Early and Late Fusion Methods for Cardiovascular Disease Prediction from Longitudinal EHR and Genetic Data

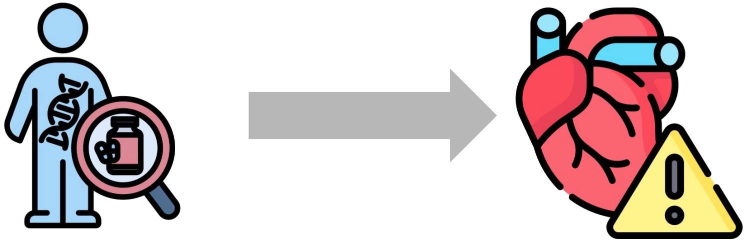
Team CoHERent
aka: Nia Abdurezak, Sophie Kearney, Ananya Rajagopalan

December 5th, 2025

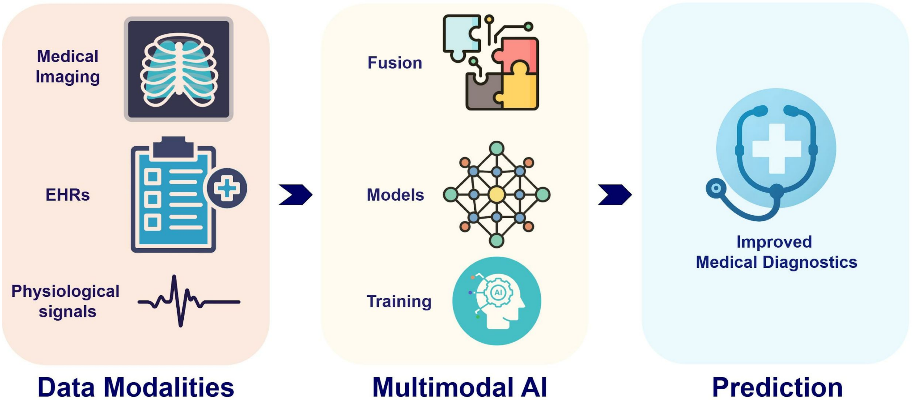
Traditional risk prediction methods for CVD do not capture genetic or longitudinal effects

- Genetic risk is excluded, despite it being a highly heritable trait (twin study estimate of ~50%[1])
- The promise of precision medicine: the right treatment for the right patient at the right time, requires incorporating biological data **(multimodal)** to comprehensively understand patient health

[1] https://doi.org/10.1016/j.jjcc.2021.09.005



Multimodal AI in Medical Diagnostics



Existing multimodal disease prediction studies have a limited focus on comparing data representation methods and interpretability

Prior Work

Zhao et al. (2019) found that XGBoost best predicted 10-year CVD risk (EHR + genetic), but only used late fusion

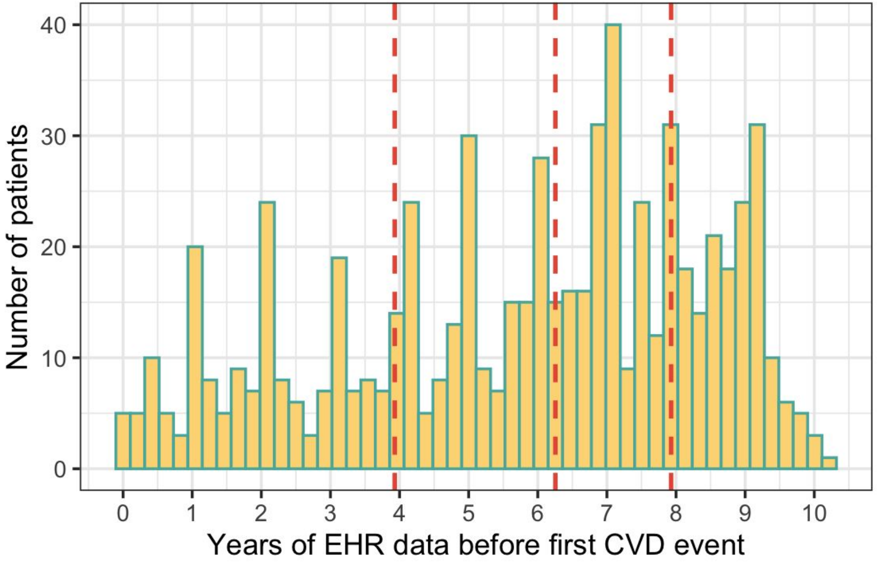
They also had an incomplete feature analysis because they used RNNs and LSTMs (black boxes)

