Statistical and Machine learning Approaches to Predicting Stroke Risk

- Stroke is a leading cause of mortality and long-term disability worldwide causing substantial burden to the NHS and massively impacting lives
- Early identification of risk reduces societal cost and enables preventative intervention, saving lives
- Growing availability of clinical, demographic, and lifestyle data creates opportunities for improving risk prediction accuracy
- Available research presents mixed evidence about whether complex machine learning models exceed previous classical statistical methods for stroke prediction, especially under class imbalance

Research Questions

1) Can routinely collected clinical, demographic, and lifestyle variables be used as accurate predictors for stroke occurrence?

2) How does predictive performance compare between:

- Classical statistical models
- Machine-learning models
- Deep-learning models ?

3) How do class imbalance and probability threshold selection affect model evaluation and clinical usefulness?

DATA OVERVIEW

The analysis uses a publicly available healthcare dataset containing information about **5110** individuals, accessed from an online repository.

Dataset includes:

demographic variables (age, gender);

clinical indicators (hypertension, heart disease, average glucose level, BMI);

lifestyle or socioeconomic factors (smoking status, work type, marital status, and residence type);

all alongside a binary stroke outcome

- Stroke occurrence is rare, with only **249** positive cases (≈4.9%). This results in a highly imbalanced classification problem
- As the data are observational rather than experimental, associations identified by the models should be interpreted as **predictive** rather than causal
- Missing BMI values were present and addressed during preprocessing, and a small number of sparse categories were identified as a potential limitation

METHODOLOGY

To address the research question, a supervised classification framework was adopted

Steps:

- Data cleaning
- Preprocessing
- Dataset split into **training (80%)** and **test (20%)** using stratified sampling (this preserves class imbalance)
- Multiple model approaches applied to predict stroke occurrence from the available predictors

Models were trained on the same feature set to ensure comparability, and evaluated on unseen test data

Model performance was assessed using **threshold-independent metrics** (e.g. ROC-AUC, precision–recall AUC) and **confusion-matrix-based measures** (e.g. sensitivity, specificity)

Given the medical screening context, sensitivity and evaluating alternative probability thresholds were prioritized rather than relying solely on default classification cut-offs

COMPARISON MODELS

- Primary models included classical statistical methods and more flexible machine-learning models
- **Logistic regression** used as a baseline due to its interpretability and widespread use in clinical risk prediction
- LASSO-penalised logistic regression was included to assess benefits of regularisation and automated feature selection

These models were compared with:
- **a random forest classifier** which can capture non-linear effects and interactions
- **a feed-forward neural network** representing a simple deep-learning approach for tabular data

Comparing these approaches allows assessment of whether increased model complexity leads to meaningful improvements in discrimination and clinical utility over standard statistical methods.