Machine & Deep Learning from Longitudinal EHRs and Genetic Data for Low CVD Risk Prediction

Matthew Hancock[1,2]

Mentor: Wei-Qi Wei, MD, PhD[2]

[1]Vanderbilt University, Department of Mathematics

[2]Vanderbilt University Medical Center, Department of Biomedical Informatics

OBJECTIVE:

Our study sought to use machine & deep learning to leverage recent advances in high-throughput genotyping & phenotyping to learn from longitudinal EHR, genomic, and expression data to predict CVD outcome in the conventionally low-risk population.

ACKNOWLEDGEMENT: This work was funded by NSF Award 1757644

Thanks for the help from Juan Zhao PhD, Qiping Feng PhD, Josh Denny MD, Patrick Wu BS, and Janey Wang MS

Introduction

Cardiovascular disease (CVD) is the leading cause of morbidity/mortality globally, responsible for 1/3 of deaths.

The **Framingham Risk Score** (FRS) is the gold standard for 10-year CVD prediction, but has key limitations:

- Predicting a complex disease using only a **small number of risk factors**
- Attempting to analyze disease outcome over time using **cross-sectional data**
- Placing the **highest correlation coefficient on age**

15-20% of first CVD-event patients had one or zero conventional CVD risk factors.

Study Design

Individuals for the study were selected from the VUMC Synthetic Derivative (SD), an anonymized copy of the whole VUMC EHR (>3 million unique individuals).

Study cohort
Age: 18 – 89
Case: ≥ 1 CVD diagnosis code(s) ; Control: no CVD code
1:8 case-control match (age, sex, race, and T_0 [first diagnosis of CVD])
EHR data within 5-year window prior to T_0 (6-month prior to diagnosis to avoid confounding issue)
FRSs at T_0 to determinate "low-risk" or "high-risk" for CVD

Features	
Demographic data	age, gender, race and smoking status
Disease phenotypes	457 phecodes
Drugs by ingredients	866 RxNorm concept codes
Physical measurement	e.g. blood pressure and body mass index (BMI)
Laboratory tests	214 tests
Genetic variants	single nucleotide polymorphisms (SNPs) from catalog
Expression data	6,697 gene expression data

Longitudinal physical measurements and laboratory tests were represented by summarized data (i.e. max, min, median, mean, and counts). Missing data was median-imputed from individuals of the same age, gender, and race. Dummy variables were used to represent missing-imputed data.

FRS was calculated using the most recent measurement, classifying subjects as high or low-risk for CVD:

- Full cohort: all cases plus matches controls
- Low-risk cohort: low-risk cases plus matched controls

Three machine learning algorithms implemented and trained for CVD classification:

- Logistic Regression (LR)
- Random Forest (RF)
- Gradient Boosting Trees (GBT)

Classification performance evaluated via 10-fold cross validation AUROC.

Results

Classification performance and the top 20 predictive features for each model are displayed below:

Model: (*Full AUROC* | *Low-Risk AUROC*)

GBT: (.657±.009 | .625±.010)

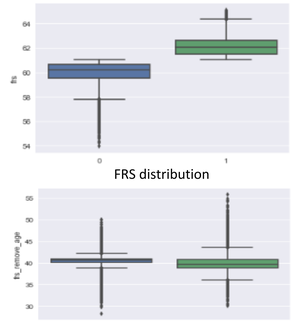
Smoking	Aspirin (med)
Essential Hyptertension (phe)	Precordial Pain (phe)
Precordial Pain(phe)	Furosemide (phe)
Aspirin (phe)	Smoking
Furosemide (phe)	Syst mean
Gluc max	Missing Blood Pressure
Missing Blood Pressure	Metoprolol (med)
Cl Counts	Gluc max
Nitroglycerin (phe)	Heparin, Porcine (med)
RDW max	Essential Hyptertension (phe)
Creat counts	NTAuto std
D250HT max	tstPBy counts
Hyperlipidemia (phe)	Multiple Sclerosis (phe)
Missing TstPBy	plt-ct Counts
Creat median	Acute Pharyngitis(phe)
QT3 max	Syst max
SUCrea std	Syst./Diast. Heart Fail. (phe)
med 284664 (proprietary)	LDH std
Dist of Skin Sensation (phe)	SGOT std
Senso. Hearing Loss (phe)	Lack of Coordination (phe)

RF: (.699±.008 | .701±.008)

Smoking	Aspirin (med)
Precordial Pain (phe)	Precordial Pain (phe)
Essential Hyptertension (phe)	Furosemide (med)
Aspirin (med)	Essential Hyptertension (phe)
Furosemide (med)	Heparin, Porcine (med)
Gluc max	Smoking
Nitroglycerin (med)	Nitroglycerin (med)
K counts	Ca min
Na min	Metoprolol (med)
Heparin, Porcine (med)	RDW max
Missing CPK	Clopidogrel (med)
Na counts	Gluc max
Wheez. & Pain Resp. (phe)	Wheez. and Pain. Resp. (phe)
Metoprolol (med)	Hgb std
Creat std	Warfarin (med)
Clopidogrel (med)	K counts
Gluc counts	Morphine (med)
Ca min	Creat counts
Acetaminophen (med)	RDW mean
Gluc mean	Syst max

