sample size is larger in the final runs. Possible explanation for this: with larger sample sizes, there are more differences between included donors, and it is harder to find a model that correctly divides into deferral/no deferral.

The plots on the next page show the marginal distribution of the previous\_Hb predictor for all different seeds. Although there are small differences, this is unlikely to be the cause of the variation in performance, because they differ between different seeds but not between the modes, and we see the performance variation there too. Distributions of other predictor variables are practically identical and therefore not included in this document, but they can be found in the github repository.

We also found differences in the mean absolute attribution (MAA) for the different seeds. However, it isn’t possible to say if that is due to the model fit or due to the effect of seed on SHAP selection. We can’t compare this fairly, because all runs have different donors in the validate/test part of the data, and therefore we can’t calculate SHAP values on the same donors.

|  |  |
| --- | --- |
| Women – distribution previous\_hb | Men – distribution previous\_Hb |
|  |  |

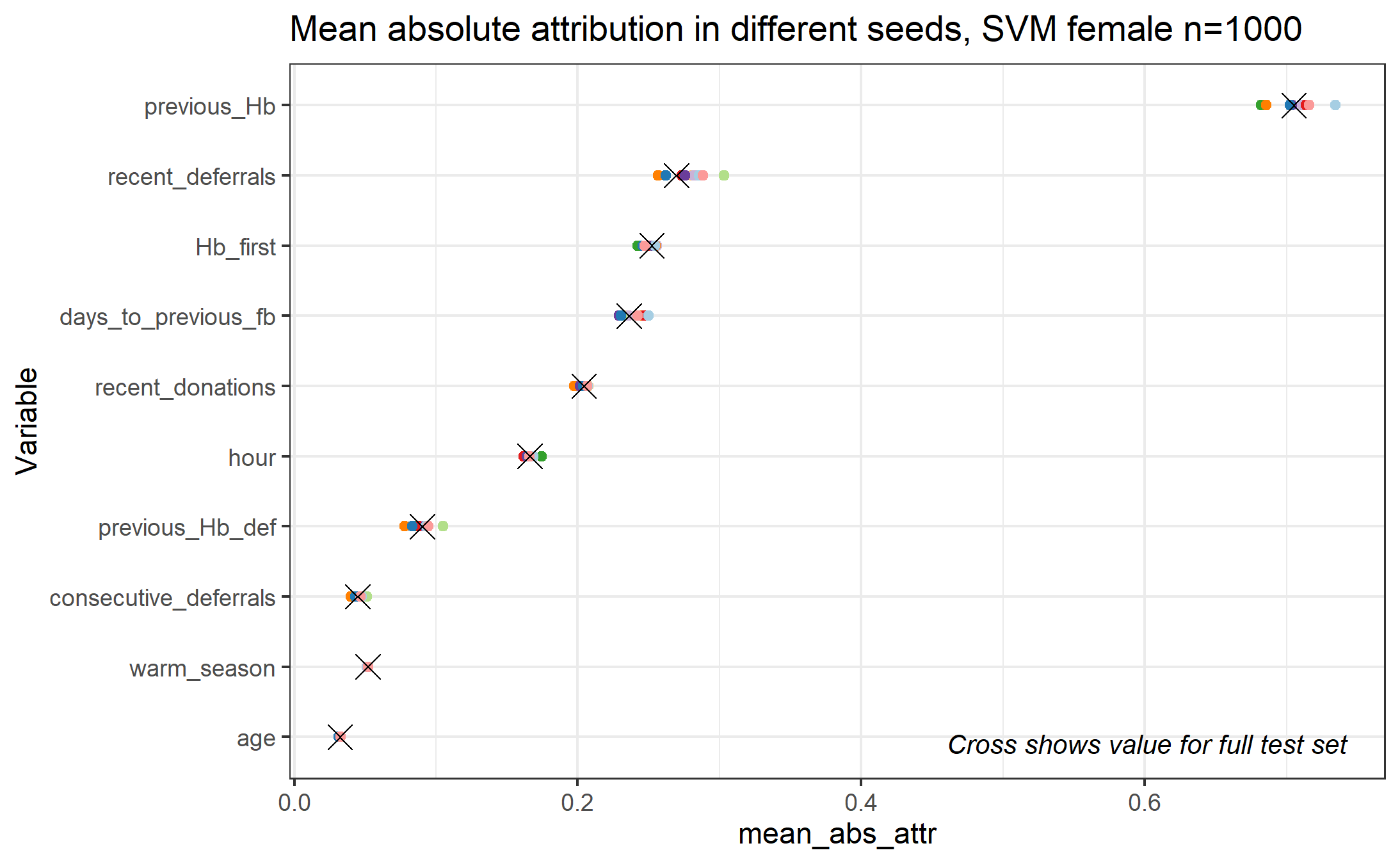
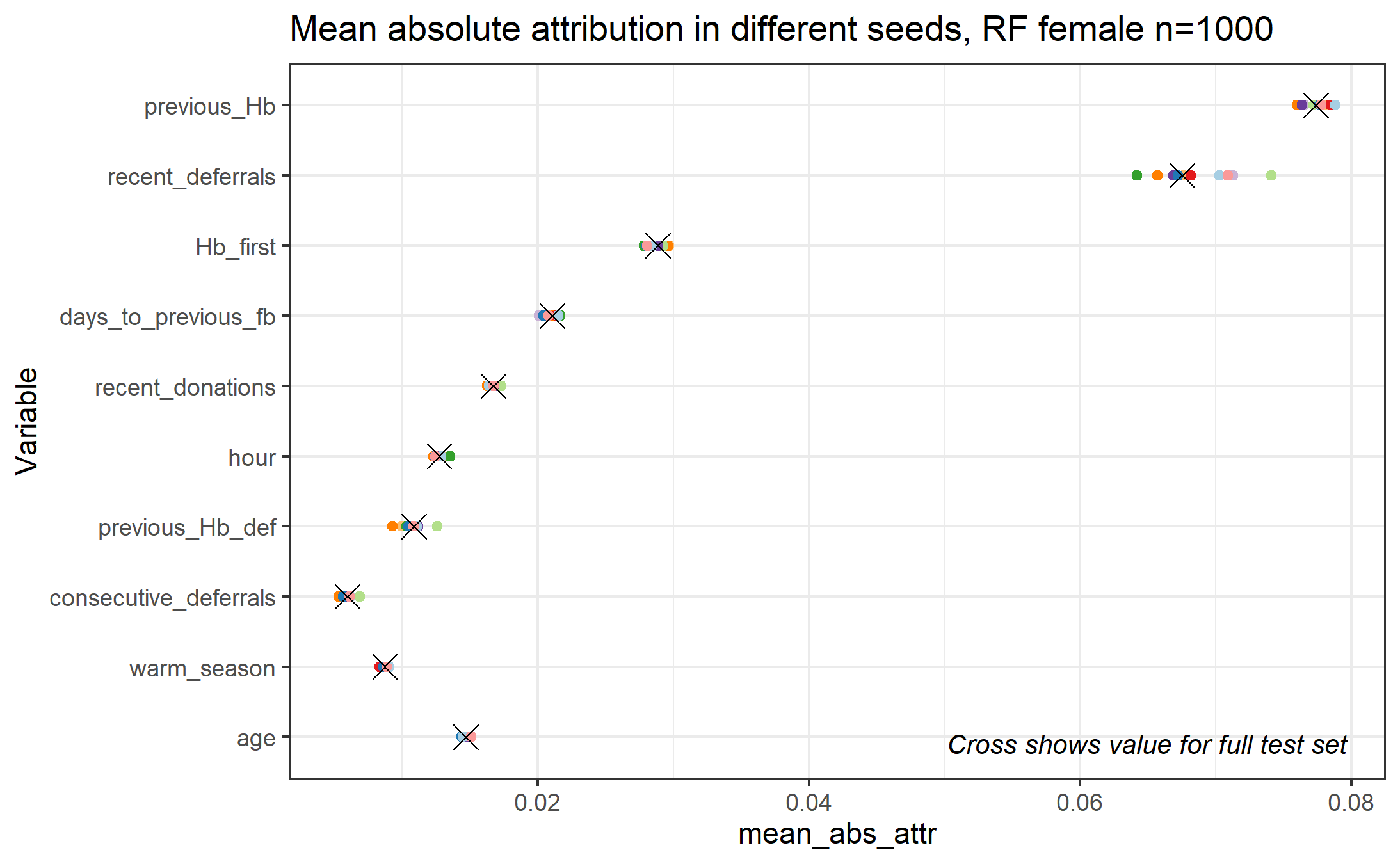
Different 10k subsamples may have differing numbers of deferrals. Classification is harder with fewer deferrals, so this may cause differences in performance. We plot the number of deferrals in each run’s test set against the AUPR:



