

Machine learning for predicting diverse stroke outcomes: binary, multi-class, and time-to-event

Josline Adhiambo Otieno



UMEÅ UNIVERSITY

This work is protected by the Swedish Copyright Act (Act 1960:729)

ISBN: 978-91-8070-890-6 (print)

ISBN: 978-91-8070-891-3 (pdf)

ISSN: 1100-8989

Statistical studies no. 61

Cover photo: Ulrika Sahlén

Cover reproduced with permission

Electronic version available at: <http://umu.diva-portal.org/>

Printed by: Scandinavian Print Group, Skarpnäck, 2026

*In loving memory of my sweet mother, gone too soon to witness my unfolding,
yet ever present in the quiet wisdom that guides my every step.*



UMEÅ UNIVERSITY

Machine learning for predicting diverse stroke outcomes: binary, multi-class, and time-to-event

Josline Adhiambo Otieno

Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå 2026

Doctoral Thesis
Department of Statistics
Umeå School of Business, Economics, and Statistics
Umeå University
SE-901 87 Umeå, Sweden

Copyright © 2026 by Josline Adhiambo Otieno (josline.otieno@umu.se)
Statistical Studies No. 61
ISBN: 978-91-8070-891-3 (digital)
ISBN: 978-91-8070-890-6 (printed)
ISSN: 1100-8989
Electronic version available at <http://umu.diva-portal.org/>

Printed by: Scandinavian Print Group, Skarpnäck, 2026
Umeå, Sweden 2026

Contents

List of papers	v
Abstract	vi
Populärvetenskaplig sammanfattning	viii
Muhtasari (Summary in Swahili)	x
Acknowledgements	xii
1 Introduction	1
2 Objectives of this thesis	3
3 Stroke and stroke registers	4
3.1 The Swedish Stroke Register (Riksstroke)	4
3.2 Sentinel Stroke National Audit Programme (SSNAP)	5
4 Theoretical background	5
4.1 Types of prediction outcomes	5
4.2 Models	6
4.2.1 Statistical models	6
4.2.2 Tree-based ensemble methods	9
4.2.3 Deep-learning models	10
4.2.4 Other methods	12
4.3 Model development and Validation	12
4.4 Evaluation measures	13
4.4.1 Binary classification metrics	13
4.4.2 Multiclass classification metrics	15
4.4.3 Survival outcome metrics	16
4.5 Model explainability	17
5 Summary of papers	17
5.1 Paper I	17
5.2 Paper II	18
5.3 Paper III	20
5.4 Paper IV	21

6 Ethical considerations	23
7 Conclusion and further research	23
Papers I-IV	

List of papers

The thesis is based on the following papers:

- I. Wang, W., Otieno, J. A., Eriksson, M., Wolfe, C. D., Curcin, V., & Bray, B. D. (2023). Developing and externally validating a machine learning risk prediction model for 30-day mortality after stroke using national stroke registers in the UK and Sweden. *BMJ open*, 13(11), e069811.
- II. Otieno, J. A., Häggström, J., Darehed, D., & Eriksson, M. (2024). Developing machine learning models to predict multi-class functional outcomes and death three months after stroke in Sweden. *PLoS One*, 19(5), e0303287.
- III. Otieno, J. A., Häggström, J., & Eriksson, M. (2026). A comprehensive simulation study evaluating the predictive performance of Cox proportional hazards model and machine learning methods for time-to-event data. *Submitted Manuscript, under review*.
- IV. Otieno, J. A., Liu X., & Eriksson, M. (2026). Predictive performance at different time horizons in the presence of competing risks: machine learning versus statistical time-to-event models. *Working paper*.

Abstract

This thesis aimed to improve the clinical prediction of short- and long-term outcomes after stroke by developing, evaluating, and comparing classical statistical and modern machine learning models. Using large, high-quality national stroke registries, specifically the Swedish national stroke register (Riksstroke) and the Sentinel Stroke National Audit Programme (SSNAP), the thesis investigates whether advanced machine learning models offer added value in predicting clinical outcomes compared to traditional models and importantly in which contexts such improvements occur and are clinically meaningful. It addresses several key gaps in the literature, including the lack of external validation for many prediction models across different health systems, the limited research on predicting multi-class functional outcomes, few comprehensive simulation-based evaluations of prediction models under different realistic data conditions, and the need for multiple-horizon evaluations of competing risk prediction models to support fair model selection.

The first paper of the thesis evaluated models for predicting short-term mortality after stroke using large national registries from two European countries (Riksstroke and SSNAP). The results showed that machine learning models offered only modest performance gains over well-specified logistic regression models, demonstrating that traditional approaches remain competitive, especially when predictors are limited and the dataset is structured.

The second paper performed multi-class prediction of functional outcomes three months after stroke, a clinically important, yet methodologically challenging outcome. All models demonstrated similar overall accuracy. However, machine learning, particularly neural networks and gradient-boosting models, indicated clearer advantages over multinomial logistic regression in distinguishing the functional dependence category. Using explainability approaches such as SHapley Additive exPlanations, the study demonstrated that complex models can still provide interpretable insights into the contribution of risk factors in predictions.

The third paper comprehensively evaluated the classical Cox proportional hazards model and machine learning models for predicting time-to-event outcomes using both simulation and real-world registry data. The Cox regression model performed better when its assumptions were satisfied or when the violations of the assumptions were minimal, while

tree-based models demonstrated better performance in the presence of non-linearity, misspecification, or large number of noise variables.

The final paper compared multiple modeling frameworks for predicting competing risks at multiple evaluation time points (horizons). The results showed that the performance of the models depended on the dataset and the evaluation time point, and no model consistently performed the best. Tree-based and deep-learning models achieved better discrimination when events were common, while pseudo-observation-based and Fine-Gray models showed better calibration, especially at longer horizons.

In summary, the thesis demonstrated that model choice should be guided not by popularity but by data structure, clinical context, and evaluation using different metrics and at multiple time horizons. Traditional and machine learning models each have strengths and rigorous validation, calibration assessment, and explainability are crucial for trustworthy clinical prediction.

KEYWORDS: Predictive modeling, Machine learning, Survival analysis, Stroke.

Populärvetenskaplig sammanfattning

Denna avhandling undersöker vilket mervärde avancerad maskininlärning kan ge vid prediktion av utfall efter stroke, jämfört med etablerade statistiska metoder, samt under vilka förutsättningar detta mervärde uppstår.

Med hjälp av det svenska kvalitetsregistret för stroke (Riksstroke), en brittisk motsvarighet (SSNAP) och simulerade data behandlar avhandlingen flera viktiga frågor. Hit hör bristen på extern validering av prediktionsmodeller i olika populationer och sjukvårdssystem, den begränsade forskningen kring prediktion av patientutfall med flera kategorier, få heltäckande och rättvisa simuleringsbaserade jämförelser mellan traditionella och maskininlärningsbaserade modeller under realistiska villkor, samt behovet av utvärderingar över flera tidshorisonter för konkurrerande risk-modeller för att möjliggöra rättvis och tillförlitlig modellselektion.

Artikel I utvecklade och externt validerade en modell för att predicera 30-dagars mortalitet efter stroke inom två nationella sjukvårdssystem. Fördelen med maskininlärning jämfört med logistisk regression var marginell, vilket visar att traditionella modeller kan prestera likvärdigt i strukturerade datamaterial.

Artikel II utvärderade om maskininlärningsmodeller förbättrar prediktionen av patienters funktionsstatus tre månader efter stroke, klassificerat som oberoende, beroende eller avliden. Den övergripande prestandan var liknande för alla modeller, men maskininlärningsmetoderna var bättre på att identifiera patienter som överlever men förblir beroende av daglig hjälp vilket är ett kliniskt viktigt utfall för rehabiliteringsplanering och resursfördelning. Studien visade också att förklaringsmetoder kan ge tolkbara insikter i hur riskfaktorer bidrar till prediktionerna.

Artikel III jämförde en klassisk överlevnadsmodell med vanligt förekommande maskininlärningsmodeller med hjälp av både registerdata och simulerade dataset. Resultaten visade att den traditionella överlevnadsmodellen presterar bättre när dess antaganden uppfylls eller när avvikelserna är små, medan maskininlärningsmodeller uppnår bättre prestanda när sambanden i datan är icke-linjära, när modellantaganden bryts eller när modellen innehåller många brusvariabler.

Artikel IV utvärderade olika modeller i situationer med konkurrerande utfall (t.ex. risk för ny stroke samtidigt som risk att avlida). Resultaten visade att ingen enskild modell konsekvent presterade bäst vid alla utvärderingstidpunkter. Modellprestandan varierade både över tid och

mellan dataset.

Sammanfattningsvis visar denna avhandling att prediktion av utfall efter stroke inte handlar om att hitta en enskild ”bästa” modell, utan om att välja modeller som är anpassade till datan och det kliniska sammanhanget, och att validera dem i olika populationer.

Muhtasari (Summary in Swahili)

Tasnifu hii imelenga kuboresha utabiri wa kitabibu wa matokeo ya muda mfupi na muda mrefu baada ya kiharusi kwa kuunda, kutathmini, na kulinganisha modeli za kitakwimu za kia jadi na modeli za kisasa za ujifunzaji wa mashine. Kwa kutumia rejista kubwa na bora za kitaifa za wagonjwa wa kiharusi, hasa rejista ya kitaifa ya kiharusi ya Uswidi (Riksstroke) na Sentinel Stroke National Audit Programme (SSNAP), tasnifu hii inachunguza iwapo modeli za hali ya juu za ujifunzaji wa mashine zina thamani ya ziada katika kutabiri matokeo ya kitabibu ikilinganishwa na modeli za kitamaduni, na muhimu zaidi, katika muktadha gani maboresho haya hutokea na kuwa na maana ya kitabibu. Inashughulikia mapungufu kadhaa muhimu katika fasihi, ikiwemo ukosefu wa uthibitisho wa nje wa modeli nyingi za utabiri katika mifumo tofauti ya afya, utafiti mdogo juu ya utabiri wa matokeo ya kiutendaji ya aina nyingi (multi-class functional outcomes), ukosefu wa tathmini za kina zinazotegemea usanisi (simulation-based evaluations) za modeli za utabiri chini ya hali mbalimbali halisi za data, na haja ya tathmini za muda mbalimbali (multiple-horizon evaluations) za modeli za utabiri wa hatari zinazoshindana (competing risk prediction models) ili kuwezesha uteuzi mzuri wa modeli.

Karatasi ya kwanza ya tasnifu ilitathmini modeli za kutabiri vifo vya muda mfupi baada ya kiharusi, ikitumia rejista kubwa za kitaifa kutoka nchi mbili za Ulaya (Riksstroke na SSNAP). Matokeo yalionyesha kuwa modeli za ujifunzaji wa mashine zilitoa uboreshaji mdogo tu wa utendaji ikilinganishwa na modeli za logistic regression zilizobainishwa vizuri, na hivyo kuonyesha kuwa mbinu za kitamaduni bado zinashindana vyema, hasa pale ambapo vihashiria ni vichache na seti ya data imepangwa vizuri.

Karatasi ya pili ilifanya utabiri wa multi-class wa matokeo ya kiutendaji miezi mitatu baada ya kiharusi, matokeo ambayo ni muhimu kitabibu lakini yenye changamoto za kimbinu. Modeli zote zilionyesha usahihi wa jumla unaofanana. Hata hivyo, modeli za ujifunzaji wa mashine, hasa neural networks na modeli za gradient boosting, zilionyesha manufaa ya wazi zaidi kuliko multinomial logistic regression katika kutofautisha kundi la utegemezi wa kiutendaji. Kwa kutumia mbinu za ufafanuzi kama SHapley Additive exPlanations (SHAP), utafiti ulionyesha kwamba hata modeli changamano bado zinaweza kutoa ufahamu unaoeleweka kuhusu mchango wa visababishi hatari katika utabiri.

Karatasi ya tatu ilitathmini kwa kina modeli ya kitamaduni ya Cox

proportional hazards na modeli za ujifunzaji wa mashine kwa ajili ya kutabiri matokeo ya muda hadi tukio, ikitumia data za usanisi na data halisi za rejista. Modeli ya Cox regression ilifanya vizuri zaidi pale ambapo dhana zake zilitimizwa au pale ambapo uvunzaji wa dhana hizo ulikuwa mdogo, ilhali modeli zinazotegemea miti ya maamuzi zilionyesha utendaji bora zaidi pale kulipokuwepo na kutokuwiana kwa mstari (non-linearity), ubainishaji usio sahihi (misspecification), au idadi kubwa ya vigezo visivyo na taarifa (noise variables).

Karatasi ya mwisho ililinganisha miundo mbalimbali ya modeli kwa ajili ya kutabiri matukio yanayoshindana (competing risks) katika nyakati tofauti za tathmini (horizons). Matokeo yalionyesha kuwa utendaji wa modeli uliathiriwa na seti ya data na muda wa tathmini, na hakuna modeli iliyofanya vizuri kwa uthabiti katika nyakati zote. Modeli zinazotegemea miti ya maamuzi na ujifunzaji wa kina (deep learning) zilipata ubora zaidi wa kutenganisha matukio pale ambapo matukio yalikuwa ya mara kwa mara, ilhali modeli zinazotumia pseudo-observations pamoja na modeli za Fine–Gray zilionyesha upangaji bora (calibration), hasa katika tathmini za muda mrefu.

Kwa muhtasari, tasnifu hii imeonyesha kwamba uchaguzi wa modeli haupaswi kuongozwa na umaarufu wake, bali na muundo wa data, muktadha wa kitabibu, na tathmini inayotumia vipimo mbalimbali na katika vipindi tofauti vya muda. Modeli za kitamaduni na modeli za ujifunzaji wa mashine kila moja ina nguvu zake, na uthibitishaji makini, tathmini ya upangaji (calibration), pamoja na ufafanuzi wa utabiri ni muhimu ili kuhakikisha utabiri wa kitabibu unaoaminika.

Acknowledgments

I sincerely appreciate my main supervisor, Marie Eriksson, whose guidance, mentorship, and support have been central to the completion of this thesis. The time and effort you have invested in me have helped me grow both professionally and personally. My sincere thanks also go to my co-supervisors, Jenny Häggström and Xijia Liu, for their valuable input and constructive discussions during our meetings. I also thank my co-authors for their excellent collaboration and support. To my colleagues at the Department of Statistics; thank you for making my time here both enriching and memorable.

To my son, Ian, you are my greatest inspiration. This achievement is dedicated to you with all my heart. Last but not least, a special thank you goes to my siblings, Joan, Charles, Veron, and Colphax, for their constant motivation throughout this journey.

Umeå, March 2026

Josline Adhiambo Otieno

1 Introduction

Prediction models are increasingly used in healthcare to support clinical decision-making by identifying high-risk patients, guiding diagnosis and prognosis, as well as facilitating patient guidance and counseling (Alowais et al., 2023). Due to the complexity of diseases such as stroke (including primary prevention, patient characteristics, comorbidities, diagnosing, acute management, secondary prevention, and organization of stroke care), the application of traditional methods such as logistic regression (LR) and Cox regression models may be insufficient to predict short- or long-term outcomes after stroke. For this reason, modern machine learning (ML) models such as support vector machines (SVM), random survival forests (RSF), eXtreme Gradient Boosting (XGBoost), among others, are becoming popular (Alhumaidi et al., 2025) as an alternative approach, as they can capture complex relationships especially in high-dimensional datasets, and offer a flexible way to learn patterns directly from data without relying on model assumptions (Srećković et al., 2022).

ML models including SVM, artificial neural networks (ANNs), and eXtreme Gradient Boosting (XGBoost) have shown promising results in different clinical classification tasks (Trabassi et al., 2022; Mufti et al., 2019; Wang et al., 2022b). Predictive modeling of survival outcomes is based on methods specifically designed to handle time-to-event data, which often include censoring. More recently, ML models including RSF, survival gradient boosting models (especially XGBoost), and neural networks (DeepSurv) have been commonly used for modeling survival data. (Teshale et al., 2024; Huang et al., 2023, 2025). Monterrubio-Gómez et al. (2024) provides a comprehensive review of competing risk models in survival analysis such as RSF, boosting models, and deep-learning models.

Despite the widespread adoption of ML models in clinical practice, several well-known challenges remain that can threaten their reliability, fairness, and trustworthiness. A common concern is overfitting, where a model performs extremely well on the training data but fails to generalize to new cases (Lynch and Liston, 2018). ML models are often considered ‘black boxes’, which hinders interpretability and transparency (Vayena et al., 2018). Another issue is that ML models can inherit the same biases as the data on which they are trained (Kumar et al., 2025).

Previous research on the performance of ML models compared to traditional approaches such as LR has produced inconsistent findings. An

evaluation of ML models on 121 public datasets reported that ML models often outperform LR in classifying categorical outcomes (Fernández-Delgado et al., 2014). However, two systematic reviews found no clear advantage of ML models over LR in predicting binary outcomes (Christodoulou et al., 2019; Mahmoudi et al., 2020). This highlights the need for more rigorous and context-specific evaluations of ML models in clinical prediction.

Concerns about model development and reporting further complicate the adoption of ML-based clinical prediction models. A systematic review of ML models for predicting stroke outcomes identified several flaws, including poor adherence to reporting standards and the absence of publicly available models for external validation (Wang et al., 2020). Moreover, many ML models rely on variables that are not consistently recorded across clinical registries (Wang et al., 2022b). An additional review of the opportunities and challenges in developing risk prediction models using electronic health record data highlights the need to evaluate the added value that more complex models may offer compared to traditional approaches (Goldstein et al., 2016). Although modern ML models may offer potential improvements, their accuracy, interpretability, and clinical relevance must be thoroughly validated and compared across different datasets (Zihni et al., 2020; Huang et al., 2023; Stahl, 2024).

In survival analysis, accurate prediction is crucial, yet selecting an appropriate model remains challenging due to the complexities of real-world survival data, such as censoring, nonlinear patterns, violations of the proportional hazards assumption, and noisy predictors. Consequently, identifying the best-suited model for a given endpoint requires comparing methods under realistic conditions (Kantidakis et al., 2021). Simulation studies facilitate such comparisons (Morris et al., 2019), but previous work has been criticized for unclear or poorly reported simulation designs, including insufficient details on data-generating mechanisms and performance metrics (Smith et al., 2022). Additionally, because absolute risk changes during follow-up, evaluating competing risks models at a single fixed time point can fail to reflect time-dependent variations in measures such as calibration, discrimination, and overall accuracy.

This thesis focuses on the development, evaluation, and reporting of the performance and feasibility of modern ML models compared to traditional regression models. Paper I develops and externally validates

a generalizable prediction model for predicting 30-day stroke mortality (binary outcome) using large nationwide stroke registers in the UK and Sweden. Previous research has largely focused on the prediction of so-called ‘hard endpoints’, such as survival, readmission, and other binary outcomes (Wang et al., 2020, 2022b, 2023; Goyal et al., 2021). Paper II aimed to develop and evaluate the performance and explainability of models in predicting multi-class outcomes after stroke using routinely-collected data. Paper III presents a comprehensive simulation study to evaluate the predictive performance of models for time-to-event data, focusing on realistic challenges such as sample size, censoring, model misspecification, and noisy variables. Finally, Paper IV comparatively evaluates models for competing risk prediction across multiple time horizons.

The remainder of the thesis is structured as follows. Section 2 outlines the objectives of the thesis. Section 3 describes the datasets used in the analyses. Section 4 presents the theoretical background of the thesis, including definitions of different types of outcomes, a brief overview of the prediction models included in the thesis, and the key steps involved in developing and validating a prediction model. Section 5 provides a summary of the papers. Section 6 highlights ethical considerations in this study, followed by a summary of the contributions of the thesis and possible areas for future research in Section 7.

2 Objectives of this thesis

The primary objective of this thesis was to develop and evaluate modern ML models, assess their performance and feasibility compared to traditional regression models in predicting stroke outcomes. Specifically,

- to adapt and externally validate pre-trained ML models to predict survival after stroke.
- to assess their performance and explainability in predicting functional outcomes (multi-class outcomes) after stroke.
- to evaluate their performance in predicting time-to-event outcomes under different simulation scenarios.
- to evaluate their predictive performance in the presence of competing risks across multiple prediction horizons.

3 Stroke and stroke registers

Stroke is a medical condition that occurs when blood flow to the brain is blocked or reduced (ischemic stroke) or a blood vessel ruptures (hemorrhagic stroke), damaging brain cells. It is a complex disease (including risk factors, stroke characteristics, primary prevention, comorbidities, acute management, secondary prevention, and organization of stroke care), which makes it a suitable example for studies of predictive modeling approaches that can capture non-linear relationships and interactions between variables. Stroke is expected to remain a major cause of mortality and long-term disability worldwide after other cardiovascular diseases, with ischemic stroke the most common subtype (Tsampasian and Bloomfield, 2025; Feigin et al., 2025). Despite increasing crude epidemiological indicators associated with stroke, age-standardized stroke mortality and prevalence are generally declining (Chong et al., 2025; Tsampasian and Bloomfield, 2025). In Sweden, a decline in incidence and an improvement in survival have been observed between 2005 and 2018 (Eriksson et al., 2021). However, more than 20,000 cases occur yearly, of which 20% are discharged to an assisted living facility, 15% die, and 14% depend on others to perform activities in daily living 3 months after stroke (Riksstroke yearly report 2024). In addition to its serious effects on patients, stroke care is costly and places a heavy burden on families and countries (Lekander et al., 2017). Accurate prediction of stroke outcomes and the identification of key prognostic features are essential for guiding individualized treatment and secondary prevention strategies, informing clinical decision-making, and optimizing the allocation of healthcare resources.

3.1 The Swedish Stroke Register (Riksstroke)

All four papers in this thesis used empirical data from Riksstroke. Riksstroke is the Swedish national quality register for stroke care and is the world's longest-running national stroke quality register, established in 1994 (Asplund et al., 2011). It includes all hospitals (around 72 hospitals) in Sweden that admit acute stroke (ischemic, primary intracerebral hemorrhagic, or unspecified type of stroke) patients. The estimated coverage is over 90% of all stroke patients treated in hospitals. Currently, about 20,000 to 23,000 new events are registered each year. The registry in-

cludes information from the entire chain of stroke care, including primary prevention, acute management, rehabilitation measures, secondary prevention, and family and community support. Basic patient characteristics (such as age, sex, living conditions, history of previous stroke, and comorbidities), diagnosis, level of consciousness on arrival, pharmaceutical treatment, complications, and sequence of care (including type of stroke care organization and department) are recorded at the hospital. A questionnaire-based follow-up that collects patient-reported outcomes is also conducted three months and sometimes one year after stroke. Detailed information about the register can be found on the Riksstroke’s website (www.riksstroke.org).

3.2 Sentinel Stroke National Audit Programme (SSNAP)

Data from SSNAP were used in Paper I for model development and internal validation. SSNAP is a national stroke care quality improvement audit commissioned by the Healthcare Quality Improvement Partnership and currently hosted at King’s College London. The registry collects prospective patient-level data on over 90% adults admitted to National Health Service hospitals with acute stroke (ischemic or primary intracerebral hemorrhage) in England, Wales and Northern Ireland. It includes a clinical audit, which records patient care, rehabilitation, and outcomes up to 6 months after stroke, and organizational audits, which describe how stroke services are structured and staffed. Data are collected and validated by clinical teams and entered into the SSNAP database via a secure web-based platform. Initial data collection began in early 2013 and continues as an ongoing national clinical audit of stroke care. More information about the registry is provided on the SSNAP website (www.strokeaudit.org).

4 Theoretical background

4.1 Types of prediction outcomes

Prediction models can target different types of outcomes depending on the clinical questions and data structure. Outcome variables are typically classified as continuous (e.g., cholesterol level) or categorical (e.g., presence or absence of disease recurrence) (Hastie et al., 2009). Contin-

uous outcomes are commonly predicted using linear regression models, whereas categorical outcomes, binary or multiclass, are predicted using classification models, such as logistic regression and SVM.

Many clinical questions, including those focused on stroke, extend beyond identifying who is at risk, and also require predicting when the event will occur. When time-to-event is part of the outcome, researchers use survival analysis (Kleinbaum and Klein, 2011). These methods account for censoring, which arises when the event has not occurred by the end of follow-up.

Sometimes, patients may experience different mutually exclusive events, where the occurrence of one event prevents the occurrence of the event of interest. These are known as competing risks (Coemans et al., 2022) and require appropriate statistical methods to obtain accurate predictions of event-specific risks.

4.2 Models

4.2.1 Statistical models

In many clinical studies, logistic regression (LR) is the most widely used statistical model for binary classification tasks based on a set of predictors (Wang, 2023; Dreiseitl and Ohno-Machado, 2002; Badawy et al., 2023). It is a standard generalized linear model that uses a logit link function. LR model models the relationship between binary outcome, $Y \in \{0, 1\}$ and a set of p predictors, $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$, $i = 1, \dots, n$ patients. The log-odds of the probability of the event, $Y = 1$, as a linear function of predictors is expressed as;

$$\log\left(\frac{P(Y = 1 \mid \mathbf{X}_i)}{1 - P(Y = 1 \mid \mathbf{X}_i)}\right) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip},$$

which is equivalent to;

$$P(Y = 1 \mid \mathbf{X}_i; \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}.$$

The model parameters, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$, are estimated by minimizing the negative log-likelihood function;

$$-\ell(\boldsymbol{\beta}) = -\sum_{i=1}^n [Y_i \log \pi(\mathbf{x}_i) + (1 - Y_i) \log(1 - \pi(\mathbf{x}_i))],$$

where $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ are the observed values, and $\pi(\mathbf{x}_i) = P(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\beta})$.

Multinomial LR (mLR) is an extension of binary LR that handles outcomes with more than two categories. An MLR model estimates the probability of each outcome category c of Y , given a set of p predictors. The model is defined as;

$$P(Y = c \mid \mathbf{X}_i; \boldsymbol{\beta}) = \frac{\exp(\beta_{0c} + \beta_{1c}X_{i1} + \dots + \beta_{pc}X_{ip})}{1 + \sum_{q=1}^{C-1} \exp(\beta_{0q} + \beta_{1q}X_{i1} + \dots + \beta_{pq}X_{ip})},$$

where $c = \{1, \dots, C\}$ ($C = 3$ levels: e.g., independent, dependent, dead), and β 's are the model parameters also estimated by minimizing the negative log-likelihood function.

In survival analysis, the Kaplan-Meier (KM) estimator and the Cox PH (Cox regression) model are used to determine the survival rate or factors affecting survival. KM is a non-parametric method for estimating the survival function when time-to-event data include right-censoring (Kaplan and Meier, 1958). It constructs a stepwise survival curve that decreases only at observed event times. For ordered event times $t_1 < t_2 < \dots < t_k$ with d_j events at time t_j , and n_j patients at risk just before t_j , the estimator is expressed as;

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

The Cox PH model is a semi-parametric survival regression model relating the time to event (possibly censored) to a set of predictors. The outcome is expressed through the hazard function (Cox, 1972). The model assumes that the predictors have an exponential effect on the hazard and the hazards remain proportional over time. For a patient i with covariates $\mathbf{X}_i \in \mathbb{R}^{p \times 1}$, the hazard function is defined as;

$$h(t \mid \mathbf{X}_i) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{X}_i),$$

where $h_0(t)$ is the baseline hazard and $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ is a vector of regression coefficients, estimated from the training data by minimizing the negative log partial likelihood function;

$$-\log \ell(\boldsymbol{\beta}) = -\sum_{i=1}^n \delta_i \left[\boldsymbol{\beta}^\top \mathbf{X}_i - \log \left(\sum_{j \in \mathcal{R}(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{X}_j) \right) \right], \quad (1)$$

where $\delta_i = 1$ if the patient experiences the event and 0 otherwise, $\mathcal{R}(t_i)$ is the set of patients still at risk just before time t_i .

In the presence of competing risks, cumulative incidence is estimated using statistical models such as the cause-specific (CS) Cox model (Cox, 1972) or the Fine-Gray subdistribution hazards model (Fine and Gray, 1999). The CS Cox model describes how predictors influence the instantaneous risk of failing from a particular event type. For a cause k , the CS hazard function is defined as the instantaneous rate of experiencing event k at time t , given no event has occurred before t ;

$$h_k^{cs}(t | \mathbf{X}_i) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T_i < t + \Delta t, \varepsilon_i = k | T_i \geq t, \mathbf{X}_i)}{\Delta t},$$

where ε_i denotes the type of event experienced by patient i . Each of the CS hazards is modeled using Cox PH model, where failures from other causes are treated as censored (Prentice et al., 1978);

$$h_k^{cs}(t | \mathbf{X}_i) = h_{k0}^{cs}(t) \exp(\boldsymbol{\beta}_k^\top \mathbf{X}_i),$$

where $h_{k0}^{cs}(t)$ is the baseline hazard for cause k , and $\boldsymbol{\beta}_k$ represents the covariate effects on the hazard of cause k , estimated by minimizing the negative log partial likelihood function in Equation (1).

Unlike the CS Cox model, the Fine-Gray model directly models the subdistribution hazard, providing a natural framework for predicting absolute risk while accounting for competing events. The subdistribution hazard for cause k for patient i is defined as;

$$h_k^{fg}(t | \mathbf{X}_i) = \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T_i < t + \Delta t, \varepsilon_i = k | T_i \geq t \cup (T_i < t \cap \varepsilon_i \neq k), \mathbf{X}_i]}{\Delta t},$$

representing the instantaneous risk of failure from cause k among patients who have not yet failed from k , where the risk set includes those who have already experienced competing events. The hazard is also modeled using a Cox PH approach;

$$h_k^{fg}(t | \mathbf{X}_i) = h_{k0}^{fg}(t) \exp(\boldsymbol{\gamma}_k^\top \mathbf{X}_i),$$

with $h_{k0}^{fg}(t)$ the baseline subdistribution hazard and $\boldsymbol{\gamma}_k$ the covariate effects, estimated by minimizing the weighted negative log partial likelihood in Equation (1) (with a modified risk set accounting for competing events through inverse probability of censoring weights).

4.2.2 Tree-based ensemble methods

Ensemble methods are based on the idea that combining the predictions of multiple single models can lead to more accurate and stable results than relying on any individual model alone. Individual models are often sensitive to random fluctuations in the training data or to model-specific biases, whereas aggregation can help average out these effects and improve generalization performance. Depending on whether all models contribute equally to the final prediction, ensemble methods are commonly divided into uniform aggregation and non-uniform aggregation approaches. A representative example of uniform aggregation is bagging, where all base learners are combined with equal weights, while boosting represents non-uniform aggregation by assigning different weights to models trained sequentially, with later models focusing more on previously mispredicted observations. In the literature, tree-based methods are the most popular choice for single models in ensemble learning. It is a non-parametric ML model that predicts an outcome by recursively partitioning the input feature space into non-overlapping regions determined by data-driven split rules (Hastie et al., 2009). Decision trees partition the feature space recursively, and the way this partitioning is performed depends on the type of the target variable. For example, classification trees split the feature space based on misclassification error rates or Gini index, whereas survival trees find the sub-feature space based on log-rank statistic, or likelihood-based criteria.

XGBoost, a gradient-boosting algorithm, is considered to have promising performance in classification tasks (Wang et al., 2022a,b). It is a scalable and regularized implementation of gradient boosting that builds an additive ensemble of decision trees in a sequential manner, where each tree corrects the errors of the previous ensemble, while preventing overfitting (Chen, 2016). For binary outcomes, it constructs an additive ensemble of decision trees, where predictions are formed as the sum of the output of each individual tree and mapped to probabilities using a logistic function. The model is trained by minimizing a regularized objective function that balances data fit and model complexity;

$$\mathcal{L} = \sum_{i=1}^n \ell(y_i, \hat{y}_i) + \sum_{m=1}^M \Omega(f_m); \quad \hat{y}_i = \sum_{m=1}^M f_m(\mathbf{X}_i), \quad (2)$$

where $\ell(\cdot)$ is the logistic loss function for binary classification, M is the

number of trees, f_m is the m^{th} regression tree, $f_m(\mathbf{X}_i)$ is the prediction from f_m for patient i , and $\Omega(f_m) = \varrho V_m + \frac{1}{2}\lambda \sum_{v=1}^{V_m} w_{mv}^2$ is the regularization term (ϱ is the complexity penalty, V_m is the number of leaves in the m^{th} tree, w_{mv} are the weights of the v^{th} leaf in the m^{th} tree, and λ is the L^2 regularization on leaf weights). The objective function is optimized using a second-order Taylor approximation of the loss, using both gradient and Hessian information to drive greedy tree splitting and parameter updates.

For multiclass classification, the loss becomes the softmax cross-entropy over all classes, computed from the model’s predicted probabilities. Each leaf stores a single weight for its class-specific tree, and the regularization term in Equation (2) extends to all leaf weights across all class-specific trees.

For application in survival analysis, the XGBoost objective function is adapted to the Cox PH model by replacing it with the Cox partial likelihood, enabling the model to handle time-to-event outcomes.

The RSF model is one of the most widely used non-parametric, ensemble, tree-based methods for modeling survival endpoints (Teshale et al., 2024; Huang et al., 2023). It extends Breiman’s RF (Breiman, 2001) to right-censored survival data by growing many survival trees on bootstrap samples and using random feature selection at each split (Ishwaran et al., 2008, 2011). At each node, candidate splits are evaluated using a survival-specific splitting criterion, such as the log-rank test. At the terminal nodes, the cumulative hazard function is estimated using the Nelson-Aalen estimator. The ensemble average gives the final cumulative hazard estimate.

