Aus dem Zentrum für Neurologie und Psychiatrie der Universität zu Köln
Klinik und Poliklinik für Neurologie
Direktor: Universitätsprofessor Dr. med. G. R. Fink

Prognostische Modellierung und Klassifikation neurologischen Outcomes nach Herz-Kreislauf-Stillstand mittels maschinellen Lernens von zerebralen Computertomographien

Inaugural-Dissertation zur Erlangung der Doktorwürde
der Medizinischen Fakultät
der Universität zu Köln

vorgelegt von
Maximilian Hubert Theodor Schmieschek
aus Köln

promoviert am 23. September 2025

Gedruckt mit Genehmigung der Medizinischen Fakultät der Universität zu Köln (2026).

Dekan: Universitätsprofessor Dr. med. G. R. Fink

1. Gutachter: Universitätsprofessor Dr. med. Ö. A. Onur
2. Gutachter: Professor Dr. med. C. Dohmen

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Dissertationsschrift ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes habe ich Unterstützungsleistungen von folgenden Personen erhalten:
Univ.-Prof. Dr. med. Oezguer A. Onur, Univ.-Prof. Dr. med. Gereon R. Fink,
Prof. Dr. med. Clemens Warnke, Dr. med. Hannes Gramespacher, Dr. med. Nils Richter,
Dr. med. Julian Dronse, Univ.-Prof. Dr. med. Stephan Baldus,
Priv.-Doz. Dr. med. Christoph Adler, Prof. Dr. med. Stefan Bittner,
Univ. Prof. Dr. med. Marc Schlamann, Dr. med. Carsten Gietzen, Dr. med. Charlotte Zaeske

Weitere Personen waren an der Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe einer Promotionsberaterin/eines Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertationsschrift stehen.

Die Dissertationsschrift wurde von mir bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die Aufbereitung der bildmorphologischen Datensätze zu dem manuellen Rating, sowie die. Verblindung, erfolgte durch Maximilian H.T. Schmieschek. Dr. med. Hannes Gramespacher und Maximilian H.T. Schmieschek trugen die vorhandenen klinischen Daten zusammen und harmonisierten diese für darauffolgende Analysen. Die Auswertung der Daten, inklusive der Vorverarbeitung, Extraktion der Features, die Modellierung des statistischen Vorhersagemodells samt der hierfür erforderlichen Python und Matlab Skripte, wurde von Maximilian H.T. Schmieschek durchgeführt. Die Interpretation der Ergebnisse erfolgte durch Univ.-Prof. Dr. med. Oezguer A. Onur, Dr. med. Hannes Gramespacher und Maximilian H.T. Schmieschek unter Beratung von Univ.-Prof. Dr. med. Gereon Fink. Die Verfassung des Manuskripts sowie das Anfertigen der Tabellen und Grafiken wurden mit der Unterstützung von Dr. med. Hannes Gramespacher durch Maximilian H. T. Schmieschek ausgeführt. Korrekturen und Verbesserungsvorschläge erfolgten durch Univ.-Prof. Dr. med. Oezguer A. Onur, Univ.-Prof. Dr. med. Gereon R. Fink, Prof. Dr. med. Clemens Warnke, Dr. med. Hannes Gramespacher, Dr. med. Nils Richter, Dr. med. Julian Dronse, Univ.-Prof. Dr. med. Stephan Baldus, Priv.-Doz. Dr. med. Christoph Adler, Prof. Dr. med. Stefan Bittner, Univ. Prof. Dr. med. Marc Schlamann, Dr. med. Carsten Gietzen, Dr. med. Charlotte Zaeske.

Die Erstellung des Studienkonzeptes erfolgte durch Univ.-Prof. Dr. med. Oezguer A. Onur, Univ.-Prof. Dr. med. Gereon Fink, Prof. Dr. med. Clemens Warnke, Priv.-Doz. Dr. med. Christoph Adler, Dr. med. Julian Dronse, Dr. med. Nils Richter, Dr. med. Hannes Gramespacher und Maximilian H.T. Schmieschek.

Die dieser Arbeit zugrunde liegenden interdisziplinären Datensätze wurden in folgenden Kliniken der Uniklinik Köln erhoben: Klinik III für Innere Medizin - Allgemeine und interventionelle Kardiologie, Elektrophysiologie, Angiologie, Pneumologie und internistische Intensivmedizin, Institut für Diagnostische und Interventionelle Radiologie und der Klinik und Poliklinik für Neurologie.

Die notfallmedizinische Versorgung der Patienten erfolgte in der Klinik III für Innere Medizin (Kardiologie) unter Leitung von Univ.-Prof. Dr. med. Stephan Baldus. Die Durchführung der Studie, die Rekrutierung der Patienten, die Untersuchung der Patienten, das Asservieren von Bioproben, die Datenerfassung und -dokumentation, der Aufbau und die Pflege der Datenbank erfolgte unter Leitung von Univ.-Prof. Dr. med. Stephan Baldus und Priv.-Doz. Dr. med. Christoph Adler. Klinische Daten zur neurologischen Prognoseeinschätzung wurden in einigen Fällen nacherfasst und bewertet durch Dr. Hannes Gramespacher und Maximilian H. T. Schmieschek. Die bildgebenden Verfahren wurden durch Univ.-Prof. Dr. med. Marc Schlamann ermöglicht. Die in dieser Arbeit zum Vergleich herangezogenen Messergebnisse von Neurofilament Light Chain (NfL) und Neuron-Specific Enolase (NSE) wurden im Labor der Klinik und Poliklinik für Neurologie an der Universitätsmedizin der Johannes-Gutenberg-Universität Mainz von Herrn Prof. Dr. med. Stefan Bittner erhoben. Die Organisation und Verarbeitung der Proben erfolgten unter der Leitung von Prof. Dr. med. Clemens Warnke und Prof. Dr. med. Özgür Onur. Die Ermittlung der Gray-White Matter Ratio (GWR) wurde verblindet durch Dr. med. Carsten Gietzen und Dr. med. Charlotte Zaeske vorgenommen.

Erklärung zur guten wissenschaftlichen Praxis

Ich erkläre hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten (Amtliche Mitteilung der Universität zu Köln AM 132/2020) der Universität zu Köln gelesen habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen.

Köln, den 20.03.2025 (M. H. T. Schmieschek)

„The world around us is of baffling complexity and the most obvious fact about it is that we cannot predict the future.“ – Wigner, 1960[1]

Inhaltsverzeichnis

ABKÜRZUNGSVERZEICHNIS 7

1. ZUSAMMENFASSUNG 8

2. EINLEITUNG 9

2.1. Architekturen von Vorhersagemodellen – eine kurze Einführung 9
2.1.1. Lineare Regression 9
2.1.2. Logistische Regression 10
2.1.3. Regularisierungen 10

2.2. Leistungsmaße und Kreuzvalidierung 11

2.3. Angewandtes Modell zur Vorhersage des neurologischen Outcomes nach Herzstillstand außerhalb des Krankenhauses 15
2.3.1. Studienziel und Forschungsfrage 15
2.3.2. Featureextraktion - Nutzung computertomographischer Daten 16
2.3.3. Erfassung der Zielgröße nach Herzkreislaufstillstand 16
2.3.4. Angewandtes Vorhersagemodell und Leistungsmaß 17

3. PUBLIKATION 20

4. DISKUSSION 57

4.1. Einordnung des prognostischen Wertes des bildbasierten Vorhersagemodells 57

4.2. Featurewichtung und Rückschlüsse auf die Vulnerabilität des Gehirns 58

4.3. Limitationen und Ausblick 59

4.4. Schlussfolgerung 61

5. LITERATURVERZEICHNIS 62

6. ANHANG 65

6.1. Tabellenverzeichnis 65

Abkürzungsverzeichnis

aCCTa: automatisiertes, CCT-basiertes Vorhersagemodell
AAL3- Hirnatlas: Automated Anatomical Labeling 3 Hirnatlas
AUC: Area Under the Curve (Fläche unter der Kurve einer ROC-Analyse)
BA: Balanced Accuracy (balancierte Genauigkeit)
CCT: Kraniale Computertomographie (CT des Kopfes)
CPC: Cerebral Performance Category (kategorisierte Einschätzung der neurologischen Leistungsfähigkeit nach Herz-Kreislauf-Stillstand)
CT: Computertomographie
∆CCT: Zeitdifferenz zwischen klinischer Vorstellung und CCT-Aufnahme
EEG: Elektroenzephalographie
F1-Score: Harmonisches Mittel aus Präzision und Recall (Sensitivität)
FP: False Positive (falsch positiv)
GMP: Graue-Materie-Wahrscheinlichkeit (Gray Matter Probability)
GWR: Gray-White-Matter-Ratio
GWR_e: Frühe (early) Gray-White-Matter-Ratio (geschätzte/erweiterte Form)
GWR_l: Späte (late) Gray-White-Matter-Ratio (spezifische Berechnungsvariante)
HU: Hounsfield Unit (Maßeinheit für Dichtewerte im CT)
L1-Regularisierung (LASSO): Least Absolute Shrinkage and Selection Operator
L2-Regularisierung (Ridge): Ridge-Regression
MAE: Mittlere absolute Abweichung
MRT: Magnetresonanztomographie
MSE: Mittlere quadratische Abweichung
NSE: Neuronenspezifische Enolase
NPV: Negative Predictive Value (negativer Vorhersagewert)
NfL: Neurofilament Light Chain
OHCA: Out-of-Hospital Cardiac Arrest (Herz-Kreislauf-Stillstand außerhalb des Krankenhauses)
OLS: Ordinary Least Squares
PPV: Positive Predictive Value (positiver Vorhersagewert)
PR: Precision-Recall (Kurve bzw. Analyse)
ROC: Receiver Operating Characteristic
ROC-AUC: Area Under the Curve der ROC (Flächenwert unter der ROC-Kurve)
ROI: Region of Interest (definierter, räumlicher Bereich in Bilddaten)
SPM: Statistical Parametric Mapping
TNR: True Negative Rate (Spezifität)
TP: True Positive (richtig positiv)
TPR: True Positive Rate (Sensitivität/Recall)
TN: True Negative (richtig negativ)

1. Zusammenfassung

Die Arbeit bietet einen Überblick über die grundlegenden methodischen Bausteine von Vorhersagemodellen, basierend auf klassischen statistischen Verfahren wie der linearen und logistischen Regression. Dabei werden im Folgenden Regularisierungsmethoden beschrieben und die Eigenschaften, die Auswahl und die Interpretation geeigneter Leistungsmaße diskutiert, die beim Training, bei der Leistungsbewertung und beim Vergleich von Vorhersagemodellen eine zentrale Rolle spielen. Bei den zuvor genannten Aspekten werden sowohl theoretische Hintergründe als auch praktische Herausforderungen und Limitationen beleuchtet.

Diese Grundlagen werden am Beispiel der veröffentlichten Arbeit zur Vorhersage des neurologischen Outcomes nach einem Herz-Kreislauf-Stillstand außerhalb des Krankenhauses mittels kontextspezifischer Designentscheidungen praktisch angewandt.

Die hierbei verwendeten bildbasierten Features basieren auf einem neuartigen, der MRT-Bildgebung entlehnten, Verfahren der Gewebssegmentierung von klinischen kranialen Computertomographien. Hierbei werden durch individuelle Abweichungen in der Segmentierung der Gewebsklassen subtile, hypoxiebedingte Veränderungen im Gehirn als Features extrahiert. Das auf diesen bildgebenden Features beruhende Vorhersagemodell prognostiziert auf einem dem Training vorenthaltenen Testdatensatz den neurologischen Outcome besser als herkömmliche manuell erhobene radiologische Kennwerte, wie etwa die klinisch etablierte Gray-White-Matter-Ratio. Zudem zeigt die Arbeit, dass das Vorhersagemodell bereits zu einem frühen klinischen Zeitpunkt nach Herz-Kreislauf-Stillstand bessere Prognosen als etablierte laborchemische Biomarker (Neurofilament Light Chain, NfL oder Neuronenspezifische Enolase, NSE) liefert. Diese Überlegenheit des bildbasierten Vorhersagemodells gilt nach einem und zwei Tagen nach Herzkreislaufstillstand gegenüber den laborchemischen Biomarkern nicht mehr.

Insgesamt legt die Arbeit dar, wie ein fundierter methodischer Unterbau, eine neuartige Nutzung von klinisch routinemäßig erhobener kraniellen Bildgebung in einem Vorhersagemodell durch maschinelles Lernen und geeignete Wahl von Leistungsmaßen, den Weg zu gut interpretierbaren, verlässlichen und klinisch relevanten Vorhersagen ebnet. Gleichzeitig werden Limitationen, wie nötige weitere Validierung mittels multizentrischer, prospektiver Erhebungen und Nutzung kombinierter Modalitäten, beleuchtet, um die Vorhersagemodelle weiter zu optimieren und langfristig in die klinische Routine integrieren zu können.

2. Einleitung

Im Folgenden wird ein Überblick über klassischen statistischen Verfahren wie der linearen und der logistischen Regression, über Regularisierungsverfahren, Evaluationskriterien und Optimierungsstrategien gegeben. Dies stellt den methodischen Grundstein für fundierte, robuste und interpretierbare Vorhersagemodelle in verschiedensten Anwendungsbereichen dar.

2.1. Architekturen von Vorhersagemodellen – eine kurze Einführung

2.1.1. Lineare Regression

Lineare Regressionsmodelle zählen seit Jahrzehnten zu den zentralen Verfahren in der statistischen Datenanalyse und haben sich insbesondere im Bereich der Vorhersage kontinuierlicher Daten als robuste, nachvollziehbare und vergleichsweise einfach zu interpretierende Modelle etabliert. Mitunter sind lineare Regressionsmodelle leistungsfähiger als komplexere nichtlineare Modelle, insbesondere bei begrenzter Datensatzgröße, im Falle von unterliegenden linearen Zusammenhängen und bei Signalen mit hohem Signal-zu-Rauschen-Verhältnis. Grundlegend beruhen einfache und multiple lineare Regressionsmodelle auf der Annahme, dass sich zu prognostizierenden Werten einer abhängigen Variablen ($\boldsymbol{y}$) durch eine lineare Kombination gewichteter ($\boldsymbol{\beta}$) unabhängiger Variablen ($\boldsymbol{X}$) mit einer meist unabhängigen normalverteilten Abweichung ($\boldsymbol{\varepsilon}$) abbilden lassen. Die unabhängigen Variablen ($\boldsymbol{X}$) werden im Feld des maschinellen Lernens und auch im Weiteren als Features bezeichnet. Folgend die allgemeine Formulierung eines linearen Regressionsmodells in der Matrixschreibweise[2] :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Die Modellparameter ($\boldsymbol{\beta}$) sind so angepasst, dass die beobachteten Zusammenhänge bestmöglich approximiert werden. D.h. das Modell wird klassischerweise durch die Minimierung der Summe der quadrierten Abweichungen zwischen den vorhergesagten und den tatsächlich beobachteten Werten bezüglich der Wichtungen ($\widehat{\boldsymbol{\beta}}$) optimiert (Ordinary Least Squares, OLS)[2] :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$$

Bei diesem simplen Vorgehen bieten lineare Regressionsmodelle sowohl ein klares theoretisches Fundament als auch ein starkes Interpretationspotenzial, da die Wichtungen ($\boldsymbol{\beta}$) den Einfluss des jeweiligen Features auf die Vorhersage ausdrücken. Zudem können durch gezielte Transformationen der Features ($\boldsymbol{X}$) vorab teils weiterführende komplexe und auch nichtlineare Zusammenhänge betrachtet werden. Damit eignen sich lineare Regressionsmodelle als Ausgangspunkt für Vorhersagen in unterschiedlichsten Disziplinen, von der Wirtschaft bis hin zur Medizin. Sie bilden den methodischen Grundstein, auf dem weiterführende, nichtlineare oder maschinelle Lernverfahren aufbauen.

