Disentanglement of Demographic and Genetic Factors on the ECG Morphology

Álvaro Morcillo, Patricia B Munroe, Ana Mincholé, Juan Pablo Martínez, Julia Ramírez

Aragón Institute of Engineering Research, University of Zaragoza Zaragoza, Spain

Background: Electrocardiogram (ECG) morphology is influenced by demographic and genetic factors. Disentangling them is crucial for cardiovascular disease (CVD) modeling to prevent misinterpreting normal variations as pathology. While Autoencoders (AEs) are widely used, previous studies have neither explicitly disentangled these factors nor included Polygenic Risk Scores (PRS) as inputs. Integrating PRSs remains challenging as weak genetic signals emerge only at distribution extremes. We aimed to develop an AE capable of disentangling these variables, yielding a confounder-corrected ECG representation (z_{rest}) while overcoming extreme PRS modeling barriers.

Materials and methods: We analyzed the 8 independent leads from standard resting ECGs from 40,039 healthy UK Biobank individuals, developing an AE using auxiliary classifiers to segment latent vectors into demographic (sex, age) and genetic (QT, QRS, PR-interval PRSs) subvectors, as well as a purified z_{rest} that captures all unexplained morphology. To capture weak genetic signals without discarding data, we applied selective masking: the model learned general morphology from the full cohort while restricting the model's training on genetic effects to extreme percentiles ($<5^{th}$, $>95^{th}$). Cross-prediction validated latent purity: vectors could not predict unassigned covariates, guaranteeing minimal information leakage.

Table 1. Cross-prediction performance of latent vectors.

Metrics	z_{sex}	z_{age}	z_{PRS}	z_{rest}
AUC Sex	**0.971**	0.533	0.541	0.539
MAE age	6.894	**5.494**	6.878	6.966
r age	0.0586	**0.544**	0.131	0.220
R^2 age	-0.209	**0.1998**	-0.207	-0.278
AUC QT	0.564	0.473	**0.865**	0.526
AUC QRS	0.546	0.519	**0.785**	0.545
AUC PR	0.515	0.556	**0.801**	0.527

Results: Target predictions from corresponding subvectors achieved high accuracy (Sex AUC=0.971, Age MAE=5.494, PRS AUCs: 0.785–0.865; Table 1). Conversely, non-targeted predictions (e.g., sex from z_{age}) fell to random level (AUC $\approx$ 0.50) and failed age regression, confirming robust disentanglement and validating z_{rest} as a latent vector free from information leakage.

Conclusions: By employing a novel selective masking strategy, our AE successfully disentangles demographic and genetic confounders. The resulting purified baseline (z_{rest}) prevents the misinterpretation of healthy inter-subject variations, paving the way for more accurate CVD risk predictions.