IPS | Translational Bioinformatics / GBDS | Data Science

Predictive Biomarker Modeling of Early Progression in Nivolumab-Treated Melanoma Patients

March 17, 2021

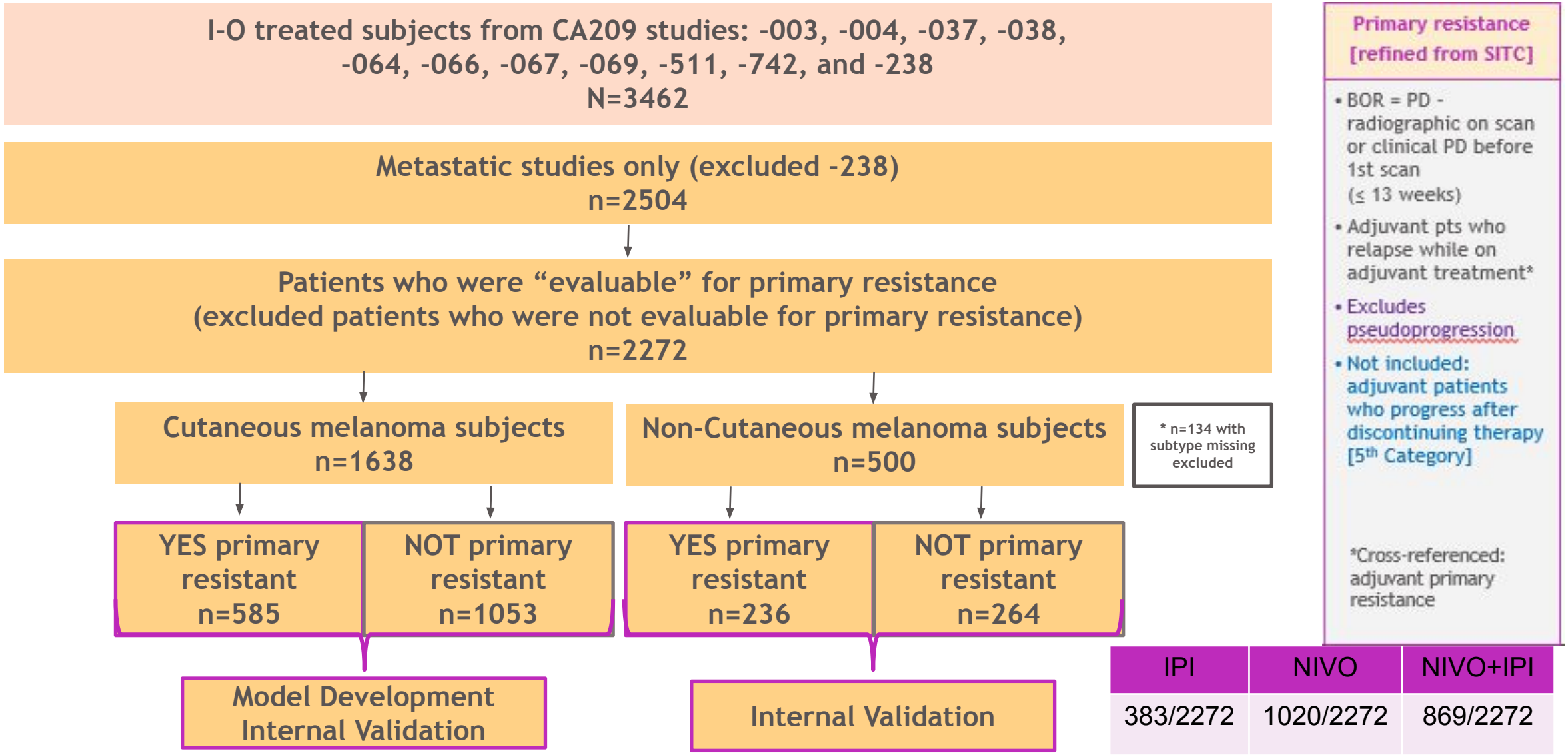
Abraham Apfel (IPS/BIOM) and David Paulucci (GBDS/Data Science)

Bristol Myers Squibb™

Goals of the analysis

- **Primary** goal: Develop a prognostic score to predict primary resistance (binary outcome) from Melanoma subjects receiving IO treatment
- *Secondary* goal: Which biomarkers are influential in determining the predicted probability of a patient being primary resistant?
- *Secondary* goal: Identify any potential predictive effects, i.e. are there predictors which can help decide on appropriate IO treatment for a given patient?
  - Assessed predictor * treatment interaction to determine if the relationship between a given biomarker and the predicted probability for being primary resistant differ across different treatment arms

Consort Diagram



Primary resistance [refined from SITC]

- BOR = PD - radiographic on scan or clinical PD before 1st scan (≤ 13 weeks)
- Adjuvant pts who relapse while on adjuvant treatment*
- Excludes pseudoprogression
- Not included: adjuvant patients who progress after discontinuing therapy [5th Category]

*Cross-referenced: adjuvant primary resistance

IPI	NIVO	NIVO+IPI
383/2272	1020/2272	869/2272

Initial candidate pre-treatment covariates used to model primary resistance

Treatment arm as randomized

Patient Characteristics:

- Age at consent
- Sex
- Race
- ECOG PS
- Continent
- BMI

Disease Characteristics:

- M stage
- AJCC stage
- Melanoma subtype
- Baseline tumor burden
- Liver metastases

Biomarkers/Labs:

- PD-L1 expression (%)
- *BRAF* mutation status
- Hemoglobin level
- LDH
- Absolute lymphocyte count
- Absolute neutrophil count
- Platelet count

Medical History:

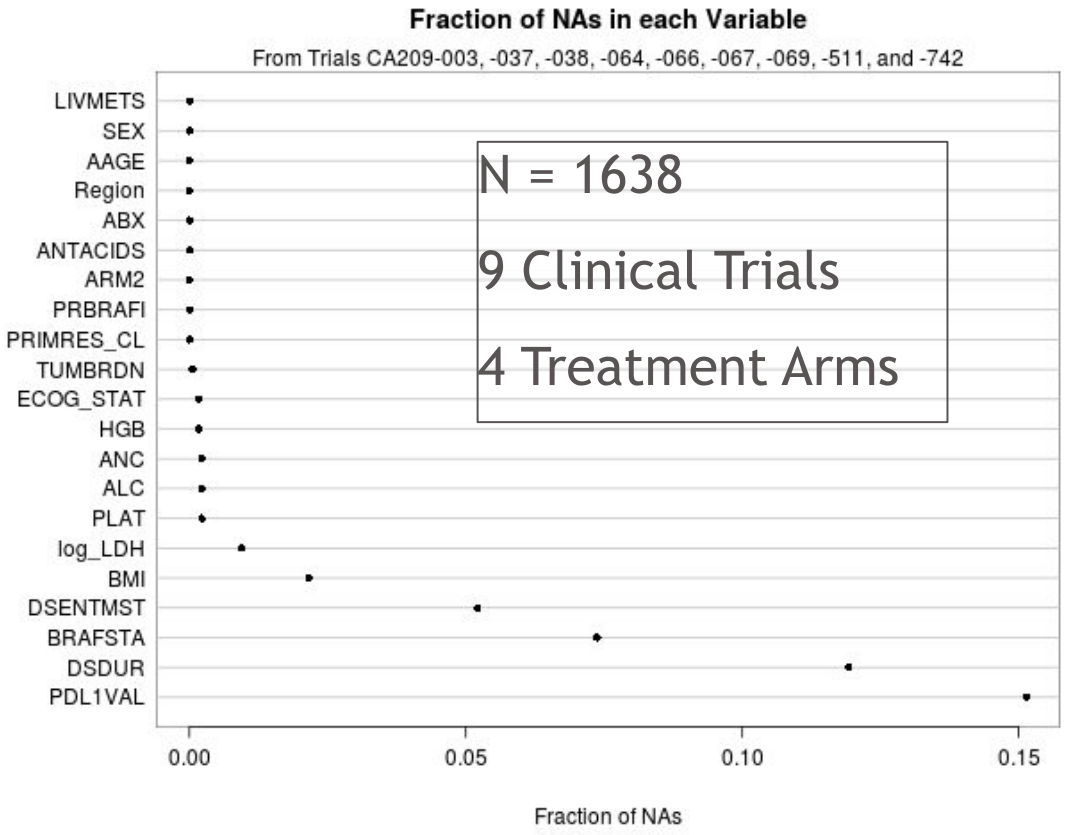
- Time from current diagnosis to treatment start date
- Antibiotic use within 30 days of treatment
- Antacid (PPI or H2 blocker) use within 30 days of treatment
- Prior BRAF inhibitor treatment
- Prior ipilimumab treatment

TM/IPS Analysis Only

- TMB
- CD8 IHC
- Gene expression (RNAseq)

Statistical Methodology Challenges

1. Missing data
2. Imputations with “block missing” (certain covariates were missing from complete trials)
3. Combining imputation with validation
4. “Trial” effect
5. Comparing multiple modeling approaches



Analysis Outline

- Two Imputation Approaches Evaluated
  - Multiply imputed 32 datasets followed by resampling.
  - Single imputation within resampling.
- Performed many modeling techniques
  - Frequentist and Bayesian logistic regression, penalized logistic regression, machine learning (Gradient Boosting Machines, Random Forests)
- Perform internal validation on each approach
  - Optimism-corrected bootstrap
  - Nested, Repeated CV
  - Internal-External CV
- Compare predictions and performance metrics of each approach
  - Prediction Matrix
  - R2, AUC
- Produce appropriate visualizations

eXtreme Gradient Boosting (XGBoost)

1. An ensemble based statistical learning method in which small classification trees are sequentially fit to the data.
2. Each iteration of tree growth uses information of the previously fit trees, by fit to the residuals of ensemble of previously fit trees.
3. Multiple hyper-parameters tuned via a comprehensive grid search, within the resampling.
   - Number of trees
   - Tree depth
   - Number of variables to grow tree at each iteration
   - Number of patients to use to grow tree at each iteration
   - Eta: Shrinkage parameter which shrinks the feature weights to make the boosting process more conservative.
   - Gamma: Minimum loss reduction required to make a further partition on a leaf node of the tree.

Logistic Regression

1. Redundancy analysis
   - See if any candidate predictor can be well predicted from other candidate predictors
2. Estimate degrees of freedom you can afford to spend in model
   - Approximate rule of thumb[1]: One degree of freedom for every 10-15 “effective” samples in your model
     - Effective samples = min(∑Y = 0, ∑Y = 1)
     - Multiple Imputation yields smaller effective sample size than if data were truly complete (but still larger than complete case analysis!)
     - Shrinkage helps
   - More rigorous estimation: van Houwelington and le Cessie[1, 2] shrinkage estimate

[1] J.C. van Houwelington and S. le Cessie. Predictive value of statistical models. *Stat Med, 9:1303-1325, 1990.*

[2] *Regression Modeling Strategies*, Frank Harrell.

Logistic Regression (cont.)

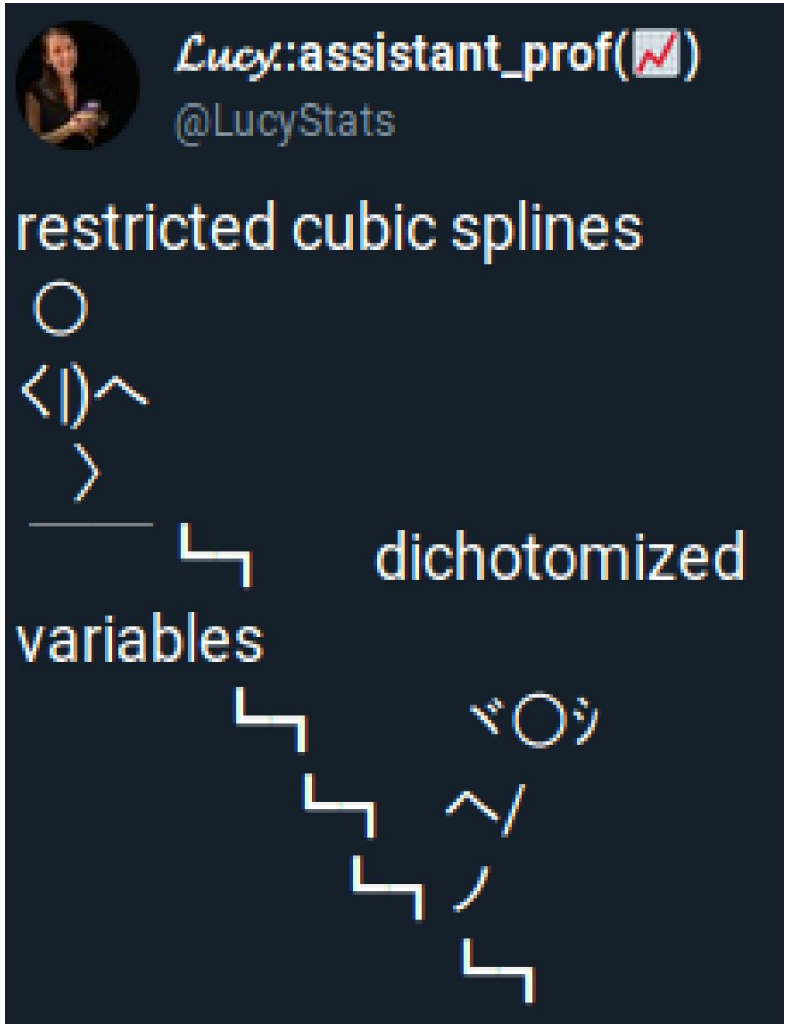
3. Figure out how best to spend degrees of freedom
   - Run saturated model (allowing non-linear relationship for all continuous markers) to determine which terms to allow non-linearity in reduced model
     - Apply cubic splines to terms with stronger associations
     - DON'T test for non-linearity
     - DON'T plot relationships

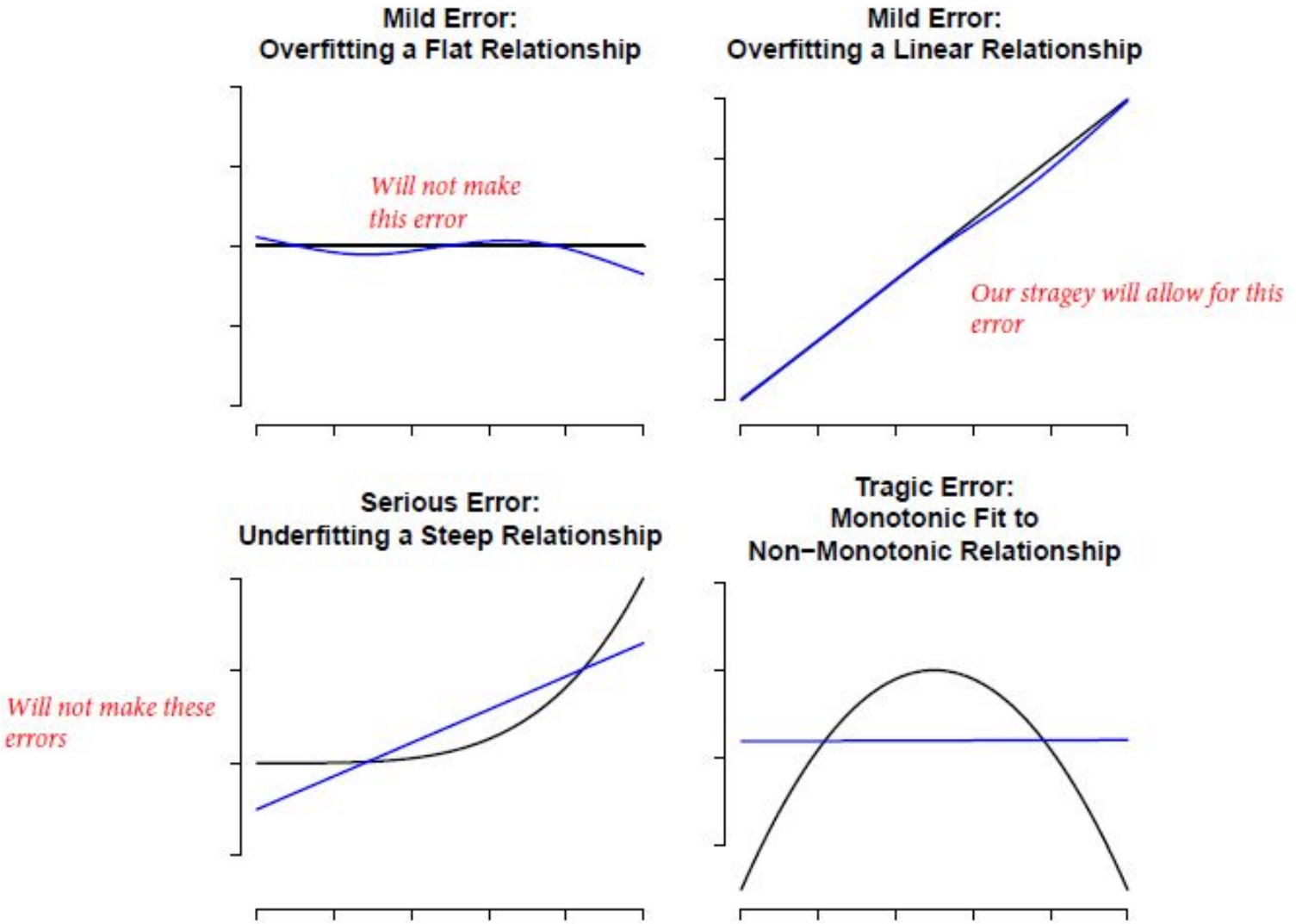
4. Explore interactions
   - *Allow for interactions between candidate variables and treatment*
   - *Perform "chunk test" (LRT) on all candidate variables simultaneously to see if interactions are worthwhile*

