Development of a Cost-Effective Predictive Model for New-Onset Atrial Fibrillation using ECG-Derived Variables

Danny Del Rosso[1], Keanna Nandlall[1], Hamda Altaf[1], Maksim Helmann[1], Mohsyn Malik[1,2], Wendy Lou[1]

[1]Dalla Lana School of Public Health, Biostatistics Division, University of Toronto, Toronto, Ontario

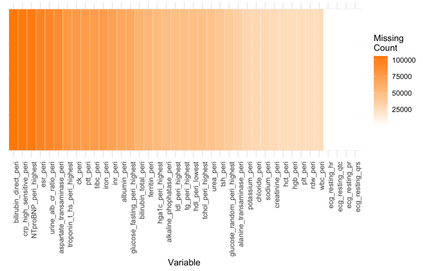
[2]Schulich School of Medicine and Dentistry, Western University, London, Ontario

Introduction

- **Background**
  - Atrial fibrillation (AF) is a common heart rhythm disorder that affects more than 46 million people globally.[1]
  - AF increases the risk of stroke four- to fivefold, with risk markedly increasing with age.[2,3]
  - Existing machine learning (ML) approaches use uncommon blood biomarkers and expensive cardiac imaging.[4]
- **Study Focus**
  - To use ML models to analyze routinely collected electrocardiogram (ECG) data, an inexpensive and evidence-based alternative.
- **Objectives**
  - Use clinical and data-driven approaches to develop a risk prediction model to predict the future occurrence of new-onset AF.
  - Evaluate the importance of aggregate ECG variables in AF prediction.

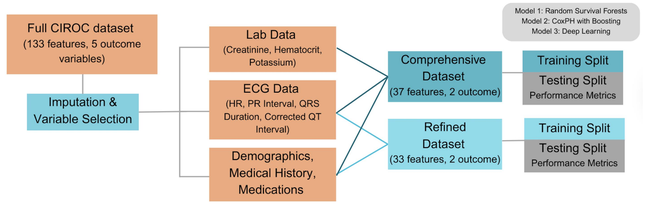
Data

- **CIROC Synthetic Dataset**
  - ~ 100,000 patients with no history of AF who had a baseline ECG performed between Jan 2010 and Jan 2023
  - Follow-up period of at least 12 months
- **Heatmap of Missingness:**

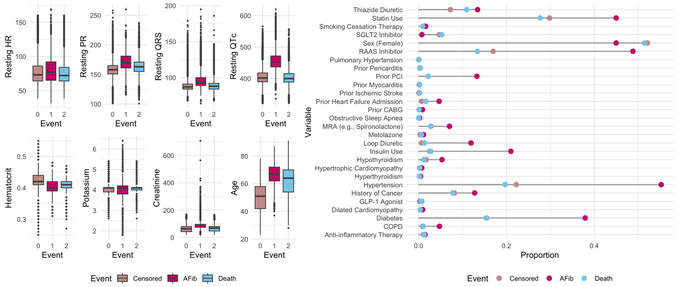


- **PTB-XL+ Publicly Available ECG Dataset[5]:**
  - 21799 ECGs from 18885 patients of 10-second length, with associated clinical data
  - 1874 patients with repeated ECG measurements free from AF at baseline

Methodology



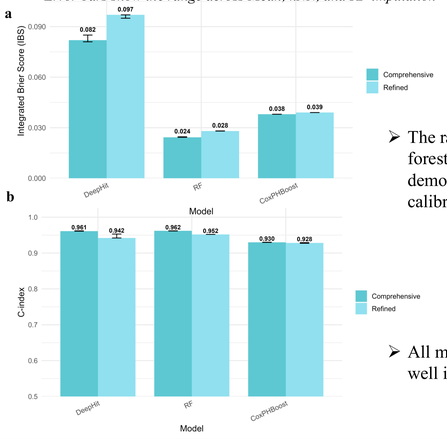
- **Variable Distribution of Clinical Measures Across Event Types**
  - Boxplots: ECG, lab, and age distributions; Dot plot: variable proportions



- **Imputation**
  - Random Forest (missranger); kNN and mean imputation for sensitivity analysis
- **Models**
  - CoxBoost (Cox Proportional Hazards with gradient boosting)
  - DeepHit (deep-learning survival model)[6]
  - Random Survival Forest (RSF)
- **Performance Metrics**
  - Harrell's C-index
  - Brier Score & Integrated Brier Score (IBS)
- **Variable Importance**
  - CoxBoost: Chooses variables which decrease the loss function the most in each boosting step
  - Random Survival Forest: Chooses variables with the greatest impact on the C-index (predictive performance)
  - DeepHit: Chooses variables that increase IBS when shuffled

Results

- **IBS (a) and C-index (b) across models for Comprehensive & Refined Data**
  - *Error bars show the range across Mean, kNN, and RF imputation*



- The random survival forest model demonstrated the best calibration.
- All models performed well in discrimination.

- **Top 5 Important Variables**

	DeepHit	Random Forest	CoxBoost
1	Corrected QT Interval	Corrected QT Interval	Corrected QT Interval
2	Age	Age	Age
3	Creatinine	Creatinine	Creatinine
4	Resting HR	Hematocrit	
5	Potassium	Potassium	

- **Results from logistic regression model predicting future AF in PTB-XL Dataset**
- $logit\big(Pr(AF = 1)\big) \sim Age + PR + HR + QRS + QTc + T$

	Coefficient	P-Value
Age (years)	0.00632	**0.0232**
Resting PR Interval (ms)	0.00580	**0.0033**
Resting Heart Rate	0.00843	**<0.0001**
QRS Duration (ms)	0.00689	0.2388
Corrected QT Interval (ms)	0.00574	0.2108
T Wave Duration	0.00851	**0.0232**

Conclusion

- **Summary of findings**
  - All models achieved strong performance (C-index > 0.92)
  - Age and corrected QT interval were the top predictors.
  - RSF performed best (C-index: 0.962, IBS: 0.024), followed by DeepHit and CoxBoost.
  - Removing lab values had minimal impact (C-index drop < 0.02).
- **Implications**
  - Results support low-cost, scalable AF risk prediction using routine ECGs and commonly collected clinical information.
  - These machine learning approaches show promise in early AF detection, highlighting their wide-spread potential.

Limitations

- Model performance may be inflated due to the synthetic dataset
- External validation with PTB-XL highlights the potential value of additional ECG features, like T-wave duration.

Next Steps

- Future work should expand ECG feature sets (e.g., T-wave) and validate models on real-world longitudinal data to enhance performance and generalizability.

Acknowledgements

The authors acknowledge the support and supervision of Dr. Wendy Lou throughout this case study.

References

Our references can be found at this QR code: