

# Aneurysmal Subarachnoid Haemorrhage

Insights through Data-Driven Population Studies



UMC Utrecht Brain Center

Jos Peter Kanning

# **Aneurysmal Subarachnoid Haemorrhage**

Insights through Data-Driven Population Studies

Jos Peter Kanning

**Title:** Aneurysmal Subarachnoid Haemorrhage: Insights through Data-Driven Population Studies

**Author:** Jos Kanning

Cover idee: Jos Kanning

Design: Proefschrift AIO

Print: Proefschrift AIO

ISBN: 978-94-93406-31-5

Copyright 2025 ©Jos Kanning

All rights reserved. No portion of this book may be reproduced in any form without prior permission from the author.

# **Aneurysmal Subarachnoid Haemorrhage**

Insights through Data-Driven Population Studies

## **Aneurysmatische Subarachnoidale Bloeding:**

Inzichten door Datagedreven Populatiestudies

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de  
Universiteit Utrecht  
op gezag van de  
rector magnificus, prof. dr. H.R.B.M. Kummeling,  
ingevolge het besluit van het College voor Promoties  
in het openbaar te verdedigen op  
dinsdag 14 januari 2025 des ochtends te 10.15 uur

door

**Jos Peter Kanning**

geboren op 1 februari 1995

te Groningen

**Promotoren:**

Prof. dr. Y.M. Ruigrok

Dr. M.I. Geerlings

**Copromotor:**

Dr. S. Abtahi

**Beoordelingscommissie:**

Prof. dr. A. Abu-Hanna

Prof. dr. H. Gardarsdottir (voorzitter)

Prof. dr. F.H. Rutten

Prof. dr. E.W. Steyerberg

Dr. M. Uyttenboogaart

The research described in this thesis has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 852173).

## TABLE OF CONTENTS

<b>Chapter 1</b>		General introduction and outline of this thesis	<b>7</b>
<b>Chapter 2</b>		Developing Clinical Prediction Models Using Primary Care Electronic Health Record Data: The Impact of Data Preparation Choices on Model Performance	<b>13</b>
<b>Chapter 3</b>		Cardiovascular Risk Prediction in Men and Women Aged Under 50 Years Using Routine Care Data	<b>31</b>
<b>Chapter 4</b>		Prediction of aneurysmal subarachnoid hemorrhage in comparison with other stroke types using routine care data	<b>49</b>
<b>Chapter 5</b>		Development and external validation of the SMA2SH2ERS risk prediction model for aneurysmal subarachnoid hemorrhage in the general population	<b>69</b>
<b>Chapter 6</b>		Identifying novel risk factors for aneurysmal subarachnoid haemorrhage using machine learning	<b>87</b>
<b>Chapter 7</b>		Prescribed Drug Use and Aneurysmal Subarachnoid Haemorrhage Incidence: A Drug-Wide Association Study	<b>107</b>
<b>Chapter 8</b>		Associations between lisinopril use and risk of aneurysmal subarachnoid haemorrhage: A UK population-based cohort study	<b>125</b>
<b>Chapter 9</b>		General discussion	<b>141</b>
<b>Chapter 10</b>		Summary	<b>153</b>
<b>Chapter 11</b>		Nederlandse samenvatting	<b>157</b>
<b>Appendices</b>		Publications by the author	<b>164</b>
		About the author	<b>166</b>
		Dankwoord	<b>167</b>



## Chapter 1

# General introduction and outline of this thesis

---

## INTRODUCTION

Big data refers to data sets that are too large or complex to be managed by traditional data-processing application software.<sup>1</sup> Big data plays an increasingly important role in biomedical and health research, driven by the expansion of data sources such as electronic health records (EHR), biobanks, and population cohorts, as well as advances in data handling techniques such as machine learning.<sup>2</sup> These advances have enabled new data-driven research paradigms,<sup>3</sup> in which large amounts of available data are used to improve disease risk prediction, discover new risk factors, and identify novel treatment options.<sup>4</sup> A big data approach offers several advantages over traditional methods, including the potential for hypothesis-free designs, exploration of multiple variables simultaneously and non-linearly, and more personalised and precise medical interventions.<sup>2,5</sup>

Despite advances in data-driven paradigms in biomedical research, their potential to enhance our understanding of aneurysmal subarachnoid haemorrhage (aSAH) has been underexplored. aSAH is a type of stroke that occurs when an intracranial aneurysm ruptures, causing bleeding into the subarachnoid space.<sup>6</sup> While aSAH accounts for only one-tenth of strokes worldwide,<sup>7</sup> it has a high impact due to its high morbidity and mortality rates and its relatively young average onset age of 55 years.<sup>7-9</sup> Early detection of unruptured aneurysms followed by preventive neurosurgical or endovascular treatment can theoretically prevent aSAH.<sup>10</sup> However, most aSAH cases are not prevented in practice because aneurysms are often undetected until they rupture.<sup>11</sup> Additionally, even when detected early, it is difficult to predict which aneurysms will rupture, as current risk estimation methods are inadequate, and we may be missing knowledge on important aSAH risk factors.<sup>12</sup> Moreover, current treatment methods pose significant risks of permanent disability or death, often outweighing the potential benefits,<sup>13</sup> leading to most patients remaining untreated. Therefore, there is a critical need for non-invasive treatment options for aSAH.

The overarching aim of this thesis is to improve our understanding of aSAH risk estimation, risk factors, and alternative treatment options using data-driven approaches. It is structured around three objectives: 1) to enhance the predictive accuracy for identifying individuals at risk of aSAH; 2) to discover and describe novel aSAH risk factors by leveraging a combination of machine learning and traditional statistics; and 3) to examine non-invasive, drug-based treatment options for aSAH through pharmacoepidemiologic methods.

## OUTLINE OF THE THESIS

Chapters **2-5** address the first objective to enhance predictive accuracy for identifying individuals at risk of aSAH. **Chapter 2** discusses the opportunities and challenges of using primary care EHR data to study cardiovascular disease and highlights how data preparation choices affect the performance of clinical prediction models. **Chapter 3** demonstrates the effectiveness of using EHR data to study cardiovascular disease by developing sex-specific prediction models based on data-driven predictor selection. In **chapter 4**, we apply the methods from chapter 3 to develop an EHR-derived prediction model specifically for aSAH and address the unique challenges of predicting aSAH in the general population. **Chapter 5** elaborates on a limitation identified in chapter 4, specifically the missingness or misclassification of important aSAH risk factors in EHR data. As a solution for this limitation, we develop a new aSAH prediction model using systematically assessed data on aSAH risk factors from the UK Biobank. **Chapter 6** focuses on the second objective by identifying and describing novel aSAH risk factors by combining machine learning and traditional statistics in the UK Biobank. Chapters **7-8** address the third and final objective by examining non-invasive drug-based treatment options for aSAH. In **chapter 7**, we use primary care data to identify commonly prescribed drugs associated with a reduced aSAH risk through a drug-wide association study. In **chapter 8**, we build on these signals by investigating whether the antihypertensive drug lisinopril is more effective at lowering aSAH risk than other drugs in its drug class, using an active comparator new-user design.

## REFERENCES

1. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. *J Big Data*. 2016;3(1):25.
2. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. *Biomed Inform Insights*. 2016;8:S31559.
3. Leonelli S. Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Stud Hist Philos Biol Biomed Sci*. 2012;43(1):1–3.
4. Hey T, Tansley S, Tolle KM, others. The fourth paradigm: data-intensive scientific discovery. Redmond, WA: Microsoft Research; 2009.
5. Cremin CJ, Dash S, Huang X. Big data: historic advances and emerging trends in biomedical research. *Curr Res Biotechnol*. 2022;4:138–51.
6. Macdonald RL, Schweizer TA. Spontaneous subarachnoid haemorrhage. *Lancet*. 2017;389(10069):655–66.
7. Nieuwkamp DJ, Setz LE, Algra A, Linn FH, Rooij NK, Rinkel GJ. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *Lancet Neurol*. 2009;8(7):635–42.
8. de Rooij NK, Linn FH, van der Plas JA, Algra A, Rinkel GJ. Incidence of subarachnoid haemorrhage: a systematic review with emphasis on region, age, gender and time trends. *J Neurol Neurosurg Psychiatry*. 2007;78(12):1365–72.
9. Rinkel GJ, Algra A. Long-term outcomes of patients with aneurysmal subarachnoid haemorrhage. *Lancet Neurol*. 2011;10(4):349–56.
10. Connolly ES, Rabinstein AA, Carhuapoma JR, Derdeyn CP, Dion J, Higashida RT, et al. Guidelines for the management of aneurysmal subarachnoid hemorrhage. *Stroke*. 2012;43(6):1711–37.
11. Harrison CH, Taquet M, Harrison PJ, Watkinson PJ, Rowland MJ. Sex and age effects on risk of non-traumatic subarachnoid hemorrhage: retrospective cohort study of 124,234 cases using electronic health records. *J Stroke Cerebrovasc Dis*. 2023;32(8):107196.
12. Etminan N, Rinkel GJ. Unruptured intracranial aneurysms: development, rupture and preventive management. *Nat Rev Neurol*. 2016;12(12):699–713.
13. Algra AM, Lindgren A, Vergouwen MD, Greving JP, van der Schaaf IC, van Doormaal TP, et al. Procedural clinical complications, case-fatality risks, and risk factors in endovascular and neurosurgical treatment of unruptured intracranial aneurysms: a systematic review and meta-analysis. *JAMA Neurol*. 2019;76(3):282–93.





## Chapter 2

# Developing Clinical Prediction Models Using Primary Care Electronic Health Record Data: The Impact of Data Preparation Choices on Model Performance

---

Hendrikus J. A. van Os\*, Jos P. Kanning\*, Marieke J. H. Wermer, Niels H. Chavannes, Mattijs E. Numans, Ynte M. Ruigrok, Erik van Zwet, Hein Putter, Ewout W. Steyerberg, Rolf H. H. Groenwold

\*Shared first authors

Front Epidemiol. 2022 Jun 2;2:871630

## ABSTRACT

### Objective

To quantify prediction model performance in relation to data preparation choices when using electronic health records (EHR).

### Study Design and Setting

Using Dutch primary care EHR data, Cox proportional hazards models were developed to predict the first-ever main adverse cardiovascular events. The reference model was based on a one-year run-in period, cardiovascular events were defined based on EHR diagnosis and medication codes, and missing values were multiply imputed. We compared data preparation choices regarding i) length of the run-in period (two- or three-year run-in); ii) outcome definition (EHR diagnosis codes or medication codes only); and iii) methods addressing missing values (mean imputation or complete case analysis) by making variations on the derivation set and testing their impact in a validation set.

### Results

We included 89,491 patients, and 6,736 first-ever main adverse cardiovascular events occurred during a median follow-up of eight years. Outcome definition based only on diagnosis codes led to systematic underestimation of risk (calibration curve intercept: 0.84; 95% CI: 0.83 – 0.84), while complete case analysis led to overestimation (calibration curve intercept: -0.52; 95% CI: -0.53 -0.51). Differences in the length of the run-in period showed no relevant impact on calibration and discrimination.

### Conclusion

Data preparation choices regarding outcome definition or methods to address missing values can substantially impact the calibration of predictions, hampering reliable clinical decision support. This study further illustrates the urgency of transparent reporting of modelling choices in an EHR data setting.

## INTRODUCTION

Electronic health records (EHRs) enable the improvement of quality of care by providing structured information stored in a digital format, straightforwardly derived from routine health care.<sup>1,2</sup> Besides advantages related to the clinical workflow, increased standardisation and pooling of EHR data lead to very large datasets that can be of great value for the development of clinical prediction models. EHR-based datasets can reach an unprecedented scale and variety of recorded data, which is practically impossible to achieve in traditional cohort research.<sup>3,4</sup> However, EHRs are designed to record data routinely collected during the clinical workflow under a time constraint, in contrast to dedicated prospective cohort studies in which data are collected by trained personnel in a highly standardized manner.<sup>5</sup> Consequently, numerous data quality problems are relatively more pronounced in EHR data.<sup>6</sup> Previous studies have already enumerated the challenges that the EHR data quality limitations pose for the development of valid clinical prediction models. To overcome these challenges, in many cases the researcher is faced with difficult or seemingly arbitrary choices in data preparation, for example regarding the handling of missing predictor values.<sup>6-8</sup> Consequently, it may occur in research practice that different data preparation choices will be made for model derivation (or validation) compared with the context of model deployment, which may impact the predictive performance of the model when deployed in clinical practice. The quantification of such choices has not received much attention. In this paper we aimed to evaluate the impact of three previously identified data preparation challenges for EHR-derived prediction models: i) using a run-in period to define predictors at time zero, ii) outcome definition, and iii) methods used to address missing values.<sup>6-8</sup> As a case study, we focused on estimating cardiovascular risk in Dutch primary care EHR data.

## METHODS

### Data source

Patient information was derived from general practitioner (GP) practice centers affiliated with the Extramural LUMC Academic Network (ELAN), Leiden, the Netherlands. From the ELAN data warehouse we defined an open cohort of patients enlisted with ELAN GP practice center within the period of January 1<sup>st</sup> 2007 to and including December 31<sup>st</sup> 2018. Patient data included anonymized prescribed medication coded according to the Anatomical Therapeutic Chemical (ATC) classification, laboratory test results performed in primary care, symptoms and diagnoses coded according to the WHO-FIC recognized International Classification

of Primary Care (ICPC).<sup>9, 10</sup> For many GP practice centers the EHR data on ATC and laboratory test result data became available shortly before or after 2007. Inclusion criteria were age between 40 and 65 years, and absence of a history of cardiovascular disease at cohort entry at the end of the run-in period (see section 2.4.1 for details on the run-in period).

## Study design

From our original dataset we derived nine datasets based on the predefined data preparation challenges. We considered the dataset with a one-year run-in period, an outcome defined as either ICPC or ATC code for first-ever main adverse cardiovascular events and multiple imputation as method for addressing missing values as the reference dataset. In addition to the reference set, we created two derivation sets with a variation in run-in time, four with varying outcome definitions, and two with different methods to address missing values. These eight variations on the reference dataset are described in more detail in the sections below. For each derived dataset, we took a random 70% to 30% sample from the original dataset IDs to generate a list of derivation- and validation IDs. Derivation IDs were joined with the derived dataset of interest in order to generate a derivation set. Validation IDs were joined with the reference set to generate a validation set. Through this approach, we ensured that no individual ID could be in both the derivation and validation sets. We subsequently performed data preparation steps on the derivation and validation sets, fitted the predictive model and recorded outcome measures. This process was repeated 50 times per derived dataset in a bootstrap procedure for a robust estimate of outcome measures. The study design is graphically displayed in Figure 1.

## Model development

A multivariable Cox proportional hazards model was developed predicting first-ever main adverse cardiovascular events. The following predictors were selected based on prior knowledge: age, sex, mean systolic blood pressure, mean total cholesterol, and smoking as predictors, conform to the European SCORE model for prediction of cardiovascular mortality.<sup>11</sup>

## Data preparation challenges at model development Defining predictors at time zero and a run-in period

Time zero (or  $t_0$ ) is usually defined as the time of enrolment or baseline assessment of covariates. The start of the recording of data in EHRs is in principle the first contact with the healthcare system, which for an individual could be birth or in the prenatal or preconception period. However, as many countries do not have a single, national

EHR, health data may be fragmented across EHRs of different healthcare providers resulting in left-truncation within an EHR database. Hence, there generally is not one clear baseline assessment of predictors. When the time of EHR entry is chosen as  $t_0$  usually no values for laboratory or vital parameter predictors are available. This initial absence of recorded data is in computer sciences also known as the 'cold start' problem.<sup>12</sup> A possible solution is to define a run-in period, in which all data routinely acquired during a predefined time interval are aggregated into summary variables at the end of this time interval.<sup>13</sup> Because of left truncation in our EHR dataset we chose the start date of our data window as January 1<sup>st</sup>, 2007. We then defined a run-in period of one year, meaning that the  $t_0$  was defined as one year after the first moment a patient entered the database since January 1<sup>st</sup>, 2007. Additional requirements were age between 40 and 65 years old at  $t_0$ . Follow-up ran until the end of the data window at 31<sup>st</sup> of Dec 2018, or until unregistering with an ELAN GP practice center, death or first-ever main adverse cardiovascular event, whichever came first. Baseline predictors were assessed based on predictor values up until the end of the run-in period. If within this period multiple measurements of systolic blood pressure or total cholesterol were present, the mean value was taken as baseline measurement. As derivation set variations we defined run-in periods of two and three years (see Table 3). The reason we chose the one year run-in period as a reference was to maximize follow-up time. We chose the mean value as aggregation method for multiple measurements during run-in, as within this one year period measurement values were relatively recent with respect to  $t_0$ . Patients who suffered from main adverse cardiovascular events during the run-in period were excluded from analyses.

### **Outcome definition**

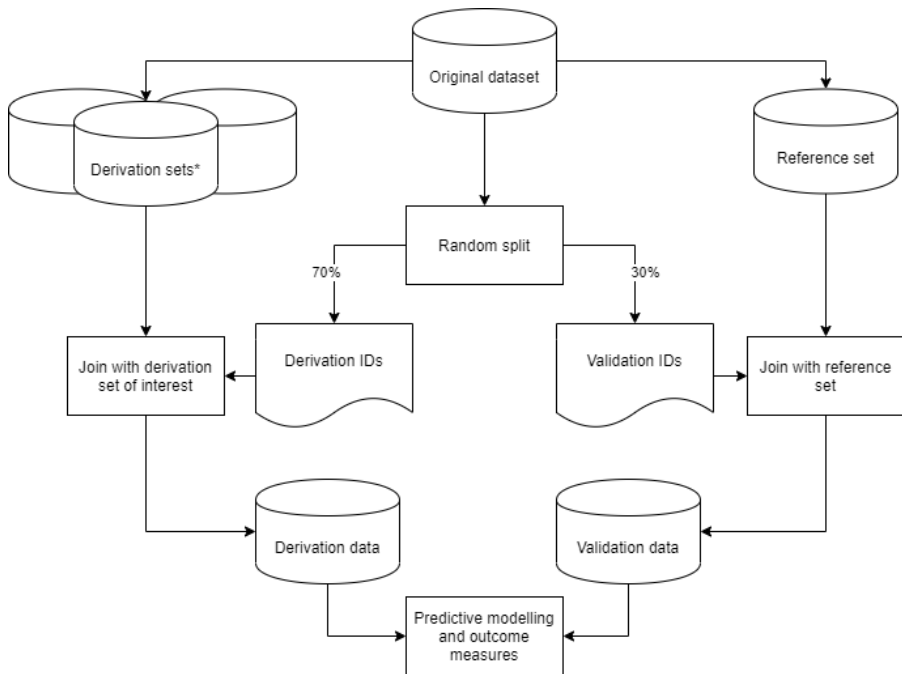
EHRs are designed to record data that are routinely collected during the clinical workflow. This is different from traditional research, where data are collected by trained personnel in a highly standardized manner.<sup>5</sup> This difference could lead to several EHR data quality issues. For instance a clinical outcome may be present in reality, but has not been recorded in the EHR at all or under a different code, possibly leading to misclassification of outcomes.<sup>14</sup> What is more, in an EHR data context one has many more options for outcome definition than in traditional cohort data, such as constructing outcome using medication or diagnosis codes, or both. Differences in outcome definition in the derivation and target population may cause poor model performance in the target population. The clinical outcome of this study was the 10-year risk of a first-ever major adverse cardiovascular event, and was based on either event specific ICPC codes for primary care diagnoses of acute stroke [K90], TIA [K89], acute myocardial infarction [K75], or the start of prescription of event specific ATC codes for thrombocyte aggregation inhibitors (ticagrelor, picotamide,

clopidogrel, dipyridamole, acetylsalicylic acid). In different derivation sets, the outcome was defined i) based on ATC codes (without acetylsalicylic acid) or ICPC codes; ii) based on ATC codes only (including acetylsalicylic acid); iii) based on ATC codes only, excluding acetylsalicylic acid; or iv) based on ICPC codes only. The reason for emitting acetylsalicylic acid from the outcome definition is that in the period of our t0 (2007) it was also prescribed as analgesic in primary care.<sup>15</sup> In addition, Dutch guidelines recommend prescription of acetylsalicylic acid for stable angina pectoris.<sup>16</sup> Consequently, although it may increase sensitivity for predicting major adverse cardiovascular events, it could come at a cost for specificity. Ticagrelor, picotamide, clopidogrel, and dipyridamole can be regarded as more specific for main adverse cardiovascular events. Although non-cardiovascular mortality could be considered as a competing event, we did not perform a competing risk analysis to limit the complexity of analyses in this paper.

### **Missing values**

Since EHR data result from routine care processes, virtually all health data are recorded during clinical contacts for a clinical reason. The missingness of a predictor value is therefore most likely related to clinical choices of the healthcare professional. In dealing with missing values it is essential to consider the mechanism of missingness.<sup>17</sup> For e.g. a missing measurement of systolic blood pressure in the EHR, missing completely at random (MCAR) is very unlikely because in clinical practice blood pressure assessment generally requires a medical indication. Missing at random (MAR) will occur if contextual information present in the EHR fully captures the clinician's motives – including those related to the outcome – to assess systolic blood pressure. Arguably, this is unlikely as clinical decision making takes a large number of biological, psychological and social factors into account. Missing not at random (MNAR) is therefore the most likely mechanism in this case. In case of MNAR commonly used imputation strategies such as multiple imputation may result in biased imputed values.<sup>18</sup> The combination of an MNAR mechanism with large extent of missingness in many predictors in EHR data may further increase risk of biased imputations.<sup>19,20</sup> One way of still leveraging information from the data without requiring sophisticated imputation is the missing indicator method. However, also in this case similarity of the missingness mechanism between the derivation and target populations is needed.<sup>21</sup> Complete case analysis in EHR data could introduce a bias towards the selection of e.g. sicker patients.<sup>22</sup> One should therefore assess how risk of bias resulting from handling missing values may affect the validity of predictions in the target population, and thus the clinical safety of future implementation of the model. Based on this assessment it may be advisable to discard predictors with a very high extent of missingness and possibly MNAR mechanism altogether.

We imputed the missing continuous predictors systolic blood pressure and cholesterol using Multivariate Imputation by Chained Equations (MICE). As input for the MICE algorithm we used the 30 most important predictors according to a Cox PH model with an elastic net penalty predicting first-ever cardiovascular events. Although missing values in systolic blood pressure or total cholesterol predictors are unlikely MAR, we multiply imputed because these are important baseline predictors which are used in virtually all cardiovascular risk prediction models. In addition, the aim of this study is not to produce prediction models that can be transported to true clinical settings, but the comparison of different data preparation choices in an EHR data context. Imputations were performed for all derivation and validation sets separately to prevent cross-contamination. We performed multiple visualizations of the complete and completed datasets. Further, we compared the results of the different imputation strategies with the Dutch population means for our age distribution.<sup>23</sup> For binary variables we assumed that absence of a registration of a clinical entity meant the clinical entity itself was absent. We defined two derivation set variations in which we addressed missing values in the continuous predictors using complete case analysis and mean imputation instead of MICE.



**Figure 1.** Graphic display of the study design

Graphic display of the study design. \*Derivation sets (nine in total: one reference and eight variations) were derived from our original data set, with data preparation steps based on the predefined data preparation challenges.

## Assessment of model performance at validation

Models based on the derivation set variations were validated on the reference dataset (see schematic overview in Figure 1). Model performance was assessed via the concepts of discrimination (ability of the model to separate individuals who develop the event versus those who do not) and calibration (the agreement between the estimated and observed number of events). For evaluation of discrimination we used the concordance index (c-index), and calibration was assessed using the calibration curve slope and -intercept. For details on these metrics we refer to the literature.<sup>24</sup> We used bootstrap validation with 50 bootstraps for internal validation, and simple bootstrap resampling to derive empirical confidence intervals. Analyses were performed using Python version 3.7.

## RESULTS

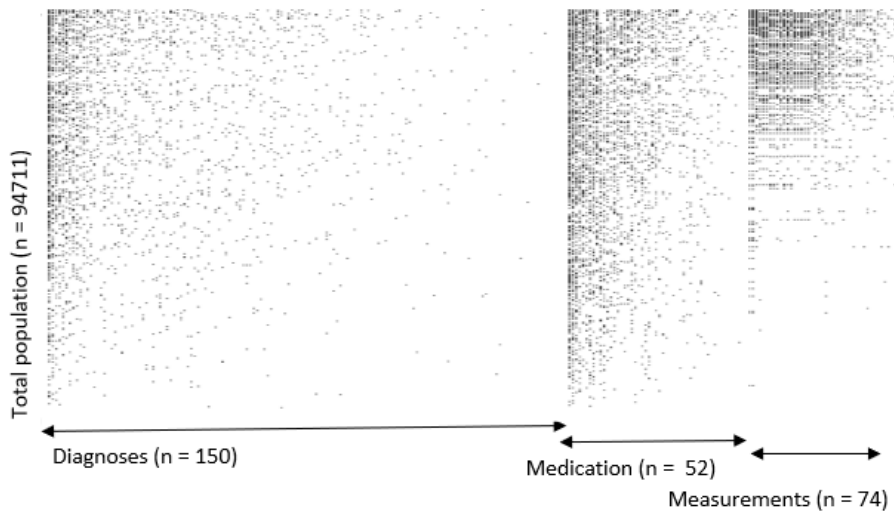
For our example case study, we included 89,491 patients for analyses in whom 6,736 first-ever cardiovascular events occurred during a median follow-up of eight years. On average, patients were 51 years old, and 51% were women. (Table 1) Visualization of the routine data recorded in the entire population showed that for the majority of patients, of the total of 150 potential diagnoses no EHR-registrations were present. Although relatively more registrations among the 52 medication and 74 measurement codes were present, for a large part of the population no information was available (Figure 2). For variations in definition of outcome, the inclusion of acetylsalicylic acid in the definition resulted in a larger number of cases (Figure 3). Differences were noted between the means in complete cases analysis, imputed by MICE and the estimated population mean. (Table 2)

**Table 1.** Baseline characteristics of participants

<b>Baseline characteristics</b>	<b>Cases (n = 6,736)</b>	<b>Controls (n = 82,755)</b>
Age, mean ( $\pm$ SD)	54.8 (6.8)	51.3 (7.3)
Women, n (%)	2849 (42.3)	42867 (51.8)
Smoking, n (%)	494 (7.3)	3760 (4.5)
<b>Presence of predictor measurement, n (%)</b>		
Systolic blood pressure	2302 (34.2)	18992 (22.9)
Total serum cholesterol	1637 (24.3)	13254 (16.0)

**Table 2.** Imputation results of systolic blood pressure and total cholesterol in Dutch primary care EHR data (n=89,491)

	Systolic blood pressure (mmHg)	Total cholesterol (mmol/l)
<b>Estimated population mean used for mean imputation (SD)</b>	130 (16)	5.7 (1.1)
<b>Sample mean of available measurements/ complete case analysis (SD)</b>	136 (17)	5.4 (1.1)
<b>Sample mean after MICE imputation (SD)</b>	132 (10)	5.4 (0.5)

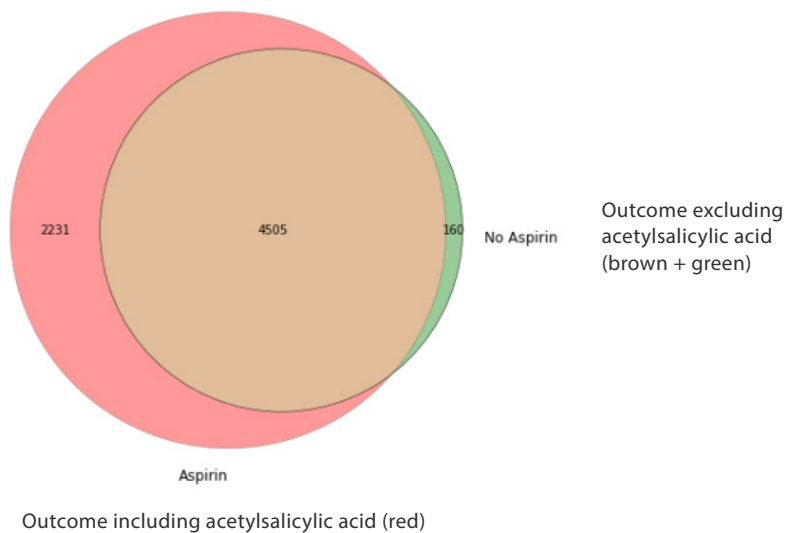


**Figure 2.** Visualization of data density in Dutch primary care EHR (n = 89,491)

This figure shows the data density in the EHR for the first year of follow-up of all included patients. The x-axis is divided into three different predictor groups: diagnoses (any type of ICPC registration), medications (any type of ATC registration), and laboratory or vital parameter measurements (any type of registration), with each dot representing an EHR registration data point. The y-axis represents the entire research population ranked from patients with most data points and descending.

Testing the reference Cox PH model predicting cardiovascular events on the validation set resulted in a c-statistic of 0.67; 95% CI: 0.67 – 0.67), a calibration curve intercept of 0.00; 95% CI: -0.01 – 0.00), and -slope of 1.00; 95% CI: 0.99 – 1.00). Discrimination and calibration were similar for the models based on derivation sets with two- or three-year run-in variations. For the derivation sets with variations in outcome definition discrimination remained the same but calibration varied greatly, especially when outcome was based only on ICPC (calibration curve intercept: 0.84; 95% CI: 0.83 – 0.84, and -slope: 2.31; 95% CI: 2.29 – 2.32). In this

derivation set variation the event rate was substantially lower compared with the validation set (3.4% versus 7.5%, respectively), and hence risk was underestimated at model validation. For models based on derivation set variations in missing data handling, again discrimination was similar to the reference model, but for complete case analysis calibration was substantially worse (calibration curve intercept: -0.52; 95% CI: -0.53 – -0.51, and -slope: 0.60; 95% CI: 0.59 – 0.60). For this variation also the total sample size was substantially smaller (around 12% of the reference derivation set) and event rate was higher (11.4% versus 7.5% of the validation set), hence risk was overestimated at model validation (Table 3).



**Figure 3.** Venn diagram with three different operationalizations for the outcome definition

This Venn diagram shows the numbers of first-ever main adverse cardiovascular event cases resulting from the different outcome definitions: ICPC only (brown; 4505 cases), ICPC and ATC codes for event specific medication (clopidogrel, ticagrelor, dipyridamole) including acetylsalicylic acid (red; 4505 + 2231 cases) and ICPC and ATC codes for event specific medication excluding acetylsalicylic acid (brown + green; 4505 + 160 cases).

**Table 3.** Performance of the models based on derivation set variations compared with the reference model in Dutch primary care EHR data (n=89,491)

Data preparation challenge	Derivation set variation description	Derivation set characteristics**				Performance metrics**			
		Sample size (range)	Percentage events (range)	Median follow-up time (days; range)	C-statistic (95% CI)	Calibration curve intercept (95% CI)	Calibration curve slope (95% CI)	Calibration curve slope (95% CI)	
Reference derivation set*	NA	62644 (62557 - 62730)	7.5 (7.5 - 7.6)	2912 (2904 - 2920)	0.67 (0.67 - 0.67)	0.00 (-0.01 - 0.00)	1.00 (1.00 - 1.01)	1.00 (1.00 - 1.01)	
Run-in variations	2 years run-in	58168 (58098 - 58236)	7.0 (7.0 - 7.1)	2832 (2832 - 2832)	0.67 (0.67 - 0.67)	0.00 (-0.01 - 0.00)	1.00 (0.99 - 1.00)	1.00 (0.99 - 1.00)	
Variations in outcome definition	3 years run-in	54958 (54884 - 55031)	6.4 (6.4 - 6.5)	2833 (2833 - 2833)	0.67 (0.67 - 0.67)	0.02 (0.01 - 0.03)	1.02 (1.01 - 1.03)	1.02 (1.01 - 1.03)	
	ATC (excl. ASA) or ICPC	63376 (63301 - 63448)	5.1 (5.1 - 5.2)	2933 (2925 - 2940)	0.67 (0.67 - 0.67)	-0.40 (-0.41 - -0.40)	0.67 (0.66 - 0.67)	0.67 (0.66 - 0.67)	
	ATC only	63518 (63436 - 63597)	7.5 (7.4 - 7.5)	2916 (2909 - 2922)	0.68 (0.68 - 0.68)	-0.01 (-0.02 - 0.00)	0.99 (0.99 - 1.00)	0.99 (0.99 - 1.00)	
Missing data method variations	ATC (excl. ASA) only	64739 (64662 - 64819)	4.6 (4.5 - 4.6)	2968 (2956 - 2979)	0.68 (0.68 - 0.68)	-0.52 (-0.53 - -0.51)	0.59 (0.59 - 0.60)	0.59 (0.59 - 0.60)	
	ICPC only	64089 (63998 - 64180)	3.4 (3.3 - 3.4)	3025 (3010 - 3040)	0.66 (0.66 - 0.66)	-0.84 (-0.85 - -0.83)	0.43 (0.43 - 0.44)	0.43 (0.43 - 0.44)	
	Complete Case Mean imputation	7601 (7573 - 7629) 62548 (62478 - 62618)	11.4 (11.3 - 11.5) 7.5 (7.5 - 7.6)	2425 (2409 - 2442) 2910 (2901 - 2918)	0.62 (0.62 - 0.62) 0.66 (0.66 - 0.66)	0.53 (0.51 - 0.54) 0.01 (0.00 - 0.02)	1.69 (1.67 - 1.71) 1.01 (1.00 - 1.02)	1.69 (1.67 - 1.71) 1.01 (1.00 - 1.02)	

ASA = acetylsalicylic acid; ICPC = International Classification of Primary Care diagnosis codes; ATC = Anatomical Therapeutic Chemical medication codes

\*The reference derivation and validation set is defined by one year run-in, imputation using MICE, and outcome definition based on ICPC or ATC codes (including aspirin)

\*\*Derivation set characteristics and performance metrics are given as average across 50 bootstrap samples

## DISCUSSION

This study shows that for the prediction of first-ever cardiovascular event risk using Dutch primary care EHR data, different data preparation choices regarding the outcome definition (first-ever cardiovascular events) and methods used to address missing values in the derivation set can have a substantial impact on model calibration, while model discrimination remains essentially the same. The large changes in calibration curve intercept and -slope could be explained by the changes in percentage of events that resulted from the different data preparation choices in the derivation set variations. A drop of the proportion of events in derivation set variations compared with the reference derivation set (e.g. defining outcome using only ICPC codes) led to a decrease in the calibration curve intercept, and a rise of the proportion of events (e.g. in case of using complete case analysis to handle missing values) led to an increase. These deteriorations of calibration may be of substantial clinical significance when a prediction model is used in clinical practice, for example within a clinical decision support tool. To evaluate a model on its utility to support clinical decisions, calibration is a more relevant performance metric than model discrimination.<sup>24, 25</sup>

Previous research already identified numerous methodological challenges for development of clinical risk prediction models using EHR data.<sup>6-8</sup> To the best of our knowledge, this is the first study that quantifies the impact that different data preparation choices in an EHR data setting have on model performance. The three data preparation challenges that are treated in this paper do relate to previous studies that focus on EHR-based data. One study used multiple methods for aggregation of baseline measurements during a run-in period and found that simple aggregations such as the mean are sufficient to improve model performance.<sup>26</sup> Further, several studies illustrate the difficulty of choosing an outcome definition in an EHR data context, especially due to the substantial variations of misclassification for different types of EHR diagnosis codes. In one example the positive predictive value (PPV) of the diagnosis code for chronic sinusitis was 34%, versus 85% for nasal polyps. With the additional information of evaluation by an otorhinolaryngologist the PPV of the latter rose to 91%.<sup>27, 28</sup> One study quantified the effect on model performance of misclassification in predictors instead of the outcome, using the CHA<sub>2</sub>DS<sub>2</sub>-VASC prediction rule as a case study. The substantial misclassification of predictors did not affect overall model performance, but it did affect the risk of the outcome with a certain CHA<sub>2</sub>DS<sub>2</sub>-VASC score.<sup>29</sup> In this study we focused on the influence of misclassification in outcome on model performance, but also misclassification in predictors should be taken into account when developing

a clinical prediction model using EHR data. Regarding the imputation of EHR predictor values that are likely MNAR, studies found that there may still be options for imputation if missingness structure is explicitly modelled. Methodologies such as Bayesian analysis may be specifically suited for this purpose.<sup>6,30</sup> However, further research into this topic is needed. One option is to discard a variable altogether, especially in case of large extent of missingness.<sup>19</sup> In the future, missingness in EHR data might be reduced by more systematic data capture, or through automated analysis of free text using natural language processing techniques.<sup>31</sup>

### **Strengths and limitations**

Several methodological limitations must be considered to interpret our study results. First, in our EHR data, no reference standard for the definition of the outcome was present, complicating the interpretation of the model results. It should also be noted that for many EHR-derived diagnoses, available reference standards may have a certain degree of misclassification.<sup>32</sup> Therefore, the researcher needs to work with the routine data that are available, often resulting in difficult or seemingly arbitrary choices regarding outcome definition. In this study, we focused on the relative impact on model performance of different outcome definitions instead of a comparison with a reference standard for outcome. We assumed that the definition used in the reference derivation set (ATC including acetylsalicylic acid or ICPC) was most sensitive because of the broad inclusion of thrombocyte aggregation inhibitors prescribed after cardiovascular events. However, in the first years of our follow-up period acetylsalicylic acid was also prescribed in a primary prevention setting, thus outcome according to ATC excluding acetylsalicylic acid is considered as most specific.

Second regarding the different choices in addressing missing data, in the reference derivation set systolic blood pressure and blood cholesterol were imputed using MICE despite the large extent of missingness in these predictors. As the predominant missingness mechanism is likely MNAR as has been argued in section 2.4.3, these imputation results are likely biased to some extent. The density of datapoints across all diagnosis, medication and measurement codes showed that for a large number of patients the lack of information often extended to the entire dataset, which also hampers reliable imputation. We compared imputation results with expected population means and indeed found a moderate difference. Although these likely biased estimates may not be a problem at internal validation, it may be at external or prospective validation when the missingness mechanism itself is not transportable to these new data environments. Third, although non-cardiovascular mortality could be considered as a competing event, we did not

perform a competing risk analysis to limit the complexity of analyses in this paper. The number of non-cardiovascular deaths recorded during follow-up was 2838, which represents only 3% of the total study population. Therefore, the effect of non-cardiovascular mortality as competing event on potential overestimation of the cumulative incidence of cardiovascular events was likely limited. Finally, the discriminative performance of our models is relatively low. An explanation for the relatively poor discrimination is the limited number of predictors selected for the model and the limited age range of 40 to 65 years, based on our conformity with the SCORE model. Discriminative performance found in our study however is not uncommon for clinical prediction models used in practice, and is comparable with that of e.g. the CHA<sub>2</sub>DS<sub>2</sub>-VASc prediction rule.<sup>33</sup> In addition, compared with discrimination calibration is of more interest to compare model performance because of the future intended use of the models to support clinical decisions.<sup>24</sup> Strengths of this study include the very large sample size of our routine care dataset, and the large number of derivation set variations (eight) that we used to assess the impact of difficult or seemingly arbitrary choices in data preparation on model performance.

### **Future considerations**

Our findings stress the importance of carefully considering differences data preparation choices between the population used for model derivation compared with the target population for model validation or deployment, because these differences may lead to substantial miscalibration. In essence this study's methodology of including multiple derivation set variations could be seen as a form of sensitivity analysis to assess transportability of the model to a clinical setting in which different data preparation choices are made. However, all data used in this study were derived from the same EHR data source (ELAN). Therefore, we could not formally test transportability across different EHR data sources. Still, this study further illustrates the need for transparent reporting of choices in model development studies and model calibration in validation studies. This could be done using e.g. the RECORD statement for reporting on data preparation choices using routinely collected health data in EHR, and the TRIPOD statement for reporting on clinical prediction model development.<sup>34, 35</sup>

## CONCLUSION

Our findings support that for developing clinical prediction models using EHR data, variations in data preparation choices regarding outcome definition and dealing with missing values may have substantial impact on model calibration, while discrimination remains essentially the same. It is, therefore, important to transparently report data preparation choices in model development studies and model calibration in validation studies.