In the competing risk setting, the RSF uses a weighted log-rank splitting rule based on Gray’s test (Fine and Gray, 1999) and estimates the cause-specific cumulative incidence function using ensemble estimates of the sub-distribution cumulative hazard (Ishwaran et al., 2014).

4.2.3 Deep-learning models

Deep-learning-based models learn by transforming input data through a sequence of layers, where each layer usually applies a weighted linear combination to the inputs and then introduces nonlinearity through an activation function. During training, the model compares its output to the target using a loss function, and the resulting error is propagated

backward through the network using backpropagation. The computed gradients are then used to update the model parameters through gradient-based optimization.

Artificial neural networks (ANNs) are widely applied to classification tasks in various fields (Molina Menéndez and Parraga Alava, 2024). ANNs are computational models inspired by biological nervous systems, consisting of layers of interconnected neurons that transform input data into output through weighted connections (Zakaria et al., 2014). Each neuron applies a nonlinear activation function, e.g., ReLU or sigmoid, to a weighted sum of its inputs, allowing the network to model complex nonlinear relationships. The network is trained using gradient-based optimization (e.g., backpropagation) to iteratively update weights and minimize a cost function such as softmax cross-entropy (for multiclass classification tasks).

DeepSurv (Katzman et al., 2018), a deep learning-based survival model, has gained popularity and shown promising performance in predicting time-to-event outcomes (Teshale et al., 2024). It extends the classical Cox PH model by using a deep neural network to capture complex relationships between features and survival outcomes. The network weights are optimized by minimizing the negative log partial likelihood with L^2 regularization term to prevent overfitting.

DeepHit is a deep learning model for survival analysis, including competing risk analysis that directly learns the joint distribution of event times and event types without relying on parametric assumptions about hazards or Gaussian processes (Lee et al., 2018). It models time to event in user-defined discrete intervals. In a competing risks setting, DeepHit uses a multitask architecture composed of a shared subnetwork, which extracts features common to all causes, and cause-specific subnetworks, which learn patterns associated with each type of event. The outputs are combined and normalized with a global softmax applied across all event-time combinations, ensuring a valid probability distribution. Summing these probabilities over intervals up to a given time point yields a cause-specific cumulative incidence function for each patient. Training uses a composite loss consisting of a log-likelihood loss that encourages high probability for the observed event while handling censoring appropriately, and a ranking loss that promotes correct risk ranking, assigning higher risk to patients who experience an event earlier.

4.2.4 Other methods

Support vector machine SVM, also known as the maximum margin classifier, is a classical linear classification method that maximally separates the classes of data points (Cortes and Vapnik, 1995). A basic condition for the SVM is that the data are linearly separable in the original feature space. However, in many real-world problems, this assumption does not hold. In such cases, we can apply the kernel trick to kernelize the SVM, implicitly mapping the data into a higher-dimensional feature space where they become linearly separable, thus obtaining a nonlinear classifier. Among various kernel functions, the Radial Basis Function (RBF) kernel is the most commonly used. The RBF kernel is capable of modeling complex nonlinear decision boundaries and is relatively robust to parameter choices, making it a popular and effective default in practice. For multiclass classification tasks, SVMs are commonly extended using strategies such as one-versus-all or one-versus-one (Hsu and Lin, 2002).

Pseudo-observation-based models A pseudo-observation-based model is a statistical approach for handling right-censored time-to-event data (Andersen and Pohar Perme, 2010). Instead of treating censored outcomes as missing, each patient is assigned a pseudo-observation, a jackknife-based estimate that approximates the expected value of a function of interest in a population (e.g., cumulative incidence function or the survival function) when survival data is censored. These pseudo-observations provide continuous outcomes suitable for direct use in standard regression or ML models.

4.3 Model development and Validation

Prediction models are developed through a structured process to ensure that they are reliable and well-calibrated (Efthimiou et al., 2024). The first step is to define the research objective, including the outcome to predict and the target population. Next, verify that the features and outcome variable in the dataset align with the objective. Perform data preparation, e.g., handling missing observations, applying transformations such as one-hot encoding, and selecting features to improve the model’s ability to capture meaningful relationships in the dataset. After data prepro-

cessing, an appropriate predictive model is chosen based on the type of task. The dataset is then partitioned for internal validation using resampling methods such as k-fold cross-validation, or by randomly splitting the dataset into separate training and test sets. The model is trained on the training dataset by iteratively updating its internal parameters to minimize the prediction error. Before final model training, hyperparameters that govern the learning process can be optimized using hyperparameter optimization strategies such as grid search or random search to maximize predictive performance and improve generalization.

In κ -fold cross-validation, the dataset is divided into κ equal folds, and the model is trained κ times by iteratively using $\kappa - 1$ folds for training and the remaining fold for validation. The overall performance of the model is computed by averaging the evaluation metrics across all cross-validation runs. For a random training-test split, the model is trained only on the training set, and its performance is evaluated once on the reserved test set. Finally, the entire modeling process and its results are documented according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) reporting guidelines to ensure transparency, reproducibility, and detailed reporting (Collins et al., 2024).

A prediction model should be externally validated before it is recommended for clinical use (Ramspek et al., 2021). This involves evaluating the model on datasets that are completely separate from the data used during model development, preferably from a different but related population, setting, or time period.

4.4 Evaluation measures

4.4.1 Binary classification metrics

Many standard classification metrics (such as accuracy, precision, recall, F1 score, and specificity) are derived from the confusion matrix, which summarizes the performance of the model by comparing predicted class labels against true class labels (Berrar, 2018). In a binary classification, the matrix contains four outcomes, including True Positives (TP), False Negatives (FN), True Negatives (TN), and False Positives (FP). The evaluation metrics include;

Sensitivity or Recall or True positive rates (TPR) This measures how well a model identifies positive classes correctly, $TPR = \frac{TP}{(TP+FN)}$.

Specificity or True negative rates (TNR) This measures how well a model identifies negative classes correctly, $TNR = \frac{TN}{(TN+FP)}$.

The complements of sensitivity and specificity include false negative rates (FNR) and false positive rates (FPR), respectively.

Precision or Positive predictive value (PPV) This measures how many positive predictions were actually correct, $PPV = \frac{TP}{(TP+FP)}$.

Precision-Recall curves The curve evaluates a binary classifier by plotting precision against recall across all possible decision thresholds, summarizing the trade-off between correctly identifying positives and avoiding false positives. The area under the curve, obtained as the integral under the precision-recall curve, provides a single measure of performance, with higher values indicating better balance between the two measures.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) AUC-ROC is a measure of a model's discrimination. The receiver operating characteristic curve shows how well a binary classification model separates two classes by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across all possible decision thresholds, with the area under the curve summarizing the model's discriminative ability (Fawcett, 2006). The decision threshold ranges from a value that classifies all patients as negative (TPR=0, FPR=0) to one that classifies all patients as positive (TPR=1, FPR=1). At each threshold, the TPR and FPR are computed from the corresponding confusion matrix. A random classifier has an AUC-ROC of 0.5, and any model with a value above 0.5 performs better than random. The higher the AUC-ROC, the better the model's discriminative ability. In binary classification outcomes, AUC-ROC is identical to C-statistics (C-index) (Steyerberg and Vergouwe, 2014).

Brier score (BS) BS is the mean squared error between the predicted probabilities p_i and the observed outcomes y_i (Glenn et al., 1950);

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n (p_i - y_i)^2.$$

Lower values of BS indicate better prediction performance.

Calibration curves This evaluates how well a risk prediction model’s estimated probabilities agree with the observed probabilities. The classical way to construct a calibration curve is to divide the dataset into groups (often deciles) of patients based on their predicted probabilities. For each group, the mean predicted probability is plotted against the observed proportion of outcomes. Alternatively, a smoothed curve (e.g., generated using locally estimated scatterplot smoothing or spline functions) can be used to estimate observed probabilities across the full range of predicted probabilities to avoid grouping (Van Calster et al., 2019). The model is well calibrated when most points or the smoothed curve lie close to the diagonal line, while deviations above or below the line indicate underestimation or overestimation of risk, respectively.

Decision curve analysis A decision curve illustrates the clinical value of a prediction model by plotting its net benefit, $\frac{\text{TP}}{n} - \frac{\text{FP}}{n} \left(\frac{p_t}{1-p_t} \right)$, across a range of threshold probabilities (p_t), the probability at which a patient or clinician would choose treatment (Vickers and Elkin, 2006; Vickers et al., 2008). At each threshold, the model classifies patients as test-positive if their predicted risk is at or above that threshold, thus they would receive treatment, and test-negative if below, implying no treatment. A model is clinically useful only in the range of thresholds where its decision curve provides a higher net benefit than the default strategies of treating all patients or treating none, which are always included for comparison in the decision curve.

4.4.2 Multiclass classification metrics

In multiclass classification, the confusion matrix extends to $C \times C$ matrix, where the diagonal entries represent correct predictions for each class, and off-diagonal entries represent different types of misclassifications across classes. Each class has its own TP, FP, TN, and FN values,

that is, TP_c (correctly classified patients for class c), FN_c (patients of class c misclassified into other classes), FP_c (patients from other classes incorrectly classified as class c), and TN_c (all correct predictions that do not involve class c). Here, the metric formulations from binary classification are applied per class in a one-vs-rest strategy (Grandini et al., 2020). The results can be reported directly or optionally aggregated (macro, micro, or weighted) to obtain overall scores.

4.4.3 Survival outcome metrics

C-index This is the widely used measure of discrimination in survival analysis, quantifying a model’s ability to correctly rank patients by their risk of an event;

$$C_{\text{index}} = P(r_i > r_j \mid t_i < t_j),$$

where (t_i, t_j) and (r_i, r_j) are the pair of observed survival times and predicted risk scores for any randomly selected pair of patients (i, j) , respectively.

The C-index can be estimated using Harrell’s approach (Harrell Jr et al., 1996), which becomes biased under heavy censoring, or Uno’s approach (Uno et al., 2011), which uses inverse probability censoring weighting to adjust for censoring. C-index values closer to 1 indicate better discrimination and 0.5 shows random performance.

Integrated Brier score In survival analysis, BS measures the mean squared error between the observed survival status and the predicted survival probability at a fixed time point τ (Graf et al., 1999). The integrated Brier score assesses the accuracy of predicted survival probabilities over a specified time interval, that is, averaging time-dependent BS over a specified time interval (Graf et al., 1999; Gerds and Schumacher, 2006).

Time-dependent C-index This is a suitable discrimination metric in the presence of competing risks because, unlike the classical C-index, it correctly accounts for competing events and time-varying risk sets, ensuring that only valid comparable pairs contribute to the evaluation. It is the probability that, in a randomly selected pair of patients (i, j) , where at least one has the event of interest before the prediction horizon τ , the patient with the higher predicted cumulative incidence also experiences

the event earlier (Wolbers et al., 2014). Censoring is also handled using inverse probability weighting.

Integrated calibration index This summarizes the average absolute difference between the predicted risks (cumulative incidence) and the corresponding observed probabilities, in the presence of competing risks (Austin et al., 2022).

4.5 Model explainability

Explainability refers to the ability of a model to make its decision-making process transparent and understandable to the user. Unlike ‘black-box’ models, whose internal operations are difficult to interpret, explainable models reveal how input features influence predictions, enabling users to trust and verify results. Explainability can be achieved through inherently interpretable models, such as linear regression or through post-hoc methods applied to models such as SHapley Additive exPlanations (Štrumbelj and Kononenko, 2014) and feature-importance visualizations. These methods quantify the contribution of each input variable, highlight the patterns a model has learned, and present explanations in easy-to-understand formats.

5 Summary of papers

5.1 Paper I

The study aimed to develop and externally validate a cross-national prediction model using large, high-quality stroke registries from the UK (SSNAP) and Sweden (Riksstroke). Data from SSNAP comprised 488,497 stroke patients in England, Wales, and Northern Ireland (2013-2019), and Riksstroke included 128,360 patients in Sweden (2015-2020). Both registries capture nearly all hospitalized stroke cases in their respective countries. Thirteen predictors such as age, sex, comorbidities, level of consciousness, pre-stroke functional status, and type of stroke were included based on clinical relevance and mutual availability. The primary endpoint was 30-day in-hospital mortality in SSNAP and both in-hospital and all-cause mortality in Riksstroke.

Three models, LR, elastic net LR with interaction terms, and XGBoost were developed using the SSNAP dataset (80% training, 20% validation, and temporal validation using 2019 data). They were externally validated on the Riksstroke dataset. Missing data were handled through a combination of rule-based imputation (missing as category, missing indicators) and multiple imputation by chained equations. Model performance was assessed based on Brier scores, AUC-ROC, precision-recall curves, and calibration plots, with clinical utility evaluated using decision curve analysis. Subgroup analyzes by stroke type (ischemic versus hemorrhagic) and sensitivity analyzes using alternative modified Rankin Scale (mRS) coding and all-cause mortality as the outcome were also performed.

Across both registries, 30-day mortality was higher among older patients and those with greater comorbidity burden, worse pre-stroke functional status, hemorrhagic stroke, or reduced consciousness on admission. XGBoost performed best overall but only marginally better than LR models. It achieved an AUC of 0.852 in the SSNAP temporal validation, while the performance was slightly higher in the Riksstroke validation (AUC of 0.861). Models were well calibrated in SSNAP but slightly overestimated the risk of mortality in Riksstroke for in-hospital deaths. Precision-recall performance was higher in Riksstroke, and all models performed better for hemorrhagic than for ischemic stroke. Sensitivity analysis showed no impact on the results. Performance was consistent in both countries, suggesting that the model is suitable for quality improvement analytics and cross-national benchmarking. Although XGBoost performed slightly better, the differences from LR were minimal, emphasizing that well-designed classical models can match ML in settings with structured data and limited predictors. The model requires further validation if applied beyond similar European health systems.

5.2 Paper II

The study developed and evaluated the performance and explainability of three ML models and mLR in predicting death and functional dependence three months after stroke. The dataset included 102,135 adult patients with ischemic, hemorrhagic, or unspecified stroke registered in the Riksstroke from 2015 to 2020, after excluding patients younger than 18 years and those lost to follow-up. The variables comprised demographic characteristics, cardiovascular risk factors, medications, pre-stroke func-

tional status, acute care, stroke type, and severity, among others. The primary outcome was measured using mRS at three months after stroke, categorized into independence (a scale of 0–2), dependence (3–5), and death (6). Missing values in variable National Institutes of Health Stroke Scale (NIHSS) ($\approx 42.6\%$) were imputed using multivariate imputation by chained equations, while other variables with minimal missingness were assigned separate missing categories.

Three ML models, XGBoost, SVM and ANNs, and two classical models mLR with and without two-way interaction terms, were developed using the same set of predictors. All models were trained on a 75/25 train-test split with preserved class proportions. Categorical variables were one-hot encoded, while continuous variables were transformed using Min-Max scaling. Hyperparameters for ML models were optimized using 5-fold cross-validated grid search based on weighted F1 scores. Model performance was evaluated using metrics for multi-class outcomes, including accuracy, precision, recall, Cohen’s Kappa, Mathew’s correlation coefficient, F1 scores, and AUC-ROC, with 95% bootstrap confidence intervals. The explainability of the models was assessed using SHapley Additive exPlanations (SHAP) values, providing visualizations of feature contributions to model predictions at the patient and population levels.

All models achieved similar overall accuracy between 69% and 70%. However, ANNs and XGBoost models demonstrated better performance compared to mLR in classifying the clinically challenging functional dependence category, with F1 scores of 0.603 and 0.577, respectively, versus 0.544 for mLR. NIHSS was consistently the strongest predictor of all outcome classes, followed by age, type of stroke, ambulance service to the hospital, atrial fibrillation, and pre-stroke functional status. The study demonstrated that ML models, particularly ANNs and XGBoost, offer modest but meaningful improvements over traditional mLR in predicting multi-class functional outcomes after stroke using routine registry data. Both models showed promising performance in predicting functional dependence, a key outcome in planning rehabilitation and resource allocation. Although ML models lack explicit parameter estimates, SHAP-based explanations provided interpretability comparable to traditional models. The study emphasized the need to externally validate these models on independent datasets before their application in clinical practice.

5.3 Paper III

Paper III provided a comprehensive and fair comparison of the Cox PH model to ML models by evaluating their performance under varying data structures and benchmarking findings on real data from Riksstroke. Analyses were conducted using both simulated and Riksstroke datasets. Data were simulated to capture some aspects of the Riksstroke data structure as follows:

- 17 independent covariates ($\mathbf{X}_i \in \mathbb{R}^{1 \times 17}$; $i = 1, \dots, n$ patients) were sampled from predefined distributions (normal, binomial, and multinomial), with specified parameters set from Riksstroke summary statistics (mean, proportions), while the main linear effects β 's estimated from a Cox PH model fitted to the Riksstroke data.
- Survival times were generated under the Cox PH model with an exponential baseline hazard, $\lambda = 0.125$, using the formulae in (Bender et al., 2005);

$$T = -\frac{\log(U)}{\lambda \exp(\boldsymbol{\beta}^\top \mathbf{X}_i)}; \quad U \sim \text{Unif}(0, 1).$$

- Censoring times were sampled independently of covariates from an exponential distribution, $C \sim \exp(\alpha)$, where the parameter α was varied to achieve the desired proportion of censoring.
- Survival outcome was defined by $t = \min(T, C)$.
- Data generating processes included varying sample sizes (500, 1000, 5000, 10000), proportion of censoring (0%, 10%, 50%, 85%), including noise variables in the fitted model (0, 15, 100, 500), and in the presence of model misspecification (weak non-linearity, moderate non-linearity, violated PH assumption).

The Riksstroke dataset comprised 70,289 adult patients with ischemic stroke registered between 2015–2020. They were observed for 2191 days, from the onset of stroke until time of death (19,374 cases) or the end of follow-up (72% right censored). Thus, the primary endpoint was overall survival. The dataset included 17 covariates on demographics, cardiovascular risk factors, acute care, secondary prevention, pre-stroke functional status, and stroke characteristics.

Three ML models, DeepSurv, RSF, XGBoost, and a classical Cox PH model were trained on 70% of each dataset and evaluated on the remaining 30% at the maximum observed follow-up time, using Harrell’s C-index, Uno’s C-index, and the integrated Brier score. Hyperparameters for the ML models were tuned on sub-samples of the Riksstroke dataset to assess whether optimization improved predictive performance compared to the default settings used for training in the simulation study.

All models performed better with larger sample sizes. However, the performance declined with heavy censoring and a greater number of noise variables. The Cox PH model demonstrated better performance when the model was either correctly specified or when the departures from its assumptions were minor, whereas tree-based models (XGBoost, RSF) performed better than Cox PH and DeepSurv models under moderate non-linearity, violations of the PH assumption or in the presence of many noise variables. Compared with default settings, tuning of ML hyperparameters yielded small performance gains, mostly observed in smaller samples, but negligible changes in larger samples, and did not change the model’s performance ranking. The findings in Paper III suggested that the Cox PH model is preferable when its assumptions approximately hold, while tree-based ML models are advantageous in settings with non-linear or misspecified data structures, emphasizing the need to choose prediction models based on research question and underlying data characteristics.

5.4 Paper IV

Paper IV comparatively evaluated the predictive performance of traditional, ML, and pseudo-observation-based competing risk models across multiple time points. Two real-world clinical datasets were used. The first dataset, Riksstroke, consisted of 50,356 patients discharged alive after a first-ever ischemic stroke between 2020–2023, followed for recurrent stroke (event of interest), all-cause mortality (competing event) or censoring. It contained 23 covariates on demographics, cardiovascular risk factors, stroke severity, acute management, and secondary prevention. The study also used the primary biliary cirrhosis (PBC) dataset (available in the R `survival` package), which included 276 patients enrolled in a Mayo Clinic trial (1974 – 1984), with death as primary event and liver transplantation as competing event, and 16 baseline demographic, clinical, and laboratory covariates.

Three modeling frameworks that included traditional models (CS Cox PH and Fine-Gray models), a tree-based model (RSF), a deep-learning-based model (DeepHit), and pseudo-observation-based models (RF, linear regression, and elastic-net linear regression models) were implemented to estimate the cause-specific cumulative incidence function at multiple prediction horizons. The hyperparameters for RSF, RF, and DeepHit were tuned based on cross-validated random search, after which all models were trained on 70% of each dataset and evaluated on the remaining 30%. The performance of the models was evaluated using the time-dependent C-index (discrimination), smoothed calibration curves and integrated calibration index (calibration), and the Brier score (overall prediction error), with 95% confidence intervals estimated by bootstrap resampling.

The results showed that model performance changed substantially with the evaluation time point and dataset. In the PBC dataset, where the event of interest was common, models demonstrated clear performance differences, with RSF and DeepHit models achieving the highest discrimination, and PO-based models showing better calibration, especially at longer horizons. All models indicated time-dependent miscalibration, typically underestimating risk in low-risk patients and overestimating in high-risk patients, with RSF and DeepHit models showing the reverse pattern. In the Riksstroke dataset, where stroke recurrence was rare, discrimination was modest and similar across all models. The FG model showed the most consistent calibration, while PO-based models, especially elastic-net linear regression model, performed comparatively better at longer horizons. In conclusion, the study showed that the predictive performance of competing risks models is dependent on both the evaluation time point and dataset. When events are common, models benefit from richer event information, resulting in clearer differences in performance metrics across models. In contrast, when outcomes of interest are rare, models achieve similar and only modest discrimination, and achieve similar overall prediction errors regardless of the modeling framework. The findings emphasize that no single model performs best across all evaluation time points, and that meaningful comparison requires assessing multiple performance metrics at multiple clinically relevant time horizons to guide fair model selection.

6 Ethical considerations

The use of Riksstroke data for the purposes of statistical model development and evaluation in this thesis have been approved by the Swedish Ethical Review Authority (ref. 2021-06152-01 for papers I-III, and ref. 2023-07750-01 for paper IV). Data was pseudonymized before it was accessed by the research group.

Patients are informed about registration in the quality registry Riksstroke that the registry aims to support high and consistent quality of care for stroke patients throughout Sweden, and that data may be used for research purposes. In accordance with the Personal Data Act (Swedish law No. SFS 1998:204), no informed consent is needed to collect data from medical charts or inpatient records for quality registries. However, patients are informed of their rights to deny participation (opt-out consent).

SSNAP has approval from the Clinical Advisory Group of the NHS Health Research Authority to collect patient-level data under section 251 of the NHS Act 2006. No patient-level data were transferred between countries.

7 Conclusion and further research

The performance of the externally validated model was comparable across two European countries (UK and Sweden). This supports cross-registry generalizability within similar European health systems. The model could strengthen hospital benchmarking by supporting fairer international comparisons of standardized mortality ratios (Roessler et al., 2021), and it may also help identify patients at high risk of 30-day mortality who could benefit from interventions. Machine learning models demonstrated modest but significant improvements compared to the multinomial logistic regression model in predicting 3-month post-stroke functional outcomes, especially the challenging outcome of functional dependence. The improved multi-class prediction can support rehabilitation planning, resource allocation, benchmarking of outcomes across hospitals, and potentially real-time clinical decision support systems. The findings demonstrated that each model has strengths depending on specific conditions, including when the model assumptions hold or not, when the dataset size is small or large, when the proportion of censoring is small or not, and the computational efficiency of the model. Additionally, the prediction

performance of competing risks models depends on the evaluation time points and the dataset, which emphasizes the need to report multiple measures evaluated at different clinically relevant time points to guide model selection.

This thesis was based on a limited set of predictive features. Extending the models to include more detailed information on diagnosing (e.g., imaging), clinical biomarkers (e.g., blood pressure levels, cholesterol), and comorbidities (e.g., cancer, heart disease) could improve their performance further, as these factors may provide additional prognostic information. Our external validation of the risk prediction model of 30-day mortality demonstrated good agreement between the UK and Swedish stroke settings. An extended external validation, including different time periods and settings, and in particular of the other models in this thesis is needed to evaluate their generalization and potential use in supporting decision making. Because these models require substantial computational resources, the thesis did not explore a wider range of hyperparameters during model tuning. Future studies could investigate a broader hyperparameter space to identify potential performance improvements. Additionally, while tree-based models demonstrated comparatively better performance in the presence of noise, tuning is needed to determine whether this reflects genuine robustness or simply unintended overfitting. The simulations used independent covariates with fixed log-hazard effects, providing a straightforward and controlled setup to assess model behavior; however, the design lacks realistic covariate correlations and may not reflect real-world effect sizes. Additional simulations could incorporate correlated covariates and more complex data-generation scenarios to evaluate whether the model's behavior is consistent under more realistic conditions.

