Machine learning and integrative 'omics for multi-cancer risk prediction

Principal supervisor's name: Dr **Siddhartha Kar**
Principal supervisor's email address: sk718@cam.ac.uk

Principal supervisor's CRUK CC theme:

- Early Cancer Institute
- Ovarian Cancer Programme

Department for student registration: Department of Oncology
Department or institute where research will take place: Early Cancer Institute

Postgraduate scheme: **MRes + PhD (1 + 3-year non-clinical applicants only)**

MRes project outline:

Multi-cancer risk prediction is the assessment of an individual's susceptibility to multiple types of cancer. Traditional cancer risk prediction models focus on one specific type of cancer, such as colorectal or prostate cancer. However, multi-cancer risk prediction has the potential to provide a more comprehensive, personalised understanding of susceptibility to a range of different cancer types. By identifying those who are at higher risk for any of multiple types of cancer, recommendations for cancer prevention, screening, and early intervention can be tailored, prioritising lifestyle modifications, or targeted and more frequent screenings, or other interventions to catch potential cancers at earlier, more treatable stages with the consequent possibility of improvement in outcomes. Multi-cancer risk prediction also allows for the ability to learn from risk profiling for common cancers, such as breast cancer, for which powerful risk prediction tools are already available and transfer these learnings to relatively rarer cancers such as pancreatic cancer where risk prediction is lagging. The remarkable convergence of computational advances and deeply-phenotyped data availability offers a unique and unprecedented opportunity now for this project that will drive the development of multi-cancer risk prediction at scale.

Screening tests (e.g., mammography) presently used in the clinic for the early detection of cancer focus on a single cancer type (in this case, breast cancer). Recent years have witnessed the rapid development of multi-cancer early detection (MCED) molecular screening tests in the research setting. As the name suggests, MCED tests aim to detect any of several cancer types using a single blood sample. However, the current crop of MCED tests under commercial development have so far suffered from poor sensitivity in identifying early-stage cancers. It is imperative to improve the detection of early-stage cancers across the spectrum of cancer types since it is precisely at these stages that timely therapeutic intervention can have the greatest impact on cancer mortality. Targeting MCED tests to those at the highest risk of developing various cancers will be vital to improving their performance and efficiency. However, we currently do not have statistical/epidemiological tools or algorithms for population-level multi-cancer risk prediction and stratification that are any more sophisticated than simply offering these MCED screening tests to older age groups or to smokers. Hence there is a pressing need for multi-cancer risk prediction models and population stratification strategies and delivering these represents the overarching ambition of this project.

The project will leverage state-of-the-art artificial intelligence/machine learning (AI/ML) techniques coupled with conventional statistical methods that have stood the test of time and tap into large-scale multi-omics and clinico-epidemiological data (lifestyle traits, family history of cancer, disease diagnoses etc.) to train, test, validate, and perform objective evaluations of multi-cancer risk prediction models. The specific aim/objective at the MRes level is:

1. Build and evaluate a circulating proteomics-based multi-cancer risk prediction model

If the student would like to continue this project for a PhD then the specific aims/objectives will additionally include:

2. Build and evaluate a circulating metabolomics-based multi-cancer risk prediction model

3. Build and evaluate multi-cancer polygenic risk scores

4. Integrate genomic, proteomic, and metabolomic data with demographic, lifestyle and clinical information into an omnibus risk prediction approach

MRes experimental plan:

The experimental plan for this computationally- and statistically-intensive project involves extensive use of the UK Biobank, a prospective longitudinal study containing in-depth genetic and health information from half a million UK participants. Specifically, the UK Biobank provides us with access to circulating protein levels for >3,000 proteins measured in >55,000 individuals using the Olink proteomics platform.

The UK Biobank will be split randomly into training, testing, and evaluation data sets for model construction, optimisation (parameter tuning by cross-validation and grid searches), and validation, respectively. Models will be developed in this way for time to incident cancer diagnoses both at the single cancer and the multi-cancer levels. The results will be explored to determine the best approach to combining predictions from the single cancer models into a pan-cancer algorithm. Cox proportional hazards with feature selection via regularisation, random survival forests, gradient boosting, and support vector machines will be the mainstay machine learning methods used and evaluated. Model performance metrics will include assessments of discrimination and calibration. Models will include age, sex, genetic ancestry, smoking status, alcohol consumption, and body mass index as covariates and will be compared to baseline models that include just these covariates.