LR: (.703±.005 | .681±.009)

Smoking	Precordial Pain (phe)
Missing Blood Pressure	Clopidogrel (med)
Precordial Pain (phe)	Missing Blood Pressure
Clopidogrel (med)	Nonrheu. Aort. Valve Dis. (phe)
Nonrheu. Aort Valve Dis. (phe)	Smoking
Aspirin (med)	Creat max
Missing TstPBy	Nonrheu. Mitral Valve Dis. (phe)
Convulsions (phe)	Convulsions (phe)
Trigs mean	LDL-C min
Missing MPV	NtAbs std
LDL-C (med)	Acute Pharyngitis (phe)
Missing Tprose	Chol max
Multiple Sclerosis (phe)	Missing TstPBy
Cl mean	Aspirin (med)
Syst mean	SUCrea std
LDL-C counts	Palpitations (phe)
Thyotropin (med)	Syst./Diast. Heart Fail. (phe)
Heart Failure (phe)	Thyotropin (med)
Ca min	CO2 max
Dry Eyes (phe)	Multiple Sclerosis (phe)



FRS distributions (with age removed)

Key Findings:

Similar classification performance and conventional CVD top features when trained on both the full and low-risk cohorts indicate that when age, race, and gender are ignored, the presentation of conventional risk factors is the best predictive indicator for low-risk CVD outcome.

As age is the largest contributor to FRS, our study results indicate that the FRS is highly biased by age, potentially classifying younger individuals presenting conventional symptoms as low-risk despite benefiting from CVD intervention.

Genetic Factors Play a Bigger Role in CVD Outcome in Low-Risk Individuals Compared to High-Risk

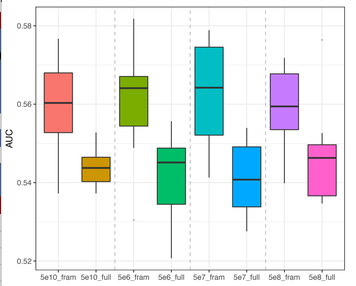
We were interested in seeing how the amount of CVD information encoded in SNPs in low-risk individuals compared to high-risk.

Individuals were selected from VUMC BioVU, Vanderbilt's biobank containing fully genotyped individuals:

- 13,900 CVD cases out of 37,573 individuals selected
- 6,553 cases classified as low-risk for CVD by FRS computed with most recent available measurements

Statistically significant SNPs identified from CVD GWAS (Genome-Wide Association Study) with various statistically significant p-value thresholds.

CVD classification results from a 3-hidden layer deep neural network trained on SNP data (no demographic information):



GWAS p-value threshold	# of SNPs	AUROC Full cohort	AUROC Low-Risk cohort	P-value
5e-6	12,339	.542±.011	.561±.014	.004
5e-7	2,644	.541±.009	.563±.013	.001
5e-8	1,859	.547±.012	.559±.010	.027
5e-10	1,198	.544±.005	.559±.012	.001

Key Findings:

1) No significant performance gains as more SNPs are included. This means that CVD relevant information is encoded in the 5e-10 GWAS SNPs.

2) There is a statistically significant performance gain in the prediction on the low-risk cohort over the full cohort pointing to unavoidable genetic causes over long-term lifestyle factors as a possible explanation for why low-risk individuals develop CVD earlier in life or present CVD without notable conventional factors.

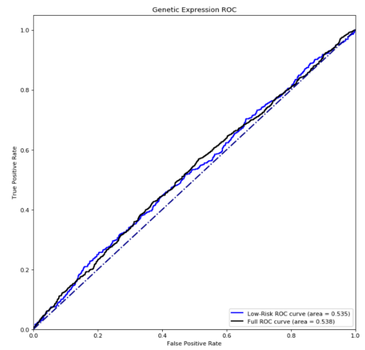
Genetic Expression Data makes no Significant Contribution

Synthetic genetic expression data generated from SNP frequency fed into a deep learning neural network

- Representative expression of 6,697 genes from blood
- 12,533 CVD cases (5489 low-risk) out of 26,722 individuals

CVD predictive performance was slightly lower than when trained on raw SNP data.

No statistically significant increase in performance for the low-risk cohort over full.



Conclusions

Longitudinal EHR & Genetic information holds predictive insight for individuals considered low-risk for CVD. Our major conclusions are that the FRS is overly influenced by age causing it to overlook younger individuals with conventional risk factors and that genetic information provides a larger contribution to CVD progression in the low-risk population. Future work includes using stricter definitions for individuals considered low-risk and late-fusion approaches for combining trained predictive models.