2.1.2. Logistische Regression

Logistische Regressionsmodelle gehören zu den grundlegenden Verfahren in der statistischen Modellierung und Vorhersageanalyse, insbesondere dann, wenn die abhängigen Variablen von nominal- oder ordinalskalierter Natur sind. Anders als bei linearen Regressionsansätzen, bei denen das Ziel darin besteht, einen stetigen Zielwert möglichst präzise zu schätzen, liegt der Fokus bei logistischen Modellen auf der Modellierung der Wahrscheinlichkeit, dass ein bestimmtes Ereignis eintritt. Hierbei wird der Zusammenhang zwischen einer oder mehreren unabhängigen Variablen und der Eintrittswahrscheinlichkeit eines Outcomes durch eine nichtlineare Funktion abgebildet. Dieser Ansatz ermöglicht es, komplexe Zusammenhänge adäquat zu erfassen und sicherzustellen, dass die geschätzten Wahrscheinlichkeiten stets im Intervall zwischen 0 und 1 liegen. In Anlehnung an P. F. Verhulst Veröffentlichung von 1838[3] :

$$\frac{dP}{dx} \simeq \theta P$$

$$\Rightarrow P(y = 1|x) = \frac{1}{1 + e^{-(\sum_{i=0}^{n} \theta_i x_i)}} | x_0 = 1$$

Der Einfluss einzelner Features ($\boldsymbol{x}$) auf die Vorhersage $P(y = 1|x)$ kann, analog zur linearen Regression, mittels der Wichtungen ($\boldsymbol{\theta}$) ermittelt werden. Dank dieser klaren Interpretierbarkeit der logistischen Regression und dem ausgereiften methodischen Fundament, das auf Maximum-Likelihood-Schätzungen basiert, ist sie in einer Vielzahl von Disziplinen weitverbreitet. Beispiele hierfür reichen von medizinischen Diagnosen, bei denen vorhergesagt wird, ob ein Patient an einer bestimmten Krankheit erkrankt, über Kreditrisikobewertungen in den Wirtschaftswissenschaften, bis hin zu Klassifizierungsproblemen in der industriellen Qualitätssicherung. Durch Erweiterungen wie multikategoriale logistische Modelle oder gemischte logistische Modelle kann die Anwendung zudem auf komplexere Szenarien ausgedehnt werden. Damit bildet die logistische Regression einen wesentlichen Baustein im Spektrum prognostisch relevanter statistischer und ökonometrischer Methoden.

2.1.3. Regularisierungen

Regularisierungsverfahren spielen eine zentrale Rolle bei der Entwicklung von robusten und verallgemeinerungsfähigen statistischen und maschinellen Lernmodellen. Sie zielen darauf ab, die Anzahl der Features zu kontrollieren und damit die Komplexität eines Modells zu vereinfachen. Dies trägt insbesondere zur Lösung multikollinearitätsbedingter Probleme bei, da stark korrelierte Features reguliert und numerische Instabilitäten reduziert werden. Zudem entstehen dadurch Modelle, die durch die Reduktion der Features einfacher und klarer interpretierbar sind. Ferner wird dadurch eine mögliche Überanpassung (Overfitting) an die vorhandenen Trainingsdaten vermieden und die Generalisierbarkeit des Modells, im Sinne der sinnvollen Vorhersageleistung neuer ungesehenen Daten, gestärkt. Zu den gängigen Formen der Regularisierung zählen vor allem die L1- und L2-Regularisierung.

Die L1-Regularisierung, auch als LASSO (least absolute shrinkage and selection operator) bezeichnet, führt durch ihren absoluten Strafterm nicht nur zu einer Verkleinerung der Koeffizienten, sondern kann diese auch exakt auf null setzen. Dadurch eignet sich die L1-Regularisierung besonders gut für die Merkmalsselektion, da irrelevante oder kaum beitragende Features effektiv aus dem Modell entfernt werden.[4]

$$\widehat{\boldsymbol{\beta}} = arg\,\min_{\beta}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|)$$

Die L2-Regularisierung, auch als Ridge-Regularisierung bekannt, bestraft große Koeffizientenquadrate und schrumpft damit alle Parameterwerte gleichmäßig, ohne sie gänzlich zu eliminieren. Diese Methode erhöht die Stabilität des Modells, reduziert jedoch nicht die Anzahl der tatsächlich verwendeten Features.[5]

$$\widehat{\boldsymbol{\beta}} = arg\,\min_{\beta}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2)$$

Das Elastische Netz (Elastic Net) stellt eine Kombination beider Verfahren dar, indem es den Strafterm aus einer gewichteten Kombination von L1- und L2-Regularisierung aufbaut, sodass gleichzeitig die Stärken beider Welten nutzbar sind.[6]

$$\widehat{\boldsymbol{\beta}} = arg\,\min_{\beta}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda_2\|\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|)$$

Die L2-Komponente verhindert, dass sich die Koeffizienten instabil verhalten, während die L1-Komponente sicherstellt, dass Features von geringer Relevanz konsequent abgeschwächt oder entfernt werden. Zusammenfassend bietet diese Form der Regularisierung ein wichtiges Handwerkszeug, um Modelle effizienter, interpretierbarer und vor allem besser generalisierbar zu gestalten.

2.2. Leistungsmaße und Kreuzvalidierung

Die Nutzung von geeigneten Leistungsmaßen für Vorhersagemodelle ist sowohl bei linearen als auch bei logistischen Regressionsverfahren von zentraler Bedeutung, um die Vorhersagemodelle an die Daten anzupassen und fundierte Aussagen über deren Leistungsfähigkeit in dem gewünschten Kontext treffen zu können. Die genutzten Leistungsmaße fungieren hierbei als quantifizierbare Größe, um die Vorhersagequalität widerzuspiegeln und mit alternativen Vorhersageverfahren vergleichen zu können.

Bei linearen Regressionsmodellen eignen sich als Leistungsmaß unter anderem folgende:

Die weit verbreitete mittlere quadratische Abweichung (MSE), welche sich aus der durchschnittlichen quadratischen Differenz zwischen den tatsächlichen Werten (y) und den vom Modell vorhergesagten Werten ($\hat{y}$) berechnet[7,8] :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Ein niedriger MSE zeigt an, dass die Vorhersagen des Modells insgesamt eng an den tatsächlichen Werten liegen und das Modell somit eine gute Anpassung erreicht. Die mittlere

absolute Abweichung (MAE), welche sich aus dem durchschnittlichen absoluten Unterschied zwischen den tatsächlichen Werten (y) und den vom Modell vorhergesagten Werten ($\hat{y}$) berechnet[9] :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Im Gegensatz zur MSE, bei der die Abweichungen quadriert werden, fließen Ausreißer hierbei nicht überproportional stark in das Leistungsmaß ein, wodurch der MAE als robuster gegenüber extremen Werten angesehen wird. Das Leistungsmaß R^2, häufig auch als Bestimmtheitsmaß benannt, bewertet, wie gut ein Modell die Streuung in den beobachteten Daten erklären kann[10,11] :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} | \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Ein R^2 nahe 1 deutet darauf hin, dass das Modell die Variation in den Zielwerten sehr gut erfasst. Je niedriger der Wert, desto weniger Variation wird durch das Modell erklärt. Ein Wert von 0 bedeutet, dass das Modell genauso gut ist wie eine einfache Schätzung basierend auf dem Mittelwert $\bar{y}$.

Für logistische Regressionsmodelle stehen Leistungsmaße im Vordergrund, die je nach Kontext unterschiedliche Aspekte einer erfolgreichen Klassifizierung abbilden. Die Auswahl geeigneter Leistungsmaße zur Bewertung von Vorhersagemodellen hängt zudem in hohem Maße von der Verteilung der Zielvariablen im Datensatz ab. Daher ist deren Anwendung entsprechend der Datenverteilung sowie der Eignung für bestimmte Problemstellungen auszuwählen.
Bei balancierten Datensätzen, in denen die Klassenverteilungen annähernd gleichverteilt sind, können Leistungsmaße wie bspw. die Genauigkeit (Accuracy) oft hinreichende Aussagen über die Modellgüte treffen. Diese Leistungsmaße spiegeln in solchen Szenarien typischerweise das Gesamtvorhersagevermögen des Modells gut wider, ohne durch ein Übergewicht einer bestimmten Klasse verzerrt zu werden. Die Kreuztabelle (Confusion Matrix) und ihre abgeleiteten Kennzahlen ermöglichen eine differenzierte Betrachtung der Modellleistung und stellen die Grundlage für ableitbare Leistungsmaße dar[11,12] :

	Tatsächlich positiv $y = 1$	Tatsächlich negativ $y = 0$
Vorhergesagt positiv $\hat{y} = 1$	richtig positiv (TP)	falsch positiv (FP)
Vorhergesagt negativ $\hat{y} = 0$	falsch negativ (FN)	richtig negativ (TN)

Tabelle 1: Kreuztabelle (Confusion Matrix)

Aufbauend auf diesen Grundbegriffen (TP, FP, TN, FN) werden häufig folgende Kennzahlen definiert und kontextspezifisch als Leistungsmaß genutzt:
Die Richtig-Positiv-Rate oder auch Trefferquote (True Positive Rate, TPR, Sensitivität, Recall) genannt, definiert sich als Anteil richtig positiver Fälle, die als positiv erkannt wurden[12] :

$$True\ Positive\ Rate\ (TPR) = \ Sensitivität = Recall = \frac{TP}{TP + FN}$$

Dies spiegelt insbesondere wider, wie gut das Modell wahre positive Fälle erkennt.
Die Richtig-negativ-Rate (True Negative Rate, TNR) oder Spezifität genannt, gibt den Anteil richtig negativer Fälle, die als negativ erkannt wurden wieder, und zeigt, wie gut das Modell wahre negative Fälle erkennt[12] :

$$True\ Negative\ Rate\ (TNR) = Spezifität = \frac{TN}{TN + FP}$$

Der positive Vorhersagewert oder die Genauigkeitsquote (Positive Predictive Value, PPV, Precision) gibt den Anteil tatsächlich positiver Fälle an allen positiv Klassifizierten (positiv prognostischer Wert) wieder und werden häufig in der Information-Retrieval-Forschung genutzt[12] :

$$Positive\ Vorhersagewert = Precision = \frac{TP}{TP + FP}$$

Der negative Vorhersagewert (Negative Predictive Value, NPV) spiegelt den Anteil tatsächlich negativer Fälle an allen negativ klassifizierten wider[12] :

$$\text{Negative Vorhersagewert} = Negative\ Predictive\ Value\ (NPV) = \frac{TN}{TN + FN}$$

Die Genauigkeit oder Accuracy stellt den Anteil korrekter Vorhersagen an allen Fällen dar[13] :

$$Genauigkeit = Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Bei stark unausgewogener Klassenverteilung kann das Leistungsmaß der Genauigkeit täuschen, da eine dominante Klasse das Ergebnis verzerren kann.
Im Gegensatz dazu sind unbalancierte Datensätze dadurch gekennzeichnet, dass eine oder mehrere Klassen deutlich unterrepräsentiert sind. Die Verwendung von auf das Gesamtergebnis orientierten Kennzahlen wie der Genauigkeit kann dann irreführend sein, da Modelle mit starker Tendenz zur häufigeren Klasse scheinbar hohe Werte erzielen. In diesem Fall kann es dazu kommen, dass seltenen Ereignisse von Interesse nicht gewertet werden. Um diese problematische Verzerrung zu adressieren, bieten sich folgende Leistungsmaße an:

Die Balanced Accuracy, welche formal als arithmetisches Mittel aus Sensitivität und Spezifität definiert ist, ist fester Bestandteil in der Forschung des maschinellen Lernens und lässt sich konzeptuell auf den Youden-Index (J) zurückführen[14,15] :

$$Balanced\ Accuracy = \frac{Sensitivität + Spezifität}{2} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) = \frac{J + 1}{2}$$

Der F1-Score bildet sich aus dem harmonischen Mittel von Präzision (Precision) und Sensitivität (Recall)[16] :

$$F1 - Score = 2\frac{Precision * Recall}{Precision + Recall} = \frac{2\,TP}{2\,TP + FP + FN}$$

Neben diesen Kennzahlen ermöglichen grafische Darstellungen weitere Einblicke in die Modellleistung und bieten insbesondere in dem medizinischen Kontext die Möglichkeit der Vergleichbarkeit mit etablierten Leistungsmaßen.

Die Receiver Operating Characteristic (ROC) Kurve visualisiert bspw. den Kompromiss zwischen Sensitivität (True Positive Rate) und Falsch-Positiv-Rate (False Positive Rate), bei variierenden Schwellenwerten des Klassifikators. Die Fläche unter der ROC-Kurve (Area Under the Curve, AUC) fasst damit die Gesamtleistung des Modells zusammen. Ein ROC-AUC-Wert von 1,0 signalisiert eine perfekte Trennung zwischen positiven und negativen Klassen, während ein Wert von 0,5 einer zufälligen Klassifikation entspricht. Diese Kennzahl ist besonders bei balancierten Datensätzen robust und bietet einen umfassenden Überblick über die Trennschärfe des Modells.

Die Precision-Recall (PR)-Kurve konzentriert sich hingegen speziell auf die positive Klasse, indem sie das Verhältnis von Precision zu Recall über verschiedene Schwellenwerte darstellt. Die Fläche unter der Precision-Recall-Kurve (AUC-PR) liefert insbesondere bei unbalancierten Datensätzen, in denen die positive Klasse selten ist, eine differenzierte Bewertung der Modellleistung. Da sie weniger von einer überwiegenden Anzahl negativer Fälle beeinflusst wird, zeigt die AUC-PR direkt, wie gut das Modell die seltenen, jedoch häufig entscheidenden positiven Fälle identifizieren kann.

Bei balancierten Datensätzen bietet die ROC-AUC ein verlässliches Maß, da sie den gesamten Zusammenhang zwischen korrekten und falschen Vorhersagen abbildet. Die PR-AUC liefert hier hingegen weniger Zusatzinformationen, da der Fokus allein auf der positiven Klasse liegt und kleine Schwankungen in Precision zu starken Kurvenänderungen führen können. Bei unbalancierten Datensätzen hat die PR-AUC klare Vorteile, da sie direkt die Qualität der Vorhersagen der seltenen positiven Klasse abbildet und nicht durch die überwältigende Anzahl negativer Fälle verzerrt wird. Die ROC-AUC kann in solchen Fällen oft zu optimistischen Bewertungen führen, da selbst bei einer schlechten Erkennung positiver Fälle die Falsch-Positiv-Rate niedrig erscheint.

Zusammenfassend zeigt sich, dass die Wahl des Leistungsmaßes, wie Sensitivität, Spezifität, Precision, Accuracy, Balanced Accuracy, F1-Score, ROC- und PR-AUC, entscheidend von der Verteilung der Klassen und der kontextspezifischen Zielformulierung abhängt. Die gezielte Wahl und differenzierte Betrachtung ist essenziell für eine sinnvolle Beurteilung und Optimierung von Klassifikationsmodellen in verschiedensten Anwendungsbereichen.

Bei maschinellen Lernalgorithmen werden Hyperparameter zur Optimierung (Minimierung oder Maximierung) eines der zuvor gewählten Leistungsmaße innerhalb des

Trainingsdatensatzes oder einer Teilmenge des Trainingsdatensatz genutzt. Ein gängiges Verfahren ist die Kreuzvalidierung, bei welcher der verfügbare Trainingsdatensatz in mehrere Teilmengen (Folds) aufgeteilt und zu mehrfachen Trainings-/Testdurchgängen genutzt wird. Als besonders geeignete Varianten haben sich dabei Fünffache-[17] und Zehnfache-[18] Kreuzvalidierung erwiesen, da sie in vielen Anwendungsfällen eine robusten, generalisierbare Modellgestaltung liefern. Dennoch ist eine anschließende Validierung der Generalisierbarkeit anhand desselben Leistungsmaßes an einem vom Training bzw. von der Kreuzvalidierung getrennten Testdatensatz erforderlich.

2.3. Angewandtes Modell zur Vorhersage des neurologischen Outcomes nach Herzstillstand außerhalb des Krankenhauses

In der vorliegenden Studie wurde ein regularisiertes logistisches Regressionsmodell zur binären Klassifikation zur Vorhersage des neurologischen Outcomes nach einem Herz-Kreislauf-Stillstand außerhalb eines Krankenhauses (Out-of-Hospital Cardiac Arrest, OHCA) eingesetzt. Das hierbei genutzte Verfahren basiert auf transformierten Werten der Gewebsdichte des gesamten Gehirns. Die genutzte Modellwahl zur kontextspezifischen Anwendung lässt sich durch eine Reihe methodischer und inhaltlicher Erwägungen begründen.

2.3.1. Studienziel und Forschungsfrage

Die vorliegende Studie verfolgt das Ziel, aus routinemäßig durchgeführten kranialen CT-Aufnahmen (CCT) verlässliche und automatisierbare Informationen zur neurologischen Prognose zu gewinnen. Ferner ist das Ziel der Arbeit, den prognostischen Wert dieser Methode in den Kontext von etablierten Prognoseverfahren, wie etablierten bildbasierten Prognosemarkern (GWR) und labormedizinischen Biomarkern im Serum (Neurofilament Light Chain, NfL oder Neuronenspezifische Enolase, NSE), zu setzen.

Die zentrale Herausforderung bei der bildbasierten Vorhersage des neurologischen Outcomes nach OHCA besteht darin, subtile, hypoxisch-ischämisch bedingte Veränderungen des Gehirns bereits frühzeitig, objektiv und zuverlässig, über das Hintergrundrauschen der im klinischen Kontext erhobenen Daten hinaus, zu erfassen. Klinisch tätige Mitarbeitende und Angehörige wünschen sich robuste Prognoseabschätzungen trotz einer heterogenen Datenlage, unterschiedlicher Erhebungszeitpunkte, variabler Bildqualitäten, verschiedener Scannerhersteller und meist kleiner Trainingsdatensätze.

