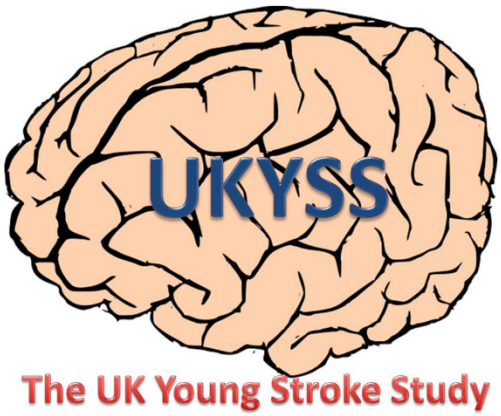

UKYSS

Statistical Analysis Plan (from UKYSS protocol version 1.3 – 8th July 2025)

Sample size calculation

A total sample size of 2466 strokes was calculated. A sample size of 1996 was calculated to produce a prediction model for recurrent ischaemic stroke at 5 years using 10 parameters . This was calculated based on the cohort from Pezzini et al (1) where recurrent events at a median follow up of 42 months were 86, with a follow up of 1867 ischaemic stroke patients equating to 86491 person-months, giving an event rate of 0.00099. The reported C-statistic for their model's performance at 1- and 5-year predictions were 0.62 and 0.67 respectively, however the Pezzini et al. model's performance is unlikely to be representative of the potential performance of a new model in the current study setting. The Pezzini et al. cohort had a lower upper limit of age and had a homogenous population enrolling only Caucasian patients, both of which limit the spread of case-mix in their population, and so we would expect different performance for the new model development study which has no such exclusion criteria. We therefore assume a conservative estimate of 15% expected explained variation (i.e. Nagelkerke's R-squared = 0.15) for the new model, which corresponds to an adjusted Cox-Snell R squared value of 0.044.

A sample size of 470 was calculated to produce a prediction model for haemorrhagic stroke mortality at 5 years using 5 parameters based on long term mortality data produced by Ekker et. al. (2), where mortality after a mean follow up of 10.16 years was 349, with a follow up of 2086 subjects equating to 21194 person-years, giving an event rate of 0.0165. Given that no previous prediction model has been created for this outcome in this patient group, a conservative value of 15% expected explained variation (i.e. Nagelkerke's R-squared = 0.15) was assumed for the new model's performance.

Sample size estimations were calculated using pmsampsize software in Stata, which implements the approach of Riley et al . (3,4) to derive the minimum sample size required for developing a multivariable new prediction model.

Statistical analysis plan

Appropriate descriptive analysis (e.g. means and standard deviations for continuous data, frequencies, and percentages for binary data) will be calculated for all study variables. Incidence of stroke will be calculated using first ever strokes as the numerator and the combined catchment population of each participating centre as the denominator. Chi-squared and Fisher Exact tests (if values <5) will be used to compare categorical variables across groups. The Student t test will be used to compare means between groups with the Mann U Whitney in instances where the data is not normally distributed. Kaplan-Meier curves will be constructed to measure the cumulative incidences for recurrent stroke, mortality, composite of other vascular events post stroke epilepsy and post stroke cancer. A Cox proportional hazards model will be fitted and hazard ratios calculated to estimate the prognostic value of individual risk factors which will be adjusted for other confounding factors .

We will develop and internally validate a multivariable prognostic model for individual outcome (risk) prediction, for stroke recurrence and mortality. To reduce concerns of overfitting during model development, we will adhere to sample size recommendations and use a set of candidate predictors defined a priori based on a combination of existing evidence of prognostic importance, clinical judgement, and availability at the point of prediction. We will follow best practice in model development and validation including not categorising continuous predictors, examining potential complex non-linear effects using splines and fractional polynomials, using multiple imputation to handle missing values, and accounting for competing risks where necessary (5). To further reduce the potential for overfitting, we will use the entire study sample for model development, as opposed to using sub-optimal data splitting approaches. We will perform internal validation of the model using bootstrap resampling, to check stability of included predictors, adjust predictors for optimism, and to produce optimism-adjusted model performance measures (including both measures of model calibration and discrimination). To examine potential clinical value of using the model, we will apply decision curve analysis, which examines the net-benefit of using the model (over a range of thresholds of risk which dictate clinical action) compared to other strategies, such as a treat all or treat none. (6) Potential external validation datasets will be sought by contacting the Chief Investigators of previously published young stroke cohorts. Our model will be reported as per TRIPD+AI reporting guidelines. (7) A separate protocol with statistical analysis plan will be produced for the prediction model and validation methods which will be written prior to undertaking this objective.

References

1. Pezzini A, Grassi M, Lodigiani C, Patella R, Gandolfo C, Zini A, et al. Predictors of long-term recurrent vascular events after ischemic stroke at young age: the Italian Project on Stroke in Young Adults. Circulation. 2014;129(16):1668-76.

2. Ekker MS, Verhoeven JI, Vaartjes I, Jolink WMT, Klijn CJM, de Leeuw FE. Association of Stroke Among Adults Aged 18 to 49 Years With Long-term Mortality. Jama. 2019;321(21):2113-23.

3. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. Statistics in medicine. 2019 Mar 30;38(7):1276-96.

4. Riley, RD, Ensor, J, Snell, KIE, et al. Calculating the sample size required for developing a clinical prediction model. Br Med J 2020; 368: m441.

5. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ. 2024 Sep 3;386:e078276. doi: 10.1136/bmj-2023-078276. PMID: 39227063; PMCID: PMC11369751.

6. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019 Oct 4;3:18. doi: 10.1186/s41512-019-0064-7. PMID: 31592444; PMCID: PMC6777022.

7. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024 Apr 18;385:q902. doi: 10.1136/bmj.q902. Erratum for: BMJ. 2024 Apr 16;385:e078378. doi: 10.1136/bmj-2023-078378. PMID: 38636956; PMCID: PMC11025451.