If the student would like to continue this project up for the PhD, the following additional experimental plan also applies:

For each additional layer of data (metabolomics, genomics, clinico-epidemiological), the methods will be analogous to those described above for the proteomics-based MRes project. For metabolomics, the UK Biobank offers circulating levels of ~250 metabolites from >120,000 individuals profiled on the Nightingale metabolomics platform. The Biobank contains whole-genome genotype and sequencing data and whole exome sequences on nearly 500,000 individuals with linked epidemiological and electronic health records, which will enable the genomics component of the PhD.

Additional models that include epidemiological (e.g., family history of cancer), lifestyle (e.g., physical activity and sleep as measured using wearables), and clinical (such as cancer-specific risk factors, e.g., endometriosis diagnoses for ovarian cancer) variables linked to the UK Biobank

will also be built and evaluated as standalone and in conjunction with the 'omics data. Recently published multi-trait polygenic scoring methods will be applied to genome-wide association study (GWAS) data from international consortia for the common cancers to develop the multi-cancer polygenic scores. Beyond the UK Biobank, the project will also offer the possibility of model validation in external cohorts such as Our Future Health, All of US, INTERVAL, and East London Genes and Health.

PhD project outline:

Multi-cancer risk prediction is the assessment of an individual's susceptibility to multiple types of cancer. Traditional cancer risk prediction models focus on one specific type of cancer, such as colorectal or prostate cancer. However, multi-cancer risk prediction has the potential to provide a more comprehensive, personalised understanding of susceptibility to a range of different cancer types. By identifying those who are at higher risk for any of multiple types of cancer, recommendations for cancer prevention, screening, and early intervention can be tailored, prioritising lifestyle modifications, or targeted and more frequent screenings, or other interventions to catch potential cancers at earlier, more treatable stages with the consequent possibility of improvement in outcomes. Multi-cancer risk prediction also allows for the ability to learn from risk profiling for common cancers, such as breast cancer, for which powerful risk prediction tools are already available and transfer these learnings to relatively rarer cancers such as pancreatic cancer where risk prediction is lagging. The remarkable convergence of computational advances and deeply-phenotyped data availability offers a unique and unprecedented opportunity now for this PhD project that will drive the development of multi-cancer risk prediction at scale.

Screening tests (e.g., mammography) presently used in the clinic for the early detection of cancer focus on a single cancer type (in this case, breast cancer). Recent years have witnessed the rapid development of multi-cancer early detection (MCED) molecular screening tests in the research setting. As the name suggests, MCED tests aim to detect any of several cancer types using a single blood sample. However, the current crop of MCED tests under commercial development have so far suffered from poor sensitivity in identifying early-stage cancers. It is imperative to improve the detection of early-stage cancers across the spectrum of cancer types since it is precisely at these stages that timely therapeutic intervention can have the greatest impact on cancer mortality. Targeting MCED tests to those at the highest risk of developing various cancers will be vital to improving their performance and efficiency. However, we currently do not have statistical/epidemiological tools or algorithms for population-level multi-cancer risk prediction and stratification that are any more sophisticated than simply offering these MCED screening tests to older age groups or to smokers. Hence there is a pressing need for multi-cancer risk prediction models and population stratification strategies and delivering these represents the overarching ambition of this PhD project.