Die sich daraus ableitende Forschungsfrage lautet: Ermöglicht ein automatisiertes Verfahren mittels klinisch erhobenen CCT-Aufnahmen nach OHCA eine ebenso verlässliche oder überlegene Vorhersage des neurologischen Outcomes im Vergleich zu etablierten manuellen CCT-Verfahren (GWR) und Laborparametern (NfL, NSE)?

2.3.2. Featureextraktion - Nutzung computertomographischer Daten

In dieser Studie wurde eine neue Methode zur prognostischen Analyse entwickelt, die auf der Extraktion von transformierten Werten der Gewebsdichte als Features aus 166 Regionen von Interesse (ROIs) des AAL3-Hirnatlas basiert.[19] Ergänzende klinische Features sind Alter und Geschlecht der Patienten, sowie der jeweilige Zeitraum zwischen klinischer Vorstellung und CCT-Aufnahme (ΔCCT).

Die Vorverarbeitungs-Pipeline umfasste eine an das CCT angepasste automatisierte, auf SPM (Statistical Parametric Mapping) aufbauende Segmentierung der nativen CCT-Daten in die Gewebsklassen graue Substanz, weiße Substanz, Schädelknochen und Liquor.[20,21] Dieser Segmentierungsprozess und die Extraktion der mittleren Wahrscheinlichkeit der grauen Substanz (GMP) aus den 166 ROIs des AAL3-Hirnatlas bilden die Basis für das individuelle Feature Set. Die damit durchgeführte nicht-lineare Transformation der CCTs geht über Hounsfield-Units-(HU)-basierte Nutzung von Kennwerten bzw. Ratios hinaus, ist anatomisch differenziert und erfasst subtile Veränderungen des gesamten Gehirns im Rahmen hypoxiebedingter Veränderungen.

2.3.3. Erfassung der Zielgröße nach Herzkreislaufstillstand

Die Cerebral Performance Category (CPC) ist eine klinische Skala zur Beurteilung des neurologischen Funktionsniveaus nach einem Herz-Kreislauf-Stillstand. Dabei wird der neurologische Zustand eines Patienten anhand von fünf Abstufungen erfasst:

Bei CPC 1 liegt eine gute zerebrale Leistungsfähigkeit vor, was bedeutet, dass die Person wach, ansprechbar und in der Lage ist, sowohl alltägliche Aktivitäten als auch berufliche Aufgaben zu erfüllen, auch wenn leichte neurologische oder psychische Defizite vorhanden sein können. Bei CPC 2 wird von einer mäßigen zerebralen Behinderung gesprochen, wobei die betroffene Person zwar bei Bewusstsein ist, aber möglicherweise in einer geschützten Umgebung oder unter angepassten Bedingungen Unterstützung benötigt, um den Alltag weitgehend selbstständig zu bewältigen. CPC 3 beschreibt eine schwere zerebrale Behinderung, bei der die Person zwar wach und ansprechbar ist, jedoch erhebliche Einschränkungen in ihren geistigen oder körperlichen Fähigkeiten bestehen, wodurch die Person bei alltäglichen Verrichtungen auf umfassende Hilfe angewiesen ist. CPC 4 steht für ein Koma oder einen vegetativen Zustand, hierbei kann die Person zwar Schlaf-Wach-Zyklen zeigen, reagiert jedoch nicht bewusst auf äußere Reize und ist nicht in der Lage, mit ihrer Umgebung zu interagieren. Schließlich steht CPC 5 für den irreversiblen Hirnfunktionsausfall („Hirntod“), der durch das vollständige Ausbleiben jeglicher Hirnaktivität, etwa fehlende Spontanatmung, Reflexe oder EEG-Signale, charakterisiert ist.[22–24] Durch diese graduelle Einteilung liefert das CPC-System eine vergleichsweise klare, übersichtliche und klinisch praktikable Einschätzung der neurologischen Funktion nach einem kardialen Ereignis.

Die dichotome Zusammenfassung in "gutes" (CPC 1–2) und "schlechtes" (CPC 3–5) neurologisches Outcome dient in wissenschaftlichen Studien vor allem der Vereinfachung und besseren Vergleichbarkeit von Patientengruppen.[25,26] Dieser Ansatz ermöglicht eine eindeutige Unterscheidung mit klinischer Relevanz und stellt auch im Rahmen dieser Studie eine robuste vorherzusagende Zielgröße dar. Durch diese dichotome Betrachtung gehen relevante Informationen für die Vorhersage verloren, da weder feine funktionelle Unterschiede noch das Potential möglicher Erholungs- oder Verschlechterungsdynamiken im Verlauf vollständig abgebildet werden. Auch kann eine solche Zusammenfassung in individuellen Fällen den tatsächlichen klinischen Zustand nur eingeschränkt widerspiegeln. Trotz dieser Einschränkungen ist die dichotome Einteilung in "gutes" vs. "schlechtes" neurologisches Outcome in der Forschungspraxis weit verbreitet und etabliert, da sie eine pragmatische, schnell interpretierbare und standardisierte Klassifikation ermöglicht, um Interventionen, Vorhersagemodelle oder Therapiealgorithmen miteinander vergleichbar zu machen.

2.3.4. Angewandtes Vorhersagemodell und Leistungsmaß

Ein wesentlicher Grund für die Verwendung einer regularisierten logistischen Regression ist die hohe Dimensionalität und die starke Korrelation der aus den CCT-Daten extrahierten Features. Anders als bei der klassischen GWR-Bestimmung, die nur wenige vorab definierte Regionen und deren HU betrachtet, wurde hier ein umfassendes, atlasbasiertes Vorgehen mit 166 ROIs gewählt. Da bildgebende Features aufgrund der räumlichen biologischen Kontinuität potenziell hoch korrelieren, besteht ein erhebliches Risiko für Multikollinearität, was die Modellkomplexität erhöhen und zu Overfitting führen kann. Overfitting äußert sich darin, dass das Modell die Trainingsdaten sehr gut abbildet, bei neuen, ungesehenen Daten jedoch nicht generalisiert und eine deutlich schlechtere Vorhersageleistung aufweist. Zudem weisen in diesem Fall die Terme höherer Ordnung meist eine breite Streuung mit hoher Wichtung auf. Regularisierungsmethoden wirken dieser Problematik effektiv entgegen, indem sie übermäßige Modellkomplexität, etwa durch die Anzahl der betrachteten Features, bestrafen. Dieses Vorgehen führt zu stabilen und generalisierbaren Vorhersagemodellen.

Das hier angewandte Verfahren basiert auf einer Regularisierung mittels elastischen Netzes, einer gewichteten Kombination aus L1- und L2- Regularisierung. Die L2-Regularisierung bestraft große Parameterwerte quadratisch, wodurch alle Koeffizienten tendenziell kleiner, jedoch nicht vollständig auf null gesetzt werden. Dies ist besonders hilfreich, um den Einfluss hoch korrelierender Features abzumildern und die Modellstabilität zu erhöhen. Die L1-Regularisierung ermöglicht es zudem, die Wichtungen einzelner Features vollständig auf null zu reduzieren. Hierdurch wird eine datengetriebene Feature-Selektion ermöglicht, indem irrelevante Features gezielt aus dem Modell entfernt werden. Damit reduziert sich die effektive Dimensionalität des Vorhersageproblems, was Interpretierbarkeit und Generalisierung verbessert. Das angewandte elastische Netz vereint diese Aspekte der Regularisierung. Dies

ist hier besonders wertvoll, da eine große Anzahl von ROIs einbezogen wird, deren Relevanz für das Outcome a priori nicht eindeutig ist, die jedoch potenziell wertvolle prädiktive Informationen enthalten können. Ferner weist das betrachtete Vorhersageproblem ungefähr die gleiche Anzahl von Features (p = 169) und einbezogenen unabhängigen Fällen (n = 132) auf, wodurch eine Selektion der Features und damit eine Reduktion der Dimensionen für eine gute Generalisierbarkeit hilfreich ist.

Ein weiterer wichtiger Grund für die Wahl eines regularisierten logistischen Regressionsmodells ist dessen intuitive und deutliche Interpretierbarkeit. Im Gegensatz zu komplexen, tiefen neuronalen Netzen, die für ähnlich anspruchsvolle Vorhersagen meist große Datenmengen benötigen und deren interne Entscheidungslogik schwer zu durchdringen ist, lassen sich bei einer logistischen Regression die Aussagen einzelner Features eindeutig zuordnen und somit die Entscheidungsfindung besser nachvollziehen. Dies unterstützt die wissenschaftliche Plausibilitätsprüfung, da geprüft werden kann, ob Regionen, die als besonders prädiktiv identifiziert wurden (z. B. thalamische oder cerebelläre Areale), mit der bekannten Pathophysiologie der hypoxisch-ischämischen Schädigung nach Herzstillstand übereinstimmen.

Die Receiver-Operating-Characteristic-(ROC)-Kurve und der daraus abgeleitete Flächenwert unter der Kurve (Area Under the Curve, AUC) sind in der biomedizinischen Forschung etablierte Leistungsmaße von Klassifikations- und Vorhersagemodellen. Die ROC-AUC zeichnet sich durch ihre Unabhängigkeit von konkreten Grenzwerten aus, wodurch eine objektive Einschätzung der Gesamtleistungsfähigkeit des Vorhersagemodells bei der Unterscheidung zwischen gutem von schlechtem neurologischem Outcome in dem Testdatensatz möglich ist. Je näher der AUC-Wert an eins liegt, desto besser gelingt es dem Modell, gutem von schlechtem neurologischem Outcome zu differenzieren. Die ROC-AUC ist auch deshalb weit verbreitet, weil sie eine gut interpretierbare, dimensionslose Maßzahl liefert, die insbesondere im klinischen Kontext über verschiedene Studien, Datensätze und Methoden hinweg angewendet und damit vergleichbar ist. Dies ist insbesondere relevant, wenn die Vorhersageleistung des Modells einerseits mit der etablierten bildbasierten GWR aus dem CCT und andererseits mit standardisierten Laborparametern, wie der NSE oder NfL, verglichen wird. Aus diesem Grund ist die ROC-AUC ein bedeutendes Leistungsmaß, um ein neues Vorhersagemodell kritisch im Kontext etablierter klinischer Verfahren zu evaluieren, seine relative Über- oder Unterlegenheit transparent darzustellen und letztlich fundierte, datenbasierte Entscheidungen für den klinischen Einsatz zu treffen.

Obwohl die ROC-AUC ein verbreitetes und vielseitiges Leistungsmaß darstellt, ist ihre Aussagekraft bei stark bis mäßig unbalancierten Datensätzen eingeschränkt. In dem vorliegenden Fall neigt die Fallzahl zwischen gutem neurologischem Outcome (49 Patienten) zu schlechtem Outcome (83 Patienten) mit einem Verhältnis von 0,59 leicht zu den Fällen mit

schlechtem Outcome. In solchen Szenarien kann die ROC-AUC dazu neigen, eine zu optimistische Sicht auf die Modellleistung zu vermitteln, da sie gleichermaßen auf Sensitivität (True Positive Rate) und Spezifität (1 - False Positive Rate) abzielt, ohne dabei zu berücksichtigen, dass falsche Klassifizierungen in der unterrepräsentierten Klasse im klinischen Kontext schwerwiegende Konsequenzen haben können. Insbesondere, wenn das Hauptinteresse darin liegt, seltene, aber klinisch besonders bedeutsame Fälle, wie etwa den günstigen neurologischen Verlauf, möglichst zuverlässig vorherzusagen, kann die ROC-AUC zu einer verzerrten Darstellung im Leistungsvergleich der Vorhersagemethoden führen. Aus diesem Grund wurden ferner die Leistungsmaße F1-Score, Balanced Accuracy und die Precision-Recall-Kurve der Vorhersagemethoden betrachtet. Insbesondere sind die Leistungsmaße F1-Score und Balanced Accuracy auch bei stark unbalancierten Datensätze aussagekräftig und liefern so eine ergänzende Einordnung der Vorhersageleistungen. Nichtsdestotrotz stellt die ROC-AUC ein in diesem Kontext geeignetes primäres Leistungsmaß dar, an dem sich Modellverbesserungen im Training orientieren und mit dem sich der vergleichende Leistungsunterschied zwischen den verschiedenen Vorhersagemethoden auf Basis bildbasierter und labormedizinischer Parameter innerhalb derselben Studie ermitteln lässt.

3. Publikation

Die folgende Publikation ersetzt die Punkte Material, Methoden und Ergebnisse der vorliegenden kumulativen Dissertation.

„Analysis of Cerebral CT Based on Supervised Machine Learning as a Predictor of Outcome After Out-of-Hospital Cardiac Arrest“

Hannes Gramespacher*; Maximilian H.T. Schmieschek*; Clemens Warnke; Christoph Adler; Stefan Bittner; Julian Dronse; Nils Richter; Charlotte Zaeske; Carsten Gietzen; Marc Schlamann; Stephan Baldus; Gereon R. Fink; Oezguer A. Onur

Erschienen in: Neurology, Volume 103, Number 1, July 9, 2024

Dies ist eine nicht final editierte Version des Artikels, der in seiner endgültigen Fassung veröffentlicht wurde in: / This is a non-final version of an article published in final form in:
Gramespacher H, Schmieschek MHT, Warnke C, et al. Analysis of Cerebral CT Based on Supervised Machine Learning as a Predictor of Outcome After Out-of-Hospital Cardiac Arrest. Neurology 2024; 103: e209583.
https://www.neurology.org/doi/abs/10.1212/WNL.0000000000209583

Abstract

Background and Objectives: In light of limited intensive care capacities and a lack of accurate prognostic tools to advise caregivers and family members responsibly, this study aims to determine whether automated cerebral computed tomography analysis allows prognostication after out-of-hospital cardiac arrest.

Methods: In this monocentric, retrospective cohort study, a supervised machine learning classifier based on an elastic net regularized logistic regression model for gray matter alterations on non-enhanced cerebral computed tomography obtained after cardiac arrest was trained using a 10-fold cross-validation and tested on a hold-out sample (random split 75% / 25%) for outcome prediction. Following the literature, a favorable outcome was defined as a cerebral performance category of 1-2 and a poor outcome of 3-5. The diagnostic accuracy was compared with established and guideline-recommended prognostic measures within the sample, i.e., gray matter-white matter ratio, neuron-specific enolase, and neurofilament light chain in serum.

Results: Of 279 adult patients, 132 who underwent cerebral computed tomography within 14 days of cardiac arrest with good imaging quality were identified. Our approach discriminated between favorable and poor outcomes with an AUC of 0.73 (95%CI 0.59-0.82). Thus, the prognostic power outperformed the gray-white matter ratio (AUC 0.66, 95%CI 0.56-0.76).

The biomarkers neurofilament light chain, measured at days 1 and 2, and neuron-specific enolase, measured at day 2, exceeded the reliability of the imaging markers derived from computed tomography (AUC NFL day 1: 0.87, 95%CI 0.75-0.99; AUC NFL day 2: 0.90, 95%CI0.79-1.00; AUC NSE day: 2 0.78, 95%CI 0.62-0.94).

Discussion: Our data show that machine learning-assisted gray matter analysis of cerebral computed tomography images offers prognostic information following out-of-

CONFIDENTIAL

hospital cardiac arrest. Thus, CCT gray matter analysis could become a reliable and time-independent addition to the standard workup with serum biomarkers sampled at predefined time points. Prospective studies are warranted to replicate these findings.

Introduction

As one of the leading causes of death in the Western world with various etiologies and low survival rates ranging between 3-16 %[1], out-of-hospital cardiac arrest (OHCA) and its consequences are of enormous importance to healthcare systems worldwide[2].

The neurological outcome and the overall prognosis after OHCA are primarily determined by the extent of hypoxic-ischemic brain injury (HIBI)[3], even more so given that most deaths resulting from OHCA are caused by a withdrawal of life support due to an assumed poor neurological prognosis.[4]

Accordingly, identifying HIBI and reliable outcome prognostication is paramount for informing relatives, making treatment decisions, and using limited intensive care capacities responsibly[5], and, given its implications, presents one of the most demanding challenges with far-reaching consequences for the patient, relatives, and treating staff.

For this reason, reliable algorithms for estimating outcomes after a cardiovascular arrest have been developed to aid decision-making.[6] These algorithms include electrophysiological, clinical, imaging, and laboratory parameters.[7]

The most common, accessible, and guideline-recommended imaging modality following OHCA is cerebral computed tomography (CCT). For outcome prediction, commonly, the gray-white matter ratio (GWR) is determined in several brain regions of interest (ROIs) prone to hypoxia[8–10] to visualize and quantify secondary brain injury[3] indicated by a loss of gray matter (GM) / white matter (WM) differentiation.

CONFIDENTIAL

Although a reduced GWR has been reported to have high specificity for predicting poor neurologic outcomes after resuscitation from cardiac arrest, its sensitivity seems to depend on the time interval between circulatory arrest and imaging, with limited and inconsistent evidence regarding the appropriate timing of imaging.[11–15] In emergencies, the timing of cerebral imaging is often determined by the clinical course and routines, with the determination of an ideal time for imaging playing a rather subordinate role.

The present study investigated the benefits of an objective and structured analysis of non-enhanced CCT imaging acquired within 14 days after admission in patients with OHCA in predicting their overall outcomes.

For this purpose, a supervised machine-learning-assisted approach for automated CCT analysis (aCCTa) was developed, assessing GM changes in the entire brain, in contrast to the GWR that draws on specific ROIs, and temporal information. The present work aimed to determine the additional value of aCCTa compared to established imaging and laboratory prognostic markers after OHCA under the hypothesis that aCCTa would provide a reliable and competitive prognostic value.

CONFIDENTIAL

Methods

Standard Protocol Approvals, Registrations, and Patient Consents

The Ethics Committee of the Medical Faculty of the University of Cologne approved the study and waived informed consent for the retrospective analysis and additional diagnostics in available blood samples (Reference No. 13–086). The study followed the ethical standards of the Declaration of Helsinki. The TRIPOD reporting guidelines were applied.[16]

Study design, setting, and patients

The present study analyzed data from a monocentric, retrospective cohort of patients following non-traumatic OHCA. OHCA patients were admitted to the Department of Cardiology, University Hospital Cologne, between September 2013 and August 2018. Within the study sample, serum NFL and NSE, collected in the context of a previous work[17], were available in a subsample. Per our routine clinical protocol, all patients received intensive care treatment, including targeted temperature management (32-34 °C) for 24 hours. In the overall cohort, imaging was not performed if the patient had no or only minor neurological deficits (CPC 1), had already deceased, or had clinical signs of brain death at admission. For our study, patients were excluded if no CCT was obtained within 14 days of OHCA, if acute primary CNS pathology was present as a possible cause of OHCA (e.g., intracranial hemorrhage), or if signs of elapsed cerebral lesions were identified that could interfere with gray matter analysis (e.g., ischemic stroke). Furthermore, patients were excluded if the image quality was insufficient (e.g., motion artifacts, hardening artifacts, contrast agent administration due to coronary angiography before CCT). As only a minority of the patients received a second CCT, we decided not to include these scans for reasons of model applicability and the trade-off between model complexity and model fit.

CONFIDENTIAL

Clinical variables and laboratory parameters determination

Cardiac arrest was defined as abnormal or absent respiration in a comatose patient with a nonpalpable pulse. Return of spontaneous circulation (ROSC) was defined as a clinically or instrumentally measurable restoration of circulation over several seconds.[18] Time to ROSC was calculated from the time of the detected event to ROSC.

Neurological outcome was assessed at the time of hospital discharge according to the Glasgow-Pittsburgh cerebral performance categories (CPC) ranging from CPC 1 (conscious and alert with good cerebral performance) to CPC 5 (brain dead, circulation preserved).[18–20] CPC categories were dichotomized into "favorable" (CPC 1-2) and "poor" (CPC 3-5) outcomes, as commonly applied.[21,22] NFL and NSE were collected within 3 hours after admission (day 0), after 24 hours (day 1), and after 48 hours (day 2). The NFL concentration was, as previously described[17], determined at the University Medical Center Mainz, Department of Neurology, in duplicates and a blinded fashion using a SiMoA HD-1 device (Quanterix, USA) and NF-Light Advantage Kits (Quanterix) from a single batch according to the manufacturer's instructions. NSE was analyzed routinely in the laboratory of the University Hospital Cologne with an Elecsys NSE assay using the Cobas E801 analytical unit (both Roche Diagnostics, USA). Other clinical variables were collected following the hospital's routine clinical practice.

CONFIDENTIAL

Computed tomography data acquisition

Non-enhanced CCT imaging, in some cases as part of a more comprehensive CT scan, was performed on Philips Brilliance scanners, Siemens Somatom Definition Flash, or Somatom Definition Force scanners per routine clinical protocol. CCT scans were converted from DICOM to NIfTI format using dcm2niix[23], correcting for skew by interpolation (image dimensions 512 x 512 x 328 mm, voxel dimensions 0. 535 x 0.535 x 0.500 mm), without correction for gantry skew, which was taken into account during the segmentation and normalization process. The time from admission to the clinic to imaging, as extracted from the DICOM header (ΔCCT), was included as an additional feature.

126

127 **Segmentation of computed tomography images**

128 Segmentation and spatial normalization of CCT images were performed using the
129 CTSeg toolbox.[24,25] CTSeg makes use of an adaptation of the Statistical Parametric
130 Mapping software's unified segmentation algorithm (SPM12, Wellcome Department of
131 Imaging Neuroscience, London, UK) implemented in MATLAB (MATLAB R2022a;
132 MathWorks Inc., Natick, MA, USA, 2000). The CCT template was used for affine
133 registration and computation of probabilistic segmentation of the brain and intracranial
134 space (gray and white matter, cerebrospinal fluid, skull, soft tissue, background) in
135 native space, as well as warped and modulated in Montreal Neurological Institute
136 (MNI) space.

137

138 **Automated cerebral computed tomography analysis (aCCTa)**

139 We used a supervised machine learning classifier based on an elastic net regularized
140 logistic regression model[26] for outcome prediction. We chose this regularization
141 technique, with a linear combination of L1 and L2 penalties, to achieve better
142 prediction accuracy and generalizability concerning the feature set than using only L1
143 or L2 penalties. In particular, the aim was to account for the high correlation between
144 imaging features extracted from adjacent atlas-based ROIs in direct contrast to the
145 spatially unconstrained biology of an integral cerebral continuum (grouping of
146 features), as well as the relatively large number of features compared to the number
147 of patients. Furthermore, we applied this architecture to allow good interpretability and
148 the possibility of spatial projection of our results onto the brain network. A pure
149 convolutional neural network (CNN) would have required a large amount of labeled

CONFIDENTIAL

patient datasets for sufficient generalizability and would, therefore, have been limited in its interpretability.

A training and hold-out test sample was created via random split (75% / 25%) of the cohort. Besides age, sex, and ΔCCT, extractions of the segmented and modulated mean GM probability (GMP) within each of the 166 ROIs of the automated anatomical labeling 3 (AAL3) brain atlas[27] were used as $n = 169$ features x for each patient (vectors are represented by the bolded $\boldsymbol{x}$. After scaling within the training sample, hyperparameters were optimized via stratified 10-fold cross-validation with elastic net mixing parameters [0, 0.0001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 0.9999, 1] and a maximum number of 1 000 000 iterations, to increase the chosen performance matrix with the python 3 module scikit-learn (version 1.4.1) with dependencies.[28] The random split, the 10-fold cross-validation, the elastic net regularization, and an appropriate quantity of AAL3 ROIs in relation to the cohort size were defined a priori to guarantee a sensible model complexity and avoid overfitting. The area under the curve (AUC) of the receiver operating characteristic (ROC) was selected as a performance matrix to ensure comparability of our approach with the clinically established, guideline-recommended prognostic measures and their performance. The SAGA solver[29] was used with a tolerance for stopping of 10^{-4} to achieve convergence. The coefficients θ (also referred to as weights) and the intercept b of the best score formed the logit z of the sigmoid decision function. This logit z was used in the logistic regression to calculate the probability $P(y = 1 \mid x)$, to binary classify with a threshold of 0.5 y ($y = 1$ for favorable outcome and $y = 0$ for poor outcome) among the features x and to assess the predictive performance of each feature within the hold-out test sample.

$$z = (\sum_{i=1}^{n} \theta_i x_i) + b \mid \theta, x, b \in \mathbb{R}$$

CONFIDENTIAL

$$P(y = 1 \mid x) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

All 166 AAL3 ROIs were scaled to the absolute feature weight θ, to visualize the spatial brain alterations, The scaled ROIs were projected on a brain surface mesh (Surf Ice[30]) and voxel-wise plotted (FSLeyes[31]).

Determination of the gray-white matter ratio

Two experienced radiologists (CZ, CG) with at least four years of training in neuroradiology assessed the GWR blinded, with only age, sex, and ΔCCT known to the raters. Following the literature[8–10], circular GM ROIs (area = 0.1 cm^2) were manually placed in the Putamen (PU) bilaterally, and mean HUs were divided by extracted mean HUs of white matter ROIs, placed in the posterior limb of the internal capsule (PIC), leading to the bilaterally summarized ratio GWR = PU / PIC.

The GWR was first calculated for both raters and their mean values for all CCT scans, regardless of the acquisition time. Besides, to analyze the time dependency of the GWR's prognostic power, the CCT scans were divided into early (GWR_e, within 24 h after admission) and late (GWR_l, after 24 h after admission) scans.

Statistical analysis

Statistical analyses of demographic, imaging, laboratory, and clinical data were performed using R (R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria) and SPSS (IBM Corp. Released 2021. IBM SPSS Statistics for Mac, Version 28.0. Armonk, NY: IBM Corp). Interval scaled values were analyzed using a 2-sample t-test considering Levene's test for equality of variances. The distribution of categorical variables was

CONFIDENTIAL

calculated using a Fisher-Yates test. Testing for normal distribution of NFL and imaging data was performed using the Shapiro-Wilk test. Significance was defined as $p < 0.05$ (two-sided). ROC curves with confidence intervals were calculated for the aCCTa test sample, GWR of both raters, and NSE and NFL at days 0, 1, and 2 via ROCit (version 2.1.1).[32] For further analysis of the supervised machine learning procedure in the context of the mildly unbalanced dataset (i.e., favorable outcome in 37% of the total population), optimal cutpoints were determined and evaluated by maximizing the F1-score via the R package cutpointr.[33] The balanced accuracy at these cutpoints was additionally calculated to compare all prognostic markers.

Data Availability

Anonymized data and the code will be made available upon request from qualified investigators to replicate procedures and results.

CONFIDENTIAL

Results

Study cohort

From 279 patients with OHCA, 145 patients who had received CCT within 14 days of the event were identified. 15 patients underwent a second CT scan, which was not used in the further analyses to increase generalizability and to decrease complexity of the model. After careful visual inspection of the images, 13 patients had to be excluded due to intracerebral hemorrhages (n=2), ischemic stroke without association to OHCA (n=1), and poor imaging quality (n=5 due to motion artifacts, n=1 due to hardening artifacts, and n=4 due to previously received contrast agent for coronary angiography) resulting in a total of 132 patients included in the further analyses (Fig. 1). The early

GWR_e group (i.e., CCT within 24 h after hospital admission) included 56 and the late GWR_l group (i.e., CCT after 24 h after admission comprised 76 scans.

From the study population, 99 patients (18 female, 18%; 81 male, 82%; 0 intersex, 0%) were randomly assigned to the training sample, and 33 patients (8 female, 24%; 25 male, 76%; 0 intersex, 0%) to the test sample. Within the study sample, serum NFL and NSE were available for 33 patients.

Favorable outcome (CPC 1-2) was noted in 36 patients of the training sample (36%) and in 13 patients of the test sample (39%). The mean age was 62 years in the training sample and thus differed from the mean age of 57 years in the test samples. The remaining clinical and demographic variables did not differ between the training and test samples (Table 1).

Comparing the groups with favorable and poor outcomes in the total study population, the mean age was significantly lower in the group with a favorable outcome. Cardiogenic causes of cardiac arrest and a shockable rhythm were significantly more common in this group. Moreover, the favorable outcome group had a significantly lower no-flow time and longer intensive care unit hospitalization.

In the poor outcome group, NFL levels were higher at 24 hours after admission, and NSE levels were higher at 24 and 48 hours after admission (Table 2).

Outcome prediction using the automated cerebral computed tomography analysis (aCCTa)

The mathematical model fitting converged in the training sample successfully within the tolerance for stopping, the averaged inverse of the regularization strength of the best scores for each fold of $C = 21.544$, and an elastic-net mixing parameter of 0.25. The best fitted logit z had an intercept of $b = 0.376$ and the coefficients θ of the

CONFIDENTIAL

corresponding features x, as listed in the supplemental content (eTable 1). In the hold-out test sample, the performance measure AUC was 0.73 (95%CI 0.59-0.82), the F1-score was 0.67, and the balanced accuracy was 0.72 (sensitivity (TPR) 0.69; specificity 0.75; positive likelihood ratio (LR+, TPR/FPR) 2.76). Among the corresponding features, areas located in the midbrain, including the substantia nigra, and in cerebellar and thalamic regions, showed the greatest explanatory power for predictive decision-making with the highest absolute weights $|\theta|$ (Fig. 2, Supplementary Table 1).

Prognostic power of the aCCTa compared to GWR, NSE, and NFL

The prognostic power of the aCCTa (AUC 0.73, 95%CI 0.59-0.82) outperformed that of the GWR of both raters (AUC Rater #1: 0.68, 95%CI 0.58-0.78; AUC Rater #2: 0.64, 95%CI 0.54-0.74), independent of the time of CT imaging (AUC GWR (mean) 0.66, 95%CI 0.56-0.76; AUC GWR_e 0.67, 95%CI 0.53-0.81, and of GWR_l 0.65, 95%CI 0.51-0.79). Moreover, the prognostic power was better than that of the neurofilament light chain (AUC NFL 0.59, 95%CI 0.39-0.78) and neuron-specific enolase (AUC NSE 0.60, 95%CI 0.40-0.79) at day 0. The prognostic power of the aCCTa was comparable to that of NSE at day 1 (AUC 0.74, 95%CI 0.57-0.91). With elapsing time from the admission, the prognostic power of the blood parameters NSE at day 2 (AUC 0.78, 95%CI 0.62-0.94), NFL at day 1 (AUC 0.87, 95%CI 0.75-0.99), and NFL at day 2 (AUC 0.90, 95%CI 0.79-1.0) increased and outperformed that of the aCCTa approach (Table 3, Fig. 3). The precision-recall curves for the prognostic markers studied are provided in the supplemental material (eFigure 1).

CONFIDENTIAL

Discussion

This pilot study provides evidence that a machine-learning-assisted analysis of GM in non-enhanced cerebral computed tomography allowed a reliable, fast, and observer-independent outcome prognostication following OHCA and could compete with established prognostic markers such as NFL, NSE, and GWR.

The present approach is based on detecting alterations in segmented gray matter, more precisely in the sense of gray matter probability (GMP). Alterations in GMP after OHCA are expected to occur in brain regions with higher energy demand and along the border zones of vascular supply areas. These areas are affected by edema following hemodynamically induced hypoperfusion, which finally leads to a loss of gray and white matter differentiation[34], causing changes that can already be identified as early as 2 hours after ROSC.[35] Our approach is therefore based on machine learning-assisted weighting of the relevance of cerebral ROIs for prognosis after OHCA, also taking into account the timing of the CT, age, and sex.

Among the ROIs with the most significant explanatory power for predictive decision-making that were identified using our approach are areas in the midbrain, particularly the substantia nigra, as well as cerebellar and thalamic areas. The cerebellum is known to be one of the brain regions particularly vulnerable to hypoxic-ischemic damage after OHCA. [36,37] The thalamus, as a metabolically highly active region, is also known to be an area at risk of hypoxia and is considered responsible for a poor outcome following cardiac arrest[38] - as well as lesions in the substantia nigra, albeit evidence is limited.[39]

Integrating the 163 AAL3 atlas regions in a whole-brain analysis differs significantly from the GWR approach, which considers only a few manually placed two-dimensional

CONFIDENTIAL

ROIs. The aCCTa approach allowed prognostication with an AUC of 0.73, exceeding the predictive power of GWR as assessed in the same study population by two independent and blinded raters with extensive expertise in neuroradiology (AUC 0.68 and 0.65). The prognostic power of GWR described in the literature varies considerably (AUCs ranging between 0.59 and 0.95) depending on the study design, the timing of imaging, and the ROIs selected.[9,15] Consequently, the identification of suitable regions and the correct manual positioning of the measuring points are crucial for determining the GWR. Regarding suitable regions, determining the GWR at the level of the basal ganglia, as also carried out in our study, seems to be the most reliable approach.[8,15] This, in turn, requires a neuroradiological expertise and is a potential source of failure. Although the radiological raters' assessment could contain a subjective component, the GWR has recently been shown to have low inter-rater variability and comparable accuracy to a computer-based method but still showed considerable inter-rater variability in some individual patients[40]. This imprecision can be of great individual significance given the consequences. First attempts aiming at automated GWR analysis as a more objective and standardized alternative to human raters are promising but largely consistent with the manual approach, given the theoretical background.[41,42] However, promising results with a different theoretical approach were recently obtained in the study of a smaller cohort using deep learning-based algorithms for the early detection of HIBI after OHCA.[43]

A confounding factor in determining the GWR is that distinct cutoffs based on HU ratios are recommended for prognostication. However, different scanner types can show considerable variability in the HUs of GM and WM of up to 23%, which limits the interpretation and transferability of the results.[44] While these differences in HUs on different CT scanners might also impact the validity of our aCCTa approach, it remains

CONFIDENTIAL

323 to be investigated whether these inter-scanner differences result in an intentionally
324 accepted loss of segmentation accuracy. However, inter-scanner differences are
325 expected to be of minor importance as the CCT scans were registered to a
326 standardized space (MNI) during the preprocessing of the CT scans.
327
328 The assumed loss of segmentation accuracy, reflecting hypoxia-induced loss of
329 differentiation of cortex and white matter, ultimately leading to changes in GMP, may
330 represent an advantage of our method, as it considers the brain tissues in a volumetric
331 perspective besides the information provided by the HUs and may, therefore,
332 contribute indirectly to the identification of areas that are particularly at risk of hypoxia.
333
334 The time point of the first (and often only) cerebral imaging is typically determined by
335 the emergency with cerebral imaging in the first few hours after admission, e.g., to rule
336 out a central nervous cause of OHCA or is performed several days after admission as
337 coma persists. The literature concerning the best prognostic power of GWR is partly
338 ambiguous regarding the timing of imaging following OHCA and the prognostic power
339 of the GWR is likely limited within 12 to 24 h after OHCA since GWR decreases with
340 increasing time from the event.[11,12,45] GWR determination in late scans (>24 h) could,
341 therefore, be of relatively low prognostic value. In our analyses, however, no significant
342 difference in the prognostic power of the GWR from early (<24 h) and late (>24 h) CCT
343 scans was found, which reflects the uncertainty regarding the ideal timing of GWR
344 determination following.
345 To consider the timing of CT imaging after hospital admission, we included ΔCCT as
346 a feature in the aCCTa approach. However, the weighting of the coefficients indicated
347 that the time of imaging ($\theta_2 = -0.04882$) was of minor importance for prognostication

CONFIDENTIAL

in our approach (eTable 1). In summary, aCCTa can provide valuable prognostic information within 14 days after the event, regardless of the exact timing, including even relatively early stages after the OHCA.

