Retrospective machine-learning stratification of a Brugada-suspect cohort into SCN5A-mediated genetic architecture classes using ICD-10 encoded hospital trajectories

by:

Jubileejoy Zirebwa

MOD004875 – Undergraduate Project (BMS)

Biomedical Science (Hons) Dissertation

SID: 2208155
School of Life Science
Anglia Ruskin University
April 2026

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Timothy Hearn, for his guidance, patience and constructive challenge throughout this project. His support helped me develop the confidence to work independently within a complex topic, particularly in learning how to frame exploratory findings carefully, make systematic methodological decisions and treat unexpected results as part of the research process.

I am especially grateful for his guidance during the early stages of the project, when the unexpectedly large proportion of individuals without tiering information initially led me to consider whether this could be recovered through participant-level VCF investigation. I came to understand that these cases represented an unexplained group within the available data. This clarification shaped my understanding of the diagnostic odyssey that the project sought to address and helped refine the work towards a more appropriate promoter-focused and methodologically cautious analysis.

I would also like to thank him for directing me towards resources that strengthened my understanding of data science, programming and machine-learning methodology. This support was important in allowing me to engage more critically with the analytical pipeline and to choose methods that were proportionate to the study design.

I am grateful to the participants and families whose data contributed to this research. Their contribution made this work possible and gives significance to the broader aim of improving understanding within unresolved inherited cardiac disease pathways.

Finally, I would like to acknowledge the Genomics England support team for their assistance in making work within the Research Environment understandable and manageable. Their help enabled my first exposure to operating within a secure genomic research setting and supported the practical completion of this project.

ABSTRACT

Brugada syndrome is a rare inherited arrhythmia syndrome in which specialist ECG interpretation, symptoms, family history and genetic context matter more than any single hospital code. This investigation used pre-sequencing Hospital Episode Statistics trajectories from a Genomics England Trusted Research Environment cohort to ask whether hospital-coded phenotype history could stratify a Brugada-suspect cohort along a promoter-focused SCN5A architecture axis. The source cohort contained 252 participants: 139 promoter non-carriers, 90 heterozygous carriers and 23 homozygous carriers. The primary analysis compared promoter-homozygous participants with eligible non-carriers after Tier 1/2 SCN5A mutation-positive exclusions, leaving 153 participants and only 23 positive cases. A compact 5-year feature set retained 10 active predictors: five ICD-derived ever flags, three grouped burden rates and two observability covariates. Across five repeated hold-out seeds, random forest was the strongest main-branch family but remained modest (median ROC AUC 0.5429, PR AUC 0.2621, Brier score 0.1350), with severe events-per-variable limitation. A secondary raw-ICD branch suggested possible residual ranking signal but was seed-unstable. The conclusion of this work is that HES trajectories may contain weak promoter-relevant signal, but not enough for reliable classification without richer phenotyping, larger event counts and external validation.

DECLARATION

This dissertation is submitted as original work for Biomedical Science (Hons). The analysis was conducted under approved data-governance arrangements inside the Genomics England Trusted Research Environment. Small phenotype-linked counts were masked in the exported summary artefacts used for this write-up. The main literature, code-derived methods and TRE export findings are acknowledged through in-text citation, tables, figures and the reference list.

LIST OF ABBREVIATIONS

Abbreviations used in the dissertation.

Abbreviation	Meaning
AE	Accident and Emergency
APC	Admitted Patient Care
AUPRC / PR AUC	Area under the precision-recall curve
AUROC / ROC AUC	Area under the receiver operating characteristic curve
BrS	Brugada syndrome
CV	Cross-validation
ECG	Electrocardiogram
EHR	Electronic health record
EPV	Events per variable
GE	Genomics England
HES	Hospital Episode Statistics
ICD-10	International Classification of Diseases, Tenth Revision
LR / RF / NN	Logistic regression / random forest / neural network
OP	Outpatient
SE	Standard Error
TRE	Trusted Research Environment

TABLE OF CONTENTS

REFERENCES

Althoff, K.N. et al. (2019) 'Mind the gap: observation windows to define periods of event ascertainment as a quality control method for longitudinal electronic health record data', Annals of Epidemiology, 33, pp. 54-63. doi:10.1016/j.annepidem.2019.01.015.

Ambroise, C. and McLachlan, G.J. (2002) 'Selection bias in gene extraction on the basis of microarray gene-expression data', Proceedings of the National Academy of Sciences, 99(10), pp. 6562-6566. doi:10.1073/pnas.102102699.

Bastarache, L. et al. (2018) 'Phenotype risk scores identify patients with unrecognized Mendelian disease patterns', Science, 359(6381), pp. 1233-1239. doi:10.1126/science.aal4043.

Bastarache, L. et al. (2019) 'Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease', Journal of the American Medical Informatics Association, 26(12), pp. 1437-1447. doi:10.1093/jamia/ocz179.

Cawley, G.C. and Talbot, N.L.C. (2010) 'On over-fitting in model selection and subsequent selection bias in performance evaluation', Journal of Machine Learning Research, 11, pp. 2079-2107.

Collins, G.S. et al. (2024) 'TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods', BMJ, 385, e078378. doi:10.1136/bmj-2023-078378.

Denaxas, S. et al. (2019) 'UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER', Journal of the American Medical Informatics Association, 26(12), pp. 1545-1559. doi:10.1093/jamia/ocz105.

Denham, N.C. et al. (2019) 'Systematic re-evaluation of SCN5A variants associated with Brugada syndrome', Journal of Cardiovascular Electrophysiology, 30(1), pp. 118-127. doi:10.1111/jce.13740.

Genomics England (2026a) Clinical and phenotype data. Available at: https://re-docs.genomicsengland.co.uk/clinical_landing/ (Accessed: 1 May 2026).

Genomics England (2026b) Labkey API. Available at: https://re-docs.genomicsengland.co.uk/labkey_api/ (Accessed: 1 May 2026).

Genomics England (2026c) Rare disease tiering. Available at: https://re-docs.genomicsengland.co.uk/tiering/ (Accessed: 1 May 2026).

Hersh, W.R. et al. (2022) 'Clinical study applying machine learning to detect a rare disease: results and lessons learned', JAMIA Open, 5(2), ooac053. doi:10.1093/jamiaopen/ooac053.

Huang, Y. et al. (2020) 'A tutorial on calibration measurements and calibration models for clinical prediction models', Journal of the American Medical Informatics Association, 27(4), pp. 621-633. doi:10.1093/jamia/ocz228.

Kawada, S. et al. (2018) 'Shanghai Score System for Diagnosis of Brugada Syndrome: Validation of the Score System and Reclassification of the Patients', JACC: Clinical Electrophysiology, 4(6), pp. 724-730. doi:10.1016/j.jacep.2018.02.009.

Matsumura, H., Nakano, Y., Oda, N. et al. (2017) ‘H558R, a common SCN5A polymorphism, modifies the clinical phenotype of Brugada syndrome by modulating DNA methylation of SCN5A promoters’, Journal of Biomedical Science, 24(1), p. 91. doi: 10.1186/s12929-017-0397-x.

NHS England (2026) Hospital Episode Statistics Data Dictionary. Available at: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary (Accessed: 24 April 2026).

Nissen, F., Quint, J.K., Morales, D.R. and Douglas, I.J. (2019) 'How to validate a diagnosis recorded in electronic health records', Breathe, 15(1), pp. 64-68. doi:10.1183/20734735.0344-2018.

Park, J.K. et al. (2012) 'Genetic variants in SCN5A promoter are associated with arrhythmia phenotype severity in patients with heterozygous loss-of-function mutation', Heart Rhythm, 9(7), pp. 1090-1096. doi:10.1016/j.hrthm.2012.02.023.

Pendergrass, S.A. and Crawford, D.C. (2019) 'Using electronic health records to generate phenotypes for research', Current Protocols in Human Genetics, 100(1), e80. doi:10.1002/cphg.80.

Postema, P.G. et al. (2023) 'Use, misuse, and pitfalls of the drug challenge test in the diagnosis of the Brugada syndrome', European Heart Journal, 44(27), pp. 2427-2437. doi:10.1093/eurheartj/ehad295.

Priori, S.G. et al. (2013) 'HRS/EHRA/APHRS expert consensus statement on the diagnosis and management of patients with inherited primary arrhythmia syndromes', Heart Rhythm, 10(12), pp. 1932-1963. doi:10.1016/j.hrthm.2013.05.014.

Riley, R.D. et al. (2024a) 'Evaluation of clinical prediction models (part 1): from development to external validation', BMJ, 384, e074819. doi:10.1136/bmj-2023-074819.

Riley, R.D. et al. (2024b) 'Evaluation of clinical prediction models (part 2): how to undertake an external validation study', BMJ, 384, e074820. doi:10.1136/bmj-2023-074820.

Saito, T. and Rehmsmeier, M. (2015) 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', PLoS ONE, 10(3), e0118432. doi:10.1371/journal.pone.0118432.

Van Calster, B. et al. (2019) 'Calibration: the Achilles heel of predictive analytics', BMC Medicine, 17, 230. doi:10.1186/s12916-019-1466-7.

Walsh, C.G., Sharman, K. and Hripcsak, G. (2017) 'Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk', Journal of Biomedical Informatics, 76, pp. 9-18. doi:10.1016/j.jbi.2017.10.008.

Yagihara, N. et al. (2016) 'Variants in the SCN5A Promoter Associated With Various Arrhythmia Phenotypes', Journal of the American Heart Association, 5(9), e003644. doi:10.1161/JAHA.116.003644.

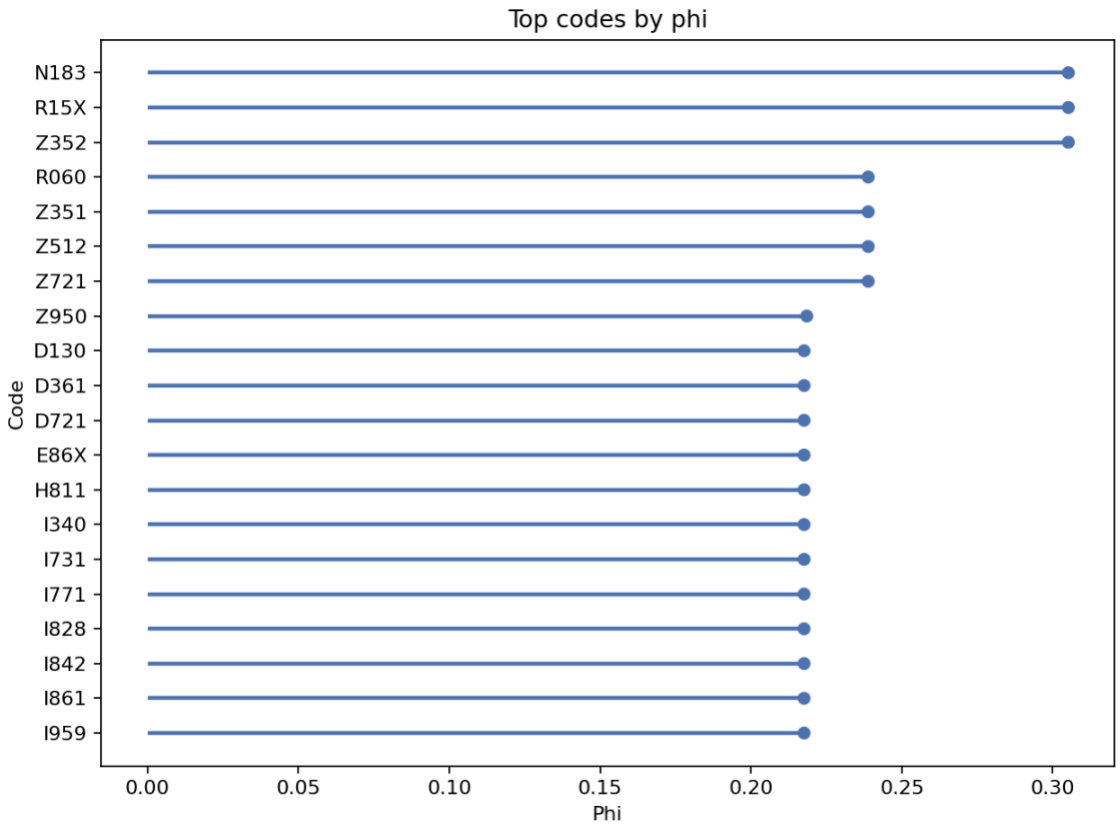
Zeppenfeld, K. et al. (2022) '2022 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death', European Heart Journal, 43(40), pp. 3997-4126. doi:10.1093/eurheartj/ehac262.

APPENDIX A: ACTIVE FEATURE AND ICD CODE LOGIC

This appendix records the active compact-primary code logic in words. ICD groups were handled as structured hospital-history features designed to capture arrhythmia, conduction, severe-event and care-pathway traces before sequencing.

Appendix Table A1. ICD code groups used to build active features.

Group	ICD-10 code logic	Meaning
Ventricular tachyarrhythmia	I47.0, I47.2	Ventricular tachycardia-related hospital coding.
Ventricular fibrillation/flutter	I49.0	High-severity ventricular electrical instability.
Cardiac arrest history	I46*	Hospital-coded cardiac arrest history.
Syncope/unspecified conduction alert	R55 or I45.9	R55 captured syncope/collapse; I45.9 captured unspecified conduction-disorder alerting.
AV block	I44*	Atrioventricular block/conduction delay.
Bundle/fascicular/intraventricular block	I45*	Intraventricular conduction-system abnormality.
Sinus-node dysfunction	I49.5	Sick sinus syndrome/sinus-node dysfunction.
Bradycardia unspecified	R00.1	Generic bradycardia; removed as a standalone compact-primary ever flag.
Pathway proxy	I49.8, R94.3, Z95.0, Z45.0	Specified arrhythmia/channelopathy, abnormal cardiac tests, device presence/follow-up.



Appendix Figure A1. *Seed 62 top raw-ICD phi signals from the exploratory branch. All 20 captured codes by phi were positive.*

APPENDIX B: NUMERICAL HANDLING AND FORMULAE

Balanced accuracy was used to reduce the misleading effect of the majority class. It was calculated as:

$$balanced\ accuracy\ =\ (sensitivity\ +\ specificity)\ /\ 2$$

where sensitivity was TP/(TP+FN) and specificity was TN/(TN+FP). ROC AUC measured ranking performance across classification thresholds. PR AUC measured precision-recall behaviour and was prioritised because the positive class was small. The Brier score was the mean squared difference between predicted probability and observed binary outcome:

$$Brier\ score\ =\ mean((predicted\ probability\ -\ observed\ outcome)\text{^}2)$$

For the final compact branch, the crude events-per-variable calculation was 17 positive training events divided by 14 post-preprocessing feature-effect rows, giving approximately 1.21. This was a severe limitation on model interpretation.

For the raw-ICD branch, the support score was defined as:

$$support\ score\ =\ log10((participant\ support\ +\ offset)\ x\ (event-volume\ proxy\ +\ offset))$$

Support thresholds were selected inside training data using inner cross-validation. Cross-validation means that training data are repeatedly split into inner training and validation folds so that tuning decisions can be compared without using the held-out test set. SE means standard error. In the support-threshold plot, mean +/- SE shows the estimated mean inner-CV balanced accuracy and its uncertainty across inner folds.

The phi coefficient was used to describe the association between each retained binary ICD code and the target class. Positive phi indicated enrichment in the promoter-homozygous group, while negative phi indicated enrichment in non-carriers. Jaccard similarity was used to measure code co-occurrence:

$$Jaccard\ similarity\ =\ shared\ participants\ with\ both\ codes\ /\ participants\ with\ either\ code$$

Codes were grouped by ICD chapter, phi sign and co-occurrence similarity. This was intended to convert sparse raw ICD codes into more stable group-level features, although the results showed that many groups were still singletons.

APPENDIX C: EXTENDED AUDIT TABLES

Appendix Table C1. Raw-ICD seed-level audit.

Model	Seed	Groups	Codes	ROC AUC	PR AUC	Balanced accuracy
LR	42	312	378	0.394	0.159	0.538
LR	52	97	121	0.492	0.166	0.470
LR	62	522	636	0.596	0.207	0.561
LR	72	94	111	0.338	0.146	0.530
LR	82	126	150	0.295	0.122	0.470
NN	42	312	378	0.369	0.140	0.455
NN	52	97	121	0.604	0.286	0.583
NN	62	522	636	0.581	0.205	0.500
NN	72	94	111	0.530	0.197	0.515
NN	82	126	150	0.381	0.170	0.500
RF	42	312	378	0.295	0.120	0.356
RF	52	97	121	0.548	0.179	0.500
RF	62	522	636	0.535	0.243	0.500
RF	72	94	111	0.641	0.226	0.492
RF	82	126	150	0.563	0.249	0.500

APPENDIX D: REPRODUCIBILITY SUMMARY

Appendix Table D1. Summary of reproducibility details

Component	Implementation detail
Secure data environment	Analyses were conducted within the Genomics England Trusted Research Environment. Clinical and phenotype data were accessed through LabKey, containing de-identified clinical, phenotypic and bioinformatics data, including secondary clinical data such as Hospital Episode Statistics (Genomics England, 2026a; Genomics England, 2026b).
Data release and access route	The cohort-building code used the Genomics England Main Programme LabKey. LabKey API supports Python-based querying of LabKey tables and the cohort code used SQL queries returned as pandas data frames (Genomics England, 2026b).
Software	The active analysis scripts were written in Python and used pandas, NumPy, SciPy, scikit-learn and Matplotlib.
Source tables	Hospital Episode Statistics Admitted Patient Care, Outpatient and Accident and Emergency sources were used as pre-index hospital-trajectory data. Diagnosis tokens were extracted from diag_all and available diag_XX fields where present, normalised by removing punctuation and mapped into ICD-derived groups.
Temporal censoring	Predictor features were derived from hospital events occurring before the sequencing/index date. The compact-primary branch used a five-year pre-index window.
Primary target	The main target was promoter_homozygous_vs_noncarrier, derived from promoter dosage. Promoter homozygotes were mapped to class 1, promoter non-carriers to class 0 and promoter heterozygotes were dropped from the primary contrast.
Tier 1/2 exclusion	The cohort filter was exclude_tier12_positive. In code, Tier 1/2 signal was resolved from tier12_positive if available and rows with Tier 1/2 signal equal to 1 were excluded. Genomics England describes rare-disease Tier 1 and Tier 2 variants as variants that should be clinically assessed by NHS Genomic Medicine Centres and that are assigned through the rare-disease tiering process (Genomics England, 2026c).
Final primary modelling sample	The final main-branch modelling set contained 153 participants: 23 promoter-homozygous positives and 130 eligible non-carrier negatives. Each repeated split contained 114 training participants and 39 held-out participants, with 17 positives in training and 6 positives in the held-out set.
Compact-primary features	The compact-primary profile retained 10 predictors: five ICD ever flags, three grouped burden rate features and two observability controls. Recency features, generic event/code intensity features, missingness flags, raw grouped ICD counts and broader raw burden-count families were excluded from the compact-primary profile.
Active ICD ever flags	The five retained binary ICD features were ventricular tachyarrhythmia ever, ventricular fibrillation/flutter ever, cardiac arrest ever, syncope/collapse ever and bundle/fascicular/intraventricular block ever within the five-year window.
Active burden-rate features	The three retained rate features were ventricular-severity burden rate, conduction-system burden rate and pathway-proxy burden rate, normalised by observable exposure time inside the five-year window.
Observability controls	exposure_years_core and short_history_indicator_5y were retained to distinguish true absence of recorded events from limited observable hospital history.
Split design	Both compact-primary and raw-ICD signalling branches used participant-level repeated stratified hold-out evaluation with seeds 42, 52, 62, 72 and 82 and a 0.25 test fraction. Splits were stratified where valid.
Preprocessing sequence	For each seed, the split was made before model fitting. Numeric imputation, categorical handling, log transformation, winsorisation, scaling and threshold selection were fitted using training data only and then applied to the held-out data.

Imputation and categorical handling	Categorical unseen levels were mapped to Unknown. Numeric imputation values were learned from the training partition. Binary features were protected from log transformation, winsorisation and scaling.
Log, winsor and scaling policy	Log transformation was applied to eligible count, duration and rate features for LR and NN only. Winsorisation used the training-side upper 0.995 quantile for eligible features, post-log for LR/NN and raw for RF. Scaling was applied to eligible continuous/ordinal/count/duration/rate features for LR and NN only.
Main LR model	Logistic regression used scikit-learn LogisticRegression with solver="liblinear", L2 regularisation by default, max_iter=2000, seed-specific random_state and no class weighting in the primary analysis.
Main RF model	Random forest used scikit-learn RandomForestClassifier with n_estimators=500, default depth, seed-specific random_state, n_jobs=-1 and no class weighting in the primary analysis.
Main NN model	The compact-primary NN comparator used scikit-learn MLPClassifier with hidden layers (64, 32), ReLU activation, Adam optimiser, alpha=0.0001, initial learning rate 0.001, max_iter=500, seed-specific random_state and early stopping disabled. It remained an exploratory stress-test comparator.
Threshold policy	Thresholds for binary classification used training-only out-of-fold probabilities. The default policy selected the threshold that maximised training balanced accuracy subject to a minimum training specificity floor of 0.55. If inner-CV support was insufficient, fixed-threshold fallback logic was available and logged; the compact-primary runs did not use threshold fallback.
Inner CV for thresholds	Compact-primary threshold selection used training-only inner cross-validation with up to five folds, reduced when the minority-class count required fewer folds.
Raw-ICD branch role	The raw-ICD branch was a secondary exploratory analysis asking whether unconstrained ICD-code space contained residual signal missed by the compact expert-grouped feature set.
Raw-ICD support score	For each ICD code c in the training split, support was s_sup(c) = log10((n_pt(c) + 0.5) * (n_evt(c) + 0.5)), where n_pt(c) was distinct participant support and n_evt(c) was the event-volume term.
Raw-ICD support-threshold tuning	Candidate support thresholds were empirical support-score quantiles from 0.10 to 0.85 in increments of 0.05. Three-fold inner CV was used where feasible. The selected threshold was the most strongly pruning candidate within one standard error of the best mean inner-CV balanced accuracy.
Raw-ICD phi and grouping	For each retained code, a target-skew phi coefficient was computed using training data only. If any 2x2 contingency-table cell was zero, 0.5 was added to all cells. Codes were grouped within ICD chapter and phi-sign partitions using Jaccard co-occurrence weighted by sqrt(abs(phi_i) * abs(phi_j)), average-linkage hierarchical clustering and an adaptive largest-gap cut.
Raw-ICD models	Raw-ICD LR used L2 logistic regression with liblinear and max_iter=2000; RF used 500 trees; NN used hidden layers (32, 16), alpha=0.001, learning rate 0.0005, max_iter=1200 and early stopping. Class weighting was not used.
Reporting convention	Aggregate performance was reported as median (IQR) across the five repeated hold-out seeds. ROC AUC, PR AUC and Brier score were used as probability/ranking summaries; sensitivity, specificity and balanced accuracy depended on the training-selected threshold and were interpreted cautiously because each test split had only six positives.