5. Run "reduced" model

6. Visualize the results

Restricted Cubic Splines





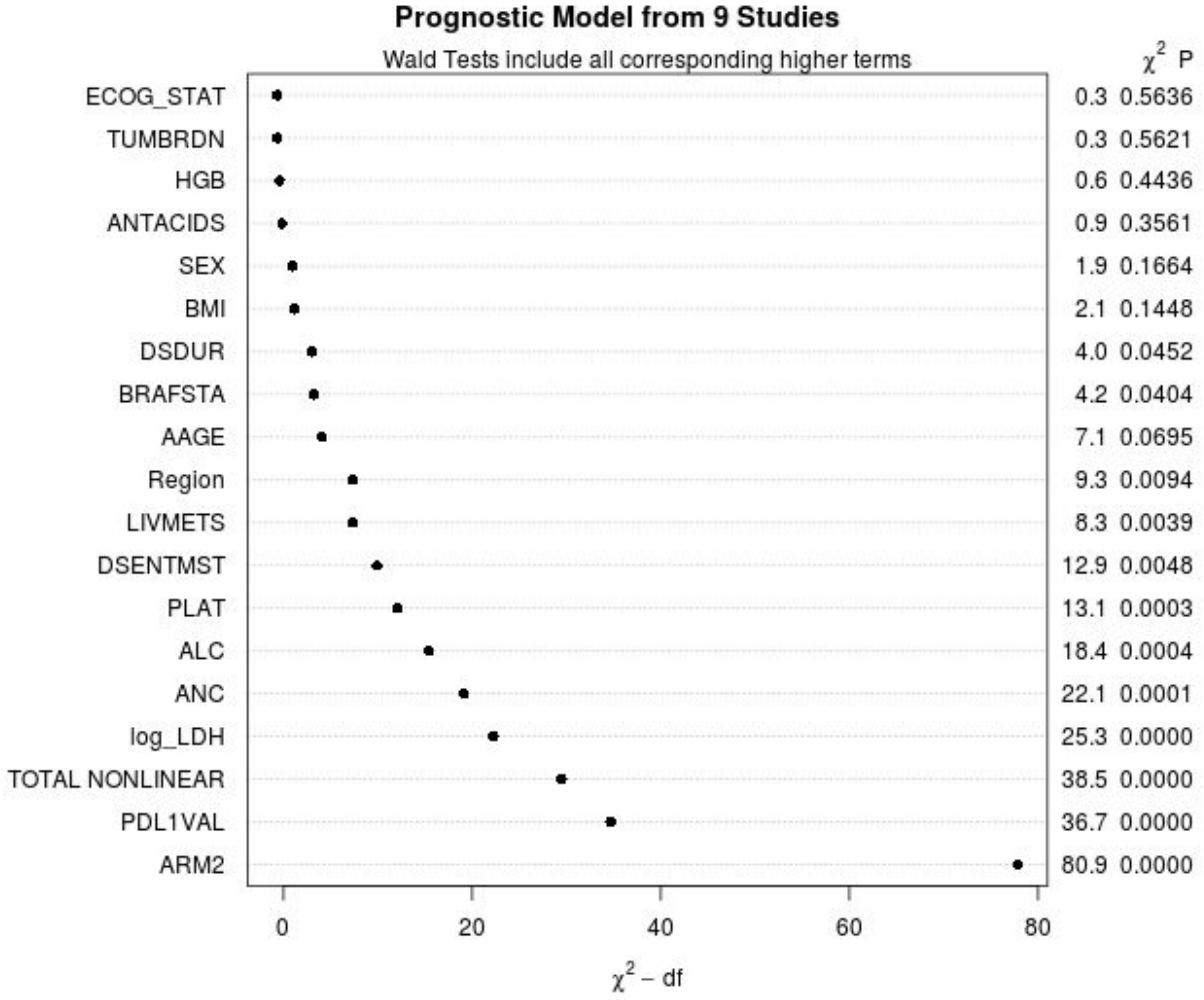
Courtesy of Frank Harrell

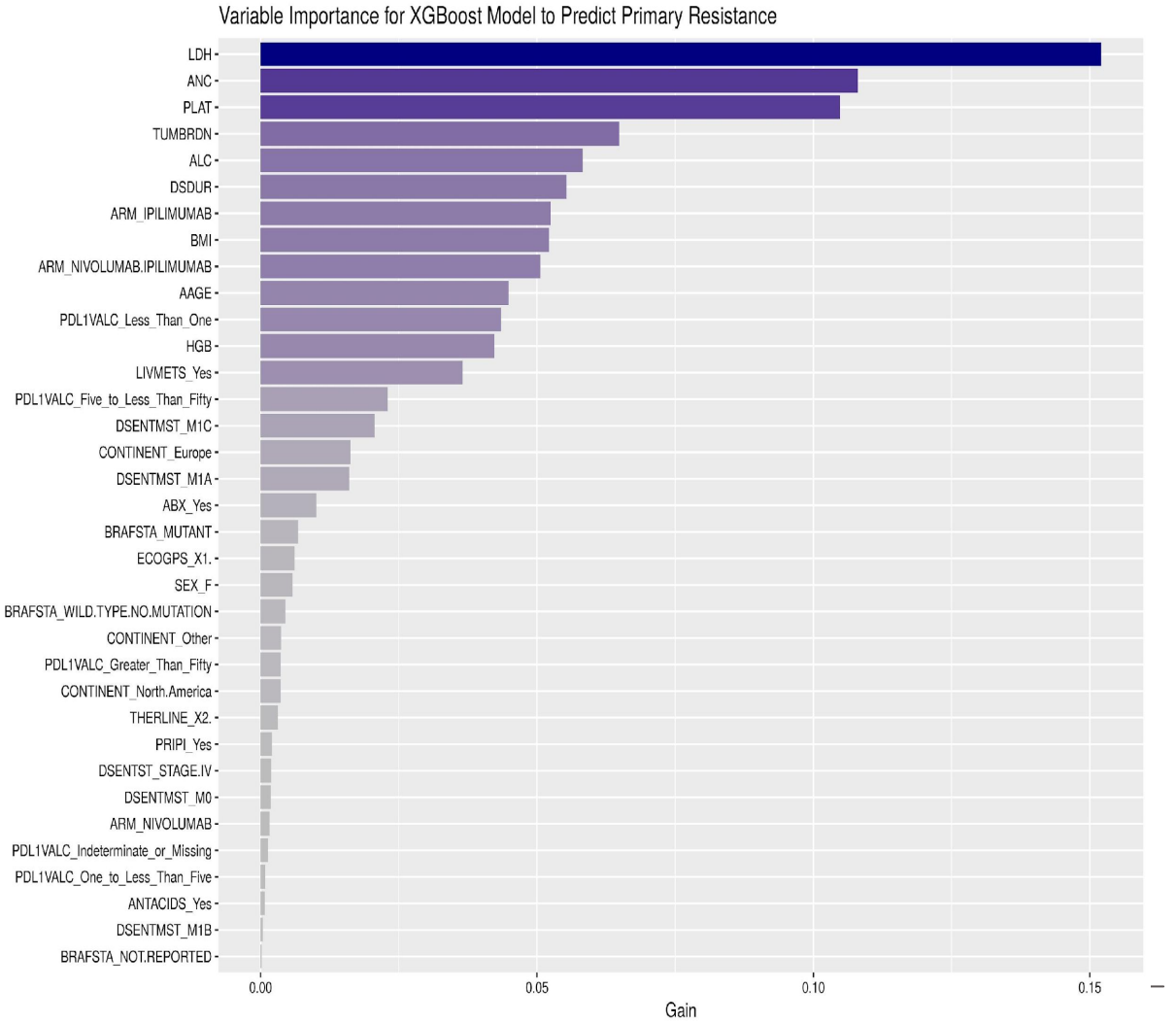
https://hbiostat.org/doc/rms1.pdf

Prognostic Model Summary for Logistic and Bayesian Models

Model: LDH + PDL1 + ALC + Arm + BRAF + ANC + PLAT + Region + DSDUR + BMI + Antacids + HGB + TUMBRDN + LIVMETS + ECOG + Sex + Age + DSENTMST + (Random effect for STUDYID – Bayesian model only)

Due to sample size constraints non-linear effects were only included for PDL1, LDH, ANC, ALC, and Age

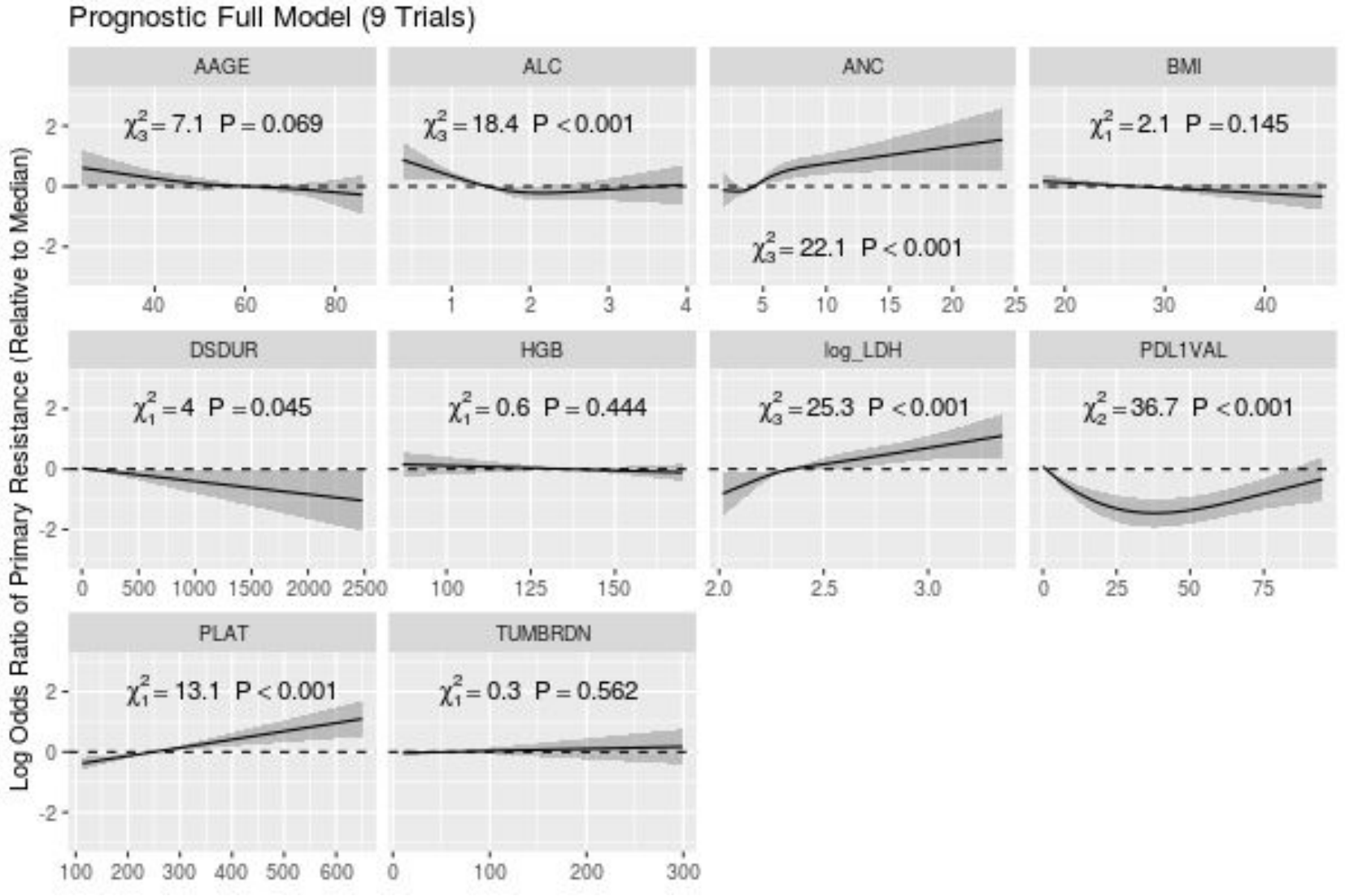




Partial Effects Plots -Relative to Median of Respective Predictor

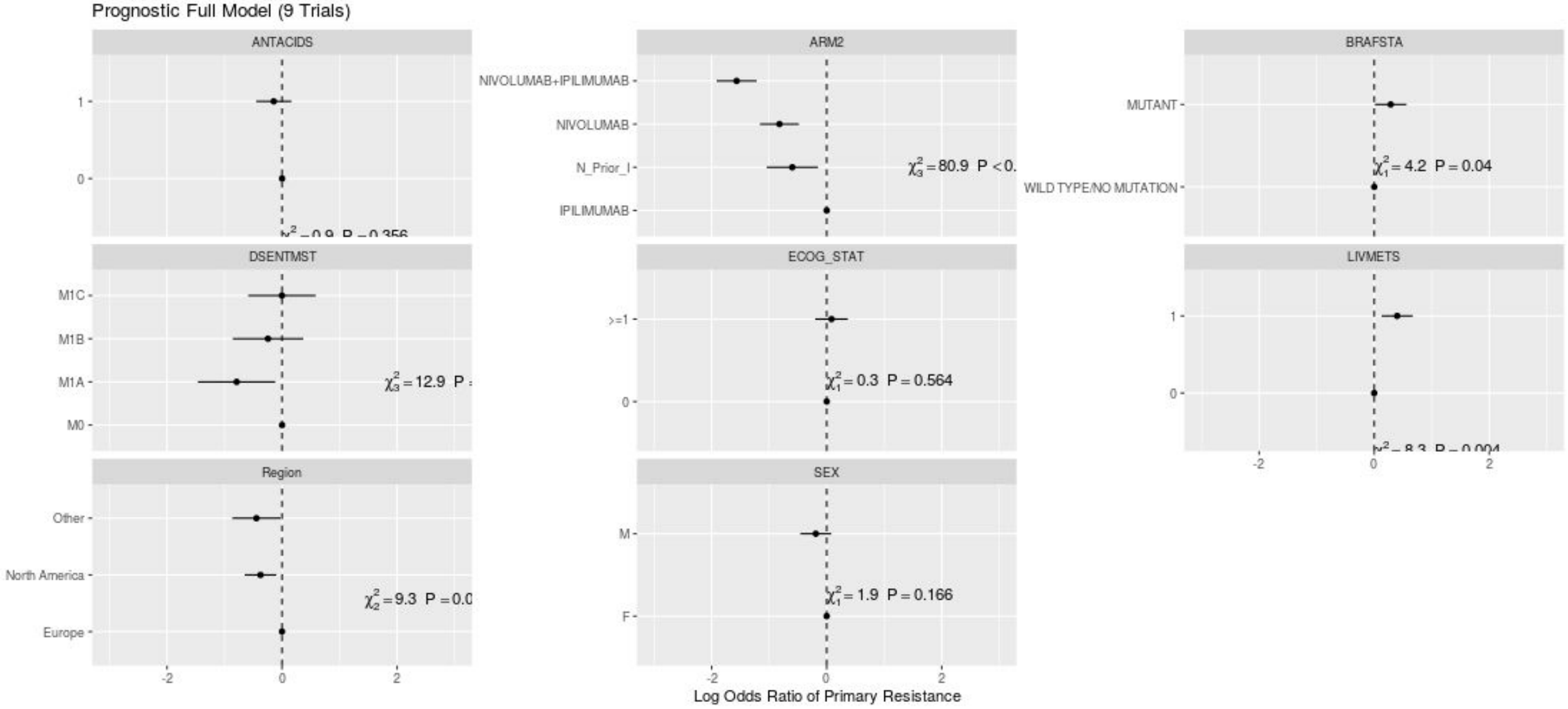
Model: LDH + PDL1 + ALC + Arm + BRAF + ANC + PLAT + Region + DSDUR + BMI + Antacids + HGB + TUMBRDN + LIVMETS + ECOG + Sex + Age + DSENTMST + (Random effect for STUDYID – Bayesian model only)

- Disclaimer: No Causal Inference applied prior to model fitting
- Forest plots can be misleading when non-linear terms are present