Thus, another advantage of our method could be the identification of more discrete signs of cerebral hypoxia occur in the first hours after OHCA. This assumption is supported by recent work, in which HIBI was reliably identified using deep transfer learning-based analysis of CCT scans acquired within 3 hours of ROSC[43] and the observation that HIBI can be detected on native CCT scans within 2 hours of ROSC.[35]

Compared to the blood parameters NSE and NFL, aCCTa outperformed these biomarkers on day 0. The prognostic value of these markers increased subsequently, with NSE outperforming aCCTa on day 2 and NFL already on day 1. Notably, the diagnostic values of NSE and NFL have only been examined at specific time points or about the dynamics between these time points.[17,46]

NSE is the most established and extensively investigated biomarker for neuroprognostication in patients with cardiac arrest.[6,47] NSE is found in neurons and neuroendocrine tumors, perineuronal tissue, red blood cells, and platelets.[48] Thus, the presence of NSE in blood components could limit its prognostic value as hemolysis is often present in critically ill patients. The superiority of NFL over NSE for neuroprognostication has been demonstrated recently.[17,46] This superiority could be because NFL, as part of the neuronal cytoskeleton, is only expressed in neurons and, therefore, more accurately reflects damage to the central nervous system. Our work does not diminish the importance of NSE and NFL for prognostication. Instead, our data suggest that clinically obtained CCT can significantly contribute to the diagnostic workup after OHCA, aiding to determine the prognosis as accurately as possible and

CONFIDENTIAL

might be even more relevant in cases where the NSE and NFL were not determined in the first couple of days.

Limitations

Given this study's retrospective nature, the features to determine the criteria of severe HIBI according to the prognostication strategy algorithm proposed by the European Resuscitation Council and the European Society of Intensive Care Medicine[49] were not available for all patients. Consequently, it is difficult to distinguish between a poor outcome due to severe brain injury and other causes. Since severe HIBI can be expected in two-thirds of patients with a poor outcome after OHCA[50], our approach primarily aimed to identify less severe HIBI in a heterogeneous collective in order to be able to transfer our results to a broad clinical population. Another limitation is that the image quality was not always suitable for our approach, especially if a contrast agent had been administered due to cardiac catheterization before CCT.

Moreover, the classification of patients into favorable and poor outcomes is a known problem in studies of prognosis after OHCA and commonly poses challenges to their interpretation. The treating physicians could, e.g., see the patient's findings (e.g., imaging and neurophysiology) and thus be influenced in their decision on continuation or withdrawal of life-sustaining therapies, which could have led to a self-fulfilling prophecy in terms of an incorporation bias. The fact that a relevant number of deaths in the context of cardiac arrest is ultimately caused by an (early) withdrawal of life support[4] is another confounding factor that could bias the classification of patients into favorable and poor outcomes, which, however, is common to almost all studies investigating prognosis after cardiac arrest. The fact that there was an age difference in the training and test sample was due to the purely random split of the study

CONFIDENTIAL

population and is, therefore, deliberately tolerated. Furthermore, the study population did not adequately consider ethnicity, sex, and age. Statistical analyses were corrected for sex and age to address these issues at least in part. When considering the number of patients studied, the results' generalizability may be limited, especially in more ethnically heterogeneous populations.

Prospects

Not only in times of global pandemics but also given massive regional differences in healthcare capacities, valid methods for determining the prognosis following OHCA and the resulting therapy decisions are of considerable importance to distribute limited intensive care resources responsibly and to advise caregivers professionally. Our data suggest that an automated analysis of CCT scans can provide an unbiased and easily accessible additional value to the prognostic framework following OHCA, thereby significantly contributing to the diagnostic workup after OHCA and accurate prognostication.

CONFIDENTIAL

Conclusion

This pilot study provides evidence that automated CCT analysis could offer a simple, observer-independent, cost-effective, and objective addition to the prognostic framework following OHCA. Notably, the suggested approach derives additional information from a resource used for clinical reasons, unlike the serum biomarkers NSE and NFL, which are sampled exclusively for prognostic purposes at predefined time points[17,46]. Further, our approach provides a better prognostic value than the established imaging approach GWR. Given the study design and its associated limitations discussed above, our results need to be considered preliminary, suggesting

that our approach warrants further validation in larger and more heterogeneous populations and prospective multicenter studies.

Acknowledgments

HG was supported by the Cologne Clinician Scientist Program (CCSP) / Faculty of Medicine / University of Cologne. Funded by the German Research Foundation (DFG, FI 773/15-1). MS received support from Alzheimer Forschung Initiative e.V. This work was also supported by a Marga and Walter Boll-Foundation grant to OO and GRF.

References

1. Chai J, Fordyce CB, Guan M, et al. The association of duration of resuscitation and long-term survival and functional outcomes after out-of-hospital cardiac arrest. Resuscitation. Epub 2022.

2. Berdowski J, Berg RA, Tijssen JGP, Koster RW. Global incidences of out-of-hospital cardiac arrest and survival rates: Systematic review of 67 prospective studies. Resuscitation. 2010;81:1479–1487.

3. Sekhon MS, Ainslie PN, Griesdale DE. Clinical pathophysiology of hypoxic ischemic brain injury after cardiac arrest: a “two-hit” model. Crit Care. 2017;21:90.

4. Dragancea I, Wise MP, Al-Subaie N, et al. Protocol-driven neurological prognostication and withdrawal of life-sustaining therapy after cardiac arrest and targeted temperature management. Resuscitation. 2017;117:50–57.

5. Sandroni C, Grippo A, Nolan JP. ERC-ESICM guidelines for prognostication after cardiac arrest: time for an update. Intens Care Med. 2020;46:1901–1903.

6. Sandroni C, D’Arrigo S, Nolan JP. Prognostication after cardiac arrest. Crit Care. 2018;22:150.

7. Moseby-Knappe M, Westhall E, Backman S, et al. Performance of a guideline-recommended algorithm for prognostication of poor neurological outcome after cardiac arrest. Intens Care Med. 2020;46:1852–1862.

8. Gentsch A, Storm C, Leithner C, et al. Outcome Prediction in Patients After Cardiac Arrest: A Simplified Method for Determination of Gray–White Matter Ratio in Cranial Computed Tomography. Clin Neuroradiol. 2015;25:49–54.

9. Wang WJ, Cui J, Lv GW, et al. Prognostic Values of the Gray-to-White Matter Ratio on Brain Computed Tomography Images for Neurological Outcomes after

CONFIDENTIAL

456 Cardiac Arrest: A Meta-Analysis. Martorana"] ["Alessandro, editor. Biomed Res Int
457 [online serial]. 2020;2020:7949516. Accessed at:
458 https://doi.org/10.1155/2020/7949516. Accessed December 13, 2023.

459 10. Zhou F, Wang H, Jian M, et al. Gray-White Matter Ratio at the Level of the Basal
460 Ganglia as a Predictor of Neurologic Outcomes in Cardiac Arrest Survivors: A
461 Literature Review. Frontiers Medicine. 2022;9:847089.

462 11. Kirsch K, Heymel S, Günther A, et al. Prognostication of neurologic outcome
463 using gray-white-matter-ratio in comatose patients after cardiac arrest. Bmc Neurol.
464 2021;21:456.

465 12. Scheel M, Storm C, Gentsch A, et al. The prognostic value of gray-white-matter
466 ratio in cardiac arrest patients treated with hypothermia. Scand J Trauma, Resusc
467 Emerg Med. 2013;21:23.

468 13. Wang WJ, Cui J, Lv GW, et al. Prognostic Values of the Gray-to-White Matter
469 Ratio on Brain Computed Tomography Images for Neurological Outcomes after
470 Cardiac Arrest: A Meta-Analysis. Biomed Res Int. 2020;2020:7949516.

471 14. Zhou F, Wang H, Jian M, et al. Gray-White Matter Ratio at the Level of the Basal
472 Ganglia as a Predictor of Neurologic Outcomes in Cardiac Arrest Survivors: A
473 Literature Review. Frontiers Medicine. 2022;9:847089.

474 15. Na MK, Kim W, Lim TH, et al. Gray matter to white matter ratio for predicting
475 neurological outcomes in patients treated with target temperature management after
476 cardiac arrest: A systematic review and meta-analysis. Resuscitation. 2018;132:21–
477 28.

478 16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a
479 multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the
480 TRIPOD Statement. BMC Med. 2015;13:1.

481 17. Adler C, Onur OA, Braumann S, et al. Absolute serum neurofilament light chain
482 levels and its early kinetics predict brain injury after out-of-hospital cardiac arrest. J
483 Neurol. 2022;269:1530–1537.

484 18. Jacobs I, Nadkarni V, Bahr J, et al. Cardiac Arrest and Cardiopulmonary
485 Resuscitation Outcome Reports. Circulation. 2004;110:3385–3397.

486 19. Langhelle A, Nolan J, Herlitz J, et al. Recommended guidelines for reviewing,
487 reporting, and conducting research on post-resuscitation care: The Utstein style.
488 Resuscitation. 2005;66:271–283.

489 20. Rittenberger JC, Raina K, Holm MB, Kim YJ, Callaway CW. Association between
490 Cerebral Performance Category, Modified Rankin Scale, and discharge disposition
491 after cardiac arrest. Resuscitation. 2011;82:1036–1040.

CONFIDENTIAL

21. Bernard SA, Gray TW, Buist MD, et al. Treatment of Comatose Survivors of Out-of-Hospital Cardiac Arrest with Induced Hypothermia. New Engl J Medicine. 2002;346:557–563.

22. Grossestreuer AV, Abella BS, Sheak KR, et al. Inter-rater reliability of post-arrest cerebral performance category (CPC) scores. Resuscitation. 2016;109:21–24.

23. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J Neurosci Meth. 2016;264:47–56.

24. Adduru V, Baum SA, Zhang C, et al. A Method to Estimate Brain Volume from Head CT Images and Application to Detect Brain Atrophy in Alzheimer Disease. Am J Neuroradiol. 2020;41:224–230.

25. Brudfors M, Balbastre Y, Flandin G, Nachev P, Ashburner J. Flexible Bayesian Modelling for Nonlinear Image Registration. Arxiv. Epub 2020.

26. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc: Ser B (Stat Methodol). 2005;67:301–320.

27. Rolls ET, Huang C-C, Lin C-P, Feng J, Joliot M. Automated anatomical labelling atlas 3. NeuroImage. 2020;206:116189.

28. Pedregosa, Varoquaux F and, Gramfort G and, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. Epub 2011.:2825–2830.

29. Defazio A, Bach F, Lacoste-Julien S. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. arXiv. Epub 2014.

30. Surf Ice [online]. Accessed at: https://www.nitrc.org/projects/surfice/). Accessed December 13, 2023.

31. FSLeyes [online]. Accessed at: https://zenodo.org/records/8376979. Accessed December 13, 2023.

32. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Statist Med. 1997;16:2143–2156.

33. Thiele C, Hirschfeld G. cutpointr : Improved Estimation and Validation of Optimal Cutpoints in R. J Stat Softw. 2021;98.

34. Kjos B, Brant-Zawadzki M, Young R. Early CT findings of global central nervous system hypoperfusion. Am J Roentgenol. 1983;141:1227–1232.

35. Schick A, Prekker ME, Kempainen RR, et al. Association of hypoxic ischemic brain injury on early CT after out of hospital cardiac arrest with neurologic outcome. Am J Emerg Medicine. 2022;54:257–262.

CONFIDENTIAL

528 36. Horn M, Schlote W. Delayed neuronal death and delayed neuronal recovery in
529 the human brain following global ischemia. Acta Neuropathol. 1992;85:79–87.

530 37. Ng T, Graham DI, Adams JH, Ford I. Changes in the hippocampus and the
531 cerebellum resulting from hypoxic insults: frequency and distribution. Acta
532 Neuropathol. 1989;78:438–443.

533 38. Sohn G, Kim SE. Measurement of thalamus and cortical damages in hypoxic
534 ischemic encephalopathy. IBRO Neurosci Rep. 2023;15:179–185.

535 39. Fujioka M, Okuchi K, Sakaki T, Hiramatsu K, Miyamoto S, Iwasaki S. Specific
536 changes in human brain following reperfusion after cardiac arrest. Stroke.
537 2018;25:2091–2095.

538 40. Kenda M, Cheng Z, Guettler C, et al. Inter-rater agreement between humans and
539 computer in quantitative assessment of computed tomography after cardiac arrest.
540 Front Neurol. 2022;13:990208.

541 41. Kenda M, Scheel M, Kemmling A, et al. Automated Assessment of Brain CT
542 After Cardiac Arrest—An Observational Derivation/Validation Cohort Study. Crit Care
543 Med. 2021;49:e1212–e1222.

544 42. Hanning U, Sporns PB, Lebiedz P, et al. Automated assessment of early hypoxic
545 brain edema in non-enhanced CT predicts outcome in patients after cardiac arrest.
546 Resuscitation. 2016;104:91–94.

547 43. Mansour A, Fuhrman JD, Ammar FE, et al. Machine Learning for Early Detection
548 of Hypoxic-Ischemic Brain Injury After Cardiac Arrest. Neurocrit Care. 2022;36:974–
549 982.

550 44. Oh JH, Choi SP, Wee JH, Park JH. Inter-scanner variability in Hounsfield unit
551 measured by CT of the brain and effect on gray-to-white matter ratio. Am J Emerg
552 Medicine. 2019;37:680–684.

553 45. Wang G-N, Chen X-F, Lv J-R, Sun N-N, Xu X-Q, Zhang J-S. The prognostic
554 value of gray–white matter ratio on brain computed tomography in adult comatose
555 cardiac arrest survivors. J Chin Med Assoc. 2018;81:599–604.

556 46. Moseby-Knappe M, Mattsson N, Nielsen N, et al. Serum Neurofilament Light
557 Chain for Prognosis of Outcome After Cardiac Arrest. Jama Neurol. 2019;76:64.

558 47. Huntgeburth M, Adler C, Rosenkranz S, et al. Changes in Neuron-Specific
559 Enolase are More Suitable Than Its Absolute Serum Levels for the Prediction of
560 Neurologic Outcome in Hypothermia-Treated Patients with Out-of-Hospital Cardiac
561 Arrest. Neurocrit Care. 2014;20:358–366.

562 48. Marangos PJ, Campbell IC, Schmechel DE, Murphy DL, Goodwin FK. Blood
563 Platelets Contain a Neuron-Specific Enolase Subunit. J Neurochem. 1980;34:1254–
564 1258.

CONFIDENTIAL

565 49. Nolan JP, Sandroni C, Böttiger BW, et al. European Resuscitation Council and
566 European Society of Intensive Care Medicine Guidelines 2021: Post-resuscitation
567 care. Resuscitation. 2021;161:220–269.

568 50. Lemiale V, Dumas F, Mongardon N, et al. Intensive care unit mortality after
569 cardiac arrest: the relative contribution of shock and brain injury in a large cohort.
570 Intensive Care Med. 2013;39:1972–1980.

