A scalable multimodal framework for unbiased risk biomarker discovery across multiple cancer types

Constantin Petrescu[1*], Jack M. Monahan[1*], Abbas Salami[1], Lisa J Schmunk[1] and Thomas M. Stubbs[1†].

[1]Hurdle.Bio, London, UK. *Contributed equally to the work. †Corresponding author with email. Presenter of the work at AACR 3.

HURDLE

1116

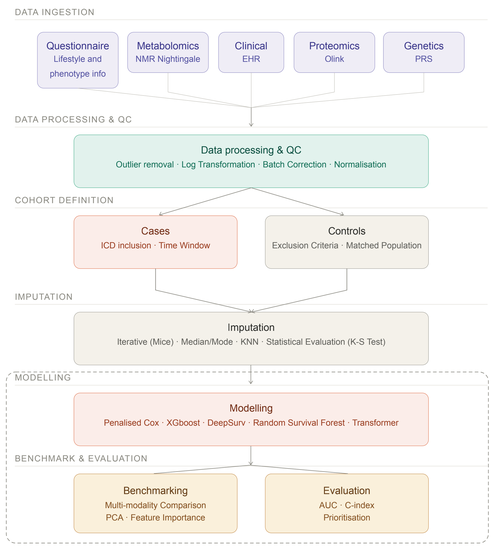
Background

Traditional cancer risk models are typically based on single data modalities, whereas recent approaches increasingly integrate multi-modal data [1]. The emergence of large-scale, comprehensive biobanks with rich, deep clinical phenotyping empowers the discovery of novel risk biomarkers across multiple cancer types. The UK Biobank (UKB) includes over 500,000 participants with longitudinal, multi-modal data (clinical, imaging, genomic, metabolomic, and proteomic), over 50,000 incident cancer cases have been identified following the initial assessment centre visit.

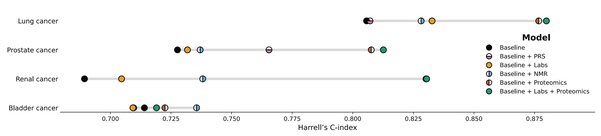
Few frameworks have leveraged this resource for cancer risk prediction. MILTON [2] applies classical machine learning to clinical and proteomics features from the UKB to predict time-agnostic or prognostic biomarkers across hundreds of phenotypes. ProteinScores [3] have been developed to demonstrate the utility of proteomics for deriving protein-based disease risk scores at scale.

Here, we present Hurdle's flexible, biomarker discovery pipeline, ***DiscoDx***, that integrates multimodal data into various machine learning architectures, enabling the identification of early cancer risk signatures across multiple primary cancer types. We present a case study on renal cancer (ICD-10 C64) within the UKB, and benchmark our results against those reported by MILTON.

Overview

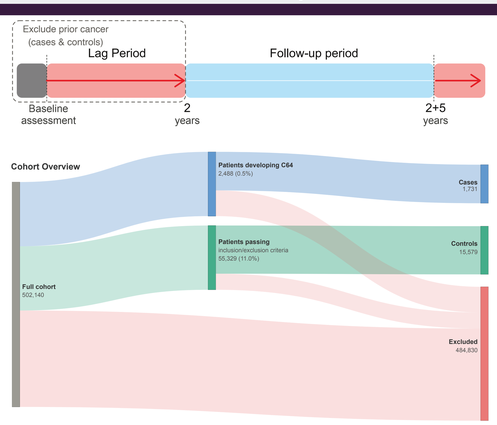


A) Overview of the Hurdle multi-modal biomarker discovery framework. The pipeline integrates multiple data types from multiple data sources (UKB, All of Us, Finnish Biobank). Cohorts can be defined using ICD outcomes, specific feature thresholds and time windows.

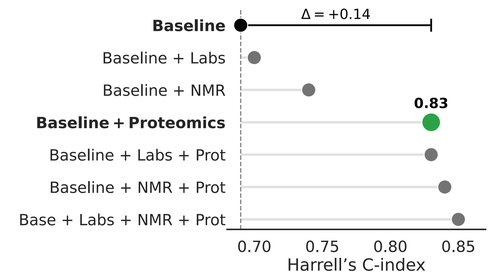


B) Comparison of Penalised Cox Regression model performances with grid search on Elastic Net Regularisation across multiple primary cancers (ICD-10). With some exceptions, multimodal models had higher discriminatory power than simpler single modality models. Models including plasma proteomic features generally outperformed those that did not. Notably, NMR metabolomic features outperformed proteomic features in bladder cancer.

