Statistical methods for maximizing the utility of electronic health records data for the development, evaluation, and implementation of real-time prognostic models

Yeji Ko

Department of Biostatistics
Vanderbilt University

Oral Candidacy Exam
July 22nd, 2024

Overview

1. **Motivation**
2. **Model Development (Topic 1)**
3. **Model Implementation (Topic 2)**
4. **Model Evaluation (Topic 3)**
5. **Future Plans**

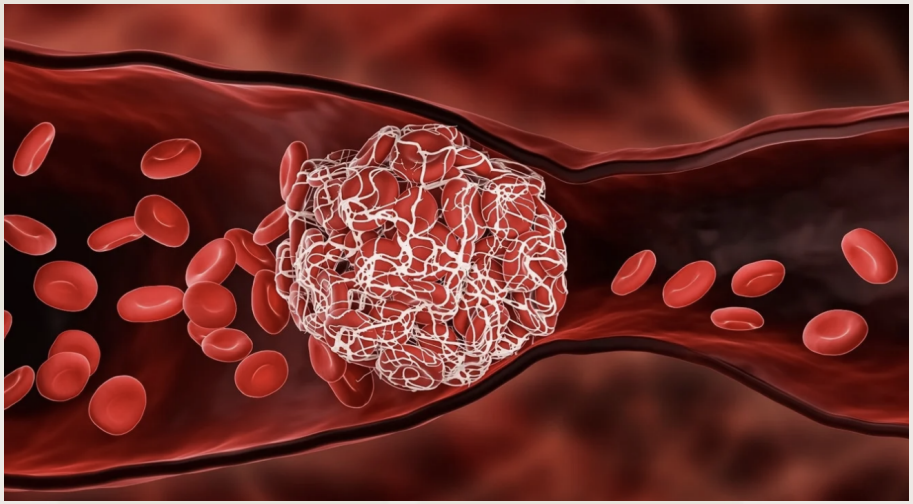
VANDERBILT UNIVERSITY MEDICAL CENTER

1. Motivation

VANDERBILT UNIVERSITY
MEDICAL CENTER

Clinical motivation: What is Venous thromboembolism?

- Venous thromboembolism (VTE) occurs when a blood clot forms in a vein.
- VTE is a common complication among hospital inpatients, contributing to longer hospital stays, increased morbidity, and higher mortality rates.
- Prompt medical attention is crucial, making real-time prediction of VTE imperative.



VANDERBILT UNIVERSITY MEDICAL CENTER

Statistical motivation: Challenge in electronic health records

- An Electronic Health Record (EHR) is a digital version of a patient's medical history, maintained by the healthcare provider over time. It may include all key administrative and clinical data relevant to the patient's care.
- EHR systems are valuable for developing and implementing prognostic models.
- A major challenge is synthesizing multiple predictors to develop accurate prediction models.
- Missing data, especially informative absences reflecting a patient's clinical status, potentially impact prognostic accuracy.
- Most imputation methods focus on accurate value estimation, leaving their effectiveness in real-time prediction modeling uncertain.

VANDERBILT UNIVERSITY MEDICAL CENTER

2. Model Development (Topic 1)

VANDERBILT UNIVERSITY
MEDICAL CENTER

Data overview

- Data from VUMC's Research Derivative (RD) database.
- Derivation cohort: Inpatients from January 1, 2018, to December 31, 2020 (n=132,330).
- Validation cohort: Inpatients from January 1, 2021, to June 30, 2022 (n=62,546).
- Used ICD-10 codes to identify HA-VTE cases; manual review for accuracy.
- Screened 82 potential risk factors: demographic/clinical characteristics, diagnostic procedures, vital signs, and lab measurements.

