**NIHR Undergraduate Data Science Internship Projects 2025**

Dr Elinor Curnow. Developing guidance software for missing data analysis

This internship will involve further development of a published R package, created by Dr Curnow ([https://CRAN.R-project.org/package=midoc](https://CRAN.R-project.org/package%3Dmidoc)). The Multiple Imputation Doctor (midoc) R package provides guidance for researchers on how best to analyse data when some information is missing. It is particularly targeted at researchers working on clinical and health studies, where missing data is a very common problem. Methods for analysing the available data must be chosen carefully, or the results of the study will be biased, or less precise than they should be, or both. Choosing the best approach is complex, involving multiple decisions which depend on the study design, the analysis model, and the available information.

The midoc package guides researchers step-by-step through the decision-making process. The package currently provides functions to describe the missing data patterns, explore model and method assumptions, and apply the most appropriate missing data strategy for the researcher’s data. The intern’s role will be to review package code for robustness and quality, and develop testing suites for package functions.

Note that this is a separate internship from the similar project offered as a Widening Participation Research Summer Internship (<https://www.bristol.ac.uk/study/postgraduate/widening-participation-and-access/summer-research-opportunities/research-summer-internships/>).

Dr Charlotte James: Understanding how hyperparameter restrictions impact the performance of machine learning algorithms

Secure Data Environments (SDEs) enable research using sensitive data, while protecting privacy and preventing disclosure of information that can identify individuals. The accessibility of detailed electronic health records through SDEs presents an opportunity for developing personalized prediction models that could improve population health through better timing and targeting of treatments. The scale and complexity of electronic health record data means machine learning (ML) is suited to this task. However, extracting trained ML models from SDEs is challenging because there is a risk that the model could be used to reveal sensitive information about individuals it was trained on.

One solution to this is to put restrictions on hyperparameters of ML models. These restrictions ensure the models are ‘safe’; information about individuals in the training data can’t be learnt from the trained ML model, and it can therefore be output from an SDE. However, we don’t know the effect these privacy-preserving restrictions have on ML model performance. If there are tasks that optimized ML models perform well on, but hyperparameter restrictions mean performance drops and the model is no longer useful, alternative definitions of ‘safe’ models are needed. Conversely, if hyperparameter restrictions do not impact model performance, this class of ‘safe’ models will enable researchers to develop personalized prediction models in SDEs and extract them from the environments. These models could have a significant, positive, impact on population health and patient outcomes in the future.

[SACRO-ML](https://github.com/AI-SDC/SACRO-ML/tree/0a34c3d7710c672ea5c1a888bb98a0fe19e0f607) is a python package designed to support the safe output of trained ML models from SDEs. It is still being developed, and currently contains ‘safe’ definitions of several common ML classifiers, including RandomForest and Support Vector Machine.

The goal of the internship project is to compare the performance of these ‘safe’ classifiers, to classifiers where hyperparameters have been optimized, rather than restricted. The intern will work with publicly available data that is like the electronic health record data in SDEs.

Dr. Chin Yang Shapland: Improving [triangulate](https://github.com/mcguinlu/triangulate) R package for evidence synthesis

The intern will contribute to improvement of the R package [triangulate](https://github.com/mcguinlu/triangulate) (<https://github.com/mcguinlu/triangulate>). triangulate is a tool that can be used in evidence synthesis to implement quantitative triangulation. Triangulation is an approach for strengthening causal inference by integrating evidence from multiple sources. Most studies using triangulation have qualitatively examined whether different studies agree upon the presence of a causal effect, rather than estimating the effect by quantitatively integrating results. triangulate systematically integrates assessments from risk-of-bias tools and adjusts for bias and relevance concerns from multiple study designs using prior distributions, then pools the adjusted estimates using standard meta-analysis.

The project goal for the intern is to make triangulate more user friendly and accessible by adding tutorials and visualisation tools to help future applied researchers to apply quantitative triangulation methods. The framework for triangulate is applicable to any research in health sciences that seeks to estimate the magnitude of the causal relationship. Quantifying relationships is important for several reasons, including the identification of safe levels of an exposure, and the development of interventions that might achieve worthwhile gains in health. With an ever-growing number of studies, triangulation will become more important, and thus robust methods are crucial. Our framework also provides guidance for the process of identifying and addressing biases and differences in relevance which will enable better practices in future public health research. triangulate uses published summary statistics and does not require individual-level data.

On completion, the intern will be given credit to contributing to *triangulate* via “collaborator” status in *GitHub,* which is an online platform where researchers can upload their analysis code or software for open access and sustainability.The undergraduate will learn how their coding skills are essential to work as a methodologist in health research.