Risk Prediction of New-Onset Atrial Fibrillation Using Tree Ensemble Methods on Synthetic EHR and ECG Data

UNIVERSITY OF TORONTO
DALLA LANA SCHOOL OF PUBLIC HEALTH

Luke Bai[1], Zhengyang Fei[1], Vanessa Yuan Liao[1], Zhaoyu Tan[1]

[1]Dalla Lana School of Public Health, Biostatistics Division University of Toronto, Toronto, ON, Canada.

Background

Atrial Fibrillation (AF)

The most common heart rhythm disorder, affecting up to 1 in 3 individuals over the age of 45[1,2], and can increase the risk of stroke over 4 times[1]. Earlier detection supports early intervention efforts.

Existing Risk Prediction Models

Traditional risk scores like C2HEST and CHARGE-AF have only shown modest predictive performance in validation datasets[6]. Additionally, these risk scores are often not generalizable to patients routinely encountered across all clinical practice environments.

Objectives

1) Develop a risk prediction model that can accurately predict the future occurrence of AF and time to occurrence
2) Evaluate the performance of two tree based ensemble methods:
   - **Random survival forest (RSF)**
   - **Gradient Boosted Cox Model (GBM)**

Models are assessed using ROC-AUC

Methods

Dataset

- Synthetic patient data from the Cardiovascular Imaging Registry of Calgary
- Outcome: AF diagnosis, time to AF diagnosis
- Predictors: Demographics, comorbidities, cardiac history and procedures, ECG variables, laboratory test results, medication prescriptions

Data Processing

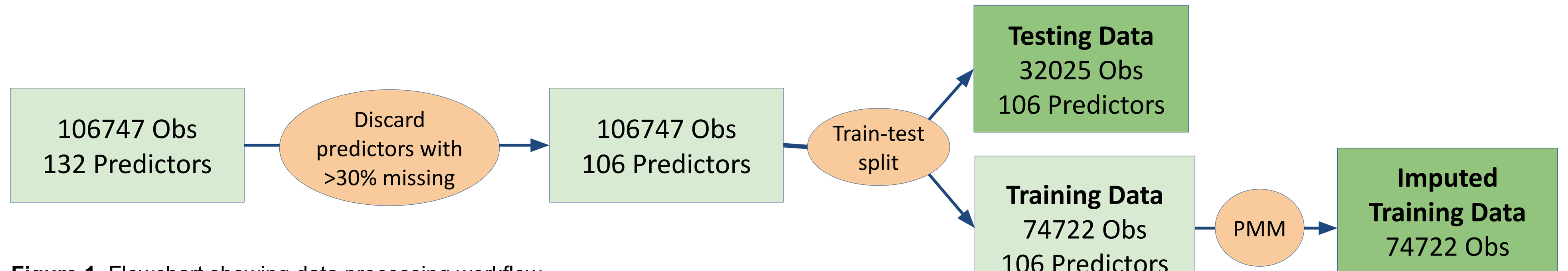

Figure 1. Flowchart showing data processing workflow.

- In predictors with ≤ 30% missing, missing values were imputed using predictive mean matching (PMM).

Model Fitting Process

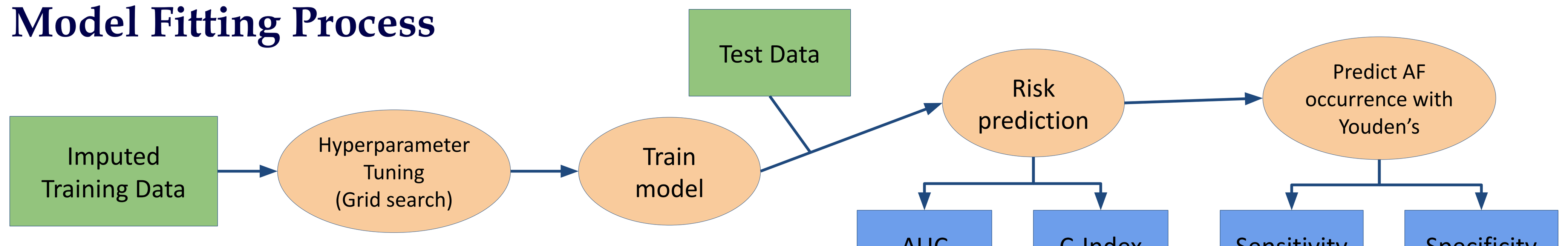

Figure 2. Flowchart showing general model fitting steps.

Random Survival Forest (RSF)

- **Core Algorithm Steps[4]:**
  - Draw 500 bootstrap samples
  - Grow a tree for each bootstrapped dataset selecting 11 variables at each split
  - Grow full tree: terminal node > 15 unique cases
  - Calculate estimators for feature vector $\mathbf{X}$ and terminal node h

$$H_h(t) = \sum_{t_{j,h} \le t} \frac{d_{j,h}}{Y_{j,h}} \qquad S_h(t) = \prod_{t_{j,h} \le t} \left(1 - \frac{d_{j,h}}{Y_{j,h}}\right)$$

In-Bag (IB) Estimator

$$H^{IB}(t|\mathbf{X}) = H_h(t)$$
$$S^{IB}(t|\mathbf{X}) = S_h(t) \text{ if } \mathbf{X} \in h$$

Out-of-Bag (OOB) Estimator

Where i is an OOB sample:

$$H^{OOB}(t|\mathbf{X}_i) = H_h(t)$$
$$S^{OOB}(t|\mathbf{X}_i) = S_h(t) \text{ if } \mathbf{X}_i \in h$$

Gradient Boosted Cox Model (GBM)

- Extends gradient boosting for right-censored survival data via Cox PH loss

$$f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m g(\mathbf{x}; \theta_m) \qquad \arg\min_f \sum_{i=1}^{n} \delta_i \left[f(\mathbf{x}_i) - \log\left(\sum_{j \in \mathcal{R}_i} \exp(f(\mathbf{x}_j)) \right) \right]$$

- **Core Algorithm Steps:**
  - Initialize: $f^{(0)}(\mathbf{x}) = 0$
  - For m = 1 to M:
    - Compute residuals $r_i = \delta_i - \dfrac{\exp\left(f^{(m-1)}(\mathbf{x}_i)\right)}{\sum_{j \in \mathcal{R}_i} \exp\left(f^{(m-1)}(\mathbf{x}_j)\right)}$
    - Fit $g(\mathbf{x}; \theta_m)$
    - Update $f^{(m)}(\mathbf{x}) = f^{(m-1)}(\mathbf{x}) + \eta \beta_m g(\mathbf{x}; \theta_m)$
  - Final model: $f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m g(\mathbf{x}; \theta_m)$

where $f(x)$: Risk score, M : Iterations, $g(x; \theta_m)$: Base learner, β_m : Step size, $\mathcal{R}_i$: Risk set, η : Learning rate

Results

Predictive Performance

Table 1. Table reporting performance statistics for prediction.

Model	AUC			C-Index		
	0.5 yr	1 yr	2 yr	0.5 yr	1 yr	2 yr
GBM	0.9461	0.9534	0.9352	0.9444	0.9540	0.9301
RSF	0.8098	0.8132	0.8431	0.8759	0.7882	0.9511

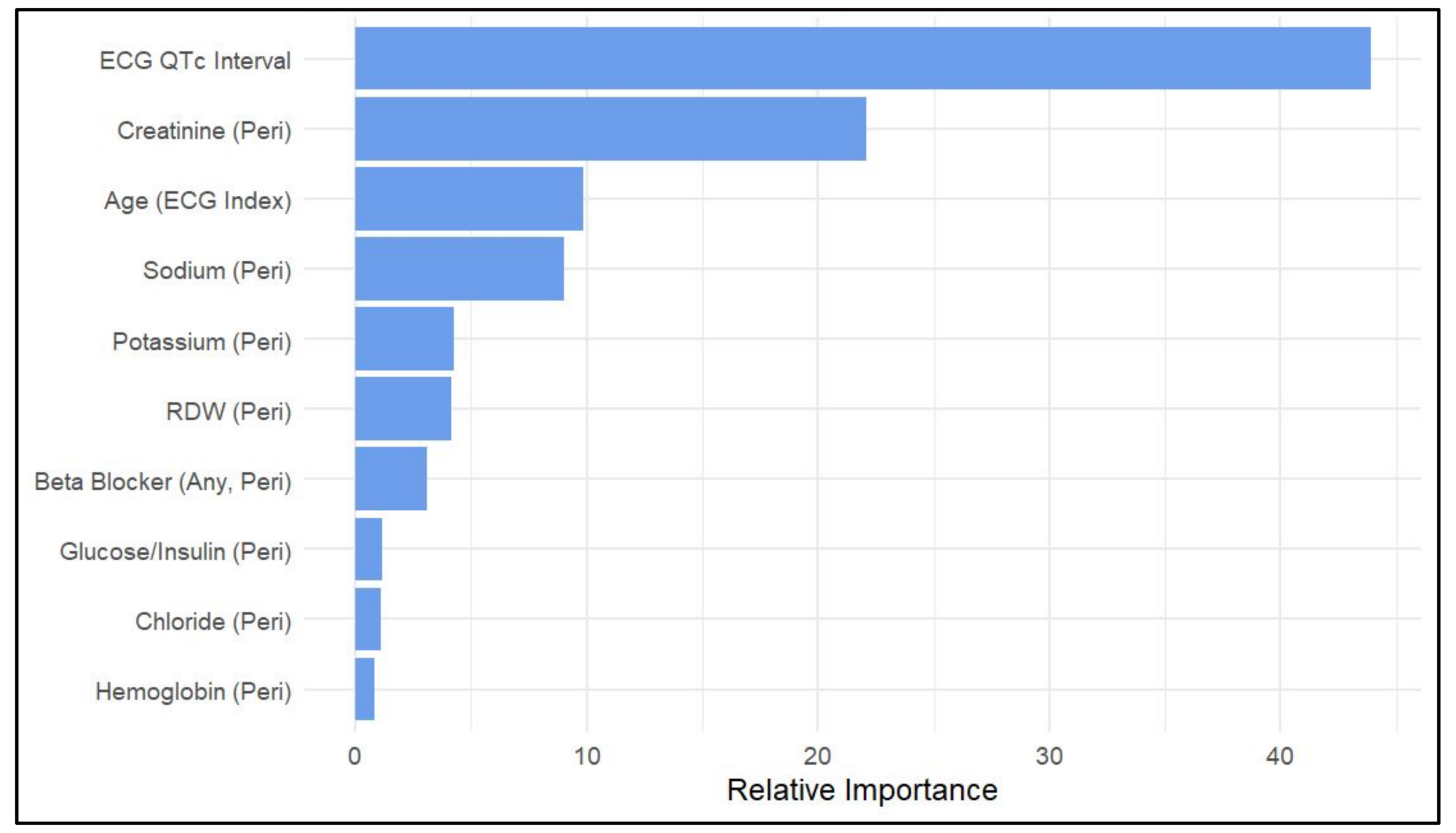

Figure 3. Graph of variable importance (GBM).

Table 2. Confusion Matrix for GBM Predicting 6-Month New-Onset Atrial Fibrillation.

Predicted \ Actual	True	False
True	353	4925
False	27	26720

Table 3. Confusion Matrix for RSF Predicting 6-Month New-Onset Atrial Fibrillation.

Predicted \ Actual	True	False
True	173	579
False	47	6258

Table 4. Table reporting sensitivity and specificity of GBM and RSF, calculated from Table 2 and Table 3.

Model	Sensitivity	Specificity
GBM	92.9%	84.4%
RSF	78.6%	91.5%

Conclusion

Our machine learning models achieved high AUC and C-Index scores and outperformed existing risk prediction models for AF.

Next steps

- Validation using real-world data for broader applicability
- Evaluate model robustness across patient subgroups

Acknowledgements

We sincerely thank Dr. Aya Mitani and Dr. Nicholas Mitsakakis for their support and guidance throughout this project.

References

1. Kornej, J., Börschel, C. S., Benjamin, E. J., & Schnabel, R. B. (2020). Epidemiology of Atrial Fibrillation in the 21st Century. *Circulation Research*, *127*(1), 4–20. https://doi.org/10.1161/circresaha.120.316340
2. Linz, D., Gawalko, M., Betz, K., Hendriks, J. M., Gregory Y.H. Lip, Nicklas Vinter, Guo, Y., & Johnsen, S. (2024). Atrial fibrillation: epidemiology, screening and digital health. *The Lancet Regional Health - Europe*, *37*, 100786–100786. https://doi.org/10.1016/j.lanepe.2023.100786
3. Nadarajah, R., Wu, J., Hogg, D., Raveendra, K., Nakao, Y. M., Nakao, K., Arbel, R., Haim, M., Zahger, D., Parry, J., Bates, C., Cowan, C., & Gale, C. P. (2023). Prediction of short-term atrial fibrillation risk using primary care electronic health records. *Heart*. https://doi.org/10.1136/heartjnl-2022-322076
4. *Random Survival Forests*. (n.d.). Www.randomforestsrc.org. https://www.randomforestsrc.org/articles/survival.html
5. van Zutphen, M., van Duijnhoven, F. J. B., Wesselink, E., Schrauwen, R. W. M., Kouwenhoven, E. A., van Halteren, H. K., de Wilt, J. H. W., Winkels, R. M., Kok, D. E., & Boshuizen, H. C. (2021). Identification of Lifestyle Behaviors Associated with Recurrence and Survival in Colorectal Cancer Patients Using Random Survival Forests. *Cancers*, *13*(10), 2442. https://doi.org/10.3390/cancers13102442
6. Dykstra, Steven et al. (2022) *Machine learning prediction of atrial fibrillation in cardiovascular patients using cardiac magnetic resonance and electronic health information*. Front. Cardiovasc. Med. 9.