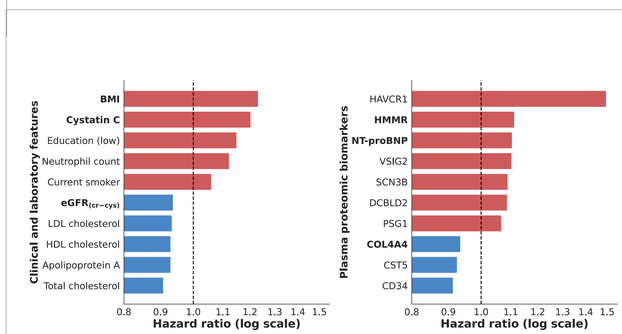
Proteomics Improves Risk Prediction of Incident Renal Cancer in UKBB



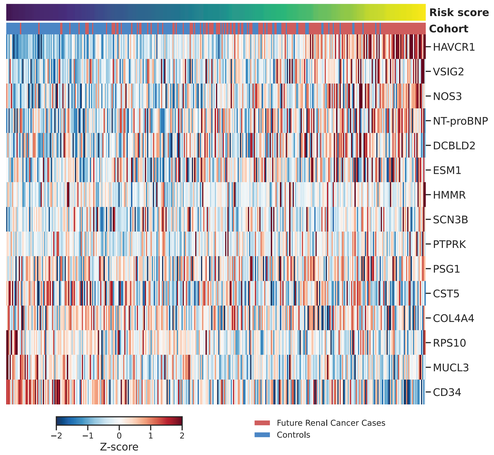
A) Study design for Cox proportional hazards analysis. Individuals with any prior cancer were excluded at baseline. A 2-year lag period was applied, followed by up to 5 years of follow-up for incident kidney cancer outcomes. Of the initial 500K participants in the UKB, the final cohort comprised **1,731 cases** and **15,579 controls.**



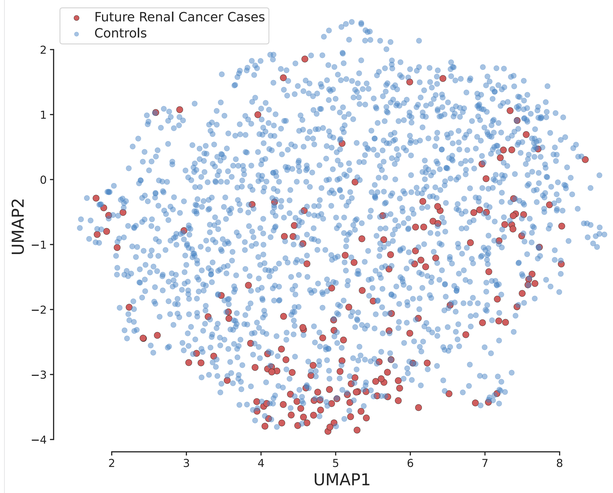
B) Concordance Index for different models. Multimodal models had the highest discriminatory power, but inclusion of **plasma proteomic features** alone **substantially improved discrimination** (+0.14) compared to the baseline model (demographic and anthropometric features).

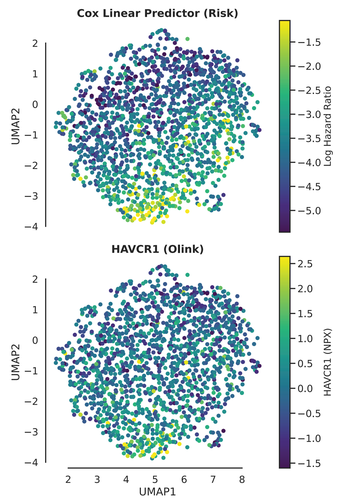


C) Hazard ratios from clinical and proteomic Cox models for renal cancer. Key proteomic predictors, including HAVCR1 (**KIM-1**) and HMMR, alongside clinical markers such as **eGFR and cystatin C, capture shared renal and glomerular pathophysiology** [4,5].