There does appear to be a correlation here!

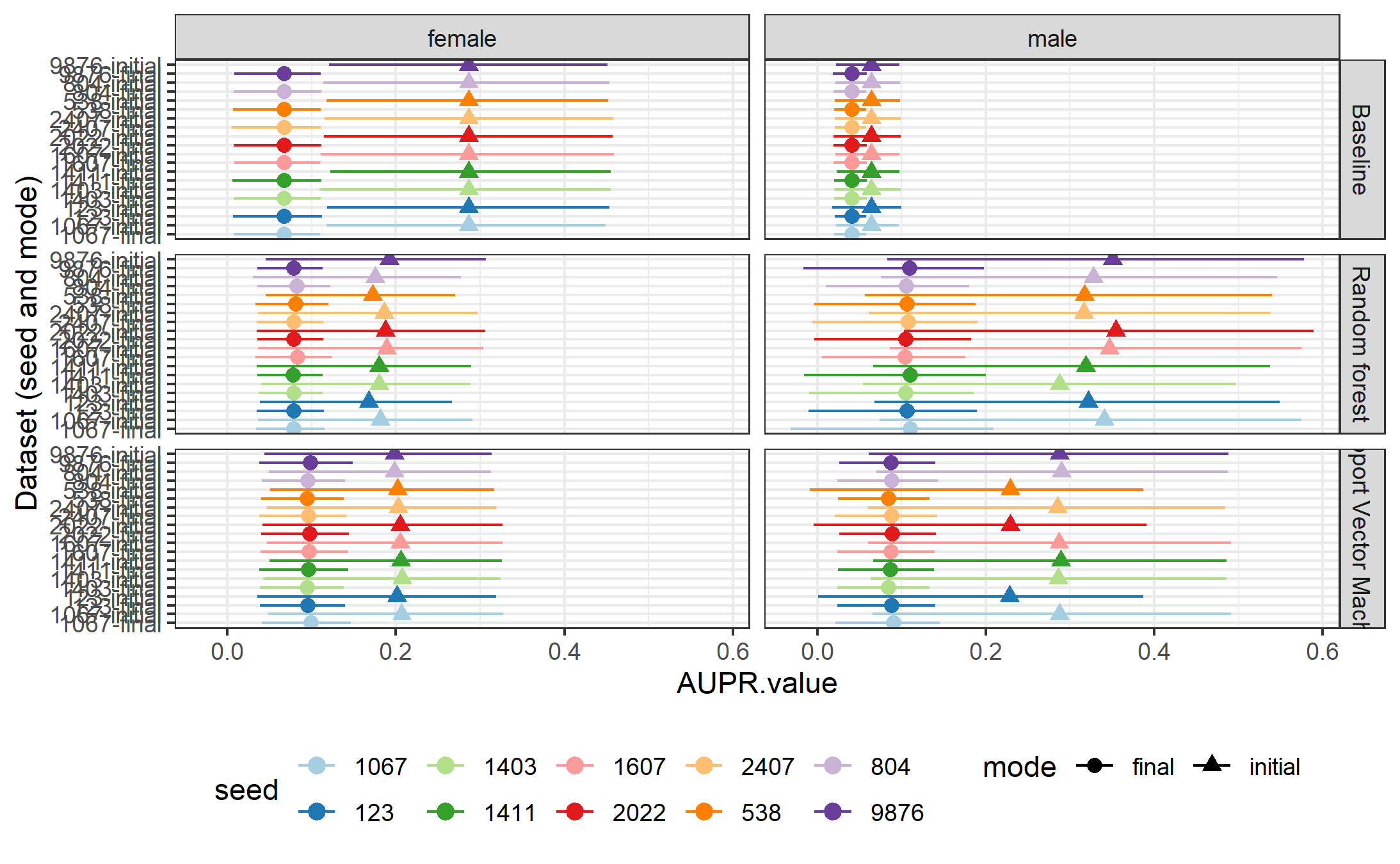
* 1. **Effect of seed on the SHAP subsample**