VANDERBILT UNIVERSITY MEDICAL CENTER

Primary objectives

- Train and validate statistical and machine learning models for real-time clinical prediction of VTE.
  - Logistic regression
  - Random forest
  - XG boost
  - Neural network
  - Ensemble method
- Compare the performance between models.
  - C statistic, Integrated calibration index (ICI), Calibration slope, Brier score

VANDERBILT UNIVERSITY MEDICAL CENTER

Model description

- **Logistic regression:** A statistical model that captures the relationship between predictors and binary outcome using a logistic function.
- **Random forest:** This machine learning algorithm operates by taking the majority vote of multiple decision trees, improving prediction accuracy and robustness.
- **XG boost**: A machine learning algorithm that enhances prediction by sequentially updating decision trees to minimize the loss function and optimize the model.
- **Neural network**: This machine learning model identifies complex relationships in a dataset by mimicking the functions of the human brain through interconnected nodes.
- **Ensemble method**: An advanced algorithm that combines the predictions of the four models above by averaging their outputs to enhance overall prediction accuracy.

VANDERBILT UNIVERSITY MEDICAL CENTER

Model characteristics

Table 2. Classifier characteristics

Classifier	R package	Command	Requires dummy coding	Hyperparameters (argument)	Selected hyperparameter values
Logistic regression	*caret*	*glm*	No	• None	• None
Random forest	*randomforestSRC*	*rfsrc*	No	• Number of tree (ntree) • Number of split (mtry) • Node size (nodesize)	• ntree = 3000 • mtry = 82 • nodesize = 30
XG boost	*caret*	*xgbTree*	No	• Learning rate (eta) • Max depth of tree (max_depth) • Min split loss (gamma) • Percentage of features for tree (colsample_bytree) • Min sum of instance weight (min_child_weight)	• eta = 0.3 • max_depth = 2 • gamma = 0 • colsample_bytree = 0.8 • min_child_weight = 1
Neural network	*keras*	*keras_model_sequential*	Yes	• Number of layers (layer_dense) • Number of neurons (units) • Drop-out rates (layer_dropout) • Activation function (activation) • Loss function (loss), Optimizer (optimizer) • Number of epochs (epochs), Batch size (batch_size)	• layer_dense = 4 • units = 84 -> 42 -> 20 -> 10 • layer_dropout = 0.3-0.4 • activation = relu and sigmoid • loss = binary_crossentropy • optimizer = adam • epochs = 10 • batch_ size = 50

AUC is optimized via bootstrap in *caret* and cross validation in *randomforestSRC* and *keras.*

