Analysis of Genetic Risk Scores, Exercise Testing Data and Clinical Data for Prediction of Sudden Cardiac Death in the Fincavas Cohort

Jussi Hernesniemi[1,2], Leo-Pekka Lyytikäinen[1,2], Teemu Pukkila[1], Esa Räsänen[1], Mika Kähönen[1,3]
Terho Lehtimäki[1]

[1]Tampere University, Tampere, Finland
[2]Tampere Heart Hospital, Tampere, Finland
[3]Tampere University Hospital, Tampere, Finland

Abstract

We explored the overlap and the independent contribution of multiple data modalities in prediction of sudden cardiac death (SCD) analysing a prospective clinical exercise testing cohort (FINCAVAS; n=2861 with SCD events (n=75). Contributions of different data domains were evaluated using a multimodal modelling framework by fitting Cox proportional hazards models to time-to-event data. Preselected domains included demographic, physiological (and anthropometric), dynamic exercise, laboratory, medication, genetic (polygenic risk scores based on published genome-wide association studies), and electronic health record data.

Cross-validated are under curve for the full model was 0.79. Most predictive information was captured by dynamic exercise, demographic, and genetic data, accounting for 29%, 28%, and 17% of total contribution. Medication data and physiological measures contributed moderately (12% and 11%), whereas laboratory variables added only 3.6%. Overlap between domains was generally low (−0.001 to 0.018).

1. Introduction

Sudden cardiac death (SCD) is hard to predict with over 80% of cases occurring in a population with presumably low risk[1]. Use of traditional risk factor data even high-risk individuals has not led to any meaningful advances in risk stratification[2].

Besides traditional clinical risk factors and electrocardiographic data, there are other features domains that have demonstrated several individual features associated with the risk of SCD or similar events[3]. These include demographic, physiological (and anthropometric), dynamic exercise, laboratory, medication, genetic, and electronic health record data domains[4].

The possible overlap and independent contribution of these different domains remain unclear. We used a prospectively collected dataset of 2861 patients undergoing clinical exercise testing (The Fincavas dataset) to address these questions.

2. Methods

2.1. Population

The patient population of FINCAVAS comprises all willing consecutive patients between October 2001 and December 2008 who were referred for an exercise test at Tampere University Hospital (n=3776 patients with technically successful exercise test). Some patients performed multiple exercise tests but only the first technically successful exercise test for each patient was included. A technically successful exercise test was defined as a test with reliably measured exercise, heart rate, blood pressure and concomitant continuous ECG data. Only patients without pacemakers or ICDs performing the exercise test on a cycle ergometer and reaching maximal or almost maximal level of exercise were included. Indications for the test were suspicion of CAD (43%), evaluation of work capacity (24%), arrhythmias (25%), adequacy of CAD drug therapy (13%) and evaluation prior to an invasive procedure (9%) or after myocardial infarction (MI) (7.6%). Some patients had more than one indication. Patients with pacemakers, those without fully available metadata on pre-existing significant comorbidities, as well as patients without one minute of resting heart rate (ECG) data or genetic data were excluded. After exclusions data for 2,861 patients was available. The Ethics Committee of Tampere University Hospital District approved the study protocol, and all patients gave written informed consent

2.2. Follow-up data collection

The primary endpoint of this study was SCD. Death certificates listing the circumstances of the deaths and specific causes of death were received from the Causes of Death Register maintained by Statistics Finland and reviewed for adjudication of events (autopsy rate of 46%). SCD was defined by the ESC definition as sudden and unexpected death by a primary cardiac cause (without any evidence or other probable cause) and occurring within approximately 1 hour from the onset of symptoms or if the death was not witnessed if the person had last been seen symptomless within 24 hours of the death. Follow-up data for SCDs was available until 31st of December 2013. There was no loss-to-follow-up.

2.3. Exposure variables and their classification to different domains

Six distinct domains were selected to represent different data modalities: demographic (e.g. age, sex, prevalent conditions/diseases evaluated at baseline by a physician, prevalence of left ventricular hypertrophy), physiological (and anthropometric such as height, with, BMI, resting blood pressure and heart rate, electrocardiographic features), dynamic exercise, laboratory (e.g. circulating serum and plasma levels for sodium, potassium, creatinine, haemoglobin), medication (e.g. use of beta-blockers, angiotensin converting enzyme inhibitors or angiotensin receptor blockers, calcium channel blockers, digitalis, nitrates, oral diabetes), genetic (polygenic risk scores based on published genome-wide association studies), and electronic health record data in ICD-10 coded structure (several conditions).

2.4. Statistical analysis

First, all variables in each feature domain were tested for their association with SCDs occurring during follow-up. Variables with a p-value of more than 0.2 were excluded from further analyses.