In the container, SHAP values are calculated for a subsample of 1000 donors from the validate/test set. This cannot explain differences in performance, as the AUPR is calculated based on the entire validate/test set, but it would lead to differences in MAA values. We took one run from the above mentioned different 10k runs: seed 1067, mode final. We then calculated SHAP values multiple times on different subsamples (different seeds). The results are below. They are shown only for female donors, they look the same for male donors and those plots can be found in the github repo. The repo also contains these plots for SHAP values calculated on 100 donors (rather than 1000): clearly the variation is much bigger there, and 1000 seems to be a good sample size where variation is present but limited.



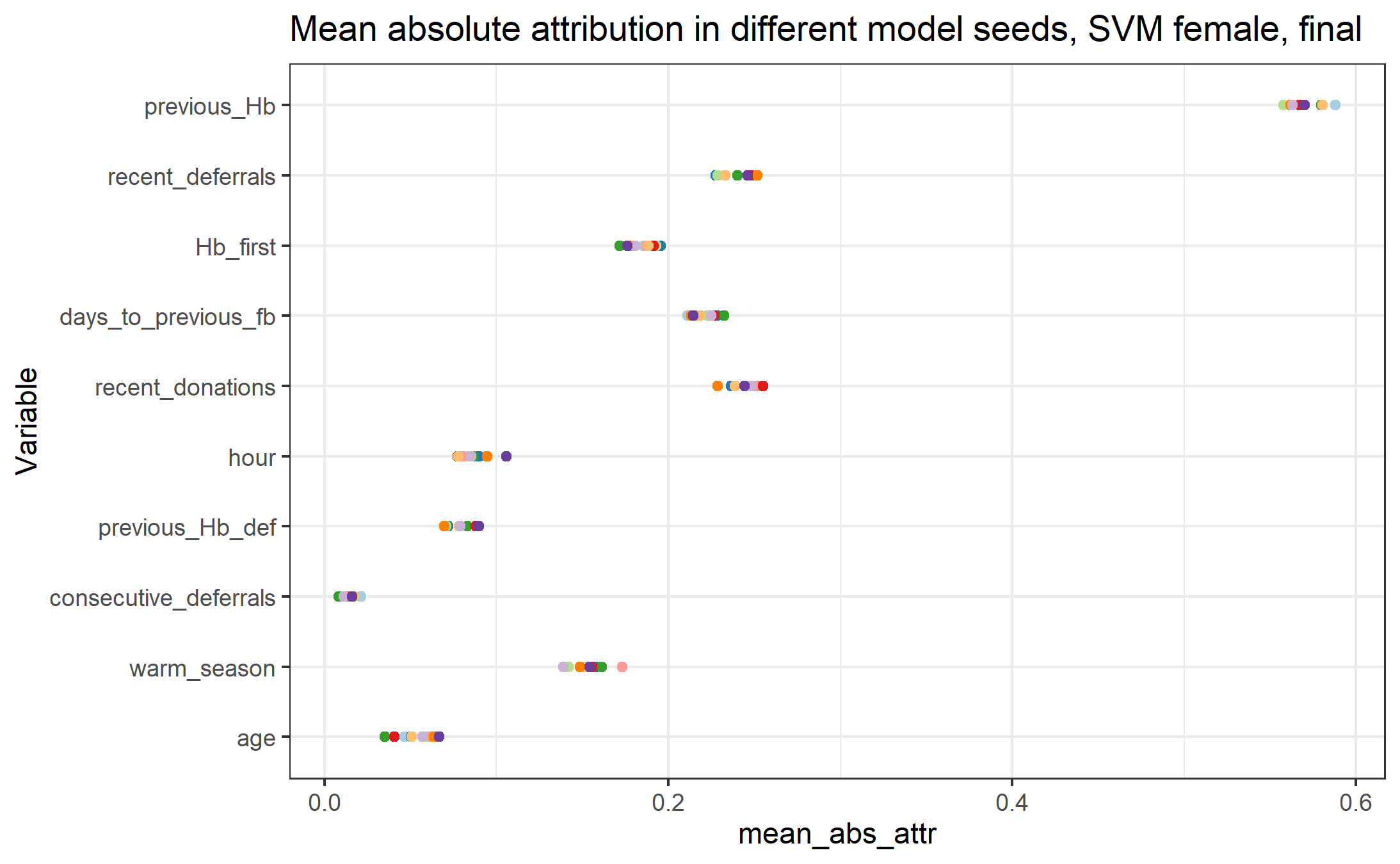
* 1. **Effect of seed on model fitting**