571

CONFIDENTIAL

Clinical characteristics	Training sample (n = 99; 75%)	Test sample (n = 33; 25%)	p-value
Age (years ± SD)	62.31 ± 14.7	57.36 ± 11.1	0.045*
Female/Male/Intersex (%)	18/81/0 (18/82/0)	8/25/0 (24/76/0)	0.456
Cause of cardiac arrest			
Cardiogenic shock	79 (80)	26 (79)	1.000
Others – n (%)	20 (20)	7 (21)	1.000`
Cardiac arrest characteristics			
No-flow-time (min ± SD)	3.30 ± 2.7 [a]	4.70 ± 4.0 [b]	0.098
BLS provided by bystander – n (%)	55 (56)	13 (39)	0.114
Shockable rhythm – n (%)	61 (62)	25 (76)	0.205
Number of shocks ± SD	3.47 ± 4.4	2.79 ± 3.0	0.320
Time to ROSC (min ± SD)	26.66 ± 17.4 [c]	25.06 ± 17.2 [d]	0.658
Computed tomography characteristics			
Time to head CT (days ± SD)	2.54 ± 3.0	1.94 ± 2.2	0.658
Outcome			
Favorable outcome (CPC 1-2)	36 (36)	13 (39)	0.836
Period of ICU hospitalization (days ± SD)	14.58 ± 11.1	14.97 ± 13.0	0.867
Ventilation time (days ± SD)	9.06 ± 7.8	9.79 ± 9.3	0.659

Table 1. Clinical characteristics of the randomly selected training and test sample. BLS = basic life support, ROSC = return of spontaneous circulation, CT = computed tomography, ICU = intensive care unit, CPC = cerebral performance category, SD = standard deviation. [a] 23 datasets missing. [b] 6 datasets missing. [c] 7 datasets missing. [d] 2 datasets missing

CONFIDENTIAL

Clinical characteristics	Study population (n = 132)	Favorable outcome (CPC 1-2) (n = 49)	Poor outcome (CPC 3-5) (n = 83)	p-value
Age (years ± SD)	61.08 ± 14.0	56.14 ± 15.4	63.99 ± 12.4	0.002*
Female/Male/Intersex (%)	26/106/0 (20/80/0)	9/40/0 (18/82/0)	17/66/0 (20/80/0)	0.824
Cause of cardiac arrest				
Cardiogenic shock	105 (80)	47 (96)	58 (70)	< 0.001*
Others – n (%)	27 (20)	2 (4)	25 (30)	< 0.001`
Cardiac arrest characteristics				
No-flow-time (min ± SD)	3.67 ± 3.2 [a]	2.93 ± 1.9 [b]	4.18 ± 3.7 [c]	0.027*
BLS provided by bystander – n (%)	68 (52)	29 (59)	39 (47)	0.209
Shockable rhythm – n (%)	86 (65)	45 (92)	41 (49)	< 0.001*
Number of shocks ± SD	3.30 ± 4.1	4.10 ± 4.7	2.83 ± 3.7	0.087
Time to ROSC (min ± SD)	26.26 ± 17.3 [d]	24.08 ± 18.6 [e]	27.65 ± 16.4 [f]	0.266
Computed tomography characteristics				
Time to head CT (days ± SD)	2.39 ± 2.8	1.96 ± 2.6	2.55 ± 2.8	0.406
NFL and NSE characteristics [g]				
Mean sNFL at Day 0 (pg/ml ± SD)	33.0 ± 41.0	32.5 ± 46.9	33.7 ± 34.2	0.937
Mean sNFL at Day 1 (pg/ml ± SD)	2285.4 ± 6132.5	65.1 ± 57.6	4949.7 ± 8486.0	0.043*
Mean sNFL at Day 2 (pg/ml ± SD)	57812.0 ± 19047.0	128.3 ± 154.5	12764.4 ± 27127.3	0.093
Mean NSE at Day 0 (pg/ml ± SD)	43.6 ± 16.5	41.7 ± 15.8	45.9 ± 17.5	0.467
Mean NSE at Day 1 (pg/ml ± SD)	57.8 ± 53.5	36.7 ± 15.2	83.0 ± 70.8	0.025*
Mean NSE at Day 2 (pg/ml ± SD)	64.2 ± 89.6	30.3 ± 14.0	104.9 ± 121.9	0.033*
Outcome				
Period of ICU hospitalization (days ± SD)	14.67 ± 11.6	17.59 ± 8.3	12.95 ± 12.9	0.013*
Ventilation time (days ± SD)	9.24 ± 8.2	8.10 ± 6.4	9.92 ± 9.0	0.180

Table 2. Clinical characteristics of the study population. CPC = cerebral performance category, BLS = basic life support, ROSC = return of spontaneous circulation, CT = computed tomography, ICU = intensive care unit, SD = standard deviation. [a] 29 datasets missing. [b] 7 datasets missing. [c] 22 datasets missing. [d] 9 datasets missing. [e] 1 dataset missing. [f] 8 datasets missing. [g] 31 datasets missing in the group with favorable outcomes and 68 missing in the group with poor outcomes, within 3h (Day 0), at 24h (Day 1), and at 48h (Day 2).

CONFIDENTIAL

Prognostic Marker	AUC	CI 5%-95%	Optimal Cutpoint	F1-score	Balanced Accuracy
aCCTa [a]	**0.731**	0.59-0.82 [b]	Decision Boundary $P(y = 1 \mid x)$	0.667	0.721
GWR Rater #1	**0.678**	0.58-0.78	1.1753	0.600	0.639
GWR Rater #2	**0.642**	0.54-0.74	1.2154	0.574	0.642
GWR (mean)	**0.660**	0.56-0.76	1.1954	0.587	0.641
GWR_e [c]	**0.670**	0.53-0.81	1.1968	0.670	0.666
GWR_l [c]	**0.653**	0.51-0.79	1.2472	0.542	0.658
NFL Day 0 [d]	**0.589**	0.39-0.78	15.6882	0.667	0.650
NFL Day 1 [d]	**0.874**	0.75-0.99	63.2934	0.823	0.828
NFL Day 2 [d]	**0.900**	0.79-1	84.3125	0.811	0.806
NSE Day 0 [d]	**0.598**	0.40-0.79	38	0.629	0.617
NSE Day 1 [d]	**0.739**	0.57-0.91	55	0.759	0.783
NSE Day 2 [d]	**0.779**	0.62-0.94	56	0.769	0.806

Table 3. Comparison of the prognostic markers. AUC = area under the curve, CI = confidence interval, acCTa = automated cCT analysis, GWR = gray-white matter ratio, NFL = neurofilament light chain in serum, NSE = neuron-specific enolase, [a] within Test Sample, [b] calculated over 10 fold distribution within the training sample, [c] mean GWR of both raters using CCTs acquired within 24h (GWR_e) and after 24h (GWR_l) after admission. [d] 31 datasets missing in the group with favorable outcomes and 68 missing in the group with poor outcomes within 3h (Day 0), at 24h (Day 1), and at 48h (Day 2).

CONFIDENTIAL

Figure legends

Fig 1. Flow of participants.

Fig 2. Spatial distribution of the feature coefficients θ. AAL3 ROIs, scaled to the absolute value of the feature coefficients. a. superior (top) view, b. basal (bottom) view, c. dorsal view, d. coronal view, e. transversal view. AAL3 = automated anatomical labeling atlas. ROI = region of interest.

Fig 3. Receiver operating characteristic (ROC) comparison of prognostic markers. a. Automated cerebral computed tomography analysis (aCCTa), b. Gray-white matter ratio of both raters (GWR Rater #1 and GWR Rater #2), c. Neurofilament light chain in serum (NFL) for day 0 (within 3h), day 1 (>3h and <24h), and day 2 (>24h and <48h), d. Neuron-specific enolase in serum (NSE) for day 0, day 1, and day 2.

CONFIDENTIAL

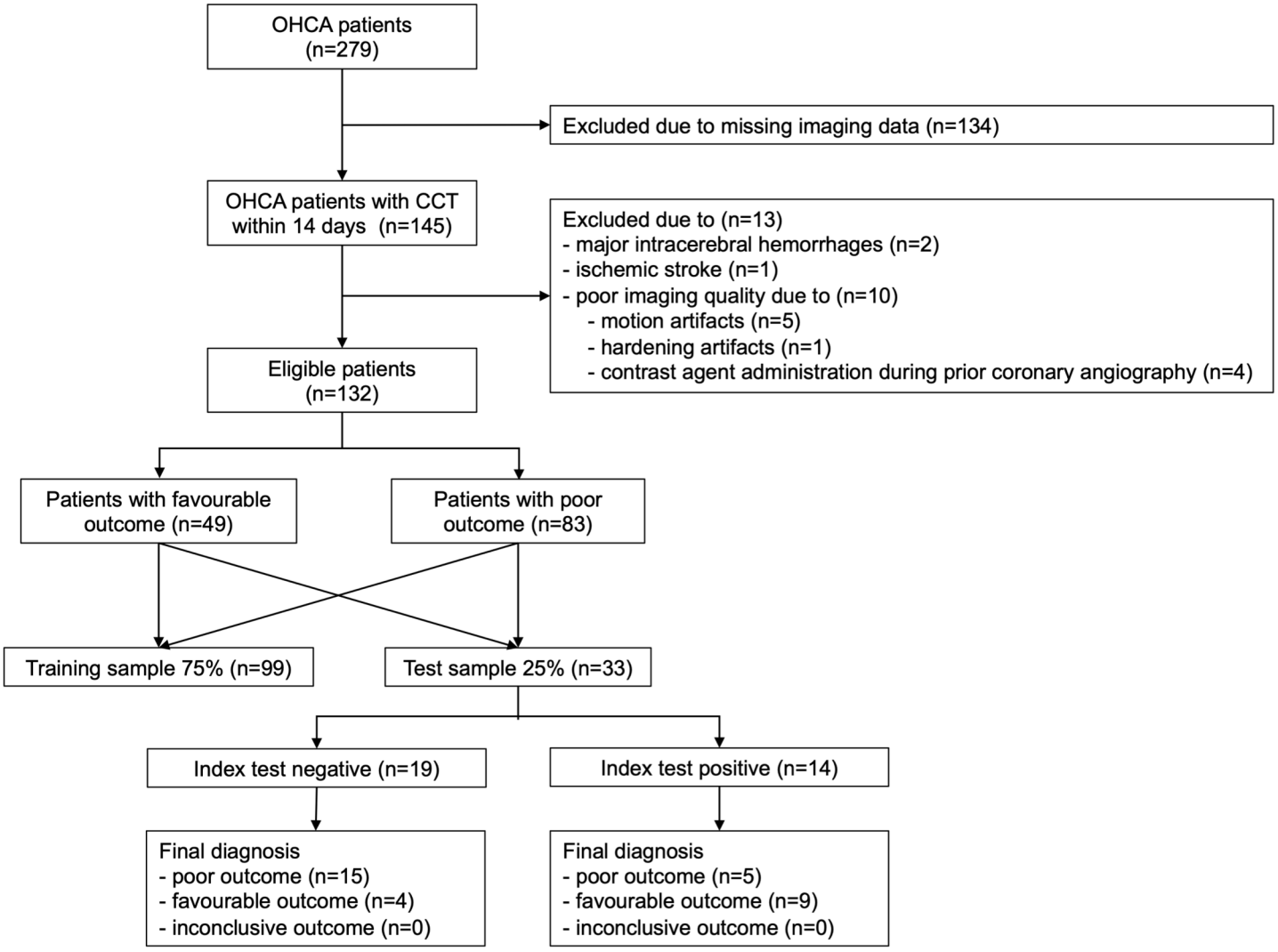


Fig 1. Flow of participants.

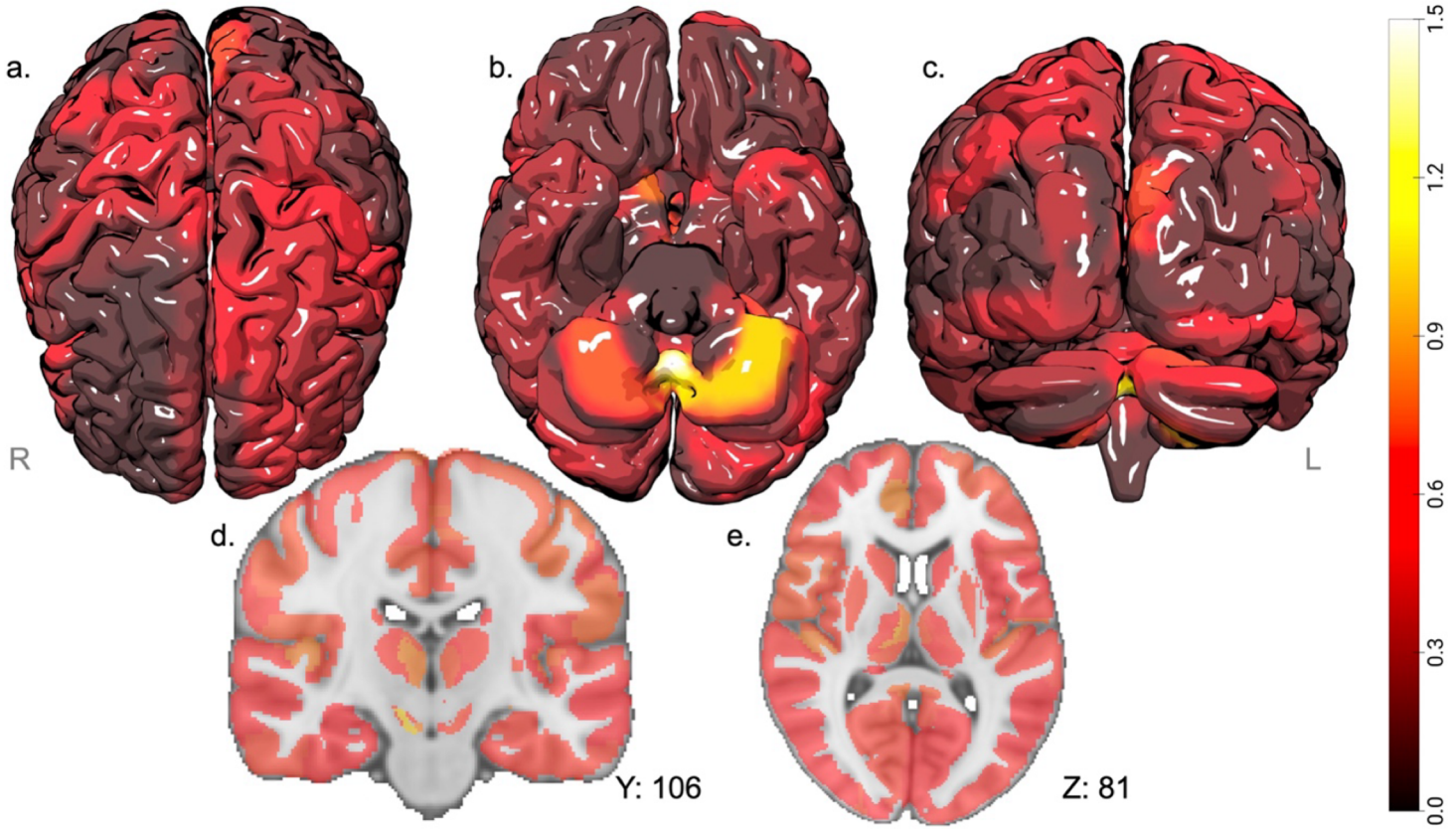


Fig 2. Spatial distribution of the feature coefficients θ. AAL3 ROIs, scaled to the absolute value of the feature coefficients. a. superior (top) view, b. basal (bottom) view, c. dorsal view, d. coronal view, e. transversal view. AAL3 = automated anatomical labeling atlas. ROI = region of interest.