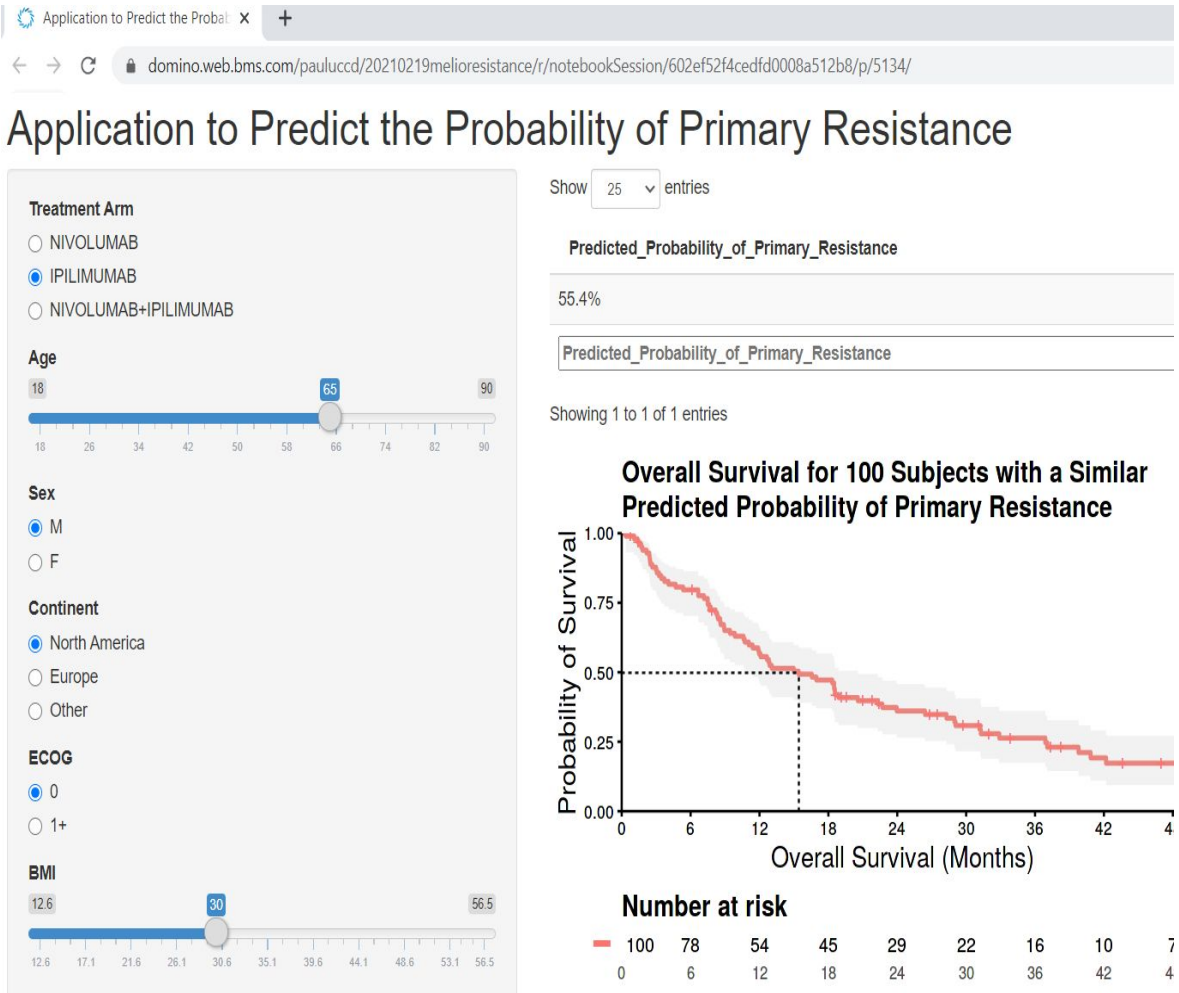


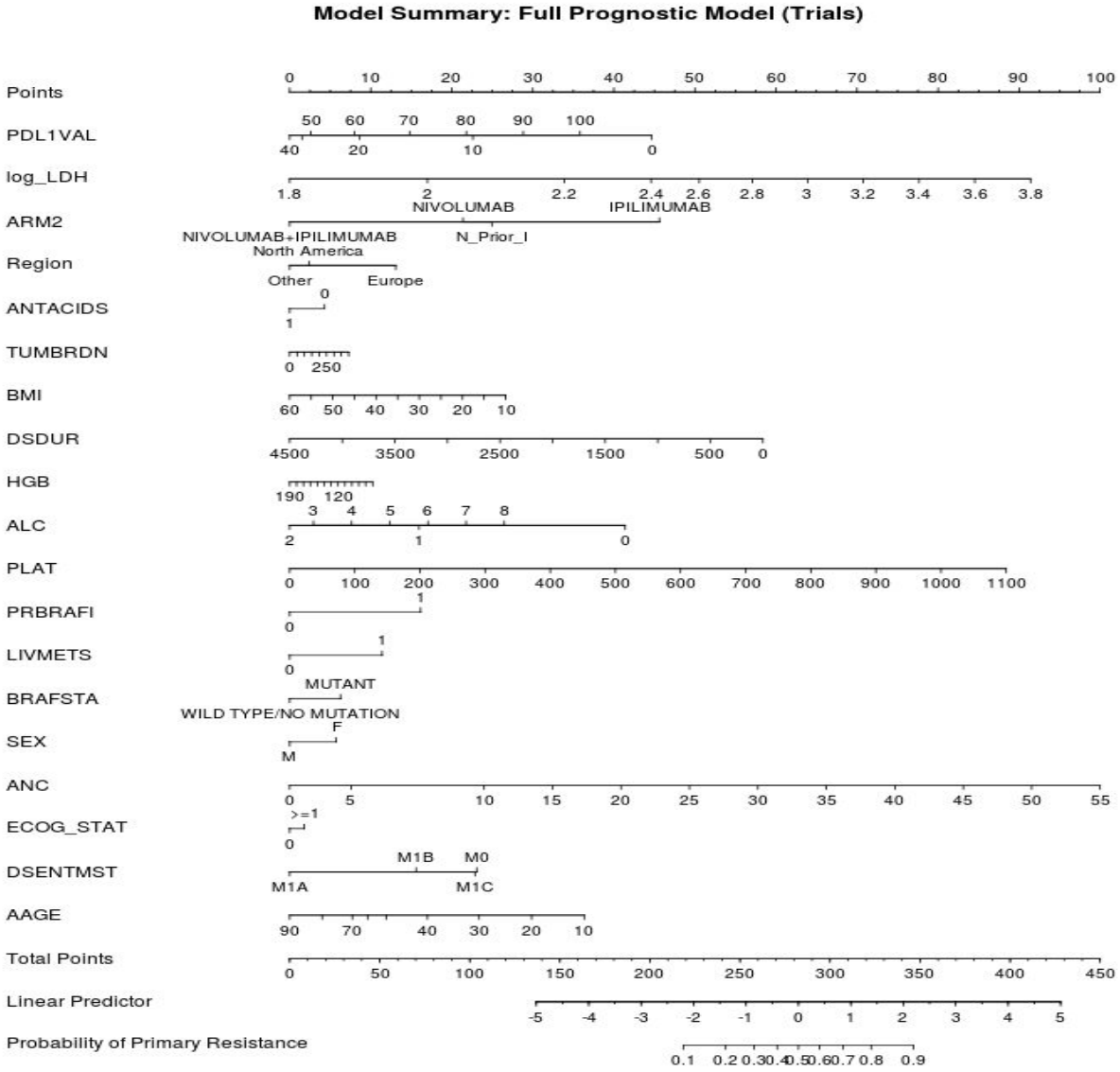
Partial Effects Plots

Model: LDH + PDL1 + ALC + Arm + BRAF + ANC + PLAT + Region + DSDUR + BMI + Antacids + HGB + TUMBRDN + LIVMETS + ECOG + Sex + Age + DSENTMST + (Random effect for STUDYID – Bayesian model only)



Model Summary



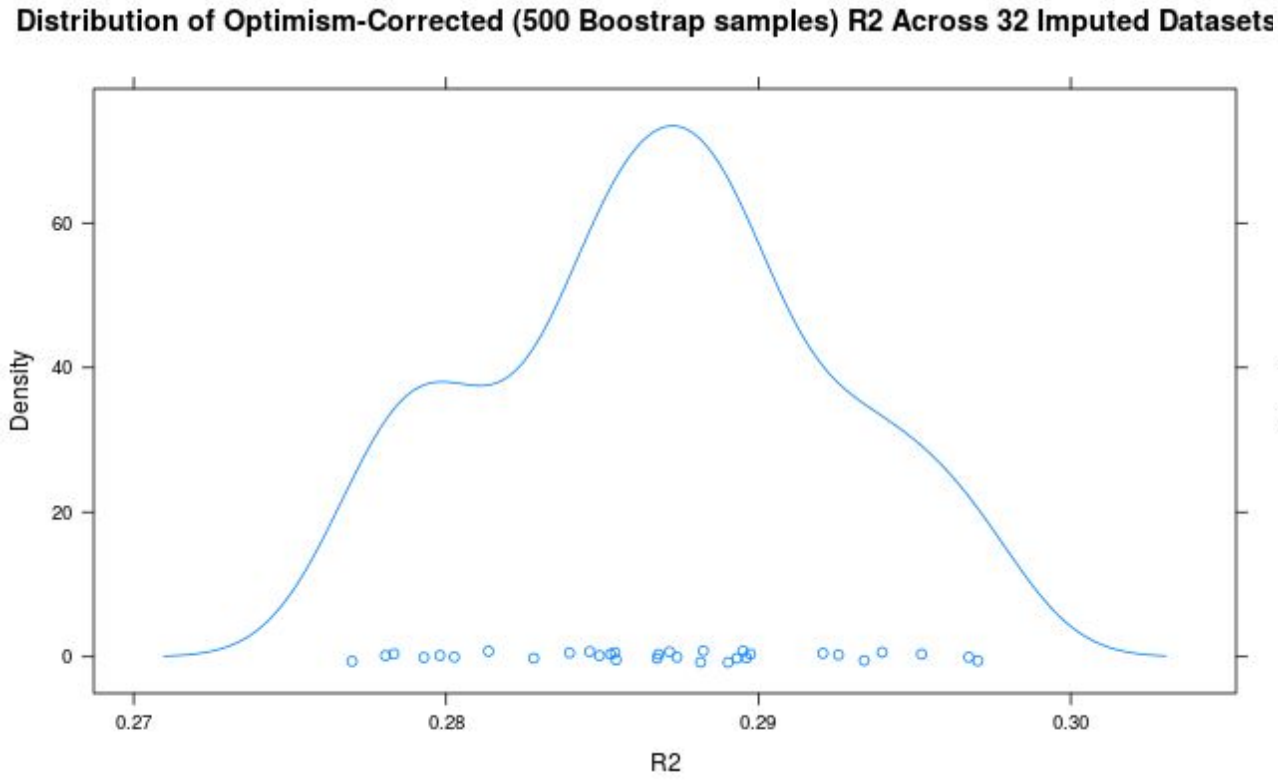


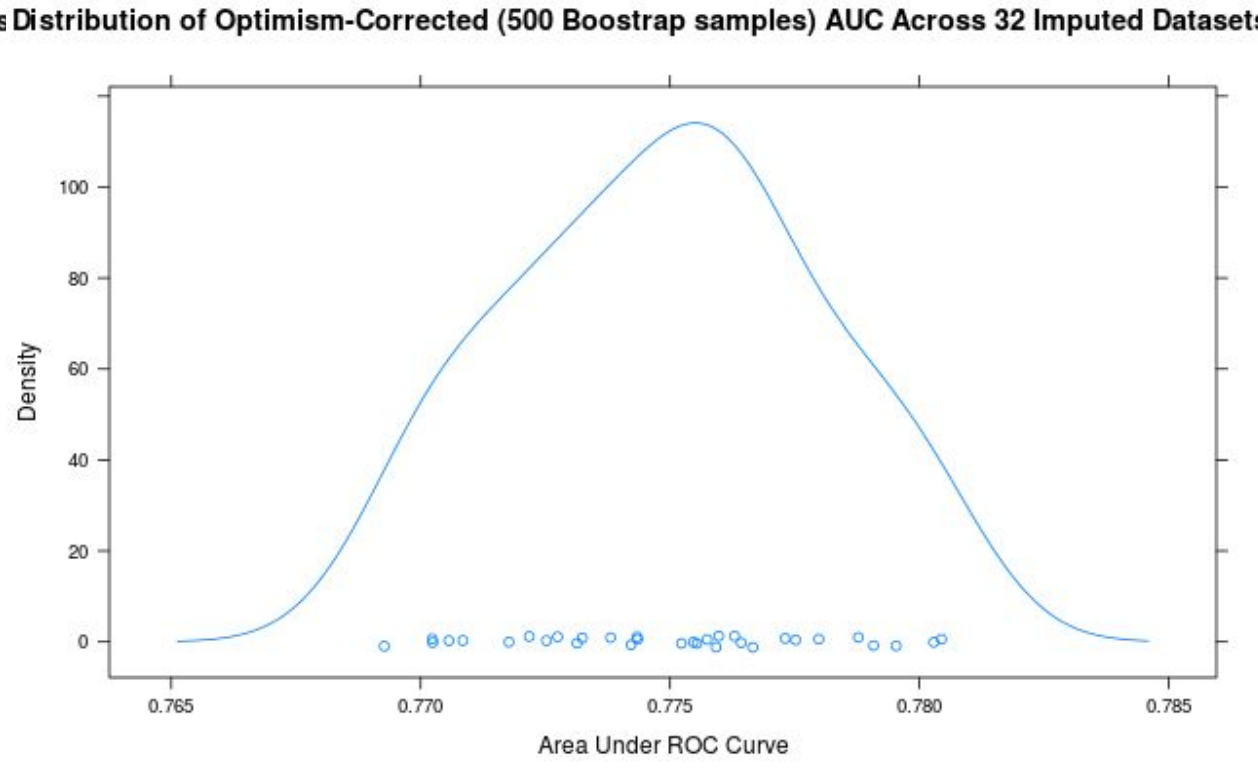
Internal Validation (Discrimination) via Multiple Imputation

