ABSTRACT

Background and Objective

Long-term outcome after subarachnoid haemorrhage, beyond the first few months, is difficult to predict but has critical relevance to patients, their families and carers. We assessed the performance of the Subarachnoid Haemorrhage International Trialists (SAHIT) prediction models, which were initially designed to predict short-term (90 day) outcome, as predictors of long-term (2-year) functional outcome after aSAH.

Methods

We included 1545 patients with angiographically-proven aSAH from the Genetic and Observational Subarachnoid Haemorrhage (GOSH) study at 22 hospitals between 2011-2014. We collected data on age, WNFS grade on admission, history of hypertension, Fisher grade, aneurysm size and location, as well as treatment modality. Functional outcome was measured by the Glasgow Outcome Scale (GOS) with GOS 1 to 3 corresponding to unfavourable and 4 to 5 to favourable functional outcome, according to the SAHIT models. The SAHIT models were assessed for long-term outcome prediction by estimating measures of calibration (calibration slope) and discrimination (Area under the receiver operating characteristic curve (AUC)) in relation to poor clinical outcome.

Results

Follow-up was standardized to 2 years using imputation methods. All three SAHIT models demonstrated acceptable predictive performance for long-term functional outcome. The estimated AUC was 0.71 (95%CI: 0.65-0.76), 0.73 (95%CI:0.68-0.77) and 0.74 (95%CI: 0.69-0.79) for the core, neuroimaging and full models, respectively; the calibration slopes were 0.86, 0.84 and 0.89 indicating good calibration.

Conclusion

The SAHIT prediction models, incorporating simple factors available on hospital admission, show good predictive performance for long-term functional outcome after aSAH.

Running Title:

Assessing the SAHIT models for long-term outcome

Keywords:

Aneurysmal subarachnoid haemorrhage, complications, functional outcome**,** long-term functional outcome**,** prognostic models, validation

INTRODUCTION

With an overall mortality of up to 50%, aneurysmal subarachnoid haemorrhage (aSAH) is a major contributor to stroke-related loss of productive life years, since it occurs at younger ages than ischemic stroke or intracerebral haemorrhage[1,2]. Functional outcome remains poor[3]. Prediction models aim to assist early decision-making regarding acute treatments. Previous risk models for aSAH have been mostly based on small sample sizes, are time-consuming to apply, and have had insufficient evaluation and reporting of discrimination and calibration, with limited external validation[4]. Moreover, findings on risk factors entering a risk score have been inconsistent and sometimes contradictory with opposed effects. More recently a large collaborative group developed and validated three different models, the Subarachnoid Haemorrhage International Trialists (SAHIT) models, to predict short-term functional outcome after aSAH, overcoming many of the these limitations[5]. However, the SAHIT models have yet to be evaluated for longer-term outcome prediction, which is of critical importance in aSAH, a disease affecting younger people, often of working age.

We therefore assessed the performance of the SAHIT models in a large multicentre UK cohort in predicting long-term functional outcome (standardized to 2 years using imputation) after aSAH, including the assessment of model calibration and discrimination. We also investigated the smoking subgroup status, which has been suggested to be important in predicting outcome after aSAH[6-9].

PATIENTS AND METHODS

Study population

For this study we used patients with aSAH who had been recruited into the Genetic and Observational Subarachnoid Haemorrhage (GOSH) study between 2011-2014. We recruited participants with acute aSAH both retrospectively and prospectively. Due to the main focus of GOSH being genetic analysis, we mainly recruited patients retrospectively during follow-up in outpatient clinic if data regarding their acute aSAH was available (1115 patients, 72.2%). All participants had angiographically-verified intracranial aneurysms. Written informed consent was obtained from all participants or, in case of lack of capacity, from a representative. A stroke research practitioner completed a standardised case report form for each participant. Data collected included information on SAH severity measured using the WFNS score, age, history of hypertension, volume of aSAH on CT on admission measured using the Fisher grade[10], size and location of ruptured aneurysm, and treatment modality (i.e. coiling, surgery, or conservative treatment). Imaging parameters were assessed locally at the individual centre by a trained

person. Rebleeding was defined as occurring during the acute hospitalisation period. Functional outcome, using the Glasgow Outcome Scale (GOS), was collected at hospital discharge and at follow-up during the next documented routine clinical assessment, up to 12 years (median 2.3 years, IQR 2.5) after aSAH.