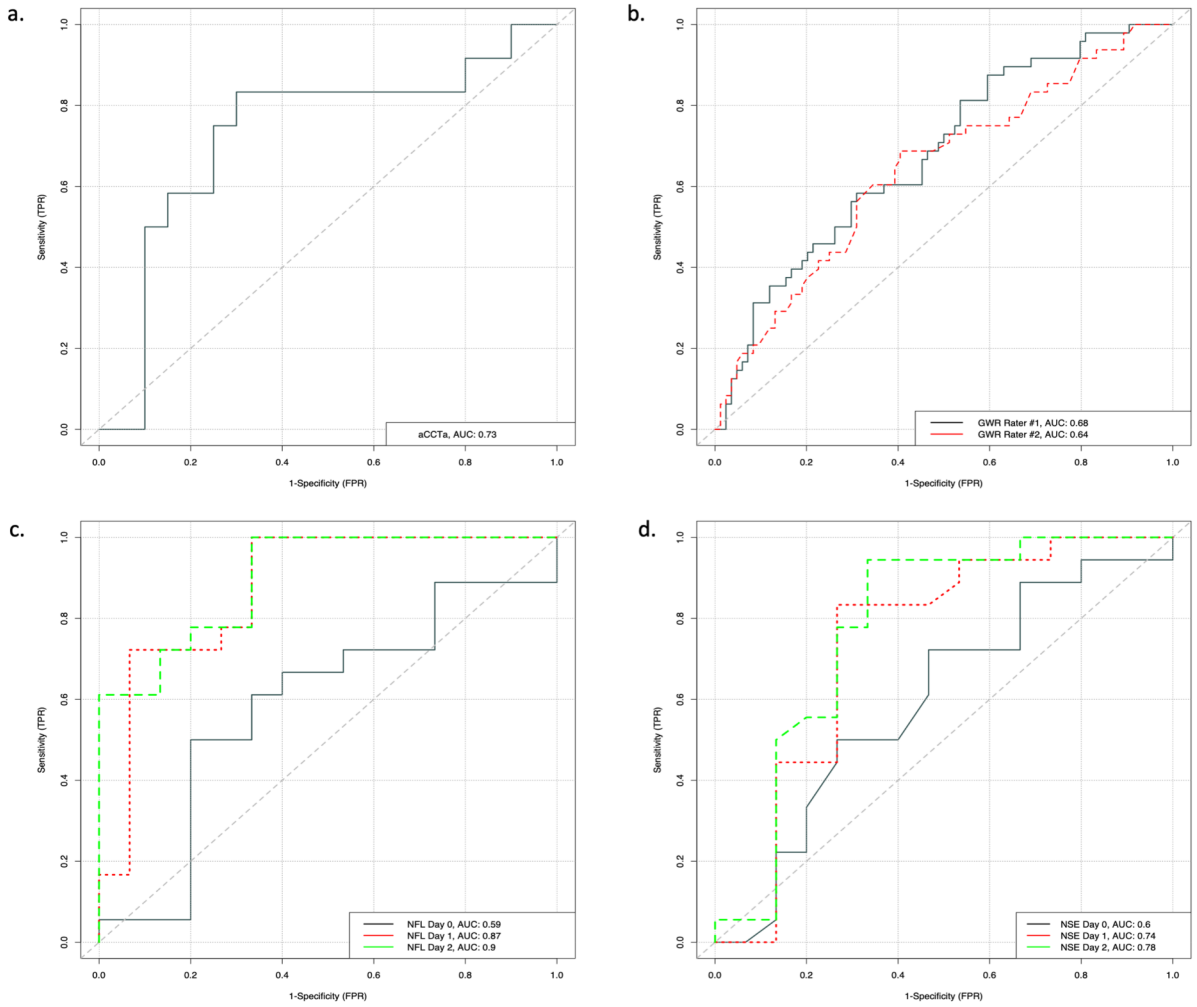


Fig 3. Receiver operating characteristic (ROC) comparison of prognostic markers.

Supplementary Content

eTable 1

Feature Number	Absolute Coefficient θ	Coefficient θ	AAL ROI Number / Feature Name	Label AAL
117	1.510771808	-1.510771808	119	Vermis_9
160	1.262403372	1.262403372	162	SN_pc_R
116	1.110206908	-1.110206908	118	Vermis_8
105	0.989278597	0.989278597	107	Cerebelum_8_L
136	0.980943046	0.980943046	138	Thal_MDl_R
156	0.77302915	-0.77302915	158	Vent_Str_R
103	0.745849039	0.745849039	105	Cerebelum_7b_L
146	0.741503865	0.741503865	148	Thal_PuA_R
152	0.728810515	-0.728810515	154	ACC_pre_R
137	0.715171135	-0.715171135	139	Thal_LGN_L
158	0.714094974	0.714094974	160	VTA_R
106	0.677014312	0.677014312	108	Cerebelum_8_R
82	0.672750898	0.672750898	84	Heschl_R
134	0.657076312	0.657076312	136	Thal_MDm_R
49	0.65007599	-0.65007599	49	Cuneus_L
122	0.645983908	0.645983908	124	Thal_LP_R
120	0.639241571	0.639241571	122	Thal_AV_R
101	0.638528185	-0.638528185	103	Cerebelum_6_L
99	0.606018167	-0.606018167	101	Cerebelum_4_5_L
140	0.592622827	-0.592622827	142	Thal_MGN_R
115	0.579659026	-0.579659026	117	Vermis_7
104	0.541010664	0.541010664	106	Cerebelum_7b_R
40	0.522908694	-0.522908694	40	Cingulate_Post_R
57	0.516213592	0.516213592	57	Occipital_Inf_L
155	0.511441289	-0.511441289	157	Vent_Str_L
97	0.476911931	-0.476911931	99	Cerebelum_3_L
93	0.475338477	-0.475338477	95	Cerebelum_Crus1_L
61	0.45980094	0.45980094	61	Postcentral_L

81	0.44504447	0.44504447	83	Heschl_L
74	0.443821248	0.443821248	74	Paracentral_Lobule_R
23	0.433320819	0.433320819	21	Frontal_Med_Orb_L
141	0.432911248	-0.432911248	143	Thal_PuI_L
114	0.432555937	-0.432555937	116	Vermis_6
17	0.429698572	0.429698572	15	Supp_Motor_Area_L
89	0.420304652	0.420304652	91	Temporal_Pole_Mid_L
10	0.415623821	0.415623821	8	Frontal_Inf_Oper_R
1	0.406186451	-0.406186451	sex	NA
70	0.399684895	-0.399684895	70	Angular_R
64	0.379159175	0.379159175	64	Parietal_Sup_R
142	0.3741403	-0.3741403	144	Thal_PuI_R
138	0.370069331	-0.370069331	140	Thal_LGN_R
113	0.349234647	-0.349234647	115	Vermis_4_5
16	0.343510549	0.343510549	14	Rolandic_Oper_R
5	0.342703652	0.342703652	3	Frontal_Sup_2_L
58	0.341113329	0.341113329	58	Occipital_Inf_R
3	0.338246799	0.338246799	1	Precentral_L
39	0.326976012	-0.326976012	39	Cingulate_Post_L
63	0.326737419	0.326737419	63	Parietal_Sup_L
62	0.325495054	0.325495054	62	Postcentral_R
133	0.319628639	0.319628639	135	Thal_MDm_L
46	0.318288358	0.318288358	46	Amygdala_R
38	0.30230918	-0.30230918	38	Cingulate_Mid_R
118	0.296549204	-0.296549204	120	Vermis_10
110	0.284322817	-0.284322817	112	Cerebellum_10_R
7	0.283846502	0.283846502	5	Frontal_Mid_2_L
69	0.279680041	-0.279680041	69	Angular_L
102	0.268425438	-0.268425438	104	Cerebelum_6_R
154	0.267118461	-0.267118461	156	ACC_sup_R
149	0.263748268	0.263748268	151	ACC_sub_L
80	0.261898753	-0.261898753	80	Pallidum_R
132	0.260789933	0.260789933	134	Thal_Re_R

144	0.2590412	0.2590412	146	Thal_PuM_R
54	0.249356217	0.249356217	54	Occipital_Sup_R
159	0.24890038	0.24890038	161	SN_pc_L
162	0.248677686	0.248677686	164	SN_pr_R
12	0.24501658	0.24501658	10	Frontal_Inf_Tri_R
52	0.244242719	0.244242719	52	Lingual_R
51	0.242522786	0.242522786	51	Lingual_L
72	0.242394767	-0.242394767	72	Precuneus_R
73	0.239594798	0.239594798	73	Paracentral_Lobule_L
90	0.238450169	0.238450169	92	Temporal_Pole_Mid_R
77	0.229008233	-0.229008233	77	Putamen_L
109	0.227864264	-0.227864264	111	Cerebelum_10_L
94	0.224638275	-0.224638275	96	Cerebelum_Crus1_R
92	0.222743513	0.222743513	94	Temporal_Inf_R
164	0.216013165	0.216013165	166	Red_N_R
22	0.214405484	0.214405484	20	Frontal_Sup_Medial_R
131	0.213677001	-0.213677001	133	Thal_Re_L
78	0.206578907	-0.206578907	78	Putamen_R
119	0.204659473	0.204659473	121	Thal_AV_L
139	0.203619503	-0.203619503	141	Thal_MGN_L
19	0.20124266	-0.20124266	17	Olfactory_L
36	0.195641414	0.195641414	34	Insula_R
71	0.191376651	-0.191376651	71	Precuneus_L
37	0.191253657	0.191253657	37	Cingulate_Mid_L
14	0.186814652	0.186814652	12	Frontal_Inf_Orb_2_R
34	0.183872114	-0.183872114	32	OFClat_R
48	0.174866544	0.174866544	48	Calcarine_R
91	0.17295969	0.17295969	93	Temporal_Inf_L
11	0.168960076	0.168960076	9	Frontal_Inf_Tri_L
168	0.163980288	0.163980288	170	Raphe_M
76	0.156249375	-0.156249375	76	Caudate_R
161	0.15564843	0.15564843	163	SN_pr_L
41	0.150009245	-0.150009245	41	Hippocampus_L

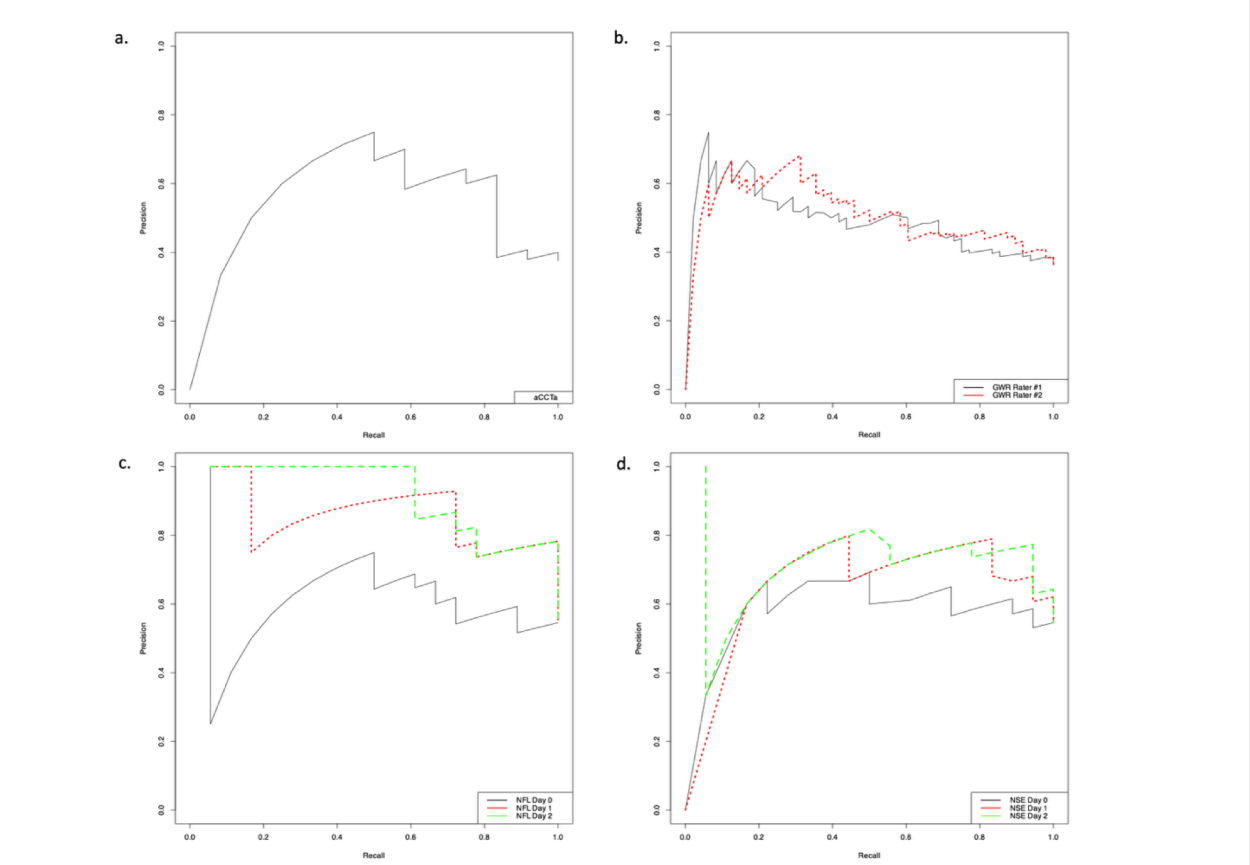
59	0.149159897	0.149159897	59	Fusiform_L
13	0.148315667	0.148315667	11	Frontal_Inf_Orb_2_L
167	0.146719084	0.146719084	169	Raphe_D
127	0.146160441	-0.146160441	129	Thal_VPL_L
98	0.14060953	-0.14060953	100	Cerebelum_3_R
150	0.14012786	0.14012786	152	ACC_sub_R
157	0.130878524	-0.130878524	159	VTA_L
135	0.130021053	-0.130021053	137	Thal_MDl_L
121	0.127299721	0.127299721	123	Thal_LP_L
83	0.122749979	0.122749979	85	Temporal_Sup_L
128	0.122561942	-0.122561942	130	Thal_VPL_R
123	0.119642013	0.119642013	125	Thal_VA_L
20	0.117596138	-0.117596138	18	Olfactory_R
18	0.117115709	0.117115709	16	Supp_Motor_Area_R
95	0.115497078	-0.115497078	97	Cerebelum_Crus2_L
153	0.114582421	0.114582421	155	ACC_sup_L
4	0.113580334	0.113580334	2	Precentral_R
130	0.103694993	0.103694993	132	Thal_IL_R
100	0.103604047	-0.103604047	102	Cerebelum_4_5_R
151	0.101943371	-0.101943371	153	ACC_pre_L
43	0.100963593	-0.100963593	43	ParaHippocampal_L
25	0.100415401	-0.100415401	23	Rectus_L
55	0.099948239	0.099948239	55	Occipital_Mid_L
15	0.099890679	-0.099890679	13	Rolandic_Oper_L
65	0.099882159	0.099882159	65	Parietal_Inf_L
33	0.096825869	-0.096825869	31	OFClat_L
45	0.095114759	0.095114759	45	Amygdala_L
27	0.093995844	-0.093995844	25	OFCmed_L
86	0.09327585	0.09327585	88	Temporal_Pole_Sup_R
66	0.089845462	-0.089845462	66	Parietal_Inf_R
32	0.086385686	-0.086385686	30	OFCpost_R
68	0.086375244	-0.086375244	68	SupraMarginal_R
28	0.083192844	-0.083192844	26	OFCmed_R

30	0.083135826	0.083135826	28	OFCant_R
148	0.076282263	0.076282263	150	Thal_PuL_R
124	0.076043961	0.076043961	126	Thal_VA_R
147	0.0712499	-0.0712499	149	Thal_PuL_L
126	0.06881239	-0.06881239	128	Thal_VL_R
31	0.067627038	0.067627038	29	OFCpost_L
84	0.066917727	0.066917727	86	Temporal_Sup_R
60	0.066328171	0.066328171	60	Fusiform_R
53	0.06617617	0.06617617	53	Occipital_Sup_L
21	0.065446558	-0.065446558	19	Frontal_Sup_Medial_L
163	0.06309725	0.06309725	165	Red_N_L
79	0.061442894	-0.061442894	79	Pallidum_L
125	0.059111931	-0.059111931	127	Thal_VL_L
143	0.057444346	0.057444346	145	Thal_PuM_L
112	0.050951869	-0.050951869	114	Vermis_3
145	0.049181981	-0.049181981	147	Thal_PuA_L
2	0.048820252	-0.048820252	delta_CT	NA
75	0.047891079	-0.047891079	75	Caudate_L
35	0.045882098	-0.045882098	33	Insula_L
26	0.040980683	-0.040980683	24	Rectus_R
47	0.040804525	0.040804525	47	Calcarine_L
0	0.039534497	-0.039534497	age	NA
85	0.038302989	-0.038302989	87	Temporal_Pole_Sup_L
88	0.037978614	-0.037978614	90	Temporal_Mid_R
44	0.029892604	-0.029892604	44	ParaHippocampal_R
56	0.029029374	0.029029374	56	Occipital_Mid_R
129	0.028521821	-0.028521821	131	Thal_IL_L
67	0.028093161	0.028093161	67	SupraMarginal_L
8	0.025314422	-0.025314422	6	Frontal_Mid_2_R
107	0.024907498	0.024907498	109	Cerebelum_9_L
24	0.01975113	0.01975113	22	Frontal_Med_Orb_R
108	0.017089162	-0.017089162	110	Cerebelum_9_R
111	0.015394774	-0.015394774	113	Vermis_1_2

87	0.015244465	0.015244465	89	Temporal_Mid_L
42	0.01402371	-0.01402371	42	Hippocampus_R
165	0.011634776	0.011634776	167	LC_L
29	0.01130333	-0.01130333	27	OFCant_L
96	0.009756427	0.009756427	98	Cerebelum_Crus2_R
50	0.003749027	-0.003749027	50	Cuneus_R
9	0.003729585	0.003729585	7	Frontal_Inf_Oper_L
166	0.002196229	0.002196229	168	LC_R
6	0.00081498	0.00081498	4	Frontal_Sup_2_R

eTable 1. Weighting θ of the corresponding features, consisting of the 166 regions of interest of the automated anatomical labeling atlas 3 (AAL3) as well as age, sex and delta cCT

eFigure 1



eFig 1. Precision-recall curves of the prognostic markers.
a. Automated cerebral computed tomography analysis (aCCTa)
b. Gray-white matter ratio of both raters (GWR Rater #1 and GWR Rater #2)
c. Neurofilament light chain in serum (NFL) for day 0 (within 3h), day 1 (>3h and <24h), and day 2 (>24h and <48h)
d. Neuron-specific enolase in serum (NSE) for day 0, day 1, and day 2

4. Diskussion

In der vorliegenden Studie wurde ein automatisiertes, CCT bildbasiertes Vorhersagemodell (aCCTa) entwickelt und hinsichtlich seiner Leistungsfähigkeit zur Vorhersage des neurologischen Outcomes nach OHCA untersucht. Die Ergebnisse zeigen, dass aCCTa sowohl im primären (ROC-AUC) als auch in den sekundären (F1-score, Balanced accuracy) Leistungsmaßen gut abschneidet und die etablierte, manuell erhobene radiologische Methode GWR sowie die laborchemischen Biomarker NfL und NSE zu einem frühen Zeitpunkt nach OHCA übertrifft, im weiteren Verlauf jedoch gegenüber den spät erhobenen laborchemischen Biomarkern unterlegen ist. Im Folgenden werden diese Befunde im Kontext der prognostischen Bedeutung von aCCTa sowie der durch das Modell identifizierten Hirnregionen diskutiert; abschließend wird ein Ausblick auf zukünftige Forschungsarbeiten gegeben.