## REFERENCES

1. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. *Ann. Intern. Med.* 2006;144:742-752
2. Canadian Electronic Library P, Canada Health Infoway. The emerging benefits of electronic medical record use in community-based care: full report. Toronto, ON: Canada Health Infoway; 2013.
3. Ohno-Machado L. Sharing data from electronic health records within, across, and beyond healthcare institutions: Current trends and perspectives. *J. Am. Med. Inform. Assoc.* 2018;25:1113
4. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA.* 2013;309:1351-1352
5. Spasoff RA. *Epidemiologic Methods for Health Policy.* New York: Oxford University Press I.
6. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *J. Am. Med. Inform. Assoc.* 2017;24:198-208
7. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J. Am. Med. Inform. Assoc.* 2018;25:969-975
8. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS (Wash DC).* 2013;1:1035
9. Lamberts H. WMOUPUI, international classification of primary care.
10. Methodology WCCfDS. Atc index with ddds. Oslo; norway. 2002
11. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in europe: The score project. *Eur. Heart J.* 2003;24:987-1003
12. Lika B, Kolomvatsos K, Hadjiefthymiades S. Facing the cold start problem in recommender systems. *Expert Systems with Applications.* 2014;41:2065-2073
13. Schneeweiss S, Rassen JA, Brown JS, Rothman KJ, Happe L, Arlett P, et al. Graphical depiction of longitudinal study designs in health care databases. *Ann. Intern. Med.* 2019;170:398-406
14. de Lusignan S, Valentin T, Chan T, Hague N, Wood O, van Vlymen J, et al. Problems with primary care data quality: Osteoporosis as an exemplar. *Inform. Prim. Care.* 2004;12:147-156
15. Pijnstilling op recept. 2008 PW, Jaargang 143 Nr 39.
16. Bouma M DGG, De Vries H, et al. NHG-Standaard Stabiele angina pectoris (M43) Versie 4.0. Nederlands Huisartsen Genootschap. 2019;12.
17. Rubin DB *Imd*, vol. 63, no. 3, pp. 581–592, 1976.
18. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. *J. Clin. Epidemiol.* 2006;59:1087-1091
19. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: Data analysis. *JMIR Med Inform.* 2018;6:e11
20. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Med. Res. Methodol.* 2010;10:7
21. Groenwold RHH. Informative missingness in electronic health record systems: The curse of knowing. *Groenwold Diagnostic and Prognostic Research* 2020;4:8

22. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: Bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med. Inform. Decis. Mak.* 2014;14:51
23. Bos G J-vdB, Ujcic-Voortman JK, Uitenbroek DG, Baan CA. Etnische verschillen in diabetes, risicofactoren voor hart- en vaatziekten en zorggebruik Resultaten van de Amsterdamse Gezondheidsmonitor 2004. RIVM rapport 260801002/2007
24. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology.* 2010;21:128-138
25. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic t, et al. Calibration: The achilles heel of predictive analytics. *BMC Med.* 2019;17:230
26. Goldstein BA, Pomann GM, Winkelmayr WC, Pencina MJ. A comparison of risk prediction methods using repeated observations: An application to electronic health records for hemodialysis. *Stat. Med.* 2017;36:2750-2763
27. Hsu J, Pacheco JA, Stevens WW, Smith ME, Avila PC. Accuracy of phenotyping chronic rhinosinusitis in the electronic health record. *Am J Rhinol Allergy.* 2014;28:140-144
28. Joan A. Casey BSS, Walter F. Stewart, Nancy E. Adler. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annual Review of Public Health* 2016 37:1, 61-81.
29. van Doorn S, Brakenhoff TB, Moons KGM, Rutten FH, Hoes AW, Groenwold RHH, et al. The effects of misclassification in routine healthcare databases on the accuracy of prognostic prediction models: A case study of the cha2ds2-vasc score in atrial fibrillation. *Diagn Progn Res.* 2017;1:18
30. E. Ford PR, P. Hurley, S. Oliver, S. Bremner, J. Cassell. Can the use of bayesian analysis methods correct for incompleteness in electronic health records diagnosis data? Development of a novel method using simulated and real-life clinical data. *Front. Publ. Health*, 8 (2020), p. 54, 10.3389/fpubh.2020.00054.
31. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One.* 2012;7:e30412
32. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: Challenges, recent advances, and perspectives. *J. Am. Med. Inform. Assoc.* 2013;20:e206-211
33. van Doorn S, Debray TPA, Kaasenbrood F, Hoes AW, Rutten FH, Moons KGM, et al. Predictive performance of the cha2ds2-vasc rule in atrial fibrillation: A systematic review and meta-analysis. *J. Thromb. Haemost.* 2017;15:1065-1077
34. Nicholls SG, Quach P, von Elm E, Guttman A, Moher D, Petersen I, et al. The reporting of studies conducted using observational routinely-collected health data (record) statement: Methods for arriving at consensus and developing reporting guidelines. *PLoS One.* 2015;10:e0125620
35. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. *BMJ.* 2015;350:g7594



## Chapter 3

# Cardiovascular Risk Prediction in Men and Women Aged Under 50 Years Using Routine Care Data

---

Hendrikus J. A. van Os, Jos P. Kanning, Tobias N. Bonten, Margot Rakers, Hein Putter, Mattijs E. Numans, Ynte M. Ruigrok, Rolf H. H. Groenwold, Marieke J. H. Wermer

J Am Heart Assoc. 2023 Apr 4;12(7):e027011

## ABSTRACT

**Background:** Prediction models for risk of cardiovascular events generally do not include young adults, and cardiovascular risk factors differ between women and men. Therefore, this study aimed to develop a prediction model for first-ever cardiovascular event risk in men and women aged 30–49, using a large Dutch electronic health record (EHR)-derived primary care population-based cohort and comparing complex data-driven models with Cox regression models.

**Methods and Results:** We included patients from the Dutch STIZON routine care database. Patients aged 30–49 years without cardiovascular disease, or prescription of statins or thrombocyte aggregation inhibitors prior to baseline were included. Outcome was defined as first-ever cardiovascular event. Our reference models were sex-specific Cox proportional hazards models based on traditional cardiovascular predictors. In addition, we developed Cox elastic net and random survival forests models, and used two other predictor subsets with the 20 or 50 most important predictors from all information available in the EHR, based on the Cox elastic net model regularization coefficients. For all models we assessed the C-index and calibration curve slopes at ten years of follow-up. We stratified our analyses based on the 30–39 and 40–49 years age groups at baseline. We included 542,141 patients (mean age 39.7 years, 51% women). During follow-up, 10,767 first-ever cardiovascular events occurred (incidence rate: 19.7 [95%CI: 19.3–20.1] per 10,000 person years). Cox elastic net predictor selection resulted in several non-traditional cardiovascular predictors that were ranked as important, including socioeconomic status score and hormonal contraceptive use in women specifically. Discrimination of reference models including traditional cardiovascular predictors for both women and men was moderate (women: C-index: 0.648; 95%CI: 0.645–0.652; men: C-index: 0.661; 95%CI: 0.658–0.664). In women and men, the Cox PH model including 50 most important predictors resulted in an increase in C-index (0.030 in women and 0.012 in men), and a net correct reclassification of 3.7% of the events in women and 1.2% in men compared with the reference model. After stratification of the 30–39 and 40–49 years age groups at baseline, discriminatory performance was attenuated for all Cox PH models in both women and men.

**Conclusions:** Sex-specific EHR-derived prediction models for first-ever cardiovascular events in the general population under 50 have moderate discriminatory performance. Data-driven predictor selection leads to identification of non-traditional cardiovascular predictors which modestly increase discriminatory performance of models and correct reclassification of events, particularly in women.

## INTRODUCTION

Cardiovascular events are a leading cause of disability and death worldwide.<sup>1</sup> In the last half century cardiovascular event-related mortality decreased continually. However, opportunities in primary prevention of cardiovascular events are still being missed.<sup>2</sup> Currently in Europe, decisions on preventive interventions in adults without prior cardiovascular disease aged 40–69 years are based on the absolute ten-year risk of cardiovascular events, resulting from the SCORE2 prediction model.<sup>3</sup> Early identification of individuals at high risk of cardiovascular events is beneficial, because atherosclerosis is a chronic process that starts early in life.<sup>4</sup> Therefore, early treatment of risk factors is beneficial, and accurate risk estimates applicable to younger persons are required.<sup>5</sup>

Evidence on sex differences between cardiovascular risk factors is mounting, which pleads for including sex-specific risk factors such as preeclampsia and combined oral contraceptive pill use in prediction models.<sup>6</sup> Derivation of sex-specific models for the prediction of cardiovascular risk in young individuals requires a large sample size. Pooling electronic health record (EHR) data results in large prospective cohorts, offering a great opportunity for the derivation of prediction models.<sup>7</sup> The QRISK3 prediction model for the risk of cardiovascular events is an example of leveraging information from the EHR, and has been successfully externally validated in the general population in the United Kingdom.<sup>8</sup> QRISK3 is a traditional regression model using predictors which are selected based on prior knowledge. However, because EHR-derived cohorts are constituted by both a large sample size and a very high number of potentially relevant predictors, complex data-driven modelling techniques may outperform traditional regression models in predicting the risk of cardiovascular event.<sup>9–11</sup>

This study aimed to develop sex-specific prediction models for first-ever cardiovascular event risk in patients aged 30–49 in a primary care setting, using data from a large Dutch EHR-derived population-based cohort. We assessed whether the data-driven selection of predictors and the use of complex prediction models offer an increase in predictive performance, compared with a Cox regression model using only traditional cardiovascular predictors.

## METHODS

### Data source

The research cohort in this study was derived from the STIZON database. STIZON directly receives data from EHRs of a large number of primary care providers throughout the Netherlands.<sup>12</sup> We only selected patients from general practice centers which were localized in catchment areas of hospitals participating in the STIZON network. This enabled us to link hospital ICD-9 and ICD-10 diagnoses to primary care data. The STIZON dataset contains ATC medication prescriptions from primary care pharmacies during follow-up time, and ICPC diagnosis codes for clinical entities in principle starting from birth.<sup>13,14</sup> ICD-9 and ICD-10 codes were available for all in-hospital diagnoses that occurred during follow-up. Inclusion criteria were an age of 30–49 at baseline, and subscription to a STIZON general practice center between January 1<sup>st</sup> 2007 and December 31<sup>st</sup> 2020 for at least one year, which was required because we defined the one-year as a run-in period. This run-in period was used for averaging the predictor values of laboratory or vital parameter assessments, if multiple of such measurements were present within this period. Exclusion criteria were cardiovascular disease, and use of statins or cardiovascular event-specific thrombocyte aggregation inhibitors at baseline. Follow-up time started at the end of the one year run-in period (January 1<sup>st</sup> 2008) or on the first general practice center subscription date after January 1<sup>st</sup> 2008. Patients were censored at the earliest date of the diagnosis of a first-ever fatal or non-fatal cardiovascular event, non-cardiovascular death, deregistration with any practice connected to the STIZON network, or the last upload of computerised data to the STIZON database (December 31<sup>st</sup> 2020). The ethics review board has provided a statement that this study was not subject to ethics review according to the Medical Research Involving Human Subjects Act (WMO). Because of the sensitive nature of the data collected for this study, data will need to be requested from a third party (STIZON).

### Outcome definition

First-ever cardiovascular events were defined using ICD-9, ICD-10 or ICPC codes for fatal and non-fatal acute myocardial infarction and stroke (including ischemic, hemorrhagic and unspecified stroke)

### Predictors

Predictors included demographics, symptoms and diagnoses other than fatal and non-fatal cardiovascular events, and were based on ICPC, ICD-9, and ICD-10 codes, prescribed medication coded according to the ATC classification, laboratory test

results performed in primary care, consultation dates and frequency.<sup>13,14</sup> In addition, the four-digit postal code area data was transformed into a socioeconomic status score based on income, education and occupation of the inhabitants.<sup>15</sup> ICPC, ICD-9, and ICD-10 codes and condition-specific ATC-codes were clustered based on clinical knowledge by two domain experts (HvO & MR) if multiple codes constituted the same clinical entity. An example is the grouping of different types of malignancy diagnoses into an overall malignancy predictor. For computational purposes, we only selected predictors that occurred in at least 0.1% of the total study population across the entire follow-up time, after clustering. All continuous predictors were standardized before analysis. Baseline information was assessed at the end of the one-year run-in period.

### **Missing value handling**

With respect to missing predictor values, we made a distinction between binary predictors – such as registration of a certain diagnosis or prescription of medication – and continuous predictors such as measurements of laboratory parameters or blood pressure. For all binary predictors, we assumed that the absence of an EHR registration meant the absence of the clinical entity itself, and therefore no imputation was performed. However, for continuous predictors such as vital parameter or laboratory assessments, imputation of missing values was required for inclusion in the prediction models. Because in routine healthcare data the majority of such assessments is only performed in a small subset of the population, the extent of missingness may be large and the underlying mechanism of missingness is likely missing not at random. Because in our dataset for all continuous laboratory or vital parameter assessments missingness exceeded 25%, we chose not to impute the missing values to limit the risk of biased predictor value imputations. We only used binary indicators in the analyses, which indicated whether the assessment had been performed or not.

### **Predictor selection**

We used two methods for the selection of predictors which were used to develop prediction models. First, for the reference models we chose the traditional cardiovascular risk factors age, sex, smoking (ever), and either an ICD-9, ICD-10 or ICPC diagnosis code or condition-specific ATC medication prescription code for hyperlipidemia, hypertension, and diabetes mellitus, based on prior evidence.<sup>16</sup> Since we excluded patients who received statin treatment at baseline, hyperlipidemia was based on diagnosis codes only. Second, we used data-driven predictor selection based on a Cox elastic net model ( $\alpha$  of 0.00058 for women,  $\alpha$  of 0.00072 for men; L1 to L2 regularization penalty ratio: 0.5) to select the most

important 20 and 50 predictors based on the absolute regularized coefficients of a sex-specific Cox elastic net model.

### **Model development**

The three different selections of predictors (traditional cardiovascular risk factors for the reference model, and the 20 and 50 most important predictors based on a Cox elastic net model) were used to develop Cox proportional hazards (PH) models, Cox elastic net models, and random survival forests. Models were developed for women and men separately. Cox elastic net models and random survival forests are more flexible than Cox PH models, because they include hyperparameters. Hyperparameters of Cox elastic net and random survival forests were optimized using predefined hyperparameter grids. To account for overfitting and internally validate our findings, we used a nested validation approach. First, the data was randomly split into a derivation and validation set, of respectively 80% and 20% of the population. Hyperparameter optimization was then performed on the derivation set, using 10-fold cross validation. Overall model performance was assessed using the hold-out validation set. We repeated this process 50 times using bootstrap resampling to assess variability in outcomes and to report empirical 95% confidence intervals. We did not take non-cardiovascular death into account as a competing event, since our population was young and non-cardiovascular mortality was expected to be very low. Model performance was defined by both model discrimination (concordance index or C-index) and calibration (calibration curve slope at ten years of follow-up). We expressed change in C-index between reference and other prediction models as difference relative to the full scale of the C-index, which is from 0.5 to 1. Further, we assessed net reclassification using the categorical net reclassification index (NRI). We chose a 2.5% ten-year absolute risk of first-ever cardiovascular events as threshold for high cardiovascular risk. This is in line with the European Society of Cardiology (ESC) guideline for prevention of CVD in individuals under 50 years, and implies that risk factor treatment should be considered. Our predefined absolute risk threshold of 2.5% is therefore of clinical importance.<sup>17</sup> In addition, we stratified our analyses based on two age groups (30–39 and 40–49 years at baseline). The 30–39 years age group is of particular interest, because the SCORE2 model starts at an age of 40. For all performance metrics we calculated empirical 95% confidence intervals (CI) by fitting a new model in each of the 50 bootstrap samples, and basing the CI on the standard deviation of the distribution of the performance metrics. Python version 3.10 was used for pre-processing and analysis of data. Our study adhered to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement for reporting.<sup>18</sup>

## RESULTS

We included 542,141 patients aged 30–49 years without prior CVD or statin use at baseline in this study, of whom 51% were women. During 5,461,316 person years of follow-up, a total of 10,767 first-ever cardiovascular events occurred. This resulted in an incidence rate of 19.7 (19.3–20.1) per 10,000 person years in the total population, 13.6 (13.2–14.0) in women and 26.2 (25.5–26.8) in men. Table 1 shows the baseline characteristics of men and women in the total study population. The average age was 39.7 years (SD  $\pm$  5.7). Systolic blood pressure was assessed in 6.6%, and total serum cholesterol in 2.4% of the total population. We, therefore, discarded continuous measurements and only included indicators of whether tests were performed.

**Table 1.** Baseline characteristics for women and men

Baseline characteristics	Women (n = 276,113)		Men (n = 266,028)	
	Cases (n = 3,800)	Controls (n = 272,313)	Cases (n = 6,915)	Controls (n = 259,113)
<b>Demographic features</b>				
Age (mean $\pm$ SD)	42.4 (5.0)	39.5 (5.7)	42.9 (4.8)	39.6 (5.6)
Socioeconomic status score (mean $\pm$ SD)	0.23 (0.75)	0.31 (0.71)	0.25 (0.74)	0.30 (0.72)
Follow-up time (median years $\pm$ IQR)	6.6 (3.8–9.4)	11.0 (8.3–13.0)	6.9 (4.0–9.6)	11.0 (8.0–13.0)
<b>Cardiovascular risk factors, n (%)</b>				
Smoking (current)	154 (4.1)	4897 (1.8)	264 (3.8)	5087 (2.0)
Hyperlipidemia	32 (0.8)	761 (0.3)	69 (1.0)	1261 (0.5)
Hypertension	157 (4.1)	3896 (1.4)	168 (2.4)	3339 (1.3)
Diabetes mellitus	43 (1.1)	1163 (0.4)	67 (1.0)	1295 (0.5)
<b>Measurements, n (%)*</b>				
Systolic blood pressure	485 (12.8)	20823 (7.6)	526 (7.6)	13907 (5.4)
Serum glucose	133 (3.5)	8245 (3.0)	171 (2.5)	4463 (1.7)
Total serum cholesterol	318 (8.4)	13585 (5.0)	468 (6.8)	12150 (4.7)

Cases = patients who suffered a first-ever cardiovascular event during follow-up; controls = all other patients

\*Any laboratory or vital parameter measurement during the one-year run-in period

Subsequently, after the data-driven selection of predictors using Cox elastic net models, the 20 most important predictors are shown in Table 2. Substantial differences in predictor importances were observed between women and men. For example, for women two female-specific risk factors (combined oral contraceptive use and intrauterine contraceptive use) are ranked in the top 20.

**Table 2.** Top 20 most important predictors for women and men separately

<b>Women (n = 276,113)</b>		<b>Men (n = 266,028)</b>	
<b>Predictor</b>	<b>Coef.*</b>	<b>Predictor</b>	<b>Coef.*</b>
Age	0.416	Age	0.533
Socioeconomic status score	0.115	Socioeconomic status score	0.101
Combined oral contraceptive use	0.070	Smoking: current	0.069
Antirheumatic medication	0.060	Antirheumatic medication	0.067
Gastroesophageal reflux medication	0.053	Diabetes mellitus	0.039
Smoking: current	0.052	Practice nurse contact for somatic complaints	0.035
Acetylsalicyc acid use	0.052	RAAS inhibitors	0.033
Comorbidity count	0.049	Psoriasis	0.031
RAAS inhibitors	0.045	Gastroesophageal reflux medication	0.027
Betablockers	0.043	Comorbidity count	0.026
Calcium channel blockers	0.040	Hyperlipidemia	0.019
Blood pressure measured last year	0.032	Epilepsia	0.019
Dermatological complaints	0.031	Calcium channel blockers	0.018
Intrauterine contraceptive use	0.030	Oral anticoagulant drugs	0.016
Hyperlipidemia	0.029	Esophageal disorders	0.014
Antibiotic use	0.028	Allergic rhinitis	0.014
Depression	0.027	Antibiotic use	0.014
HIV/AIDS	0.024	Alcohol use	0.014
Female sex organ complaints and symptoms	0.023	Kidney failure	0.014
Diabetes mellitus	0.023	Male sex organ complaints	0.014

\*Absolute, regularized coefficient of Cox elastic net models (women: alpha = 0.00058; men: alpha = 0.00062)

\*\*Comorbidity count: simple count of chronic conditions per patient.

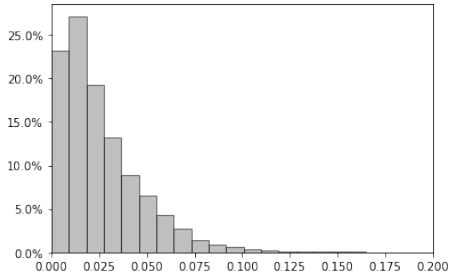
Discrimination of Cox PH reference models including traditional cardiovascular predictors for both women and men was moderate (women: C-index: 0.648; 95% CI: 0.645–0.652; men: C-index: 0.661; 95% CI: 0.658–0.664) and calibration was good (calibration curve slope in women: 0.999; 95% CI: 0.998–1.001; and in men: 1.001; 95% CI: 0.998–1.004; Table 3). In women, the Cox PH model including 50 most important predictors resulted in an increase in C-index of 0.030 compared with the reference model (20% difference with the reference model relative to the full scale of the C-index). In men, Cox PH model including 50 most important predictors also resulted in the relatively largest increase in C-index, although to a lesser extent compared with women (0.012 increase in C-index; 7% difference with the reference model relative to the full scale of the C-index). The more flexible modelling approaches (Cox elastic net and random survival forests) did not perform better than the Cox PH models across any of the different predictor subsets.

**Table 3.** Discrimination and calibration of sex-specific prediction models for different predictor subsets, stratified by age groups

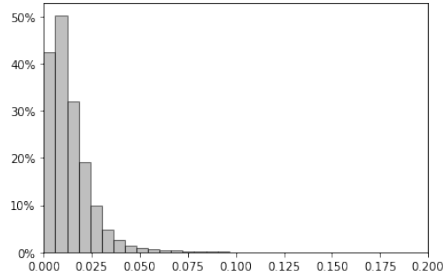
		Women (n = 276, 113)				Men (n = 266, 028)			
Age range	Predictors	Performance metrics (95% CI)		Calibration curve slope at 10 years	Δ C-stat.**	Performance metrics (95% CI)		Calibration curve slope at 10 years	Δ C-stat.**
		C-index	Δ C-stat.*			C-index	Δ C-stat.*		
30-49	Baseline	0.648 (0.645-0.652)	Ref.	0.999 (0.998-1.001)	Ref.	0.661 (0.658-0.664)	Ref.	1.001 (0.998-1.004)	Ref.
	20	0.674 (0.671-0.677)	0.026	1.000 (0.998-1.003)	18%	0.673 (0.670-0.676)	0.012	1.000 (0.998-1.002)	7%
	50	0.678 (0.675-0.681)	0.03	1.000 (0.997-1.002)	20%	0.673 (0.671-0.675)	0.012	1.001 (0.998-1.004)	7%
30-39	Baseline	0.605 (0.601-0.609)	Ref.	1.000 (0.998-1.003)	Ref.	0.608 (0.604-0.612)	Ref.	1.000 (0.998-1.003)	Ref.
	20	0.651 (0.646-0.654)	0.049	1.000 (0.997-1.003)	47%	0.629 (0.625-0.633)	0.021	1.001 (0.998-1.004)	19%
	50	0.658 (0.654-0.663)	0.053	0.999 (0.998-1.002)	50%	0.629 (0.626-0.633)	0.021	0.999 (0.996-1.002)	19%
40-49	Baseline	0.572 (0.568-0.576)	Ref.	0.999 (0.998-1.002)	Ref.	0.578 (0.574-0.583)	Ref.	1.001 (0.998-1.004)	Ref.
	20	0.619 (0.615-0.623)	0.047	1.000 (0.997-1.003)	65%	0.600 (0.596-0.605)	0.022	1.000 (0.997-1.003)	28%
	50	0.624 (0.619-0.628)	0.052	1.000 (0.997-1.002)	72%	0.601 (0.597-0.605)	0.023	1.001 (0.998-1.004)	29%

Baseline traditional cardiovascular predictors: age, hypertension, antihypertensive medication, diabetes mellitus, hyperlipidemia, with Cox PH model using baseline predictors as reference model

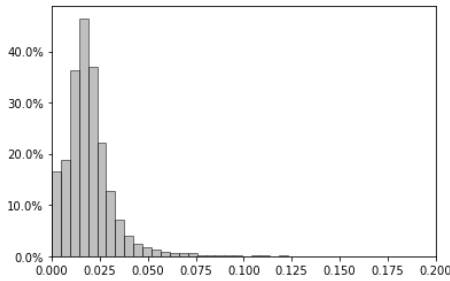
\*Difference in C-statistic compared with the reference model; \*\*Difference in C-statistic compared with the reference model relative to full scale



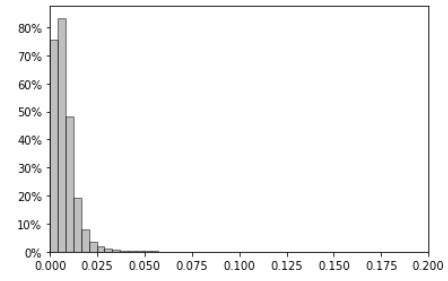
4a. Women aged 30-49 at baseline



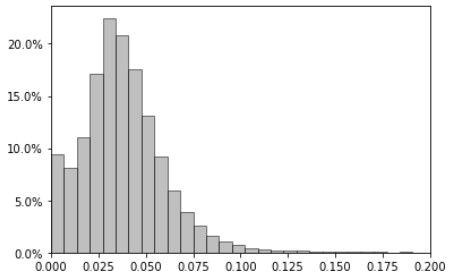
4b. Men aged 30-49 at baseline



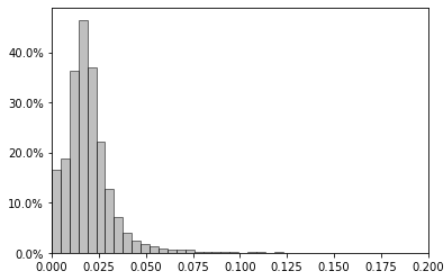
4c. Women aged 30-39 at baseline



4d. Men aged 30-39 at baseline



4e. Women aged 40-49 at baseline



4f. Men aged 40-49 at baseline

**Figure 1.** Absolute ten-year risk predictions of first-ever cardiovascular events including the 50 most important predictors, for women and men stratified by age groups.

On the X-axis the predicted probabilities from prediction models including the 50 most important predictors are shown, and on the Y-axis the fraction (%) of the total population in each bin. All histograms have a bin size of 100.

For women and men, the categorical NRI was assessed for the Cox PH model with 50 most important predictors versus the reference Cox PH model. For women, net correct reclassification was for events 3.7% (95% CI: 3.2%–4.2%), and for non-events 0.0% (-0.1%–0.1%); and for men, net correct reclassification for events was 1.2% (0.8%–1.6%), and for non-events was -0.8% (-1.1%–0.4%). Absolute risks for the Cox PH model with 50 most important predictors is shown for women and men (Figure 1).

After stratification of the 30–39 and 40–49 years age groups at baseline, discriminatory performance was attenuated in the 30–39 years age group, and further decreased in the 40–49 years age group, for all Cox PH models in both women and men (Table 3).

## DISCUSSION

We found that in an EHR-derived population-based cohort of primary care patients aged between 30–49, sex-specific prediction models for first-ever cardiovascular events had moderate discriminatory performance and were well calibrated. Compared with the reference Cox PH models, the Cox PH models based on the 50 most important predictors had better discriminatory performance in both women and men, and were well calibrated. In women the improvement in discrimination was more substantial as compared with men, and the net correct reclassification of events was 3.7%. The more complex modelling methods Cox elastic net and random survival forests did not result in improvements in discrimination or calibration compared with the reference model, regardless of the predictor subset that was chosen. After stratification of the age groups at baseline, we found that discriminatory performance was attenuated in the 30–39 years age group, and further decreased in the 40–49 years age group. This was as expected, because we restricted the range of age, which is the most important predictor for cardiovascular events.

Several previous studies reported on the prediction of cardiovascular events using large EHR-derived datasets and complex data-driven models. One study which used data from the CPRD database (n = 378,256 patients between 30–84 years at baseline) found that a neural network substantially outperformed a reference logistic regression model (C-index: 0.764 versus 0.728), and correctly reclassified 7.6% of events. However, no survival models were used which limits the possibilities for valid clinical implementation. Another study included 423,604 UK Biobank participants, and deployed an automated machine learning

pipeline named AutoPrognosis. Compared with a Cox PH reference model which included only traditional cardiovascular predictors, a machine learning ensemble method including all 473 predictors resulted in a C-index of 0.774 versus 0.734 of the reference model, and a net correct reclassification of events of 12.5%. An important difference with our study is that the UK Biobank contained relatively complete information on continuous predictors such as systolic blood pressure and total cholesterol.

In general, improvement in model performance may be due to (i) information gain resulting from including more predictors, or (ii) modelling gain which is the ability of models to capture non-linear associations or interactions among predictors.<sup>19</sup> In our study, the gain of complex (random survival forests) versus simple (Cox PH) models appeared to be limited. Random survival forests performed slightly more poorly compared with Cox regression models, potentially because of random forests methods are prone to overfitting.<sup>20</sup> We do seem to find information gain by including predictors which are ranked as most important according to Cox elastic net models. This indicates that data-driven predictor selection results in the identification of valuable non-traditional cardiovascular predictors which increase predictive performance, such as socioeconomic status score and hormonal contraceptive use in women specifically. Because Cox PH and Cox elastic net models have a similar performance, Cox PH models would be preferred for clinical use since they can be interpreted more easily.<sup>21</sup>

### **Limitations and strengths**

Our study has several limitations. First, EHRs are designed to record data that are routinely collected during the clinical workflow to streamline patient care, and not for the purpose of research.<sup>22</sup> Despite standardization using universal ICPC, ICD and ATC coding, previous research shows substantial underreporting in clinical diagnosis codes and large variability in inter-practice data quality.<sup>23</sup> Underreporting leads to misclassification in predictors and outcome. Misclassification is not a problem in prediction research if the measurement error is similar in development compared with the deployment setting. Misclassification of the outcome may, however, lead to a biased estimation of absolute risk.<sup>24</sup> Fatal cardiovascular events could only be identified if they occurred in-hospital using ICD-9 or ICD-10 codes. It is possible that in our study incidence of these events has been underestimated. Cardiovascular mortality comprises a quarter of all total CVD events. Prior research shows that the discriminating ability of prediction models did not differ between the fatal and non-fatal cardiovascular events.<sup>25</sup> Further, to optimally exclude patients with a history of cardiovascular events at baseline, we excluded patients

with prescriptions of thrombocyte aggregation inhibitors which were specific for cardiovascular events (clopidogrel, dipyridamole, ticagrelor) at baseline. We did not include acetylsalicylic acid in this definition because of its prescription as analgesic in the study period, hence specificity for cardiovascular events was low.<sup>26</sup> In addition, we did not develop lifetime risk models in this cohort of young patients, because of the risk of misclassification in predictors and outcome may aggravate cohort effects. Second, we did not take non-cardiovascular death into account as a competing risk because we assessed a young patient cohort at a maximum of 49 years at baseline. In this population, the cumulative incidence of non-cardiovascular death was very small (0.6%) compared with the entire population, limiting the competing risk effect on the estimation of stroke risk. It should however be noted that registration of mortality in our EHR data is of suboptimal quality. Third, the reference Cox PH model did not include continuous laboratory or vital parameter measurements such as systolic blood pressure and total serum cholesterol, which limits the head to head comparison with commonly used models such as SCORE2.<sup>3</sup> However, such a comparison was not the purpose of this study. In addition, because we use data-driven selection of predictors, we identified predictor representations other than continuous measurements of blood pressure and cholesterol that did not require imputation. This is an advantage because of the often very high extent of missingness of measurement data in the EHR. Fourth, our study population excluded patients receiving statin at baseline, which limits its use in patients already receiving statin treatment. However, our prediction models are specifically suited to support preventive interventions such as initiation of statin treatment, similar to the QRISK3 study in the United Kingdom, which is also based on EHR data.<sup>8</sup> We did not choose to exclude patients who received antihypertensive but not statin treatment at baseline, since in these patients the clinical decision on the initiation of statin treatment is also relevant and our models could be used for this decision. Fifth, although the continuous NRI is a more sensitive measure to assess model reclassification, we chose the categorical NRI because the 10-year risk threshold of 2.5% represents a clinically relevant threshold.

Strengths of this study includes the very large sample size of a cohort of patients under 50 years at baseline which is to our best knowledge among the largest to date. This offered a unique possibility to study data driven methods for the prediction of cardiovascular events in young patients. Further, all predictors used in our models are directly available in the EHR, which facilitates implementation of the models directly into the EHR. In addition, the linking of primary care and hospital diagnosis codes in the STIZON cohort enables validation of the cardiovascular outcome. Further, the data-driven predictor selection procedure results in that our

models leverage predictive information from predictors other than continuous measurements of traditional cardiovascular predictors. Therefore, it is not necessary to impute these continuous measurements, which were missing in the vast majority of patients in our population.

### **Clinical implications**

Our EHR-derived models will not replace traditional models such as SCORE2, but could be used in a two-step population health approach. First, at any given time point our models can automatically identify patient subgroups at increased risk for first-ever cardiovascular events above the absolute ten-year risk cut-off as specified by the ESC prevention guideline. Second, these patients subgroups could be invited to the primary care practice center for further cardiovascular risk assessment including measurement of systolic blood pressure and total- and HDL-cholesterol, after which traditional models such as SCORE2 could be used to estimate individualised risk. A previous modelling study found that such stepped strategy may result in more cost-effective cardiovascular risk management than the current opportunistic screening.<sup>27</sup> The ESC guideline states 2.5% ten-year risk of cardiovascular events as the threshold between moderate and high risk for women and men under 50 years, high risk being an indication for preventive pharmacotherapeutics. Although for patients under 50 years in our cohort absolute ten-year risks are generally low, our data-driven models can be used to automatically identify patients whose absolute risk reaches the 2.5% risk cut-off. In women, we found that the Cox PH model with 50 most important predictors resulted in a net correct reclassification of events (3.7%) around this risk cut-off compared with the reference model. Although this percentage is low, application on a large scale could lead to sufficient clinical impact to justify the use of a relatively more complex model. After stratification based on the 30–39 and 40–49 year age groups, we found that men and women between the age of 30–39 years at baseline had substantially lower absolute risks of cardiovascular events compared with those aged between 40–49 years. However, since the ESC guideline uses the SCORE2 model which does not include patients under 40 years, the absolute risk threshold of 2.5% likely is too high for individuals between the age of 30–39 years. Therefore, to define meaningful thresholds that can guide preventive therapy, we call for further research into the age group of 30–39 years. The focus may in this context not be pharmacotherapeutic, but rather on lifestyle interventions for prevention of cardiovascular disease. In addition, for the 30–39 years age group lifetime risk estimation may further help in risk communication and interpretation. However, we should first invest in the creation of higher quality longitudinal data sources to derive valid lifetime risk prediction models. In addition, data-driven predictor

selection has led to the identification of important non-traditional cardiovascular predictors such as socioeconomic status score and NSAID use. After stratifying for age subgroups, we found differences in the ranking of the 20 predictors that were most important in our prediction models. For example, in both women and men aged 30–39 years at baseline, the relative importance of NSAID use further increased compared with the 40–49 years age group.

## CONCLUSION

Sex-specific EHR-derived prediction models for first-ever cardiovascular events in the general population under 50 have moderate discriminatory performance and are well calibrated. Data-driven predictor selection leads to identification of non-traditional cardiovascular predictors, which modestly increase discriminatory performance of models and correct reclassification of events, mostly in women.

## REFERENCES

1. Mendis S. Global status report on noncommunicable diseases 2014: World Health Organization hawibhepjACABC.
2. van der Ende MY, Sijtsma A, Snieder H, van der Harst P. Letter to editor: Reply on question of marques jr et al. Regarding the paper entitled: "The lifelines cohort study: Prevalence and treatment of cardiovascular disease and risk factors". *Int. J. Cardiol.* 2019;294:57
3. Score working group. Score2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in europe. *Eur. Heart J.* 2021;42:2439-2454
4. Ference BA, Ginsberg HN, Graham I, Ray KK, Packard CJ, Bruckert E, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the european atherosclerosis society consensus panel. *Eur. Heart J.* 2017;38:2459-2472
5. Graham IM, Di Angelantonio E, Visseren F, De Bacquer D, Ference BA, Timmis A, et al. Systematic coronary risk evaluation (score): Jacc focus seminar 4/8. *J. Am. Coll. Cardiol.* 2021;77:3046-3057
6. Appelman Y, van Rijn BB, Ten Haaf ME, Boersma E, Peters SA. Sex differences in cardiovascular risk factors and disease prevention. *Atherosclerosis.* 2015;241:211-218
7. Ohno-Machado L. Sharing data from electronic health records within, across, and beyond healthcare institutions: Current trends and perspectives. *J. Am. Med. Inform. Assoc.* 2018;25:1113
8. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ.* 2017;357:j2099
9. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One.* 2018;13:e0202344
10. Weng SF, Reys J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One.* 2017;12:e0174944
11. Alaa AM vdSM. Autoprognosis: Automated clinical prognostic modeling via bayesian optimization with structured kernel learning. International conference on machine learning. 2018.
12. Kuiper JG, Bakker M, Penning-van Beest FJA, Herings RMC. Existing data sources for clinical epidemiology: The pharma database network. *Clin. Epidemiol.* 2020;12:415-422
13. Lamberts H. WM. Oxford university press; USA: 1987. Icdpc, international classification of primary care.
14. WHO. Collaborating centre for drug statistics methodology. Atc index with ddds. Oslo; norway. 2002
15. Sociaal Cultureel Planbureau Nssbph, [www.scp.nl/Onderzoek/Lopend\\_onderzoek/A\\_Z\\_alle\\_lopende\\_onderzoeken/Statusscores](http://www.scp.nl/Onderzoek/Lopend_onderzoek/A_Z_alle_lopende_onderzoeken/Statusscores), (Updated). Adj.
16. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ.* 2016;353:i2416
17. Visseren FLJ, Mach F, Smulders YM, Carballo D, Koskinas KC, Back M, et al. 2021 esc guidelines on cardiovascular disease prevention in clinical practice. *Eur. Heart J.* 2021;42:3227-3337
18. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. *BMJ.* 2015;350:g7594
19. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants. *PLoS One.* 2019;14:e0213653

20. Ishwaran H KU, Blackstone EH, Lauer MS. Random survival forests, the annals of applied statistics, 2008, vol. 2 (pg. 841-860).
21. James G, Witten, D., Hastie, T., & Tibshirani, R. An introduction to statistical learning (1st ed.). Springer. 2013
22. Spasoff RA. Epidemiologic Methods for Health Policy. New York: Oxford University Press I.
23. de Lusignan S, Valentin T, Chan T, Hague N, Wood O, van Vlymen J, et al. Problems with primary care data quality: Osteoporosis as an exemplar. *Inform. Prim. Care.* 2004;12:147-156
24. Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J. Clin. Epidemiol.* 2019;105:136-141
25. van Dis I, Geleijnse JM, Boer JM, Kromhout D, Boshuizen H, Grobbee DE, et al. Effect of including nonfatal events in cardiovascular risk estimation, illustrated with data from the netherlands. *Eur J Prev Cardiol.* 2014;21:377-383
26. Pijnstilling op recept. 2008;Pharmaceutisch Weekblad, Jaargang 143 Nr 39
27. Crossan C, Lord J, Ryan R, Nherera L, Marshall T. Cost effectiveness of case-finding strategies for primary prevention of cardiovascular disease: A modelling study. *Br. J. Gen. Pract.* 2017;67:e67-e77



## Chapter 4

# Prediction of aneurysmal subarachnoid hemorrhage in comparison with other stroke types using routine care data

---

Jos P. Kanning, Hendrikus J.A. van Os, Margot Rakers, Marieke J.H. Wermer, Mirjam I. Geerlings, Ynte M. Ruigrok

PLoS One. 2024 May 31;19(5):e0303868

## ABSTRACT

Aneurysmal subarachnoid hemorrhage (aSAH) can be prevented by early detection and treatment of intracranial aneurysms in high-risk individuals. We investigated whether individuals at high risk of aSAH in the general population can be identified by developing an aSAH prediction model with electronic health records (EHR) data. To assess the aSAH model's relative performance, we additionally developed prediction models for acute ischemic stroke (AIS) and intracerebral hemorrhage (ICH) and compared the discriminative performance of the models. We included individuals aged  $\geq 35$  years without history of stroke from a Dutch routine care database (years 2007-2020) and defined outcomes aSAH, AIS and ICH using International Classification of Diseases (ICD) codes. Potential predictors included sociodemographic data, diagnoses, medications, and blood measurements. We cross-validated a Cox proportional hazards model with an elastic net penalty on derivation cohorts and reported the c-statistic and 10-year calibration on validation cohorts. We examined 1,040,855 individuals (mean age 54.6 years, 50.9% women) for a total of 10,173,170 person-years (median 11 years). 17,465 stroke events occurred during follow-up: 723 aSAH, 14,659 AIS, and 2,083 ICH. The aSAH model's c-statistic was 0.61 (95%CI 0.57-0.65), which was lower than the c-statistic of the AIS (0.77, 95%CI 0.77-0.78) and ICH models (0.77, 95%CI 0.75-0.78). All models were well-calibrated. The aSAH model identified 19 predictors, of which the 10 strongest included age, female sex, population density, socioeconomic status, oral contraceptive use, gastroenterological complaints, obstructive airway medication, epilepsy, childbirth complications, and smoking. Discriminative performance of the aSAH prediction model was moderate, while it was good for the AIS and ICH models. We conclude that it is currently not feasible to accurately identify individuals at increased risk for aSAH using EHR data.