- Relatively small amount of optimism in discrimination statistics implies minimal overfitting
  - Can be confident in strength of effects from model

	Optimism	Corrected
R2	0.03	0.29
AUC	0.04	0.79

- Validation statistics are based off of 500 bootstrap samples from each of 32 imputed datasets[1]



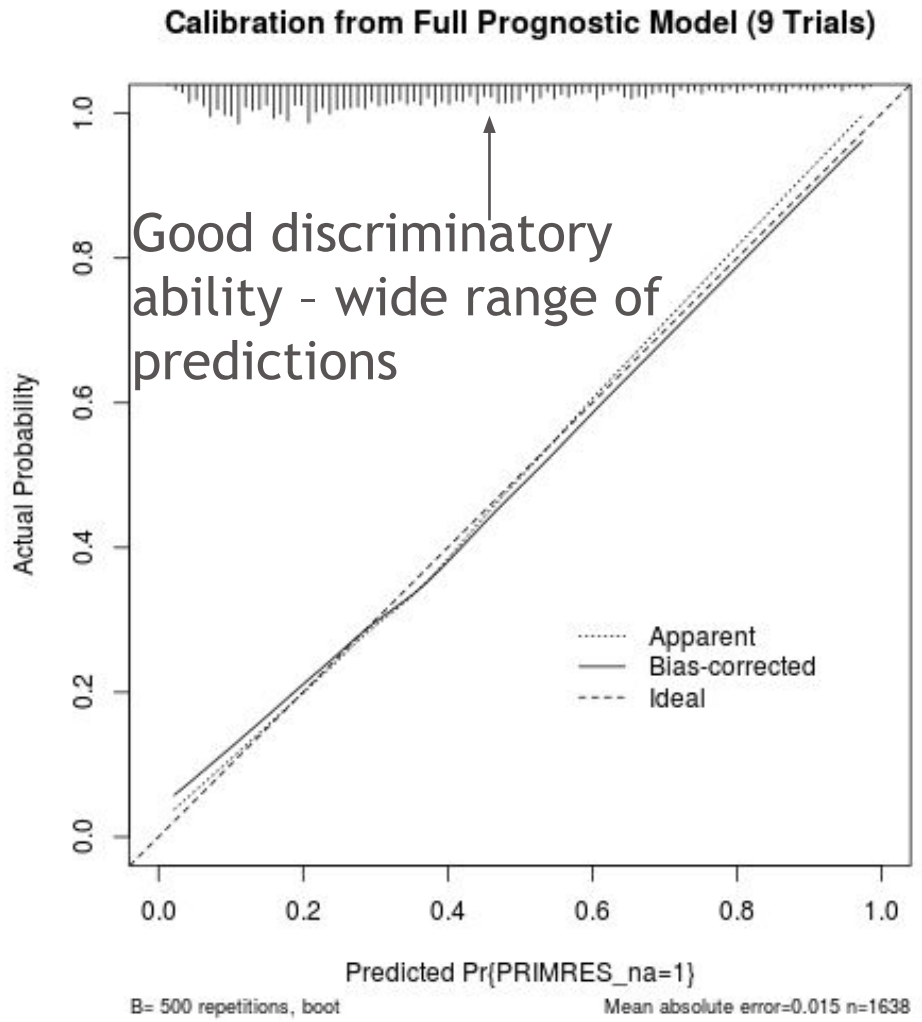


[1] *Clinical Prediction Models*, Ewout Steyerberg.

Internal Validation (Calibration) via Multiple Imputation

- Calibration assesses how precise predictions are
  - As opposed to discrimination, which assesses ability to rank subjects correctly
- Perfect calibration would have intercept = 0 and slope = 1

	Optimism	Corrected
Calibration Intercept	0.05	-0.05
Calibration Slope	0.1	0.9

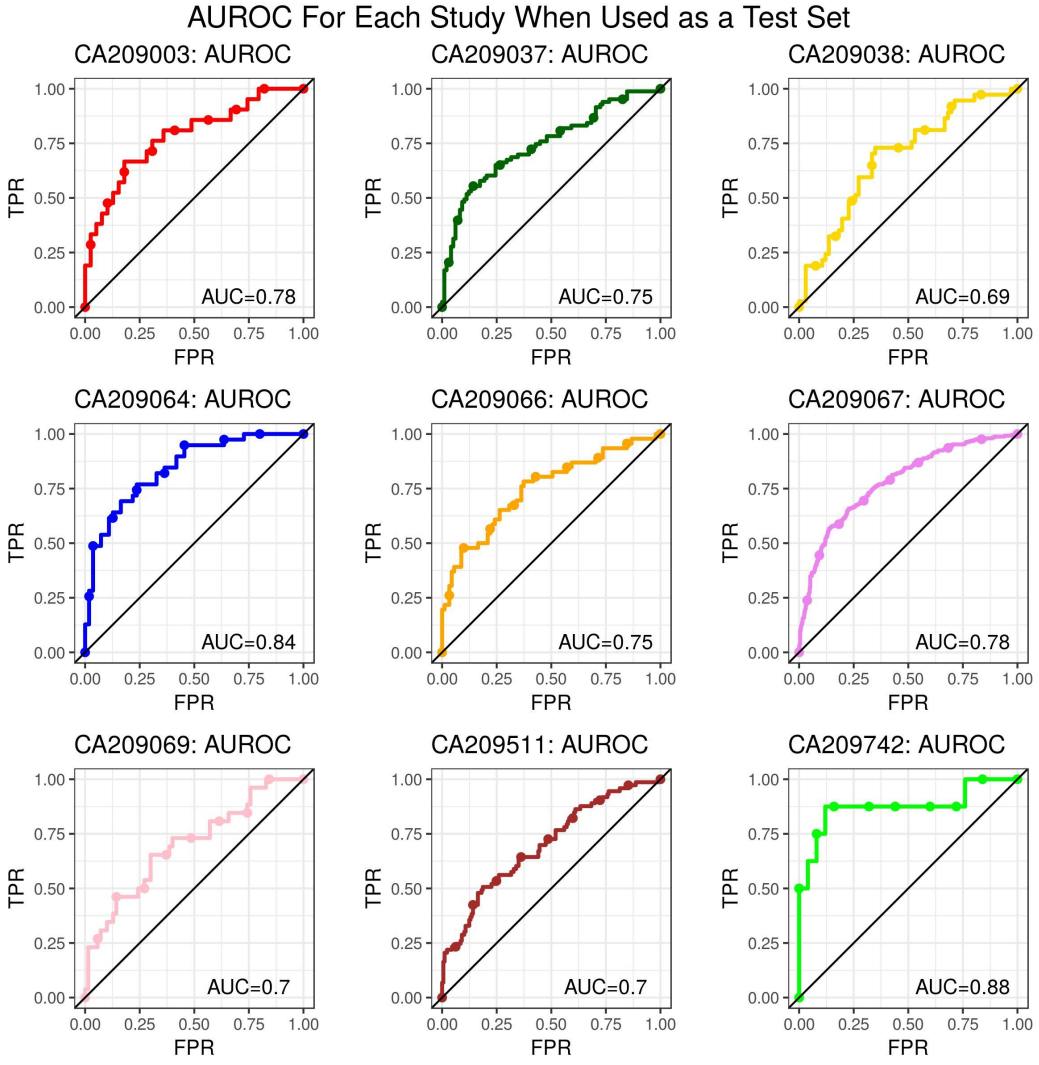


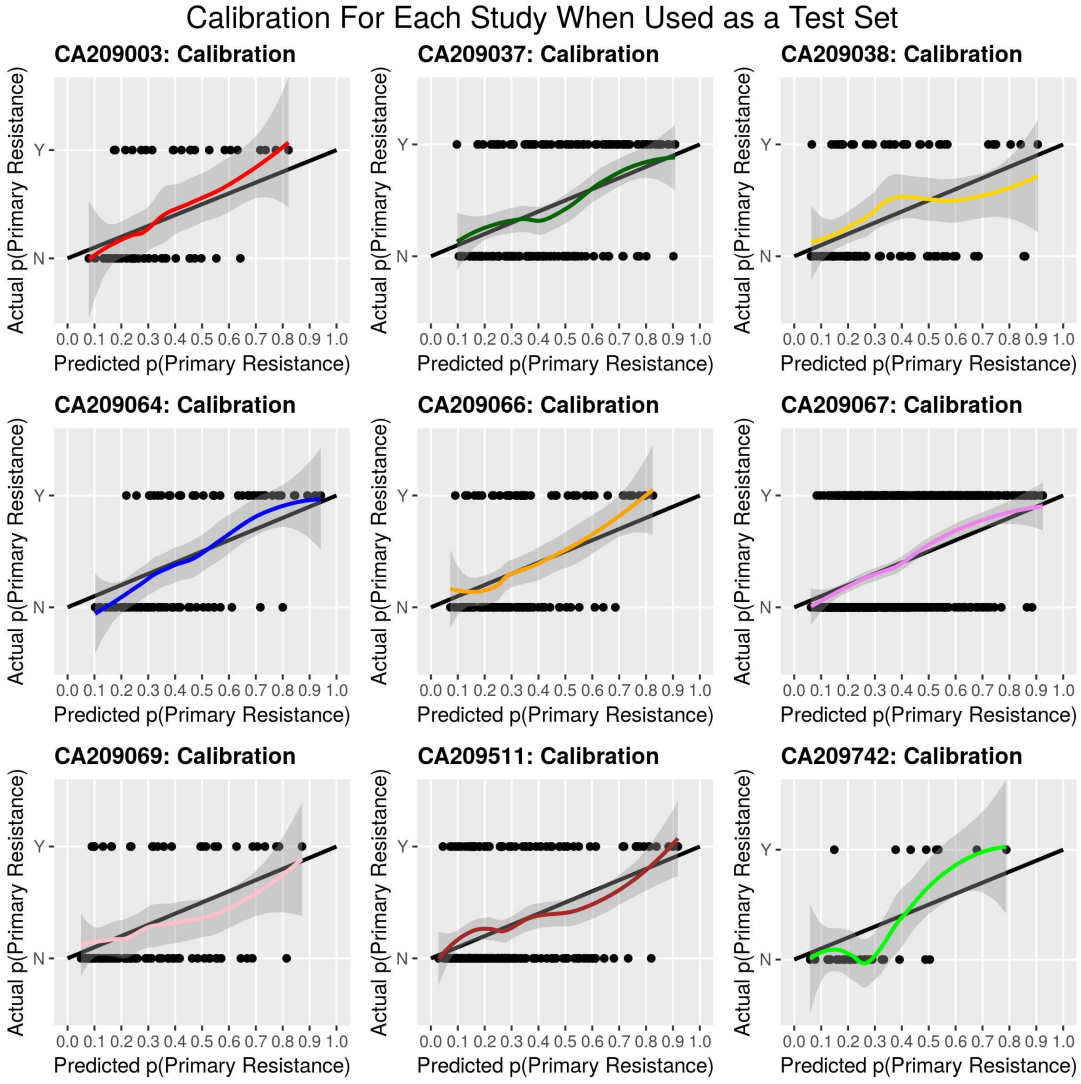
Sample calibration plot from randomly chosen imputed dataset