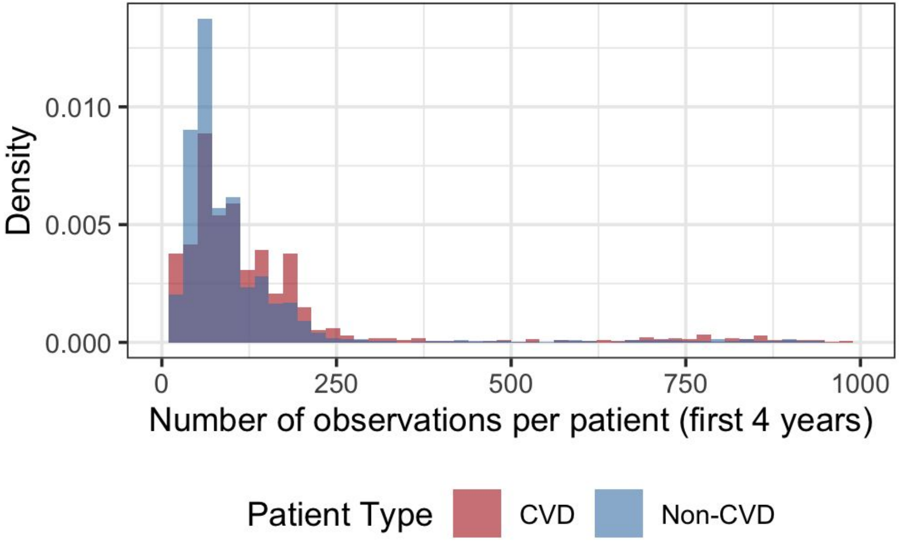
Our Contributions

1) **evaluating early and late-stage** data fusion approaches for multimodal CVD prediction

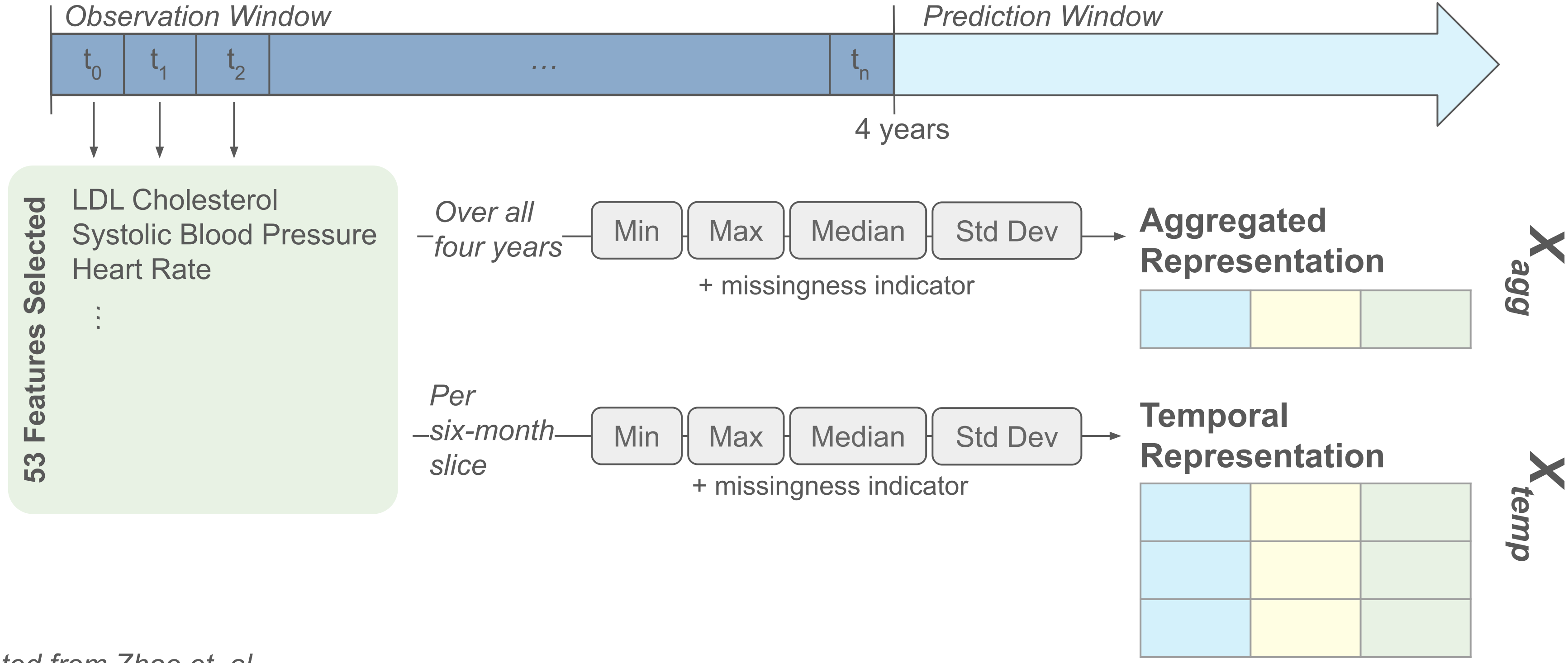
2) **exploring interpretable ML methods** (including transformers) to model multimodal EHR data longitudinally