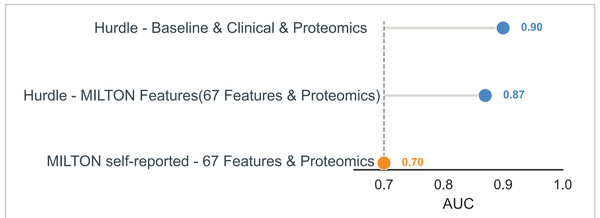
D) Heatmap of the top 20 proteomic predictors from the renal cancer model. Subset of cohort ordered by predicted risk score. Expression (Z-score) patterns show coordinated variation across key biomarkers, including HAVCR1 (KIM-1), CD34 and NT-proBNP, highlighting **risk-associated proteomic signatures consistent with cardiorenal and vascular pathways** [6].



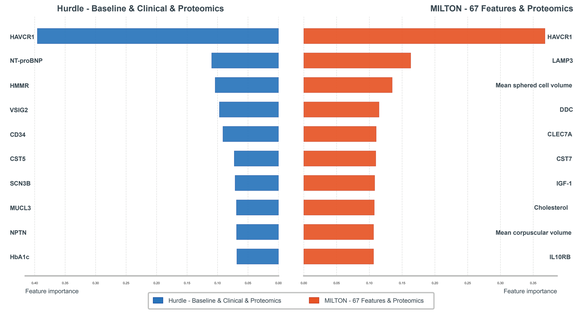


E) UMAP embedding of individuals based on proteomic profiles. On left, future renal cancer cases (red) are interspersed among controls (blue), showing some local enrichment but no distinct clustering, suggesting is a heterogeneous, diffuse risk signal across the proteomic space. Paired UMAP projections on right are coloured by Cox model risk score (top) and HAVCR1 (KIM-1) NPX (bottom). A shared gradient across the embedding indicates that **HAVCR1 tracks the spatial distribution of predicted renal cancer risk.**

Outperforms state-of-the-art



A) AUC comparison of best performing Hurdle and MILTON models. MILTON's reported features [2] were trained within our pipeline to remove the confounding effect of Hurdle's additional inclusion/exclusion criteria and feature QC.



B) Top 10 prognostic features from Hurdle and reported by Milton for incident renal cancer (C64) from the best performing models.

Conclusions

Feature agnostic, multimodal models had the highest discriminatory power in predicting incident renal cancer, but the inclusion of plasma proteomic features alone substantially improved discrimination. Years before clinical manifestation, subtle multimodal differences in cardiorenal and clinical features already distinguish individuals who later develop renal cancer from controls. Further optimisations are possible, feature selection and prioritisation would further sharpen model performances and improve cost-effectiveness in a health economic context.

Hurdle's DiscoDx pipeline outperforms MILTON, the current gold-standard in biomarker discovery. In addition to risk biomarkers, DiscoDx's flexible framework is capable of identifying novel diagnostic, monitoring and prognostic marker that can be used at any stage in a patient's disease journey. DiscoDx's flexibility facilitates federated learning across multiple biobanks and other large cohorts.

References

[1] - Javaid, H. (2025). The impact of artificial intelligence on biomarker discovery. Emerging Topics in Life Sciences, 8(2). https://doi.org/10.1042/ETLS20243003

[2] - Garg, M. (2024). Disease prediction with multi-omics and biomarkers empowers case–control genetic discoveries in the UK Biobank. Nature Genetics, 56(9). https://-doi.org/10.1038/s41588-024-01898-1

[3] - Gadd, D.A. (2024). Blood protein assessment of leading incident diseases and mortality in the UK Biobank. Nature Aging. https://doi.org/10.1038/s43587-024-00655-7

[4] - Lee, J.C. (2021). Kidney injury molecule-1 inhibits metastasis of renal cell carcinoma. Scientific Reports, 11. https://doi.org/10.1038/s41598-021-90919-8

[5] - Lees, J.S. (2021). Kidney function and cancer risk: an analysis using creatinine and cystatin C in a cohort study. EClinicalMedicine, 38. https://-doi.org/10.1016/j.eclinm.2021.101030

[6] Chaikijurajai, T. (2023). Prognostic value of natriuretic peptide levels for adverse renal outcomes in patients with moderate to severe acute kidney injury. Journal of the American Heart Association, 12(21). https://doi.org/10.1161/JAHA.123.031453