The project will leverage state-of-the-art artificial intelligence/machine learning (AI/ML) techniques coupled with conventional statistical methods that have stood the test of time and tap into large-scale genomics, proteomics, metabolomics, and clinico-epidemiological data (lifestyle traits, family history of cancer, disease diagnoses, etc.) to train, test, validate, and perform objective evaluations of multi-cancer risk prediction models. The specific aims/objectives of the PhD are:

1. Build and evaluate circulating proteomics- and metabolomics-based multi-cancer risk prediction models
2. Build and evaluate multi-cancer polygenic risk scores
3. Integrate genomic, proteomic, and metabolomic data with demographic, lifestyle and clinical information into an omnibus risk prediction approach

PhD experimental plan:

The experimental plan for this computationally- and statistically-intensive PhD project involves extensive use of the UK Biobank, a prospective longitudinal study containing in-depth genetic and health information from half a million UK participants. Specifically, the UK Biobank provides us with access to circulating protein levels for >3,000 proteins measured in >55,000 individuals using the Olink proteomics platform. For metabolomics, the UK Biobank offers circulating levels of ~250 metabolites from >120,000 individuals profiled on the Nightingale metabolomics platform. The Biobank contains whole-genome genotype and sequencing data and whole exome sequences on nearly 500,000 individuals with linked epidemiological and electronic health records, which will enable the genomics component of the PhD.

For each layer of data (proteomics, metabolomics, genomics, clinico-epidemiological), the UK Biobank will be split randomly into training, testing, and evaluation data sets for model construction, optimisation (parameter tuning by cross-validation and grid searches), and validation, respectively. Models will be developed in this way for time to incident cancer diagnoses both at the single cancer and the multi-cancer levels. The results will be explored to determine the best approach to combining predictions from the single cancer models into a pan-cancer algorithm. Cox proportional hazards with genomic/proteomic/metabolomic feature selection via regularisation, random survival forests, gradient boosting, and support vector machines will be the mainstay machine learning methods used and evaluated. Model performance metrics will include assessments of discrimination and calibration. Models will include age, sex, genetic ancestry, smoking status, alcohol consumption, and body mass index as covariates and will be compared to baseline models that include just these covariates. Additional models that include epidemiological (e.g., family history of cancer), lifestyle (e.g., physical activity and sleep as measured using wearables), and clinical (such as cancer-specific risk factors, e.g., endometriosis diagnoses for ovarian cancer) variables linked to the UK Biobank will also be built and evaluated as standalone and in conjunction with the ‘omics data. Recently published multi-trait polygenic scoring methods will be applied to genome-wide association study (GWAS) data from international consortia for the common cancers to develop the multi-cancer polygenic scores. Beyond the UK Biobank, the project will also offer the possibility of model validation in external cohorts such as Our Future Health, All of US, INTERVAL, and East London Genes and Health.

Main techniques:

This project will offer wide-ranging and in-depth training in modern computational and statistical methods in biomedical data science. The main techniques include the handling of big multi-omic and electronic health record data, regression modelling strategies, and artificial intelligence/machine learning (AI/ML). Specifically, but not limited to, Cox proportional hazards modelling, regularized regression, random survival forests, gradient boosting, support vector machines, methods for dealing with missing data and class imbalance, methods for evaluating the discrimination and calibration of AI/ML models, multi-omic integration, statistical genetics and genetic epidemiology (GWAS and polygenic risk scoring), and R/Python programming (especially tidymodels or Keras/TensorFlow frameworks).

Key references:

1. Jung, A. W. et al. Multi-cancer risk stratification based on national health data: A retrospective modelling and validation study. (2022) doi:10.1101/2022.10.12.22280908.
2. Kim, E. S. et al. Potential utility of risk stratification for multicancer screening with liquid biopsy tests. NPJ Precis Oncol 7, 39 (2023).

3. Albiñana, C. et al. Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. Nat Commun 14, 4702 (2023).
4. Raoof, S. et al. Multicancer Early Detection Technologies: A Review Informed by Past Cancer Screening Studies. Cancer Epidemiol Biomarkers Prev 31, 1139–1145 (2022).
5. Kar, S. P. et al. Genome-wide meta-analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types. Cancer Discov 6, 1052–1067 (2016).
6. Yang, X., Kar, S., Antoniou, A. C. & Pharoah, P. D. P. Polygenic scores in cancer. Nat Rev Cancer 23, 619–630 (2023).
7. Patel, A. V. et al. Key risk factors for the relative and absolute 5-year risk of cancer to enhance cancer screening and prevention. Cancer 128, 3502–3515 (2022).