References

- Alhumaidi, N. H., Dermawan, D., Kamaruzaman, H. F., and Alotaiq, N. (2025). The use of machine learning for analyzing real-world data in disease prediction and management: Systematic review. *JMIR Medical Informatics*, 13(1):e68898.
- Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K.,

- Badreldin, H. A., et al. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, 23(1):689.
- Andersen, P. K. and Pohar Perme, M. (2010). Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19(1):71–99.
- Asplund, K., Hulter Åsberg, K., Appelros, P., Bjarne, D., Eriksson, M., Johansson, Å., Jonsson, F., Norrving, B., Stegmayr, B., Terént, A., et al. (2011). The riks-stroke story: building a sustainable national register for quality assessment of stroke care. *International Journal of Stroke*, 6(2):99–108.
- Austin, P. C., Putter, H., Giardiello, D., and Van Klaveren, D. (2022). Graphical calibration curves and the integrated calibration index (ici) for competing risk models. *Diagnostic and prognostic research*, 6(1):2.
- Badawy, M., Ramadan, N., and Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, 10(1):40.
- Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.
- Berrar, D. (2018). Performance measures for binary classifications. *S Ranganathan, K Nakai, C, eds. Schonbach Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1:546–70.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Chen, T. (2016). Xgboost: A scalable tree boosting system. *Cornell University*.
- Chong, B., Jayabaskaran, J., Jauhari, S. M., Chan, S. P., Goh, R., Kueh, M. T. W., Li, H., Chin, Y. H., Kong, G., Anand, V. V., et al. (2025). Global burden of cardiovascular diseases: projections from 2025 to 2050. *European journal of preventive cardiology*, 32(11):1001–1015.

- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22.
- Coemans, M., Verbeke, G., Döhler, B., Süsal, C., and Naesens, M. (2022). Bias by censoring for competing events in survival analysis. *bmj*, 378.
- Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., et al. (2024). Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.
- Dreiseitl, S. and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359.
- Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., and Salanti, G. (2024). Developing clinical prediction models: a step-by-step guide. *Bmj*, 386.
- Eriksson, M., Åsberg, S., Sunnerhagen, K. S., von Euler, M., and Collaboration, R. (2021). Sex differences in stroke care and outcome 2005–2018: observations from the swedish stroke register. *Stroke*, 52(10):3233–3242.
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.
- Feigin, V. L., Brainin, M., Norrving, B., Martins, S. O., Pandian, J., Lindsay, P., F Grupper, M., and Rautalin, I. (2025). World stroke organization: global stroke fact sheet 2025. *International Journal of Stroke*, 20(2):132–144.
- Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181.

- Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509.
- Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.
- Glenn, W. B. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.
- Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. P. (2016). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 24(1):198.
- Goyal, M., Ospel, J. M., Kappelhof, M., and Ganesh, A. (2021). Challenges of outcome prediction for acute stroke treatment decisions. *Stroke*, 52(5):1921–1928.
- Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545.
- Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.
- Harrell Jr, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.
- Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multi-class support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425.
- Huang, Y., Bazzazzadehgan, S., Li, J., Arabshomali, A., Li, M., Bhattacharya, K., and Bentley, J. P. (2025). Comparison of machine learning methods versus traditional cox regression for survival prediction in

- cancer using real-world data: a systematic literature review and meta-analysis. *BMC Medical Research Methodology*, 25(1):243.
- Huang, Y., Li, J., Li, M., and Aparasu, R. R. (2023). Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC medical research methodology*, 23(1):268.
- Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. (2014). Random survival forests for competing risks. *Biostatistics*, 15(4):757–773.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3).
- Ishwaran, H., Kogalur, U. B., Chen, X., and Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(1):115–132.
- Kantidakis, G., Biganzoli, E., Putter, H., and Fiocco, M. (2021). A simulation study to compare the predictive performance of survival neural networks with cox models for clinical trial data. *Computational and Mathematical Methods in Medicine*, 2021(1):2160322.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.
- Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24.
- Kleinbaum, D. and Klein, M. (2011). *Survival Analysis: A Self-Learning Text, Third Edition*. Statistics for Biology and Health. Springer New York.
- Kumar, A., Dhanka, S., Sharma, A., Sharma, A., Nain, M., Kumar, P., Gupta, A. R., Bansal, J., Saxena, N. K., and Pant, R. (2025). A comprehensive review of bias in ai, ml, and dl models: Methods, impacts, and future directions. *Archives of Computational Methods in Engineering*, pages 1–31.

- Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. (2018). Deephit: a deep learning approach to survival analysis with competing risks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.
- Lekander, I., Willers, C., Von Euler, M., Lilja, M., Sunnerhagen, K. S., Pessah-Rasmussen, H., and Borgström, F. (2017). Relationship between functional disability and costs one and two years post stroke. *PLoS one*, 12(4):e0174861.
- Lynch, C. J. and Liston, C. (2018). New machine-learning technologies for computer-aided diagnosis. *Nature medicine*, 24(9):1304–1305.
- Mahmoudi, E., Kamdar, N., Kim, N., Gonzales, G., Singh, K., and Waljee, A. K. (2020). Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ*, 369:m958.
- Molina Menéndez, E. and Parraga Alava, J. (2024). Artificial neural networks for classification tasks: A systematic literature review. *Enfoque UTE*, pages 1–10.
- Monterrubio-Gómez, K., Constantine-Cooke, N., and Vallejos, C. A. (2024). A review on statistical and machine learning competing risks methods. *Biometrical Journal*, 66(2):2300060.
- Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102.
- Mufti, H. N., Hirsch, G. M., Abidi, S. R., and Abidi, S. S. R. (2019). Exploiting machine learning algorithms and methods for the prediction of agitated delirium after cardiac surgery: models development and validation study. *JMIR medical informatics*, 7(4):e14993.
- Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554.

- Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., and van Diepen, M. (2021). External validation of prognostic models: what, why, how, when and where? *Clinical kidney journal*, 14(1):49–58.
- Roessler, M., Schmitt, J., and Schoffer, O. (2021). Can we trust the standardized mortality ratio? a formal analysis and evaluation based on axiomatic requirements. *PLoS One*, 16(9):e0257003.
- Smith, H., Sweeting, M., Morris, T., and Crowther, M. J. (2022). A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagnostic and Prognostic Research*, 6(1):10.
- Srećković, S., Berber, A., and Filipović, N. (2022). The automated laplacean demon: How ml challenges our views on prediction and explanation. *Minds and Machines*, 32(1):159–183.
- Stahl, D. (2024). New horizons in prediction modelling using machine learning in older people’s healthcare research. *Age and ageing*, 53(9):afae201.
- Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an abcd for validation. *European heart journal*, 35(29):1925–1931.
- Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.
- Teshale, A. B., Htun, H. L., Vered, M., Owen, A. J., and Freak-Poli, R. (2024). A systematic review of artificial intelligence models for time-to-event outcome applied in cardiovascular disease risk prediction. *Journal of medical systems*, 48(1):68.
- Trabassi, D., Serrao, M., Varrecchia, T., Ranavolo, A., Coppola, G., De Icco, R., Tassorelli, C., and Castiglia, S. F. (2022). Machine learning approach to support the detection of parkinson’s disease in imu-based gait analysis. *Sensors*, 22(10):3700.
- Tsampsian, V. and Bloomfield, G. S. (2025). The evolving global burden of cardiovascular diseases: what lies ahead.

- Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L.-J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117.
- Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):230.
- Vayena, E., Blasimme, A., and Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. *PLoS medicine*, 15(11):e1002689.
- Vickers, A. J., Cronin, A. M., Elkin, E. B., and Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC medical informatics and decision making*, 8(1):53.
- Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.
- Wang, L. (2023). Logistic regression for stroke prediction: An evaluation of its accuracy and validity. *Highlights in Science, Engineering and Technology*, 39:1086–1092.
- Wang, R., Zhang, J., Shan, B., He, M., and Xu, J. (2022a). Xgboost machine learning algorithm for prediction of outcome in aneurysmal subarachnoid hemorrhage. *Neuropsychiatric Disease and Treatment*, 18:659.
- Wang, W., Kiiik, M., Peek, N., Curcin, V., Marshall, I. J., Rudd, A. G., Wang, Y., Douiri, A., Wolfe, C. D., and Bray, B. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PloS one*, 15(6):e0234722.
- Wang, W., Otieno, J. A., Eriksson, M., Wolfe, C. D., Curcin, V., and Bray, B. D. (2023). Developing and externally validating a machine learning risk prediction model for 30-day mortality after stroke using national stroke registers in the uk and sweden. *BMJ open*, 13(11):e069811.

- Wang, W., Rudd, A. G., Wang, Y., Curcin, V., Wolfe, C. D., Peek, N., and Bray, B. (2022b). Risk prediction of 30-day mortality after stroke using machine learning: a nationwide registry-based cohort study. *BMC neurology*, 22(1):195.
- Wolbers, M., Blanche, P., Koller, M. T., Wittteman, J. C., and Gerds, T. A. (2014). Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539.
- Zakaria, M., Mabrouka, A., and Sarhan, S. (2014). Artificial neural network: a brief overview. *neural networks*, 1:2.
- Zihni, E., Madai, V. I., Livne, M., Galinovic, I., Khalil, A. A., Fiebach, J. B., and Frey, D. (2020). Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *Plos one*, 15(4):e0231166.