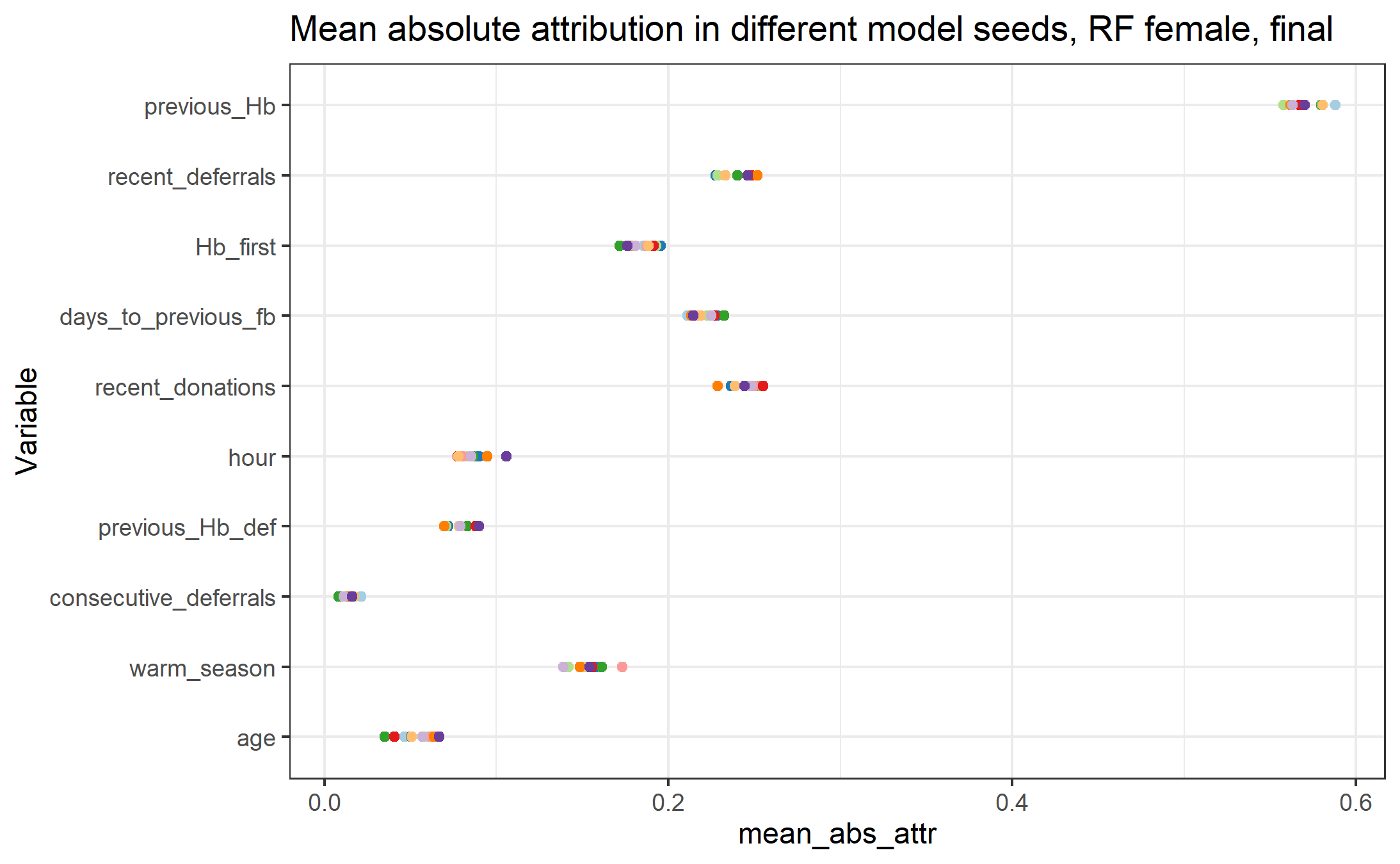
To find out if – and if so, how much – the seed has any effect on the model fitting outside of selecting the 10k subsample and the 1000 subsample for the SHAP values, we manually created a 10k subsample dataset outside of the container. This was then used for runs with different seeds. We first compared model performance, which is not dependent on the seed, and thus if there are differences, this must be due to differences in the model fitting. This is the result:



In the baseline models, there are no differences in AUPR values between the different seeds, only between the different modes. This means that for the baseline model, seed does not affect model fitting, as expected. However, for both RF and RVM, AUPR values differ between seeds. Interestingly, the differences are much smaller in mode final than in mode initial. The performance is especially unstable for male donors. In the SVM, it is clearly visible that seeds 538, 2022 and 123 form one group that differs from the rest, as its AUPRs are lower than the others. In the RF, this pattern is not visible, and variation seems more random.

We computed the SHAP values for all these models, using the same seed each time so the same donors are selected. To save computation time, these SHAP values are only calculated on 50 donors rather than 1000. The point here is to show that using the exact same data set, the seed affects the fitted model that is learned, and there are differences in SHAP values. Only the plots for women, mode final are shown, but the same pattern is visible in the other groups, and those plots can be found on github.





# Preliminary conclusions

The seed that is specified by the user in the container user interface affects model performance in several ways.

First, it decides which subsample of donors to use, if the sample ratio specified in the user interface is less than 1. With a sample size of 10k donors, different selections of donors can lead to quite different AUPR values. The most likely explanation for this is the different number of deferrals in the selection, particularly in the part of the data that will be used for testing rather than model fitting. With more deferrals in the data set, model performance is higher.

Second, the specified seed affects the actual model that is learned. For the random forest, this is expected, because as the name implies there is a random part to the algorithm when decision trees are fitted to random samples (with replacement) of the data. Some variation is expected here, and we see from the plots in section 3.3 that this variation is quite small. It is likely even smaller than shown here, because SHAP values were computed only on 50 donors for these plots, while they are calculated on 1000 donors in the container.

Third and last, the seed that is specified is also used to select the 1000 donor subsample to calculate SHAP values on. Again, we expect some variation here, but the plots in section 3.2 indicate that n=1000 is a large enough sample that variation is limited.

**Overall, we learn the following lessons:**

* **The proportion of deferrals in the data is correlated with the model performance.** We already knew this was important when comparing between countries, so this is not surprising.
* **There is a random part to the model fitting**. However, this does not cause a worrying amount of variation.
* **SHAP values are seed-dependent**. Calculating SHAP values on a subset of 1000 donors does not cause a worrying amount of variation.

And we should mention the following in the manuscript:

* If computation power allows, use the full data rather than a 10k subsample to avoid variations in deferral proportions. Runs with increasing sample size that Tinus is doing now will tell us how variation decreases as the sample size increases. Running on full data is obviously always optimal.
* To get the best estimate for variable importance, SHAP values should be calculated on the full test set rather than a subsample.
* There are multiple sources of variation that are controlled by the seed. All these small variations add up to larger total variation in performance. Part of this variation is unavoidable. The user should be aware that this variation exists.