VANDERBILT UNIVERSITY MEDICAL CENTER

Key results

Table 2. Performance comparison between models

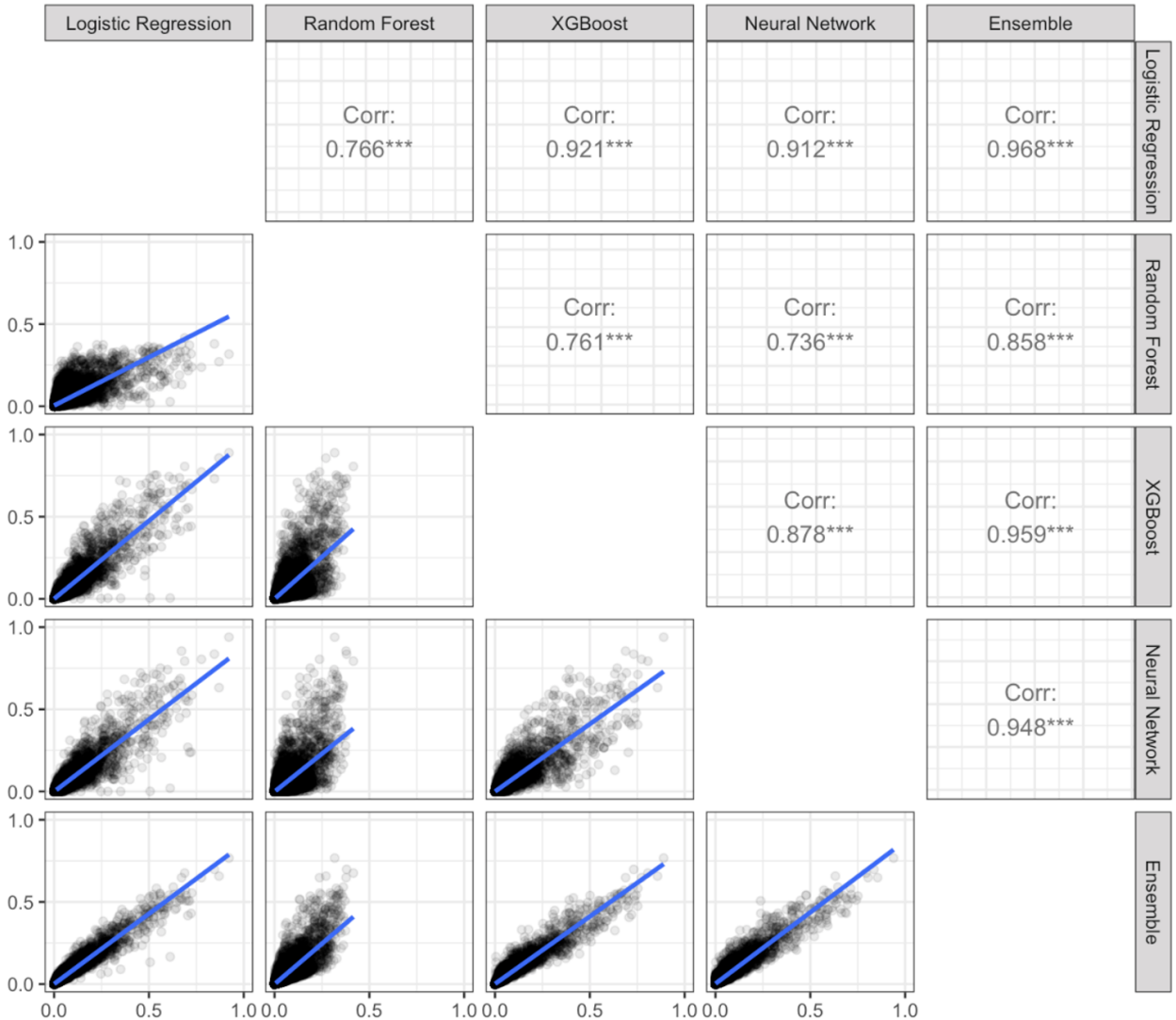
Classifier	Derivation cohort	Validation cohort			
	C statistic	C statistic	ICI[1]	Calibration slope	Brier score
Logistic regression	0.896	0.894	0.0037	0.983	0.0122
Random forest	0.993	0.886	0.0030	0.951	0.0124
XG boost	0.908	0.893	0.0044	0.971	0.0122
Neural network	0.923	0.885	0.0048	0.969	0.0122
Ensemble method	0.975	0.901	0.0042	1.024	0.0121

[1]ICI: Integrated Calibration Index

VANDERBILT UNIVERSITY MEDICAL CENTER

Key results

Figure 3. Correlation in predicted values between models



VANDERBILT UNIVERSITY MEDICAL CENTER

Key results

Table 3. Model accuracy measures at risk cut-off of 0.02

Classifier	Sensitivity	Specificity	PPV[1]	NPV[2]	F1[3]
Logistic regression	0.690	0.904	0.092	0.995	0.162
Random forest	0.710	0.880	0.077	0.995	0.138
XG boost	0.670	0.914	0.098	0.995	0.171
Neural network	0.703	0.898	0.088	0.995	0.156
Ensemble method	0.719	0.900	0.092	0.996	0.163

[1]PPV: Positive Predictive Value, [2]NPV: Negative Predictive Value

[3]F1 score is the harmonic mean of PPV (precision) and sensitivity (recall).

F1 scores can range from 0 to 1, with 1 indicating a model with perfect classification.

VANDERBILT UNIVERSITY MEDICAL CENTER

Summary of the results

- All classifiers perform similarly well in terms of discrimination and calibration.
- More complex classifiers, known to handle interactions better, do not show improved performance compared to logistic regression.
- Random forest has a relatively low correlation with all other models, except for ensemble method. This is because its predicted probabilities cannot exceed 0.5.
- Model accuracy measures are more balanced when a cut-off of 0.02 (clinically meaningful cut-point) is used. Still, the sensitivity, PPV, and F1 are low due to a low prevalence of VTE.

VANDERBILT UNIVERSITY MEDICAL CENTER

3. Model Implementation (Topic 2)

VANDERBILT UNIVERSITY
MEDICAL CENTER

Primary objectives

- Conduct a simulation study to evaluate imputation performance.
- Imputation involves replacing a missing value in the data with a reasonable estimate:
  - Fixed-value imputation: Median
  - Multiple imputation:
    - Data augmentation
    - Multivariate imputation by chained equations (MICE)
    - Conditional modeling
    - Joint modeling

VANDERBILT UNIVERSITY MEDICAL CENTER

Primary objectives

- Evaluate the performance of real-time VTE prediction models with imputed missing values.
  - Logistic regression
  - Random forest
  - XG boost
  - Neural network
  - Two-stage logistic regression model conditioning on a missing indicator M
    - Stage 1: Regression model conditioning on $M = 1$
    - Stage 2: Regression model conditioning on $M = 0$
    - This model does not require data imputation.

VANDERBILT UNIVERSITY MEDICAL CENTER

Simulation set-up

Define outcome Y and five predictors of interest, denoted as X_1, X_2, X_3, X_4, X_5.

- Y: Fully observed binary outcome (VTE status)
- X_1: Fully observed binary non-lab variable (history of VTE)
- X_2: Fully observed five-level categorical non-lab variable (type of admission)
- X_3: Fully observed continuous lab variable (BUN)
- X_4: Partially observed continuous lab variable (CRP)
- X_5: Fully observed continuous (auxiliary) lab variable associated with missingness in X_4 (CL)
- M: Missingness in X_4, where $M = 1$ if X_4 is absent

VANDERBILT UNIVERSITY MEDICAL CENTER

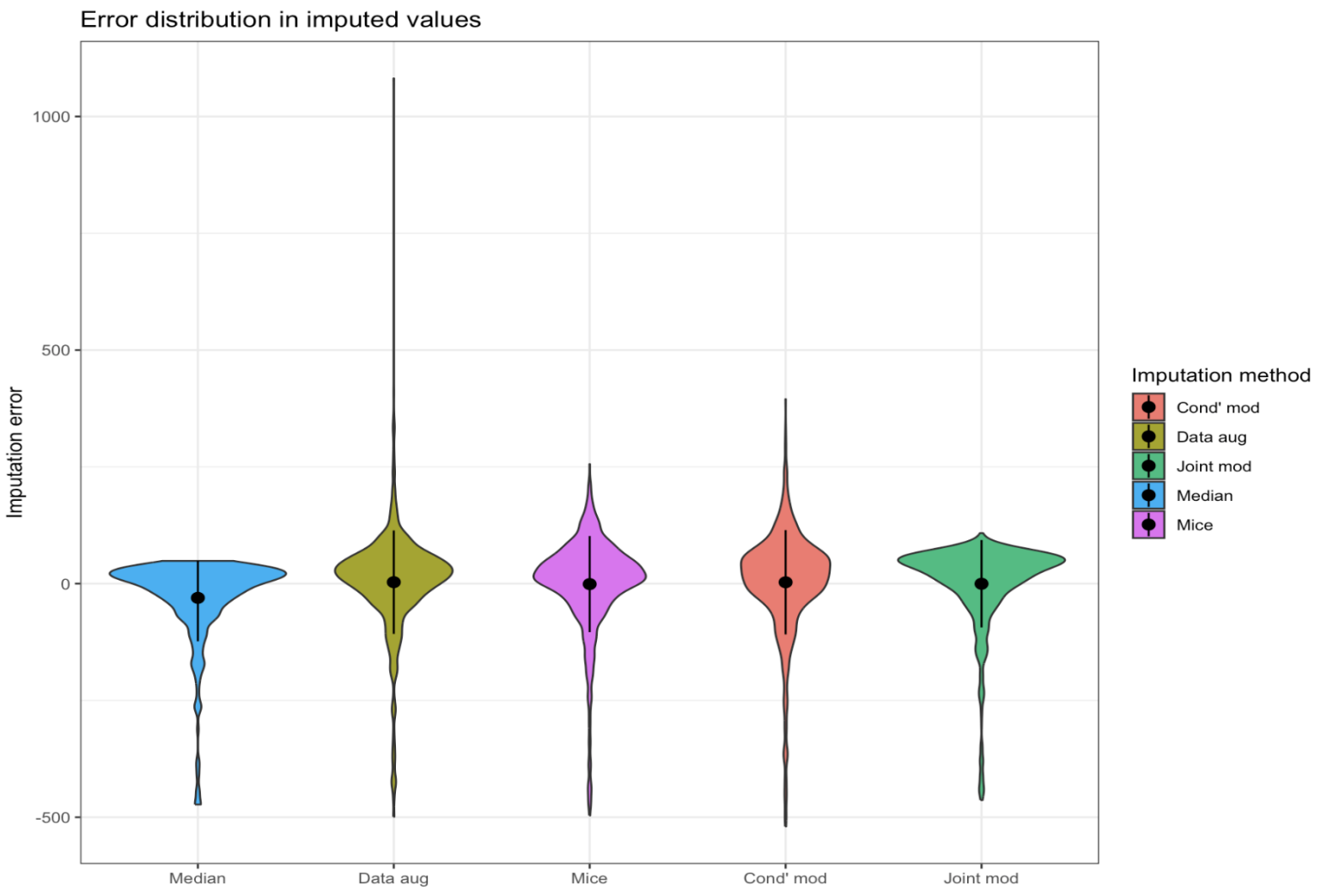
Simulation set-up

- Step 1: Generate data for a total population of 10,000 patients and compute true risks of each individual. Generate missing indicator (M) for all patients from a logistic distribution under MAR.
- Step 2: Develop five models on the true population with predictors X_1,X_2,X_3,X_4 and outcome Y.
- Step 3: Draw a sample of 100 patients (n_{sample}) from the total population. Draw 99 patients from non-missing group ($M = 0$) and draw 1 patient from missing group ($M = 1$).
- Step 4: Define a patient from the missing group as a new patient D^i. Impute the missing value of a new patient, D^i, using various imputation methods.
- Step 5: Predict risk of VTE for the patient i with the five models developed in Step 2.
- Step 6: Evaluate the performance of the imputation methods.
- Step 7: Compare the true risks obtained from the true population (Step 1) with the risk predictions based on imputation (Step 5). Compare performance and system time between different imputation methods and between five models. Repeat the step 3-7 1000 times.

VANDERBILT UNIVERSITY MEDICAL CENTER

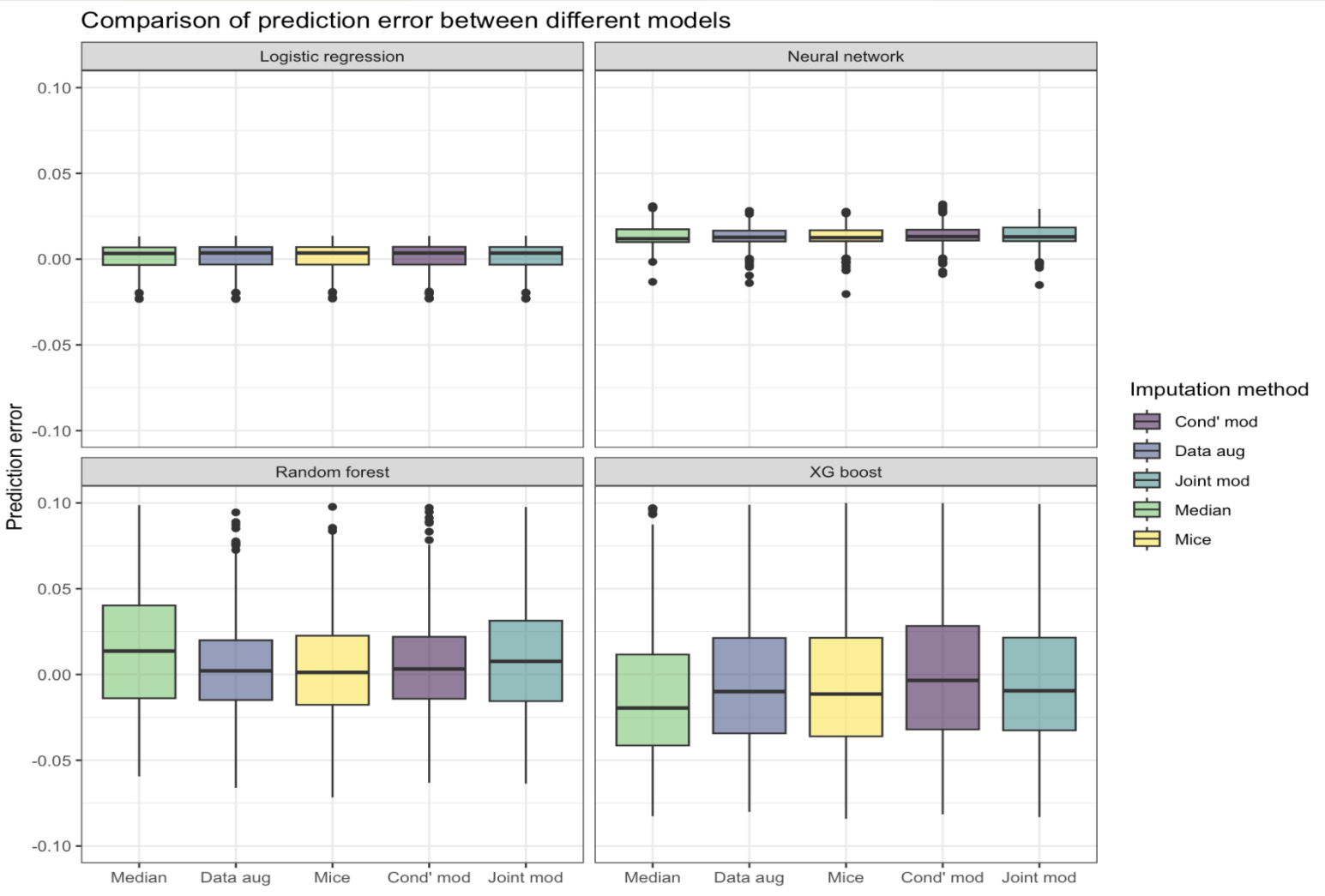
Simulation results

Event rate (0.1), Absence rate (0.1), MAR
Imputation error = Imputed value - True value



Simulation results

Event rate (0.1), Absence rate (0.1), MAR
Prediction error = Predicted risk - True risk



Simulation results

Event rate (0.1), Absence rate (0.1), MAR
The units for time spent is seconds.

Mean time spent for each imputation method				
Median	Data augmentation	MICE	Cond' modeling	Joint modeling
0	1.382	1.588	1.393	1.393

VANDERBILT UNIVERSITY MEDICAL CENTER

Summary of the results

- **Median imputation:** Errors from median imputation tend to center around 0 and exhibit much less variability compared to other imputation methods. However, median imputation does not work well with tree-based methods.
- **Performance of imputation methods:** While all imputation methods perform similarly, data augmentation shows the highest variability with some outliers. Despite this, its performance in prediction models is quite good.
- **Logistic regression:** Logistic regression has the lowest bias and variability, since it is the model used to generate the outcome in the simulation.
- **Neural network:** This model has low variability but tends to overestimate the probability of VTE.
- **Random forest and XG boost:** These tree-based models show higher variability in predictions overall. In particular, XG boost has the highest variability because it is more likely to overfit compared to robust bagging methods. Still, prediction errors in any imputation method other than median imputation are centered around 0.

VANDERBILT UNIVERSITY MEDICAL CENTER

4. Model Evaluation (Topic 3)

VANDERBILT UNIVERSITY
MEDICAL CENTER

Motivation

- Prediction models for VTE ignore potentially important information regarding the event.
  - Type of VTE: Pulmonary embolism, Deep vein thrombosis, Obstetric thromboembolism, etc.
  - Severity of VTE
- Additional information could facilitate more accurate prediction.
  - A model could have limited accuracy for all VTE.
  - The model might accurately predict a certain type of VTE.
  - The model might accurately predict more severe VTE.

VANDERBILT UNIVERSITY MEDICAL CENTER

Sketch of statistical approach

- Develop ROC-based summaries for predication accuracy
- Considering event timing
- Considering event severities on a continuous scale (marks)
- Extension of time-dependent prediction accuracy summaries introduced by Heagerty and colleagues

VANDERBILT UNIVERSITY
MEDICAL CENTER

5. Future Plans

VANDERBILT UNIVERSITY
MEDICAL CENTER

Current progress and future plans

- Topic 1: Analysis is complete, and a manuscript is in preparation.
- Topic 2: Conduct a comprehensive simulation study considering a range of simulation settings.
- Topic 3: Development of an estimator and evaluate its properties.

VANDERBILT UNIVERSITY
MEDICAL CENTER

References I

Alzubaidi, L., J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data 8*, 1–74.

Azur, M. J., E. A. Stuart, C. Frangakis, and P. J. Leaf (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research 20*(1), 40–49.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Collins, G. S., J. A. de Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L.-M. Yu, K. G. Moons, et al. (2014). External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology 14*, 1–11.

French, B. and P. J. Heagerty (2009). Marginal mark regression analysis of recurrent marked point process data. *Biometrics 65*(2), 415–422.

Groenhof, T., Z. Rittersma, M. Bots, M. Brandjes, J. Jacobs, D. Grobbee, W. Van Solinge, F. Visseren, S. Haitjema, F. Asselbergs, et al. (2019). A computerised decision support system for cardiovascular risk management 'live'in the electronic health record environment: development, validation and implementation—the utrecht cardiovascular cohort initiative. *Netherlands Heart Journal 27*, 435–442.

VANDERBILT UNIVERSITY MEDICAL CENTER

References II

Haber, G., J. Sampson, and B. Graubard (2021). Bias due to berkson error: issues when using predicted values in place of observed covariates. *Biostatistics 22*(4), 858–872.

Hadeed, S. J., M. K. O'rourke, J. L. Burgess, R. B. Harris, and R. A. Canales (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of the Total Environment 730*, 139140.

Harton, J., N. Mitra, and R. A. Hubbard (2022). Informative presence bias in analyses of electronic health records-derived data: a cautionary note. *Journal of the American Medical Informatics Association 29*(7), 1191–1199.

Heagerty, P. J., T. Lumley, and M. S. Pepe (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics 56*(2), 337–344.

McCulloch, C. E., J. M. Neuhaus, and R. L. Olin (2016). Biased and unbiased estimation in longitudinal studies with informative visit processes. *Biometrics 72*(4), 1315–1324.

Moons, K. G., R. A. Donders, T. Stijnen, and F. E. Harrell Jr (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology 59*(10), 1092–1101.

Neal, P. and T. Kypraios (2015). Exact bayesian inference via data augmentation. *Statistics and Computing 25*, 333–347.

VANDERBILT UNIVERSITY MEDICAL CENTER

References III

Nijman, S. W. J., T. K. J. Groenhof, J. Hoogland, M. L. Bots, M. Brandjes, J. J. Jacobs, F. W. Asselbergs, K. G. Moons, and T. P. Debray (2021). Real-time imputation of missing predictor values improved the application of prediction models in daily practice. *Journal of Clinical Epidemiology 134*, 22–34.

Perkins, N. J., S. R. Cole, O. Harel, E. J. Tchetgen Tchetgen, B. Sun, E. M. Mitchell, and E. F. Schisterman (2018). Principled approaches to missing data in epidemiologic studies. *American Journal of Epidemiology 187*(3), 568–575.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks 61*, 85–117.

Scornet, E. (2016). Random forests and kernel methods. *IEEE Transactions on Information Theory 62*(3), 1485–1500.

Sisk, R., L. Lin, M. Sperrin, J. K. Barrett, B. Tom, K. Diaz-Ordaz, N. Peek, and G. P. Martin (2021). Informative presence and observation in routine health data: a review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association 28*(1), 155–166.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*(398), 528–540.

Thrun, S., W. Burgard, and D. Fox (2005). *Probabilistic Robotics* (1st ed.). The MIT Press.

VANDERBILT UNIVERSITY MEDICAL CENTER

References IV

Van der Heijden, G. J., A. R. T. Donders, T. Stijnen, and K. G. Moons (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of Clinical Epidemiology 59*(10), 1102–1109.

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (mice) package. *Annals of Translational Medicine 4*(2).

Zhang, Z., M. W. Beck, D. A. Winkler, B. Huang, W. Sibanda, H. Goyal, et al. (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine 6*(11).

VANDERBILT UNIVERSITY
MEDICAL CENTER

Thank You!

VANDERBILT UNIVERSITY

MEDICAL CENTER