## INTRODUCTION

Aneurysmal subarachnoid hemorrhage (aSAH) is caused by the rupture of an intracranial aneurysm.<sup>1</sup> The age-standardized incidence rate of aSAH (14.5 per 100,000 people) is lower than that of acute ischemic stroke (AIS; 94.5 per 100,000 people) and intracerebral hemorrhage (ICH; 41.8 per 100,000 people).<sup>2</sup> However, because of aSAH's early onset (mean age of 50 years) and high morbidity and mortality rates, aSAH's loss of productive life-years is comparable to that of AIS.<sup>3</sup>

Unlike other stroke types, aSAH incidence can be reduced by early identification of individuals at increased risk of aSAH, followed by preventive endovascular or neurosurgical treatment of any aneurysms found.<sup>4</sup> Such early identification is already possible through screening individuals who have a known increased risk of aSAH, such as individuals with a first-degree relative with aSAH.<sup>5</sup> Other high-risk individuals who may be eligible for aneurysm screening could be identified by estimating the absolute risks of developing aSAH in the general population.

Prediction models constructed with electronic health records (EHRs) may provide a novel and cost-effective strategy of identifying individuals at increased risk of aSAH. Such prediction models generate individual risk estimates by integrating readily available data such as patient demographics, medical history, and clinical characteristics.<sup>6-8</sup> Although there are currently no such EHR-derived prediction models for aSAH, prediction models for cardiovascular risk (including AIS and transient ischemic attack) do exist and are widely used.<sup>9-13</sup>

We aimed to identify individuals at increased risk of aSAH by developing an aSAH prediction in the general population using data from a large Dutch EHR-derived population-based cohort. We additionally developed models for stroke outcomes with different incidence rates (AIS and ICH) and compared the discriminative performance of the three models in order to test the aSAH model's relative performance.

## METHODS

### Cohort definition

The cohort was derived on the 2nd of June, 2021, from the STIZON (Stichting Informatievoorziening voor Zorg en Onderzoek) dataset. STIZON is a foundation that acquires, manages, anonymizes, and processes EHRs from a large number of Dutch primary care physicians.<sup>14</sup> These data include diagnoses, blood measurements (e.g.

blood pressure and cholesterol levels) and prescriptions for medications. Primary care diagnoses are encoded with International Classification of Primary Care (ICPC) codes, whereas drug prescriptions are encoded with Anatomical Therapeutic Chemical (ATC) codes. For all in-hospital diagnoses occurring during follow-up International Classification of Diseases (ICD) codes versions 9 and 10 are available.

We selected all patients from general practice centers located within the hospital catchment locations of STIZON network hospitals. This allowed us to link primary care and hospital data, resulting in a completely linked EHR for each individual. Additional inclusion criteria were as follows:  $\geq 35$  years of age at baseline (as stroke before this age is rare),<sup>15</sup> no history of stroke prior to baseline assessment and a minimum follow-up period of one year between. For each individual, we defined a baseline date as the first recorded entry between January 1<sup>st</sup>, 2007 and January 1<sup>st</sup>, 2021. We used a one-year run-in period after the baseline date to aggregate numerical values, counting discrete values (e.g. number of consultations) and averaging continuous measurements (e.g. blood pressure). The individual follow-up time started at the conclusion of the one year run-in period and ended at the earliest date of outcome registration, non-outcome-related mortality, deregistration, or the end of study period.

### **Outcome definition**

The stroke types aSAH, AIS and ICH were defined by ICD-9 and ICD-10 hospital codes. ICD-9: 430; ICD-10: I60.\* for aSAH, ICD-9: 433.\*, 434.\*, 436; ICD-10: 163.\*, I64 for AIS and ICD-9: 431; ICD-10: I61.\* for ICH, with an asterisk indicating that all subcodes were included. Individuals who had a primary care ICPC code for stroke (aSAH: K90.01; AIS: K90.03; ICH: K90.02) without an associated ICD code were excluded.

### **Predictors**

We included all information available to the GP as potential predictors: Diagnoses other than stroke (e.g. hypertension, diabetes), medication use, demographics, lifestyle measurements (e.g. smoking status), symptoms (e.g. fever,, headache), primary care laboratory blood measurements (e.g. glucose, creatinine), and general practitioner consultation frequencies. Diagnoses were clustered based on a combination of ICPC, ICD-9, ICD-10 codes and ATC-coded medications. Two clinicians (H.v.O. and M.R.) clustered ICPC, ICD-9, ICD-10, and condition-specific ATC-codes based on clinical knowledge when multiple codes referred to the same medical entity. For instance, the presence of hypertension could be defined as either ICPC codes K86/K87, ICD-9 codes 401/402 or ICD-10 codes I10/I11. Comorbidity scores, based on the clustered medical entities, were calculated using the Charlson comorbidity index.<sup>16</sup> In addition, we used the four-digit postal code

of each patient to derive neighborhood population density and socioeconomic status (based on income, education, and occupation of the inhabitants).<sup>17</sup> For computational purposes, we only selected predictors that were present in at least 0.1% of the study population. All predictors were assessed at the end of the one-year run-in period and were held static during follow-up.

### Missing values

Regarding missing predictor values, we distinguished between binary (e.g., diagnoses, prescriptions) and continuous (e.g., blood pressure) values. For missing binary values, we assumed that their absence in the EHR indicated their absence in the patient, requiring no imputation. For missing continuous values, we decided not to impute in order to minimize bias due to the large amount of missingness (for example, over 50% of patients had no blood pressure measurements available) and the likely non-random nature of missing data.<sup>18</sup> Instead, we created a binary variable (so-called missingness indicator) to indicate whether the measurement had been performed or not. The missingness indicator was incorporated as a predictor in the statistical model, under the assumption that EHR-data is missing not at random and thus provides predictive information about the patient and the missing variable itself.<sup>19</sup>

### Model development

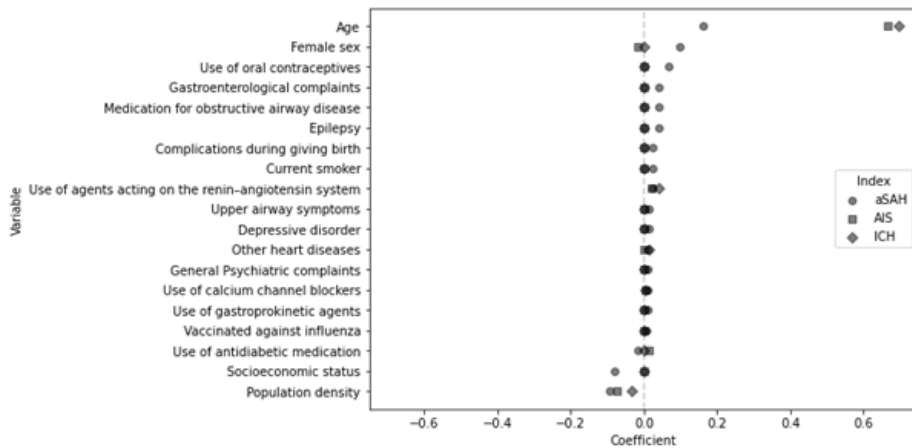
We developed a Cox proportional hazard model with an elastic net penalty for each of the three stroke outcomes (aSAH, AIS and ICH). All potential predictors were included in the model. The Cox elastic net model efficiently handles multicollinearity among predictors by blending features of Lasso and Ridge regression.<sup>20</sup> By shrinking coefficients of less predictive predictors towards zero, it emphasizes the most informative variables, ensuring model robustness and minimizing overfitting in datasets with many predictors.<sup>21</sup> To further account for overfitting, we separated the original dataset into a derivation set of 70% and a validation set of the remaining 30%. Stratified splitting was used to maintain comparable outcome proportions between the two sets. We converted continuous values to z-scores separately for both sets. We used five-fold cross validation to tune the alpha parameter in the derivation set (number of alphas: 10, minimum alpha ratio of 0.3) and selected the model with the lowest average error across folds to generate predictions on the withheld validation set. We used the predictions on the validation set to assess discrimination and calibration. Discrimination was assessed by a bootstrapped c-statistic, whereas calibration was assessed by generating Kaplan-Meier plots for the predicted versus observed 10-year probabilities of survival. We additionally reported the important predictors (i.e. the non-zero coefficients) for each outcome. We did not report confidence intervals for these coefficients because penalized

regression techniques result in biased coefficient estimates, making traditional confidence intervals potentially misleading or inappropriate for conveying uncertainty. All analyses were performed in Python version 3.<sup>22</sup> Our study adhered to both the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement,<sup>23</sup> and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement.<sup>24</sup>

The ethics review board has provided a statement that this study was not subject to ethics review according to the Medical Research Involving Human Subjects Act wet medisch onderzoek. Because of the sensitive nature of the data collected for this study, data will need to be requested from a third party (STIZON).

## RESULTS

We included 1,040,855 individuals with a mean age of 54.6 (SD 13.3) of whom 50.9% were women. During a total of 10,029,958 follow-up years (median 11.0 years, IQR 5.0 years) 17,465 stroke events occurred, which included 723 aSAH (4.2%), 14,659 AIS (83.9%), and 2,083 (11.9%) ICH events. Table 1 shows the baseline characteristics of the aSAH, AIS and ICH patients and the reference group.



**Figure 1.** All predictors of the aneurysmal subarachnoid hemorrhage (aSAH) prediction model in relation to the corresponding coefficients of these predictors in the acute ischemic stroke (AIS) and intracerebral hemorrhage (ICH) prediction models.

Each coefficient corresponds to the log hazard ratios after applying elastic net penalties. 0 indicates that the predictor is not predictive for that outcome. No confidence intervals are reported for these coefficients because penalized regression techniques result in biased coefficient estimates, making traditional confidence intervals potentially misleading or inappropriate for conveying uncertainty.

The elastic net Cox model identified 19 predictors for aSAH (Table 2 and Fig 1). Age was the strongest predictor, followed by female sex, population density, socioeconomic status, oral contraceptive use, gastroenterological complaints, obstructive airway medication, epilepsy, childbirth complications, smoking, angiotensin-converting enzyme (ACE) inhibitors use, antidiabetic medication use, upper airway symptoms and depressive disorder, heart disease, general psychiatric complaints, calcium channel blockers use, gastroprokinetic agent use, and influenza vaccination. The model for AIS identified 11 predictors while the ICH model identified 10 predictors (S1 and S2 Tables and S1 and S2 Figs). Identified predictors common to all stroke types included age, population density, ACE inhibitors use, and calcium channel blockers use. The coefficient of age for predicting AIS and ICH was approximately four times higher than that for the aSAH model (Table 2 and Fig 1).

**Table 1.** Baseline characteristics.

	aSAH	AIS	ICH	No stroke
<b>Demographic features</b>				
N	723	14,659	2,083	1,008,617
Age, mean (SD)	56.5 (12.0)	65.7 (12.2)	66.5 (12.6)	54.3 (13.2)
Women, n (%)	449 (62)	6,544 (45)	965 (46)	515,916 (51)
Socioeconomic status score, mean (SD)	0.2 (0.7)	0.2 (0.7)	0.3 (0.7)	0.3 (0.7)
Follow-up time in days, median (IQR)	2,148 (2,287)	2,398 (2,168)	2,413 (2,241)	4,018 (1,827)
Charlson Comorbidity Index (SD)	0.4 (0.7)	0.5 (0.8)	0.6 (0.9)	0.3 (0.6)
<b>Cardiovascular risk factors, n (%)</b>				
Smoking, current	22 (3)	397 (3)	32 (2)	15,924 (2)
Hyperlipidemia	25 (3)	459 (3)	60 (3)	19,474 (2)
Hypertension	49 (7)	1,378 (9)	199 (10)	47,941 (5)
Diabetes	19 (3)	801 (5)	93 (4)	23,119 (2)
<b>Measurements, mean (SD)</b>				
Systolic blood pressure (mmHg)	144.0 (14.1)	144.8 (15.9)	144.2 (16.2)	139.7 (16.2)
Serum glucose (mmol/L)	7.1 (1.9)	7.2 (1.5)	6.9 (1.6)	7.0 (1.7)
HbA1c (mmol/mol)	45.9 (11.1)	49.0 (11.2)	48.4 (9.6)	46.6 (11.0)
Creatinine ( $\mu$ mol/L)	78.2 (16.5)	83.3 (18.1)	83.7 (20.3)	79.2 (16.9)

aSAH = aneurysmal subarachnoid hemorrhage, AIS = acute ischemic stroke, ICH = intracerebral hemorrhage, SD = standard deviation, IQR = interquartile range.

Discrimination of the aSAH prediction model was moderate with a c-statistic of 0.61 (95%CI 0.57-0.65) whereas discrimination of the models for AIS and ICH was good;

the AIS model had a c-statistic of 0.77 (95%CI: 0.77 - 0.78) and the ICH model had a c-statistic of 0.77 (95%CI 0.75-0.78). The calibration showed good correspondence between predicted and observed risk for all models, with slight overprediction of the risk of AIS and ICH (S3 Fig).

**Table 2.** All predictors of the aneurysmal subarachnoid hemorrhage (aSAH) prediction model. The corresponding coefficients for these predictors in the intracerebral hemorrhage (ICH) and acute ischemic stroke (AIS) models are also shown.

	aSAH	AIS	ICH
Age	0.163	0.665	0.695
Female sex	0.098	-0.018	0.000
Use of oral contraceptives	0.068	0.000	0.000
Gastroenterological complaints	0.042	0.000	0.000
Medication for obstructive airway disease	0.040	0.000	0.000
Epilepsy	0.039	0.000	0.000
Complications during giving birth	0.025	0.000	0.000
Current smoker	0.024	0.000	0.000
Use of angiotensin-converting enzyme (ACE) inhibitors	0.024	0.022	0.039
Upper airway symptoms	0.014	0.000	0.000
Depressive disorder	0.013	0.000	0.000
Other heart diseases	0.012	0.000	0.015
General psychiatric complaints	0.011	0.000	0.000
Use of calcium channel blockers	0.010	0.005	0.003
Use of gastroprokinetic agents	0.010	0.000	0.000
Vaccinated against influenza	0.006	0.000	0.000
Use of antidiabetic medication	-0.016	0.013	0.000
Socioeconomic status	-0.080	0.000	0.000
Population density	-0.095	-0.073	-0.034

## DISCUSSION

Within our large Dutch EHR-derived population-based cohort individuals at increased risk for aSAH appeared more difficult to identify than individuals at risk for AIS and ICH. The prediction model for aSAH had moderate discriminatory performance, whereas the prediction models for AIS and ICH models both had good discriminatory performance. Age was the most important predictor in all three models, but its coefficient was notably larger in the model for AIS and ICH

than in the aSAH model. Besides age, other predictors identified by the aSAH model included already established risk factors (e.g. female sex, smoking) and potentially new prognostic factors (e.g. oral contraceptive use, gastroenterological complaints, obstructive airway medication, epilepsy, childbirth complications).

While there are no comparable aSAH prediction models in the general population to which our findings can be directly compared, a number of studies have developed prediction models for stroke (including AIS and ICH). A recent machine learning model trained on EHR data reported stroke prediction accuracy scores between 0.68 and 0.77.<sup>25</sup> QStroke, a scoring model for predicting the risk of AIS in primary care, generated ROC statistics between around 0.81 for patients older than 35.<sup>26</sup> A similar model developed to predict AIS and transient ischemic stroke in a retrospective cohort produced AUC scores ranging from 0.68 to 0.77.<sup>27</sup> The similarity between these outcome metrics and the c-statistics found for our AIS model suggests that our findings are representative of the predictive performance of stroke models developed for use in the general population.

Several factors may account for the aSAH model's poorer discriminative performance compared to other stroke outcomes. First, aSAH has the lowest incidence among stroke types.<sup>2</sup> With an age-standardized incidence rate of 14.5 per 100,000 reported in a recent study,<sup>2</sup> aSAH has a considerably lower incidence than AIS (94.5 per 100,000) and ICH (41.8 per 100,000). This relative rarity leads to less available data for model training and validation, thereby possibly affecting the performance of the aSAH prediction model compared to those for AIS and ICH. Second, the lower performance of the aSAH model may be attributable to aSAH's young age of onset. Consistent with the findings of our study, aSAH occurs most frequently before the age of 60 with approximately half of the patients being younger than 50,<sup>1</sup> whereas the incidence of AIS and ICH peaks after the age of 60.<sup>28,29</sup> Age was thus a less useful discriminator between aSAH (mean age 56.5) and the reference group (54.3) in our dataset than it was for AIS (65.7) and ICH (66.5). While increased age correlates with various stroke risk factors,<sup>30</sup> it also correlates with increased opportunities for interaction with healthcare, increasing the likelihood of identifying possible stroke risk factors. In our dataset, younger individuals were less likely than older individuals to have available data on known stroke risk factors such as smoking status, alcohol use, blood pressure, and glucose levels.<sup>31-35</sup>

While there are no models for predicting aSAH in the general population, models have been developed for individuals with known aneurysm presence. The PHASES score, which incorporates both aneurysm (size, location) and patient characteristics

(ethnicity, hypertension, age, history of aSAH) to predict aneurysmal rupture risk, achieved a c-statistic of 0.82 (95%CI 0.79–0.85).<sup>36</sup> This discriminative performance is most likely attributable to the inclusion of patients with aneurysm presence for whom detailed information on the risk factors (e.g., smoking, hypertension) is available, coupled with the fact that the prior chance of aSAH was higher within this specific patient population compared to the general population.

The Cox model identified a number of predictors for aSAH. Age, female sex, lower socioeconomic status, and smoking are aSAH predictors that are consistent with previous studies.<sup>32,37</sup> The model did not identify known predictors of aSAH such as hypertension or diabetes,<sup>32,33</sup> but we did find a non-zero coefficient for medications prescribed for these indications. The negative association between aSAH and antidiabetic medications may reflect the lower incidence of aSAH among diabetics,<sup>38,39</sup> while the positive association between aSAH and antihypertensives (e.g. ACE-inhibitors, calcium channel blockers) may reflect the higher incidence of aSAH among people with hypertension.<sup>32</sup> The model's identification of women-specific predictors, such as childbirth complications and the use of oral contraceptives, likely reflects the higher incidence of aSAH in women and suggests a promising direction for future research into sex-specific predictors.<sup>40</sup> Epilepsy was identified as a predictor in our model. While epilepsy itself isn't directly linked to aSAH, the observed effect might be due to the association between antiepileptic medication and aSAH.<sup>41</sup> Other identified predictors (i.e. gastroenterological complaints, medication for obstructive airway disease, depression, and influenza vaccination) have, to the best of our knowledge, not been previously studied in relation to aSAH and require further investigation.

This study has several strengths. First, given the low aSAH incidence,<sup>2</sup> we used a large Dutch dataset with over a million individuals and a long follow-up time to identify a relatively large number of aSAH cases to develop our prediction model with. Second, we developed our model using routine care data, which means that the variables identified by the model are routinely available or easily ascertainable by general practitioners during a standard consultation.

This study also has several limitations. First, we used data from EHRs and as EHRs reflect routine care, EHR-based data are not systematically collected, but rather reflect standard care practice.<sup>42</sup> Most individuals therefore lack a comprehensive and systematic medical history, e.g. blood pressure is rarely measured in relatively healthy individuals.<sup>43</sup> This lack of systematic data collection may explain why we did not find associations between well-known risk factors (such as hypertension)<sup>32</sup>

and the stroke outcomes. The lack of systematic measurements likely also explains the relative importance given to variables that were available for each individual (e.g. age, sex, socioeconomic status). Second, we assessed the general population as a whole, whereas in practice an aSAH prediction model may have a better performance in a population with a known established increased risk, such as in smokers with hypertension.<sup>44</sup> Due to the rarity of aSAH and the non-systematic collection of EHR data for smoking and hypertension, we unfortunately lacked sufficient cases in our dataset to evaluate aSAH risk in this subpopulation. Thirdly, although we utilized calibration plots to evaluate the accuracy of our model, it is important to note that in datasets where the outcome of interest is extremely rare, these plots can be relatively uninformative.<sup>45</sup> Finally, the findings of our analysis are based on the Dutch primary care system. Other systems may employ distinct methods of encoding, data acquisition, and diagnosis; therefore, our findings may not be generalizable to other care settings.

We showed that the aSAH prediction model performed worse than the AIS and ICH prediction models, which can be attributable to several factors. First, because aSAH is more rare than AIS and ICH,<sup>2</sup> there is less data available for model training and validation. Second, because aSAH tends to occur at a younger age,<sup>1,28,29</sup> age was a much stronger predictor in the AIS and ICH models than in the aSAH model. In addition, a younger age corresponds to fewer opportunities to measure established risk factors for aSAH (e.g. smoking status, hypertension) in an EHR-context. As a result, fewer predictors are available for inclusion in the aSAH prediction model, limiting its performance. We conclude that using EHR data to accurately identify individuals at increased risk for aSAH in the general population is currently not feasible. Instead, our model identified potential prognostic factors for aSAH that can be investigated in future studies.

## REFERENCES

1. van Gijn J, Kerr RS, Rinkel GJE. Subarachnoid haemorrhage. *Lancet*. 2007 Jan 27;369(9558):306–18.
2. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol*. 2021 Oct;20(10):795–820.
3. Johnston SC, Selvin S, Gress DR. The burden, trends, and demographics of mortality from subarachnoid hemorrhage. *Neurology*. 1998 May;50(5):1413–8.
4. Hoh BL, Ko NU, Amin-Hanjani S, Chou SHY, Cruz-Flores S, Dangayach NS, et al. 2023 Guideline for the Management of Patients With Aneurysmal Subarachnoid Hemorrhage: A Guideline From the American Heart Association/American Stroke Association. *Stroke*. 2023 Jul;54(7):e314–70.
5. Rinkel GJ, Ruigrok YM. Preventive screening for intracranial aneurysms. *Int J Stroke*. 2022 Jan;17(1):30–6.
6. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2017 Jan;24(1):198–208.
7. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012 Jun;13(6):395–405.
8. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*. 2014 Feb 7;2:3.
9. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017 May 23;357:j2099.
10. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*. 2007 Jul 21;335(7611):136.
11. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ*. 2012 Jun 21;344:e4181.
12. Teoh D. Towards stroke prediction using electronic health records. *BMC Med Inform Decis Mak*. 2018 Dec 4;18:127.
13. Robson J, Dostal I, Sheikh A, Eldridge S, Madurasinghe V, Griffiths C, et al. The NHS Health Check in England: an evaluation of the first 4 years. *BMJ Open*. 2016 Jan 1;6(1):e008840.
14. Available from: <https://stizon.nl/>
15. Ekker MS, Verhoeven JI, Vaartjes I, Nieuwenhuizen KM van, Klijn CJM, Leeuw FE de. Stroke incidence in young adults according to age, subtype, sex, and time trends. *Neurology*. 2019 May 21;92(21):e2444–54.
16. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373–83.
17. Ministerie van Volksgezondheid, Welzijn en Sport; Available from: <https://www.scp.nl/actueel/nieuws/2022/06/07/cijfers-over-welvaart-opleidingsniveau-en-arbeid-per-wijk-voortaan-te-vinden-op-cbs.nl>
18. Rubin DB. Inference and Missing Data. *Biometrika*. 1976;63(3):581–92.

19. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and Prognostic Research*. 2020 Jul 2;4(1):8.
20. Wu Y. Elastic net for Cox's proportional hazards model with a solution path algorithm. *Stat Sin*. 2012;22:27–294.
21. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2005;67(2):301–20.
22. Python documentation [Internet]. [cited 2023 Apr 5]. The Python Language Reference. Available from: <https://docs.python.org/3/reference/index.html>
23. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine*. 2015 Jan 6;13(1):1.
24. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *Ann Intern Med*. 2007 Oct 16;147(8):573–7.
25. Dev S, Wang H, Nwosu CS, Jain N, Veeravalli B, John D. A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics*. 2022 Nov 1;2:100032.
26. Hippisley-Cox J, Coupland C, Brindle P. Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. *BMJ*. 2013 May 2;346:f2573.
27. Yuan Z, Voss EA, DeFalco FJ, Pan G, Ryan PB, Yannicelli D, et al. Risk Prediction for Ischemic Stroke and Transient Ischemic Attack in Patients Without Atrial Fibrillation: A Retrospective Cohort Study. *Journal of Stroke and Cerebrovascular Diseases*. 2017 Aug 1;26(8):1721–31.
28. Yousufuddin M, Young N. Aging and ischemic stroke. *Aging (Albany NY)*. 2019 May 5;11(9):2542.
29. Camacho E, LoPresti MA, Bruce S, Lin D, Abraham M, Appelboom G, et al. The role of age in intracerebral hemorrhages. *J Clin Neurosci*. 2015 Dec;22(12):1867–70.
30. Kelly-Hayes M. Influence of Age and Health Behaviors on Stroke Risk: Lessons from Longitudinal Studies. *Journal of the American Geriatrics Society*. 2010;58(s2):S325–8.
31. Ruigrok YM, Buskens E, Rinkel GJ. Attributable risk of common and rare determinants of subarachnoid hemorrhage. *Stroke*. 2001 May;32(5):1173–5.
32. Feigin VL, Rinkel GJE, Lawes CMM, Algra A, Bennett DA, van Gijn J, et al. Risk factors for subarachnoid hemorrhage: an updated systematic review of epidemiological studies. *Stroke*. 2005 Dec;36(12):2773–80.
33. Isaksen J, Egge A, Waterloo K, Romner B, Ingebrigtsen T. Risk factors for aneurysmal subarachnoid haemorrhage: the Tromsø study. *Journal of Neurology, Neurosurgery & Psychiatry*. 2002 Aug 1;73(2):185–7.
34. Müller TB, Vik A, Romundstad PR, Sandvei MS. Risk Factors for Unruptured Intracranial Aneurysms and Subarachnoid Hemorrhage in a Prospective Population-Based Study. *Stroke*. 2019 Oct;50(10):2952–5.
35. Ohkuma H, Tabata H, Suzuki S, Islam MS. Risk factors for aneurysmal subarachnoid hemorrhage in Aomori, Japan. *Stroke*. 2003 Jan;34(1):96–100.

36. Greving JP, Wermer MJH, Brown RD, Morita A, Juvela S, Yonekura M, et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. *Lancet Neurol*. 2014 Jan;13(1):59–66.
37. Mariajoseph FP, Huang H, Lai LT. Influence of socioeconomic status on the incidence of aneurysmal subarachnoid haemorrhage and clinical recovery. *J Clin Neurosci*. 2022 Jan;95:70–4.
38. Kim JH, Jeon J, Kim J. Lower risk of subarachnoid haemorrhage in diabetes: a nationwide population-based cohort study. *Stroke Vasc Neurol* [Internet]. 2021 Sep 1 [cited 2023 May 23];6(3). Available from: <https://svn.bmj.com/content/6/3/402>
39. Adams HP Jr, Putman SF, Kassell NF, Torner JC. Prevalence of Diabetes Mellitus Among Patients With Subarachnoid Hemorrhage. *Archives of Neurology*. 1984 Oct 1;41(10):1033–5.
40. Rehman S, Sahle BW, Chandra RV, Dwyer M, Thrift AG, Callisaya M, et al. Sex differences in risk factors for aneurysmal subarachnoid haemorrhage: Systematic review and meta-analysis. *Journal of the Neurological Sciences*. 2019 Nov 15;406:116446.
41. Bakker MK, van der Spek RAA, van Rheenen W, Morel S, Bourcier R, Hostettler IC, et al. Genome-wide association study of intracranial aneurysms identifies 17 risk loci and genetic overlap with clinical risk factors. *Nat Genet*. 2020 Dec;52(12):1303–13.
42. Haneuse S, Arterburn D, Daniels MJ. Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task. *JAMA Network Open*. 2021 Feb 26;4(2):e210184.
43. van Os HJA, Kanning JP, Wermer MJH, Chavannes NH, Numans ME, Ruigrok YM, et al. Developing Clinical Prediction Models Using Primary Care Electronic Health Record Data: The Impact of Data Preparation Choices on Model Performance. *Frontiers in Epidemiology* [Internet]. 2022 [cited 2022 Oct 26];2. Available from: <https://www.frontiersin.org/articles/10.3389/fepid.2022.871630>
44. Vlak MHM, Rinkel GJE, Greebe P, Greving JP, Algra A. Lifetime risks for aneurysmal subarachnoid haemorrhage: multivariable risk stratification. *J Neurol Neurosurg Psychiatry*. 2013 Jun;84(6):619–23.
45. Wallace BC, Dahabreh IJ. Class Probability Estimates are Unreliable for Imbalanced Data (and How to Fix Them). In: 2012 IEEE 12th International Conference on Data Mining. 2012. p. 695–70

## SUPPLEMENTARY MATERIALS

**Table S1.** All predictors of the acute ischemic stroke (AIS) prediction model. The corresponding coefficients for these predictors in the aneurysmal subarachnoid hemorrhage (aSAH) and intracerebral hemorrhage (ICH) models are also shown.

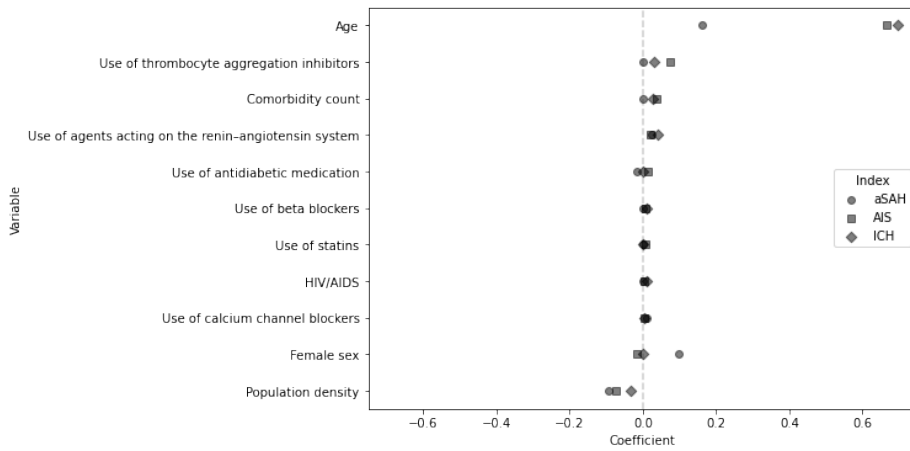
	aSAH	AIS	ICH
Age	0.163	0.665	0.695
Use of thrombocyte aggregation inhibitors	0.000	0.074	0.031
Comorbidity count	0.000	0.038	0.027
Use of agents acting on the renin–angiotensin system	0.024	0.022	0.039
Use of antidiabetic medication (excl. Insulin)	-0.016	0.013	0.000
Use of beta blockers	0.000	0.008	0.009
Use of statins	0.000	0.007	0.000
HIV/AIDS	0.000	0.005	0.011
Use of calcium channel blockers	0.010	0.005	0.003
Female sex	0.098	-0.018	0.000
Population density	-0.095	-0.073	-0.034

Each coefficient corresponds to the log hazard ratios after applying elastic net penalties. 0 indicates that the predictor is not predictive for that outcome.

**Table S2.** All predictors of the intracerebral hemorrhage (ICH) prediction model. The corresponding coefficients for these predictors in the aneurysmal subarachnoid hemorrhage (aSAH) and acute ischemic stroke (AIS) models are also shown.

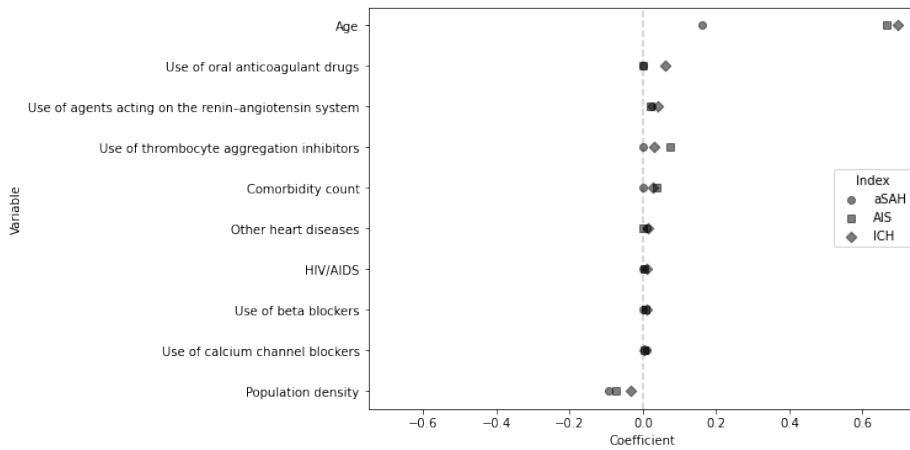
	aSAH	AIS	ICH
Age	0.163	0.665	0.695
Use of oral anticoagulant drugs	0.000	0.000	0.060
Use of agents acting on the renin–angiotensin system	0.024	0.022	0.039
Use of thrombocyte aggregation inhibitors	0.000	0.074	0.031
Comorbidity count	0.000	0.038	0.027
Other heart diseases	0.012	0.000	0.015
HIV/AIDS	0.000	0.005	0.011
Use of beta blockers	0.000	0.008	0.009
Use of calcium channel blockers	0.010	0.005	0.003
Population density	-0.095	-0.073	-0.034

Each coefficient corresponds to the log hazard ratios after applying elastic net penalties. 0 indicates that the predictor is not predictive for that outcome.



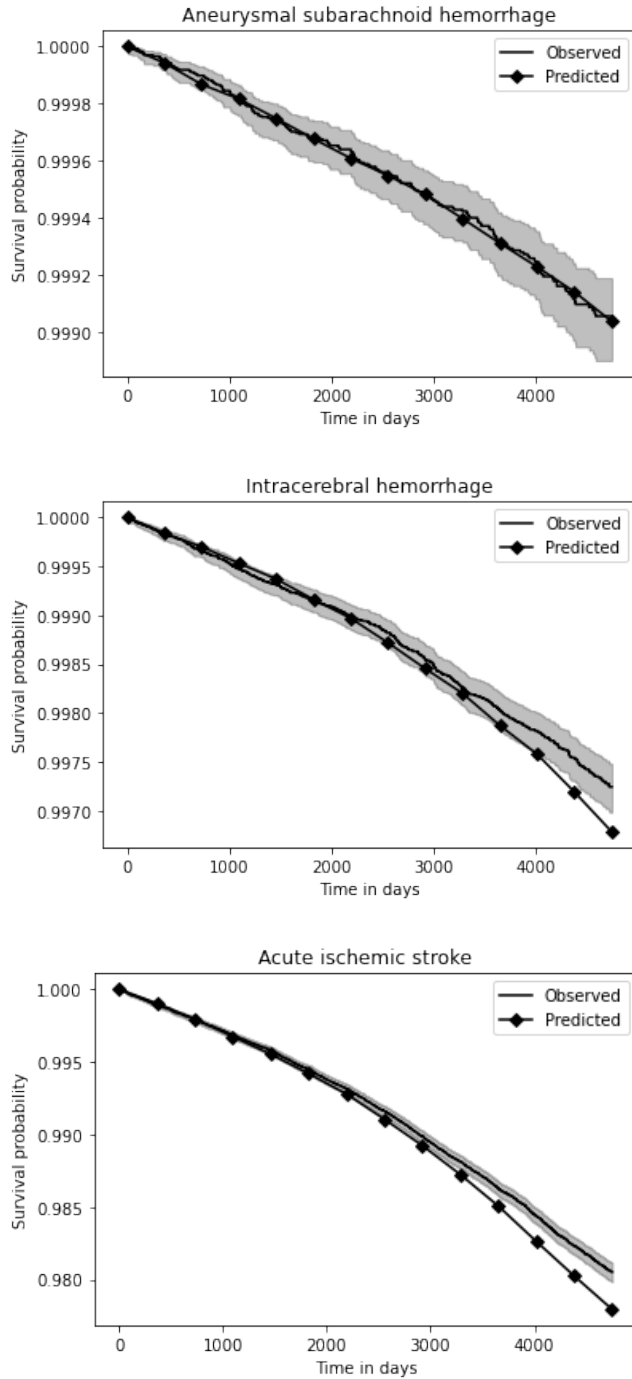
**Figure S1.** All predictors of the acute ischemic stroke (AIS) prediction model in relation to the corresponding coefficients of these predictors in the aneurysmal subarachnoid hemorrhage (aSAH) and intracerebral hemorrhage (ICH) prediction models.

Each coefficient corresponds to the log hazard ratios after applying elastic net penalties. 0 indicates that the predictor is not predictive for that outcome.



**Figure S2.** All predictors of the intracerebral hemorrhage (ICH) prediction model in relation to the corresponding coefficients of these predictors in the aneurysmal subarachnoid hemorrhage (aSAH) and acute ischemic stroke (AIS) prediction models.

Each coefficient corresponds to the log hazard ratios after applying elastic net penalties. 0 indicates that the predictor is not predictive for that outcome.



**Figure S3.** 10-year Kaplan-Meier curves for observed and predicted survival probability for each outcome.



## Chapter 5

# Development and external validation of the SMA2SH2ERS risk prediction model for aneurysmal subarachnoid hemorrhage in the general population

---

Vita M. Klieverik, Jos P. Kanning, Ina L. Rissanen, Kristiina Rannikmäe, Amy E. Martinsen, Bendik S. Winsvold, Mirjam I. Geerlings, Ynte M. Ruigrok

Submitted

## ABSTRACT

**Background:** Aneurysmal subarachnoid hemorrhage (ASAH) is a severe stroke type, preventable by screening for intracranial aneurysms followed by treatment in high-risk individuals. We aimed to develop and validate a risk prediction model for ASAH in the general population to identify high-risk individuals.

**Methods:** UK Biobank data were used for model development and Trøndelag Health (HUNT) Study data for model validation. Participants with prior ASAH before baseline or missing data were excluded. Employing multivariable Cox regression, predictors for ASAH were identified, while correcting for overfitting using bootstrapping techniques. Predictive performance was assessed using discrimination and calibration.

**Results:** In the development cohort, ASAH occurred in 738 (0.2%) of 456,856 individuals during 5,407,909 person-years of follow-up. Predictors for ASAH were Sex (S), diabetes mellitus (M), Age and Alcohol consumption (A<sup>2</sup>), Smoking (S), Hypertension and Hypercholesterolemia (H<sup>2</sup>), Educational attainment (E), Regular physical activity (R) and family history of Stroke (S; SMA<sup>2</sup>SH<sup>2</sup>ERS), and multiple interactions between these predictors. The c-statistic of the model in the development cohort was 0.62 (95%CI 0.60–0.64). Predicted absolute 10-year ASAH risk varied from 0.042% to 0.52%. In the validation cohort, 220 of 46,483 individuals developed ASAH during 2,077,927 person-years of follow-up and the c-statistic of this model was 0.64 (95%CI 0.58–0.69). Both models showed fair to good calibration.

**Conclusions:** Our SMA<sup>2</sup>SH<sup>2</sup>ERS model provides ASAH risk estimates between 0.042% and 0.52% for the general population. While overall ASAH risk is low, the model identifies individuals with up to 12 times increased risk compared to those at lowest risk.

## INTRODUCTION

Unruptured intracranial aneurysms (UIAs) affect 3% of the general population.<sup>1</sup> When an UIA ruptures, it causes an aneurysmal subarachnoid hemorrhage (ASAH), striking at a mean age of 55 years), and more often in women than men.<sup>2</sup> The incidence of ASAH is 6.1 per 100,000 person-years corresponding to a lifetime risk of 0.2%.<sup>3</sup> Approximately one-third of ASAH patients die, and half of survivors require continuous care,<sup>4</sup> often with severe cognitive impairments affecting functionality and quality of life.<sup>2</sup> ASAH presents a significant socio-economic burden, comparable to ischemic stroke in potential life years lost.<sup>5</sup>

A recent study showed that approximately 24% of ASAH patients die before receiving medical attention, and the early effects of ASAH are the leading cause of death among those admitted to hospital.<sup>2,6</sup> As a result, the opportunities to improve prognosis after ASAH are limited and therefore prevention of ASAH is essential to reduce disease burden. Non-invasive screening for UIAs followed by endovascular or neurosurgical treatment can prevent ASAH.<sup>7</sup> Screening is already proven cost-effective for first-degree relatives of ASAH patients.<sup>8,9</sup> Individuals with one affected first-degree relative with ASAH have an estimated lifetime risk of ASAH of up to 0.4%<sup>10</sup> and in 4% of these individuals UIAs can be identified at first screening.<sup>11</sup> In individuals with two or more affected first-degree relatives the estimated lifetime risk increases to up to 10%<sup>10</sup> with UIAs identified in 11% at first screening.<sup>12</sup> Whether additional high-risk individuals in the general population may benefit from preventive screening is unclear.

Risk factors for ASAH include female sex, older age, hypertension, smoking, and alcohol consumption,<sup>13-16</sup> but a comprehensive risk prediction model applicable to the general population is lacking. Estimating individual absolute ASAH risk in primary care settings could aid in identifying high-risk candidates for UIA screening. Thus, we aimed to develop and externally validate a risk prediction model for ASAH in the general population to estimate individual absolute ASAH risk.

## METHODS

This study was performed in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement.

### **Development cohort**

For model development, we utilized data from the United Kingdom (UK) Biobank Prospective Cohort Study, a large population-based study with over 500,000 participants aged 37 to 73, recruited between 2006 and 2010.<sup>17</sup> Data from the UK Biobank were linked to International Classification of Diseases (ICD) codes ICD-9 and ICD-10 to record diagnosis information, alongside national death and cancer registries. Participants with prior ASAH (please see next paragraph 'Outcome and candidate predictors' for the definition of ASAH) at baseline assessment were excluded. Follow-up data were available until March 28, 2021. All participants provided written informed consent, and the study was approved by the North West Multicenter Research Ethics Committee (MREC) and the National Health Service (NHS) National Research Ethics Service (ref 11/NEW/0382). Official approval for the present study was considered unnecessary.

### **Validation cohort**

For model validation, we utilized data from the Trøndelag Health (HUNT) Study, a general population-based cohort comprising over 240,000 Norwegian participants aged 20 or older, recruited between 1984 and 2008.<sup>18</sup> HUNT Study data were linked to Hospital Episode Statistics (HES) and national health registries. We focused on data from the HUNT2 (recruitment 1995–1997) and HUNT3 (recruitment 2006–2008) studies, previously employed in a study on ASAH and UIA genetic risk.<sup>19</sup> Participants with prior ASAH were excluded. Follow-up data were available until June 6, 2018. All participants provided written consent, and the study was approved by the Regional Committee for Medical Research Ethics (REC) (ref #2015/578).

### **Outcome and candidate predictors**

The primary outcome, incident ASAH, was identified using ICD-9 430 and ICD-10 I600 to I609 codes in both the UK Biobank and the HUNT Study. We decided to include the ICD-10 I608 and I609 codes, which are expected to encompass solely non-aneurysmal cases, comprising approximately 10-15% of all subarachnoid hemorrhage cases.<sup>2</sup> However, it's probable that these codes also encompass ASAH cases in the UK Biobank, given that the number of cases represented by these codes constitutes 56.6% of the total cases (418 out of 738 ) in the UK Biobank. Candidate predictors were chosen based on literature review and limited to factors easily accessible to general practitioners during routine consultations, including sex, age, family history of stroke, hypertension, smoking status, hypercholesterolemia, regular physical activity, hormone replacement therapy (HRT), diabetes mellitus (DM), alcohol consumption, and educational attainment.<sup>13-16</sup> While most predictors increase risk, some, like hypercholesterolemia and DM, may decrease it.<sup>13-16</sup> Family

history of ASAH was omitted due to unavailability in the UK Biobank. Interaction terms between age and HRT, smoking status and alcohol consumption, and regular physical activity with hypertension, hypercholesterolemia, and DM were included, alongside sex interaction terms with other predictors, to explore potential effect modification by sex.<sup>20</sup> Definitions for these predictors are detailed in the Supplementary Material's methods section.

### Statistical analysis

Statistical analysis involved expressing normally distributed continuous variables as means  $\pm$  standard deviations (SD) and skewed distributed continuous variables as medians with corresponding interquartile ranges (IQR). Categorical variables were presented as counts with percentages. In the development cohort, missing data were minimal (ranging from 0.001% to 5.5%) and participants with missing data were excluded.<sup>21</sup> No missing data were observed in the validation cohort.

For model development, multivariable Cox proportional hazards regression analysis was conducted using follow-up time as the time scale. Follow-up data were censored at the time of incident ASAH, date of death, or last follow-up assessment on March 28, 2021, whichever came first. The functional form of continuous candidate predictor age was assessed using martingale residuals.<sup>22</sup> Candidate predictors were considered for model entry regardless of their univariable association with incident ASAH. Backward selection based on the Akaike Information Criterion (AIC) was performed to evaluate predictor contributions. Proportional hazard assumption was assessed visually and numerically using scaled Schoenfeld residuals plots and tests. To address overfitting, a shrinkage factor was applied to regression coefficients determined by bootstrapping procedures.<sup>23</sup> Hazard ratios (HR) with corresponding 95% confidence intervals (CI) represented the estimated effect sizes of independent predictors.

Model performance was evaluated using discrimination and calibration. Discrimination, measured by the concordance statistic (c-statistic), was corrected for overoptimism through bootstrapping.<sup>23</sup> Calibration, an indicator for the agreement between predicted and observed probability of incident ASAH, visually using 5-year and 10-year calibration plots.<sup>24</sup> Statistical analyses were conducted using R statistical software, version 4.0.2, with an online interactive risk calculator developed using the Shiny R package (1.7.1). The risk calculator predicts an individual's absolute ASAH risk at 5- and 10-year follow-up based on provided predictors, with each predictor's contribution calculated by dividing regression coefficients by the smallest coefficient and rounding to the nearest integer.

## RESULTS

### Baseline characteristics

Baseline characteristics are detailed in Table 1. From the UK Biobank, 493,650 participants were recruited, with 456,856 included after excluding those with missing data or prior ASAH ( $n = 36,794$ ). Among UK Biobank participants, 54.0% were women, mean age was  $56.4 \pm 8.1$  years, and 738 (0.2%) developed ASAH during 5,407,909 person-years of follow-up (13.6 per 100,000 person-years, median follow-up 12.1 years, range 2 days to 14.3 years). In the HUNT Study, 46,483 participants were included, with 53.1% women, mean age  $59.1 \pm 14.1$  years, and 220 (0.5%) developing ASAH during 2,077,927 person-years of follow-up (10.6 per 100,000 person-years, median follow-up 1.0 years, range 0 days to 24.0 years).

**Table 1.** Baseline characteristics of the development and validation cohorts

Characteristic	Development cohort (n = 456,856)		Validation cohort (n = 46,483)	
	n	%	n	%
Women	246,771	54.0	24,661	53.1
Age (years)				
< 50	109,062	23.9	12,996	28.0
≥ 50	347,794	76.1	33,534	72.0
Mean ± SD	56.4 ± 8.1		59.1 ± 14.1	
Family history of stroke	119,690	26.2	10,224	22.0
Hypertension	227,823	49.9	17,332	37.3
Smoking status				
Never smoking	250,870	54.9	18,808	40.5
Former smoking	159,174	34.8	16,207	34.9
Current smoking	46,812	10.2	11,468	24.7
Hypercholesterolemia	77,510	17.0	13,152	28.3
Regular physical activity	148,892	32.6	2,075	4.5
HRT	93,223	20.4	3,393	7.3
Diabetes mellitus	22,903	5.0	1,357	2.9
Alcohol consumption				
Never	34,201	7.5	670	1.4
On special occasions	327,445	71.7	37,813	81.3
Daily or almost daily	95,210	20.8	8,000	17.2
Educational attainment				
Low	71,427	15.6	16,790	36.1
Intermediate	230,438	50.4	20,595	44.3
High	154,991	33.9	9,098	19.6

SD = standard deviation, HRT = hormone replacement therapy, NA = not available.

**Table 2.** Univariable and multivariable Cox proportional hazards regression analysis of predictors of incident aneurysmal subarachnoid hemorrhage (ASAH)

Predictor	Univariable	Multivariable*
	HR (95% CI)	HR (95% CI)
Female sex	1.50 (1.29 – 1.75)	0.40 (0.14 – 1.12)
Age per year	1.03 (1.02 – 1.04)	1.01 (1.00 – 1.03)
Family history of stroke	1.26 (1.08 – 1.47)	1.14 (0.99 – 1.31)
Hypertension	1.33 (1.15 – 1.54)	1.45 (1.15 – 1.84)
Smoking status		
Never smoking	Reference	
Former smoking	1.07 (0.91 – 1.26)	1.13 (0.87 – 1.47)
Current smoking	2.14 (1.76 – 2.60)	1.70 (1.22 – 2.37)
Hypercholesterolemia	1.05 (0.86 – 1.27)	0.98 (0.79 – 1.21)
Regular physical activity	0.90 (0.77 – 1.06)	1.02 (0.88 – 1.18)
Diabetes mellitus	0.88 (0.61 – 1.26)	0.70 (0.47 – 1.06)
Alcohol consumption		
Never	1.56 (1.23 – 1.98)	1.41 (1.07 – 1.85)
On special occasions	Reference	
Daily or almost daily	1.14 (0.95 – 1.36)	1.13 (0.88 – 1.45)
Educational attainment		
Low	1.29 (1.07 – 1.56)	1.05 (0.88 – 1.24)
Intermediate	Reference	
High	0.75 (0.63 – 0.89)	0.84 (0.72 – 0.98)
Interactions		
Former smoking * Never alcohol consumption	1.32 (0.78 – 2.24)	1.32 (0.84 – 2.10)
Former smoking * Daily alcohol consumption	0.93 (0.62 – 1.40)	0.92 (0.64 – 1.31)
Current smoking * Never alcohol consumption	0.44 (0.19 – 1.00)	0.47 (0.23 – 0.98)
Current smoking * Daily alcohol consumption	0.97 (0.61 – 1.54)	1.02 (0.68 – 1.54)
Regular physical activity * Hypercholesterolemia	0.75 (0.48 – 1.19)	0.68 (0.44 – 1.04)
Regular physical activity * DM	1.68 (0.79 – 3.58)	1.93 (0.96 – 3.89)
Female sex * Age per year	1.02 (1.00 – 1.04)	1.03 (1.01 – 1.04)
Female sex * Hypertension	0.82 (0.60 – 1.13)	0.75 (0.56 – 1.00)
Female sex * Former smoking	0.82 (0.58 – 1.15)	0.86 (0.63 – 1.16)
Female sex * Current smoking	1.45 (0.96 – 2.20)	1.54 (1.06 – 2.22)

\*The initial regression coefficients were corrected for overfitting with a shrinkage factor of 0.88.

HR = hazard ratio, CI = confidence interval

## Model development and performance

Predictors were Sex (S), DM (M), Age and Alcohol consumption (A2), Smoking (S), Hypertension and Hypercholesterolemia (H2), Educational attainment (E), Regular physical activity (R), and family history of Stroke (S; SMA<sup>2</sup>SH<sup>2</sup>ERS), with multiple predictor interactions, including smoking-alcohol and sex with age, hypertension, and smoking (Table 2). HRT and the interactions between age and HRT, regular physical activity and hypertension and sex with other predictors than age, hypertension and smoking were excluded due to limited predictive value (Table 2). Age was linearly analyzed, and proportional hazard assumptions were met (Supplementary Figure 1). We inspected the scaled Schoenfeld residuals plots and tests for each independent predictor and detected no deviations from the proportional hazard assumption (Supplementary Figure 2 and Supplementary Table 1). Following shrinkage of the regression coefficients, the c-statistic of the model in the development cohort was 0.62 (95% CI 0.60–0.64). The c-statistic of the model in the validation cohort was 0.64 (95% CI 0.58–0.69). The 5-year and 10-year calibration plots for the development and validation cohorts showed good correspondence between predicted and observed risk (Supplementary Figure 3).

## Individual risk prediction

To determine an individual's absolute risk of ASAH, one can utilize the original regression equation from Supplementary Table 2. However, due to the complex nature of multiple predictor interactions, manual calculation of absolute risk is challenging. Therefore, we developed an online interactive SMA<sup>2</sup>SH<sup>2</sup>ERS risk calculator, accessible at <https://asah-prediction.shinyapps.io/app-1/> enabling calculation of individual ASAH risks at 5- and 10-year follow-ups based on provided predictors.

The mean predicted 5-year absolute ASAH risk was 0.05%, ranging from 0.018% to 0.22%. Predicted 10-year cumulative absolute risk averaged 0.13%, ranging from 0.042% to 0.52%. In the UK Biobank data, only 0.006% of participants (28 out of 456,856) had a predicted 10-year cumulative absolute risk exceeding 0.40% (i.e. the risk of ASAH in individuals with one first-degree relative with ASAH<sup>11</sup>).

Among UK Biobank participants, the lowest absolute risk was for a 38-year-old non-smoking man with diabetes mellitus and hypercholesterolemia, who consumes alcohol occasionally, exercises regularly, and has high educational attainment. Conversely, the highest absolute risk was for a 73-year-old woman with hypertension and a family history of stroke, who is a former smoker, abstains from alcohol, exercises regularly, and has low educational attainment.

## DISCUSSION

We developed the SMA<sup>2</sup>SH<sup>2</sup>ERS risk prediction model to estimate individuals' absolute ASAH risk using readily available predictors in primary healthcare. Independent predictors included Sex (S), DM (M), Age and Alcohol consumption (A2), Smoking (S), Hypertension and Hypercholesterolemia (H2), Educational attainment (E), Regular physical activity (R), and family history of Stroke (S; SMA<sup>2</sup>SH<sup>2</sup>ERS), with multiple predictor interactions, including three with sex (age, hypertension, and smoking). Predicted 5-year absolute ASAH risk ranged from 0.018% to 0.22%, and 10-year cumulative absolute risk ranged from 0.042% to 0.52%. While overall ASAH risk is low, the SMA<sup>2</sup>SH<sup>2</sup>ERS model identifies individuals with up to 12 times increased risk compared to those at lowest risk.

To our knowledge this is the first study to develop and validate a risk prediction model to predict ASAH in the general population. Two previous prediction studies did examine the risk of ASAH but in patients with already proven UIAs. In these studies, like in our study, age, sex and hypertension were also identified as predictors.<sup>14,24</sup>

The finding that female sex is a predictor for ASAH aligns with prior studies indicating a higher risk of developing ASAH in women, both in a general population-based settings and among patients diagnosed with an UIA.<sup>3,25</sup> UIAs are also more prevalent in women, particularly after age 50, coinciding with increased ASAH incidence in women beyond this age.<sup>1,2</sup> This observation underscores the interaction between age and sex in our study, potentially linked to hormonal fluctuations during and post-menopause, theorized to elevate UIA risk.<sup>26</sup> However, the precise role of female hormones in ASAH pathogenesis remains uncertain.<sup>26,27</sup> Our examination of HRT revealed no independent association with ASAH risk. Literature on HRT's role in ASAH yields conflicting findings, with some studies suggesting risk reduction while others indicate increased or neutral effects.<sup>27,28</sup> These inconsistent results emphasize the need for further investigation into HRT's impact, alongside other female hormonal factors, on ASAH risk. Alternatively, differential effects of risk factors based on sex may contribute to this disparity. Previous research has shown that women with hypertension or who smoke are at greater ASAH risk than men with similar risk factors, a phenomenon corroborated by our study.<sup>19,26</sup>

We observed an increased risk of ASAH associated with familial stroke, a novel finding not previously demonstrated. We used this predictor as a proxy for familial ASAH, given its absence in our data. This aligns logically with prior research

indicating familial ASAH as a risk factor, given ASAH's classification as a stroke subtype.<sup>7,10</sup> The increased ASAH risk of in familial stroke is likely explained by genetic risk factors in combination with clinical risk factors including hypertension and smoking.<sup>28,29</sup> We also found an elevated risk of ASAH associated with alcohol abstinence. This result may be connected to individuals experiencing conditions that not only prevent them from consuming alcohol but also elevate the risk of ASAH.

While the calibration plot demonstrated good correspondence between predicted and observed ASAH risk in the development cohort, calibration slightly declined in the validation cohort. This discrepancy may arise from differences in ASAH epidemiology and candidate predictors in the validation cohort. It's possible that the predictor set in the model is inadequate for accurately predicting ASAH risk in the validation cohort.

A significant strength of our study lies in the large prospective cohort with follow-up data derived from ICD codes, enabling robust model development. The extensive sample size facilitated the examination of numerous candidate predictors and potential interactions. Furthermore, we conducted external validation in an independent population cohort, enhancing the model's generalizability and practical utility. Notably, the predictors in our model are readily accessible to general practitioners during routine consultations, facilitating easy integration into daily practice. Consequently, we opted against incorporating genetic risk factors into our model, as previous research has demonstrated their limited added value over clinical data for ASAH prediction.<sup>30</sup>

Limitations include missing data in the development cohort, necessitating exclusion of affected participants. However, the small proportion of missing data was anticipated to have minimal impact. Another limitation stems from the use of the UK Biobank as our development cohort, which may skew toward more women, older individuals, and higher socioeconomic status compared to non-participants.<sup>31,32</sup> Nonetheless, we mitigated this by incorporating educational attainment as a proxy for socioeconomic status. Moreover, the UK Biobank's limited ethnic diversity precluded subgroup analyses to assess model validity across ethnicities, although validation in the HUNT study, which includes more participants with low education, partially addressed this limitation.<sup>18</sup> Additionally, the accuracy of incident ASAH identification in population cohorts like the UK Biobank remains uncertain, with limited data available. While one study assessing a limited number of 24 ASAH cases reported a positive predictive value for ASAH

ICD codes in the UK Biobank being 71% (95% CI, 49%–87%),<sup>33</sup> further research with larger ASAH patient cohorts is warranted to confirm these findings. Furthermore, uncertainty persists regarding whether the ICD codes used encompass solely ASAH or also include non-aneurysmal cases.<sup>34</sup> We deliberately opted to include the ICD-10 I608 and I609 codes, indicative of non-aneurysmal cases, expecting them to account for 10–15% of all subarachnoid hemorrhage cases.<sup>2</sup> However, in our study, their prevalence constituted a significantly larger proportion (418/738, 56.6%), suggesting they likely encompass ASAH cases as well. This supports our decision to include these codes. Sensitivity analysis excluding these ICD-10 I608 and I609 codes yielded a slightly higher c-statistic (0.72 [95% CI, 0.69–0.75] versus 0.62 [95% CI, 0.60–0.64]), albeit with reduced statistical power due to fewer cases. Lastly, the relatively low ASAH incidence<sup>3</sup> contrasts with the prevalent predictors for this disease,<sup>13–16</sup> which may limit the attainment of a very high c-statistic.

Our SMA<sup>2</sup>SH<sup>2</sup>ERS risk prediction model offers insights into ASAH predictors in the general population, identifying individuals with up to a 12-fold increased risk compared to those at lowest risk. Prior research indicates cost-effectiveness of screening for UIAs for individuals with two or more first-degree relatives with ASAH who have an estimated ASAH lifetime risk of up to 10%.<sup>10</sup> Screening for UIAs is also likely to be cost-effective for individuals with only one first-degree relative with ASAH who have an estimated ASAH lifetime risk of 0.4%.<sup>7,10</sup> Given our model's ability to predict 10-year cumulative absolute risks up to 0.52%, it may therefore aid in identifying high-risk individuals for preventive UIA screening. However, future studies must evaluate the cost-effectiveness of such screening in individuals identified as high risk by the SMA<sup>2</sup>SH<sup>2</sup>ERS model. Additionally, considering interactions between sex and other predictors, future investigations should explore the potential for separate risk prediction models for men and women. Lastly, further research on the differential effects of risk factors and female-specific ASAH risk factors is warranted.

## REFERENCES

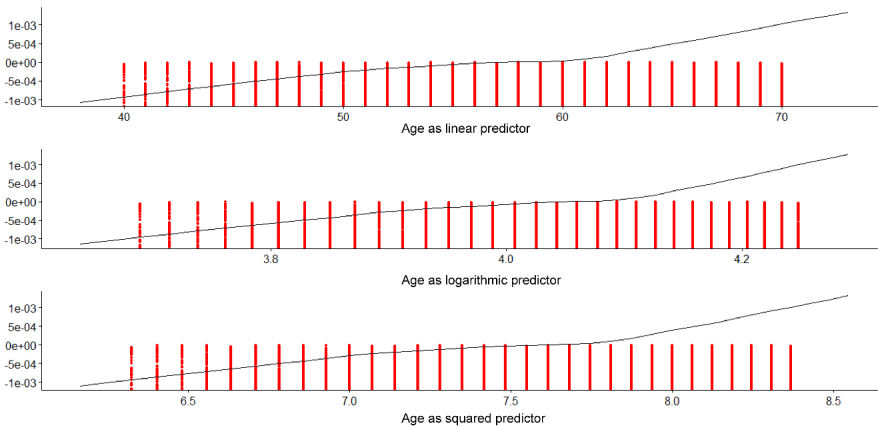
1. Vlak MH, Algra A, Brandenburg R, Rinkel GJ. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. *Lancet Neurol* 2011;10(7):626-36.
2. Macdonald RL, Schweizer TA. Spontaneous subarachnoid haemorrhage. *Lancet*. 2017; 389(10069):655-66.
3. Etminan N, Chang HS, Hackenberg K, et al. Worldwide incidence of aneurysmal subarachnoid hemorrhage according to region, time period, blood pressure, and smoking prevalence in the population: a systematic review and meta-analysis. *JAMA Neurol* 2019;76(5):588-97.
4. Nieuwkamp DJ, Setz LE, Algra A, Linn FH, de Rooij NK, Rinkel GJ. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *Lancet Neurol* 2009;8(7):635-42.
5. Johnston SC, Selvin S, Gress DR. The burden, trends, and demographics of mortality from subarachnoid hemorrhage. *Neurology* 1998;50(5):1413-8.
6. Asikainen A, Korja M, Kaprio J, Rautalin I. Case fatality in patients with aneurysmal subarachnoid hemorrhage in Finland: a nationwide register-based study. *Neurology* 2023;100(3):e348-e356.
7. Rinkel GJ, Ruigrok YM. Preventive screening for intracranial aneurysms. *Int J Stroke* 2022;17(1):30-6.
8. Takao H, Nojo T, Ohtomo K. Screening for familial intracranial aneurysms: decision and cost-effectiveness analysis. *Acad Radiol* 2008;15(4):462-471.
9. Bor AS, Koffijberg H, Wermer MJH, Rinkel GJE. Optimal screening strategy for familial intracranial aneurysms: a cost-effectiveness analysis. *Neurology* 2010;74(21):1671-1679.
10. Bor AS, Rinkel GJ, Adami J, et al. Risk of subarachnoid haemorrhage according to number of affected relatives: a population based case-control study. *Brain* 2008;131(Pt 10):2662-5.
11. Raaymakers TWM, Rinkel GJE, van Gijn J, et al. Risks and benefits of screening of intracranial aneurysms in first-degree relatives of patients with sporadic subarachnoid hemorrhage. *N Engl J Med* 1999;341(18):1344-50.
12. Bor AS, Rinkel GJ, van Norden J, Wermer MJ. Long-term, serial screening for intracranial aneurysms in individuals with a family history of aneurysmal subarachnoid haemorrhage: a cohort study. *Lancet Neurol* 2014;13(4):385-92.
13. Feigin VL, Rinkel GJ, Lawes CM, et al. Risk factors for subarachnoid hemorrhage: an updated systematic review of epidemiological studies. *Stroke* 2005;36(12):2773-80.
14. Greving JP, Wermer MJ, Brown RD, Jr., et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. *Lancet Neurol* 2014;13(1):59-66.
15. Sundstrom J, Soderholm M, Soderberg S, et al. Risk factors for subarachnoid haemorrhage: a nationwide cohort of 950 000 adults. *Int J Epidemiol* 2019;48(6):2018-25.
16. Vlak MH, Rinkel GJ, Greebe P, Algra A. Independent risk factors for intracranial aneurysms and their joint effect: a case-control study. *Stroke* 2013;44(4):984-7.
17. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12(3):e1001779.
18. Krokstad S, Langhammer A, Hveem K, et al. Cohort Profile: the HUNT Study, Norway. *Int J Epidemiol* 2013;42(4):968-77.

19. Bakker MK, Kanning JP, Abraham G, et al. Genetic risk score for intracranial aneurysms: Prediction of subarachnoid hemorrhage and role in clinical heterogeneity. *Stroke* 2023;54(3):810-818.
20. Lindeklev H, Sandvei MS, Njolstad I, et al. Sex differences in risk factors for aneurysmal subarachnoid hemorrhage: a cohort study. *Neurology* 2011;76(7):637-43.
21. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Med Res Methodol* 2017;17(1):162.
22. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;338:b604.
23. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.
24. Tominari S., Morit A., Ishibashi T., et al. 2015 Prediction model for 3-year rupture risk of unruptured cerebral aneurysms in Japanese patients. *Ann Neurol* 2015;77(6):1050-9.
25. Zuurbier CCM, Molenberg R, Mensing LA, et al. Sex Difference and rupture rate of intracranial aneurysms: An individual patient data meta-analysis. *Stroke* 2022;53(2):362-9.
26. Fuentes AM, Stone McGuire L, Amin-Hanjani S. Sex Differences in cerebral aneurysms and subarachnoid hemorrhage. *Stroke* 2022;53(2):624-633.
27. Algra AM, Klijn CJ, Helmerhorst FM, Algra A, Rinkel GJ. Female risk factors for subarachnoid hemorrhage: a systematic review. *Neurology* 2012;79(12):1230-6.
28. Bakker MK, Ruigrok YM. Genetics of intracranial aneurysms. *Stroke* 2021;52(9):3004-12.
29. Zuurbier CCM, Mensing LA, Wermer MJH, et al. Difference in rupture risk between familial and sporadic intracranial aneurysms: an individual patient data meta-analysis. *Neurology* 2021;97(22):e2195-e203.
30. Bakker MK, Kanning JP, Abraham G, et al. Genetic risk score for intracranial aneurysms: prediction of subarachnoid hemorrhage and role in clinical heterogeneity. *Stroke* 2023;54(3):810-818.
31. Janssen KJ, Donders AR, Harrell FE, Jr., et al. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol* 2010;63(7):721-7.
32. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am J Epidemiol* 2017;186(9):1026-34.
33. Rannikmäe K, Ngho K, Bush K, et al. Accuracy of identifying incident stroke cases from linked health care data in UK Biobank. *Neurology* 2020;95(6):e697-e707.
34. Roark C, Wilson MP, Kubes S, et al. Assessing the utility and accuracy of ICD10-CM non-traumatic subarachnoid hemorrhage codes for intracranial aneurysm research. *Learn Health Syst* 2021;5(4):e10257.

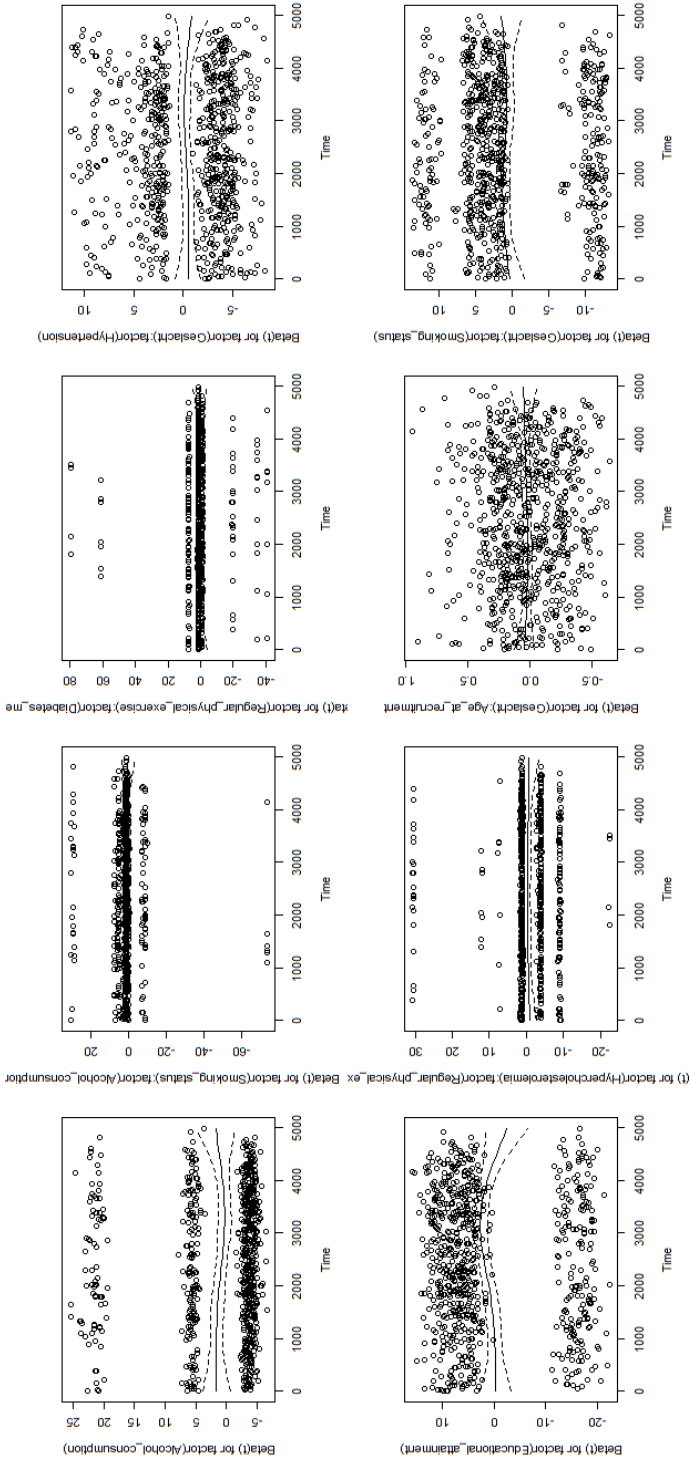
## SUPPLEMENTAL MATERIALS

### Supplementary methods

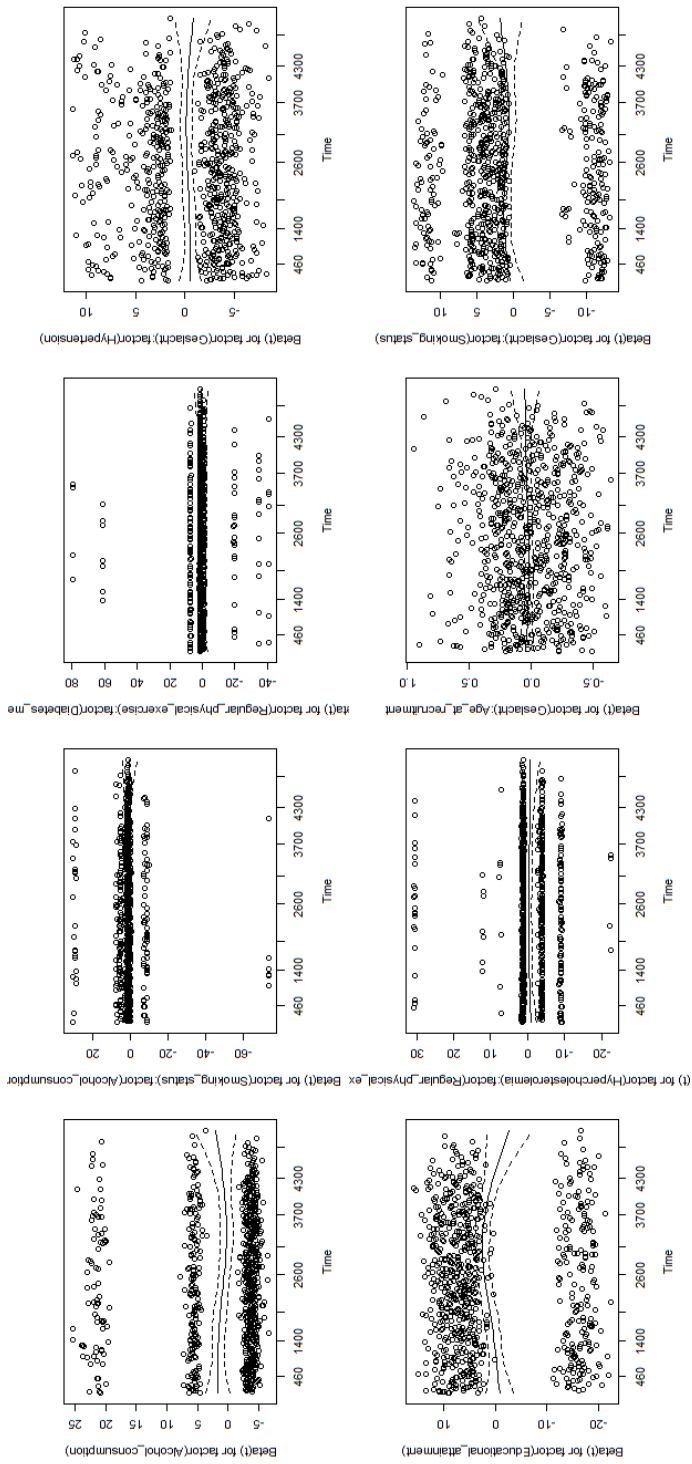
A family history of stroke was defined as at least one first-degree relative being affected with the disease. Hypertension was defined as systolic blood pressure  $\geq 140$  mmHg or diastolic blood pressure  $\geq 90$  mmHg and/or use of antihypertensive medication. We categorized smoking status into (1) never smokers, (2) former smokers, and (3) current smokers. Hypercholesterolemia was defined as use of cholesterol lowering medication. Regular physical activity was defined as vigorous physical activity  $\geq$  three times per week. We grouped HRT into (1) never users and (2) former and current users combined. DM was defined based on a past medical history of DM and/or use of antidiabetic medication. We categorized alcohol consumption into (1) no alcohol consumption, (2) alcohol consumption on special occasions, and (3) daily or almost daily alcohol consumption. We grouped educational attainment into (1) high, (2) intermediate, and (3) low educational attainment. High educational attainment was defined as having a university or college degree, intermediate educational attainment as having either an Advanced Level qualification, Ordinary Level qualification, Certificate of Secondary Education (CSE), National Vocational Qualification (NVQ), Higher National Diploma (HND), Higher National Qualification (HNC), or other professional qualification, and low educational attainment as having no degree.



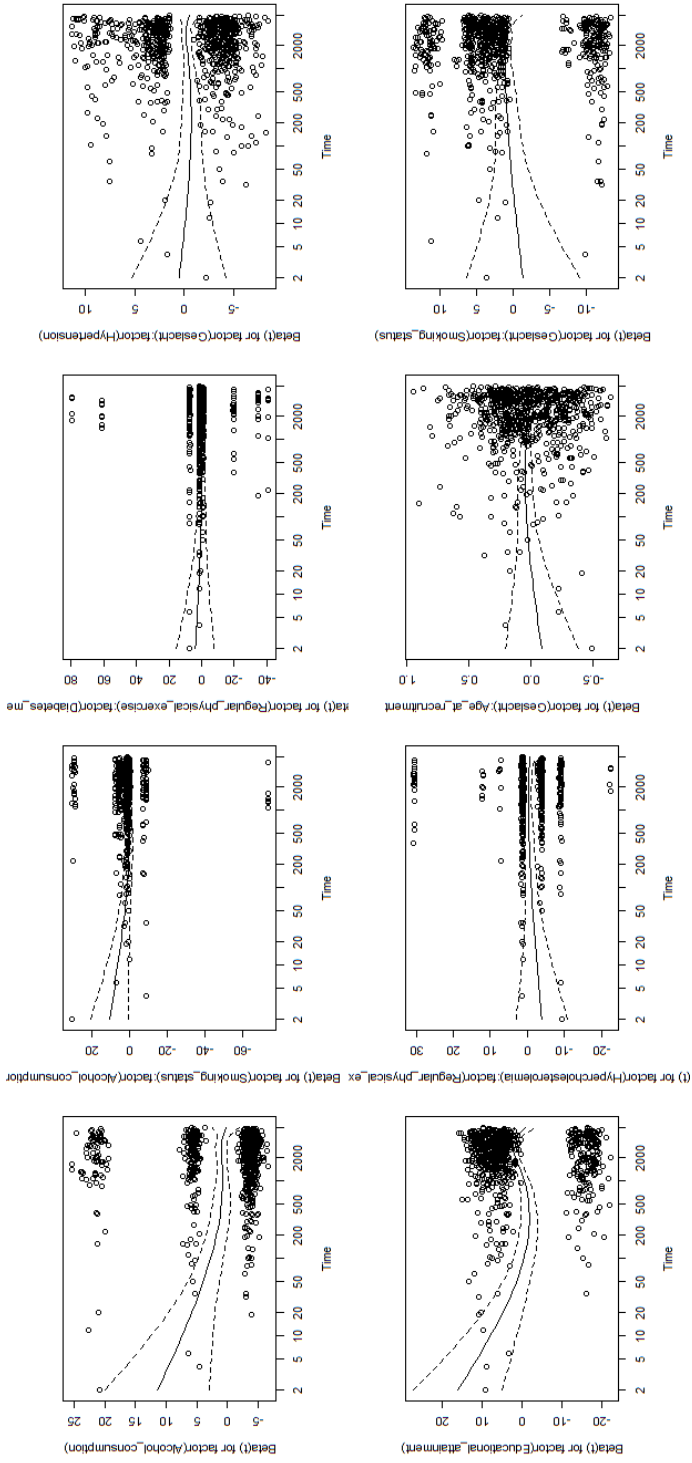
**Supplementary Figure 1.** Martingale residuals for the continuous predictor age.



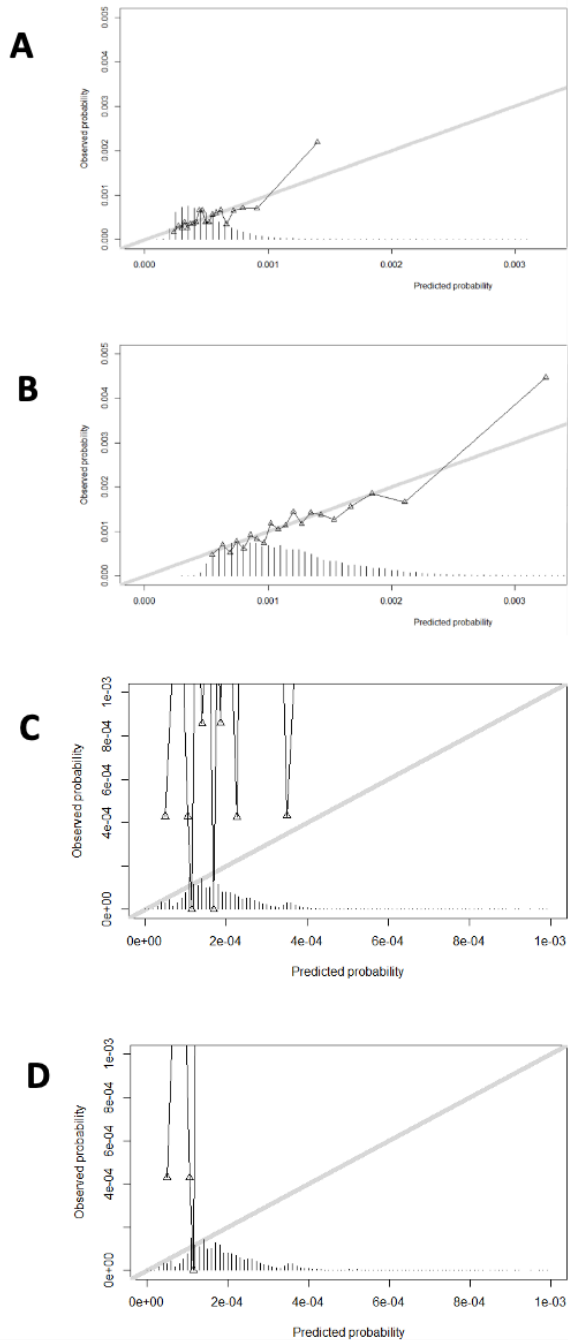
Supplementary Figure 2A. Scaled Schoenfeld residuals plot with survival time on linear time scale.



Supplementary Figure 2B. Scaled Schoenfeld residuals plot with survival time on Kaplan-Meier-transformed time scale.



**Supplementary Figure 2C.** Scaled Schoenfeld residuals plot with survival time on logarithmic-transformed time scale.



**Supplementary Figure 3.** Calibration plots of predicted and observed probabilities for the development cohort at A) five years; and B) ten years and for validation cohort at C) five years; and D) ten years.