Exclusion criteria: patients with missing follow-up (Figure 1).

The SAHIT models were determined as previously[5]:

- Core model: consisting of patients age (continuous variable), WFNS grade on admission, and premorbid history of hypertension (yes or no)
- Neuroimaging model: adding volume of aSAH on CT on admission (measured by the Fisher scale), size (<12, 13-24 and >25 mm) and aneurysm location (anterior, internal middle cerebral artery and posterior circulation)
- Full model: incorporation of treatment modality (clipping, coiling or none) into the neuroimaging model

Our primary outcome was functional outcome at 2 years. Following the SAHIT models, we dichotomized the outcome GOS with GOS 1 to 3 corresponding to unfavourable and 4 to 5 to favourable functional outcome[5]. There was a wide time range for outcome data, so we standardized the outcome at 2 years using imputation. Briefly, the original outcomes were used if they were recorded within 1 year of the 2-year time-point. Otherwise, outcomes were either linearly interpolated or extrapolated using outcomes recorded at discharge and follow-up.

This study follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines and Transparent Reporting of a multivariable prediction model for Individual Prediction or Diagnosis (TRIPOD) statement[11].

This UK study was approved by the corresponding local Research Ethics committee (ethics reference number: 09/H0716/54).

Assessment of the SAHIT models

Any missing data in the predictor variables were imputed using multiple imputation by chained equations (ICE)[12]. Outcome, all pre-specified potential predictors, and predictors of missingness were included in the imputation model. Calibration was assessed using the calibration slope and Hosmer-Lemeshow test and discrimination was quantified by the area under the receiver-operating characteristic curve (AUC)[13]. A calibration slope of 1 indicates perfect calibration, a calibration slope of less than one indicatives model overfitting. An AUC of 0.5 indicates no discriminative ability, whereas an AUC of >0.7 indicates acceptable, AUC of >0.8 good discriminative ability, AUC of >0.9 excellent, and an AUC of 1 perfect discriminative ability[14]. We used the AUC since it is a common measure to quantify how well

the models discriminate between patients at high and patients at low risk of unfavourable outcome and to compare it to the SAHIT validation cohort. Functional outcome at 2 years was obtained for every patient, using imputation (linear interpolation/extrapolation) if necessary. We conducted a sensitivity analysis examining the models including only patients without imputation. The original outcome was used for 501 patients since this was recorded within 1 year. However, imputation was used for the remaining patients although most of these (967 of 1044 patients) had the same outcomes at discharge and follow-up (563 were measured before 1 year and 404 after 3 years).

Performance of the models in clinically relevant subgroups of patients

We measured the calibration and discrimination of the SAHIT models separately for smokers and non-smokers, based on the potential influence of smoking status on functional outcome after aSAH[6-9].

In addition, we examined the association of the predictors in the SAHIT models in our cohort by refitting the SAHIT models and comparing the re-estimated regression coefficients with the original coefficients. Additionally, we investigated whether adding smoker status might improve the prediction model.

Statistical analysis was performed by two biostatisticians and one neurosurgical trainee using STATA 15 (StataCorp. 2011. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LP) and R version 3.2 (The R Foundation).

RESULTS

Of 1729 patients recruited to the GOSH study we included 1545 patients with available follow-up. See Table 1 for summary of baseline and initial treatment characteristic. Most patients were female (71.3%) and mean age of the whole cohort was 53 years (range 18 to 92; 12.7 SD). 1219 patients (79.9%) suffered from a low-grade SAH with a WFNS of 1 or 2 and 326 (21.1%) of patients suffered from high-grade SAH with a WFNS of 3 to 5. With regards to treatment 1177 (76.2%) were treated with coiling, 275 (17.8%) with clipping, 55 (3.5%) with a combination of coiling and clipping and 38 (2.5%) did not receive any intervention.

We observed unfavourable outcome in 8.5% of our cohort. Mean follow-up time before imputation was 2.7 years with a SD of 3.6 years. Patient characteristics comparing the dichotomized outcome variable GOS (in line with the SAHIT models) are summarized in Table 1. Patients with poor functional outcome after aSAH were more frequently male, had more comorbidities (hypertension, hypercholesterolaemia, and diabetes mellitus), were more

frequently smokers, had higher WFNS and Fisher score, more frequently had aneurysms in the posterior circulation and larger aneurysms, were less frequently coiled and more frequently not treated, more frequently rebled, developed DCI and infarcted.

Assessment of the SAHIT models for long-term functional outcome

Figure 2A-C shows the calibration plots for each of the three prognostic scores divided into approximately equally sized groups. The cut-off points were the quintiles of the predicted risk. Agreement between observed and predicted risks was reasonable for all three models. This was supported by calibration slopes of 0.86 (95% CI 0.66-1.05), 0.84 (95% CI 0.66-1.03) and 0.89 (95% CI 0.71-1.08) for the core, neuroimaging and full model respectively as well as p-values in the Hosmer-Lemeshow test of 0.14, 0.05 and 0.11, respectively, though perhaps there is some evidence of lack of fit for the neuro model. We note that the calibration slopes in the original validation cohort of the SAHIT models were 1.06 for the core, 1.07 for the neuroimaging and 1.05 for the full model[5]. The SAHIT models predict long-term functional outcome with acceptable accuracy in our cohort: AUC was 0.71 (95% CI 0.65-0.76) for the core model, 0.73 (95% CI 0.68-0.77) for the neuroimaging model and 0.74 (95%CI 0.69-0.79) for the full model. The respective AUCs in the original SAHIT validation cohort was 0.80 (95% CI 0.78-0.82) for the core, 0.81 (95% CI 0.79-0.84) for the neuroimaging and 0.81 (95% CI 0.79-0.83) for the full model[5].

As a sensitivity analysis, we assessed the performance of the SAHIT models using those patients whose outcomes were not imputed (501 patients). Performance improved, with an increase in AUC values and improved calibration (supplementary Table 1). We then included all patients whose follow-up times were recorded within 5 years (1282 patients). Performance of the models also improved in this sample (supplementary Table 1).

Performance of the models by smoking status

The AUC values were slightly higher for two of the models when applied to smokers only: 0.75 (95% CI 0.68-0.81) for the core, 0.76 (95% CI 0.69-0.83) for the neuroimaging and 0.76 (95% CI 0.7-0.83) for the full model respectively (Figure 1A-C, supplemental material for the ROC curve). In contrast, all of the AUC values were lower when the models were applied to non-smokers only: 0.66 (95% CI 0.58-0.74), 0.69 (95% CI 0.62-0.76) and 0.72 (95% CI 0.65-0.79). Agreement between observed and predicted was good for non-smokers but poor for smokers. For smokers the risk was underestimated for the highest risk groups (see Figure 2A-C, supplemental material, for the calibration plots by smoking group).

Importance of the predictors of the SAHIT models in the GOSH cohort

Due to the number of events in our cohort we only evaluated the predictors in the core model. Overall, the results were similar (Table 2): age and WFNS were significant predictors of long-term functional outcome. The re-estimated regression coefficients were similar to those from the original SAHIT model but there was some evidence that the core model might be improved for 2-year outcomes through re-estimation of the regression coefficients (p=0.01).

We then added the potential predictor smoking status to the core model (Table 3). This improved the fit of the model (p=0.01) and suggests that adding smoker status might improve the SAHIT core model for 2-year outcomes.

DISCUSSION

Key results

We assessed the performance of the SAHIT models in predicting long-term outcome after aSAH and have demonstrated adequate prediction by their good discriminative abilities in a large UK cohort. Accuracy of the SAHIT models in our cohort was acceptable measured by an AUC of 0.71, 0.73 and 0.74 of the core, neuroimaging and full model, respectively. The respective pooled AUC in the SAHIT validation cohort was 0.8, 0.81 and 0.81[5]. When taking into account that the lowest AUC for one of the included cohorts in the SAHIT models for the core, neuroimaging and full model were 0.66, 0.72 and 0.7 in the development and 0.76, 0.75 and 0.75 in the validation dataset, the models performed reasonably well for long-term functional outcome prediction in our cohort[5]. This suggests that all three models perform reasonably well for prediction of long-term outcome after aSAH measured by GOS at 2 years.

Interpretation and Generalizability

Although there was no significant difference between the three models in our cohort, we see the advantage of having three models in their potential application at different time points with improved accuracy when adding additional information. The core model is useful in the acute setting as the required variables are usually known on admission of the patient even without available imaging (e.g. in poor-resource areas or in non-specialised centres). The neuroimaging and the full model both are valuable adjuncts for potentially more accurate prediction (albeit not statistically superior to the core model). The neuroimaging model is helpful after the hyperacute arrival phase once a scan is available. The application of the full model is likely to be more relevant at later time points as treatment can be delayed. Although in the SAHIT cohort the full model, including the treatment option, is not better in functional outcome prediction compared to the neuroimaging model, in our cohort for long-term outcome prediction the full model is indeed slightly superior compared to the neuroimaging model. A sensitivity analysis

evaluating patients where no imputation was required showed that our results regarding performance of the SAHIT models may be slightly conservative, and the true predictive value of the models slightly higher than shown in our cohort.

As an additional step we evaluated the performance of the models according to smoker status as this variable was significantly associated with functional outcome in our cohort. Previous studies have indicated an influence of smoking status on functional outcome after aSAH, although interestingly showing smokers having a better outcome than non-smokers[7,8]. When adding smoking status to the core model we found weak evidence for this trend. When dividing the models by smoking status the model did not perform as well for the subgroup of smokers compared to non-smokers. This indicates a potential problem with prediction in the group of high-risk patients, which most likely is due to small predicted risks due to bias towards survivors in our cohort. Nevertheless, our findings suggest that smoking should be evaluated as a further factor to include in the original SAHIT models. However, although including many factors improves the prediction ability of a model, it also makes the application more complicated and time-consuming and thus decreases the likelihood of it being applied.

There were significant differences between our cohort and the SAHIT cohorts: our cohort had a significantly lower rate of unfavourable outcome. As previously described, individuals could be included prospectively as well as retrospectively. This inevitably creates survivor bias. Indeed, 8.5% of our cohort had unfavourable outcome compared to the combined dataset of the SAHIT models in which 29% would suffer unfavourable outcome[5]. It is reassuring that despite these differences, we demonstrate good predictability of the SAHIT models for long-term functional outcome. This difference on the other hand, might explain why the SAHIT models performed slightly worse in our cohort. A further difference exists in treatment modalities. Overall, in the SAHIT cohorts clipping was significantly more frequent compared to coiling, this difference being larger in the development compared to the validation population[5]. In our cohort patients were more frequently coiled, which reflects the current treatment standard more accurately. In the SAHIT cohort 48% of the patients underwent clipping compared to 17.29% in our cohort and 34% underwent coiling as opposed to 77.57% in our cohort. As coiling has become more common compared to clipping in recent years, this difference is most likely due to the enrolment period of the included cohorts in the SAHIT models[5]. In our cohort patients who underwent coiling more frequently had a favourable outcome whereas patients who underwent clipping more frequently had unfavourable outcome. The higher coiling and lower clipping rate in our cohort could partially contribute to the lower rate of unfavourable outcome

in the GOSH compared to the SAHIT cohort, although we acknowledge a bias towards survivors due to some of our participants being included retrospectively.

Another difference is noted in Fisher grades: the GOSH cohort demonstrated a higher frequency of Fisher grade 4 compared to the SAHIT cohorts, where Fisher grade 3 was the most common grade. Although again, when comparing the development with the validation population in the SAHIT cohorts it appears that the validation cohort resembles the GOSH cohort more closely[5]. The difference here could be due to how people classify patients into Fisher grade 3 and 4 depending on whether or not they have intraventricular or intracerebral blood. The other variables had similar distributions compared to the whole dataset of the SAHIT models (development and validation together).

Despite these differences, all three models performed well in our independent large cohort, which clearly supports the usefulness and applicability of these three models.

Other prognostic models include the SAFIRE grading scale[15], used to predict poor functional outcome at two months (modified Rankin Scale of 4-6). This model includes age, aneurysm size and Fisher grade and WFNS after resuscitation such as insertion of an extraventricular drain or haematoma evacuation[5,15,16]. Although SAFIRE also showed good predictive performance, SAHIT has the advantages of application on admission and prediction of long-term functional outcome.

Our study has important strengths: the analysis was conducted on a large sample of patients recruited to a multicentre study conducted in the UK. As the different centres participated in one single retrospective as well as prospective study patients were recruited using a standardized patient questionnaire ensuring the use of uniform definitions of demographic data and risk factors. Although we did observe differences between the SAHIT and our cohort, this is a common finding in independently collected cohorts. Our cohort is a realistic example of an independent cohort and as such reflects realistic performance of the SAHIT models for prediction of functional outcome in patients with aSAH.

Limitations

Our study also has limitations. First and foremost, due to the possible retrospective inclusion we acknowledge a bias towards survivors after aSAH. Although this is a common problem as patients not reaching the primary care centre or dying within the first few hours after arrival can seldomly be recruited into a study, this bias might be more pronounced in our cohort due to 72.2% being recruited retrospectively. Additionally, the follow-up time was not standardized and therefore had a wide range. With the application of imputation methods, we have standardized the follow-up time to our time-point of interest, 2 years. Despite using imputation

methods, we were not able to get a GOS for 184 patients (10.6%) so they were therefore excluded. However, these patients did not vary significantly in their baseline characteristics from the included patients (data not shown). Our models showed evidence of mild overfitting indicating increased model complexity and decreased generalizability, but in a sensitivity analysis we have compared patients with and without follow-up within the range and did not find significant differences.

CONCLUSION

We successfully demonstrate that the SAHIT models accurately predict long-term functional outcome after aSAH, measured by the dichotomized GOS at 2 years in a multicentre cohort.

REFERENCES

1. Suarez JI, Tarr RW, Selman WR. Aneurysmal subarachnoid hemorrhage. *The New England journal of medicine.* 2006;354(4):387-396.
2. Johnston SC, Selvin S, Gress DR. The burden, trends, and demographics of mortality from subarachnoid hemorrhage. *Neurology.* 1998;50(5):1413-1418.
3. Connolly ES, Jr., Rabinstein AA, Carhuapoma JR, et al. Guidelines for the management of aneurysmal subarachnoid hemorrhage: a guideline for healthcare professionals from the American Heart Association/american Stroke Association. *Stroke; a journal of cerebral circulation.* 2012;43(6):1711-1737.
4. Jaja BN, Cusimano MD, Etminan N, et al. Clinical prediction models for aneurysmal subarachnoid hemorrhage: a systematic review. *Neurocritical care.* 2013;18(1):143-153.
5. Jaja BNR, Saposnik G, Lingsma HF, et al. Development and validation of outcome prediction models for aneurysmal subarachnoid haemorrhage: the SAHIT multinational cohort study. *BMJ.* 2018;360:j5745.
6. Nath S, Koziarz A, Badhiwala JH, Almenawer SA. Predicting outcomes in aneurysmal subarachnoid haemorrhage. *BMJ.* 2018;360:k102.
7. Pobereskin LH. Influence of premorbid factors on survival following subarachnoid hemorrhage. *Journal of neurosurgery.* 2001;95(4):555-559.
8. Dasenbrock HH, Rudy RF, Rosalind Lai PM, et al. Cigarette smoking and outcomes after aneurysmal subarachnoid hemorrhage: a nationwide analysis. *Journal of neurosurgery.* 2018;129(2):446-457.
9. Friedrich V, Bederson JB, Sehba FA. Gender influences the initial impact of subarachnoid hemorrhage: an experimental investigation. *PloS one.* 2013;8(11):e80101.
10. Fisher CM, Kistler JP, Davis JM. Relation of cerebral vasospasm to subarachnoid hemorrhage visualized by computerized tomographic scanning. *Neurosurgery.* 1980;6(1):1-9.
11. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine.* 2015;162(1):W1-73.
12. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine.* 1999;18(6):681-694.
13. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.*: Springer Science & Business Media; 2001.
14. Simundic AM. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC.* 2009;19(4):203-211.
15. van Donkelaar CE, Bakker NA, Birks J, et al. Prediction of Outcome After Aneurysmal Subarachnoid Hemorrhage. *Stroke; a journal of cerebral circulation.* 2019;50(4):837-844.
16. van Donkelaar CE, Bakker NA, Veeger NJ, et al. Prediction of outcome after subarachnoid hemorrhage: timing of clinical assessment. *Journal of neurosurgery.* 2017;126(1):52-59.

328

FIGURE LEGEND

Figure 1. Patient selection flow diagram

Figure 2A-C, Calibration plot observed versus predicted values with according ROC, A core model, B neuroimaging model, C full model

SUPPLEMENTARY DATA LEGEND

Supplementary Figure 1A-C, ROC by smoker group, A core model, B neuroimaging model, C full model

Supplementary Figure 2A-C, Calibration plot observed versus predicted values by smoker group, risk groups combined for A core model, B neuroimaging model, C full model

Supplementary Table 1, Model performance of the SAHIT models for the main analysis and two sensitivity analyses. AUC = Area Under Curve, CS = Calibration Slope, N = Number. Values in parentheses are 95% confidence intervals.