4.1. Einordnung des prognostischen Wertes des bildbasierten Vorhersagemodells

Die Leistung des hier untersuchten, bildbasierten Vorhersagemodells ist als vielversprechend zu bewerten. Im primären Leistungsmaß ROC-AUC übertraf aCCTa die konventionelle GWR-Bestimmung, die von neuroradiologisch ausgebildeten Fachärzten anhand definierter ROIs manuell und verblindet gegenüber dem klinischen Outcome durchgeführt wurde. Darüber hinaus zeigte sich aCCTa den laborchemischen Biomarkern NSE und NfL zu einem frühen Zeitpunkt nach OHCA (Tag 0) überlegen. Dies unterstreicht den prognostischen Mehrwert eines datengetriebenen Ansatzes, der über einzelne manuell erhobene HU-Quotienten im Sinne von GWR hinaus den gesamten zerebralen Kontext einbezieht. Zu späteren Zeitpunkten (Tag 1, Tag 2) waren die etablierten laborchemischen Marker NSE und NfL dem bildbasierten Modell aCCTa hingegen in allen betrachteten Leistungsmaßen überlegen. Tabelle 3 der Publikation fasst die Ergebnisse für das primäre Leistungsmaß ROC AUC sowie die sekundären Leistungsmaße F1-Score und Balanced Accuracy für sämtliche untersuchten Modalitäten zusammen.

Insgesamt zeigt der Vergleich der verschiedenen Vorhersagemethoden, dass aCCTa, trotz der ab Tag 1 zunehmenden Überlegenheit der laborchemischen Biomarker NfL und NSE, eine durchweg robuste Leistung erbringt, die insbesondere gegenüber der klinisch etablierten, bildbasierten GWR überzeugt. Dies zeigt, dass maschinelle Lernverfahren, die CCT-basierte Features der grauen Substanz spezifischer Hirnregionen berücksichtigen, geeignet sind, relevante Informationen über den späteren neurologischen Zustand nach einem Herz-Kreislauf-Stillstand zu extrahieren. Im Vergleich zur herkömmlichen, manuell ermittelten GWR bietet das Modell einen objektiveren und automatisierten Ansatz, der auch subtilere Veränderungen des gesamten Gehirns für die Vorhersage nutzbar macht. Im Vergleich zu alternativen automatisierten Vorhersagemodellen, die auf weniger umfangreich vorverarbeiteter CCT-Bildgebung basieren, zeigte sich weder eine deutliche Über- noch Unterlegenheit.[27]

Gleichwohl ist festzuhalten, dass die bildbasierte Vorhersageleistung in ihrer gegenwärtigen Form für eine routinemäßige klinische Entscheidungsfindung nur mit Vorbehalten zu betrachten ist, insbesondere angesichts der heterogenen Datenlage und der Überlegenheit der laborchemischen Biomarker im späteren Verlauf. Im Einklang mit der einschlägigen Literatur bestätigte sich auch in dieser Studie die deutliche Überlegenheit von NfL für die Vorhersage des neurologischen Outcomes.[28] Perspektivisch besteht dennoch das Potenzial, dass bildgestützte, automatisierte Verfahren wie aCCTa einen wertvollen Baustein der frühen Vorhersage des neurologischen Outcomes darstellen, indem sie klinisch ohnehin erhobene, hochauflösende Bildgebung objektiver, reproduzierbarer und schneller auswerten als manuelle Verfahren.

4.2. Featurewichtung und Rückschlüsse auf die Vulnerabilität des Gehirns

Die räumliche Verteilung und Ausprägung hypoxisch-ischämischer Veränderungen im Gehirn wird maßgeblich von der individuellen Vulnerabilität einzelner Hirnregionen gegenüber Sauerstoffmangel bestimmt, die wiederum von Blutversorgung, Stoffwechselaktivität und Gewebeeigenschaften abhängt. Angesichts des hohen Energiebedarfs des Gehirns reagieren insbesondere metabolisch besonders aktive Strukturen wie der Thalamus[29], die Basalganglien[30], die Hippocampi[30] sowie Teile des Kleinhirns[31,32] empfindlich auf eine reduzierte Sauerstoffzufuhr. Auch Mittelhirnareale, darunter die Substantia nigra[30], sowie vaskuläre Grenzzonen gelten als besonders vulnerabel gegenüber Hypoxie.

Die vorliegenden Ergebnisse untermauern diese anatomisch-physiologische Beobachtung. Das Vorhersagemodell aCCTa identifizierte über die automatisierte Analyse der vorverarbeiteten CCT-Datensätze insbesondere Regionen des Mittelhirns, des Kleinhirns und des Thalamus als relevant für die Vorhersage des neurologischen Outcomes nach OHCA. Diese Befunde spiegeln bekannte Muster der Hypoxievulnerabilität wider: Strukturen mit hoher Vulnerabilität reagieren bereits auf kurzzeitige hypoxische Phasen mit Zellveränderungen sowie Wasser- und Ionenverschiebung, die sich in CT-morphologisch nachweisbaren Dichteveränderungen niederschlagen. Während etablierte, vergleichsweise grobe Kennwerte wie die GWR vor allem globale Veränderungen der Mark-Rinden-Differenzierung erfassen, ermöglicht die atlasbasierte Analyse der GMP die Detektion subtiler, regionenspezifischer Gewebsveränderungen im gesamten Gehirn und macht damit jene Areale sichtbar, in denen sich hypoxische Schädigungen typischerweise am frühesten und deutlichsten manifestieren. Die Gewichtungen der einzelnen Features, einschließlich der drei klinischen Variablen (Alter, Geschlecht, ΔCCT), sind im Supplementary Content (eTable 1) der Publikation in absteigender Reihenfolge dargestellt.

Die Gewichtung der Hirnregionen und klinischen Features (Alter, Geschlecht, ΔCCT) in der automatisierten Analyse liefert wertvolle Einblicke in die regionale Vulnerabilität des Gehirns

gegenüber Hypoxie. Den Feature-Koeffizienten zufolge sind die hypoxiebedingten Veränderungen nicht gleichmäßig über das Gehirn verteilt, sondern konzentrieren sich auf spezifische, besonders empfindliche Regionen. Ein positiver Koeffizient ist dabei mit der Vorhersage eines günstigen, ein negativer Koeffizient mit der Vorhersage eines ungünstigen neurologischen Outcomes assoziiert. Koeffizienten nahe null tragen kaum zur Vorhersage bei. Besonders ausgeprägt trugen Kleinhirnregionen (nach AAL3-Hirnatlas: Cerebellum 8 und 7b) sowie Thalamus und Substantia nigra zur Vorhersage eines günstigen Outcomes bei, während weitere Kleinhirnregionen (Cerebellum 6, Vermis 8 und 9) sowie der mediolaterale Geniculatkern des Thalamus und das rechte ventrale Striatum mit der Vorhersage eines ungünstigen Outcomes assoziiert waren. Die Regionen des Kleinhirns[31,32] , Thalamus[29] und Substantia nigra[30] gelten aufgrund ihres hohen Energiebedarfs und ihrer metabolischen Aktivität als besonders anfällig für hypoxische Schädigung, sodass die Ergebnisse hinsichtlich der Feature-Gewichtung mit der bestehenden Literatur übereinstimmen. Demgegenüber übten die klinischen Variablen Alter, Geschlecht und ΔCCT, erkennbar an ihren deutlich geringeren Koeffizienten, einen vergleichsweise geringen Einfluss auf die Vorhersage aus. Dies legt nahe, dass für die Prognose die regionale Verteilung und das Ausmaß der zerebralen Schädigung von größerer Bedeutung sind als die drei untersuchten klinischen Variablen.
Zusammenfassend erlernte das Vorhersagemodell aCCTa datengetrieben biologisch plausible Feature-Gewichtungen und ermöglicht damit eine differenzierte, automatisierte Analyse sämtlicher Hirnregionen im Rahmen der Outcome-Prädiktion nach OHCA.

4.3. Limitationen und Ausblick

Eine wesentliche Limitation der Studie ergibt sich aus der durch die klinische Versorgungsrealität bedingten Heterogenität der Datenlage. Da OHCA eine lebenskritische Situation darstellt, die eine unmittelbare klinische Versorgung erfordert, ist davon auszugehen, dass die erhobenen CCT-Bildgebungen im klinischen Alltag zur GWR-basierten Prognoseabschätzung herangezogen wurden und in einzelnen Fällen in Therapieentscheidungen, etwa eine Anpassung des Therapieziels oder eine Begrenzung intensivmedizinischer Maßnahmen, eingeflossen sein könnten. Da die GWR mit den Ergebnissen des aCCTa-Modells assoziiert ist, könnte ein solches Vorgehen im Sinne einer selbsterfüllenden Prophezeiung das beobachtete Outcome beeinflusst und damit die Aussagekraft des Modells sowie den Modellvergleich verzerrt haben. Ein solcher Effekt lässt sich bei der Untersuchung lebenskritischer klinischer Phänomene kaum sicher ausschließen und stellt eine Form der Data Leakage dar. Da die Daten zwar prospektiv erhoben, jedoch teilweise auch in der klinischen Versorgung genutzt wurden, bleibt unklar, inwieweit die Ergebnisse durch diesen strukturellen Bias beeinflusst sind. Dem steht die Real-World-Charakteristik der Daten gegenüber, die mit unregelmäßigen Erhebungszeitpunkten und einem gegenüber kontrollierten Studienbedingungen höheren Hintergrundrauschen

einhergeht. Dies ist ein Umstand, der zugleich ein Vorhersagemodell begünstigt, das näher an der klinischen Realität operiert.

Eine weitere Limitation stellt die begrenzte Fallzahl dar, die in Kombination mit der Heterogenität der Bilddaten das Risiko von Verzerrungen und Überanpassungen (Overfitting) durch Zufallseffekte oder Ausreißer erhöht. Zusätzlich variiert die technische Bildqualität aufgrund unterschiedlicher CT-Scannermodelle und Aufnahmeparameter, was die Segmentierung und damit die Genauigkeit der extrahierten Features beeinträchtigen kann. Die Variabilität der Erhebungszeitpunkte nach OHCA erschwert zudem die Vergleichbarkeit der CT-Aufnahmen; ob einheitlichere Untersuchungsfenster zu zuverlässigeren Vorhersagen führen würden, bleibt offen.

Einen weiteren limitierenden Faktor stellt das Fehlen einer externen Validierung dar. Ohne Überprüfung an unabhängigen Kohorten bleibt unklar, ob das Modell auch unter realen Bedingungen und in unterschiedlichen klinischen Kontexten seine Leistungsfähigkeit beibehält. Um dies zu ermöglichen, wurde Kontakt zu weiteren international tätigen, universitären Maximalversorgern aufgenommen; eine gemeinsame Metaanalyse mit einer Forschungsgruppe aus Taiwan befindet sich derzeit in Vorbereitung.

Darüber hinaus fokussiert sich die vorliegende Studie ausschließlich auf CCT-basierte Informationen, während weitere diagnostisch relevante Modalitäten wie EEG, MRT-Sequenzen, klinische Befunde oder zusätzliche laborchemische Biomarker nicht berücksichtigt wurden. Eine multimodale Datenintegration könnte künftig zu präziseren und robusteren Vorhersagen beitragen.

Für zukünftige Arbeiten sind daher strukturierte, prospektive, multizentrische Studien zu empfehlen, um Generalisierbarkeit und Verlässlichkeit automatisierter Vorhersageverfahren zu erhöhen. Die Integration weiterer Modalitäten sowie der Einsatz fortschrittlicher maschineller Lernmethoden, insbesondere Transformer-Architekturen, tiefer neuronaler Netze und Autoencoder, bieten Potenzial für eine präzisere Vorhersage. Spieltheoretisch fundierte Erklärungsverfahren wie Shapley-Werte[33] könnten dabei helfen, die Entscheidungsprozesse auch komplexer Modelle transparent und nachvollziehbar zu gestalten.

Ein weiterer vielversprechender Ansatzpunkt liegt in der systematischen Untersuchung spezifischer Zeitfenster nach OHCA, innerhalb derer bildgebende Informationen besonders aussagekräftig sind. Dynamische, generative Modelle, die zeitliche Veränderungen im Gehirn abbilden, könnten die Vorhersage weiter verfeinern und zugleich das pathophysiologische Verständnis hypoxischer Prozesse vertiefen. Nicht zuletzt sind auch klinische, ökonomische und ethische Fragestellungen zu berücksichtigen, um eine verantwortungsvolle Implementierung solcher Verfahren in die klinische Praxis zu gewährleisten.

Zusammenfassend bieten bildbasierte Verfahren zur Vorhersage des neurologischen Outcomes nach OHCA ein vielversprechendes Fundament, erfordern jedoch weitere

Validierungen und Optimierungen, bevor sie Eingang in die klinische Routine finden können. Ein multimodaler Ansatz, der verschiedene diagnostische Feature Sets kombiniert, könnte künftig zu einer individuelleren und klinisch relevanteren Outcome-Vorhersage beitragen.

4.4. Schlussfolgerung

Die Ergebnisse dieser Studie zeigen, dass das bildbasierte Vorhersagemodell aCCTa eine vielversprechende Leistungsfähigkeit zur Vorhersage des neurologischen Outcomes nach OHCA aufweist. Es übertrifft die manuell erhobene GWR und bietet damit eine objektivere, automatisierte Methode zur Analyse von CCT-Bildern, die subtile zerebrale Veränderungen erfasst. Während die Leistung von aCCTa zu frühen Zeitpunkten nach OHCA über derjenigen etablierter Biomarker wie NfL und NSE liegt, zeigt sich zu späteren Zeitpunkten eine zunehmende Überlegenheit dieser Laborparameter. Langfristig könnte aCCTa als wertvolle Ergänzung der Outcome-Prädiktion nach OHCA dienen; weitere prospektive Studien sind jedoch erforderlich, um seine Generalisierbarkeit und klinische Anwendung zu bestätigen.

5. Literaturverzeichnis

1 Wigner EP. The unreasonable effectiveness of mathematics in the natural sciences. Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Commun Pure Appl Math* 1960; **13**: 1–14.

2 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer New York, 2009 DOI:10.1007/978-0-387-84858-7.

3 Verhulst PF. Notice sur la loi que la population poursuit dans son accroissement. *Corresp Math Phys* 1838; : 113–21.

4 Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc: Ser B (Methodol)* 1996; **58**: 267–88.

5 Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 1970; **12**: 55–67.

6 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc: Ser B (Stat Methodol)* 2005; **67**: 301–20.

7 Legendre A-M. Nouvelles méthodes pour la ́determination des orbites des comètes. Paris, France: F. Didot, 1805.

8 Gauß CF. Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium. Hamburg, Germany: Friedrich Perthes und I.H. Besser, 1809.

9 Laplace P-S. Mémoire sur la probabiliٰte des causes par les \'ev\'enements. *Mémoires de l'Acad́emie Royale des Sciences de Paris* 1774; **6**: 621–56.

10 Wright S. Correlation and Causation. *Journal of Agricultural Research* 1921; **20**: 557–85.

11 Pearson K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London Series A* 1896; **187**: 253–318.

12 Yerushalmy J. Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques. *Public Heal Rep (1896-1970)* 1947; **62**: 1432.

13 Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958; **65**: 386–408.

14 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**: 32–5.

15 Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. *2010 20th Int Conf Pattern Recognit* 2010; : 3121–4.

16 Rijsbergen CV. Information retrieval: theory and practice. In Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems, 1979: 1–14.

17 Breiman L, Spector P. Submodel Selection and Evaluation in Regression. The X-Random Case. *Int Stat Rev Rev Int Stat* 1992; **60**: 291.

18 Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: International Joint Conference on Artificial Intelligence. 1992. https://api.semanticscholar.org/CorpusID:2702042.

19 Rolls ET, Huang C-C, Lin C-P, Feng J, Joliot M. Automated anatomical labelling atlas 3. *NeuroImage* 2020; **206**: 116189.

20 Adduru V, