

SAMuEL-2: Stroke Audit with Machine Learning 2

Investigating variation in clinical decision-making with
explainable AI

Anna Laws, Kerry Pearn, and Michael Allen

University of Exeter Medical School, PenCHORD

November 2022

Outline

1 Background

- Stroke and treatment

2 Explainability of machine learning

- Simplify the model
- SHAP

3 Outcome modelling

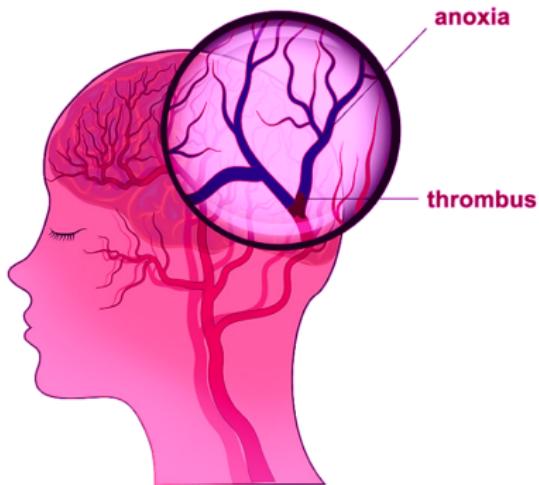
- Streamlit app

Background

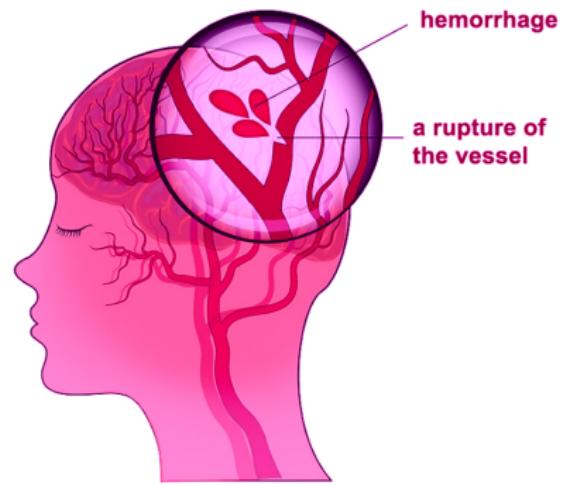
Background:

Stroke and treatment

Two types of stroke



Ischemic Stroke
(Infarction)



Hemorrhagic Stroke

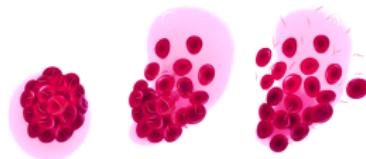
Brain scans reveal which type of stroke a patient has.

Treatments for ischaemic stroke

Thrombolysis aims to break down a clot by activating the body's own clot breakdown mechanisms.

Thrombolysis is given as an injection followed by an infusion (*drip*).

Thrombolysis is available at most hospitals.



Thrombectomy aims to grab and remove the blockage.

Thrombectomy uses a mesh device that enters the blocked blood vessel and physically removes the clot.

Thrombectomy is available at relatively few hospitals.



Downsides: not everyone is eligible for treatment; the treatments become less effective the later they are given; and there is a small chance of death.

SAMueL-1 focussed on thrombolysis use.

What is the problem?

There is a gap between target thrombolysis use (20%) and actual thrombolysis use (11–12%) in emergency stroke care

Clinical expert opinion on what *should be* happening



What is happening?



Unknown onset time or
arrived too late to treat



Not suitable for treatment
with thrombolysis



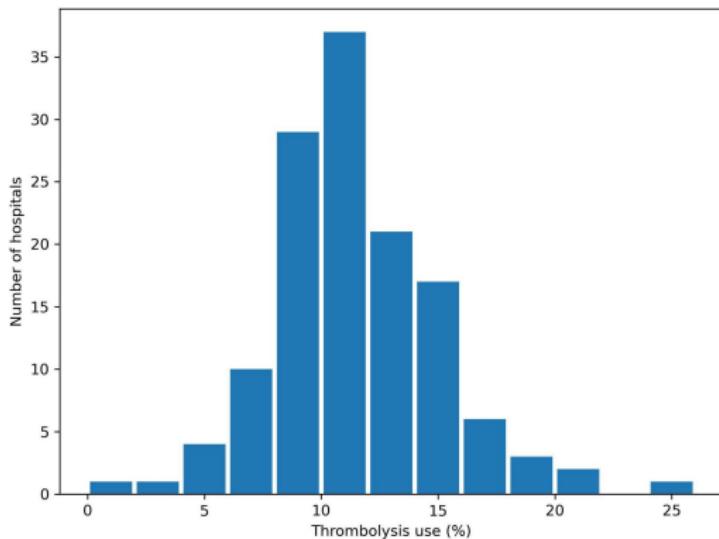
Treated with thrombolysis



Potentially treatable, but not
treated with thrombolysis

Thrombolysis rates in England and Wales have been stable at 11–12% for 10 years.

Use of thrombolysis varies considerably between hospitals



https://samuel-book.github.io/samuel-1/pathway_sim/explained_variance_in_thrombolysis.html

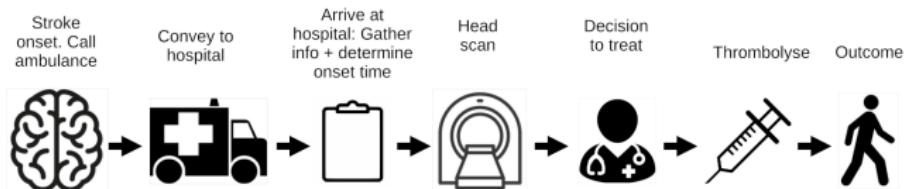
https://samuel-book.github.io/samuel-1/introduction/scientific_summary.html

A high level view of learning clinical decision-making

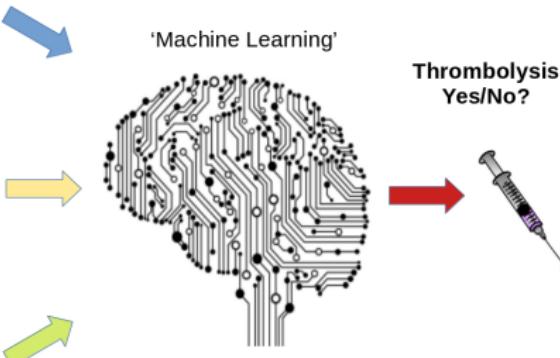
We accessed 240,000 emergency stroke admissions in England and Wales over three years.

Patient pathway:

The onset time might not be known precisely.



Hospital ID
Patient info, e.g. <ul style="list-style-type: none">• Age• Gender• Disability before stroke• Stroke symptoms• Co-morbidities• Medications• Known contra-indications
Pathway info, e.g. <ul style="list-style-type: none">• Time to arrival• Time to scan



Machine learning models include logistic regression, random forest, XGBoost, and neural networks.

Here we are using XGBoost models.

"What if?" questions from SAMueL-1

- What if arrival-to-treatment time was 30 minutes?
- What if all hospitals determined stroke onset time as frequently as an *upper quartile* hospital (ranked 25 out of 100 for determining stroke onset time).
- What if decisions to thrombolysis were made according to a majority vote of 30 benchmark hospitals?

For each hospital we use their own patients to ask these questions, to allow for differences in local patient populations.

We found that making all these changes would increase thrombolysis use in England and Wales to 18–19%. Out of every 10 patients who were potentially treatable but did not receive treatment, we found the cause to be:



Hospital processes were **too slow**

Stroke onset time was not determined when it potentially could have been



Doctors chose not to use thrombolysis when other higher-thrombolysing hospitals would have done



What questions are we asking in SAMueL-2?

What causes this variation in thrombolysis rates, and what could reasonably be achieved at each hospital (allowing for each hospital's own patient population)?

- Which patients do clinicians agree and disagree on, when considering when they should receive thrombolysis?
- How are the thrombolysis pathway and decision-making affected by *organisational factors* (such as use of specialist stroke nurses)?
- How best can we engage clinicians in our work, and prompt them to reconsider their emergency stroke pathway and/or decision-making?
 - Communication of general findings.
 - Web application for individual hospitals.
 - A 'hospital profile' for each hospital.

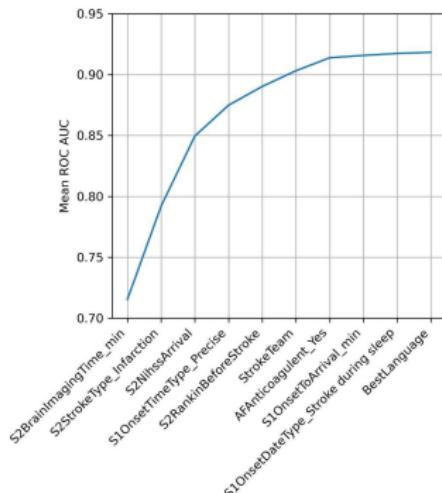
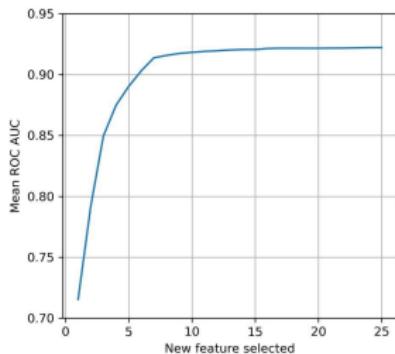
Explainability of machine learning

Explainability of machine learning:

Simplify the model

Simplifying the model with feature selection

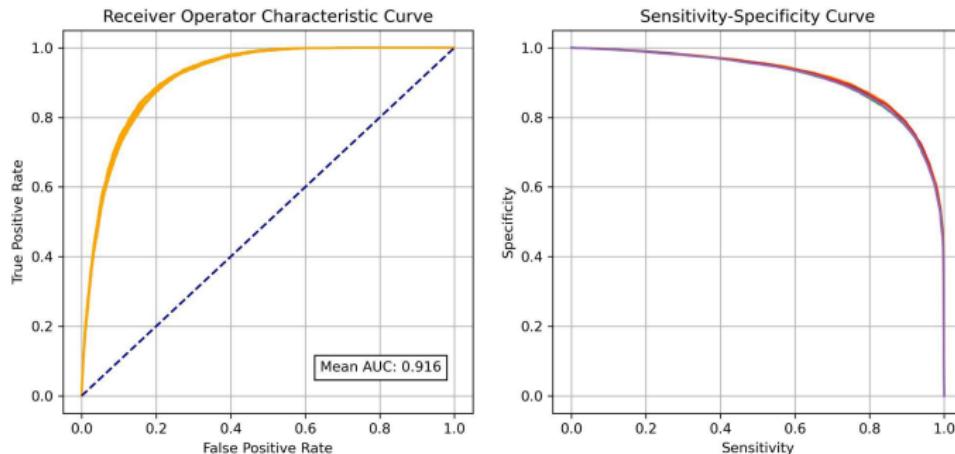
- Our full data set has 85 feature - but which are most important?
- A model with fewer features is easier to understand and explain
- We build features by selecting them one at a time according to which new feature adds most accuracy
- We measure accuracy with ROC-AUC (Receiver Operating Characteristic Curve - Area Under the Curve). But don't worry about that - it is just a robust accuracy measure.
- We find 8 features give us nearly as much accuracy as all features



Features selected:

- Arrival-to-scan time
- Infarction
- Stroke severity
- Precise onset time
- Prior disability level
- Stroke team
- Use of AF anticoagulants
- Onset-to-arrival time

Model accuracy after selection of 8 features



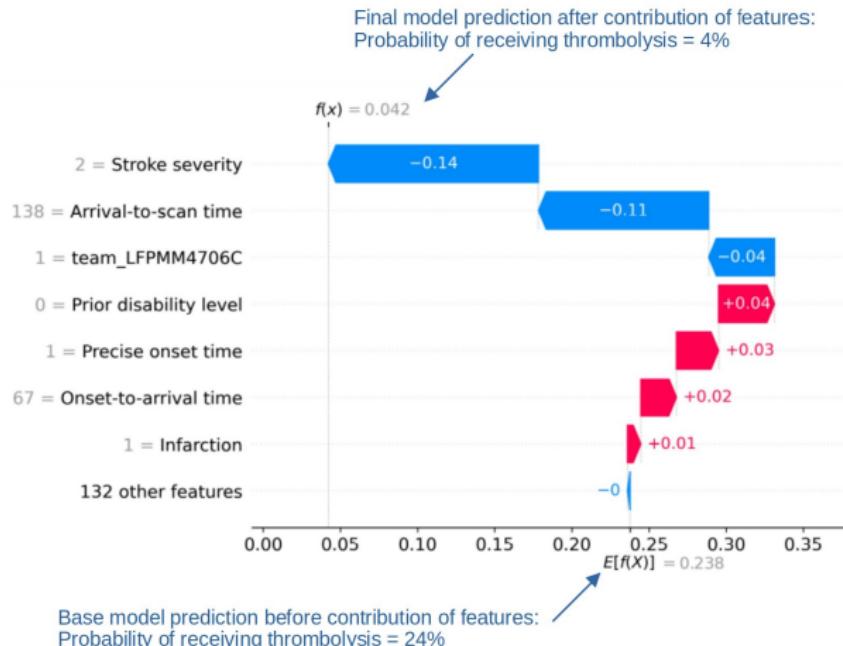
- Accuracy = 85%
- ROC AUC = 0.916
- The model can achieve 84% sensitivity and specificity simultaneously
 - Sensitivity: The proportion of patients who receive thrombolysis who are correctly classified
 - Specificity: The proportion of patients who do not receive thrombolysis who are correctly classified

https://samuel-book.github.io/samuel_shap_paper_1/xgb_with_feature_selection/02_xgb_combined_fit_accuracy_key_features.html

Explainability of machine learning: SHAP

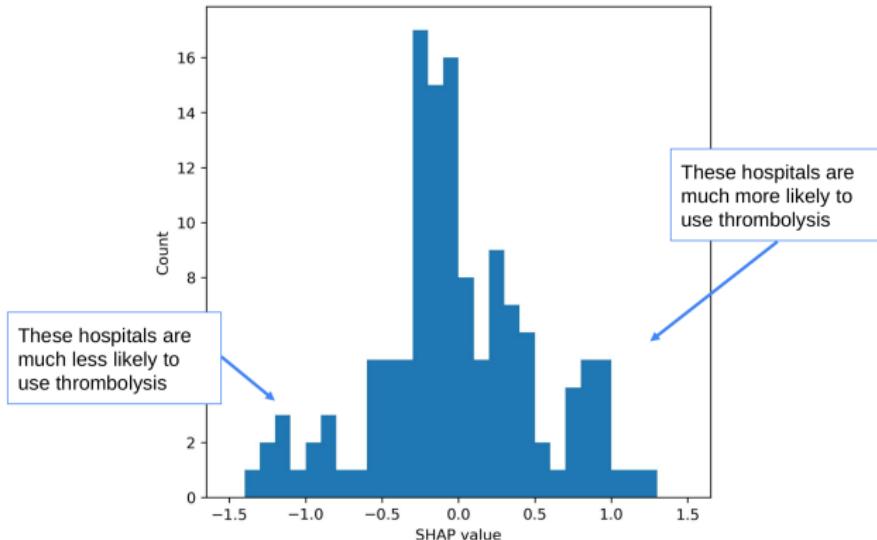
SHAP output for an individual patient

SHAP values show the influence of features (even for '*black box*' models).



For more on odds, probabilities and SHAP see: https://samuel-book.github.io/samuel_shap_paper_1/introduction/odds_prob.html

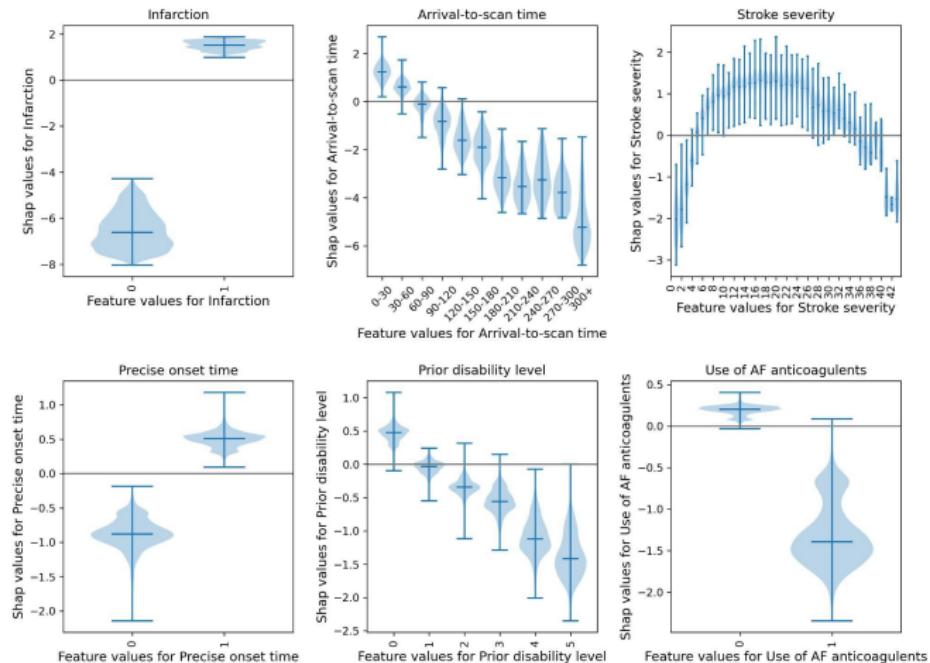
Average SHAP values for each hospital



Average SHAP values for each hospital range from -1.3 to +1.3.

https://samuel-book.github.io/samuel_shap_paper_1/xgb_with_feature_selection/03_xgb_combined_shap_key_features.html

SHAP values for predicting use of thrombolysis across all hospitals



Note: SHAP values here are *log odds*. Each step-change in value of ± 1 changes the chances of receiving thrombolysis about 3-fold. (Plots are in order of feature importance.)

https://samuel-book.github.io/samuel_shap_paper_1/xgb_with_feature_selection/03_xgb_combined_shap_key_features.html

Outcome modelling

How do we define outcomes?

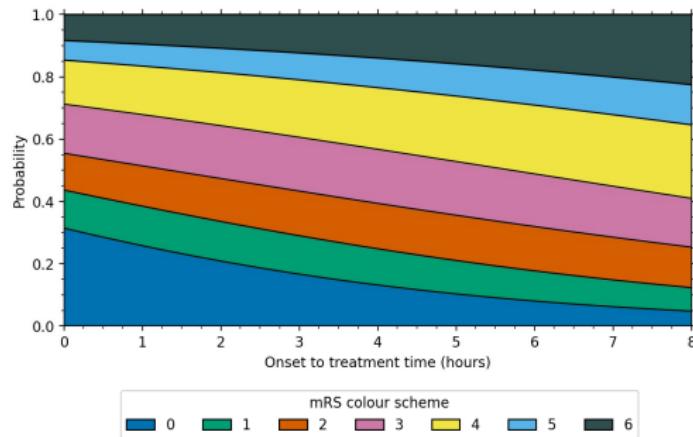
Modified Rankin Scale (mRS) is a measure of disability or dependence for people with stroke. Because mRS bands are not linear, we also use utility as a measure of quality of life.

	mRS	Utility	Description
"Good" outcome	0	0.97	<i>No symptoms.</i>
	1	0.88	<i>No significant disability.</i> Able to carry out all usual activities, despite some symptoms.
"Bad" outcome	2	0.74	<i>Slight disability.</i> Able to look after own affairs without assistance, but unable to carry out all previous activities.
	3	0.55	<i>Moderate disability.</i> Requires some help, but able to walk unassisted.
	4	0.20	<i>Moderately severe disability.</i> Unable to attend to own bodily needs without assistance, and unable to walk unassisted.
	5	-0.19	<i>Severe disability.</i> Requires constant nursing care and attention, bedridden, incontinent.
	6	0.00	<i>Dead.</i>

One of our aims is to calculate changes using all of the mRS bands, so the results are more finely-grained than the previous "good" or "bad" outcome options.

Outcome variation with time

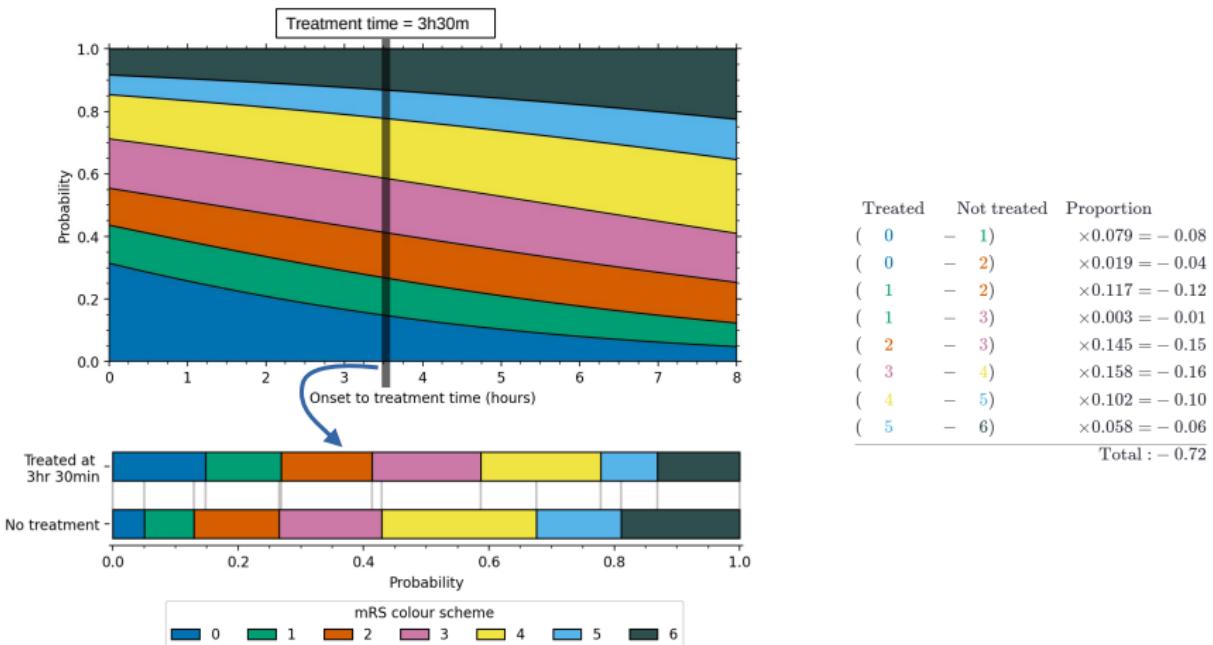
Expected outcomes of patients with large-vessel occlusions treated with thrombectomy:



The variation with time is a logistic function.

Outcome variation with time

Expected outcomes of patients with large-vessel occlusions treated with thrombectomy:

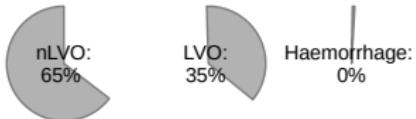


Patient population

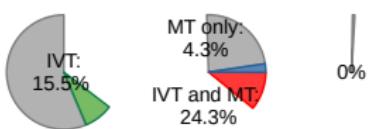


Patient population

- 1: Decide what proportion of the population has each stroke type.



- 2: Decide what proportion of each stroke type will receive each treatment.



Result: Multiply these proportions by these changes to find the mean population change in mRS and in utility.



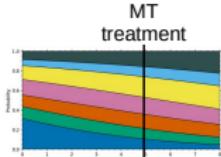
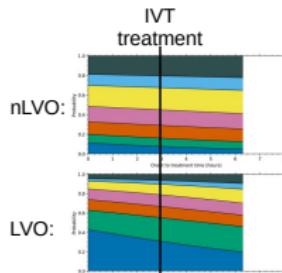
Treatment time

- 1: Pick a treatment time for IVT and for MT.

IVT: 3 hours

MT: 5 hours

- 2: Find the mRS probability distributions at the chosen times.



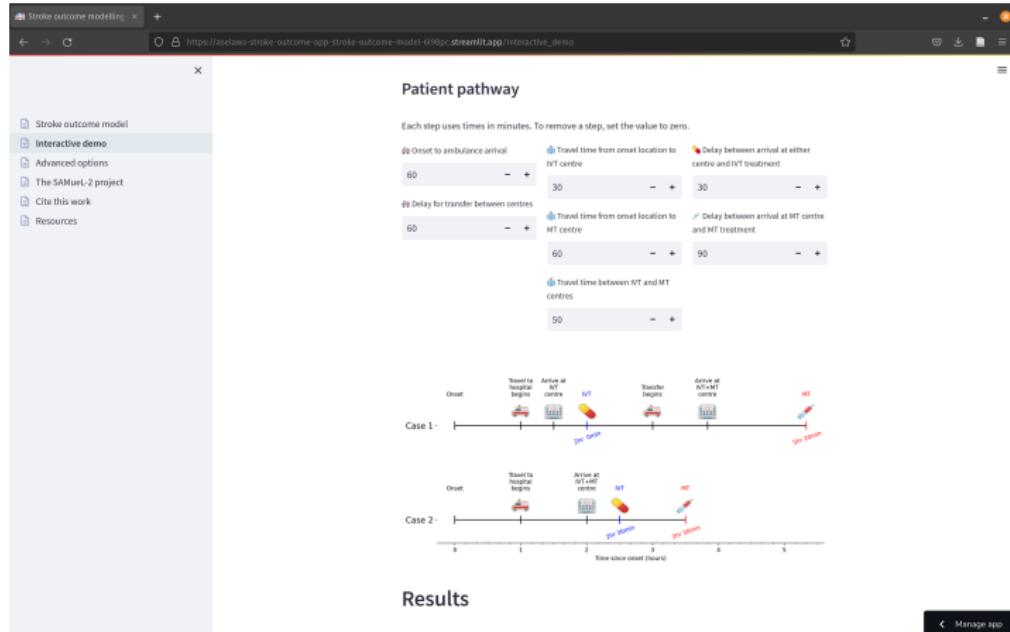
- 3: Find the mean changes in mRS and utility at these times compared with the non-treated population.

Outcome modelling:

Streamlit app

Streamlit app

Available here: <https://aselaws-stroke-outcome-app-stroke-outcome-model-690cc.streamlit.app/>



The app is linked to this (messy!) GitHub repository: https://github.com/aselaws/stroke_outcome_app

Summary

SHAP:

- An XGBoost model with 8 features has 85% accuracy at predicting use of thrombolysis
- Shapley (or SHAP) values show how much each feature influences the model prediction
- The probability of receiving thrombolysis is increased with:
 - Ischaemic stroke
 - Short arrival to scan time
 - Stroke severity 6-35
 - Precisely known onset time
 - Low prior disability
 - No use of anticoagulants for AF
 - Short onset-to-arrival times
 - Attending hospitals with high propensity to use thrombolysis
- We find that low-thrombolysing hospitals are more sensitive to these features.

Stroke outcome modelling:

- Many data sources have been combined to create modified Rankin scale (mRS) distributions as a function of time
- These can be used to find the expected changes in mRS and utility of a patient population
- An application of this is comparing the population outcomes depending on choice of hospital

Thank you!!

Thank you for your time and attention!

Reserve slides

Who are we?

- Quantitative researchers
- Qualitative researchers
- Clinicians
- Patients and Carers Involvement (PCI) group

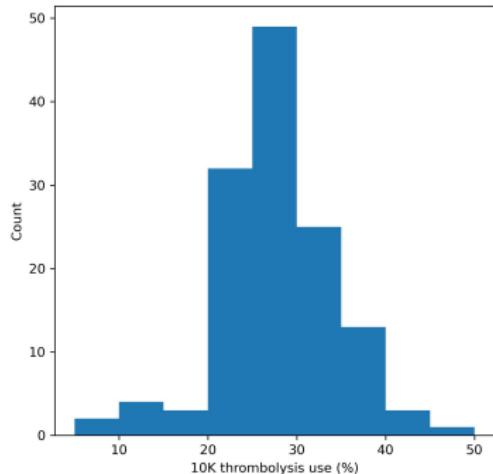
Our outputs have to be explainable to the clinicians and PCI group who are not experts in computer science.

We like Free and Open Source Software (FOSS)!

This project, SAMueL-2, is the follow-up to SAMueL-1 with more data and more detail to dig into.

10k cohort thrombolysis, and benchmark hospitals

- Separate out 10k patients
- Train model on remaining data (78,792 patients)
- Pass 10k cohort through model, changing hospital coding each time to mimic same patients going to each of 132 hospitals
 - All other patient and pathway data the same each time
- Model then predicts the thrombolysis rate across hospitals if they all saw the same patients



Note: Here we are looking at the predicted use of thrombolysis in patients who arrive within 4 hours of known stroke onset (this is about 40% of all emergency stroke admissions).

Learning differences in decision-making between hospitals with low and high propensity to thrombolysis

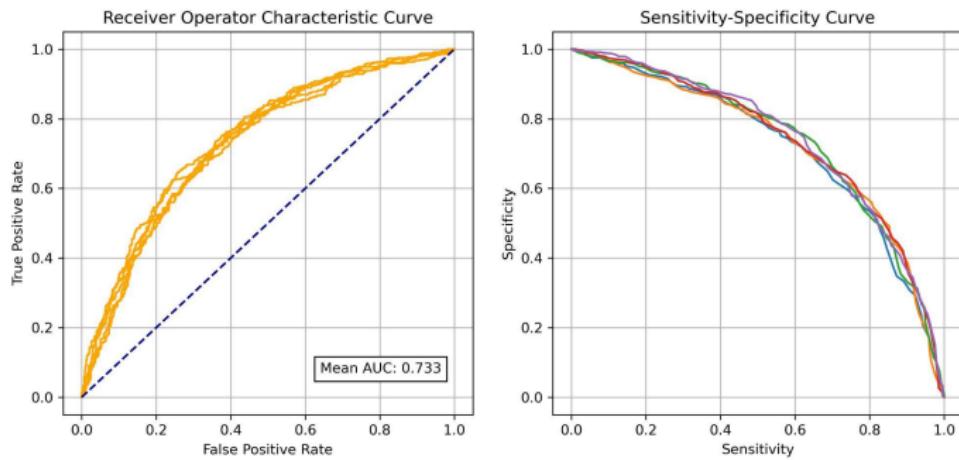
- Low-thrombolysing hospitals = lowest predicted 10k thrombolysis use
- High-thrombolysing hospitals = highest predicted 10k thrombolysis use

Motivation: “What patients do low-thrombolysing hospitals not thrombolyse, when high-thrombolysing hospitals would?”

Model:

- Get data for all patients attending low-thrombolysing hospitals
- Find those patients who would be given thrombolysis by a majority of the high-thrombolysing hospitals
- Train a model to predict which of those patients would be treated differently at the low-thrombolysing hospitals (i.e. would not receive thrombolysis)

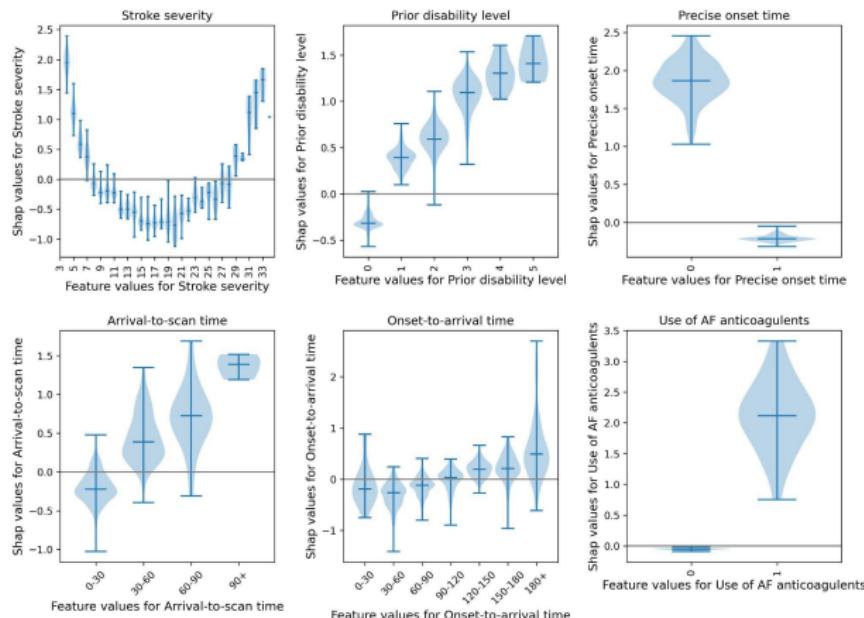
Accuracy of model to predict differences in clinical decision-making



- Accuracy = 67%
- ROC AUC = 0.733
- The model can achieve 67% sensitivity and specificity simultaneously
 - Sensitivity: The proportion of patients who receive thrombolysis who are correctly classified
 - Specificity: The proportion of patients who do not receive thrombolysis who are correctly classified

SHAP values for predicting when a low thrombolysis use hospital would **not** use thrombolysis when a high thrombolysis use hospital would

Here, a high SHAP shows when a low-thrombolysing unit will reject use of thrombolysis when a higher thrombolysing hospital would use thrombolysis. (Plots are in order of feature importance.)



We find that low thrombolysing hospitals are less likely to give thrombolysis to patients with:

- Low or very high stroke severity
- An estimated (not precise) stroke onset time
- Prior disability
- A longer arrival-to-scan time
- Use of AF anticoagulants
- A longer onset-to-arrival time

These are the same patterns as we see in general thrombolysis decision-making, but low-thrombolysing hospitals appear more sensitive to these features.

How do different SHAP values affect a starting probability of 25%?

SHAP values are usually reported as *log odds shift in odds*.

Let's build more intuition on the effect of different magnitudes of SHAP.

Starting data		SHAP value		Adjusted data	
Starting P	Starting O ($P / (1 - P)$)	SHAP	Shift in odds ($\exp(\text{SHAP})$)	Shifted O ($O * \text{Shift}$)	Shifted P (%) ($O / (1 + O)$)
0.25 (25%)	0.333	0.5	1.65	0.550	0.3547 (36%)
0.25 (25%)	0.333	1	2.72	0.907	0.4754 (48%)
0.25 (25%)	0.333	2	7.39	2.46	0.7112 (71%)
0.25 (25%)	0.333	3	20.1	6.70	0.8700 (87%)
0.25 (25%)	0.333	4	54.6	18.2	0.9479 (95%)
0.25 (25%)	0.333	5	148	49.5	0.9802 (98%)



Shapley (SHAP) values

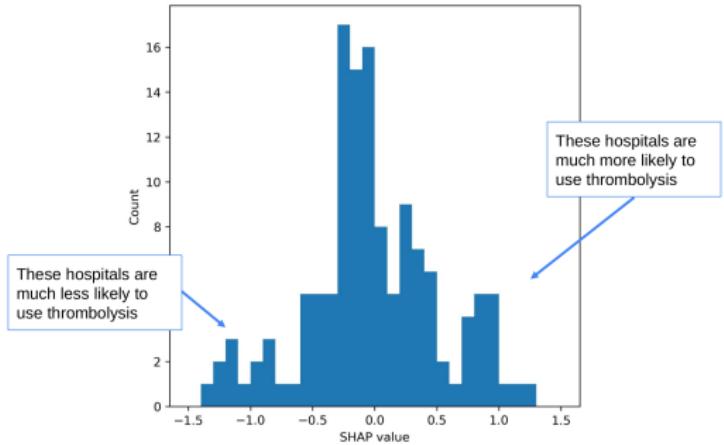
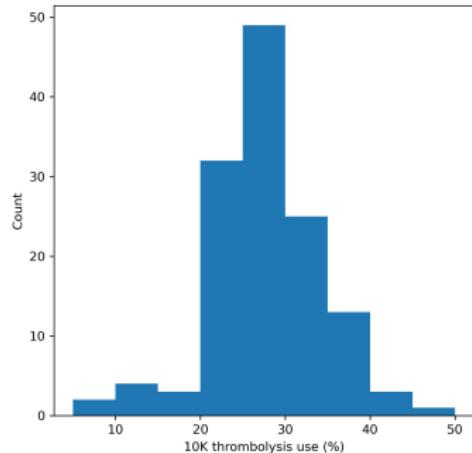
SHAP values report how individual patient features affect the model prediction.

SHAP values are usually reported as log odds shift in odds. That is not very intuitive!
But let's look at an example of how that feeds into probabilities.

Patient feature and value	SHAP	Shift ($\exp(\text{SHAP})$)	Odds (Previous odds * Shift)	Probability ($O / (1 + O)$)
Base odds	N/A	N/A	0.33	25%
Stroke type = infarction	1.8	6.05	2	67%
Stroke severity (NIHSS) = 20	1.5	4.482	8.95	90%
Prior disability (mRS) = 3	-0.7	0.497	4.44	82%
Precise onset time = Yes	0.6	1.822	8.1	89%
Arrival-to-scan time (mins) = 30	0.5	1.649	13.35	93%
Use of AF anticoagulants = No	0.3	1.35	18.02	95%

For more on odds, probabilities and SHAP see: https://samuel-book.github.io/samuel_shap_paper_1/introduction/odds_prob.html

Average SHAP values for each hospital



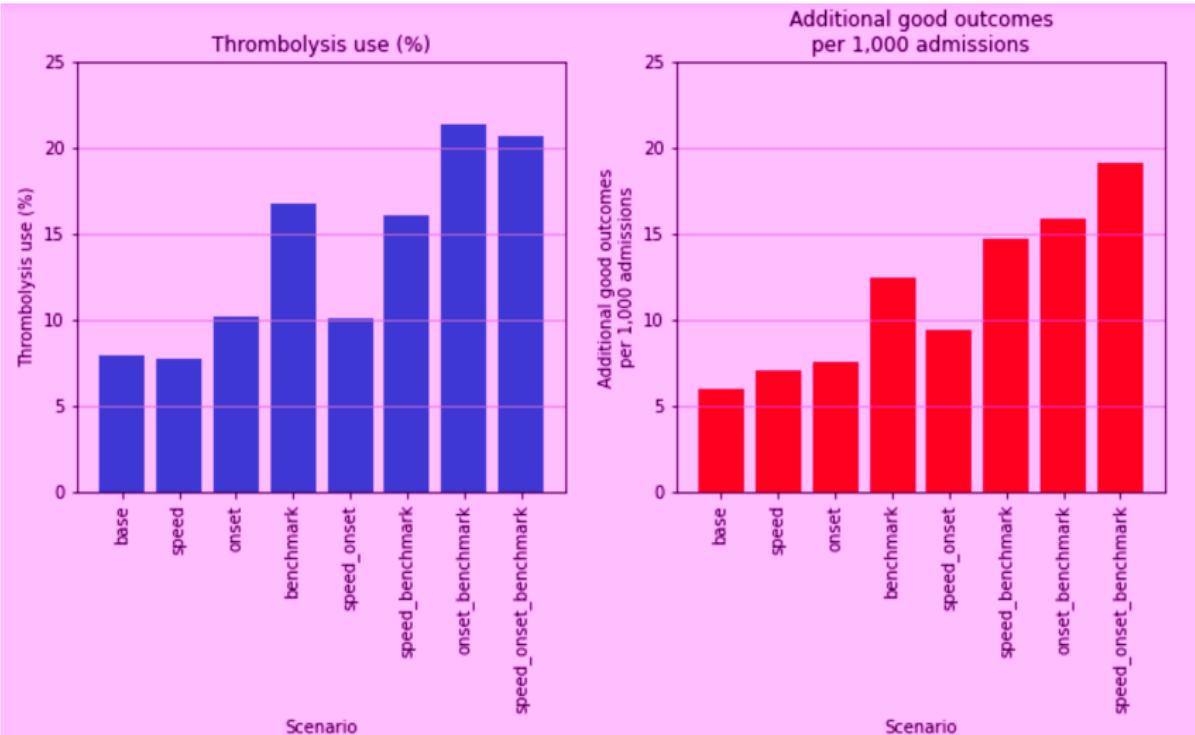
Average SHAP values for each hospital range from -1.3 to +1.3.

Note: these SHAP values were produced using the full cohort of patients, not just the 10K subset.

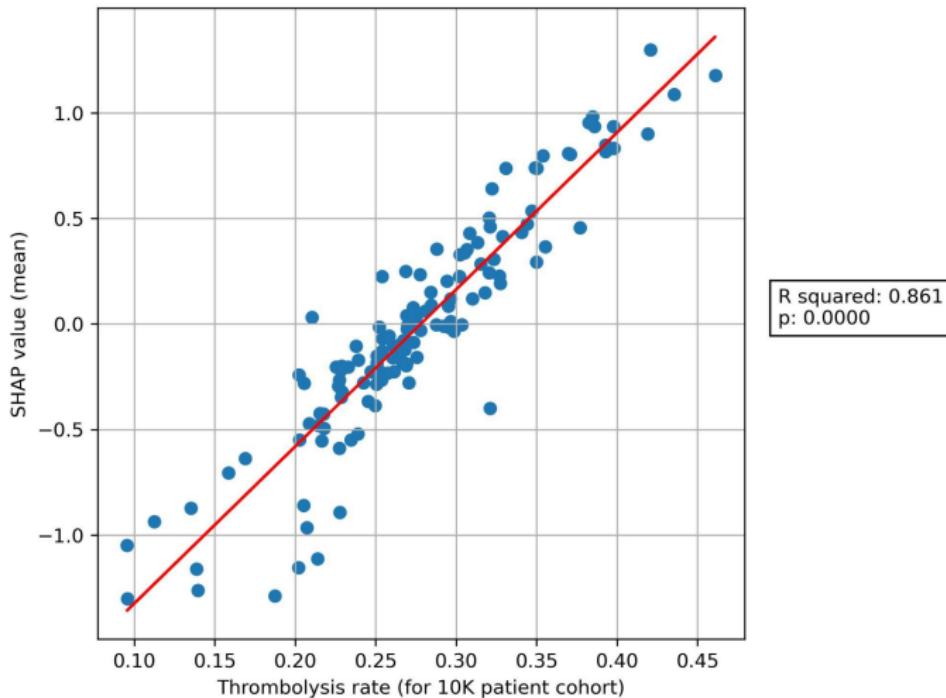
https://samuel-book.github.io/samuel_shap_paper_1/xgb_with_feature_selection/03_xgb_combined_shap_key_features.html

https://samuel-book.github.io/samuel_shap_paper_1/xgb_with_feature_selection/04_compare_10k_cohort_key_features.html

Applying our models at hospital level

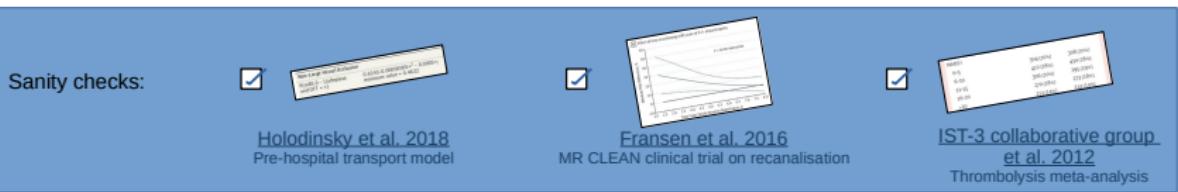
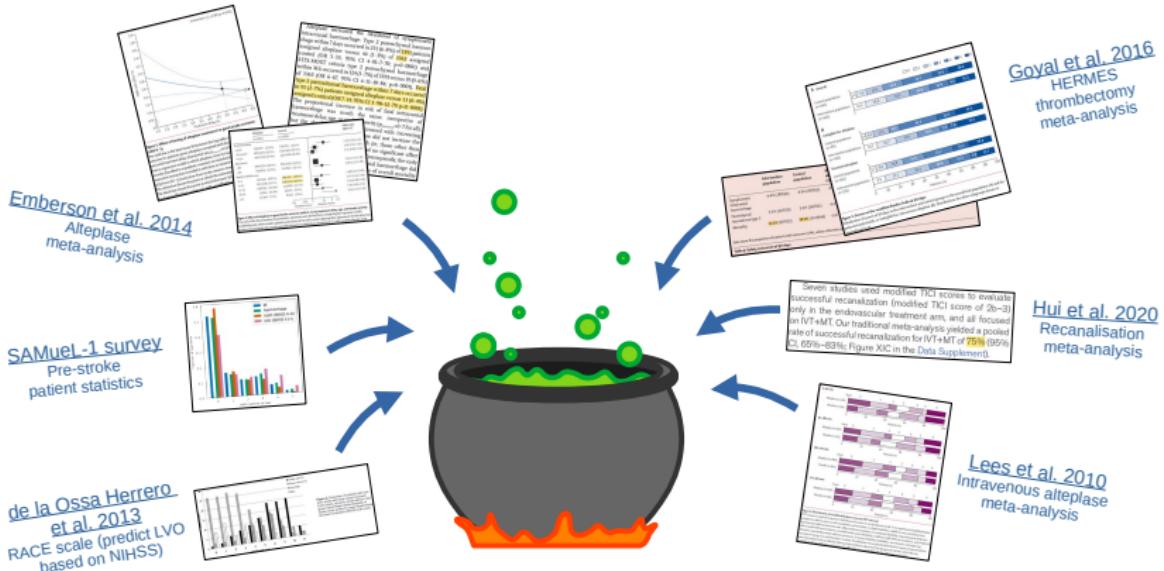


How does the predicted 10k thrombolysis use compare with the SHAP values learned for each hospital?

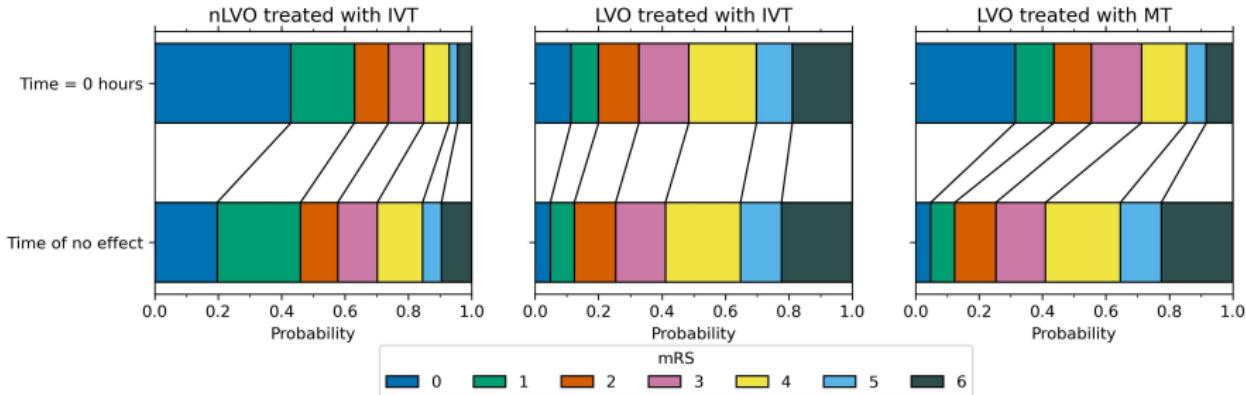


86% of the variance in predicted 10k thrombolysis use can be explained by the learned hospital SHAP.

Lots of data sources...!



The result: mRS probability distributions



We also create probability distributions for the pre-stroke population and the population that receives no treatment.

IVT Intravenous thrombolysis

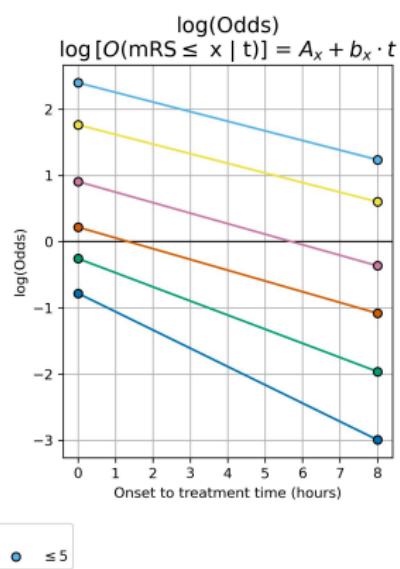
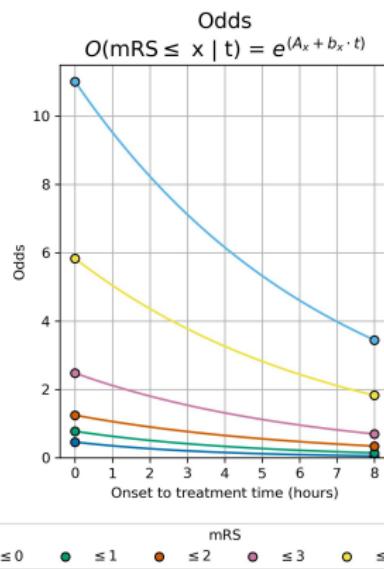
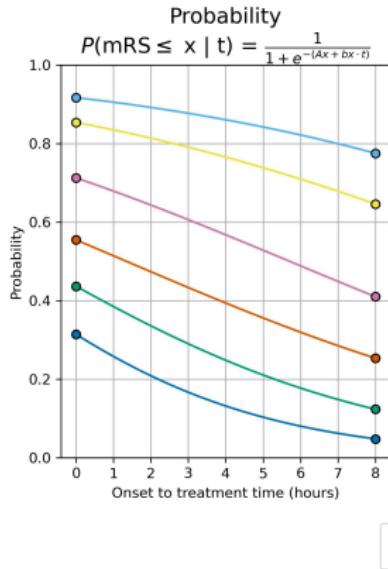
MT Mechanical thrombectomy

LVO Large-vessel occlusion

nLVO Non-large-vessel occlusion

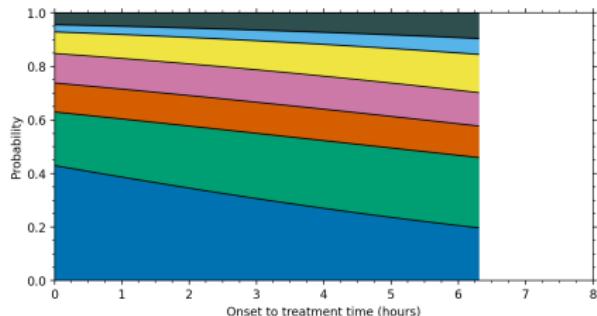
mRS Modified Rankin scale (disability level where 0=no symptoms and 6=dead)

mRS variation with time

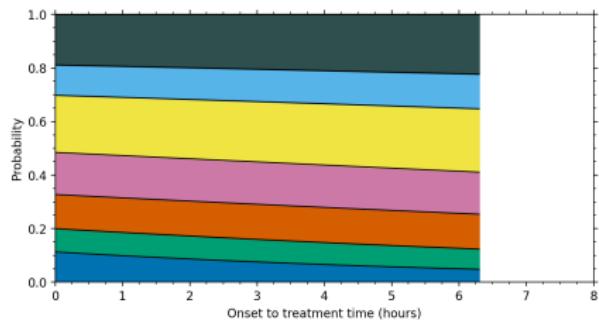


Variation with time for different stroke types and treatment types.

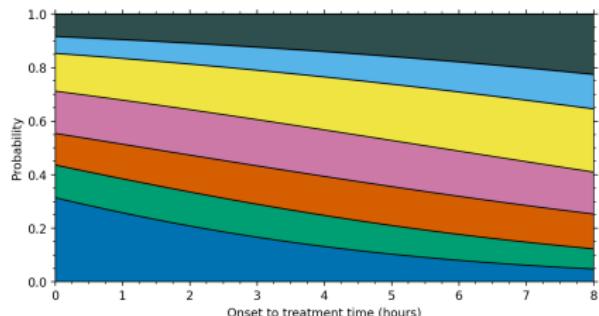
nLVO treated with IVT



LVO treated with IVT



LVO treated with MT



[https:](https://)

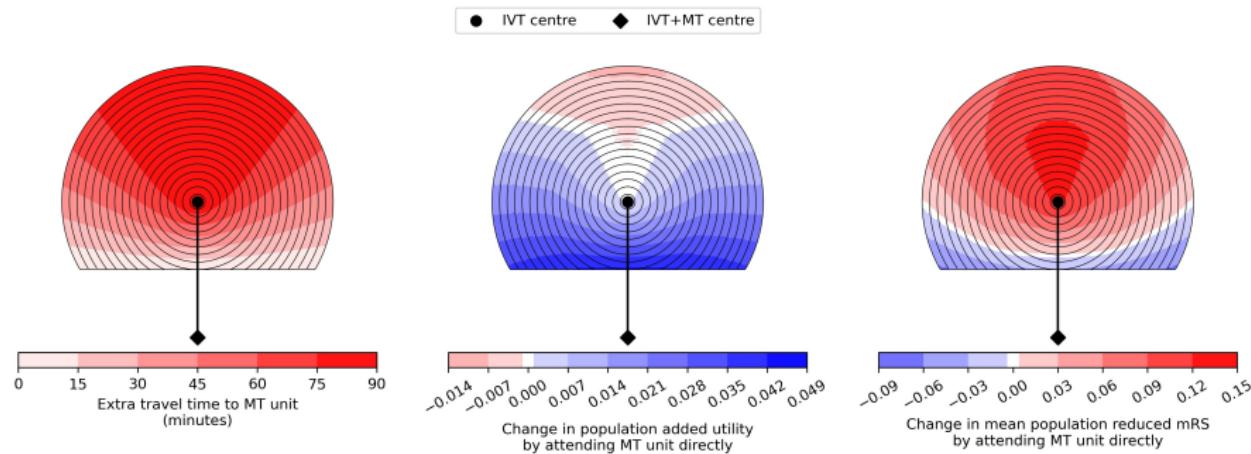
//github.com/samuel-book/stroke_outcome/blob/main/mRS_outcomes_maths.ipynb

Basic geography with stroke outcome modelling

What difference does it make if everyone in the patient population:

- ① travels to an IVT-only centre, and then transfers to a combined IVT+MT centre?
- ② travels directly to a combined IVT+MT treatment centre?

For the case when the centres are 90 minutes apart and travel begins 90 minutes after the stroke onset:



These are early results and could change. There is a large difference in which locations have zero gain depending on the outcome measure.

Based on similar geographic modelling by Holodinsky et al. 2018.

https://github.com/samuel-book/stroke_outcome/blob/main/geography.ipynb