Profiling performance of high-resolution HLA imputation with complementary metrics and visualizations using an open-source validation framework

Alyssa N. Paynter[1], Malek Kamoun[2], Nicholas K. Brown[2], Ryan J. Urbanowicz[3], Keith McCullough[4], Gerald Morris[5], Martin Maiers[6], Massimo Mangiola[7], Brendan Keating[8], Bonnie Lonze[8], Michal Mankowski[8], Loren Gragert[1]

1. Division of Biomedical Informatics and Genomics, Deming Department of Medicine, Tulane University School of Medicine, New Orleans, LA, United States. 2. Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, United States. 3. Department of Computational Biomedicine, Cedars Sinai Medical Center, Los Angeles, CA, United States. 4. Arbor Research Collaborative for Health, Ann Arbor, MI, United States. 5. Department of Pathology, UC San Diego Health, San Diego, CA, United States. 6. NMDP, Minneapolis, MN, United States. 7. Department of Pathology, NYU Grossman School of Medicine, NYU Langone Health, New York, NY, United States. 8. Department of Surgery, NYU Grossman School of Medicine, NYU Langone Health, New York, NY, United States.

INTRODUCTION

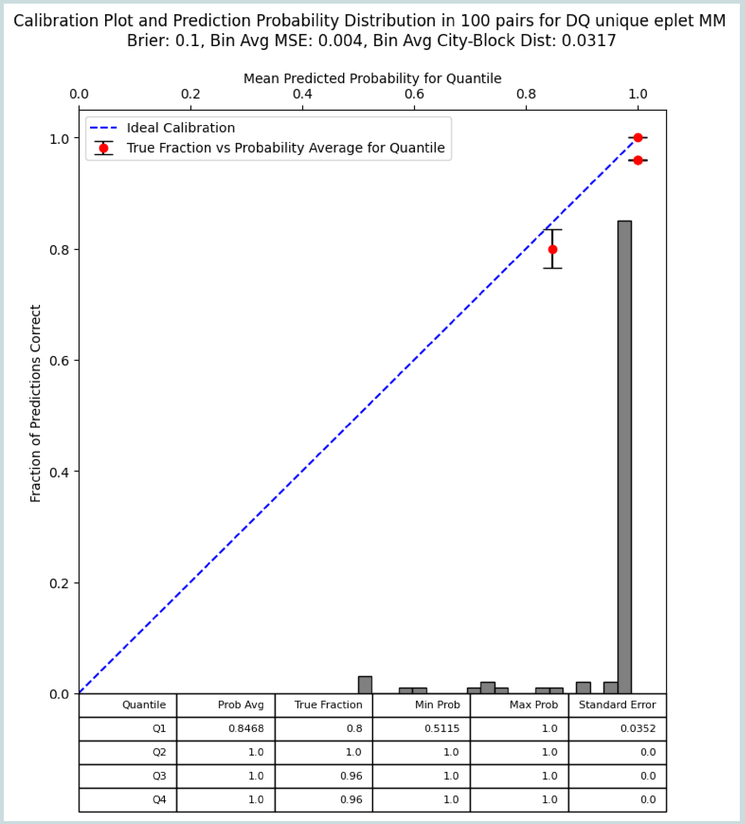
- Research studies evaluating high-resolution (HR) HLA imputation in transplantation have varied substantially in their approaches to measuring prediction performance.
- We aimed to develop a package for computing comprehensive imputation performance metrics and visualizations for predictions of HLA alleles, amino acids, and molecular mismatch categories for donor-recipient pairs with ambiguous typing.

METHODS

- We developed a Python framework for imputation validation where model performance is evaluated using built-in functions from the *scikit-learn* package, a best-in-class machine learning framework.
- We tested the framework on deceased kidney donors (N=217) typed at antigen level from the national registry, and then had HR typing preformed.

		Description
Within the Calibration Plot	**Brier Score**	Measures the distance of individual-level prediction probabilities and correctness for the dataset from an ideal predictor with 100% correct predictions each with 100% probability. (lower scores are better)
	Bin Average Mean Squared Error (MSE)	Used for fitting regression lines, are computed based on squared distance from the perfect diagonal and penalize poor calibration more heavily than city-block distance. (lower scores are better)
	Bin Average City Block Distance	Measures the distance between fraction of correct predictions and the quantile-averaged prediction probabilities, with perfect calibration falling on the dashed-line diagonal. (lower scores are better)
	Histogram of imputation probabilities	Shows the full prediction probability distribution.
	Table for the Quantiles	Imputation probabilities turned into quantiles (true fraction vs probability average) and the table gives summary statistics of each one.
ROC-AUC	**Receiver Operating Characteristic (ROC)-Area Under the Curve (AUC)**	Plots the true positive rate against the false positive rate for the most probable imputation prediction and quantify performance using an area under the curve. (higher AUC is better)

RESULTS



Quantile	Prob Avg	True Fraction	Min Prob	Max Prob	Standard Error
Q1	0.8468	0.8	0.5115	1.0	0.0352
Q2	1.0	1.0	1.0	1.0	0.0
Q3	1.0	0.96	1.0	1.0	0.0
Q4	1.0	0.96	1.0	1.0	0.0

Figure 1. Calibration curves compare prediction probability averages for quantiles with the true fraction. Red points close to diagonal indicate excellent calibration for DQ unique eplet mismatch count predictions.