Coherent: Synthetic, Multimodal, Longitudinal Dataset for CVD



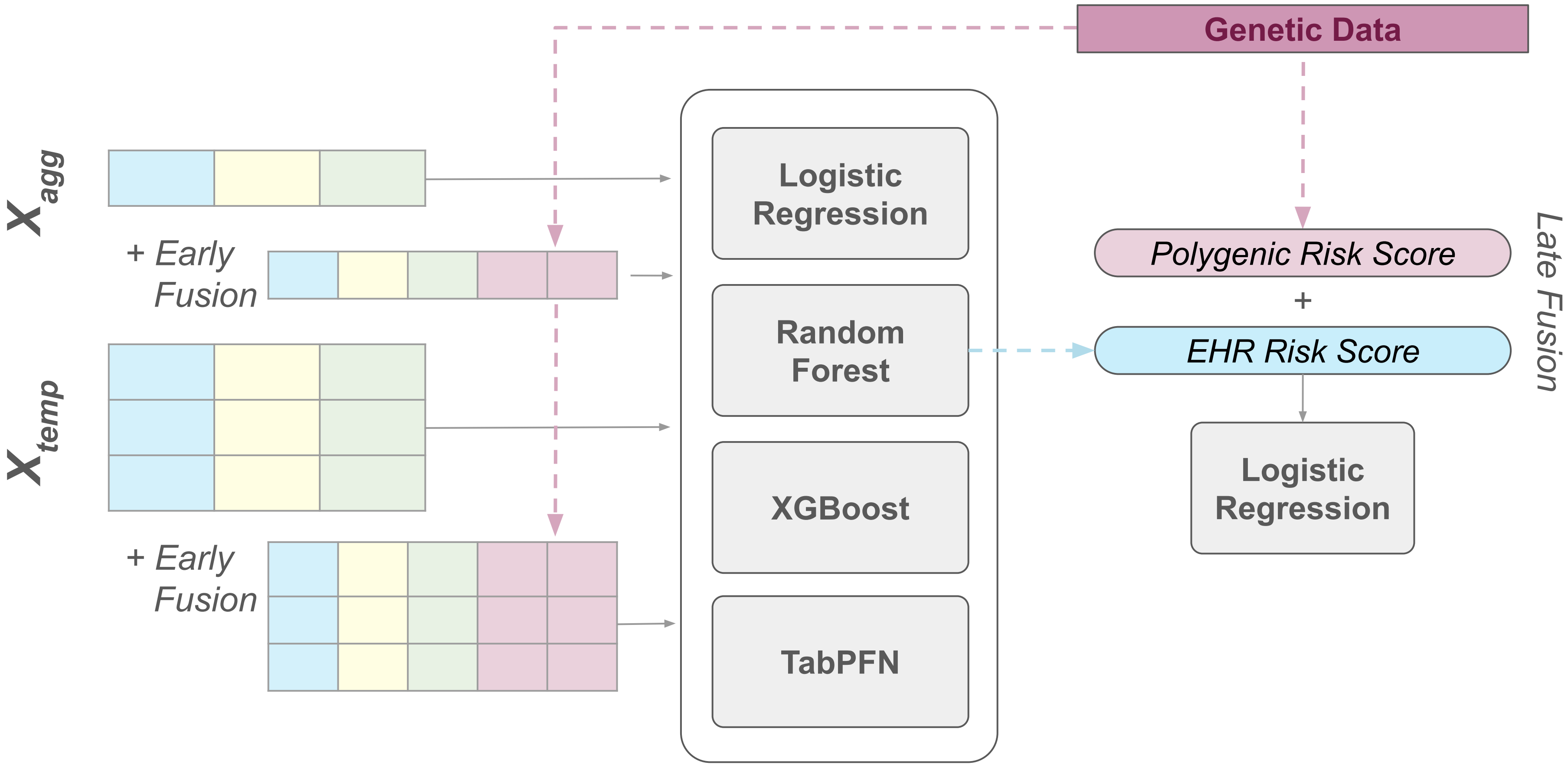


Processing Longitudinal EHR Data into Aggregated and Temporal Representations

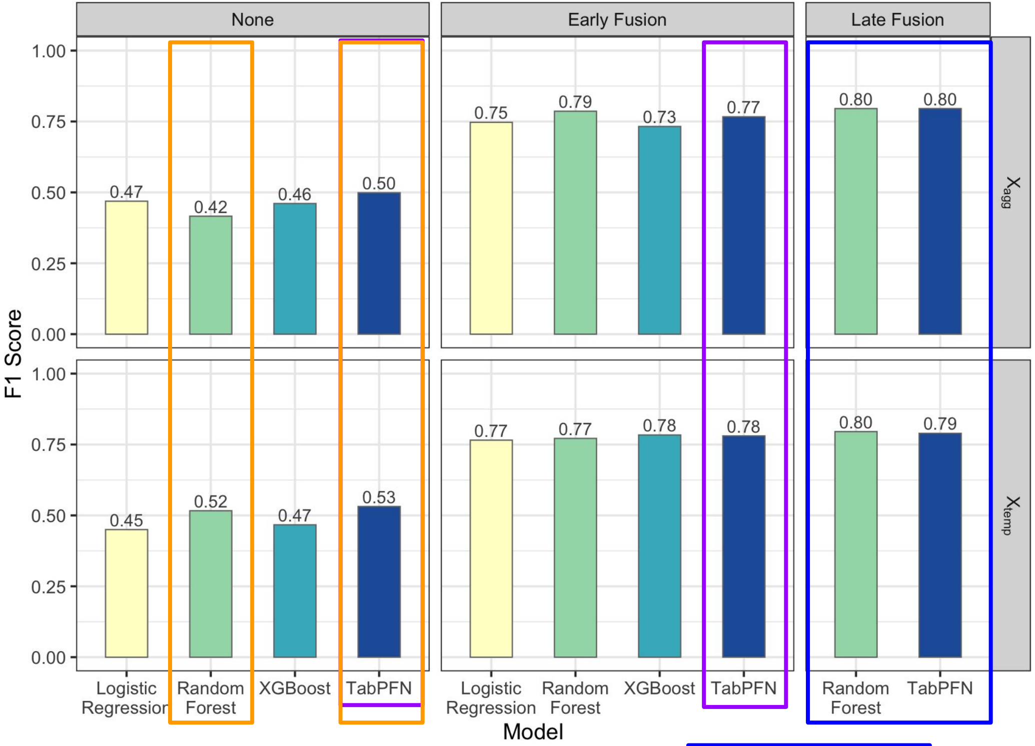


Adapted from Zhao et. al

Integrating EHR and Genetic Data for Prediction

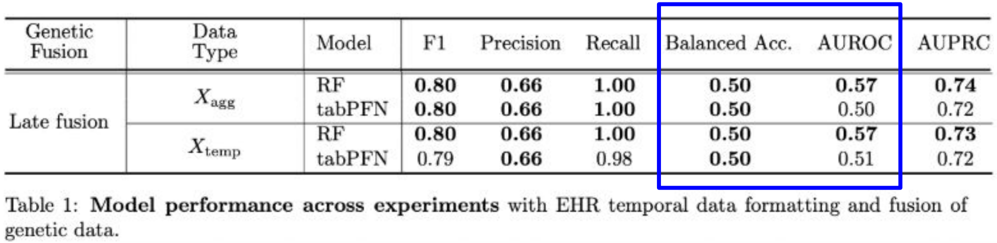


Early Fusion shows significant improvement compared to EHR only, in all models



+.27

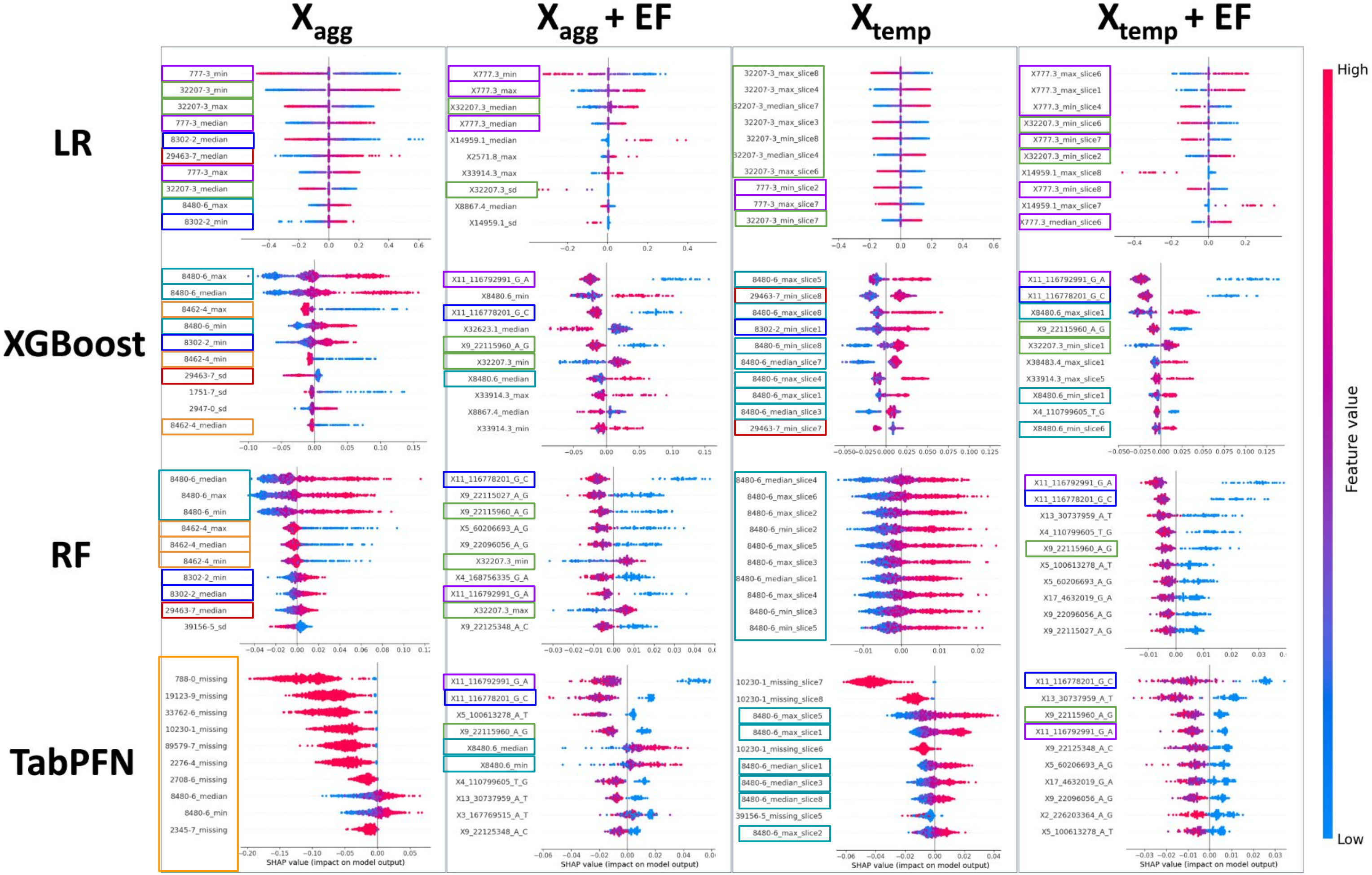
+.25



Genetic Fusion	Data Type	Model	F1	Precision	Recall	Balanced Acc.	AUROC	AUPRC