Multimodal predictive performance was then evaluated by fitting all possible non-empty combinations of the retained domains. Proportional hazards regression models were fitted for the time-to-event endpoint and cross-validated discrimination values (five-fold with three repeats) were calculated. Cox models were summarized using concordance index (C-index), and a likelihood-based pseudo-R^2 derived from the Cox model likelihood ratio statistic.

To quantify the relative contribution of each domain to model performance, Shapley decomposition was applied across all possible domain combinations using the model R^2 metric as the value function. This yielded the average marginal contribution of each domain to total explained information across all possible modeling contexts. In addition, domain importance was examined by drop-one analyses, in which the loss in model R^2 after removing each domain from the full multimodal model was quantified. Pairwise overlap indices were also calculated to evaluate the degree of shared versus complementary information between domains.

To improve interpretability and to address the possibility that domains with a larger number of predictors could appear more important simply because of greater model flexibility, we further performed domain-wise variable prioritization. Within each domain, forward selection using Akaike's information criterion (AIC) was applied separately in logistic and Cox models, and a capped sensitivity analysis was then conducted using at most five selected variables per domain. The full multimodal analyses were repeated using these domain-capped predictor sets to evaluate whether domain importance rankings were robust to differences in the number of candidate variables per domain.

3. Results

The mean age of population at baseline was 55.4 years (SD 12.8). The proportion of women was 40.3% (n=1152). The median follow-up time was 8.6 years (6.7-10.6) during which 75 patients suffered SCD.

2.3. Overall model performance

The full multimodal Cox model combining all six data domains had a pseudo-R^2 of 0.080. Highest cross-validated discrimination was approximately 0.797 as assessed by the concordance index. The discrimination of each domain are presented in Table 1.

Table 1. Cross-validated discrimination and of each domain for SCD events.

Domain	C-index	Overall Contribution
Demographics	0.741	27.6%
Physiological Features	0.701	10.5%
Laboratory Values	0.600	3.6%
Dynamic Exercise Data	0.752	28.9%
Medication data	0.725	11.9%
Genetic data	0.671	17.6%

Shapley decomposition demonstrated marked differences in the relative contribution of the individual data domains. Dynamic exercise data provided the largest contribution, accounting for 28.9% of the total model

information, followed closely by demographic and clinical data (27.6%). Genetic data accounted for a further 17.6%, whereas medication and physiological data contributed 11.9% and 10.5%, respectively. Laboratory variables provided the smallest contribution, accounting for 3.6% of the total. The drop-one analysis showed a broadly similar hierarchy. Removal of the exercise domain resulted in the largest deterioration in model fit, reducing the full-model pseudo-R^2 from 0.0800 to 0.0638 (ΔR^2=0.0162). Removal of demographic data reduced R^2 to 0.0653 (ΔR^2=0.0147), whereas removal of genetic information reduced R^2 to 0.0673 (ΔR^2=0.0127). In contrast, removal of physiological, medication, or laboratory data resulted in substantially smaller reductions in model fit (ΔR^2=0.0049, 0.0038, and 0.0026, respectively) (Figure 1).

Pairwise overlap between domains was generally small, ranging from −0.0008 to 0.0177 R^2 units. The greatest shared predictive information was observed between demographic and exercise data (overlap 0.0177), followed by exercise and medication data (0.0151) and demographic and medication data (0.0150). Genetic data showed consistently little overlap with the other modalities. Overlap with demographic, physiological, laboratory, exercise, and medication data was 0.0039, 0.0037, 0.0004, 0.0025, and 0.0047 R^2 units, respectively, suggesting that genetic information captured predictive information that was largely distinct from that available from the other domains.

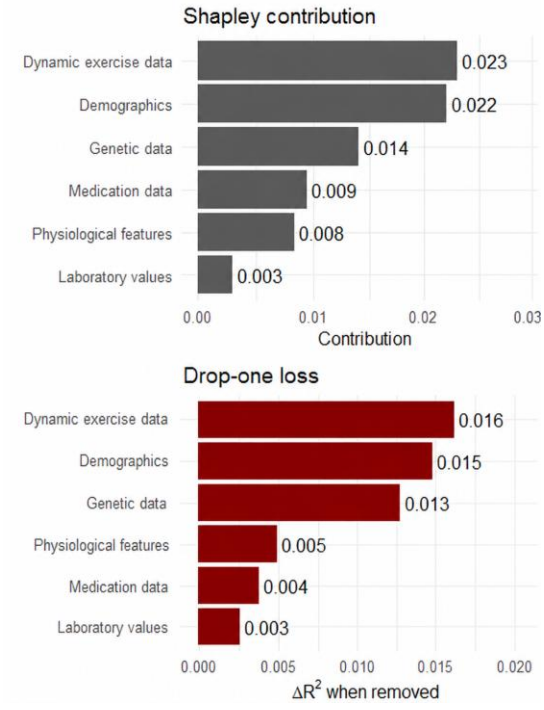

Figure 1. The individual Shapley contributions and the effect of drop-one loss of each domain explaining SCDs occurring during follow-up and the