**Supplementary Table 1.** Scaled Schoenfeld residuals tests p-values, tested against different time scales

Predictor	Linear	Kaplan-Meier-transformed	Logarithmic-transformed
Sex	0.845	0.658	0.270
Age	0.010	0.010	0.027
Family history for stroke	0.515	0.427	0.834
Hypertension	0.595	0.549	0.479
Smoking status	0.040	0.038	0.072
Hypercholesterolemia	0.890	0.933	0.826
Regular physical activity	0.847	0.856	0.332
DM	0.411	0.477	0.303
Alcohol consumption	0.567	0.606	0.121
Educational attainment	0.318	0.370	0.827
Interactions			
Smoking status*Alcohol consumption	0.270	0.225	0.383
Regular physical activity*Hypercholesterolemia	0.902	0.985	0.406
Regular physical activity*DM	0.949	0.919	0.428
Sex*Age	0.751	0.942	0.142
Sex*Hypertension	0.872	0.714	0.651
Sex*Smoking status	0.122	0.094	0.450
Global	0.482	0.430	0.522

DM = diabetes mellitus.

**Supplementary Table 2.** The original regression equation of the SMA<sup>2</sup>SH<sup>2</sup>ERS risk prediction model

<b>Linear predictor (LP)</b> -0.916 (if woman) +
0.025*age if woman +
0.013*age if man +
0.130 (if family history of stroke) +
-0.292 (if woman and hypertension) +
0.372 (if man and hypertension) +
0.341 (if never smoker and no alcohol consumption) +
0.123 (if never smoker and daily or almost daily alcohol consumption) +
0.281 (if former smoker and no alcohol consumption) +
-0.156 (if woman, former smoker and alcohol consumption on special occasions) +
0.121 (if man, former smoker and alcohol consumption on special occasions) +
-0.085 (if former smoker and daily or almost daily alcohol consumption) +
-0.744 (if current smoker and no alcohol consumption) +
0.430 (if woman, current smoker and alcohol consumption on special occasions) +
0.530 (if man, current smoker and alcohol consumption on special occasions) +
0.021 (if current smoker and daily or almost daily alcohol consumption) +
-0.388 (if regular physical activity and hypercholesterolemia) +
0.658 (if regular physical activity and diabetes mellitus) +
0.017 (if regular physical activity and no hypercholesterolemia or diabetes mellitus) +
-0.022 (if hypercholesterolemia and no regular physical activity) +
-0.352 (if diabetes mellitus and no regular physical exercise) +
0.047 (if low educational attainment) +
-0.173 (if high educational attainment) +
<b>Mean LP</b>
1.194258



## Chapter 6

# Identifying novel risk factors for aneurysmal subarachnoid haemorrhage using machine learning

---

Jos P. Kanning, Junfeng Wang, Shahab Abtahi, Mirjam I. Geerlings, Ynte M. Ruigrok

Submitted

## ABSTRACT

Aneurysmal subarachnoid haemorrhage (aSAH) is a type of stroke type with high mortality and morbidity. This study aimed to identify novel aSAH risk factors by combining machine learning (ML) and traditional statistical methods. Using the UK Biobank, we identified aSAH cases via hospital-based ICD codes and analysed 618 baseline variables covering demographics, lifestyle, medical history, and physical measurements. The CatBoost ML algorithm and SHapley Additive Explanations (SHAP) identified the top 25 variables most influential in predicting aSAH. Logistic regression further described these variables while adjusting for established aSAH risk factors. Among 501,847 participants, 893 aSAH cases were identified. ML identified 214 variables with non-zero SHAP values. Logistic regression of the top 25 variables revealed six potential novel aSAH risk factors. Increased aSAH risk was associated with mean spheroid cell volume (OR 1.11, 95% CI 1.04-1.19), urea levels (OR 1.11, 95% CI 1.04-1.18), and tea intake (OR 1.09, 95% CI 1.03-1.15). Decreased aSAH risk was associated with peak expiratory flow (OR 0.91, 95% CI 0.84-0.98), insulin-like growth factor 1 (OR 0.93, 95% CI 0.87-1.00), and haematocrit percentage (OR 0.91, 95% CI 0.84-0.99). Future research should validate these findings and explore the potential non-linear relationships and interactions indicated by the ML models.

## INTRODUCTION

Aneurysmal subarachnoid haemorrhage (aSAH) is a type of stroke that occurs when an intracranial aneurysm ruptures.<sup>1</sup> Despite accounting for only 10% of all strokes,<sup>2</sup> aSAH is particularly devastating due to its early age of onset and high mortality rate, leading to a number of potential life years lost comparable to those lost to ischaemic stroke, the most common type of stroke.<sup>3</sup> Current understanding of the pathogenesis of aSAH remains limited.<sup>4</sup> Established risk factors include age, female sex, hypertension, smoking, and excessive alcohol consumption.<sup>5</sup> However, prediction models incorporating these established risk factors are only moderately able to discriminate between aSAH cases and controls.<sup>6,7</sup> Thus, knowledge of additional risk factors is required to identify people at risk of aSAH.

Machine learning has emerged as a promising strategy for identifying novel risk factors for various medical conditions.<sup>8-10</sup> Unlike traditional statistical methods, machine learning can process large and diverse datasets, uncover complex non-linear associations and their interactions, and does not require precise model specification before analysis.<sup>11-13</sup> However, one limitation of machine learning is that the models it generates can be difficult to interpret and explain in human terms.<sup>14</sup> Integrating machine learning with traditional statistical methods could provide a comprehensive solution, by leveraging the depth and complexity of machine learning analysis while preserving the interpretability of traditional models.<sup>8</sup>

In this study, we aimed to identify new risk factors for aSAH by combining machine learning and statistical methods, using data from the United Kingdom (UK) Biobank's population-based cohort.

## METHODS

### Data source

The UK Biobank, an ongoing large-scale prospective population-based cohort study, has collected health-related data from over 500,000 participants, recruited between 2006 and 2010, who were aged 37 to 73 years at baseline.<sup>15</sup> A systematic medical history was taken for each participant on their assessment date, including touchscreen questionnaires, verbal interviews, physical measurements, and biological sample assays. Additionally, the UK Biobank performs linkage to external hospital inpatient records, which include details on admission dates, diagnoses (including underlying conditions), procedures, and discharge information.

## Outcome

We defined all aSAH cases between January 1, 1997, and October 31, 2022, using the hospital-based International Classification of Diseases, 9th Revision (ICD-9) code '430' and 10th Revision (ICD-10) codes I60.0-I60.9. We specifically included codes I60.8 and I60.9, usually associated with non-aneurysmal cases, because there is a high likelihood that these codes also include aSAH cases within the UK Biobank data. Typically, non-aneurysmal subarachnoid haemorrhages account for about 10-15% of all subarachnoid haemorrhage cases.<sup>16</sup> However, in the UK Biobank, codes I60.8 and I60.9 represent 59.4% of all subarachnoid haemorrhage cases, with 530 out of 893 cases, suggesting a probable inclusion of aSAH cases under these codes. Individuals diagnosed with aSAH before their assessment date were excluded from further analysis.

## Predictors

We included all variables that were systematically assessed on the baseline assessment visit and available for at least 80% of the total cohort. This selection encompassed 618 variables, spanning categories such as patient characteristics (e.g., age, sex), sociodemographic factors (e.g., education, ethnicity), lifestyle factors (e.g., smoking status, diet), family and medical history, medication use, physical measurements (e.g., weight, blood pressure), blood assays, and environmental factors (e.g., noise and air pollution of residence area).

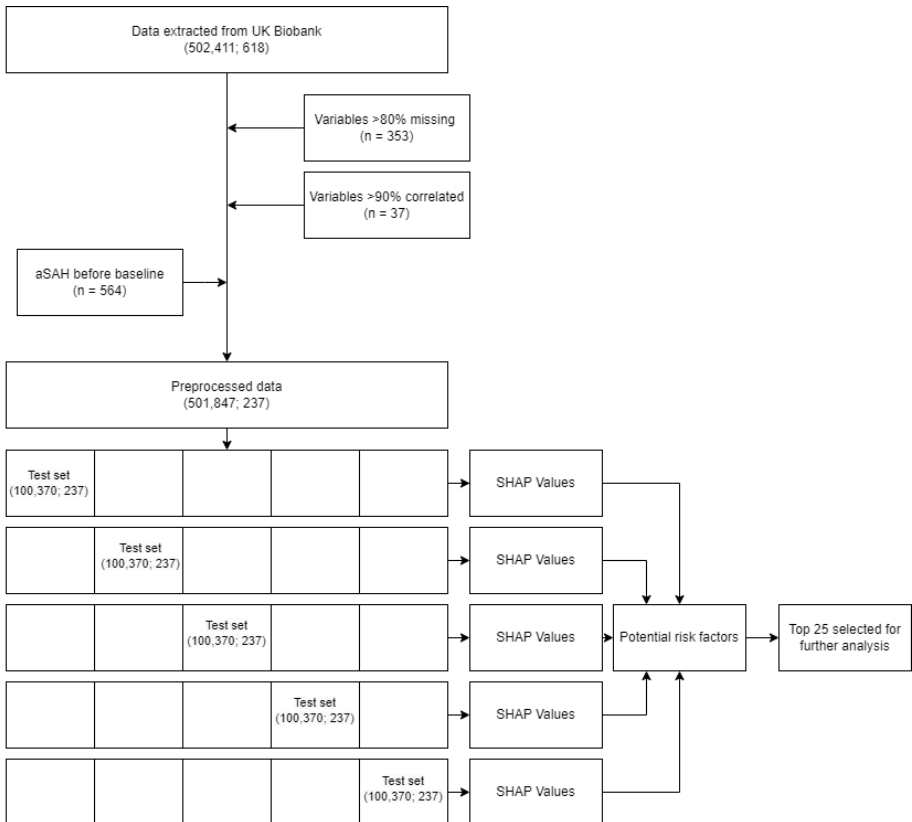
In the pre-processing of data, variables that allowed multiple responses from participants were converted into binary vectors, with each potential response encoded as a separate variable. For example, in response to the original question, "Vascular/heart problems diagnosed by a doctor," participants could select from four options: heart attack, angina, stroke, and high blood pressure. Consequently, we created four distinct binary (yes/no) variables. For variables involving multiple same-day measures, such as systolic blood pressure, the values were averaged to create a single variable. In cases of highly correlated variables (absolute Pearson correlation greater than 0.90), we retained the variable with the fewest missing data points. We considered answers such as "Do not know" and "Prefer not to answer" as missing values. We refrained from removing outliers and instead relied on the outlier checks already implemented by the UK Biobank.<sup>17</sup> For instance, responses to "How many cups of tea do you drink each day?" indicating more than 99 per day were automatically excluded, while responses over 20 prompted verification from participants.

Based on the literature, we identified five established risk factors for aSAH: age at baseline, female sex, hypertension, smoking, and alcohol use.<sup>5</sup> Hypertension

was defined as meeting one or more of the following criteria: an average systolic blood pressure over 140, diastolic blood pressure over 90, taking blood pressure medication, or having a doctor-diagnosed condition. Smoking status was categorised as 'never,' 'previous,' or 'current.' Alcohol consumption was categorised as 'never,' 'rarely' (defined as one to three times a month, or on special occasions), 'often' (defined as one to four times a week), and 'daily.'

### Statistical analysis

Our analysis consisted of two parts (Figure 1). In the first part, we used a machine learning algorithm to identify potential risk factors for aSAH without predefined hypotheses. In the second part, we used logistic regression to examine and quantify these potential risk factors, while adjusting for established risk factors.



**Figure 1.** Flowchart of study design.

Numbers in parentheses indicate the number of rows (i.e. participants) and columns (i.e. variables) respectively. aSAH = aneurysmal subarachnoid haemorrhage, SHAP = SHapley Additive exPlanations.

### **Machine Learning algorithm**

In the first part, we used the CatBoost machine learning algorithm to identify potential aSAH risk factors.<sup>18</sup> CatBoost is a gradient-boosting algorithm that operates on decision trees and is designed to process both numerical and categorical data without extensive pre-processing. A key advantage of CatBoost is its capability to automatically handle missing values without the need for imputation. We randomly divided the dataset into five equal-sized folds, each containing a similar proportion of aSAH cases. We trained a CatBoost model on four folds and reserved the fifth for validation. This cross-validation process was repeated until each fold served as the validation fold once, maximising the area under the curve (AUC) score for each validation fold. To assess variable importance, we calculated SHapley Additive Explanations (SHAP) values for each of the five models.<sup>19</sup> We then computed the average of the mean absolute SHAP values across the folds to identify the 25 variables with the highest mean absolute SHAP values. These variables represent the variables that had the greatest influence on the predicted probability of aSAH.

### **Traditional statistical model**

In the second part, we used traditional statistical methods to further analyse the 25 variables with the highest mean absolute SHAP values, identified in the first part. We did not further analyse established risk factors when identified by the Catboost model. We additionally removed variables with a variance inflation factor higher than 5 to account for multicollinearity. We addressed missing data by using multiple imputation by chained equations (MICE) with five iterations and five imputed datasets.<sup>20</sup> We then visually examined the distribution of continuous variables and applied a logarithmic transformation to those variables that exhibited severe right-skewness. We proceeded to fit both univariable (unadjusted) logistic regression models and models adjusted for established risk factors to the imputed datasets. In the adjusted models, we designated "never" as the reference category for smoking status and "rarely" for alcohol use. We present differences in summary statistics for the established risk factors, using two-sample t-tests for numerical variables and chi-squared tests for the categorical variables. Finally, we presented the unadjusted and adjusted odds ratios (OR), 95% confidence intervals (95% CI), and p-values for each potential risk factor for aSAH by pooling the datasets and using Rubin's rules.<sup>21</sup> We report the ORs as the increased aSAH risk associated with a 1 standard deviation increase of that variable to account for different variable scales. The machine learning model was developed in Python (version 3.13),<sup>22</sup> using the packages pandas,<sup>23</sup> numpy,<sup>24</sup> catboost,<sup>18</sup> shap,<sup>19</sup> and scikit-learn.<sup>25</sup> The traditional statistical models were developed in R (version 4.4.0),<sup>26</sup> using the ggplot2,<sup>27</sup> dplyr,<sup>28</sup> and mice packages.<sup>29</sup>

## Reporting standards

All results are reported according to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.<sup>30</sup>

The UK Biobank received ethics approval from the North West Multi-Center Research Ethics Committee (REC No. 16/NW/0274), and all participants provided written electronic informed consent.

## RESULTS

We identified 893 aSAH patients among 501,847 participants (Table 1). The aSAH patients tended to be older, predominantly female, more frequently current smokers, and more likely to have hypertension compared to the individuals who did not develop aSAH. Finally, aSAH patients were less likely to drink alcohol, but if they did, they drank more often.

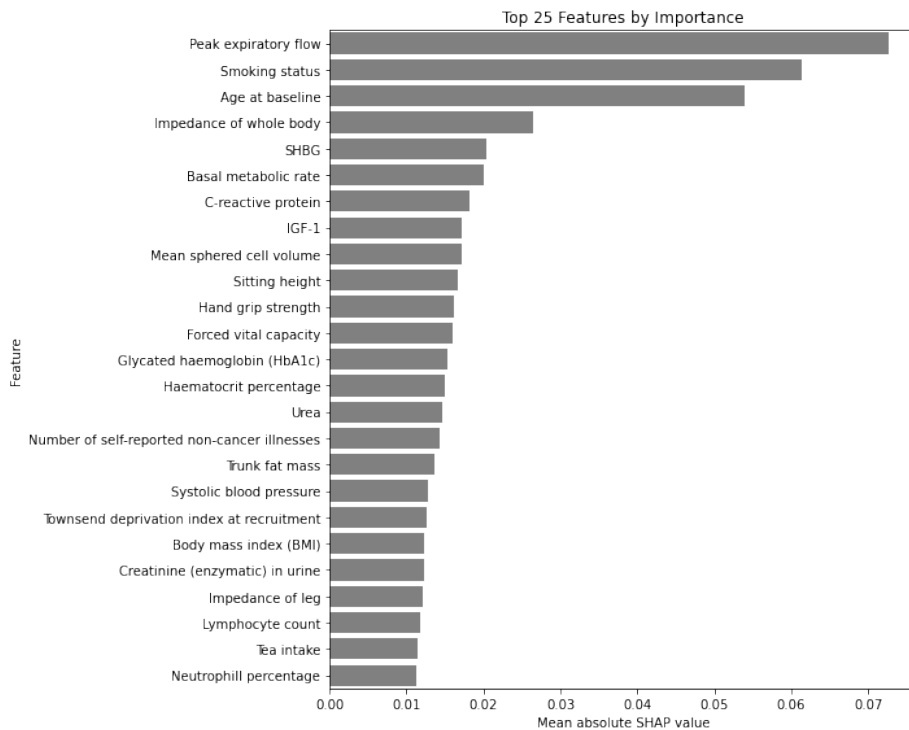
**Table 1.** Baseline characteristics.

Variable	No aSAH	aSAH	p-value
n, (%)	500,954 (99.82)	893 (0.18)	NA
Age at baseline, mean (SD)	56.53 (8.1)	58.44 (7.64)	<0.001
Female, n (%)	272,407 (54.49)	567 (63.50)	<0.001
Hypertension, n (%)	261,542 (52.22)	520 (58.24)	<0.001
Smoking status, n (%)			<0.001
Current	52,689 (10.52)	165 (18.48)	
Previous	172,481 (34.43)	298 (33.37)	
Never	272,845 (54.47)	423 (47.37)	
Unknown	2,939 (0.59)	7 (0.78)	
Alcohol use, n (%)			0.001
Daily	101,449 (20.25)	192 (21.50)	
Often	244,069 (48.72)	388 (43.45)	
Rarely	113,475 (22.65)	210 (23.52)	
Never	40,463 (8.08)	101 (11.31)	
Unknown	1,498 (0.30)	2 (0.22)	

P-values are derived from a two-sample t-test for age and Chi-squared tests for the other variables. aSAH = Aneurysmal subarachnoid haemorrhage, OR = Odds-ratio, CI = Confidence interval, SD = Standard deviation.

## Machine learning model

Initial pre-processing resulted in 235 variables to be used by the CatBoost model of which 214 variables were identified with a mean absolute SHAP value greater than 0. The 25 variables with the highest mean absolute SHAP values were: peak expiratory flow, smoking status, age at baseline, impedance of whole body, sex hormone binding globulin (SHBG), basal metabolic rate, C-reactive protein (CRP), insulin-like growth factor 1 (IGF-1), mean spheroid cell volume, sitting height, hand grip strength, forced vital capacity, glycosylated haemoglobin (HbA1c), haematocrit percentage, urea, number of self-reported non-cancer illnesses, trunk fat mass, systolic blood pressure, Townsend deprivation index at recruitment, body mass index (BMI), creatinine (enzymatic) in urine, impedance of leg, lymphocyte count, tea intake, and neutrophil percentage (Figure 2).



**Figure 2.** The 25 most important potential risk factors identified by the CatBoost algorithm.

SHAP = SHapley Additive exPlanations, SHBG = Sex hormone binding globulin, IGF-1 = Insulin-like growth factor 1.

**Table 2.** The potential aneurysmal subarachnoid haemorrhage (aSAH) risk factors analysed by traditional statistical methods.

Potential risk factors	No aSAH	aSAH	Unadjusted OR (95% CI)	Adjusted* OR (95% CI)
Log(Peak expiratory flow), mean (SD)	5.87 (0.40)	5.77 (0.45)	<b>0.81** (0.77-0.86)</b>	<b>0.91** (0.84-0.98)</b>
Log(SHBG), mean (SD)	3.81 (0.51)	3.91 (0.51)	<b>1.22** (1.13-1.30)</b>	<b>1.14** (1.06-1.24)</b>
Log(C-reactive protein), mean (SD)	0.31 (1.04)	0.42 (1.03)	<b>1.10** (1.03-1.17)</b>	1.01 (0.95-1.09)
IGF-1, mean (SD)	21.37 (5.58)	20.70 (5.64)	<b>0.88** (0.82-0.95)</b>	<b>0.93** (0.87-1.00)</b>
Mean spheroid cell volume, mean (SD)	82.79 (5.14)	83.52 (5.17)	<b>1.15** (1.08-1.23)</b>	<b>1.11** (1.04-1.19)</b>
Sitting height, mean (SD)	89.15 (4.88)	88.26 (4.86)	<b>0.83** (0.77-0.88)</b>	<b>0.92** (0.84-0.99)</b>
Hand grip strength, mean (SD)	30.61 (11.03)	28.60 (11.17)	<b>0.83** (0.77-0.89)</b>	0.92 (0.84-1.01)
Log(Forced vital capacity), mean (SD)	1.26 (0.29)	1.19 (0.31)	<b>0.82** (0.78-0.87)</b>	0.97 (0.89-1.06)
Glycated haemoglobin (HbA1c), mean (SD)	35.49 (4.50)	35.74 (4.30)	1.05 (0.98-1.12)	0.99 (0.93-1.06)
Haematocrit percentage, mean (SD)	41.08 (3.54)	40.63 (3.35)	<b>0.88** (0.82-0.93)</b>	<b>0.91** (0.84-0.99)</b>
Urea, mean (SD)	5.37 (1.28)	5.47 (1.37)	<b>1.07** (1.00-1.14)</b>	<b>1.11** (1.04-1.18)</b>
Number of self-reported non-cancer illnesses, mean (SD)	1.86 (1.87)	2.16 (2.26)	<b>1.15** (1.09-1.22)</b>	<b>1.09** (1.03-1.16)</b>
Systolic blood pressure, mean (SD)	137.81 (18.68)	140.25 (18.62)	<b>1.14** (1.06-1.20)</b>	<b>1.12** (1.02-1.23)</b>
Townsend deprivation index at recruitment, mean (SD)	-1.29 (3.09)	-1.15 (3.16)	1.05 (0.98-1.12)	0.99 (0.93-1.06)
Body mass index (BMI), mean (SD)	27.43 (4.80)	26.94 (4.79)	<b>0.89** (0.83-0.96)</b>	<b>0.87** (0.81-0.93)</b>
Log(Creatinine (enzymatic) in urine), mean (SD)	8.88 (0.68)	8.83 (0.68)	0.94 (0.88-1.00)	1.01 (0.94-1.09)
Impedance of leg, mean (SD)	247.36 (35.27)	252.11 (36.28)	<b>1.15** (1.07-1.24)</b>	<b>1.07** (1.00-1.15)</b>
Lymphocyte count, mean (SD)	1.94 (0.60)	1.97 (0.65)	1.05 (0.99-1.12)	0.97 (0.90-1.04)
Tea intake, mean (SD)	3.41 (2.90)	3.72 (3.36)	<b>1.10** (1.04-1.16)</b>	<b>1.09** (1.03-1.15)</b>
Neutrophil percentage, mean (SD)	60.88 (8.52)	60.89 (9.21)	1.00 (0.93-1.07)	0.99 (0.93-1.06)

All odds ratios for numerical variables indicate the increased aSAH risk associated with a 1 standard deviation increase of that variable. \*: Adjusted for age at baseline, female sex, hypertension, smoking status, and alcohol use. \*\*: Statistically significant at a threshold of  $p < 0.05$  (indicated in bold). OR = Odds-ratio, CI = Confidence interval, SD = Standard deviation, SHBG = Sex hormone binding globulin, IGF-1 = Insulin-like growth factor 1.

### Traditional statistical model

We did not further investigate age at baseline and smoking status, as they were already included as established risk factors. We also excluded whole body impedance, basal metabolic rate, and trunk fat mass from our study due to their high variance inflation factors, leaving 20 potential risk factors for further analysis (Table 2).

We identified 8 of the 20 variables that were univariably associated with an increased risk of aSAH. These included log-transformed SHBG levels (OR 1.22, 95% CI 1.13-1.30), log-transformed CRP levels (OR 1.10, 95% CI 1.03-1.17), mean sphered cell volume (OR 1.15, 95% CI 1.08-1.23), urea (OR 1.07, 95% CI 1.00-1.14), number of self-reported non-cancer illnesses (OR 1.15, 95% CI 1.09-1.22), systolic blood pressure (OR 1.14, 95% CI 1.06-1.20), impedance of leg (OR 1.15, 95% CI 1.07-1.24), and tea intake (OR 1.10, 95% CI 1.04-1.16). After adjusting for established risk factors, the associations remained statistically significant for log-transformed SHBG levels (OR 1.14, 95% CI 1.06-1.24), mean sphered cell volume (OR 1.11, 95% CI 1.04-1.19), urea (OR 1.11, 95% CI 1.04-1.18), number of self-reported non-cancer illnesses (OR 1.09, 95% CI 1.03-1.16), systolic blood pressure (OR 1.12, 95% CI 1.02-1.23), impedance of leg (OR 1.07, 95% CI 1.00-1.15), and tea intake (OR 1.09, 95% CI 1.03-1.15). A sensitivity analysis revealed that compared to low tea intake (less than 2 cups a day), medium tea intake (between 2 and 5 cups) was associated with a decreased aSAH risk (OR 0.92, 95% CI 0.78 -1.09), whereas high intake was associated with an increased risk (OR 1.14, 95% CI 0.96 -1.36).

We also found 7 of the 20 variables to be univariably associated with a decreased aSAH risk: log-transformed peak expiratory flow (OR 0.81, 95% CI 0.77-0.86), IGF-1 (OR 0.88, 95% CI 0.82-0.95), sitting height (OR 0.83, 95% CI 0.77-0.88), hand grip strength (OR 0.83, 95% CI 0.77-0.89), log-transformed forced vital capacity (OR 0.82, 95% CI 0.78-0.87), haematocrit percentage (OR 0.88, 95% CI 0.82-0.93), and BMI (OR 0.89, 95% CI 0.83-0.96). After adjusting for established risk factors, the associations remained statistically significant for log-transformed peak expiratory flow (OR 0.91, 95% CI 0.84-0.98), IGF-1 (OR 0.93, 95% CI 0.87-1.00), sitting height (OR 0.92, 95% CI 0.84-0.99), haematocrit percentage (OR 0.91, 95% CI 0.84-0.99), and BMI (OR 0.87, 95% CI 0.81-0.93).

For the remaining 5 variables—HbA1c, Townsend deprivation index, log-transformed creatinine in urine, lymphocyte count, and neutrophil percentage—we did not find any statistically significant associations with aSAH.

## DISCUSSION

Using a combination of machine learning and traditional statistical approaches, we identified seven variables associated with an increased risk of aSAH. These included log-transformed SHBG, mean spheroid cell volume, urea levels, the number of self-reported non-cancer illnesses, systolic blood pressure, leg impedance, and tea consumption. In contrast, five variables were linked to a decreased risk of aSAH: log-transformed peak expiratory flow, IGF-1, sitting height, haematocrit percentage, and BMI.

Our analysis identified three new potential risk factors for aSAH. The first, mean spheroid cell volume, measures red blood cells in a spherical state. Although no prior studies have linked mean cell spheroid volume to aSAH, it could be associated with aSAH via macrocytosis, which leads to increased blood viscosity and possibly a higher risk of rupture.<sup>31</sup> However, the small effect size we observed for this risk factor may suggest alternative mechanisms. For example, the effect may be mediated through other risk factors associated with high mean cell spheroid volume, such as vitamin B12 deficiency.<sup>32</sup> In turn, vitamin B12 deficiency may be associated with extreme forms of the established aSAH risk factors of alcohol abuse and smoking.<sup>33</sup> Although we have adjusted for alcohol use and smoking in our analysis, there remains a risk of residual confounding of their extreme forms. The second risk factor, urea, is a waste product synthesised by the kidneys to eliminate excess nitrogen. Urea has not been directly associated with stroke, although conditions known to be associated with aSAH (e.g. polycystic kidney disease) may explain the association between urea levels and aSAH risk.<sup>34</sup> Unmanaged hypertension, an established risk factor for aSAH,<sup>5</sup> can similarly lead to kidney damage, elevating urea levels. High urea levels might also damage the vascular endothelium,<sup>35</sup> leading to dysfunction and potentially contributing to the formation and rupture of cerebral aneurysms by weakening the vessel walls. Finally, our research suggests an increased risk of aSAH with high tea consumption, contrasting with studies indicating no association or a reduced risk.<sup>36–38</sup> This discrepancy might be due to differences in the amount of tea consumed. For instance, tea intake in the UK Biobank averages 3.5 cups a day, considerably higher than the at least 1 cup a day defined in other studies. Our sensitivity analysis confirmed this, with a reduced aSAH risk observed for moderate tea intake, and an increased aSAH risk for high intake. Moderate tea consumption may reduce aSAH risk due to antioxidants and anti-inflammatory compounds in tea that improve vascular health and lower blood pressure.<sup>36</sup> These compounds may strengthen blood vessel walls, reducing the likelihood of aneurysms. However, excessive tea intake could increase aSAH risk because high levels of caffeine can elevate blood pressure and induce vascular stress.<sup>39</sup>

Our analysis also identified three new variables potentially associated with a decreased aSAH risk. We found an inverse relationship between peak expiratory flow and the risk of aSAH. Peak expiratory flow is an indicator of lung function which measures the fastest speed at which a person can exhale air after a maximal inhalation. Similar inverse associations between stroke risk and peak expiratory flow have been documented in previous studies.<sup>40,41</sup> Although peak expiratory flow is commonly linked with cardiovascular risk factors like hypertension and smoking,<sup>42</sup> our findings indicate an association even after adjusting for these factors. One possible explanation is that low peak expiratory flow may reflect diminished lung function and possibly chronic hypoxia, which can contribute to vascular remodelling and alterations in blood pressure regulation.<sup>43</sup> This condition could lead to changes in brain blood vessels, increasing their susceptibility to aneurysm formation or exacerbating existing arterial weaknesses. We also found an inverse relationship between IGF-1 and aSAH risk. Evidence suggests that a deficiency in IGF-1 may contribute to the development of intracerebral haemorrhages.<sup>44</sup> IGF-1 has been shown to exert anti-inflammatory effects in various tissues.<sup>45</sup> It can reduce the expression of pro-inflammatory cytokines such as tumour necrosis factor-alpha and interleukin 6, which are implicated in the weakening of the arterial wall in aneurysms.<sup>46</sup> IGF-1 may help stabilise the arterial wall by dampening the inflammatory response, reducing the risk of aneurysm formation and rupture. Finally, we identified an inverse relationship between the risk of aSAH and haematocrit percentage, which is the proportion of red blood cells in the bloodstream. Despite previous research reporting no link between aSAH occurrence and haematocrit levels,<sup>47</sup> these levels are often low in patients at hospital admission and can indicate a higher risk of death.<sup>48</sup>

Our results have similarly highlighted several potential risk factors for aSAH that were previously suggested by research but not conclusively established. These include BMI, sitting height, and leg impedance, all relating to body size. The data on the relationship between body size and aSAH risk is inconsistent,<sup>5,49,50</sup> similar to our study. We found a decreased aSAH risk associated with BMI and sitting height and a small increased risk associated with leg impedance. It has been speculated that very lean individuals might have nutritional deficiencies predisposing them to aSAH.<sup>47</sup> Alternatively, the association between aSAH and body size might reflect epidemiological biases such as unmeasured confounding or selection bias.<sup>51</sup> Additionally, our findings suggest that individuals with multiple non-cancer illnesses are at an increased risk for aSAH, which may be due to overall disease vulnerability or the inclusion of established aSAH risk factors such as hypertension in this variable.

Our analysis also corroborates prior research. The CatBoost machine learning algorithm identified age and smoking status as important predictors, a finding supported by our statistical analysis. Although the machine learning model did not directly identify female sex, a well-established risk factor,<sup>5</sup> it did identify SHBG as important. SHBG is a liver-produced protein that binds to sex hormones such as testosterone and oestrogen, regulating their availability in the body. There is evidence that the elevated aSAH risk in women may be hormonally driven,<sup>52</sup> making SHBG levels, or other hormone-related variables, potentially more relevant predictors of aSAH risk than merely biological sex. Similarly, the Catboost model favoured numerical variables such as systolic blood pressure over binary ones like the presence or absence of hypertension. Our analysis shows that the association of systolic blood pressure remains statistically significant even when accounting for a history of hypertension, indicating possible untreated hypertension. Contrary to initial expectations, only a slight increase in aSAH risk was observed among heavy drinkers, which lost statistical significance after adjusting for other risk factors. The association between alcohol use and aSAH is still under investigation,<sup>49</sup> and there is some evidence that the association only exists for excessive use.<sup>53</sup> Our categorical definition of alcohol use based on frequency of use may have missed important information on current or past amounts of alcohol intake. Intriguingly, those who reported never drinking alcohol showed the highest aSAH risk, which could reflect prior excessive consumption or other health issues leading them to abstain.

The CatBoost machine learning algorithm identified several variables that did not show statistically significant associations in the traditional statistical model. These included CRP, hand grip strength, forced vital capacity, HbA1c, Townsend deprivation index, creatinine, lymphocyte count, and neutrophil percentage. These variables might exhibit non-linear associations with aSAH or depend on interaction terms, making them detectable by CatBoost but not by logistic regression. For example, HbA1c may only relate to aSAH beyond a certain threshold, as aSAH is associated with diabetes.<sup>5</sup> Similarly, variables such as forced vital capacity, which were univariably associated with aSAH but not after adjustment for established risk factors, may be either confounded or indicative of an unknown mediator role. For example, the observed increased aSAH risk in smokers might be partially explained by elevated CRP levels.<sup>54</sup> These are observational findings, however, and further validation is necessary to substantiate these claims.

This study has several strengths. Firstly, the use of a large dataset, including numerous individuals and variables, facilitated the identification of sufficient cases and variables for a hypothesis-generating approach. Additionally, variables in the

UK Biobank are systematically assessed for each individual at baseline, enabling us to evaluate the prognostic significance of each variable. This systematic assessment of variables also allowed for proper adjustment of established risk factors, thereby reducing false positives. Finally, by integrating machine learning with statistical methods, we were able to both filter and characterise the variables associated with aSAH. Relying solely on machine learning would not have permitted quantification of predictor effects, just as relying solely on statistical methods would have precluded analysis of the entire dataset.

This study also had several limitations. First, the rarity of aSAH resulted in poor model performance in the traditional statistical models. Consequently, automatic model selection procedures often yielded an empty null model, restricting our statistical analysis. This limitation confined our statistical modelling to linear associations, as exploring non-linear relationships or interactions was not feasible. Efforts to address class imbalance through sampling or weighting methods produced unreliable SHAP values, failing to identify established aSAH risk factors such as age and smoking. As a result, we chose to present our findings without adjustments for class imbalance. Moreover, the rarity of the condition necessitated analysing the entire dataset rather than dividing it into development and validation cohorts, as is typical in machine learning studies.<sup>55</sup> This limitation meant we could conduct only minimal internal validation, and our results might not be generalisable to an external dataset. Another limitation is the demographic composition of the UK Biobank participants, who are predominantly older, Caucasian, and from upper-middle-class backgrounds.<sup>56</sup> This demographic skew may limit the generalizability of our findings to other populations. Finally, we did not adjust for multiple comparisons, as the study aimed to explore risk factors without a pre-defined hypothesis for the statistical tests. Furthermore, the selection of potential risk factors was not based on statistical significance.

In conclusion, we have identified six new potential risk factors for aSAH. Mean spheroid cell volume, urea levels, and tea intake were associated with increased aSAH risk. In contrast, peak expiratory flow, IGF-1, and haematocrit percentage were associated with decreased aSAH risk. Our research builds upon previous studies by providing further evidence that body size is associated with aSAH risk, by suggesting that hormonal levels may partially explain the higher aSAH risk among women, and by confirming previously established risk factors. Future studies should use larger sample sizes to validate these preliminary findings. Additionally, special attention should be given to factors identified by the Catboost algorithm that were not statistically significant in the logistic regression model. This could indicate potential non-linearities or interaction effects of these factors which are detectable by CatBoost but not by logistic regression.

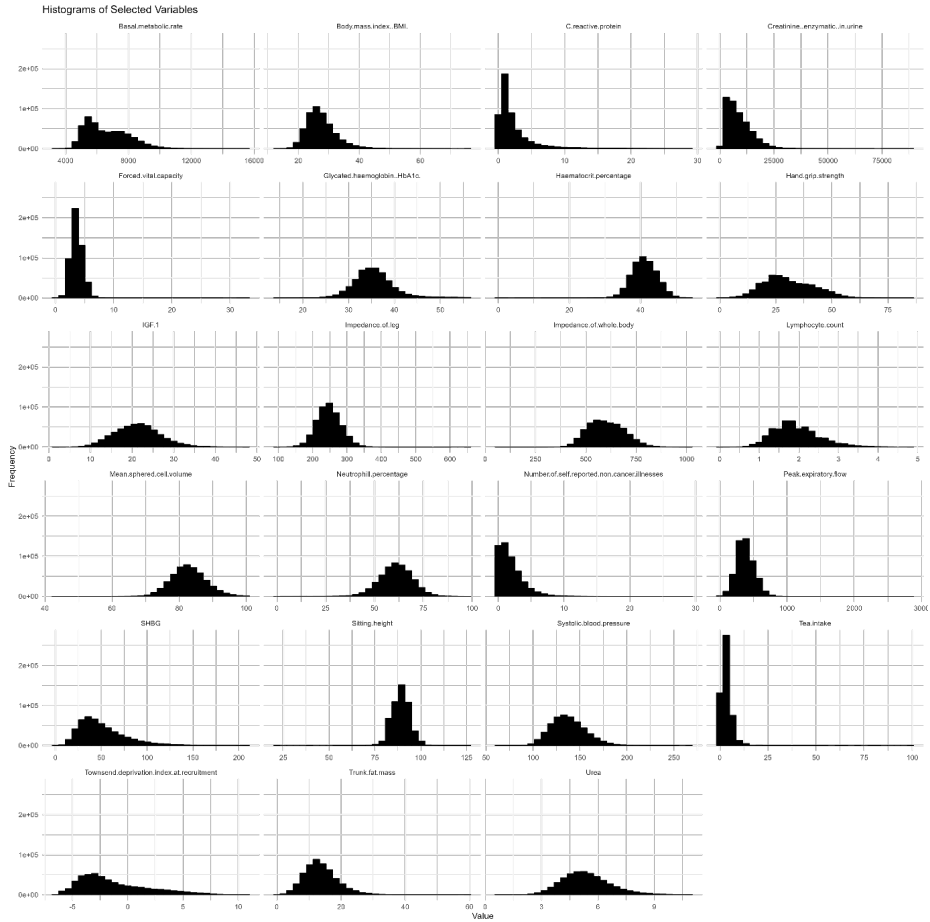
## REFERENCES

1. Macdonald, R. L. & Schweizer, T. A. Spontaneous subarachnoid haemorrhage. *The Lancet* **389**, 655–666 (2017).
2. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol* **20**, 795–820 (2021).
3. Johnston, S. C., Selvin, S. & Gress, D. R. The burden, trends, and demographics of mortality from subarachnoid hemorrhage. *Neurology* **50**, 1413–1418 (1998).
4. Etminan, N. & Rinkel, G. J. Unruptured intracranial aneurysms: development, rupture and preventive management. *Nat Rev Neurol* **12**, 699–713 (2016).
5. Feigin, V. L. *et al.* Risk factors for subarachnoid hemorrhage: an updated systematic review of epidemiological studies. *Stroke* **36**, 2773–2780 (2005).
6. Bakker, M. K. *et al.* Genetic Risk Score for Intracranial Aneurysms: Prediction of Subarachnoid Hemorrhage and Role in Clinical Heterogeneity. *Stroke* **54**, 810–818 (2023).
7. Kanning, J. P. *et al.* Prediction of aneurysmal subarachnoid hemorrhage in comparison with other stroke types using routine care data. *PLOS ONE* (2024).
8. Madakkatel, I., Zhou, A., McDonnell, M. D. & Hyppönen, E. Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. *Sci Rep* **11**, 22997 (2021).
9. García de la Garza, Á., Blanco, C., Olfson, M. & Wall, M. M. Identification of Suicide Attempt Risk Factors in a National US Survey Using Machine Learning. *JAMA Psychiatry* **78**, 398–406 (2021).
10. Lee, K.-S., Jha, N. & Kim, Y.-J. Risk factor assessments of temporomandibular disorders via machine learning. *Sci Rep* **11**, 19802 (2021).
11. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
12. Bi, Q., Goodman, K. E., Kaminsky, J. & Lessler, J. What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology* **188**, 2222–2239 (2019).
13. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science* **16**, 199–231 (2001).
14. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1**, 206–215 (2019).
15. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
16. Claassen, J. & Park, S. Spontaneous subarachnoid haemorrhage. *Lancet* **400**, 846–862 (2022).
- 17.: Resource 113241. <https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=113241>.
18. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulina, A. CatBoost: unbiased boosting with categorical features. Preprint at <https://doi.org/10.48550/arXiv.1706.09516> (2019).
19. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at <https://doi.org/10.48550/arXiv.1705.07874> (2017).
20. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* **20**, 40–49 (2011).
21. Rubin, D. B. Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Ann Intern Med* **127**, 757–763 (1997).

22. The Python Language Reference. *Python documentation* <https://docs.python.org/3/reference/index.html>.
23. API reference — pandas 2.2.2 documentation. <https://pandas.pydata.org/docs/reference/index.html>.
24. NumPy Documentation. <https://numpy.org/doc/>.
25. scikit-learn: machine learning in Python — scikit-learn 1.5.0 documentation. <https://scikit-learn.org/stable/>.
26. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* (2021).
27. Wickham, H. *et al.* ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. (2024).
28. Wickham, H. *et al.* dplyr: A Grammar of Data Manipulation. (2023).
29. Buuren, S. van *et al.* mice: Multivariate Imputation by Chained Equations. (2023).
30. von Elm, E. *et al.* Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* **335**, 806–808 (2007).
31. Breedveld, F. C., Bieger, R. & van Wermeskerken, R. K. A. The Clinical Significance of Macrocytosis. *Acta Medica Scandinavica* **209**, 319–322 (1981).
32. Blundell, E. L. *et al.* Importance of low serum vitamin B12 and red cell folate concentrations in elderly hospital inpatients. *J Clin Pathol* **38**, 1179–1184 (1985).
33. Green, R. *et al.* Vitamin B12 deficiency. *Nat Rev Dis Primers* **3**, 17040 (2017).
34. Gieteling, E. W. & Rinkel, G. J. E. Characteristics of intracranial aneurysms and subarachnoid haemorrhage in patients with polycystic kidney disease. *J Neurol* **250**, 418–423 (2003).
35. D'Apollito, M. *et al.* Urea-induced ROS cause endothelial dysfunction in chronic renal failure. *Atherosclerosis* **239**, 393–400 (2015).
36. Zhang, C. *et al.* Tea consumption and risk of cardiovascular outcomes and total mortality: a systematic review and meta-analysis of prospective observational studies. *Eur J Epidemiol* **30**, 103–113 (2015).
37. Larsson, S. C. *et al.* Coffee and Tea Consumption and Risk of Stroke Subtypes in Male Smokers. *Stroke* **39**, 1681–1687 (2008).
38. Okamoto, K. Habitual green tea consumption and risk of an aneurysmal rupture subarachnoid hemorrhage: a case-control study in Nagoya, Japan. *Eur J Epidemiol* **21**, 367–371 (2006).
39. Mesas, A. E., Leon-Muñoz, L. M., Rodriguez-Artalejo, F. & Lopez-Garcia, E. The effect of coffee on blood pressure and cardiovascular disease in hypertensive individuals: a systematic review and meta-analysis. *Am J Clin Nutr* **94**, 1113–1126 (2011).
40. Söderholm, M., Zia, E., Hedblad, B. & Engström, G. Lung Function as a Risk Factor for Subarachnoid Hemorrhage. *Stroke* **43**, 2598–2603 (2012).
41. Persson, C. *et al.* Peak expiratory flow and risk of cardiovascular disease and death. A 12-year follow-up of participants in the population study of women in Gothenburg, Sweden. *Am J Epidemiol* **124**, 942–948 (1986).
42. Cook, N. R. *et al.* Peak expiratory flow rate in an elderly population. *Am J Epidemiol* **130**, 66–78 (1989).
43. Lim, C. S., Kiriakidis, S., Sandison, A., Paleolog, E. M. & Davies, A. H. Hypoxia-inducible factor pathway and diseases of the vascular wall. *J Vasc Surg* **58**, 219–230 (2013).
44. Fulop, G. A. *et al.* IGF-1 Deficiency Promotes Pathological Remodeling of Cerebral Arteries: A Potential Mechanism Contributing to the Pathogenesis of Intracerebral Hemorrhages in Aging. *J Gerontol A Biol Sci Med Sci* **74**, 446–454 (2019).

45. Alehagen, U., Johansson, P., Aaseth, J., Alexander, J. & Brismar, K. Increase in insulin-like growth factor 1 (IGF-1) and insulin-like growth factor binding protein 1 after supplementation with selenium and coenzyme Q10. A prospective randomized double-blind placebo-controlled trial among elderly Swedish citizens. *PLoS ONE* **12**, (2017).
46. Middleton, R. K. *et al.* The pro-inflammatory and chemotactic cytokine microenvironment of the abdominal aortic aneurysm wall: a protein array study. *J Vasc Surg* **45**, 574–580 (2007).
47. Knekt, P. *et al.* Risk factors for subarachnoid hemorrhage in a longitudinal population study. *J Clin Epidemiol* **44**, 933–939 (1991).
48. Giller, C. A., Wills, M. J., Giller, A. M. & Samson, D. Distribution of hematocrit values after aneurysmal subarachnoid hemorrhage. *J Neuroimaging* **8**, 169–170 (1998).
49. Feigin, V. *et al.* Smoking and Elevated Blood Pressure Are the Most Important Risk Factors for Subarachnoid Hemorrhage in the Asia-Pacific Region. *Stroke* **36**, 1360–1365 (2005).
50. Hebert, P. R. *et al.* Height and incidence of cardiovascular disease in male physicians. *Circulation* **88**, 1437–1443 (1993).
51. Banack, H. R. & Kaufman, J. S. Does selection bias explain the obesity paradox among individuals with cardiovascular disease? *Annals of Epidemiology* **25**, 342–349 (2015).
52. Mhurchu, C. N. *et al.* Hormonal factors and risk of aneurysmal subarachnoid hemorrhage: an international population-based, case-control study. *Stroke* **32**, 606–612 (2001).
53. Larsson, S. C., Wallin, A., Wolk, A. & Markus, H. S. Differing association of alcohol consumption with different stroke types: a systematic review and meta-analysis. *BMC Medicine* **14**, 178 (2016).
54. Madsen, C. *et al.* Association between tobacco smoke exposure and levels of C-reactive protein in the Oslo II Study. *Eur J Epidemiol* **22**, 311–317 (2007).
55. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLOS ONE* **14**, e0224365 (2019).
56. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026–1034 (2017).

# SUPPLEMENTARY MATERIALS



**Supplementary Figure 1.** Distributions of potential risk factors.

SHBG = Sex hormone binding globulin, IGF-1 = Insulin-like growth factor.





## Chapter 7

# Prescribed Drug Use and Aneurysmal Subarachnoid Haemorrhage Incidence: A Drug-Wide Association Study

---

Jos P. Kanning, Shahab Abtahi, Christian Schnier, Olaf H. Klungel, Mirjam I. Geerlings, Ynte M. Ruigrok

Neurology. 2024 Jun 25;102(12):e209479

## ABSTRACT

**Background:** Current benefits of invasive intracranial aneurysm treatment to prevent aneurysmal subarachnoid haemorrhage (aSAH) rarely outweigh treatment risks. Most intracranial aneurysms thus remain untreated. Commonly prescribed drugs reducing aSAH incidence may provide leads for drug repurposing. We performed a drug-wide association study (DWAS) to systematically investigate the association between commonly prescribed drugs and aSAH incidence.

**Methods:** We defined all aSAH cases between 2000 and 2020 using International Classification of Diseases (ICD) codes from the Secure Anonymised Information Linkage (SAIL) databank. Each case was matched with nine controls based on age, sex, and year of database entry. We investigated commonly prescribed drugs (>2% in study population) and defined three exposure windows relative to the most recent prescription before index date (i.e. occurrence of aSAH): current (within 3 months), recent (3-12 months), and past (>12 months). A logistic regression model was fitted to compare drug use across these exposure windows versus never use, controlling for age, sex, known aSAH risk factors, and healthcare utilisation. The family-wise error rate was kept at  $p < 0.05$  through Bonferroni correction.

**Results:** We investigated exposure to 205 commonly prescribed drugs between 4,879 aSAH cases (mean age 61.4, 61.2% women) and 43,911 matched controls. We found similar trends for lisinopril and amlodipine, with a decreased aSAH risk for current use (lisinopril OR 0.63 95%CI 0.44-0.90, amlodipine OR 0.82 95%CI 0.65-1.04), and an increased aSAH risk for recent use (lisinopril OR 1.30 95%CI 0.61-2.78, amlodipine OR 1.61 95%CI 1.04 -2.48). A decreased aSAH risk in current use was also found for simvastatin (OR 0.78, 95%CI 0.64-0.96), metformin (OR 0.58, 95%CI 0.43-0.78), and tamsulosin (OR 0.55, 95%CI 0.32-0.93). In contrast, an increased aSAH risk was found for current use of warfarin (OR 1.35 95%CI 1.02-1.79), venlafaxine (OR 1.67, 95%CI 1.01-2.75), prochlorperazine (OR 2.15 95%CI 1.45-3.18), and co-codamol (OR 1.31, 95%CI 1.10-1.56).

**Conclusions:** We identified several drugs associated with aSAH, of which five drugs (lisinopril and possibly amlodipine, simvastatin, metformin, and tamsulosin) showed a decreased aSAH risk. Future research should build on these signals to further assess the effectiveness of these drugs in reducing aSAH incidence.

**Classification of evidence:** This study provides class III evidence that some commonly prescribed drugs are associated with subsequent development of aSAH.

## INTRODUCTION

Aneurysmal subarachnoid haemorrhage (aSAH) is a type of stroke which occurs when an intracranial aneurysm ruptures, causing bleeding in the subarachnoid space.<sup>1</sup> Despite the fact that aSAH only accounts for 10% of all strokes,<sup>2</sup> the early age of onset and high mortality rate make the years of potential life lost due to aSAH comparable to those lost due to ischemic stroke, the most common type of stroke.<sup>3</sup> In patients with a known unruptured intracranial aneurysms, aSAH can be prevented by endovascular or neurosurgical treatment.<sup>4</sup> However, these surgical treatment options carry a risk of permanent disability or mortality of up to 8%.<sup>5,6</sup> As the potential benefit of these treatment options often do not outweigh the risk of treatment complications, most intracranial aneurysms remain untreated.<sup>7</sup> Thus, a non-invasive drug compound that can prevent aneurysm rupture would be highly beneficial. Such drugs, however, have yet to be identified.

Conducting a drug-wide association study (DWAS) is a novel paradigm for drug discovery. A DWAS uses large electronic healthcare databases (EHDs) to generate hypotheses about the relationship between drugs and disease.<sup>8</sup> Previous DWAS have effectively identified drug-outcome associations for diverse medical conditions, including myocardial infarction,<sup>8</sup> cancer,<sup>9,10</sup> dementia,<sup>11,12</sup> and COVID-19.<sup>13</sup> Signals derived from DWAS can be further investigated and validated, paving the way for potential drug repurposing.<sup>14</sup>

In this study, we conducted a hypothesis-generating DWAS using a large nationwide EHD. Our primary objective was to identify commonly prescribed drugs that were associated with a lower incidence of aSAH and could be used to prevent aneurysm rupture.

## METHODS

### Setting

This study used data from the Secure Anonymised Information Linkage (SAIL) databank. The SAIL databank works with healthcare providers and government agencies to obtain a diverse array of anonymized datasets, and it currently includes anonymized, individual-level, linked routinely-collected healthcare data (e.g. diagnoses, treatments, medical histories, hospital admissions and discharges, outpatient visits, and prescriptions) for approximately 80% of the population of Wales, UK.<sup>15</sup> For this investigation, we linked data from primary care (Welsh

Longitudinal General Practice dataset [WLGP]), hospital admissions (Patient Episode Database for Wales [PEDW]), and mortality records (Annual District Death Extract [ADDE]). Primary care records in SAIL are catalogued using Read codes (version 2), while hospital admissions and mortality records employ International Classification of Diseases versions 9 (ICD-9) and 10 (ICD-10) codes.

### **Study population & outcome**

We included all patients born before 1 January 1982 with an aSAH hospital diagnosis (ICD-9 code 430 and ICD-10 codes I60.0 - I60.9) and a date of hospital admission (from now on referred to as the index date) between 1 January 2000 and 31 December 2019 (Figure 1). Patients with an observation window shorter than 365 days or a record of aSAH (ICD codes above or Read codes Gyu61, Gyu60, Gyu6E, Gyu64) before January 2000 were excluded from further analysis. For each aSAH case, we randomly matched up to 9 controls without replacement based on year of birth and sex. In addition, we matched based on year of database entry in order to ensure comparable observation windows for cases and controls. Database entry was defined as whichever came first: 1 January 2000, or the first primary care record between January 2000 and December 2019. Similarly, database exit was defined as whichever came first: 31 December 2019, date of death, date of SAIL databank exit (e.g. due to migration), or date of aSAH diagnosis. Controls with a database exit before a case's diagnosis date were no longer eligible to be matched to that case. In addition, controls were required to have a minimum observation window of 365 days. Each matched control was given the same index date as the case they were matched with.

### **Exposure definition**

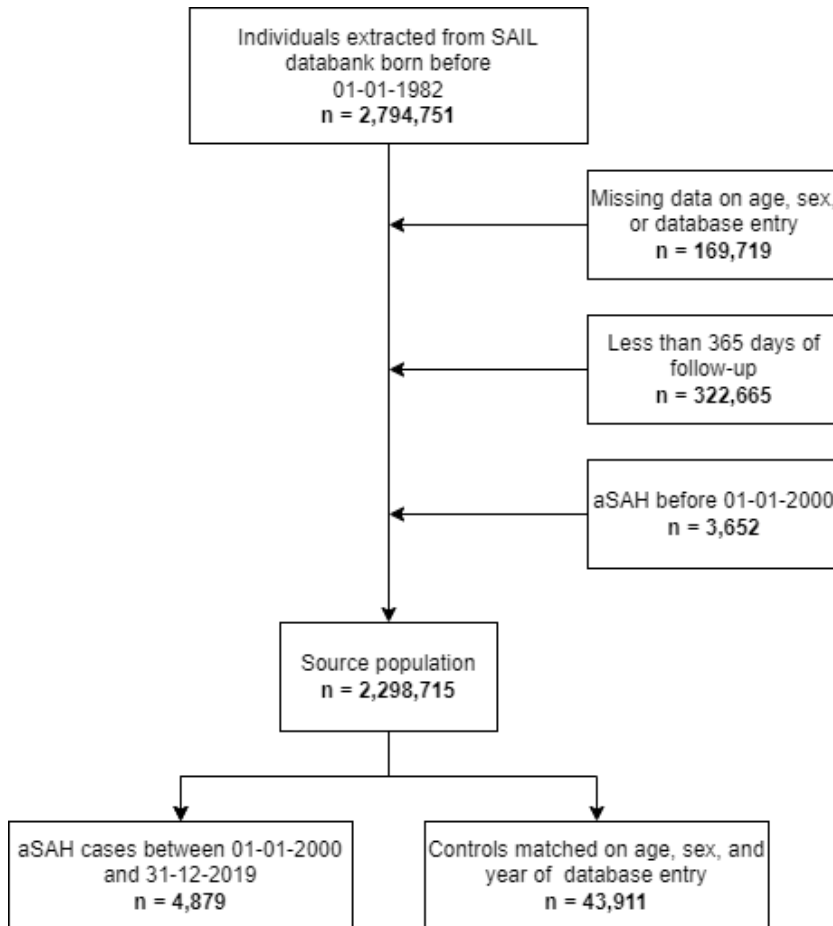
In SAIL, primary care records are encoded using 5-digit Read codes, where codes starting with a lowercase letter indicate drug prescriptions. We included all Read codes with the first character ranging from 'a' to 'o' (Table 1) and clustered them based on their first three characters. For example, we considered 'a136' (600mg sodium bicarbonate) and 'a137' (500mg sodium bicarbonate) to be the same drug: 'a13' (sodium bicarbonate). If the initial clustering produced non-descriptive clusters, we chose clustering based on the first four characters of the Read code instead. For example, we used the specific 'dia7' (co-codamol) instead of the generic 'dia' (compound analgesics A-L). Clustering resulted in a total of 2,023 drugs. We only included drugs that were prescribed in at least 2% of our study population in order to have sufficient statistical power to reliably identify associations between drug exposure and aSAH.

**Table 1.** Clustering read codes and commonly prescribed drugs in Wales between 2000-2019 using the SAIL databank.

Term	Read Code	Clustered Drugs	Commonly Prescribed*
Gastro-intestinal drugs	a....	108	19
Cardiovascular drugs	b....	223	27
Respiratory drugs	c....	110	18
Central nervous system drugs	d....	283	29
Drugs used in infections	e....	204	18
Endocrine drugs	f....	149	9
OBS/GYN/UTI drugs	g....	103	7
Chemotherapy/immunosuppressants drugs	h....	160	0
Haematology/dietetic drugs	i....	172	6
Musculoskeletal drugs	j....	75	14
Eye drugs	k....	109	8
ENT drugs	l....	71	10
Skin drugs	m....	190	32
Immunology drugs	n....	35	6
Anaesthetic drugs	o....	31	2
<b>TOTAL</b>		2,023	205

\*: Prescribed for at least 2% of patients within the case-control cohort. ENT: Ear, nose, and throat; OBS/GYN/UTI: Obstetrics, gynaecologic and urinary tract infection.

We assessed drug exposure over three non-overlapping time periods: Current (within 3 months before index date), recent (between 1 year before index date and 3 months before index date), and past (between 1 January 2000 and 1 year before index date). This method was applied to capture a comprehensive view of how drug exposure influences aSAH incidence across temporal contexts (Figure 2), revealing short-term and long-term effects that might not be apparent within a single time period. We iterated through each commonly prescribed drug and mapped the patient's most recent drug prescription date (i.e. closest to their index date) to one of these three windows. We considered a patient to have never used the drug if they did not obtain a prescription for the drug or if the prescription date occurred after their index date.



**Figure 1.** Study flow diagram.

aSAH: aneurysmal subarachnoid haemorrhage; SAIL: Secure Anonymised Information Linkage databank.

## Covariates

To adjust for confounders, we considered routinely available covariates that may be related to both aSAH and drug exposure. These included smoking status, hypertension, alcohol abuse, body mass index (BMI), and an overall comorbidity score. In addition, we included healthcare utilisation and socioeconomic status to address healthy user bias. We assessed healthcare utilisation by counting the number of visits to a general practitioner in the year preceding the index date. For socioeconomic status, we assessed the Welsh Index of Multiple Deprivation (WIMD) of a patient's area of residence. We used code lists to cluster Read codes in order to define smoking status (defined either current, former, or never), hypertension (based on a GP diagnosis), and alcohol abuse. Each of these variables was measured

as close to the index date as possible and kept constant throughout the analysis. For BMI, we retrieved the most recent measurement before the index date. We used multivariate imputation to approximate BMI values where BMI data was missing. Similarly, for missing smoking status, we assumed 'never'. Other categorical variables with missing values, such as hypertension, were imputed with a value of 0 to indicate the condition's absence. As a measure of overall comorbidity, we used Read-based diagnoses codes to define a modified version of the Elixhauser score,<sup>16</sup> in which already incorporated covariates (i.e., hypertension, alcohol abuse, and obesity) were excluded. All analyses were performed in Python.<sup>17</sup>

### Statistical analysis

We developed a binomial logistic regression model for each commonly prescribed drug with aSAH as the outcome. The model included drug exposure as a nominal variable with four levels (current, recent, past, never), with 'never' acting as the reference category. All potential confounders described above were included as covariates in the model. For smoking, 'never smoked' acted as the reference category. We calculated odds ratios (OR) with 95% confidence intervals (95% CI) for each commonly prescribed drug, and applied Bonferroni correction to account for the number of commonly prescribed drugs tested. Following reviewer comments, we conducted a post-hoc analysis in which we controlled for two additional covariates: a history of depression and a history of anxiety

## RESULTS

We identified 4,879 aSAH cases and matched them to 43,911 controls without aSAH. While our study sample was balanced at the index date in terms of sex (61.2% women in both cases and controls) and mean age (61.4, SD 15.4), aSAH cases visited their GP more frequently than controls (mean visits: 23 versus 19, respectively) (Table 2). In addition, more aSAH cases were current smokers (37% vs. 21%), or had a history of hypertension (42% vs. 37%) before the index date. Despite this, the mean modified Elixhauser score for cases was only marginally higher than for controls (1.8 vs 1.4).

Clustering resulted in 2,023 unique drugs, of which 205 (10.1%) were commonly prescribed (Table 1). A total of nine drugs had a statistically significant association with aSAH after Bonferroni correction (Figure 3, Table 3).

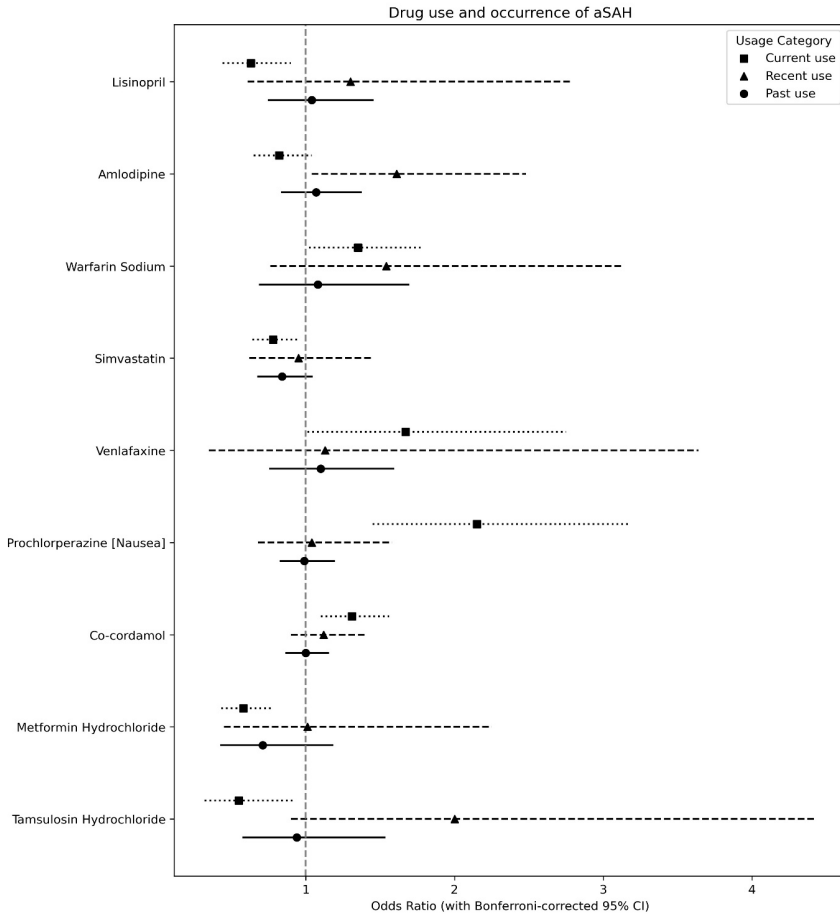
**Table 2.** Baseline characteristics of cases of aneurismal subarachnoid haemorrhage and controls in Wales between 2000-2019.

	<b>Cases</b>	<b>Controls</b>
<b>N</b>	<b>4,879</b>	<b>43,911</b>
Matching criteria		
Observation period in days, median (range)	3,554 (365 - 7,304)	3,550 (365 - 7,304)
Age*, mean (SD)	61.4 (15.4)	61.4 (15.4)
Female*, n (%)	2,988 (61.2)	26,892 (61.2)
Smoking status, n (%)		
Current	1,787 (37)	9,005 (21)
Former	978 (20)	9,511 (22)
Never	1,840 (38)	22,413 (51)
Other		
Hypertension, n (%)	2,030 (42)	16,335 (37)
Number of general practitioner consultations, mean (SD)	23.3 (20.0)	18.9 (17.4)
Deprivation index, mean (SD)	22.4 (15.7)	20.9 (15.1)
Alcohol abuse, n (%)	333 (7)	1,784 (4)
BMI, mean (SD)	26.9 (5.6)	27.6 (5.8)
Modified Elixhauser Score, mean (SD)	1.8 (1.7)	1.4 (1.6)

\* Matching criteria for nested-case control. Reported values at baseline (i.e. index date).  
SD = standard deviation, BMI = Body Mass Index.

We found a significant decrease in aSAH incidence for current use of lisinopril (OR 0.63, 95% CI 0.44 – 0.90, Figure 3) versus non-use, with a trend towards an increased risk in recent use (OR 1.30, 95% CI 0.61 – 2.78). We found a similar pattern for amlodipine, where there was a marginally nonsignificant decrease in aSAH incidence for current use (OR 0.82, 95% CI 0.65 – 1.04) versus non-use, with an increased risk in recent use (OR 1.61, 95% CI 1.04 - 2.48). Notably, we did not find this trend for other antihypertensives.

Current use of three additional drugs were associated with a decreased risk of aSAH. Specifically, we found a reduced incidence of aSAH in current users of simvastatin (OR 0.78, 95% CI 0.64 – 0.96), metformin (OR 0.58, 95% CI 0.43 – 0.78), and tamsulosin (OR 0.55, 95% CI 0.32 – 0.93). No statistically significant associations were found between these drugs and aSAH in recent or past users. Notably, other drugs within the classes of angiotensin-converting enzyme (ACE) inhibitors, statins, antidiabetics, and alpha-blockers did not show significant associations with aSAH.



**Figure 3.** Commonly prescribed drugs associated with aneurysmal subarachnoid haemorrhage (aSAH) incidence stratified by recency of use and in comparison with non-use of the same drug. CI: Confidence interval.

Conversely, current use of four other drugs were associated with an increased risk of aSAH. Specifically, we found an increased incidence of aSAH in current users of warfarin (OR 1.35 95% CI 1.02 - 1.79), venlafaxine (OR 1.67, 95% CI 1.01 - 2.75), prochlorperazine (OR 2.15 95% CI 1.45 - 3.18), and co-codamol (OR 1.31, 95% CI 1.10 - 1.56). No statistically significant associations were found between these drugs and aSAH in recent or past users. Other drugs within the drug classes of vitamin K antagonists, serotonin reuptake inhibitors (SRIs), conventional antipsychotics, and compound analgesics did not show an association with aSAH.

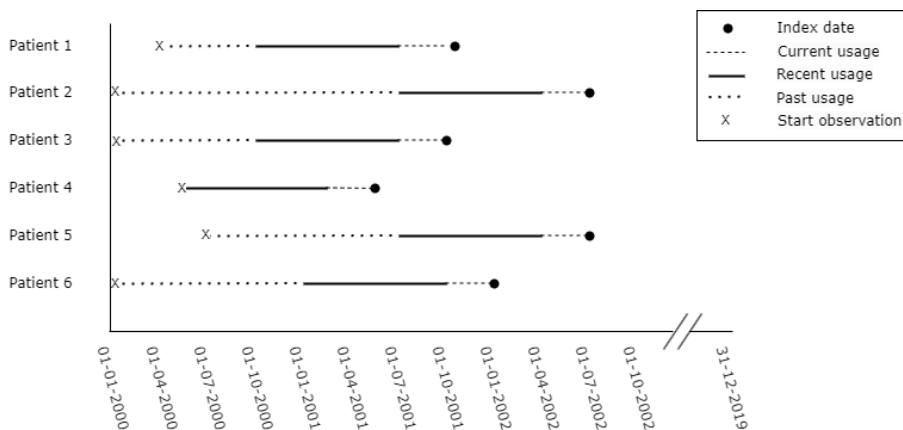
The post-hoc analyses in which we controlled for a history of depression and anxiety yielded no significant changes to the effects described above.

This study provides class III evidence that some commonly prescribed drugs are associated with subsequent development of aSAH.

**Table 3.** Drugs associated with a changed incidence of aneurysmal subarachnoid haemorrhage incidence (aSAH) using a binomial logistic regression model and Bonferroni corrected 95% confidence intervals for use versus non-use of the same drug.

Read code	Drug name	Current use OR (95% CI)	Recent use OR (95% CI)	Past use OR (95% CI)
<b>Decreased incidence</b>				
bi3	Lisinopril	0.63* (0.44 - 0.90)	1.30 (0.61 - 2.78)	1.04 (0.75 - 1.45)
bx4	Simvastatin	0.78* (0.64 - 0.96)	0.95 (0.62 - 1.45)	0.84 (0.68 - 1.04)
f41	Metformin	0.58* (0.43 - 0.78)	1.01 (0.45 - 2.25)	0.71 (0.43 - 1.18)
gc7	Tamsulosin	0.55* (0.32 - 0.93)	2.0 (0.90 - 4.43)	0.94 (0.58 - 1.53)
<b>Increased incidence</b>				
blb	Amlodipine	0.82 (0.65 - 1.04)	1.61* (1.04 - 2.48)	1.07 (0.84 - 1.37)
bs1	Warfarin	1.35* (1.02 - 1.79)	1.54 (0.76 - 3.13)	1.08 (0.69 - 1.69)
da7	Venlafaxine	1.67* (1.01 - 2.75)	1.13 (0.35 - 3.64)	1.1 (0.76 - 1.59)
dhe	Prochlorperazine [Nausea]	2.15* (1.45 - 3.18)	1.04 (0.68 - 1.58)	0.99 (0.83 - 1.19)
dia2	Co-cordamol	1.31* (1.10 - 1.56)	1.12 (0.90 - 1.4)	1.00 (0.87 - 1.15)

\* Statistically significant (p-value <0.05) after Bonferroni correction. Time windows are defined as current (within 3 months before index date), recent (between 1 year before index date and 3 months before index date), and past (between 1 January 2000 and 1 year before index date).



**Figure 2.** The definition of three non-overlapping time periods in order to capture the current, recent, and past effects of drug use on aneurysmal subarachnoid haemorrhage (aSAH) incidence. Time periods are defined as follows: current (within 3 months before index date), recent (between 1 year before index date and 3 months before index date), and past (between 1 January 2000 and 1 year before index date).

## DISCUSSION

In this DWAS we found nine drugs that were associated with aSAH incidence. We found a similar trend for lisinopril and amlodipine, with a decreased aSAH risk in current use, and an increased aSAH risk in recent use. Three additional drugs were associated with a decreased incidence of aSAH: simvastatin, metformin, and tamsulosin. In contrast, we found an increased incidence of aSAH in recent users of amlodipine and current users of warfarin, venlafaxine, prochlorperazine, and co-codamol.

We found a similar trend for lisinopril (an ACE inhibitor) and amlodipine (a calcium channel blocker), with increased aSAH incidence in recent use and decreased aSAH risk in current use (Figure 3). Other studies, while not unanimous,<sup>18</sup> generally found a protective effect of antihypertensive use on aSAH incidence and outcome.<sup>19-21</sup> However, the majority of these studies were cross-sectional and could therefore not distinguish between current and recent use. Our findings indicate a trend, with current antihypertensive use associated with a lower risk of aSAH (significant for lisinopril but not for amlodipine) and recent antihypertensive use associated with an increased risk of aSAH (significant for amlodipine but not lisinopril). An increased aSAH risk in recent of antihypertensive use may be due to confounding by indication, specifically caused by hypertension.<sup>22</sup> However, we corrected for hypertension in our analysis, and confounding by indication would not explain a protective effect in current use. Alternatively, the increased aSAH risk observed in recent use may be related to drug treatment discontinuation (e.g. resistant hypertension, side-effects, or contraindications). In contrast, a lower incidence in current use may be due to the antihypertensive's immediate effects, such as lowering blood pressure, or inhibiting the local renin-angiotensin system<sup>23</sup>. Why we found this trend for lisinopril and amlodipine, but not for other antihypertensives remains an open question for further research.

We found a reduced aSAH incidence for current users of simvastatin (a statin), and metformin (an antidiabetic). Statins have been associated with a decreased incidence of aSAH in general, although the evidence is not consistent.<sup>18,24,25</sup> A Dutch case-control study found that current statin use was associated with a lower risk of SAH (OR 0.77, 95% CI 0.55 - 1.07), while recent statin withdrawal was associated with an increased risk of SAH when compared to continued use (OR 2.34, 95% CI 1.35 - 4.05).<sup>18</sup> Simvastatin may reduce the risk of aSAH by improving endothelial functions, or by its anti-inflammatory effects on vascular walls.<sup>26</sup> Antidiabetics, similar to statins, have been shown to reduce the risk of aSAH.<sup>27</sup> Antidiabetics

may lower aSAH risk by reducing the vascular complications associated with hyperglycaemia (e.g. endothelial damage, cerebral tight junction protein expression).<sup>28,29</sup> Alternatively, our results with simvastatin and metformin could be explained by previously reported protective effects of hypercholesterolaemia and diabetes on aSAH.<sup>22</sup> However, we found no statistically significant results for other drugs that affect hypercholesterolaemia or diabetes, implying that the lower incidence of aSAH in current simvastatin and metformin users may be drug-specific.

Finally, we found a reduced incidence of aSAH in current tamsulosin users. Specifically, we found a trend similar to that found for lisinopril and amlodipine, with an increased risk in recent use and a decreased risk in current use. Alpha blockers have not been studied for their association with aSAH, although we can speculate that alpha blockers affect aSAH risk by altering blood pressure or local vasodilation.<sup>30</sup> It is currently unclear why we found an increased aSAH incidence in recent tamsulosin users, given that indications for tamsulosin prescription (e.g., prostatic hyperplasia, chronic prostatitis) are not currently known risk factors for aSAH. Future studies should further investigate these signals.

We found an increased risk of aSAH in current users of warfarin, an anticoagulant that functions by antagonising vitamin K. Although the relationship between aSAH and warfarin has not been studied specifically, vitamin K antagonists in general have been linked to an increased risk of aSAH.<sup>31,32</sup> As warfarin is known to elevate bleeding risk,<sup>33</sup> it may consequently heighten the risk of aSAH. It is unknown whether warfarin causes an aneurysm to burst directly, or indirectly by inducing tearing in the aneurysm wall. In this DWAS, detecting an increased risk associated with current use of warfarin can serve as a 'positive control', validating our methods and lending credibility to our other findings.

We additionally found an increased risk of aSAH in current users of venlafaxine (a serotonin and norepinephrine reuptake inhibitor, SNRI) and prochlorperazine (a conventional antipsychotic). SNRIs have, to the best of our knowledge, not been studied for their association with aSAH. SNRIs work by increasing noradrenergic activity,<sup>3</sup> which can raise blood pressure and heart rate,<sup>34,35</sup> potentially increasing the risk of aSAH. In addition, selective serotonin reuptake inhibitors (SSRIs), which are functionally similar to SNRIs, are known to increase the risk of bleeding.<sup>36</sup> Prochlorperazine's effects on aSAH have not been studied before and require further investigation. A recently proposed relationship between psychiatric diseases and aSAH could potentially explain our findings for both venlafaxine and prochlorperazine.<sup>37</sup> However, such residual confounding would not explain why

we were unable to find an association for aSAH and other antidepressants and antipsychotics. Furthermore, we found that the association between aSAH and both venlafaxine and prochlorperazine remained when controlling for a history of depression and anxiety. This suggests that the observed relationships were likely due to the drugs themselves rather than underlying psychiatric diseases.

Finally, we found an increased risk of aSAH in current users of co-codamol, a compound analgesic consisting of codeine and paracetamol. We were unable to investigate the individual effects of each active substance due to limitations in our clustering of read codes. Thus, the observed association could be attributed to the use of either of these medications or a synergistic effect of their concomitant use. The relationship between aSAH and co-codamol has not been studied before, and their mechanisms (e.g., blood pressure changes, vasodilation) require further investigation.

This study had several strengths. First, identifying nearly 5,000 aSAH cases from a large nationwide dataset, and focusing solely on commonly prescribed drugs, increased our statistical power to detect even relatively subtle effects. Second, we were able to differentiate acute from chronic effects by examining recency of use with non-overlapping windows. Finally, we mitigated healthy-user bias and confounding by adjusting for healthcare utilisation and a variety of characteristics related to overall health and aSAH. Although we used records from an EHD, where data was not primarily recorded for research use, the SAIL databank has a reputation for good coverage and linkage.<sup>38</sup> For instance, only 6% of individuals in the SAIL databank lacked smoking status data.

This study also had several limitations. First, observed findings may be due to associations between the indications and aSAH (i.e., confounding by indication). For example, a lower incidence of aSAH in metformin users may simply reflect a lower incidence of aSAH in diabetics. However, we aimed to reduce confounding by indication by controlling for known aSAH risk factors. Second, we assumed that a drug prescription translates to proper drug use. In reality, however, a patient may not take their drugs or use them incorrectly, which results in exposure misclassification. Thus, our results could be skewed if the discrepancy between prescribed and actual drug use differs between cases and controls. Third, by clustering the drugs investigated in our study, we lost the specificity provided by a Read code. As a result, we were unable to study dose-response relationships, but were able to study drug-specific (rather than class-specific) effects. Fourth, at this stage we were unable to study duration of use of each medication and differentiate

between first-time and long-term users. In addition, drug exposure windows were defined identically for each drug, ignoring drug-specific duration of effects. We thus may have missed differences between acute and long-term effects of drugs on aSAH. Another limitation relates to the definition of aSAH. We chose to include ICD-10 codes I60.8 ('Other nontraumatic subarachnoid haemorrhage') and I60.9 ('Nontraumatic subarachnoid haemorrhage, unspecified') because misclassification can occur when using ICD codes in general,<sup>39</sup> and for aSAH specifically.<sup>40</sup> By including these codes we may have included non-aneurysmal subarachnoid haemorrhage cases, potentially introducing outcome misclassification. However, non-aneurysmal subarachnoid haemorrhage cases are relatively rare,<sup>41</sup> and we decided that the potential cost to specificity was worth the increase in sample size and sensitivity, which was in line with our primary objective of hypothesis generation.

In conclusion, our results suggest that nine drugs may be associated with aSAH, where current use of lisinopril, simvastatin, metformin, tamsulosin, and potentially amlodipine showed a decreased risk. Using these signals as a starting point, future research should use a more hypothesis-driven approach to further investigate these associations and differentiate between drug class and specific drug substance effects. In addition, this research may help identify additional risk factors for aSAH, potentially leading to new pharmacologic therapy options for aneurysmal management.

## REFERENCES

1. Macdonald RL, Schweizer TA. Spontaneous subarachnoid haemorrhage. *Lancet*. 2017;389(10069):655-666.
2. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol*. 2021;20(10):795-820.
3. Johnston SC, Selvin S, Gress DR. The burden, trends, and demographics of mortality from subarachnoid hemorrhage. *Neurology*. 1998;50(5):1413-1418.
4. Connolly ES Jr, et al. Guidelines for the Management of Aneurysmal Subarachnoid Hemorrhage. *Stroke*. 2012;43(6):1711-1737.
5. Algra AM, et al. Development of the SAFETEA Scores for Predicting Risks of Complications of Preventive Endovascular or Microneurosurgical Intracranial Aneurysm Occlusion. *Neurology*. 2022;99(18):1725-1737.
6. Algra AM, et al. Procedural Clinical Complications, Case-Fatality Risks, and Risk Factors in Endovascular and Neurosurgical Treatment of Unruptured Intracranial Aneurysms: A Systematic Review and Meta-analysis. *JAMA Neurol*. 2019;76(3):282-293.
7. Etminan N, Rinkel GJ. Unruptured intracranial aneurysms: development, rupture and preventive management. *Nat Rev Neurol*. 2016;12(12):699-713.
8. Ryan P, Madigan D, Stang P, Schuemie M, Hripcsak G. Medication-Wide Association Studies. *CPT Pharmacometrics Syst Pharmacol*. 2013;2:e76.
9. Pottegård A, et al. Identification of Associations Between Prescribed Medications and Cancer: A Nationwide Screening Study. *EBioMedicine*. 2016;7:73-79.
10. Patel CJ, Ji J, Sundquist J, Ioannidis JP, Sundquist K. Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study. *Sci Rep*. 2016;6:31308.
11. Wilkinson T, et al. Drug prescriptions and dementia incidence: a medication-wide association study of 17000 dementia cases among half a million participants. *J Epidemiol Community Health*. 2021;76(3):223-229.
12. Cheng Y, et al. Medication-Wide Association Study Plus (MWAS+): A Proof of Concept Study on Drug Repurposing. *Med Sci*. 2022;10(1):48.
13. Bejan CA, et al. DrugWAS: Drug-wide Association Studies for COVID-19 Drug Repurposing. *Clin Pharmacol Ther*. 2021;110(6):1537-1546.
14. Pushpakom S, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov*. 2019;18(1):41-58.
15. Ford DV, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res*. 2009;9:157.
16. Metcalfe D, et al. Coding algorithms for defining Charlson and Elixhauser co-morbidities in Read-coded databases. *BMC Med Res Methodol*. 2019;19:115.
17. The Python Language Reference. *Python documentation* <https://docs.python.org/3/reference/index.html>.
18. Risselada R, et al. Withdrawal of Statins and Risk of Subarachnoid Hemorrhage. *Stroke*. 2009;40(9):2887-2892.

19. Shimizu K, et al. Candidate drugs for preventive treatment of unruptured intracranial aneurysms: A cross-sectional study. *PLOS ONE*. 2021;16(2):e0246865.
20. Pickard JD, et al. Effect of oral nimodipine on cerebral infarction and outcome after subarachnoid haemorrhage: British aneurysm nimodipine trial. *BMJ*. 1989;298(6674):636-642.
21. Hoh BL, et al. 2023 Guideline for the Management of Patients With Aneurysmal Subarachnoid Hemorrhage: A Guideline From the American Heart Association/American Stroke Association. *Stroke*. 2023;54(3):e314-e370.
22. Feigin VL, et al. Risk factors for subarachnoid hemorrhage: an updated systematic review of epidemiological studies. *Stroke*. 2005;36(12):2773-2780.
23. Tada Y, et al. Roles of hypertension in the rupture of intracranial aneurysms. *Stroke*. 2014;45(2):579-586.
24. Yoshimura Y, et al. Statin Use and Risk of Cerebral Aneurysm Rupture: A Hospital-based Case-control Study in Japan. *J Stroke Cerebrovasc Dis*. 2014;23(2):343-348.
25. Hostettler IC, et al. Characteristics of Unruptured Compared to Ruptured Intracranial Aneurysms: A Multicenter Case-Control Study. *Neurosurgery*. 2018;83(1):43.
26. Wang C-Y, Liu P-Y, Liao JK. Pleiotropic effects of statin therapy. *Trends Mol Med*. 2008;14(1):37-44.
27. Can A, et al. Antihyperglycemic Agents Are Inversely Associated With Intracranial Aneurysm Rupture. *Stroke*. 2018;49(1):34-39.
28. Li W, et al. Adaptive cerebral neovascularization in a model of type 2 diabetes: relevance to focal cerebral ischemia. *Diabetes*. 2010;59(1):228-235.
29. Tamura T, et al. Endothelial damage due to impaired nitric oxide bioavailability triggers cerebral aneurysm formation in female rats. *J Hypertens*. 2009;27(6):1284-1292.
30. Frishman WH, Charlap S. Alpha-adrenergic blockers. *Med Clin North Am*. 1988;72(2):427-440.
31. Shimizu K, et al. Associations Between Drug Treatments and the Risk of Aneurysmal Subarachnoid Hemorrhage: a Systematic Review and Meta-analysis. *Transl Stroke Res*. 2022;13(1):59-67.
32. Risselada R, et al. Platelet aggregation inhibitors, vitamin K antagonists and risk of subarachnoid hemorrhage. *J Thromb Haemost*. 2011;9(3):517-523.
33. White RH, et al. Management and prognosis of life-threatening bleeding during warfarin therapy. National Consortium of Anticoagulation Clinics. *Arch Intern Med*. 1996;156(11):1197-1201.
34. Cashman JR, Ghirmai S. Inhibition of serotonin and norepinephrine reuptake and inhibition of phosphodiesterase by multi-target inhibitors as potential agents for depression. *Bioorg Med Chem*. 2009;17(21):6890-6897.
35. Lexicomp. Venlafaxine: Drug information. *UpToDate*. Retrieved December 1, 2023, from <https://www.uptodate.com/contents/venlafaxine-drug-information>.
36. Edinoff AN, et al. Selective Serotonin Reuptake Inhibitors and Associated Bleeding Risks: A Narrative and Clinical Review. *Health Psychol Res*. 2022;10:4.
37. Cooke DL, et al. Association of select psychiatric disorders with incident brain aneurysm and subarachnoid hemorrhage among veterans. *Front Integr Neurosci*. 2023;17:1207610.
38. Lyons RA, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak*. 2009;9:3.
39. Horsky J, Drucker EA, Ramelson HZ. Accuracy and completeness of clinical coding using ICD-10 for ambulatory visits. In *AMIA Annu Symp Proc*. 2017;2017:912.
40. Roark C, et al. Assessing the utility and accuracy of ICD10-CM non-traumatic subarachnoid hemorrhage codes for intracranial aneurysm research. *Learning Health Syst*. 2021;5(4):e10288.
41. Macdonald RL, Schweizer TA. Spontaneous subarachnoid haemorrhage. *Lancet*. 2017;389(10069):10069.





## Chapter 8

# Associations between lisinopril use and risk of aneurysmal subarachnoid haemorrhage: A UK population-based cohort study

---

Jos P. Kanning, Patrick C. Souverein, Olaf H. Klungel, Mirjam I. Geerlings,  
Ynte M. Ruigrok, Shahab Abtahi

In preparation

## ABSTRACT

Aneurysmal subarachnoid haemorrhage (aSAH) is a stroke caused by the rupture of an intracranial aneurysm. Invasive treatments to prevent aSAH often pose higher risks than benefits, highlighting the need for a non-invasive drug alternative. Previous research suggests that lisinopril, an angiotensin-converting-enzyme (ACE) inhibitor, may be more effective at lowering aSAH risk than other ACE inhibitors. This study uses an active comparator new user design, using a large pharmacoepidemiologic dataset from the UK Clinical Practice Research Datalink (CPRD) GOLD database, to compare lisinopril initiators with those starting other ACE inhibitors. Among 975,316 hypertensive patients, we identified 108,717 lisinopril initiators and 276,113 initiators of other ACE inhibitors. The primary analysis followed a per-protocol approach, tracking patients from the index date until aSAH onset, end of treatment, death, database transfer, last CPRD data collection date, or 31 December 2022. Incident aSAH cases were identified through primary care records, excluding those diagnosed before the index date. Propensity score matching controlled for confounding variables, and Cox proportional hazard models calculated hazard ratios (HRs) and 95% confidence intervals (CIs) for aSAH risk. During median follow-ups of 1,044 days for lisinopril users and 1,093 days for other ACE inhibitor users, we identified 75 and 107 aSAH cases respectively, with incidence rates of 0.161 and 0.229 per 1,000 person-years. Lisinopril users had a significantly lower aSAH risk than other ACE inhibitor users (HR 0.71; 95% CI, 0.53-0.96). Kaplan-Meier and cumulative incidence plots indicated early differences in aSAH incidence. In conclusion, lisinopril is more effective at reducing aSAH risk than other ACE inhibitors, warranting further research to understand its protective mechanisms.

## INTRODUCTION

Aneurysmal subarachnoid haemorrhage (aSAH) is a type of stroke caused by the rupture of an intracranial aneurysm.<sup>1</sup> Although aSAH constitutes only 9.7% of strokes worldwide,<sup>2</sup> it represents 27.3% of all stroke-related years of potential life lost before age 65 due to its early onset and high mortality rate.<sup>3</sup> aSAH can be prevented once an intracranial aneurysm has been identified, typically through endovascular methods or surgery.<sup>4</sup> However, physicians often hesitate to recommend surgical interventions because of the inherent risks and the often low probability of rupture.<sup>5,6</sup> As a result, most aneurysms currently remain untreated.<sup>7</sup>

A hypothetical drug therapy that could safely and effectively reduce the risk of aneurysm rupture would provide a much-needed alternative, allowing clinicians to mitigate rupture risk in a minimally invasive manner. Antihypertensives are a promising drug class for lowering aneurysmal rupture risk by controlling hypertension, a known aSAH risk factor<sup>8,9</sup>. However, it is unclear if the beneficial effects of these medications on reducing aSAH risk are solely due to their hypertension-lowering properties or if there might be an additional drug-specific effect.

A recent drug-wide association study found a decreased aSAH incidence among current users of lisinopril, an angiotensin-converting enzyme (ACE) inhibitor, but not for other ACE inhibitors.<sup>10</sup> These findings suggest that lisinopril may have specific protective effects against aSAH alongside its antihypertensive effects. This initial signal requires further epidemiologic, molecular, and genetic investigation.

In this study, we investigated the association of lisinopril use with aSAH incidence among hypertensive patients in a large real-world pharmacoepidemiologic dataset from the United Kingdom (UK). Specifically, we used an active-comparator new-user design to determine whether lisinopril use decreased the risk of aSAH compared to other ACE inhibitors.

## METHODS

### Data source

We used data from the UK Clinical Practice Research Datalink (CPRD) GOLD, linked to the Hospital Episodes Statistics (HES) Admitted Patient Care (HES APC). The CPRD-GOLD database contains electronic medical records for 21.4 million people from 984 general practitioner practices in the UK. As of December 2023, CPRD included

data on around 3 million currently enrolled patients, representing 4.4% of the UK population and is generalisable to the entire UK population.<sup>11</sup> CPRD data includes patients' demographics, medical history, laboratory test results, prescription details, specialist referrals, hospital admissions, and primary outcomes, encoded using Read codes. HES statistics cover admission and discharge information for all inpatient hospital admissions in England and Wales from 1997, with diagnoses encoded using ICD-10 codes. We also included Lower Layer Super Output Areas (LSOA) data to capture patient socioeconomic status.

### **Study population**

This retrospective cohort study included all adults in CPRD practices linked to HES data with a hypertension diagnosis in CPRD-GOLD between 1 August 2004 and 31 December 2022. We used the active comparator new user design to define a cohort of lisinopril initiators and initiators of any other ACE inhibitor. This active comparator new user design, frequently used in pharmacoepidemiological studies, aims to minimise confounding and selection bias by exclusively comparing outcomes between new users of the studied drug (the exposure) and new users of a comparable treatment (the active comparator).<sup>12</sup>

We defined the baseline date as the latest of the following dates: date of hypertension diagnosis, current general practitioner registration date, date at which the general practitioner data was deemed to be of research quality, or 1 August 2004. The first prescription of an exposure or an active comparator after the baseline date was defined as the index date (start of follow-up). We included only individuals without prior prescriptions for the exposure or its active comparator before their baseline date. Individuals who simultaneously initiated an exposure and an active comparator were excluded. Patients could enter each cohort only once, and those with less than a year of continuous enrolment before the index date were excluded.

### **Exposure and outcome**

The study examined the effects of lisinopril, comparing it with other ACE inhibitors—benazepril, captopril, cilazapril, enalapril, fosinopril, imidapril, moexipril, perindopril, quinapril, ramipril, and trandolapril—as active comparators. The primary analysis followed a per-protocol approach, tracking patients from the index date until the first occurrence of any of the following events: end of treatment, the onset of aSAH, death, transfer out of the database, the last data collection date of the CPRD practice, or 31 December 2022. The end of treatment was defined as the final prescription date plus the duration of the prescribed course, extended by

a 30-day grace period. Incident aSAH cases were identified through primary care records, and individuals diagnosed with aSAH before the index date were excluded.

### **Propensity score matching**

We used propensity score matching to control for potential confounding. The propensity score was the predicted probability of initiating an exposure versus an active comparator, based on covariates associated with aSAH and antihypertensive prescription. These covariates included age, sex, BMI, smoking status, alcohol abuse, blood pressure, diabetes mellitus, hypercholesterolemia, socioeconomic status, ethnicity, and several medical histories (Supplementary Table 1).<sup>9</sup>

### **Statistical analysis**

Baseline population characteristics were summarised using descriptive analysis. Continuous variables were reported as mean and standard deviation (SD), and categorical variables as counts and percentages.

The approach to missing data varied by variable type. For categorical variables, such as smoking status, missing data were categorised into a separate 'missing' category and included in the propensity score model. Numerical variables with missing values were imputed using multiple imputation, employing five iterations and predictive mean matching.<sup>13</sup>

All reported results are derived from a single, randomly selected imputed dataset. Multivariable logistic regression was used to compute the propensity score. We matched one lisinopril initiator with one active comparator initiator using the nearest neighbour algorithm, incorporating a calliper of 0.02. Standardised mean differences were used to assess balance post-matching, and individuals with propensity scores outside the 2.5th and 97.5th percentiles were excluded based on the asymmetrical trimming principle.<sup>14</sup>

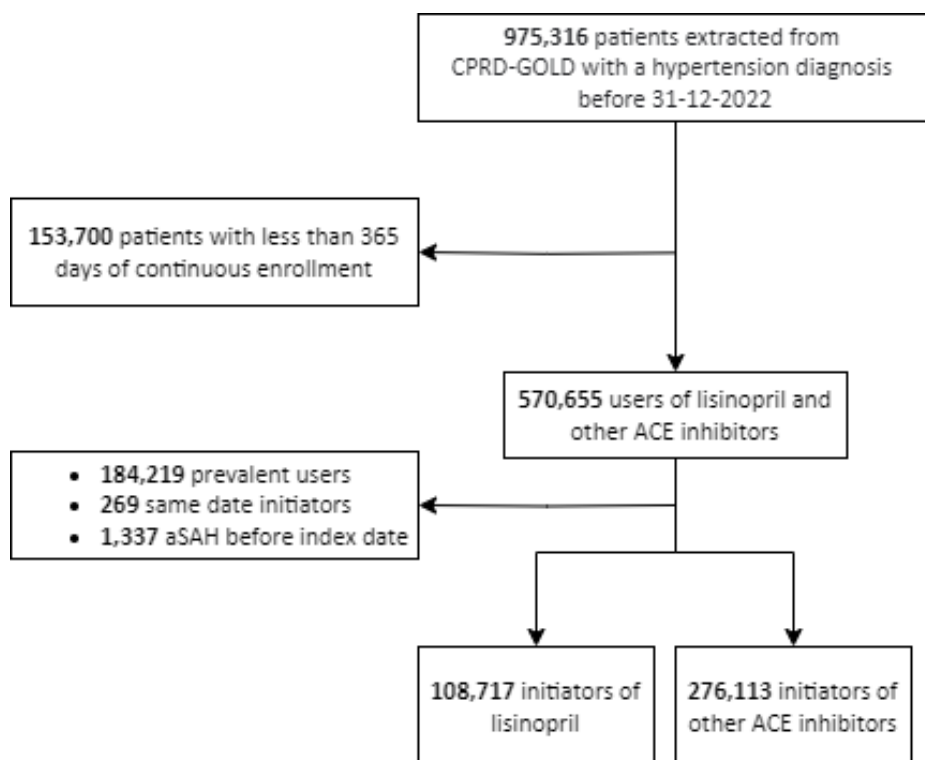
We estimated the incidence of aSAH in the matched group using Cox proportional hazards models. These models calculated hazard ratios (HRs) and 95% confidence intervals (CIs) for the risk of aSAH associated with lisinopril use compared to other ACE inhibitors. Additionally, we reported incidence rates, Kaplan-Meier plots, and cumulative incidence plots to track the incidence of aSAH over time.

### **Reporting standards**

All results are reported according to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.<sup>15</sup>

## RESULTS

We identified 975,316 patients with hypertension from the CPRD-GOLD dataset (Figure 1). After excluding 153,700 patients with fewer than 365 days of continuous enrolment, we retained 570,655 ACE inhibitor users. After excluding prevalent users, users who initiated an exposure and an active comparator simultaneously, and patients with an aSAH diagnosis before their index date, we were left with 108,717 users of lisinopril and 276,113 users of other ACE inhibitors.



**Figure 1.** Attrition flowchart showing the number of patients included and excluded at every step. CPRD = Clinical Practice Research Datalink, ACE = angiotensin-converting enzyme, aSAH = aneurysmal subarachnoid haemorrhage.

Lisinopril users were more likely to have a history of alcohol abuse, less likely to have diabetes, more likely to be smokers, more likely to be female, and slightly younger compared to users of other ACE inhibitors (Table 1). Propensity score matching confirmed a good balance between the two groups, with standardised mean differences for all covariates below 0.02 (Supplementary Table 1).

**Table 1.** Baseline characteristics for the angiotensin-converting enzyme (ACE) inhibitor cohort before and after propensity score matching.

Variable	Lisinopril (Pre-matching)	Other ACE inhibitor (Pre-matching)	SMD (Pre-matching)	Lisinopril (Post-matching)	Other ACE inhibitor (Post-matching)	SMD (Post-matching)
n (%)	108,717	276,113		104,126	104,126	
aSAH, n (%)	81 (0.1)	268 (0.1)		75 (0.1)	107 (0.1)	
Follow-up time in days, median (range)	1,040 (1-6,725)	1,083 (1-6,725)		1,042 (1-6,725)	1,032 (1-6,725)	
Age, mean (SD)	59.6 (12.8)	60.4 (13.3)	<b>0.06*</b>	59.6 (12.8)	59.6 (12.9)	<0.01
Male, n (%)	55,296 (50.9)	142,189 (51.5)	<b>0.01*</b>	52,434 (50.4)	52,632 (50.5)	<0.01
Smoking Status, n (%)			<b>0.05*</b>			<0.01
Current	23,795 (21.9)	56,748 (20.6)		21,835 (21.0)	21,530 (20.7)	
Former	29,278 (26.9)	76,171 (27.6)		28,062 (27.0)	28,065 (27.0)	
Never	54,664 (50.3)	141,413 (51.2)		54,063 (51.9)	54,379 (52.2)	
Unknown	980 (0.9)	1,781 (0.6)		166 (0.2)	152 (0.1)	
Alcohol Abuse, n (%)	7,856 (7.2)	16,010 (5.8)	<b>0.06*</b>	5,067 (4.9)	4,848 (4.7)	<b>0.01*</b>
BMI, mean (SD)	29.6 (6.4)	29.6 (6.6)	<0.01	29.6 (6.4)	31.4 (7.7)	<0.01
Diabetes, n (%)	10,650 (9.8)	31,078 (11.3)	<b>0.05*</b>	10,412 (10.0)	10,248 (9.8)	<0.01

\*, Standardized means differ significantly at  $p < 0.05$  (indicated in bold). SMD = standardised mean differences, ACE = angiotensin-converting enzyme, BMI = Body Mass Index.

During a median follow-up of 1,044 days for lisinopril users and 1,093 days for other ACE inhibitor users, we identified 75 and 107 aSAH cases respectively, corresponding to incidence rates of 0.161 and 0.229 cases per 1,000 person-years (Figure 1). The Cox proportional hazards model showed that lisinopril users had a statistically significantly lower risk of aSAH than other ACE inhibitor users, with a hazard ratio (HR) of 0.71 (95% CI, 0.53 - 0.96). Kaplan-Meier curves and cumulative incidence plots suggested that differences in aSAH incidence emerged shortly after initiating lisinopril or other ACE inhibitors (Supplementary Figure 1).

## DISCUSSION

In this study, we investigated the effects of lisinopril use on the incidence of aSAH among hypertensive patients using real-world data from the UK. We discovered that lisinopril users showed a statistically significant 29% lower risk of aSAH compared to users of other ACE inhibitors, matched by propensity score. Additionally, Kaplan-Meier and cumulative incidence plots for lisinopril and other ACE inhibitors diverge shortly after the initial prescription.

Our findings both support and expand upon previous research. Although few studies have explored the association between antihypertensive use and aSAH incidence and outcomes, existing evidence suggests a beneficial impact.<sup>4,16,17</sup> However, these studies often group all ACE inhibitors, overlooking differences between individual drugs. In a previous drug-wide association study (DWAS) using a different database and a case-control design, we found a similar 37% reduction in aSAH risk among current lisinopril users.<sup>10</sup> This study, combined with our previous DWAS results, underscores lisinopril's efficacy in reducing aSAH risk compared to other ACE inhibitors.

Several factors might explain the observed reduced risk of aSAH with lisinopril. First, lisinopril has a longer half-life than other ACE inhibitors and does not require hepatic activation, providing more stable blood pressure control.<sup>18</sup> Lisinopril uniquely binds to the N- and C-terminals of the somatic ACE domains, showing greater inhibitory potency for the C domain of ACE.<sup>19,20</sup> Unlike most other ACE inhibitors that are prodrugs converted into potent diacid metabolites, lisinopril resembles enalaprilat with effective oral bioavailability.<sup>21,22</sup> It is the only ACE inhibitor that bypasses hepatic metabolism, making it suitable for patients with severe liver disease, as it is predominantly eliminated by the kidneys.<sup>24</sup> Lisinopril may also be more effective at crossing the blood-brain barrier than other ACE inhibitors.<sup>25</sup> Finally, there may be other unknown mechanisms enhancing lisinopril's efficacy.

This study had several strengths. First, we used a large dataset, allowing the identification of sufficient lisinopril users with adequate follow-up time to detect a relatively large number of cases of the rare aSAH outcome. Second, the CPRD dataset has relatively few missing values for important lifestyle parameters (e.g., smoking, and BMI), which was a big advantage when using data from an electronic medical record database. Third, to address potential confounding by indication, the analysis was limited to patients recently diagnosed with hypertension and included only new users of lisinopril and other ACE inhibitors. Additionally, propensity score matching was employed, balancing the exposure groups on a large set of potential confounders.

This study also had several limitations. It assumes patient adherence to prescribed medications without data on actual medication dispensation and use. Despite using propensity score matching and the active comparator new user design, residual confounding could still influence results due to unmeasured variables such as diet and physical activity. Finally, there is a risk of outcome misclassification, although it is unlikely to differ between lisinopril users and other ACE inhibitor users.

In conclusion, we found that lisinopril is associated with a 29% reduced risk of aSAH as compared to other ACE inhibitors among patients with hypertension. Future research, including multi-database observational studies, genetic epidemiologic studies and eventually randomised controlled trials is needed to investigate how lisinopril acts differently on aSAH pathogenesis compared to other ACE inhibitors. In the future, lisinopril may be considered more often for hypertensive patients presenting with an intracranial aneurysm.

## REFERENCES

1. Macdonald RL, Schweizer TA. Spontaneous subarachnoid haemorrhage. *The Lancet*. 2017;389:655–666.
2. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol*. 2021;20:795–820.
3. Johnston SC, Selvin S, Gress DR. The burden, trends, and demographics of mortality from subarachnoid hemorrhage. *Neurology*. 1998;50:1413–1418.
4. Connolly ES, Rabinstein AA, Carhuapoma JR, Derdeyn CP, Dion J, Higashida RT, Hoh BL, Kirkness CJ, Naidech AM, Ogilvy CS, et al. Guidelines for the Management of Aneurysmal Subarachnoid Hemorrhage. *Stroke*. 2012;43:1711–1737.
5. Algra AM, Greving JP, de Winkel J, Kurtelius A, Laban K, Verbaan D, van den Berg R, Vandertop W, Lindgren A, Krings T, et al. Development of the SAFETEA Scores for Predicting Risks of Complications of Preventive Endovascular or Microneurosurgical Intracranial Aneurysm Occlusion. *Neurology*. 2022;10.1212/WNL.000000000200978.
6. Algra AM, Lindgren A, Vergouwen MDI, Greving JP, van der Schaaf IC, van Doormaal TPC, Rinkel GJE. Procedural Clinical Complications, Case-Fatality Risks, and Risk Factors in Endovascular and Neurosurgical Treatment of Unruptured Intracranial Aneurysms: A Systematic Review and Meta-analysis. *JAMA Neurol*. 2019;76:282–293.
7. Etminan N, Rinkel GJ. Unruptured intracranial aneurysms: development, rupture and preventive management. *Nat Rev Neurol*. 2016;12:699–713.
8. Shimizu K, Aoki T, Etminan N, Hackenberg KAM, Tani S, Imamura H, Kataoka H, Sakai N. Associations Between Drug Treatments and the Risk of Aneurysmal Subarachnoid Hemorrhage: a Systematic Review and Meta-analysis. *Transl Stroke Res*. 2022;
9. Feigin VL, Rinkel GJE, Lawes CMM, Algra A, Bennett DA, van Gijn J, Anderson CS. Risk factors for subarachnoid hemorrhage: an updated systematic review of epidemiological studies. *Stroke*. 2005;36:2773–2780.
10. Kanning JP, Abtahi S, Klungel OH, Geerlings MI, Ruigrok YM. Prescribed Drug Use and Aneurysmal Subarachnoid Haemorrhage Incidence: A Drug-Wide Association Study. *Neurology* (in press). 2024;
11. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *International Journal of Epidemiology*. 2015;44:827–836.
12. Lund JL, Richardson DB, Stürmer T. The Active Comparator, New User Study Design in Pharmacoepidemiology: Historical Foundations and Contemporary Application. *Curr Epidemiol Rep*. 2015;2:221–228.
13. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20:40–49.
14. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28:3083–3107.
15. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335:806–808.
16. Shimizu K, Imamura H, Tani S, Adachi H, Sakai C, Ishii A, Kataoka H, Miyamoto S, Aoki T, Sakai N. Candidate drugs for preventive treatment of unruptured intracranial aneurysms: A cross-sectional study. *PLOS ONE*. 2021;16:e0246865.

17. Pickard JD, Murray GD, Illingworth R, Shaw MD, Teasdale GM, Foy PM, Humphrey PR, Lang DA, Nelson R, Richards P. Effect of oral nimodipine on cerebral infarction and outcome after subarachnoid haemorrhage: British aneurysm nimodipine trial. *BMJ*. 1989;298:636–642.
18. Olvera Lopez E, Parmar M, Pendela VS, Terrell JM. Lisinopril [Internet]. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2024 [cited 2024 May 24]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK482230/>
19. Fernandez JH, Hayashi MAF, Camargo ACM, Neshich G. Structural basis of the lisinopril-binding specificity in N- and C-domains of human somatic ACE. *Biochem Biophys Res Commun*. 2003;308:219–226.
20. Wei L, Clauser E, Alhenc-Gelas F, Corvol P. The two homologous domains of human angiotensin I-converting enzyme interact differently with competitive inhibitors. *J Biol Chem*. 1992;267:13398–13405.
21. Kelly JG, O'Malley K. Clinical pharmacokinetics of the newer ACE inhibitors. A review. *Clin Pharmacokinet*. 1990;19:177–196.
22. Natesh R, Schwager SLU, Sturrock ED, Acharya KR. Crystal structure of the human angiotensin-converting enzyme–lisinopril complex. *Nature*. 2003;421:551–554.
23. Piepho RW. Overview of the angiotensin-converting-enzyme inhibitors. *Am J Health Syst Pharm*. 2000;57 Suppl 1:S3-7.
24. White CM. Pharmacologic, pharmacokinetic, and therapeutic differences among ACE inhibitors. *Pharmacotherapy*. 1998;18:588–599.
25. Ouk M, Wu C-Y, Rabin JS, Jackson A, Edwards JD, Ramirez J, Masellis M, Swartz RH, Herrmann N, Lanctôt KL, et al. The use of angiotensin-converting enzyme inhibitors vs. angiotensin receptor blockers and cognitive decline in Alzheimer's disease: the importance of blood-brain barrier penetration and APOE ε4 carrier status. *Alz Res Therapy*. 2021;13:43.

## SUPPLEMENTARY MATERIALS

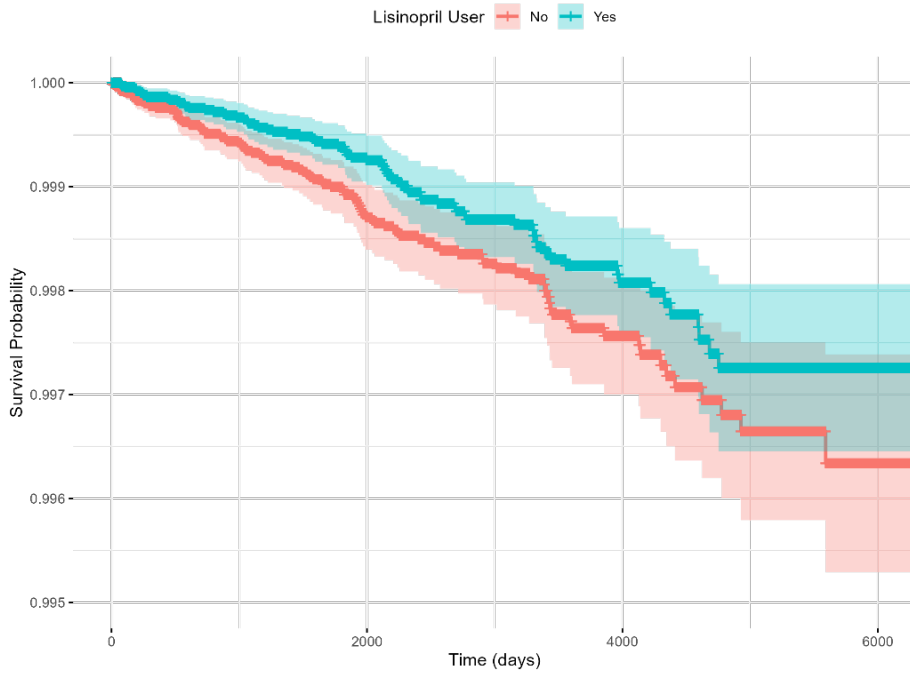
**Supplementary Table 1.** Baseline characteristics for the angiotensin-converting enzyme (ACE) inhibitor cohort before and after propensity score matching.

Variable	Lisinopril (Pre-matching)	Other ACE inhibitor (Pre-matching)	SMD (Pre-matching)	Lisinopril (Post-matching)	Other ACE inhibitor (Post-matching)	SMD (Post-matching)
n (%)	108,717	276,113		104,126	104,126	
Acute Renal Failure, n (%)	416 (0.4)	1,244 (0.5)	<b>0.01*</b>	391 (0.4)	339 (0.3)	<0.01
Anal Fissures, n (%)	1,712 (1.6)	5,287 (1.9)	<b>0.03*</b>	1,686 (1.6)	1,629 (1.6)	<0.01
Angina Pectoris, n (%)	2,475 (2.3)	9,008 (3.3)	<b>0.06*</b>	2,110 (2.0)	2,054 (2.0)	<0.01
Anxiety Disorders, n (%)	15,526 (14.3)	41,406 (15.0)	<b>0.02*</b>	15,053 (14.5)	14,818 (14.2)	<0.01
Arrhythmia, n (%)	3,626 (3.3)	13,961 (5.1)	<b>0.09*</b>	3,392 (3.3)	3,410 (3.3)	<0.01
Ascites, n (%)	63 (0.1)	160 (0.1)	<0.001	54 (0.1)	46 (0.0)	<0.01
Cancer, n (%)	9,619 (8.8)	26,755 (9.7)	<b>0.03*</b>	9,335 (9.0)	9,220 (8.9)	<0.01
Chronic Kidney Disease, n (%)	6,569 (6.0)	19,990 (7.2)	<b>0.05*</b>	6,355 (6.1)	6,167 (5.9)	<0.01
Cirrhosis, n (%)	202 (0.2)	533 (0.2)	<0.01	179 (0.2)	159 (0.2)	<0.01
Conduction Disorders, n (%)	365 (0.3)	1,148 (0.4)	<b>0.01*</b>	350 (0.3)	288 (0.3)	<b>0.01*</b>
COPD, n (%)	3,718 (3.4)	11,010 (4.0)	<b>0.03*</b>	3,599 (3.5)	3,413 (3.3)	<b>0.01*</b>
Deep Vein Thrombosis, n (%)	1,536 (1.4)	4,528 (1.6)	<b>0.02*</b>	1,486 (1.4)	1,461 (1.4)	<0.01
Depression, n (%)	25,442 (23.4)	66,672 (24.1)	<b>0.02*</b>	24,467 (23.5)	24,112 (23.2)	<0.01
Diabetic Nephropathy, n (%)	24 (0.0)	52 (0.0)	<0.01	12 (0.0)	9 (0.0)	<0.01
Essential Tremors, n (%)	1,114 (1.0)	3,412 (1.2)	<b>0.02*</b>	1,096 (1.1)	1,046 (1.0)	<0.01
Glaucoma, n (%)	2,314 (2.1)	6,813 (2.5)	<b>0.02*</b>	2,257 (2.2)	2,174 (2.1)	<0.01
Glomerular Diseases, n (%)	125 (0.1)	377 (0.1)	<0.01	124 (0.1)	126 (0.1)	<0.01
Heart Failure, n (%)	342 (0.3)	2,914 (1.1)	<b>0.09*</b>	135 (0.1)	154 (0.1)	<0.01
Hypercalciuria, n (%)	168 (0.2)	531 (0.2)	<b>0.01*</b>	166 (0.2)	149 (0.1)	<0.01

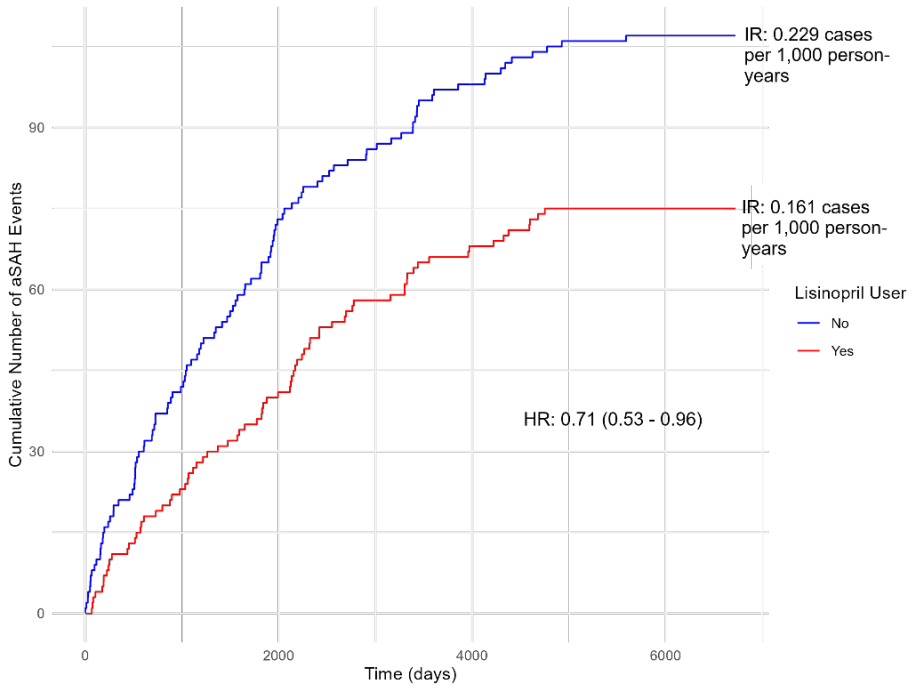
Supplementary Table 1. Continued

Variable	Lisinopril (Pre-matching)	Other ACE inhibitor (Pre-matching)	SMD (Pre-matching)	Lisinopril (Post-matching)	Other ACE inhibitor (Post-matching)	SMD (Post-matching)
Hypercholesterolaemia, n (%)	18,360 (16.9)	51,239 (18.6)	<b>0.04*</b>	17,677 (17.0)	17,600 (16.9)	<0.01
Hypertrophic Cardiomyopathy, n (%)	724 (0.7)	2,037 (0.7)	<b>0.01*</b>	694 (0.7)	611 (0.6)	<b>0.01*</b>
Hyperthyroidism, n (%)	1,485 (1.4)	4,004 (1.5)	<b>0.01*</b>	1,435 (1.4)	1,322 (1.3)	<b>0.01*</b>
Liver Failure, n (%)	12 (0.0)	28 (0.0)	<0.01	8 (0.0)	7 (0.0)	<0.01
Migraines, n (%)	8,088 (7.4)	20,411 (7.4)	<0.01	7,847 (7.5)	7,656 (7.4)	<0.01
Myocardial Infarction, n (%)	940 (0.9)	8,001 (2.9)	<b>0.15*</b>	104 (0.1)	105 (0.1)	<0.001
Nephrotic Syndrome, n (%)	126 (0.1)	375 (0.1)	<0.01	121 (0.1)	115 (0.1)	<0.01
Osteoporosis, n (%)	4,581 (4.2)	13,297 (4.8)	<b>0.03*</b>	4,460 (4.3)	4,336 (4.2)	<0.01
Parkinson's Disease, n (%)	170 (0.2)	563 (0.2)	<b>0.01*</b>	167 (0.2)	161 (0.2)	<0.01
Pericarditis, n (%)	224 (0.2)	659 (0.2)	<0.01	209 (0.2)	180 (0.2)	<0.01
Peripheral Vascular Disease, n (%)	1,705 (1.6)	4,812 (1.7)	<b>0.01*</b>	1,521 (1.5)	1,398 (1.3)	<b>0.01*</b>
Portal Hypertension, n (%)	39 (0.0)	60 (0.0)	<b>0.01*</b>	5 (0.0)	8 (0.0)	<0.01
Proteinuria and Albuminuria, n (%)	1,207 (1.1)	3,719 (1.3)	<b>0.02*</b>	1,165 (1.1)	1,144 (1.1)	<0.01
Pulmonary Hypertension, n (%)	44 (0.0)	196 (0.1)	<b>0.01*</b>	39 (0.0)	37 (0.0)	<0.01
Pulmonary Oedema, n (%)	39 (0.0)	307 (0.1)	<b>0.03*</b>	29 (0.0)	25 (0.0)	<0.01
Raynaud's Disease, n (%)	1,040 (1.0)	2,842 (1.0)	<b>0.01*</b>	1,014 (1.0)	938 (0.9)	<0.01
Scleroderma, n (%)	41 (0.0)	118 (0.0)	<0.01	40 (0.0)	38 (0.0)	<0.01
Stroke, n (%)	3,183 (2.9)	13,254 (4.8)	<b>0.01*</b>	3,042 (2.9)	3,119 (3.0)	<0.01
Venous Thromboembolism, n (%)	2,186 (2.0)	6,530 (2.4)	<b>0.02*</b>	2,113 (2.0)	2,086 (2.0)	<0.01

\* Standardized means differ significantly at p<0.05 (indicated in bold). SMD = standardised mean differences, ACE = angiotensin-converting enzyme, COPD = Chronic obstructive pulmonary disease, BMI = Body Mass Index.



**Supplementary Figure 1.** Kaplan-Meier survival probability for aSAH risk over time for lisinopril users compared to other angiotensin-converting enzyme (ACE) inhibitor users.



**Supplementary Figure 2.** Cumulative incidence of aneurysmal subarachnoid haemorrhage (aSAH) for lisinopril users compared to other angiotensin-converting enzyme (ACE) inhibitor users.

HR = Hazard ratio. IR = Incidence rate



Chapter 9

## General discussion

---

This thesis showed how data-driven approaches can improve our understanding of aneurysmal subarachnoid haemorrhage (aSAH) risk estimation, risk factors, and alternative treatment options. The following general discussion presents important aspects of data-driven aSAH research: the factors contributing to big data's prominent role in contemporary biomedical and healthcare research, the different worlds of clinicians and data scientists, and the future of aSAH prevention.

## **9.A WHAT FACTORS CONTRIBUTE TO THE PROMINENT ROLE OF BIG DATA IN CONTEMPORARY BIOMEDICAL AND HEALTHCARE RESEARCH?**

The increasingly prominent role of big data in biomedical and healthcare research is transforming how medical diseases are diagnosed, understood, and treated.<sup>1</sup> Big data is defined as data that is too large or complex to be managed by traditional data-processing tools due to the challenges imposed by the so-called five V's. These are: Volume (the extensive amount of data generated), value (the relevant insights derived from data), variety (different types of data such as images, text, and lab results), velocity (the rapid rate at which data is collected and processed), and veracity (the accuracy and reliability of data).<sup>2</sup> Despite these challenges,<sup>3</sup> big data offers opportunities that traditional data sources do not have, including the potential for hypothesis-free designs and simultaneous exploration of multiple variables.<sup>4</sup> In biomedical sciences, using big data has improved personalised risk prediction,<sup>5</sup> enabled the discovery of novel biomarkers,<sup>6</sup> and facilitated novel treatments.<sup>7,8</sup>

The growing interest in big data is primarily driven by two concurrent trends: the broader availability of large data sources, including electronic health records (EHRs) and biobanks, and the expansion of machine learning algorithms to analyse this data. These expanding trends are commonly believed to signify advancements; the more sophisticated the algorithm or the larger the data source, the better.<sup>9</sup> However, each new algorithm and data source offers unique opportunities and limitations tailored to answering specific research questions. The following section discusses some prominent new algorithms and data sources and identifies the research questions they are most suitable to answer.