Late fusion	X_{agg}	RF	**0.80**	**0.66**	**1.00**	**0.50**	**0.57**	**0.74**
		tabPFN	**0.80**	**0.66**	**1.00**	**0.50**	0.50	0.72
	X_{temp}	RF	**0.80**	**0.66**	**1.00**	**0.50**	**0.57**	**0.73**
		tabPFN	0.79	**0.66**	0.98	**0.50**	0.51	0.72

Table 1: **Model performance across experiments** with EHR temporal data formatting and fusion of genetic data.

Feature analysis demonstrates clinical relevance



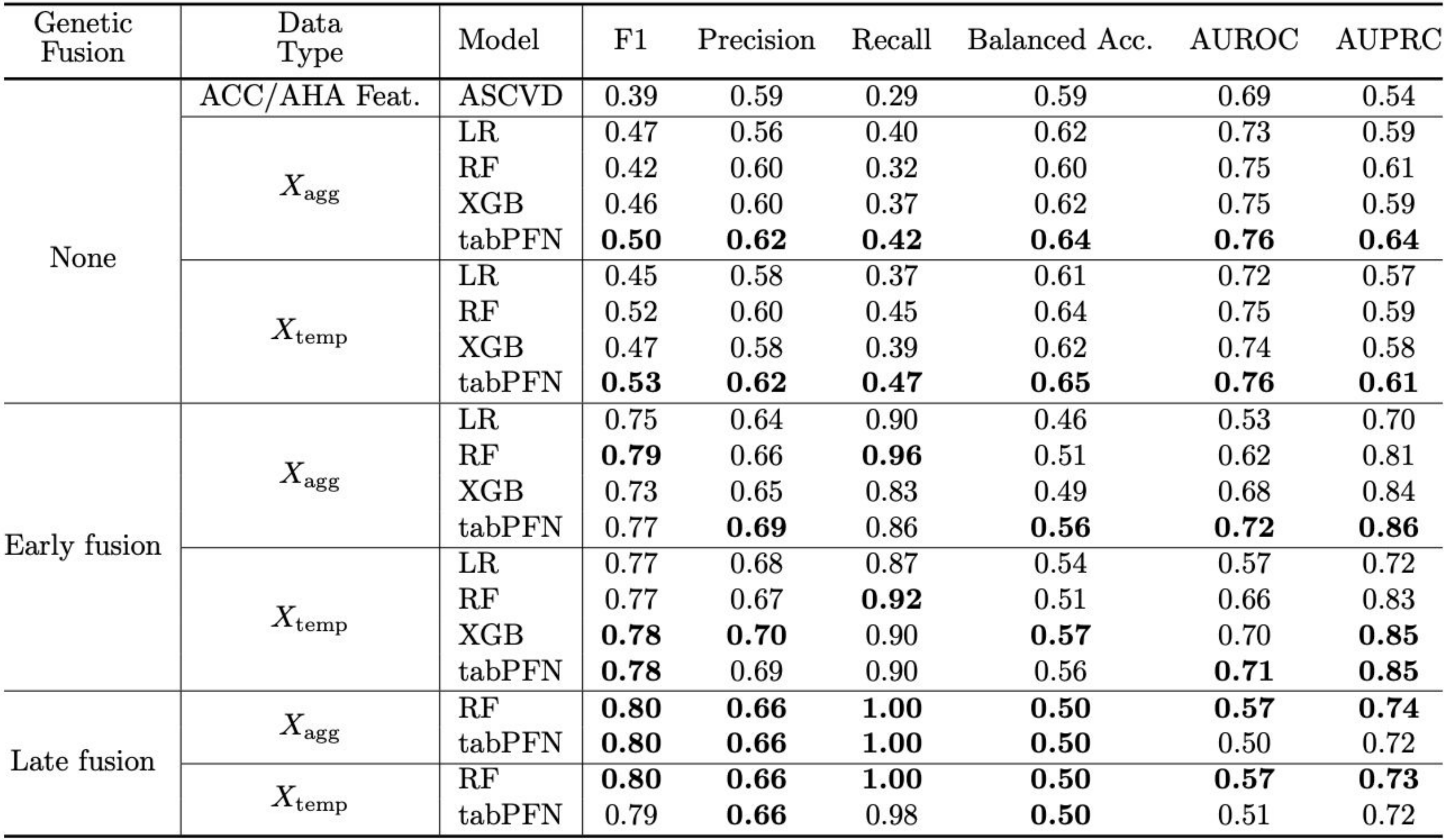
Genetic data is valuable, and the representation method matters