- 500 Bootstrap samples for each imputed dataset
- **Model seems to be well calibrated – minimal overfitting**
  - Implies confidence that predictions from our modeling can be reproduced in new data

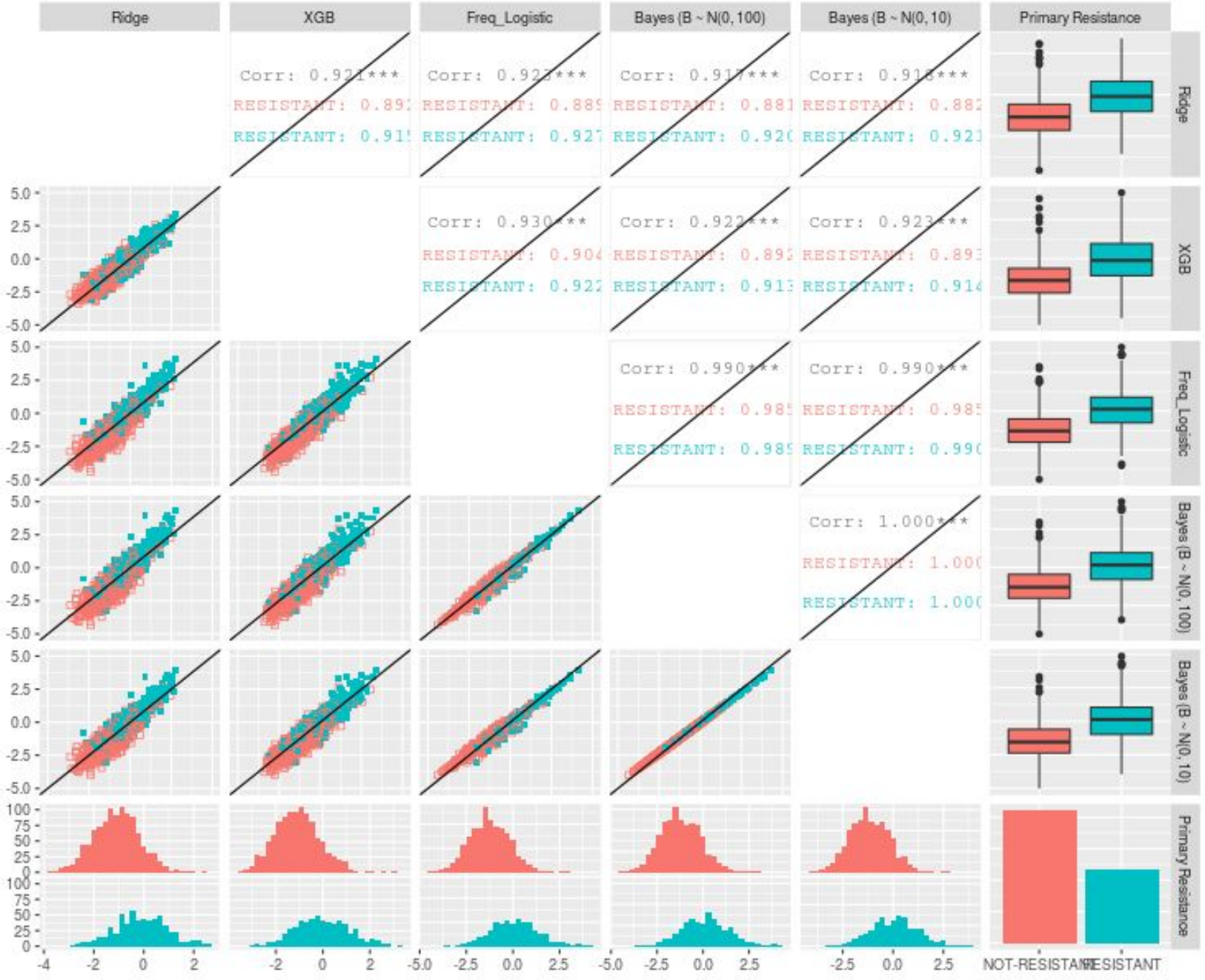
Internal-External Validation Holding Each Study Out Once as a Test Set

- Consistently Good performance Across Studies via internal-external Cross-Validation in which each study was held out once as a test set: AUC's: CM067=0.78, CM038=0.69 , CM742=0.88





Model Comparison: Prediction Matrix (9 Trials, N = 1638)



- If Resistant is above diagonal, “row model” has better predictions
- If Non-Resistant is below diagonal, “row model” has better predictions

Current Methodology Challenges

- **External validation of a Model Developed with Multiple Imputation**
  - With multiple imputation, there are 32 models developed (1 per imputed dataset). How do we combine information across multiple imputed datasets? Should we average coefficients across all imputations? Should we compute separate test statistic for each imputed dataset?
  - How do we combine imputations with external validation? Do we use current imputation model to impute external missing data? Do we create new imputation model with all data?
- **Internal validation with Multiple Imputation**
  - Potential for overly optimistic performance estimates since test data is being used to impute training data.
  - Computationally not feasible to do multiple imputation within resampling.
- **Is there a more efficient way to model Primary Resistance other than binary endpoint?**
- **Balancing trade-off of patients and potentially important variables into models.**
  - e.g., if we include CRP which is strongly patients, we will lose patients as this is not collected in several trials.
- **Having multiple models from multiple modelers creates practical challenges for application in future clinical trials.**
  - We recommend apriori agreeing on the same internal validation and performance metrics
  - Many benefits as well!

Acknowledgements

- Scott Chasalow (IPS/BIOM)
- Shu-Pang Huang (GBDS)

Backup

Multiple Imputation

- Approximate rule of thumb[1]: Number of imputations (M) should equal the proportion of rows in your data which are missing in at least one covariate
- Had trouble including trial in imputation model due to block missing, left it out
- Included outcome in imputation model[1]

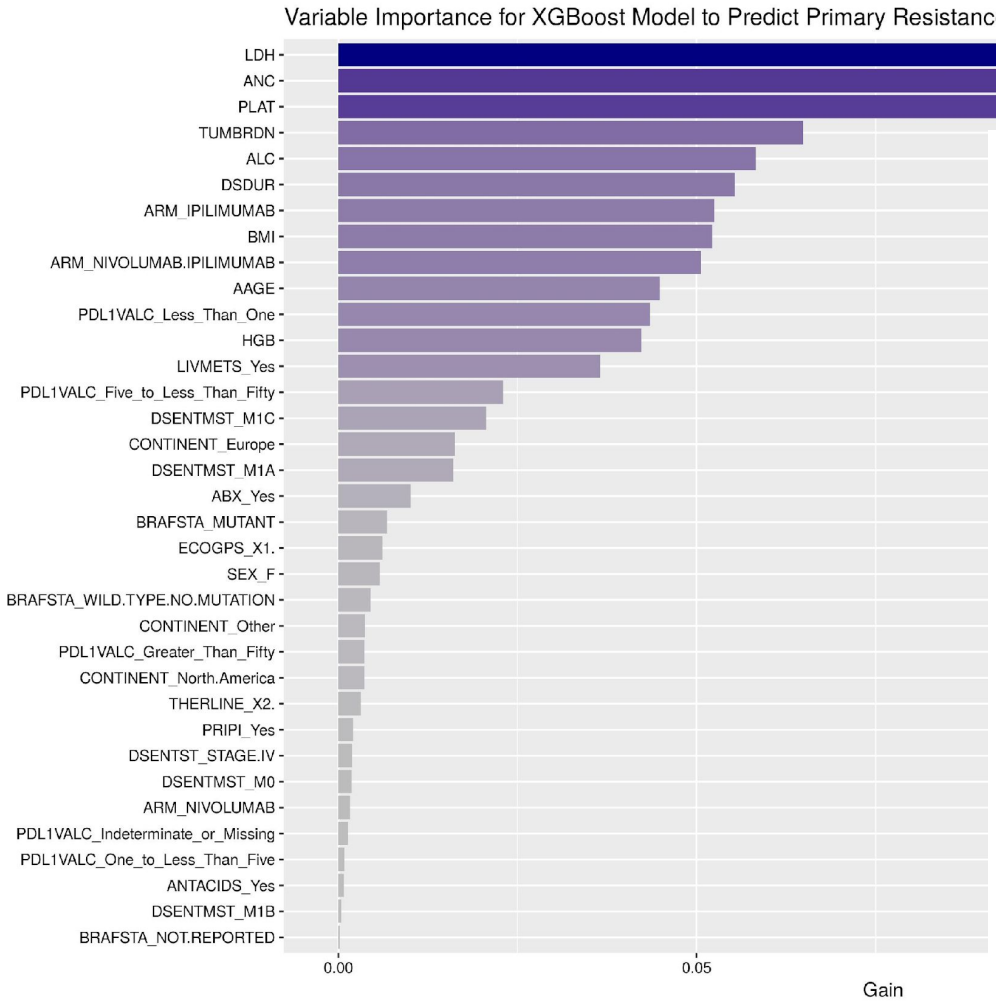
1. *Regression Modeling Strategies*, Frank Harrell