Pairwise overlap between domains was generally small, ranging from −0.0008 to 0.0177 R^2 units. The greatest shared predictive information was observed between demographic and exercise data (overlap 0.0177), followed by exercise and medication data (0.0151) and demographic and medication data (0.0150). Genetic data showed consistently little overlap with the other modalities. Overlap with demographic, physiological, laboratory, exercise, and medication data was 0.0039, 0.0037, 0.0004, 0.0025, and 0.0047 R^2 units, respectively, suggesting that genetic information captured predictive information that was largely distinct from that available from the other domains.

Restricting each domain to its five strongest predictors yielded a similar distribution of predictive contributions. Demographic, exercise, and genetic data remained the dominant domains, accounting for 27.4%, 26.0%, and 17.4% of the total contribution, respectively. Medication, physiological, and laboratory data accounted for 14.0%, 11.6%, and 3.6%. The reduced models retained approximately 80% of the full model pseudo-R^2 (0.064 vs. 0.080), indicating that the principal findings were not driven by differences in the number of variables across domains. Adding left ventricular ejection fraction (available in 71%) to the analyses resulted in modest increase in the proportion of total model information the physiological measurement domain had (from 11.9% to 13.2%).

4. Discussion

The results of this study show that when using multimodal data for the prediction of SCD, the dynamic features recorded during maximal or near maximal exercise testing seem to have the strongest proportional influence on the model. Furthermore, although genetic information comprised in polygenetic risk scores have a more modest overall predictive value, the genetic data domains stand out as a very strong individual component with very little overlap with other domains. Overall, these data domains still explain a very modest proportion of the SCD risk with pseudo-R^2 value reaching only about 8%.

There are several limitations that need to be acknowledged. First, although many clinically relevant ECG features were included in this analysis (in the physiological measurement domain) such as the presence of left bundle branch block, right bundle branch block, left ventricular hypertrophy and QRS duration, the standard 12-channel electrocardiogram holds a lot more information if used as a waveform feature[5].

Second, we did not have imaging data available to

incorporate to our analysis.

Interestingly, we did have data of measured left ventricular ejection fraction, which is considered the foremost clinical discriminator when considering primary prevention of SCD by an implantable cardioverter-defibrillator device. Adding this data improved the discriminatory value of the physiological measurement domain only modestly – a finding that is on par with the results of the larger PROFID consortium[2].

It is likely that using other machine learning algorithms, such as extreme gradient boosting or random forest classifier instead of simple Cox-regression analysis could have resulted in higher overall performance of the prediction model.[6]. However, the foremost purpose of this work was to evaluate the relations between different domains rather than achieve the highest possible model performance.

References

[1] H. V Huikuri, A. Castellanos, and R. J. Myerburg, "Sudden death due to cardiac arrhythmias," *N. Engl. J. Med.*, vol. 345, no. 20, pp. 1473–1482, Sep. 2001, doi: 10.1056/NEJMRA000650.

[2] N. Peek *et al.*, "Sudden cardiac death after myocardial infarction: individual participant data from pooled cohorts," *Eur. Heart J.*, vol. 45, no. 43, pp. 4616–4626, Sep. 2024, doi: 10.1093/EURHEARTJ/EHAE326.

[3] H. N. Pham *et al.*, "Dynamic electrocardiogram changes are a novel risk marker for sudden cardiac death," *Eur. Heart J.*, vol. 45, no. 10, pp. 809–819, Sep. 2024, doi: 10.1093/EURHEARTJ/EHAD770.

[4] R. K. Sandhu *et al.*, "Polygenic score for coronary disease predicts sudden death in coronary disease without severe systolic dysfunction," *J. Am. Coll. Cardiol.*, 2022.

[5] Z. Obermeyer, A. Schubert, J. Ross, S. Mullainathan, and M. Lingman, "An ECG biomarker for sudden cardiac death discovered with deep learning," *Nature*, vol. 655, no. 8121, Jul. 2026, doi: 10.1038/S41586-026-10674-6.

[6] J. A. Hernesniemi *et al.*, "Performance of the 12-lead ECG in predicting short- and long-term risk of sudden cardiac death," *NPJ Digit. Med.*, Sep. 2026, doi: 10.1038/S41746-026-02456-1.

Address for correspondence:

Jussi Hernesniemi
Elämänaukio 1, 33520 Tampere, Finland
jussi.hernesniemi@tuni.fi