



BMJ Open FAST-IT: Find A Simple Test – In TIA (transient ischaemic attack): a prospective cohort study to develop a multivariable prediction model for diagnosis of TIA through proteomic discovery and candidate lipid mass spectrometry, neuroimaging and machine learning – study protocol

Austin G Milton ¹, Stephan Lau ^{2,3}, Karlea L Kremer ⁴,
Sushma R Rao ^{5,6}, Emilie Mas ^{6,7}, Marten F Snel ^{5,6}, Paul J Trim ^{5,6},
Deeksha Sharma ^{3,4}, Suzanne Edwards ⁸, Mark Jenkinson ^{2,3},
Timothy Kleinig ^{6,9}, Erik Noschka ^{4,10}, Monica Anne Hamilton-Bruce ^{1,4},
Simon A Koblar ⁴

To cite: Milton AG, Lau S, Kremer KL, *et al*. FAST-IT: Find A Simple Test – In TIA (transient ischaemic attack): a prospective cohort study to develop a multivariable prediction model for diagnosis of TIA through proteomic discovery and candidate lipid mass spectrometry, neuroimaging and machine learning—study protocol. *BMJ Open* 2022;**12**:e045908. doi:10.1136/bmjopen-2020-045908

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-045908>).

MAH-B and SAK are joint senior authors.

Received 19 October 2020
Accepted 26 January 2022



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Austin G Milton;
austin.milton@adelaide.edu.au

ABSTRACT

Introduction Transient ischaemic attack (TIA) may be a warning sign of stroke and difficult to differentiate from minor stroke and TIA-mimics. Urgent evaluation and diagnosis is important as treating TIA early can prevent subsequent strokes. Recent improvements in mass spectrometer technology allow quantification of hundreds of plasma proteins and lipids, yielding large datasets that would benefit from different approaches including machine learning. Using plasma protein, lipid and radiological biomarkers, our study will develop predictive algorithms to distinguish TIA from minor stroke (positive control) and TIA-mimics (negative control). Analysis including machine learning employs more sophisticated modelling, allowing non-linear interactions, adapting to datasets and enabling development of multiple specialised test-panels for identification and differentiation. **Methods and analysis** Patients attending the Emergency Department, Stroke Ward or TIA Clinic at the Royal Adelaide Hospital with TIA, minor stroke or TIA-like symptoms will be recruited consecutively by staff-alert for this prospective cohort study. Advanced neuroimaging will be performed for each participant, with images assessed independently by up to three expert neurologists. Venous blood samples will be collected within 48 hours of symptom onset. Plasma proteomic and lipid analysis will use advanced mass spectrometry (MS) techniques. Principal component analysis and hierarchical cluster analysis will be performed using MS software. Output files will be analysed for relative biomarker quantitative differences between the three groups. Differences will be assessed by linear regression, one-way analysis of variance, Kruskal-Wallis H-test, χ^2 test or Fisher's exact test. Machine learning methods will also be applied including deep learning using neural networks.

Strengths and limitations of this study

- This prospective cohort study of transient ischaemic attack (TIA), minor stroke and TIA-mimics recruits patients through the main capture paths in the largest public hospital in South Australia at which TIA patients usually arrive, providing a representative hospital sample population.
- Find A Simple Test—In TIA uses mass spectrometry discovery proteomics with candidate lipid analysis, a combination not used in previous studies, with the potential to identify novel biomarkers of TIA and minor stroke, and to differentiate these from TIA-mimics.
- Machine learning will also be applied to interpret the quantitative biomarkers, clinical and demographic data to further distinguish patterns for the development of more powerful diagnostic algorithms.
- Advanced radiological imaging, increasingly used in treatment decisions, will be both included and excluded in the machine learning models to explore development of algorithms appropriate not only for advanced centres with these capabilities, but also for rural and remote regions without radiological imaging capacity.
- Although a limited number of patients may be excluded due to inability to give informed consent, the planned sample size provides adequate power to apply conventional statistical analysis and predict diagnosis.

Ethics and dissemination Patients will provide written informed consent to participate in this grant-funded study. The Central Adelaide Local Health Network Human



Research Ethics Committee approved this study (HREC/18/CALHN/384; R20180618). Findings will be disseminated through peer-reviewed publication and conferences; data will be managed according to our Data Management Plan (DMP2020-00062).

INTRODUCTION

Transient ischaemic attack (TIA) is defined as a transient episode of neurological dysfunction caused by focal brain, spinal cord or retinal ischaemia without acute infarction (tissue death due to inadequate blood supply).¹ Following TIA, there is a high risk of stroke and a large proportion of strokes occur in the first 48 hours following TIA.² Correct diagnosis of TIA and early treatment reduces the subsequent risk of stroke by up to 80%.³

Stroke is a major cause of death and adult disability globally, and up to 23% of strokes are preceded by TIA.⁴ The diagnosis of TIA is based on clinical presentation but is highly subjective with disagreement of diagnosis between trained neurologists.⁵ This variability is due partly to the myriad of possible clinical presentations depending on which areas of the brain are affected.^{6,7} A number of mimic conditions exist including migraine and seizures.⁸ The heterogeneity and transience of symptoms, often absent on presentation, creates a heavy reliance on patient history, which may be distorted by poor observation of the symptoms or difficulties in recalling the event (recall bias), making the diagnosis of TIA imprecise.⁹

Numerous clinical studies have clearly demonstrated that even among neurologists a diagnosis of TIA has at best 70% agreement, depending on the measurement scales used.⁵ Following neurological review and investigation, from 30% to 60% of patients referred to TIA clinics may be diagnosed as having other causes including migraine, seizure, psychiatric disturbance, peripheral vertigo, presyncope or a metabolic condition.^{8,9} It would be of major clinical significance to be able to differentiate TIA from TIA-mimic conditions. A plasma biomarker for the diagnosis of TIA would provide an objective measure to differentiate TIA from TIA-mimics rapidly. Enhanced diagnostic accuracy would allow us to take early preventative measures against stroke and reduce harm to the patient from misdiagnosis and subsequent unsuitable management. An ideal biomarker or biomarker panel would be rapidly measurable, reproducible, reliable and accurate¹⁰ and have a scientifically plausible association with TIA. It would also have high sensitivity and specificity, could be efficiently implemented clinically, and would provide cost-effective benefit when incorporated into diagnostic algorithms. An example of such a biomarker, although cardiac, that optimises diagnostic ability is the use of troponin to diagnose acute heart attack rapidly.¹¹

Mass spectrometry

Proteins, lipids and panels of these are important biomarkers of disease. With recent improvements in proteomic and lipidomic analysis technology, it is now possible to quantify hundreds of proteins and lipids from plasma in large numbers of patients, thus opening up

the possibility of discovering new circulating biomarkers and developing a diagnostic tool.^{7,12} Mass spectrometry (MS)-based proteomic and lipidomic assays already play a prominent role in the diagnosis of disease conditions. Of note is the MS-based amyloid-typing assay that has been accredited clinically and is routinely used to differentiate between the various types of amyloidosis. Amyloid typing involves identifying amyloid-associated proteins and the core constituents of fibrils to enable diagnostic characterisation for the correct treatment of systemic amyloidosis.^{13,14}

Many candidate blood biomarkers for TIA have been investigated but none have been used in routine clinical practice due to inherent study or protocol limitations.¹⁵ Discovery research using more sensitive high-throughput MS assays allows improved identification of plasma proteins.¹⁶ Standard methods of analysis include peptide/protein identification of MS/MS spectra by database searches, protein and lipid quantitation in combination with multivariate statistical analysis and conventional descriptive statistical analyses using protein and lipid abundance data. Liquid chromatography-tandem MS (LC-MS/MS) has relatively short run times, higher sensitivity and selectivity and has been shown to be powerful for simultaneous measurement, for example, isoprostane isomers which are biologically active lipids that appear to be indicative of oxidative stress in stroke.¹⁷ While accurate isoprostane quantities at extremely low concentrations are commonly measured in plasma by well-established and sensitive gas chromatography-MS (GC-MS) or ELISA, the procedure is considered tedious and time-consuming.¹⁸⁻²⁰

The inclusion of machine learning in stroke and TIA can apply more sophisticated modelling, allowing for non-linear interactions, while adapting to the nature of the dataset used, enabling the development of multiple specialised test panels for identification and differentiation.

Machine learning

The application of machine learning techniques is growing rapidly and recent applications of both classical machine learning techniques and neural networks have been reported in the area of TIA and minor strokes.²¹⁻²³ These have shown promising results for a range of clinical and radiological data from small to medium-sized datasets.^{24,25} They have been applied for a range of different purposes (eg, risk stratification, screening and short-term mortality prediction) and indicate how flexible the machine learning approach is and how much it has to offer in the study and treatment of TIA.

Depending on the size and complexity of the data, different machine learning methods will have different advantages and disadvantages, especially when it comes to robust generalisation. Classical machine learning techniques such as random forests, support vector machines, Orthogonal Projections to Latent Structures (OPLS)²⁶ and Canonical Correlation Analysis (CCA)²⁷ are evaluated alongside neural networks, which span the realms of

small to huge datasets and are at the core of deep learning applications. Given the nature and size of the data, it is important to cover a suitably wide range of techniques to ensure the greatest chance of finding one that best suits the available data and problem, while carefully employing cross-validation methodology to ensure unbiased performance estimates and maximise the generalisability of the outcome. This is important for future applications to data from different hospitals, countries and demographics.

Preliminary work and new study

An earlier pilot project in our laboratory tested a small number of patient plasma samples and found several likely candidate TIA proteins including lipid-binding proteins and blood-clotting factors that appeared to be differentially expressed in TIA patients.^{28 29} We now propose to use large datasets with a global biomarker discovery proteomic approach and also include candidate lipid profiling, using MS as a sensitive, reproducible and robust tool for identification of TIA biomarkers and patterns. Our investigations are distinguished from previous TIA biomarker research in that we also examine lipid biomarkers of the isoprostane family and include the use of machine learning to further analyse our datasets, including neuroimaging data, to distinguish patterns and develop diagnostic algorithms.

Study objectives

Primary objective

The primary objective of this study is to identify plasma protein, lipid and/or radiological biomarkers that change in response to a TIA, assessing these diagnostic methods using a range of techniques including formal statistical analysis and machine learning.

Secondary objective

To develop an algorithm of the most significant biomarkers for use as a diagnostic tool to distinguish TIA from minor stroke and TIA-mimics.

METHODS AND ANALYSIS

Find A Simple Test-In TIA (FAST-IT) is a prospective cohort protein discovery and candidate lipid biomarker study using MS to identify novel TIA and minor stroke biomarkers, and incorporating radiology. Machine learning is applied in addition to conventional statistical analysis, in order to explore development of algorithms appropriate not only for advanced neuroimaging centres, but also for rural and remote regions without radiological imaging capacity.

The design, conduct and reporting of this study are guided by the Strengthening the Reporting of Observational Studies in Epidemiology³⁰ and Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis³¹ statement checklists. Project collection started on 23 July 2019 and will run through to the end of 2024.

Participants

We will recruit patients consecutively by staff-alert from the Royal Adelaide Hospital (RAH) Emergency Department, Stroke Unit or TIA Clinic, aiming for a minimum of 518 patients. These will be distributed between three groups with up to three expert vascular neurologists in diagnostic agreement: (1) a TIA study group, (2) a minor stroke (clinically definite, positive control) group with transient neurological symptoms and (3) a negative control group of patients with transient neurological symptoms, being TIA-mimics. We aim to enrol mimic and minor stroke cohorts that are age, gender and vascular risk-matched to the presenting TIA group. A standard control group with the absence of neurological symptoms is not included in the study as the TIA-mimic control group provides our comparative clinical standard. Staffing limitations and temporary halts during COVID-19 restrictions have interrupted this consecutive collection. We will exclude patients who are outside a 48-hour window since the start of their medical episode (TIA, minor stroke or TIA-mimic).

Inclusion criteria

- ▶ Patients who attend either the RAH Emergency Department or Stroke Ward or TIA Clinic with TIA or minor stroke or TIA-like symptoms.

Exclusion criteria

- ▶ Pregnancy or age <18 years.
- ▶ Patients with haemorrhagic stroke (intracerebral or subarachnoid haemorrhage).
- ▶ Patients without imaging studies (non-contrast CT and CT-angiogram or MRI scans).
- ▶ Patients who are outside the 48-hour window since their episode.
- ▶ Unable to give informed consent or inability to comprehend English (we will not engage translators for non-English-speakers).
- ▶ Patients who are highly dependent on medical care who may be unable to give consent.

Clinical assessment

Demographic and clinical data will be collected at the Emergency Department, Stroke Unit or TIA clinic at the time of presentation. Basic demographic data includes age, gender, ethnic background, smoking, alcohol history and diabetes history. Clinical data from initial assessment includes medical history (presenting symptoms, cardiovascular risk factors and medications), modified Rankin Score (mRS)³² and ABCD2 stroke-risk stratification score.³³ Diagnosis of TIA will be performed independently by two experienced expert vascular neurologists from the Stroke Unit at the RAH on the basis of clinical presentation, radiological imaging (baseline non-contrast CT and CT-angiogram and/or multi-modal MRI (diffusion, T1, T2 and MR angiography)) and the neuroradiologist's report, using the RAH Stroke Unit's Code Stroke diagnostic protocol.³⁴ Routine blood biochemistry results

and radiological data (non-contrast CT and CT-angiogram and/or MRI scans) will be collected to assist diagnosis. Follow-up data are collected at or shortly after 3 months (accepting occasional logistical limitations) by telephoning the patient and using a scripted telephone survey to determine their mRS status in their recovery.³⁵

Radiological assessment

All patients will undergo, as a minimum, baseline non-contrast CT and CT-angiogram (from the aortic arch to vertex) on admission, followed by multi-modal MRI (diffusion, T1, T2 and MR angiography) within the first 48 hours of their event, with acceptance of occasional logistical limitations. Not all patients will have an MRI because of contraindication (ie, known contrast allergy, unstable or rapidly deteriorating renal function) and occasional resource time constraints, but approximately 95% of our patients will have both CT/CTA and MRI. Images will be achieved digitally and viewed on local servers by an expert neuroradiologist and vascular neurologist. The imaging is reported by the radiologist and reviewed by the vascular neurologist with reference to a senior neuroradiologist as required. We will assess the standard radiological measures for infarction or vascular abnormalities at the presumed site of the lesion, which provide aetiological evidence for a possible TIA, minor stroke or TIA-mimic, allowing for recruitment into FAST-IT subject to informed consent.

Final diagnostic group classification

For our research analysis purposes, a panel of two experienced vascular neurologists will subsequently independently assess the images of all patients consented in the FAST-IT study as well as all available clinical data including follow-up data if available and the neuroradiologist's report for the final study classification as either TIA, minor stroke or TIA-mimic, with any disagreement resolved by a third, expert senior vascular neurologist (TK), the Head of the Stroke Unit at the RAH, to obtain consensus. The endorsed definition of TIA is per the AHA/ASA Scientific Statement: 'TIAs are transient episodes of neurological dysfunction caused by focal brain, spinal cord, or retinal ischaemia, without evidence of acute infarction'.¹ Minor stroke is diagnosed per the Stroke Unit's Code Stroke diagnostic protocol with a National Institutes of Health Stroke Scale score less than or equal to 3.^{34 36} Minor stroke can be either imaging-proven stroke or imaging negative stroke but with persisting neurological deficit (for practical purposes greater than 24 hours).^{34 36} The final group classification of TIA versus TIA-mimic is based on these studies, all available clinical data and expert opinion.

Blood sample collection, preparation and analysis

Sample collection and preparation

Venous blood samples will be collected within 48 hours of symptom onset into ice-chilled plastic blood-collection tubes containing EDTA in accordance with Human Proteome Organisation Proteomics Standard

Initiative standard protocols.³⁷ Samples will be inverted 10×, placed on ice and transferred to the laboratory immediately. Subsequent steps will be performed at 4°C or on ice with all samples handled with gloved hands. Samples will be centrifuged at 1300g for 10 min (Heraeus Megafuge 40R, Thermo Scientific, USA) to separate the plasma. Low-protein-bind pipette tips and tubes (Eppendorf, Australia) will be used for all sample handling and storage. Plasma supernatant will be collected into a single 5 mL tube and then stored as 500 µL aliquots in five or more separate tubes (marked with sample number+'A' to 'E'). Remaining blood and plasma will be sealed and discarded appropriately. All samples will be stored at -80°C in designated freezer storage. For every sample, we record the time of collection postevent and the time before sample freezing. Subsequent laboratory analysis for plasma proteomics and lipids will use only deidentified sample numbers; laboratory scientists are blinded to any group classification.

Plasma proteomics MS analysis and data acquisition

A tube containing 500 µL of patient plasma sample will be thawed on ice and used for MS analysis. Proteins will be precipitated with ice-cold acetone using 10 µL of sample, then denatured, reduced and alkylated before digestion with trypsin to generate peptides. Resulting peptides will be processed and analysed by MS.

- ▶ Peptide samples will be analysed by nano-liquid chromatography (nLC) using a Dionex Ultimate 3000 RSLCnano system (ThermoFischer Scientific, USA) coupled online to a timsTOF (trapped ion mobility spectrometry Time of Flight) *Pro* mass spectrometer (Bruker Daltonics, Germany).
- ▶ Proteomic analysis using MS yields large amounts of spectral data that need to be de-coded reliably to extract protein abundance information. Peak picking and interpretation of mass spectra are complex and time-consuming, but there are several software packages that automate and expedite this process. We will use Peaks X+ (Bioinformatics Solutions Inc, Waterloo, Ontario, Canada) and MaxQuant (Max-Planck-Institute of Biochemistry) for raw data processing and databank searching combined with Perseus for basic multivariate statistical analyses between sample groups (Max-Planck-Institute of Biochemistry).³⁸⁻⁴² We will conduct differential protein abundance analysis to identify proteins or panels of proteins that are differentially expressed between patient groups.
- ▶ Data will be searched against the UniProt Homo sapiens reference proteome (<https://www.uniprot.org/>). Only proteins with a false discovery rate of ≤1% will be reported. Principal component analysis and hierarchical cluster analysis will be performed using Perseus software. Output files will also be analysed for relative protein quantitative differences between the three patient groups.

Plasma isoprostane LC-MS/MS analysis and data acquisition

A modified method from Dupuy⁴³ and Sánchez-Illana⁴⁴ will be followed. A tube containing 500 µL of plasma sample will be thawed on ice and used for MS analysis. Conjugated isoprostanes will be base-hydrolysed prior to extraction. To thawed plasma we will add BHT (1% v/v in methanol), internal standard mixture (1000 ng/mL), potassium hydroxide solution (1 M) and, when preparing a standard curve for quantitation, standards mix of varying concentrations. Tubes will be vortexed briefly, incubated at 40°C for 30 min and centrifuged at 1300g for 5 min. Supernatant will be loaded onto a preconditioned SPE cartridge (Bond Elut Plexa PAX 60 mg, 3 mL, Agilent) inserted in a suitable SPE vacuum manifold (24-port VISIPREP, Supelco). Conditioning consists of washing the cartridge with methanol followed by water. Loaded extract will be washed with water, then methanol, then hexane/ethyl acetate (75:25) and briefly suction dried (2 min). Sample elution into 10 mL glass test tubes will be with hexane/ethyl acetate (25:75) and methanol, both acidified with formic acid (5%). Eluant fractions will be combined, dried under nitrogen and reconstituted for UPLC (ultra-performance liquid chromatography)-MS/MS analysis.

- ▶ The following lipid and isoprostane standards will be used; 5-F2t-IsoP; 2,3-dinor-15-F2t-IsoP; 4(RS)-F4t-NeuroP; D4-10-F4t-NeuroP; 14(RS)-14-F4t-NeuroP; 20-F4t-NeuroP; 17-F2t-dihomo-IsoP; Ent-7(RS)-F2t-dihomo-IsoP; 7(RS)-ST-Δ8-11-dihomo-IsoF; 5-F3t-IsoP; 8-F3t-IsoP and 18-F3t-IsoP—all gifted by Dr Thierry Durand, Institut Des Biomolécules Max Mousseron (IBMM, France). 8-Iso PGf2α-d4; 5-iPF2α-Vi-d11 and 8-iso PGf2α were purchased from Cayman Chemical (Ann Arbor, Michigan, USA).
- ▶ UPLC will be performed on the Sciex Exion-LC UPLC system, consisting of an AD Multiplate autosampler, Exion-LC Degassing unit, Communications Bus module, Column Oven and x2 UPLC Binary Pumps. Samples will be kept at 4°C in the autosampler unit. Separation is performed with an Agilent Poroshell 120 EC-C18 column with a flow rate of 400 µL/min and column oven temperature of 50°C. Mobile Phase-A (MP-A) consists of 5% of acetonitrile while Mobile Phase-B (MP-B) consists of 95% of acetonitrile. Both solvents include formic acid and ammonium formate. Runs will start at 25% MP-B and will be held for 20 s then ramped to 50% MP-B for 6 min then further to 65% for 6.5 min; then wash at 100% MP-B from 6.8 to 8.5 min. The column will be equilibrated at 25% between sample injections.
- ▶ MS will be performed with the 6500+ Triple Quadrupole Mass Spectrometer (Applied Biosystems Sciex, Australia) using Analyst Software. Ionisation will be performed by an Electrospray Ionisation in negative ion mode. MS recordings are conducted in multiple reaction monitoring scan mode. Duration of data

acquisition will be 10 min. Data will be analysed with MultiQuant Software.

Data monitoring body

A Data Management Plan (DMP2020-00062) was lodged with the University of Adelaide Research Grants Office on 31 October 2018; amended to include lipids on 6 August 2020.

Primary outcomes

- ▶ Identification of plasma protein, lipid and/or radiological biomarkers that change in response to a TIA and minor stroke.
- ▶ Development of a panel of the most significant plasma, lipid and radiological biomarkers of TIA and minor stroke for use as a diagnostic tool to distinguish TIA from minor stroke and TIA-mimics.

Secondary outcome

- ▶ Determine the best diagnostic method using a range of techniques including formal statistical analysis and machine learning.

Sample size

A two-sample pooled t-test of mean ratio with lognormal data was used for the sample size calculation since biomarker outcomes are likely to have a skewed distribution for which a logarithmic transformation would create normally distributed residuals. The coefficient of variation was estimated from preliminary work in our laboratory to have a value of 0.2.^{28 29} To detect a 10% change in biomarker abundance (mean1 is 10% greater than mean2 such that the mean ratio between the two groups is 1.1) with coefficient of variation 0.2, 80% power and alpha=0.05, each group requires 69 samples. The vascular neurologists will assign patients to one of the three groups based on all diagnostic information and the study will continue to recruit until there is a minimum of 69 participants in the smallest group. With three groups in our study, the total sample size will be a minimum of 3×69, that is, n=207 for adequate power to achieve statistical significance. We double this number to provide more data to train machine learning models and to compensate for any unforeseen deviations from our assumptions on the effect size and distribution. Allowing 20% to account for potential mismatches through recruitment, withdrawal or loss to follow-up, a minimum of 518 patients need to be enrolled. Statistical software: SAS 9.4 (SAS Institute).

Statistical analysis and machine learning methods

Figure 1 gives an overview of the data flow for statistical analysis and machine learning.

Statistical analysis methods

Descriptive statistics of clinical and demographic variables will be presented in data tables by group (TIA, minor stroke or TIA-mimic). The mean and SD will be given for normally distributed continuous variables, median and IQR will be given for non-normally distributed

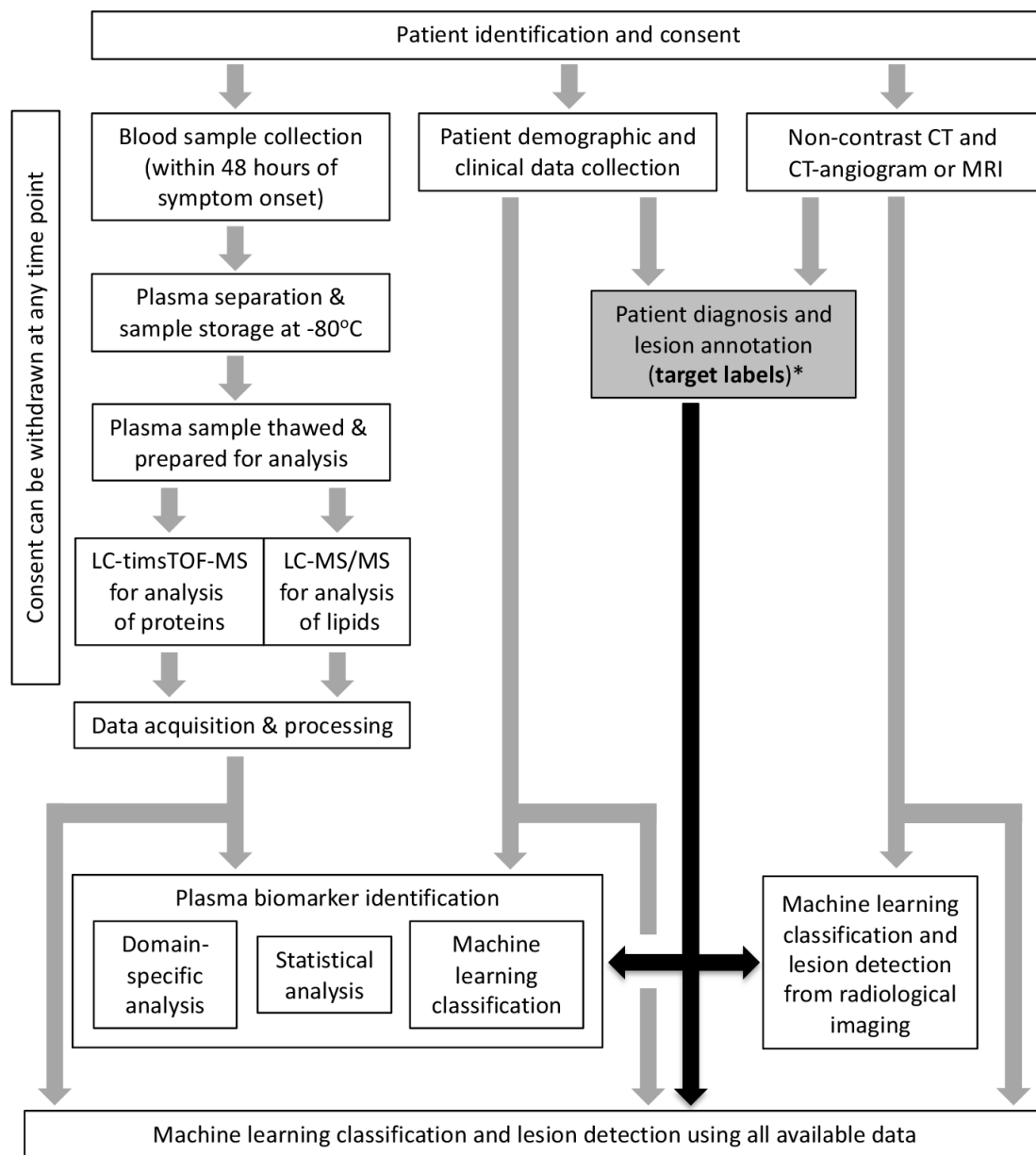


Figure 1 Data flow diagram. *Target labels are both the diagnostic groups and the labels on the images. Target label data flow is indicated by black arrows. LC-MS/MS: liquid chromatography-tandem mass spectrometry; LC-timsTOF-MS, liquid chromatography-trapped ion mobility spectrometry-time of flight-mass spectrometry.

continuous variables, and frequency and percentage will be given for categorical variables. A one-way Analysis of variance (ANOVA), Kruskal-Wallis H-test, χ^2 test or Fisher's exact test will be performed as appropriate to identify any significant differences in these variables between the three groups. We will also combine TIA and minor stroke as part of our analysis, as previously reported in Dolmans *et al.*¹⁵

As the biomarker variables are likely to have a skewed distribution, median plasma biomarker levels by group (and IQR) will also be presented in a data table. A univariate linear regression with logarithmic transformation of the outcome will be performed for each biomarker outcome, with the predictor being group: TIA, minor

stroke or TIA-mimic. Post-hoc comparisons will be calculated when the global p value is less than 0.05.

The multiple-comparison correction (eg, Bonferroni or similar) will take into account that the protein/lipid abundances are not truly independent variables. We will perform factor analysis on the data, determine the number of substantial independent factors and correct for this number.

If statistically significant linear regression results are found for a given biomarker, the receiver operating characteristic (ROC) curve can be used to determine each biomarker's use as a diagnostic accuracy tool. Logistic regression will estimate the probability of having TIA or minor stroke (or just TIA), with the logarithmic

transformation of the biomarker being included as a predictor in the model. The area under the curve (AUC), and specificity and sensitivity for predicting the correct diagnosis for each patient as well as calibration plots will be calculated for each biomarker. From this, the most appropriate cut-off levels for diagnosis can be found. Biomarker ROC curves can also be combined within the above model and graph. Sensitivity analyses including only imaging-proven stroke will also be performed.

Machine learning methods

While FAST-IT uses conventional statistics for optimal comparability, machine learning (classical and deep learning) will also be applied to interpret the quantitative demographic, clinical and biomarker (including protein, lipid and radiological imaging) data to further distinguish patterns for the development of more powerful diagnostic algorithms.

Three different applications of machine learning methods will be pursued, based on the primary input data: (1) proteomic and lipid data; (2) radiological imaging data and (3) proteomic, lipid and radiological imaging data. In each case, relevant demographic and clinical data will also be made available to the methods. For smaller or highly noisy datasets, methods based on smaller, less complex models are often advantageous as they have improved generalisation to new datasets and demographics. Therefore we will perform a model comparison across a range of methods, including random forests, support vector machines, OPLS,²⁶ CCA²⁷ and neural networks, in conjunction with (or without) dimensionality reduction methods such as principal component analysis and manifold learning. In order to avoid biases and improve performance through hyperparameter optimisation, suitable nested cross-validation will be performed in all cases. Performance of the methods will be assessed in the same way as for the statistical analyses, using ROC curves, the AUC metric and calibration plots along with additional reporting of accuracies and confusion matrices.

Another consideration for machine learning methods is the degree to which the model yields additional interpretable information. This is often a trade-off between the simpler, linear models that provide highly interpretable information versus more complex, non-linear models that are harder to interpret but often perform better. State-of-the-art methods, such as attention gates and saliency maps are available for probing the information captured by the models, thus moving away from a 'black box' model. This will be important in order to understand the nature of image-related features that are most relevant as well as the impact that the different proteins and lipids have in the discrimination and predictions made by these methods.

Missing data

During statistical analysis, participants with missing data for variables of interest in a given regression will be excluded only when that missing data would be required

for an analysis. With machine learning, small amounts of missing data will be labelled as such and handled by method-specific solutions. During deep learning, drop-out layers can be used to train neural networks to be missing-value tolerant. If a radiological or other biomarker section is missing, then this participant dataset will only be used in the analysis streams that do not require the missing data section. The number of missing values or data sections will be reported.

ETHICS AND DISSEMINATION

Project approved by the CALHN Human Research Ethics Committee (HREC/18/CALHN/384 on 19 October 2018) and Royal Adelaide Hospital (site-specific approval R20180618 on 17 April 2019); amendment for lipids approved by both on 1 May 2020. Signed informed consent will be obtained from all participants. Findings will be disseminated through peer-reviewed publication and conference presentations. Data (including deposition and curation) will be managed according to our Data Management Plan (DMP2020-00062), lodged with the University of Adelaide Research Grants Office on 31 October 2018 and amended to include lipid analysis on 6 August 2020.

Author affiliations

¹Stroke Research Programme, Central Adelaide Local Health Network, Adelaide, South Australia, Australia

²Faculty of Engineering, Computer and Mathematical Sciences, Australian Institute for Machine Learning, The University of Adelaide, Adelaide, South Australia, Australia

³South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia

⁴Adelaide Medical School, Stroke Research Programme, The University of Adelaide Faculty of Health and Medical Sciences, Adelaide, South Australia, Australia

⁵Proteomics, Metabolomics and MS-imaging Core Facility, South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia

⁶Adelaide Medical School, The University of Adelaide Faculty of Health and Medical Sciences, Adelaide, South Australia, Australia

⁷SA Pathology - Genetics and Molecular Pathology, Women's and Children's Hospital Adelaide, North Adelaide, South Australia, Australia

⁸Adelaide Health Technology Assessment, The University of Adelaide Faculty of Health and Medical Sciences, Adelaide, South Australia, Australia

⁹Department of Neurology, Royal Adelaide Hospital, Adelaide, South Australia, Australia

¹⁰School of Animal and Veterinary Sciences, The University of Adelaide, Adelaide, South Australia, Australia

Acknowledgements We thank our clinical collaborator at the Royal Adelaide Hospital, Vascular Neurologist Associate Professor Jim Jannes, Head of Neurology, CALHN, Adelaide, South Australia 5000, for initial study outline discussion and clinical guidance, and Dr Thierry Durand, Institut Des Biomolécules Max Mousseron (IBMM, France) for gifting isoprostane standards. We also thank Monique Pisaniello, who undertook her fourth-year Adelaide Medical School placement with the Stroke Research Programme, studying TIA biomarkers and protocols.

Contributors All authors contributed intellectually to the study design, research methodology and manuscript, guided by MAH-B and SAK; SAK also provided clinical leadership. MAH-B, AGM, KLK and SAK contributed to consent form development and processes. TK provided diagnostic criteria and pathways and contributed to study design. SRR, MFS and PJT designed the proteomics protocol section, EM and EN designed the lipid analysis protocol section. SL and MJ designed the machine learning protocol and contributed to study design. SE conducted power calculations and developed the conventional statistical analysis plan. DS undertook further intellectual research into methodology and laboratory procedures. All authors

reviewed the manuscript critically and approved the response to reviewers. MAH-B and SAK are joint senior authors.

Funding This study was supported by the Health Services Charitable Gifts Board: Waltham Estate grant numbers 104-05-47-06-18, 90-05-47-05-19, 62-05-47-05-20 and 63-05-47-05-20 and The Hospital Research Foundation: 'Cure for Stroke Australia' grant numbers C-PJ-03-2018 and C-PJ-01-C4S-2019. MAH-B, AGM, KLK, SAK, TK, JJ, MS and EN secured study funding.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Austin G Milton <http://orcid.org/0000-0002-3746-4138>
 Stephan Lau <http://orcid.org/0000-0002-5952-0516>
 Karlea L Kremer <http://orcid.org/0000-0002-6733-2865>
 Sushma R Rao <http://orcid.org/0000-0002-1773-9747>
 Emilie Mas <http://orcid.org/0000-0003-2848-9613>
 Marten F Snel <http://orcid.org/0000-0002-8502-7274>
 Paul J Trim <http://orcid.org/0000-0001-8734-3433>
 Deeksha Sharma <http://orcid.org/0000-0002-7775-6098>
 Suzanne Edwards <http://orcid.org/0000-0003-2074-1685>
 Mark Jenkinson <http://orcid.org/0000-0001-6043-0166>
 Timothy Kleinig <http://orcid.org/0000-0003-4430-3276>
 Erik Noschka <http://orcid.org/0000-0002-6058-3549>
 Monica Anne Hamilton-Bruce <http://orcid.org/0000-0002-5222-620X>
 Simon A Koblar <http://orcid.org/0000-0002-8667-203X>

REFERENCES

- Easton JD, Saver JL, Albers GW, *et al.* American Heart Association/American Stroke Association Stroke Council; Council on Cardiovascular Surgery and Anesthesia; Council on Cardiovascular Radiology and Intervention; Council on Cardiovascular Nursing; and the Interdisciplinary Council on Peripheral Vascular Disease. The American Academy of Neurology affirms the value of this statement as an educational tool for neurologists. *Stroke* 2009;40:2276–93.
- Johnston SC, Rothwell PM, Nguyen-Huynh MN, *et al.* Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. *Lancet* 2007;369:283–92.
- Rothwell PM, Giles MF, Chandratheva A, *et al.* Effect of urgent treatment of transient ischaemic attack and minor stroke on early recurrent stroke (EXPRESS study): a prospective population-based sequential comparison. *Lancet* 2007;370:1432–42.
- Gállego J, Muñoz R, Martínez-Vila E. Emergent cerebrovascular disease risk factor weighting: is transient ischemic attack an imminent threat? *Cerebrovasc Dis* 2009;27 Suppl 1:88–96.
- Castle J, Mlynash M, Lee K, *et al.* Agreement regarding diagnosis of transient ischemic attack fairly low among stroke-trained neurologists. *Stroke* 2010;41:1367–70.
- Brouns R, De Deyn PP. The complexity of neurobiological processes in acute ischemic stroke. *Clin Neurol Neurosurg* 2009;111:483–95.
- Penn AM, Bibok MB, Saly VK, *et al.* Verification of a proteomic biomarker panel to diagnose minor stroke and transient ischaemic attack: phase 1 of SpecTRA, a large scale translational study. *Biomarkers* 2018;23:392–405.
- Fonseca AC, Canhão P. Diagnostic difficulties in the classification of transient neurological attacks. *Eur J Neurol* 2011;18:644–8.
- Prabhakaran S, Silver AJ, Warrior L, *et al.* Misdiagnosis of transient ischemic attacks in the emergency room. *Cerebrovasc Dis* 2008;26:630–5.
- Jickling GC, Sharp FR. Biomarker panels in ischemic stroke. *Stroke* 2015;46:915–20.
- Christenson RH, Duh S-H, Apple FA, *et al.* Pivotal findings for a high-sensitivity cardiac troponin assay: results of the HIGH-US study. *Clin Biochem* 2020;78:32–9.
- Geyer PE, Holdt LM, Teupser D, *et al.* Revisiting biomarker discovery by plasma proteomics. *Mol Syst Biol* 2017;13:942–57.
- Lavatelli F, di Fonzo A, Palladini G, *et al.* Systemic amyloidoses and proteomics: the state of the art. *EuPA Open Proteom* 2016;11:4–10.
- Abildgaard N, Rojek AM, Møller HE, *et al.* Immunoelectron microscopy and mass spectrometry for classification of amyloid deposits. *Amyloid* 2020;27:1–8.
- Dolmans LS, Rutten F, Bartelink M-LEL, *et al.* Serum biomarkers in patients suspected of transient ischaemic attack in primary care: a diagnostic accuracy study. *BMJ Open* 2019;9:e031774.
- Sandow JJ, Infusini G, Dagley LF, *et al.* Simplified high-throughput methods for deep proteome analysis on the timsTOF pro. *bioRxiv* 2019:657908.
- Lorenzano S, Rost NS, Khan M, *et al.* Oxidative stress biomarkers of brain damage: hyperacute plasma F2-isoprostane predicts infarct growth in stroke. *Stroke* 2018;49:630–7.
- Haschke M, Zhang YL, Kahle C, *et al.* HPLC-atmospheric pressure chemical ionization MS/MS for quantification of 15-F2t-isoprostane in human urine and plasma. *Clin Chem* 2007;53:489–97.
- Liu X, Whitefield PD, Ma Y. Quantification of F(2)-isoprostane isomers in cultured human lung epithelial cells after silica oxide and metal oxide nanoparticle treatment by liquid chromatography/tandem mass spectrometry. *Talanta* 2010;81:1599–606.
- Janicka M, Kot-Wasik A, Paradziej-Łukowicz J, *et al.* Lc-Ms/Ms determination of isoprostanes in plasma samples collected from mice exposed to doxorubicin or tert-butyl hydroperoxide. *Int J Mol Sci* 2013;14:6157–69.
- Çelik G, Baykan Ömer K, Kara Y, *et al.* Predicting 10-day mortality in patients with strokes using neural networks and multivariate statistical methods. *J Stroke Cerebrovasc Dis* 2014;23:1506–12.
- Abedi V, Goyal N, Tsvigoulis G, *et al.* Novel screening tool for stroke using artificial neural network. *Stroke* 2017;48:1678–81.
- Chan KL, Leng X, Zhang W, *et al.* Early identification of high-risk TIA or minor stroke using artificial neural network. *Front Neurol* 2019;10:171.
- Pinto A, Mckinley R, Alves V, *et al.* Stroke lesion outcome prediction based on MRI imaging combined with clinical information. *Front Neurol* 2018;9:1060.
- Chauhan S, Vig L, De Filippo De Grazia M, *et al.* A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from MRI lesion images. *Front Neuroinform* 2019;13:53.
- Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemom* 2002;16:119–28.
- Miller KL, Alfaro-Almagro F, Bangerter NK, *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* 2016;19:1523–36.
- Djukic M, Milton AG, Hamilton-Bruce MA. Targeted peptide quantification of candidate plasma proteins to diagnose transient ischaemic attack (TIA). *Int J Stroke* 2014;9:P43.
- Djukic M. Proteomic investigations and biomarker discovery in transient ischaemic attack. PhD thesis, University of Adelaide library, 2017. Available: <http://hdl.handle.net/2440/112817> [Accessed 20 Sept 2021].
- von Elm E, Altman DG, Egger M, *et al.* The strengthening of reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007;370:1453–7.
- Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.
- Bonita R, Beaglehole R. Recovery of motor function after stroke. *Stroke* 1988;19:1497–500.
- Koton S, Rothwell PM. Performance of the ABCD and ABCD2 scores in TIA patients with carotid stenosis and atrial fibrillation. *Cerebrovasc Dis* 2007;24:231–5.
- Health SA. Government of South Australia. 'Stroke Management Procedures and Protocols' version 3.1, Clinical Guideline No.: CG002, 2019. Available: https://www.sahealth.sa.gov.au/wps/wcm/connect/ae53950047066243b403fc22d29d99f6/Clinical+Guideline+Stroke+Management_Procedures+and+Protocols_final+Oct14pdf?MOD=AJPERES&CACHE [Accessed 20 Sept 2021].
- Dennis M, Mead G, Doubal F, *et al.* Determining the modified Rankin score after stroke by postal and telephone questionnaires. *Stroke* 2012;43:851–3.
- Fischer U, Baumgartner A, Arnold M, *et al.* What is a minor stroke? *Stroke* 2010;41:661–6.
- Rai AJ, Gelfand CA, Haywood BC, *et al.* HUPO plasma proteome project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. *Proteomics* 2005;5:3262–77.

- 38 Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 2016;11:2301–19.
- 39 Tyanova S, Temu T, Sinitcyn P, *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 2016;13:731–40.
- 40 Tran NH, Rahman MZ, He L, *et al.* Complete de novo assembly of monoclonal antibody sequences. *Sci Rep* 2016;6:31730.
- 41 Tran NH, Zhang X, Xin L, *et al.* De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A* 2017;114:8247–52.
- 42 Tran NH, Qiao R, Xin L, *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods* 2019;16:63–6.
- 43 Dupuy A, Le Faouder P, Vigor C, *et al.* Simultaneous quantitative profiling of 20 isoprostanoids from omega-3 and omega-6 polyunsaturated fatty acids by LC-MS/MS in various biological samples. *Anal Chim Acta* 2016;921:46–58.
- 44 Sánchez-Illana Ángel, Thayyil S, Montaldo P, *et al.* Novel free-radical mediated lipid peroxidation biomarkers in newborn plasma. *Anal Chim Acta* 2017;996:88–97.