### **The expansion of machine learning algorithms**

A machine learning algorithm in big data is a mathematical formula or process used to analyse data and make predictions.<sup>10</sup> Over the last few decades, various machine

learning algorithms have emerged, differing in accuracy (how closely the model predictions simulate reality), robustness (how sensitive a model is to violations of model assumptions), computational complexity (the amount of computational resources required to train and run the model), scalability (how much data a model can handle), and interpretability (the degree to which a human can understand the predictions made by the model).<sup>10</sup>

Algorithms are commonly divided into traditional and modern ones. Traditional logistic and Cox regression algorithms are relatively accurate, interpretable, and have minimal computational complexity.<sup>9</sup> However, they are sensitive to model assumption violations and typically handle only a limited number of variables, limiting their robustness and scalability. In contrast, modern algorithms such as penalised logistic regression, random decision forests, gradient boosting, and neural networks are designed to be scalable by managing many variables and observations while maintaining high accuracy, often at the cost of computational complexity. Modern algorithms are also robust because they require fewer model assumptions than traditional ones.<sup>11</sup> However, a major limitation of modern algorithms is that they are difficult to interpret, resulting in so-called black-box models.<sup>12</sup>

It is a misconception that more modern algorithms are always superior to traditional ones.<sup>13</sup> Traditional and modern algorithms rarely differ in accuracy when datasets are relatively small and in a structured format. For example, modern machine learning models seldom outperform logistic or Cox regression in clinical prediction settings.<sup>13–16</sup> In addition, modern algorithms' lack of interpretability means that modern algorithms are rarely implemented in practice.<sup>17</sup>

The choice of an algorithm should be determined by the suitability of the algorithm characteristics to answer a specific research question. Traditional algorithms are preferable if researchers need to report interpretable and quantifiable findings on a relatively small set of variables. For instance, we used logistic regression analyses in chapters 7 and 8 to quantify the relationship between commonly prescribed drugs and aSAH. Consequently, we were able to conclude that lisinopril users had a 29% lower aSAH risk than users of other angiotensin-converting-enzyme (ACE) inhibitors. In contrast, modern algorithms are preferable when researchers need to study many variables simultaneously. For instance, in chapters 3 and 4, we used a penalised Cox regression algorithm to integrate nearly all EHR variables into a single prediction model. A combination of algorithms can be used when both scalability and interpretability are required. For example, in chapter 6, we combined modern (i.e. Catboost) and traditional (i.e. logistic regression) algorithms to cover each

other's limitations. We first used the scalability of the Catboost model to integrate all relevant UK Biobank variables into a single model and subsequently quantified the results using the interpretability of logistic regression.

### **Broader availability of large data sources**

Another important development in the increasingly prominent role of big data in biomedical and healthcare research is the availability of EHRs for research purposes and the creation of biobanks. These two data sources have their strengths and weaknesses, similar to the abovementioned algorithms.

EHRs contain patient data collected during routine clinical care, including signs and symptoms, diagnoses, laboratory tests, imaging, procedures, and physician notes.<sup>18</sup> While primarily collected for clinical and insurance purposes, EHRs are also valuable for biomedical research. They offer an extensive collection of retrospective data that is both readily accessible and relatively cheap to acquire, especially compared to the expensive establishment of prospective cohorts and the conduction of clinical trials.<sup>19,20</sup> Another advantage of EHRs, particularly those derived from general practitioners, is their ability to capture data from large populations over extended periods. This allows researchers to study rare events (such as adverse drug reactions) and special populations (such as children, the elderly, and pregnant women) that are not commonly included in standard randomised controlled trials.<sup>19,21</sup> However, using EHRs for biomedical research also has limitations. As described in chapter 2, EHRs do not come in a standard structured format, and relatively minor methodological choices on how to format these data can substantially affect study results. Additionally, since EHRs are not primarily collected for research purposes, they often lack systematic assessment of exposures, leading to data missingness (which is likely not missing at random) and potential exposure or outcome misclassification.<sup>18,20,22</sup>

Biobanks can help to overcome some of the challenges present in EHRs. Biobanks, defined as extensive collections of medical data and tissue samples collected for research purposes,<sup>23</sup> enable large-scale analysis of well-annotated clinical and biological data from patients and healthy individuals.<sup>24</sup> Unlike EHRs, biobanks were developed for scientific purposes, meaning biomedical data is systematically collected for each participant. However, the benefits of biobank-derived data are balanced by high acquisition and maintenance costs. In addition, individuals who register for biobank participation tend to be relatively healthy and upper-middle class and may, therefore, not accurately represent the general population.<sup>25</sup>

Similar to a data algorithm, the choice of a data source should be determined by its suitability to answer a specific research question. EHRs are suitable when many patients, an extended follow-up duration, or an underrepresented population are needed. For example, we used EHRs to develop clinical cardiovascular prediction models in young men and women, a group generally underrepresented in clinical trials.<sup>26</sup> Similarly, we were able to use EHRs to identify a large number of drug users to identify commonly prescribed drugs associated with aSAH in chapter 7. Additionally, the long follow-up duration of EHRs allowed us to identify nuances related to the recency of drug use, such as increased aSAH risk with recent use of antihypertensives and reduced aSAH risk with current use. EHRs are unsuitable when answering a research question that requires risk factors to be systematically assessed for each individual. For example, we could not use EHRs to develop an aSAH prediction model in chapter 4 due to the lack of systematic data collection on important aSAH risk factors such as smoking and hypertension in people at risk of aSAH. In such scenarios, Biobanks are preferable. For example, we were able to develop a more accurate aSAH prediction model using UK Biobank data, as described in chapter 5. Similarly, we used the rigorous data collection of the UK Biobank to identify novel potential risk factors for aSAH, as described in chapter 6.

### **Combining algorithms and data sources**

In conclusion, the increasingly prominent role of big data in biomedical and healthcare is driven by a wider availability of algorithms and data sources. Contrary to common belief, these advancements are not primarily due to the superior performance of modern algorithms over traditional ones, nor are they because larger data sources are inherently better than smaller ones.<sup>9</sup> Instead, as we have shown throughout this thesis, each specific algorithm and data source possess characteristics that make them suitable to answer a specific research question.

Efforts are underway to address the limitations of the algorithms and data sources discussed previously. There is a growing emphasis on enhancing the interpretability of machine learning models and enabling them to explain their predictions.<sup>22</sup> In return, traditional models such as logistic and Cox regression are being adapted to achieve the scalability and robustness of contemporary machine learning techniques.<sup>23</sup> Additionally, the standardisation of EHRs is progressing.<sup>24</sup> Some countries are implementing more routine health checkups, which would partly mitigate the issue of insufficient exposure information in EHRs.<sup>27</sup> Biobanks are also increasingly integrating with other data sources, combining systematic exposure assessments with long-term EHR-based outcomes.<sup>28</sup>

Despite these advancements, the inherent strengths and weaknesses of each algorithm and data source are likely to remain. Therefore, understanding the strengths and weaknesses remains crucial for effectively using big data in biomedical and health sciences.

## **9.B THE DIFFERENT WORLDS OF CLINICIANS AND DATA SCIENTISTS**

Thousands of clinical prediction models are developed each year to aid diagnosis and prognosis in healthcare.<sup>29</sup> In theory, these models can help clinicians predict outcomes, guide therapeutic decisions, and automate routine tasks.<sup>30</sup> However, most of these models are never implemented in clinical settings due to clinical irrelevance, poor description, lack of validation, or remaining closed off from public use.<sup>31-34</sup> Although some issues may stem from poor study design, the disconnect between model developers and users is a major barrier to implementation.<sup>35</sup> Without bridging this gap, most models will likely remain unused.

The gap between model developers and users mainly arises due to different priorities. Developers often focus on technical accuracy, innovation, and complexity, using advanced algorithms that may not be readily applicable in everyday clinical practice.<sup>36</sup> In contrast, model users generally prefer practical models that are easy to integrate into existing workflows and directly benefit patients.<sup>37</sup> Consequently, clinicians often favour simpler, score-based clinical risk models over more accurate but opaque 'black box' machine learning models.<sup>38,39</sup> Discrepancies in stated goals aggravate the disconnect between model developers and users. The developer literature commonly presents machine learning models as competing with or surpassing human clinicians, implying that these models could eventually replace clinicians.<sup>40</sup> This notion naturally fosters mistrust and resistance among clinicians,<sup>41</sup> whose literature often emphasises human-computer interaction instead.<sup>42</sup> Moreover, the different professional languages and literature developers and intended users use further increase these issues. Each group typically publishes in different academic journals, creating informational silos and using different terminology, complicating mutual understanding and collaboration.<sup>43</sup>

Recent attempts have been made to bridge the gap between developers and model users. Several data science tools have been developed to enhance the interpretability of machine learning models. For example, tools such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive

exPlanations (SHAP) help visualise the reasoning behind complex model predictions.<sup>44,45</sup> In chapter 6, we used such SHAP values to identify which variables the Catboost prediction model used to make its predictions. Tools such as these can improve clinicians' understanding and eventual trust in clinical prediction models, thereby increasing their likelihood of use in clinical decision-making.<sup>17</sup> Similar efforts to increase mutual understanding include conference workshops, academic primers, and new programs such as Health Data Science.

However, such attempts are unlikely to be successful without addressing the underlying disconnect between model developers and users. Interpretability tools are unlikely to be effective if they remain in journals that model users will never read. Similarly, conference workshops and programs are only effective if model developers and users attend these sessions. A more systemic solution involves fostering direct, continuous collaboration between clinicians and data scientists from the beginning of a project.<sup>46</sup> This collaboration would likely involve clinicians identifying relevant applications for prediction models and data scientists selecting suitable machine learning algorithms and data sources. Establishing a flexible workflow with open communication between clinicians and data scientists would enhance the likelihood of practical implementation of these models.

## **9.C THE FUTURE OF ASAH PREVENTION: AN INTEGRATED DATA-DRIVEN APPROACH**

In this thesis, we have contributed to the prevention of aSAH by developing risk prediction models, identifying new risk factors, and exploring potential non-invasive pharmacotherapy modalities. While these findings represent a step forward, they form only a tiny piece of a larger puzzle. The following steps must include expanding these initial results and integrating them with other aSAH study fields to create a comprehensive approach to aSAH prevention.

Several factors currently limit the prevention of aSAH: a lack of understanding about aneurysm formation and rupture and the absence of non-invasive treatment options. At present, identifying individuals at risk of aneurysm formation and subsequent rupture is based on screening. However, screening is exclusively recommended for individuals at high risk of aSAH, such as direct family members of aSAH patients and patients with autosomal dominant polycystic disease (ADPKD), a condition associated with an increased risk of aSAH.<sup>47</sup> This means that most aneurysms are either discovered incidentally or identified after rupture.<sup>48</sup> Even

when an aneurysm is discovered incidentally, predicting the aneurysmal rupture risk remains challenging. Existing models for assessing the risk of aneurysm rupture are inadequate,<sup>49</sup> and the available treatment options (e.g. neurosurgical, endovascular) are limited and often do not provide a clear benefit over their associated risks.<sup>50</sup> Consequently, most aneurysms remain untreated.<sup>51</sup>

Our research can help address the challenges previously mentioned. The aSAH prediction model developed in chapter 5 can help identify individuals at risk of aSAH within the general population. In chapter 6, we identified six potentially novel aSAH risk factors: mean sphered cell volume, urea levels, tea intake, peak expiratory flow, insulin-like growth factor 1, and haematocrit percentage. Upon validation, these factors could enhance screening processes or assist in better evaluating the risk of aneurysmal rupture. In chapter 7, we identified five commonly prescribed drugs that might serve as non-invasive treatments to lower aSAH risk. In Chapter 8, our studies on lisinopril revealed a 29% reduction in aSAH risk compared to other ACE inhibitors, suggesting its potential as the preferred antihypertensive for aneurysm patients, offering a safer, non-invasive treatment alternative. Finally, our results should be integrated with advances in other fields. Integrating our findings with genetics, imaging, and clinical research, advancements could further refine aSAH risk assessments, uncover additional risk factors, and develop personalised treatment strategies.

## REFERENCES

1. Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomed Inform Insights*. 2016;8:BII.S31559.
2. Bailly S, Meyfroidt G, Timsit J-F. What's new in ICU in 2050: big data and machine learning. *Intensive Care Med*. 2018;44:1524–1527.
3. Fan J, Han F, Liu H. Challenges of Big Data analysis. *National Science Review*. 2014;1:293–314.
4. Toga AW, Foster I, Kesselman C, Madduri R, Chard K, Deutsch EW, Price ND, Glusman G, Heavner BD, Dinov ID, et al. Big biomedical data as the key resource for discovery science. *Journal of the American Medical Informatics Association*. 2015;22:1126–1131.
5. Obermeyer Ziad, Emanuel Ezekiel J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*. 2016;375:1216–1219.
6. Lin Y, Qian F, Shen L, Chen F, Chen J, Shen B. Computer-aided biomarker discovery for precision medicine: data resources, models and applications. *Briefings in Bioinformatics*. 2019;20:952–975.
7. Hey T, Tansley S, Tolle KM, others. The fourth paradigm: data-intensive scientific discovery. Microsoft research Redmond, WA; 2009.
8. Cahan EM, Hernandez-Boussard T, Thadaney-Israni S, Rubin DL. Putting the data before the algorithm in big data addressing personalized healthcare. *npj Digit. Med*. 2019;2:1–6.
9. Levy JJ, O'Malley AJ. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC Medical Research Methodology*. 2020;20:171.
10. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*. 2015;349:255–260.
11. Efron B, Tibshirani R. Statistical Data Analysis in the Computer Age. *Science*. 1991;253:390–395.
12. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1:206–215.
13. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
14. Cerasa A, Tartarisco G, Bruschetta R, Ciancarelli I, Morone G, Calabrò RS, Pioggia G, Tonin P, Iosa M. Predicting Outcome in Patients with Brain Injury: Differences between Machine Learning versus Conventional Statistics. *Biomedicines*. 2022;10:2267.
15. Wu X, Yuan X, Wang W, Liu K, Qin Y, Sun X, Ma W, Zou Y, Zhang H, Zhou X, et al. Value of a Machine Learning Approach for Predicting Clinical Outcomes in Young Patients With Hypertension. *Hypertension*. 2020;75:1271–1278.
16. Nusinovići S, Tham YC, Chak Yan MY, Wei Ting DS, Li J, Sabanayagam C, Wong TY, Cheng C-Y. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*. 2020;122:56–69.
17. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*. 2019;19:146.
18. Katehakis DG, Tsiknakis M. Electronic Health Record [Internet]. In: Wiley Encyclopedia of Biomedical Engineering. John Wiley & Sons, Ltd; 2006 [cited 2024 May 3]. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780471740360.ebs1440>

19. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol.* 2017;106:1–9.
20. Bots SH, Groenwold RHH, Dekkers OM. Using electronic health record data for clinical research: a quick guide. *Eur J Endocrinol.* 2022;186:E1–E6.
21. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, Ferris TA, Shah NH. Pharmacovigilance Using Clinical Notes. *Clinical Pharmacology & Therapeutics.* 2013;93:547–555.
22. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery.* 2019;9:e1312.
23. Dumitrescu E, Hué S, Hurlin C, Tokpavi S. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research.* 2022;297:1178–1192.
24. Jha AK. Meaningful Use of Electronic Health Records: The Road Ahead. *JAMA.* 2010;304:1709–1710.
25. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol.* 2017;186:1026–1034.
26. Carcel C, Harris K, Peters SAE, Sandset EC, Balicki G, Bushnell CD, Howard VJ, Reeves MJ, Anderson CS, Kelly PJ, et al. Representation of Women in Stroke Clinical Trials. *Neurology.* 2021;97:e1768–e1774.
27. Liss DT, Uchida T, Wilkes CL, Radakrishnan A, Linder JA. General Health Checks in Adult Primary Care: A Review. *JAMA.* 2021;325:2294–2306.
28. Grobbee D, Hoes A, Verheij T, Schrijvers A, Ameijden EV, Numans M. The Utrecht Health Project: Optimization of routine healthcare data for research. *European Journal of Epidemiology.* 2004;20:285–290.
29. Riley RD, Pate A, Dhiman P, Archer L, Martin GP, Collins GS. Clinical prediction models and the multiverse of madness. *BMC Medicine.* 2023;21:502.
30. Ranstam J, Cook JA, Collins GS. Clinical prediction models. *British Journal of Surgery.* 2016;103:1886.
31. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, de Ridder MAJ, Seinen TM, Williams RD, Rijnbeek PR. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. *J Am Med Inform Assoc.* 2022;29:983–989.
32. Arshi B, Wynants L, Rijnhart E, Reeve K, Cowley LE, Smits LJ. What proportion of clinical prediction models make it to clinical practice? Protocol for a two-track follow-up study of prediction model development publications. *BMJ Open.* 2023;13:e073174.
33. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ.* 2009;338:b606.
34. Wessler BS, Lai YH L, Kramer W, Cangelosi M, Raman G, Lutz JS, Kent DM. Clinical Prediction Models for Cardiovascular Disease. *Circulation: Cardiovascular Quality and Outcomes.* 2015;8:368–375.
35. Celi LA, Davidzon G, Johnson AE, Komorowski M, Marshall DC, Nair SS, Phillips CT, Pollard TJ, Raffa JD, Saliccioli JD, et al. Bridging the Health Data Divide. *Journal of Medical Internet Research.* 2016;18:e6400.
36. Wagstaff K. Machine Learning that Matters [Internet]. 2012 [cited 2024 May 7]; Available from: <http://arxiv.org/abs/1206.4656>
37. Labarère J, Bertrand R, Fine M. How to derive and validate clinical prediction models for use in intensive care medicine. *Intensive Care Medicine.* 2014;40:513–527.

38. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, Andreini D, Budoff MJ, Cademartiri F, Callister TQ, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *European Heart Journal*. 2017;38:500–507.
39. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*. 2017;12:e0174944.
40. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, He Z, Wong SY, Fang P-H, Ming W-K. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR Medical Informatics*. 2019;7:e10010.
41. Huisman M, Ranschaert E, Parker W, Mastrodicasa D, Koci M, Pinto de Santos D, Coppola F, Morozov S, Zins M, Bohyn C, et al. An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. *Eur Radiol*. 2021;31:7058–7066.
42. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, Janda M, Lallas A, Longo C, Malvey J, et al. Human–computer collaboration for skin cancer recognition. *Nat Med*. 2020;26:1229–1234.
43. Shouval R, Fein JA, Savani B, Mohty M, Nagler A. Machine learning and artificial intelligence in haematology. *British Journal of Haematology*. 2021;192:239–250.
44. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier [Internet]. 2016 [cited 2024 May 7]; Available from: <http://arxiv.org/abs/1602.04938>
45. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions [Internet]. 2017 [cited 2024 Mar 9]; Available from: <http://arxiv.org/abs/1705.07874>
46. Bastian G, Baker GH, Limon A. Bridging the divide between data scientists and clinicians. *Intelligence-Based Medicine*. 2022;6:100066.
47. Rinkel GJ, Ruigrok YM. Preventive screening for intracranial aneurysms. *International Journal of Stroke*. 2022;17(1):30–36.
48. Harrison CH, Taquet M, Harrison PJ, Watkinson PJ, Rowland MJ. Sex and age effects on risk of non-traumatic subarachnoid hemorrhage: Retrospective cohort study of 124,234 cases using electronic health records. *Journal of Stroke and Cerebrovascular Diseases*. 2023;32:107196.
49. Pettersson SD, Skrzypkowska P, Pietrzak K, Och A, Siedlecki K, Czaplá-Iskrzycka A, Klepinowski T, Fodor T, Filo J, Meyer-Szary J, et al. Evaluation of PHASES Score for Predicting Rupture of Intracranial Aneurysms: Significance of Aneurysm Size. *World Neurosurg*. 2024;184:e178–e184.
50. Algra AM, Lindgren A, Vergouwen MDI, Greving JP, van der Schaaf IC, van Doormaal TPC, Rinkel GJE. Procedural Clinical Complications, Case-Fatality Risks, and Risk Factors in Endovascular and Neurosurgical Treatment of Unruptured Intracranial Aneurysms: A Systematic Review and Meta-analysis. *JAMA Neurol*. 2019;76:282–293.
51. Etminan N, Rinkel GJ. Unruptured intracranial aneurysms: development, rupture and preventive management. *Nat Rev Neurol*. 2016;12:699–713.



## Chapter 10

### Summary

---

Aneurysmal subarachnoid haemorrhage (aSAH) is a devastating type of stroke caused by the rupture of an intracranial aneurysm. Although accounting for only a tenth of strokes globally, aSAH has high morbidity and mortality rates and typically affects relatively younger individuals. The overarching aim of this thesis was to improve our understanding of aSAH risk estimation, risk factors, and alternative treatment options using data-driven approaches. It was structured around three objectives: 1) to enhance the predictive accuracy for identifying individuals at risk of aSAH; 2) to discover and describe novel aSAH risk factors; and 3) to explore non-invasive, drug-based treatment options for aSAH.

Chapters **2** to **5** addressed the first objective to enhance predictive accuracy for identifying individuals at risk of aSAH. **Chapter 2** discussed the effects of various data preparation methods on the calibration and discrimination of predictive models based on electronic health records (EHRs) from the Extramural LUMC Academic Network (ELAN) in the Netherlands. We analysed nine datasets, varying the run-in periods, outcome definitions, and strategies for handling missing data. Our multivariable Cox proportional hazards models incorporated predictors like age, sex, blood pressure, cholesterol, and smoking habits. We observed consistent model discrimination across datasets, but calibration varied significantly depending on different outcome definitions and methods of addressing missing data.

**Chapter 3** describes the development of sex-specific prediction models for first-ever cardiovascular events in adults aged 30 to 49 using Dutch EHR data, covering 542,141 patients without prior cardiovascular disease. We used Cox proportional hazards models with both traditional cardiovascular predictors and the 20 or 50 most significant predictors identified through data-driven methods. Traditional models showed moderate performance (C-index 0.65 for women, 0.66 for men), with data-driven models modestly enhancing performance, especially for women (C-index improved by 0.030 for women and 0.012 for men). Data-driven methods also improved the net reclassification index, particularly for women (3.7% versus 1.2% for men).

**Chapter 4** describes our attempts to develop a predictive model for aSAH using EHR data, comparing it to predictive models for acute ischemic stroke (AIS) and intracerebral haemorrhage (ICH). Data from Dutch routine care databases of individuals aged 35 and above were analysed, involving 1,040,855 individuals over 10,173,170 person-years. A penalized Cox regression model identified 19 predictors for aSAH and found the model's discriminative performance to be moderate (c-statistic of 0.61), whereas AIS and ICH models performed better (c-statistics of

0.77 each). We conclude that identifying high-risk individuals for aSAH using EHR data is challenging due aSAH being rare and occurring at a relatively young age.

**Chapter 5** details the SMA2SH2ERS aSAH risk prediction model development and validation in the general population using data from the UK Biobank and the Trøndelag Health Study. We developed a Cox regression model including known aSAH predictors such as sex, diabetes, age, alcohol consumption, smoking, hypertension, hypercholesterolemia, educational attainment, regular physical activity, and family history of stroke. The model's ability to discriminate and calibrate risk predictions yielded c-statistics of 0.62 and 0.64 in the development and validation cohorts, respectively.

**Chapter 6** focuses on the second objective by identifying and describing novel aSAH risk factors by combining machine learning and traditional statistics in the UK Biobank. We identified 893 aSAH cases among 501,847 UK Biobank participants and implemented a Catboost machine learning algorithm to identify the 25 most influential predictors using Shapley Additive Explanations. We further explored these predictors using logistic regression, correcting for already established aSAH risk factors such as age, sex, hypertension, smoking status, and alcohol use. This established previously reported findings and identified potentially novel aSAH risk factors such as mean spheroid cell volume, urea levels, and tea intake, along with peak expiratory flow, insulin-like growth factor 1, and haematocrit percentage.

Chapters **7** and **8** address the third and final objective by examining non-invasive drug-based treatment options for aSAH. **Chapter 7** describes a drug-wide association study that investigated the association between commonly prescribed drugs and the incidence of aSAH using the Secure Anonymised Information Linkage (SAIL) databank. Analysing 4,879 aSAH cases and 43,911 controls from 2000 to 2020, we investigated 205 commonly prescribed drugs across different usage windows. We found that the current use of lisinopril, simvastatin, metformin, and tamsulosin was associated with a decreased risk of aSAH. In contrast, current use of warfarin, venlafaxine, prochlorperazine, and co-codamol was associated with an increased risk of aSAH.

**Chapter 8** follows-up on the relationship between lisinopril use and aSAH risk using an active comparator new user design in the UK Clinical Practice Research Datalink (CPRD) GOLD database. We included 108,717 lisinopril users and 276,113, with a median follow-up time of three years. Using propensity score matching and Cox proportional hazards models, we found that lisinopril users had a significantly lower risk of aSAH compared to other ACE inhibitor users, with a hazard ratio of 0.71, indicating a 29% reduced risk of aSAH for lisinopril users.



Chapter 11

## Nederlandse samenvatting

---

Aneurysmatische subarachnoïdale bloeding (aSAH) is een verwoestend type beroerte veroorzaakt door het scheuren van een intracranieel aneurysma. Hoewel aSAH wereldwijd slechts een tiende van de beroertes uitmaakt, heeft het hoge morbiditeits- en mortaliteitscijfers en treft het doorgaans relatief jongere individuen. Het overkoepelende doel van dit proefschrift was om ons begrip van aSAH risicoschatting, risicofactoren en alternatieve behandelingsopties te verbeteren met behulp van datagedreven benaderingen. Het was gestructureerd rond drie doelstellingen: 1) de voorspellende nauwkeurigheid voor het identificeren van individuen met risico op aSAH verbeteren; 2) nieuwe aSAH-risicofactoren ontdekken en beschrijven; en 3) niet-invasieve, medicijngebaseerde behandelingsopties voor aSAH verkennen.

**Hoofdstukken 2** tot en met **5** richtten zich op de eerste doelstelling om de voorspellende nauwkeurigheid voor het identificeren van individuen met risico op aSAH te verbeteren. **Hoofdstuk 2** besprak de effecten van verschillende methoden voor datavoorbereiding op de kalibratie en discriminatie van voorspellende modellen op basis van elektronische medische dossiers (EHRs) van het Extramurale LUMC Academisch Netwerk (ELAN) in Nederland. We analyseerden negen datasets, waarbij we variaties aanbrachten in de *run-in* periodes, uitkomstdefinities en strategieën voor het omgaan met missende gegevens. Onze multivariabele Cox proportionele hazard modellen omvatten voorspellers zoals leeftijd, geslacht, bloeddruk, cholesterol en rookgewoonten. We observeerden consistente modeldiscriminatie over datasets, maar de kalibratie varieerde aanzienlijk afhankelijk van de verschillende uitkomstdefinities en methoden voor het aanpakken van ontbrekende gegevens.

**Hoofdstuk 3** beschrijft de ontwikkeling van geslachtsspecifieke voorspellingsmodellen voor eerste cardiovasculaire gebeurtenissen bij volwassenen van 30 tot 49 jaar, gebruikmakend van Nederlandse EHR-gegevens, waarbij 542.141 patiënten zonder eerdere cardiovasculaire ziekte werden onderzocht. We gebruikten Cox proportionele hazard modellen met zowel traditionele cardiovasculaire voorspellers als de 20 of 50 meest significante voorspellers geïdentificeerd via data-gedreven methoden. Traditionele modellen toonden een matige prestatie (C-index 0,65 voor vrouwen, 0,66 voor mannen), waarbij data-gedreven modellen de prestaties bescheiden verbeterden, vooral voor vrouwen (C-index verbeterde met 0,030 voor vrouwen en 0,012 voor mannen). Data-gedreven methoden verbeterden ook de net reclassification index, met name voor vrouwen (3,7% versus 1,2% voor mannen).

**Hoofdstuk 4** beschrijft onze pogingen om een voorspellingsmodel voor aSAH te ontwikkelen met behulp van EHR-gegevens, waarbij we dit vergeleken met voorspellingsmodellen voor acute ischemische beroerte (AIS) en intracerebrale bloeding (ICH). Gegevens van Nederlandse routinezorgdatabases van individuen van 35 jaar en ouder werden geanalyseerd, waarbij 1.040.855 individuen over 10.173.170 persoonsjaren werden betrokken. Een gepenaliseerd Cox regressiemodel identificeerde 19 voorspellers voor aSAH en vond dat de discriminerende prestatie van het model matig was (c-statistiek van 0,61), terwijl de modellen voor AIS en ICH beter presteerden (c-statistieken van elk 0,77). We concluderen dat het identificeren van individuen met hoog risico voor aSAH met behulp van EHR-gegevens uitdagend is vanwege de zeldzaamheid van aSAH en het feit dat het relatief op jonge leeftijd voorkomt.

**Hoofdstuk 5** beschrijft de ontwikkeling en validatie van het SMA2SH2ERS aSAH risicovoorspellingsmodel in de algemene bevolking met behulp van gegevens van de UK Biobank en de Trøndelag Gezondheidsstudie. We ontwikkelden een Cox regressiemodel met bekende aSAH voorspellers zoals geslacht, diabetes, leeftijd, alcoholconsumptie, roken, hypertensie, hypercholesterolemie, opleidingsniveau, regelmatige fysieke activiteit en familiegeschiedenis van beroerte. Het model's vermogen om risico-voorspellingen te discrimineren en kalibreren leverde c-statistieken op van respectievelijk 0,62 en 0,64 in de ontwikkelings- en validatiecohorten.

**Hoofdstuk 6** richt zich op de tweede doelstelling door nieuwe aSAH risicofactoren te identificeren en beschrijven met behulp van machine learning en traditionele statistieken in de UK Biobank. We identificeerden 893 aSAH-gevallen onder 501.847 UK Biobank deelnemers en implementeerden een Catboost machine learning algoritme om de 25 meest invloedrijke voorspellers te identificeren met behulp van Shapley Additive Explanations. We onderzochten deze voorspellers verder met behulp van logistische regressie, waarbij we corrigeerden voor reeds vastgestelde aSAH risicofactoren zoals leeftijd, geslacht, hypertensie, rookstatus en alcoholgebruik. Dit bevestigde eerder gerapporteerde bevindingen en identificeerde potentieel nieuwe aSAH risicofactoren zoals gemiddelde celvolume, ureumspiegels en thee-inname, samen met piekstroom, insuline-achtige groeifactor 1 en hematocrietpercentage.

**Hoofdstukken 7 en 8** behandelen de derde en laatste doelstelling door niet-invasieve medicijngebaseerde behandelingsopties voor aSAH te onderzoeken.

**Hoofdstuk 7** beschrijft een *drug-wide association study* die de associatie tussen vaak voorgeschreven medicijnen en de incidentie van aSAH onderzocht met behulp

van de Secure Anonymised Information Linkage (SAIL) databank. Door 4.879 aSAH-gevallen en 43.911 controles van 2000 tot 2020 te analyseren, onderzochten we 205 vaak voorgeschreven medicijnen over verschillende gebruikperiodes. We vonden dat het huidige gebruik van lisinopril, simvastatine, metformine en tamsulosine geassocieerd was met een verminderd risico op aSAH. Daarentegen was het huidige gebruik van warfarine, venlafaxine, prochlorperazine en co-codamol geassocieerd met een verhoogd risico op aSAH.

**Hoofdstuk 8** volgt de relatie tussen lisinoprilgebruik en aSAH-risico verder op met behulp van een *active comparator new user* design in de UK Clinical Practice Research Datalink (CPRD) GOLD-database. We includeerden 108.717 lisinoprilgebruikers en 276.113 andere ACE-remmergebruikers, met een mediane follow-up tijd van drie jaar. Met behulp van propensity score matching en Cox proportionele hazard modellen vonden we dat lisinoprilgebruikers een significant lager risico op aSAH hadden vergeleken met andere ACE-remmergebruikers, met een hazard ratio van 0,71, wat wijst op een 29% verminderd risico op aSAH voor lisinoprilgebruikers..





Appendices

Publications by the author

About the author

Dankwoord

## PUBLICATIONS BY THE AUTHOR

### Accepted Publications

- Hendrikus J. A. van Os\*, **Jos P. Kanning\***, Marieke J. H. Wermer, Niels H. Chavannes, Mattijs E. Numans, Ynte M. Ruigrok, Erik W. van Zwet, Hein Putter, Ewout W. Steyerberg, Rolf H. H. Groenwold. Developing Clinical Prediction Models Using Primary Care Electronic Health Record Data: The Impact of Data Preparation Choices on Model Performance. *Front Epidemiol.* 2022 Jun 2;2:871630.
- Hendrikus J. A. van Os, **Jos P. Kanning**, Tobias N. Bonten, Margot M. Rakers, Hein Putter, Mattijs E. Numans, Ynte M. Ruigrok, Rolf H. H. Groenwold, Marieke J. H. Wermer. Cardiovascular Risk Prediction in Men and Women Aged Under 50 Years Using Routine Care Data. *J Am Heart Assoc.* 2023 Apr 4;12(7):e027011.
- **Jos P. Kanning**, Hendrikus J.A. van Os, Margot Rakers, Marieke J.H. Wermer, Mirjam I. Geerlings, Ynte M. Ruigrok. Predicting Aneurysmal Subarachnoid Haemorrhage using Routine Care Data: A Comparison with other Stroke Types. *PLoS One.* 2024 May 31;19(5):e0303868.
- **Jos P. Kanning**, Shahab Abtahi, Christian Schnier, Olaf H. Klungel, Mirjam I. Geerlings, Ynte M. Ruigrok. Prescribed Drug Use and Aneurysmal Subarachnoid Haemorrhage Incidence: A Drug-Wide Association Study. *Neurology.* 2024 Jun 25;102(12):e209479.
- Bakker MK, **Kanning JP**, Abraham G, Martinsen AE, Winsvold BS, Zwart JA, Bourcier R, Sawada T, Koido M, Kamatani Y, Morel S. Genetic risk score for intracranial aneurysms: prediction of subarachnoid hemorrhage and role in clinical heterogeneity. *Stroke.* 2023 Mar;54(3):810-8.
- van Os HJ, **Kanning JP**, Ferrari MD, Bonten TN, Kist JM, Vos HM, Vos RC, Putter H, Groenwold RH, Wermer MJ. Added Predictive Value of Female-Specific Factors and Psychosocial Factors for the Risk of Stroke in Women Under 50. *Neurology.* 2023 Aug 22;101(8):e805-14.

### Submitted Manuscripts

- Vita M. Klieverik, **Jos P. Kanning**, Ina L. Rissanen, Kristiina Rannikmäe, Amy E. Martinsen, Bendik S. Winsvold, Mirjam I. Geerlings, Ynte M. Ruigrok. Development and external validation of the SMA2SH2ERS risk prediction model for aneurysmal subarachnoid hemorrhage in the general population.
- **Jos P. Kanning**, Junfeng Wang, Shahab Abtahi, Mirjam I. Geerlings, Ynte M. Ruigrok. Identifying novel risk factors for aneurysmal subarachnoid hemorrhage using machine learning.
- Junfeng Wang, Haoyue Wang, **Jos P. Kanning**, Shengfeng Wang. Comparison of different strategies in using Lasso in clinical prediction models for rare outcomes: a simulation study.

### **In Preparation**

- Ina Rissanen, Vita M. Klieverik, **Jos P. Kanning**, Mirjam I. Geerlings, Ynte M. Ruigrok. Sex-specific SMA2SH2ERS -risk prediction models for aneurysmal subarachnoid haemorrhage.
- **Jos P. Kanning**, Patrick C. Souverein, Olaf H. Klungel, Mirjam I. Geerlings, Ynte M. Ruigrok, Shahab Abtahi. Associations between lisinopril and amlodipine use and risk of aneurysmal subarachnoid haemorrhage: A UK population-based cohort study.

## ABOUT THE AUTHOR

Jos Peter Kanning was born on February 1, 1995, in Groningen, the Netherlands. He completed his pre-university education (VWO) at Het H.N. Werkman College in Groningen in 2014. He then obtained his Bachelor's degree in Psychology at the University of Groningen in 2016 and his Master's degree in Cognitive Neuroscience with honors at Maastricht University in 2018.

Jos worked as a data scientist for two years through a traineeship program, where he was primarily assigned to the international software and services company TDGlobal. In this role, Jos was responsible for diverse data science projects, including projects in Singapore (Barghest Building Performance) and Indonesia (Telkomsel).

Jos started his PhD at the Department of Neurology and Neurosurgery at the University Medical Center Utrecht, the Netherlands. In a project supervised by Prof. dr. Ynte Ruigrok, dr. Mirjam Geerlings, and later dr. Shahab Abtahi, he applied big data techniques to stroke research, with a focus on subarachnoid haemorrhage. During his PhD, Jos completed a Master's in Epidemiology, and one of his papers was selected for the American Academy of Neurology's (AAN) press release program.

Jos is currently working as a data scientist at Amsterdam UMC, focusing on optimizing the GP-based ANHA database and consulting with researchers to facilitate medical research.

## DANKWOORD

Deze thesis was niet tot stand gekomen zonder de hulp van velen. Het is onmogelijk om iedereen afzonderlijk te bedanken, maar ik wil een aantal personen en groepen in het bijzonder benoemen.

Allereerst mijn (co)promotoren en begeleiders. Ynte, ik had me geen betere promotor kunnen wensen. Je constante steun, zowel academisch als persoonlijk, heeft enorm veel voor me betekend. Je was altijd bereikbaar voor vragen, en vaak had ik binnen minuten een antwoord. Ondanks dat het grootste deel van mijn PhD zich afspeelde tijdens de COVID-pandemie, zorgde jouw betrokkenheid – van online pubquizzes en SAB-diners tot het regelen van fysieke werkplekken – ervoor dat ik me nooit eenzaam voelde. Heel veel succes in je nieuwe rol als professor! Ik ben ervan overtuigd dat we in de toekomst nog vaker zullen samenwerken. Mirjam, jouw expertise op het gebied van epidemiologie heeft deze thesis verrijkt. Hoewel je kritische vragen soms uitdagend waren, resulteerden ze altijd in een beter eindresultaat. Dank voor het openen van mijn academische wereld, van interessante contacten bij BRAINLAB tot mijn nieuwe baan bij het Amsterdam UMC. Ik kijk ernaar uit om samen verder te werken aan een sterkere wetenschappelijke wereld. Shahab, ook al sloot je later aan bij het team, jouw bijdrage resulteerde in enkele van de mooiste artikelen in mijn these. Je kennis, expertise en positieve houding maakten onze samenwerking ontzettend aangenaam. Ik zie ernaar uit om in de toekomst verder samen onderzoek te doen.

Daarnaast wil ik de leden van de beoordelingscommissie bedanken: Prof. dr. A. Abu-Hanna, Prof. dr. H. Gardarsdottir, Prof. dr. F.H. Rutten, Prof. dr. E.W. Steyerberg, en Dr. M. Uyttenboogaart. Dank voor jullie tijd en inzet. Sommigen van jullie heb ik al mogen ontmoeten; anderen hoop ik in de toekomst te spreken en mee samen te werken.

Mijn collega's wil ik ook graag bedanken: De SAB Groep, ondanks de beperkingen door COVID heb ik veel geleerd van deze diverse en gezellige groep. Het was fijn om op de valreep nog een congres bij te wonen, inclusief de beroerte-borrel. Mark, bedankt voor de prettige samenwerking en de goeie dosis humor. Vita, onze samenwerking groeide al snel uit tot een vriendschap. Het was fijn om niet de enige Groninger te zijn. Hine, dank voor jouw enthousiaste ideeën en talent om altijd weer nieuwe connecties en middelen te vinden. Ook bedankt voor de goeie lunches. Pieter, Frank en Hugo, jullie maakten thuiswerken een stuk gezelliger. Ik ben dankbaar dat ik vrienden ook collega's mocht noemen.

Ook mijn studievrienden wil ik bedanken. Harmen, Fadime en CP, ondanks dat psychologie een ver verleden lijkt, ben ik blij dat we elkaar nog steeds zien. CP, in het bijzonder dank voor de vele inspirerende wandelingen in het Wilhelminapark. Lennie en Caoimhe, bijzonder dat we allemaal in de academische wereld zijn beland. Dank voor de fijne PBL-sessies, vakanties en gezelligheid. Megh, Joey, Valat, Victor, Lotte, Anne Jasmijn, Chiara, Aline, Eric en alle andere leden van de *party office*: bedankt voor de goede discussies, feestjes en het creëren van een werkplek waar ik altijd met plezier naartoe ging. Emma, jouw enthousiasme en nieuwsgierigheid werken aanstekelijk. Ontzettend bedankt dat ik jouw paranimf mocht zijn en dat we niet alleen collega's, maar ook vrienden zijn gebleven.

Anouk en Marina, bedankt dat jullie mijn paranimfen willen zijn. Jullie maakten mijn start in Utrecht, midden in de pandemie, een stuk aangenamer. Grotendeels dankzij jullie kijk ik terug op deze periode als een waardevolle tijd. Ik kijk ernaar uit om deze PhD samen met jullie af te ronden.

Tot slot wil ik mijn dierbaren bedanken. Mijn ouders, Klaas en Fenny, als nuchtere Groningers zeggen we het te weinig, maar ik ben dankbaar voor jullie steun en vertrouwen. Ik had me geen betere ouders kunnen wensen. Janka, bedankt voor je onvoorwaardelijke steun de afgelopen jaren. Dank dat je altijd interesse toonde in mijn werk en me bleef motiveren om het beste uit mezelf te halen. Sinds het einde van mijn PhD wonen we samen, wat dit afscheid tevens een prachtig nieuw begin maakt. *Köszönöm, hogy vagy!*



**UMC Utrecht**



**Universiteit Utrecht**

ISBN 978-94-93406-31-5