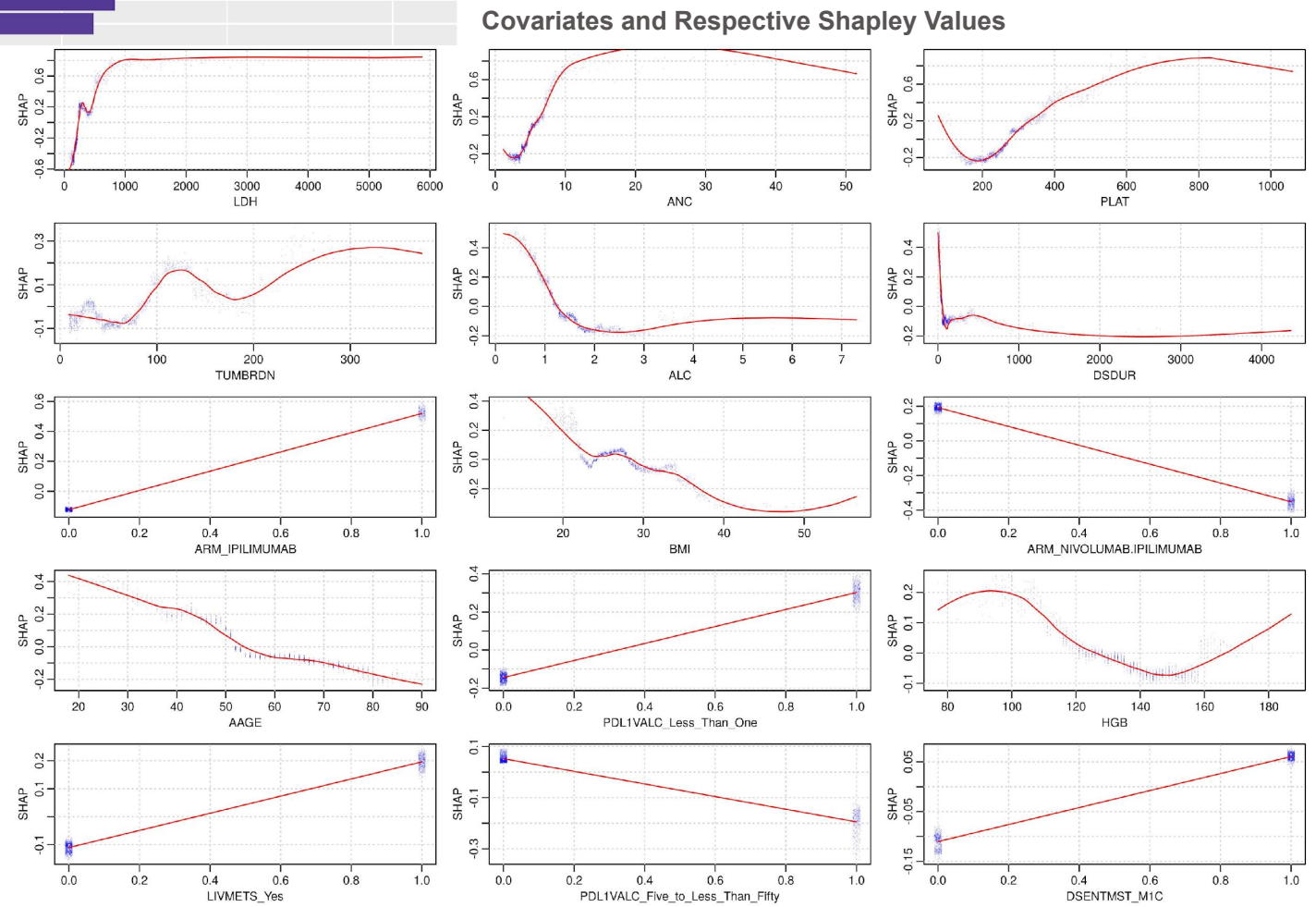
Single Imputation

- Missingness for each covariate handled with Bagged CART models (1 per covariate with missing data) using available data on other covariates.
- These models were developed on the held-in training sets of the cross-validation after resampling to impute the missing training covariates, then applied to the held out sets of the cross-validation to impute missing test set covariates.
- Eliminates potential for data leaking between held-in and held-out sets.

Prognostic Model Summary for XGBoost Model

Top 10 Drivers of Primary Resistance: LDH (+), Platelets (-/+), Neutrophils (+), Lymphocytes (-), Tumor Burden (+), Time to Treatment (-), Treatment Arm (+/-), BMI (-), Age (-), PD-L1 <1% (+)





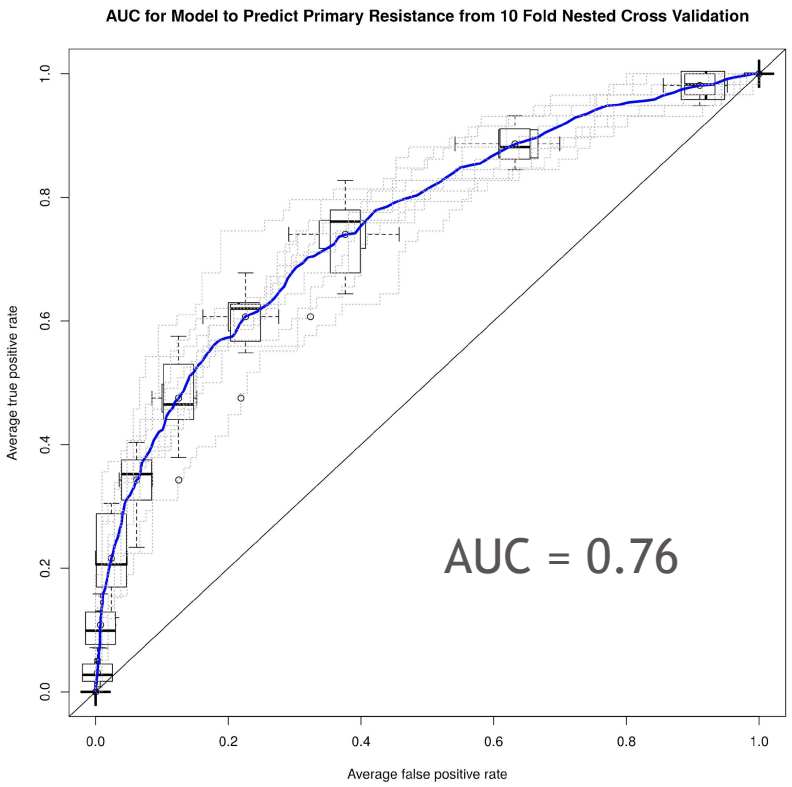
Optimism-Corrected Bootstrap[1, 2]

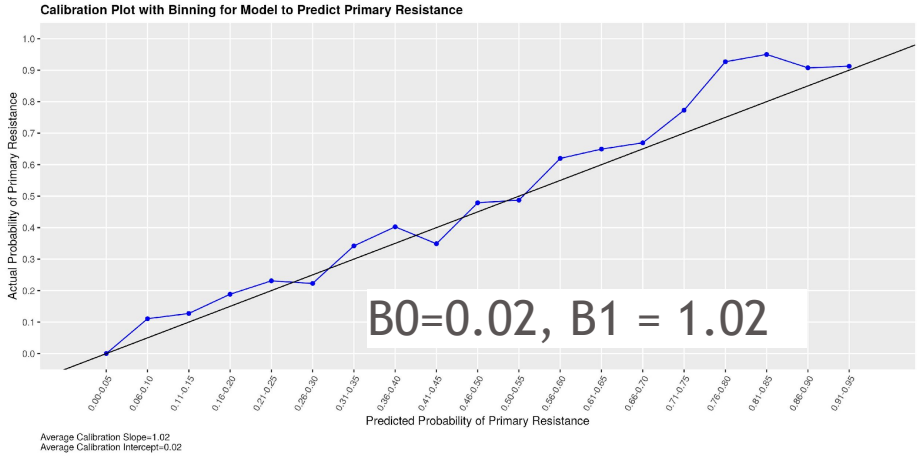
- 2 reasons to prefer over repeated CV:
  1. Easier to validate with imputed data
  2. Uses full data
- Procedure:
  1. Run full model on all of data
  2. Determine performance metrics on full model
  3. Draw bootstrap sample (with replacement) Build model on bootstrap sample and calculate “apparent” performance on same bootstrap sample
  4. Apply model from bootstrap sample on original (full) data
  5. Calculate Optimism by subtracting difference between performance metric calculated in step 3 from that calculated in step 4
  6. After repeating steps 3 – 5 many times, calculate average amount of optimism
  7. Subtract average optimism from performance metric calculated in step 2
  8. Although intuitively there is some leakage in step 4 since on average 63.2% of data from full model is in each bootstrap sample as well, nonetheless both Harrell and Steyerberg assert this method of internal validation has strong basis both in theory and based on many simulations

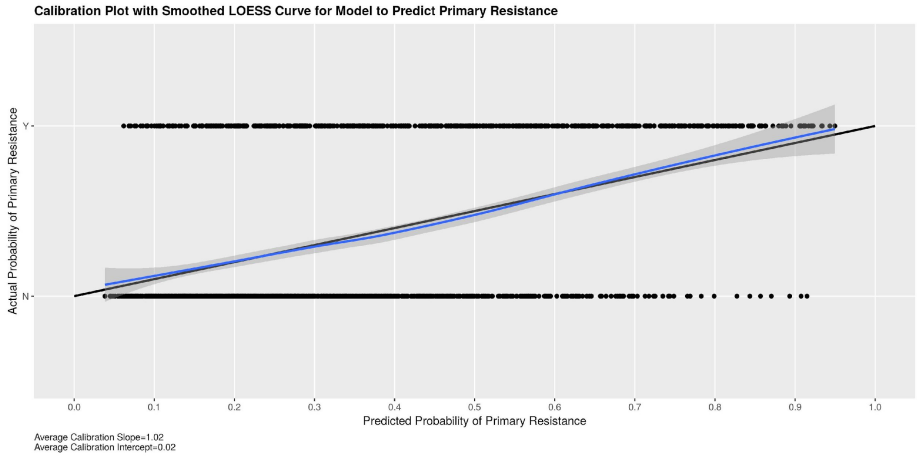
[1] *Clinical Prediction Models*, Ewout Steyerberg.
[2] *Regression Modeling Strategies*, Frank Harrell.

Internal Validation (Discrimination) via Single Imputation

1. Good Discrimination (AUC=0.76) and Excellent Calibration (β0=0.02; β1=1.02) via internal validation (10 Fold Nested Cross-Validation) on the cutaneous melanoma subjects used in model development.

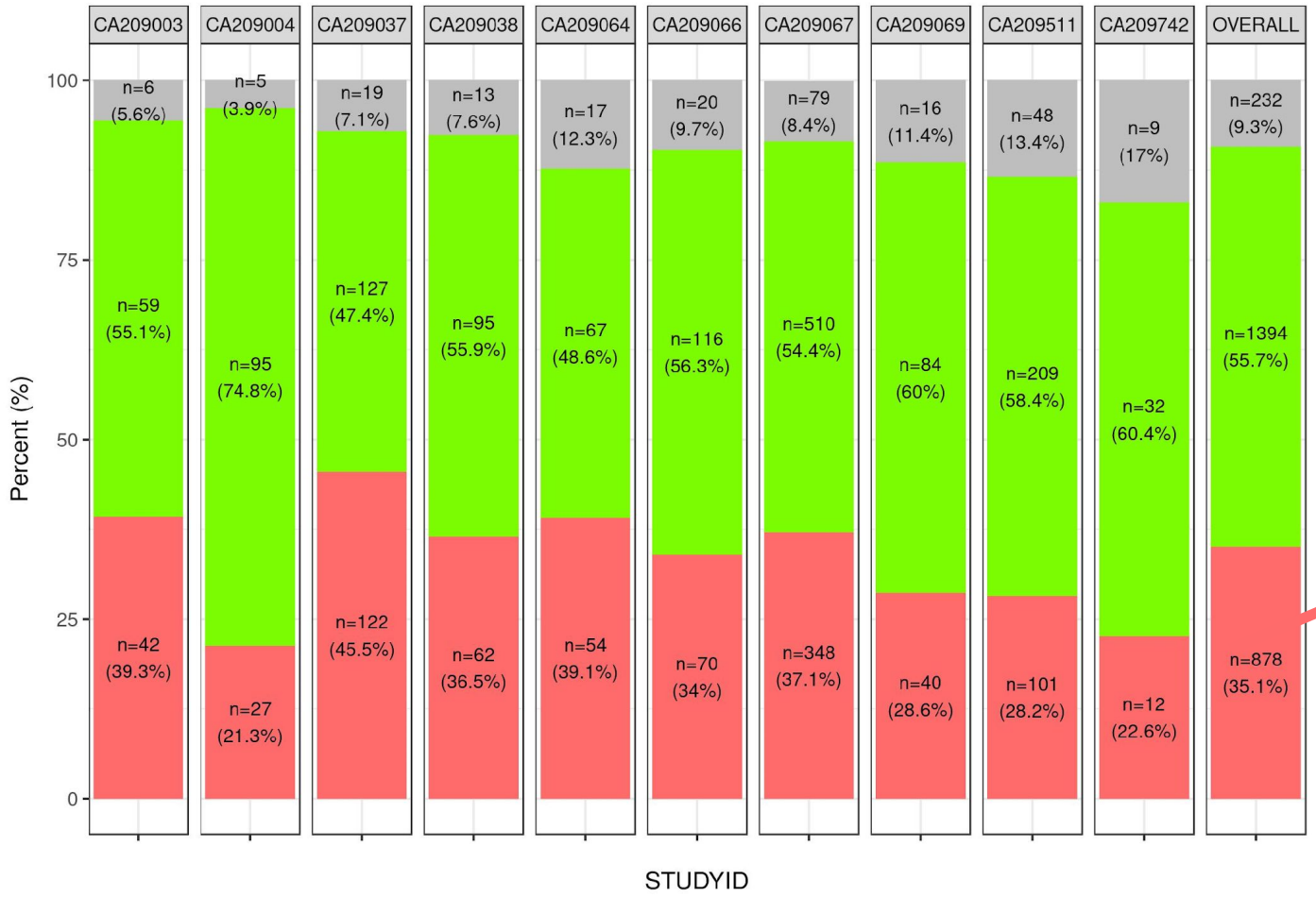






Primary resistance across metastatic melanoma studies: Similar fraction of resistant patients across various studies