Across all model architectures, **adding genetic data** alongside EHR data (regardless of the fusion method) **improved model performance** compared to EHR data alone

Early fusion drives better model performance compared to late fusion → supports our hypothesis that early fusion provides richer information for the model

Feature importance analysis reveals a **mix of traditional risk factors + molecular biomarkers + key genetic variants** being most predictive of future CVD

Thank you for your attention!
Questions?



Genetic Fusion	Data Type	Model	F1	Precision	Recall	Balanced Acc.	AUROC	AUPRC
None	ACC/AHA Feat.	ASCVD	0.39	0.59	0.29	0.59	0.69	0.54
	X_{agg}	LR	0.47	0.56	0.40	0.62	0.73	0.59
		RF	0.42	0.60	0.32	0.60	0.75	0.61
		XGB	0.46	0.60	0.37	0.62	0.75	0.59
		tabPFN	**0.50**	**0.62**	**0.42**	**0.64**	**0.76**	**0.64**
	X_{temp}	LR	0.45	0.58	0.37	0.61	0.72	0.57
		RF	0.52	0.60	0.45	0.64	0.75	0.59
		XGB	0.47	0.58	0.39	0.62	0.74	0.58
		tabPFN	**0.53**	**0.62**	**0.47**	**0.65**	**0.76**	**0.61**
Early fusion	X_{agg}	LR	0.75	0.64	0.90	0.46	0.53	0.70
		RF	**0.79**	0.66	**0.96**	0.51	0.62	0.81
		XGB	0.73	0.65	0.83	0.49	0.68	0.84
		tabPFN	0.77	**0.69**	0.86	**0.56**	**0.72**	**0.86**
	X_{temp}	LR	0.77	0.68	0.87	0.54	0.57	0.72
		RF	0.77	0.67	**0.92**	0.51	0.66	0.83
		XGB	**0.78**	**0.70**	0.90	**0.57**	0.70	**0.85**
		tabPFN	**0.78**	0.69	0.90	0.56	**0.71**	**0.85**
Late fusion	X_{agg}	RF	**0.80**	**0.66**	**1.00**	**0.50**	**0.57**	**0.74**
		tabPFN	**0.80**	**0.66**	**1.00**	**0.50**	0.50	0.72
	X_{temp}	RF	**0.80**	**0.66**	**1.00**	**0.50**	**0.57**	**0.73**
		tabPFN	0.79	**0.66**	0.98	**0.50**	0.51	0.72

Table 1: **Model performance across experiments** with EHR temporal data formatting and fusion of genetic data.