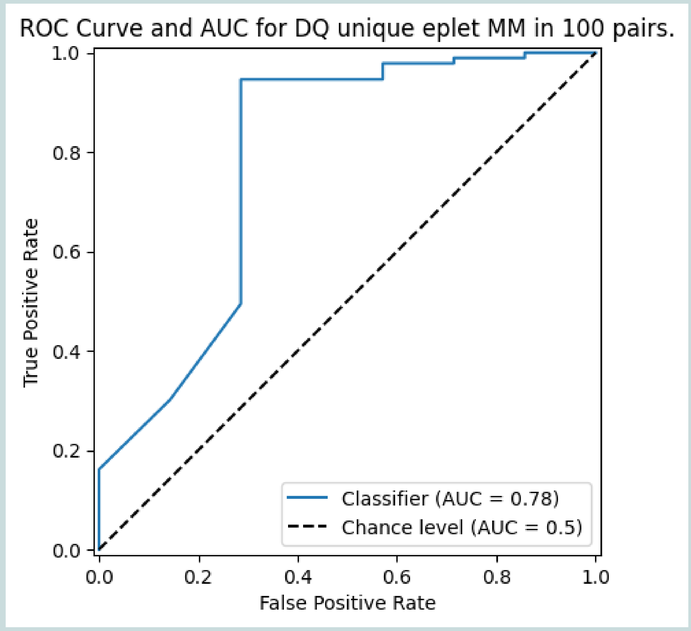
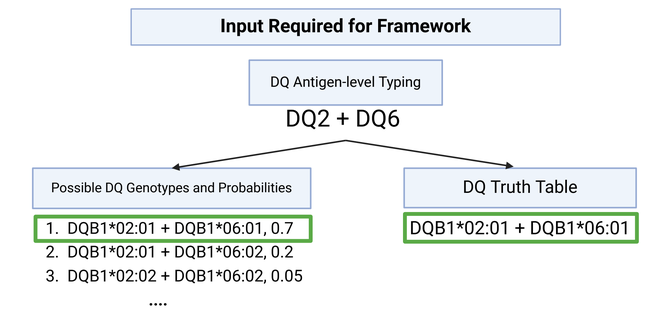


Figure 2. ROC curves that are towards the upper left indicate better performance, however ROC curves are more useful with a more even distribution of prediction probabilities. Most predictions were near to 100%, causing class imbalance.

DISCUSSION



- Imputation distribution and truth table are input
- Mismatched metrics derived from high resolution predictions on transplant pairs.
- *When HLA imputation probabilities are shown to be well calibrated, uncertainty in HLA allele assignments can be incorporated into statistical models for HLA association studies.*

Different Levels of Analysis

Mutli-loci Unphased Genotype: DRB1 + DQA1 + DQB1	**Single Locus Unphased Genotype:** DRB1 + DRB1
Eplet Mismatch: Unique eplet MMs Counts of eplet MM Risk categories	**Amino Acid Mismatch:** Unique AAMMs Counts of AAMMs Risk categories [2]

- The framework is available on GitHub at https://github.com/lgragert/imputation-validation/. Python PyPi package forthcoming.

CONCLUSION

- Calibration plots provide advantages over using ROC metrics and simple accuracy measures, especially in datasets with class imbalance.
- This framework could support high quality HLA imputation studies across immunogenetics and transplant settings.
- Planned features include typing resolution score visualization at amino acid level and support for more mismatch metrics *to more clearly separate the concepts of uncertainty and inaccuracy which are often conflated [1].*

References:

1.Engen, R. M., Jedraszko, A. M., Conciatori, M. A. & Tambur, A. R. Substituting imputation of HLA antigens for high-resolution HLA typing: Evaluation of a multiethnic population and implications for clinical decision making in transplantation. *American Journal of Transplantation***21**, 344–352 (2021).
2.Dasariraju, S. *et al.* HLA amino acid Mismatch-Based risk stratification of kidney allograft failure using a novel Machine learning algorithm. *Journal of Biomedical Informatics***142**, 104374 (2023).

Contact
Alyssa Paynter
Email: apaynter@tulane.edu

Funding
NIDDK R01 DK139240, NIAID U01 AI152960,
NIAID R01 AI173095, UM1TR004771, T32TR4767

Tulane University