Figure 1. Primary Resistance in CA209 Melanoma Trials



35.1% (878/2504) primary resistant

Range

Low of 21.3% in CA209-004

High of 45.5% in CA209-037

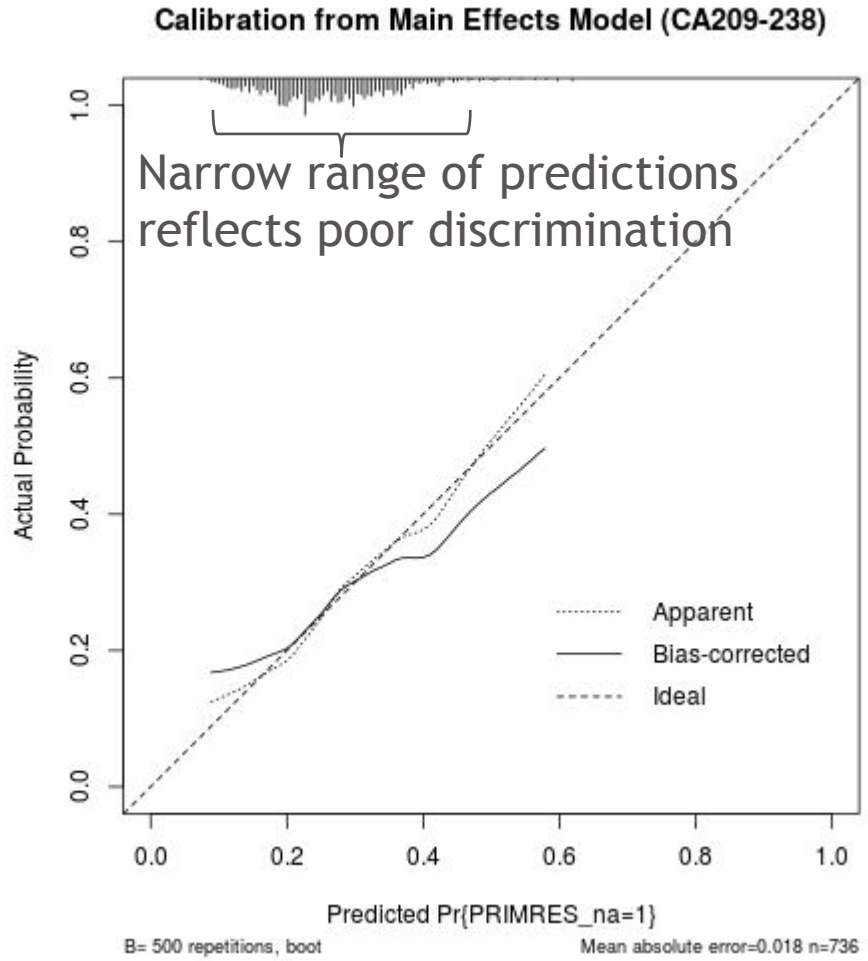
9.3% (232/2504) Not evaluable

Example of Poor Calibration: Full Adjuvant data (n = 872)

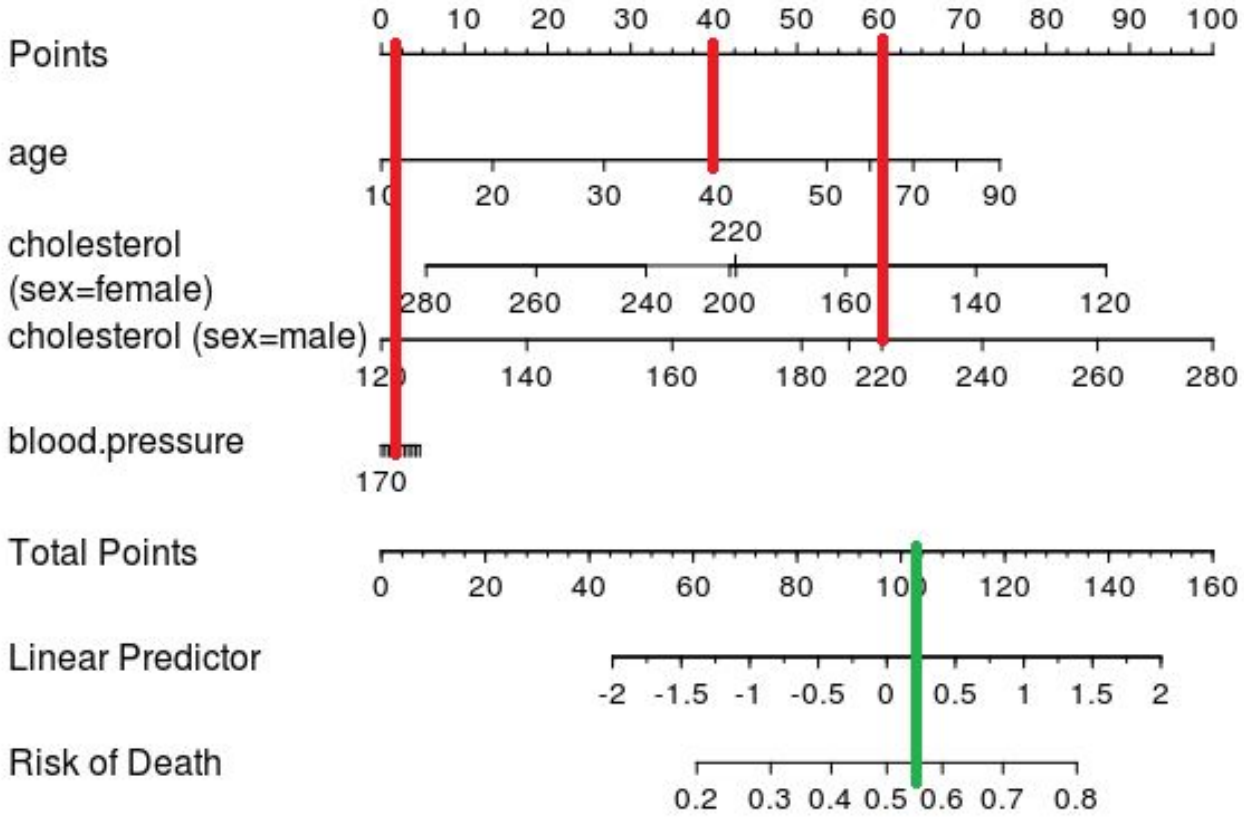
	Optimism	Corrected
Calibration Intercept	0.30	-0.30
Calibration Slope	0.28	0.72

- 500 Bootstrap samples for each imputed dataset
- **Model seems to be poorly calibrated**

Sample calibration plot from randomly chosen imputed dataset



Nomogram Illustration



Obtained from stack exchange:

https://stackoverflow.com/questions/38276973/r-how-to-read-nomograms-to-predict-the-desired-variable/38317862

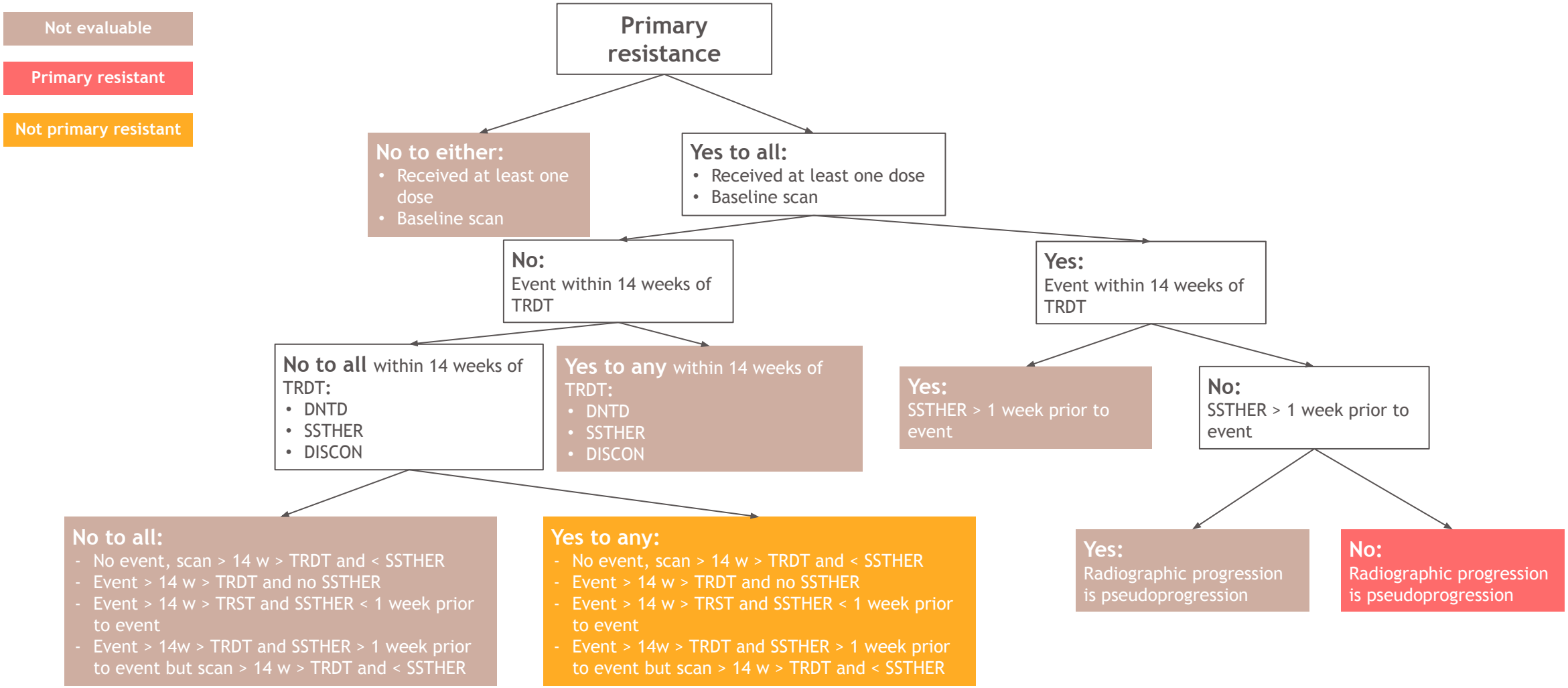
Improvements (for next time)

- Use more informative priors in Bayesian modeling
- If interested in making inference, first draw Directed Acyclic Graph (DAG) to avoid potential collider bias (does not affect predictions)

David Paulucci,Sr. Data Scientist | GBDS

In consultation with the Melanoma Clinical Team

Defining primary resistance (metastatic setting)



Event: Radiographic or clinical progression or death due to disease.

DISCON; discontinuation from study; DNTD, death not due to disease; TRDT, treatment start date; SSTHER, subsequent therapy.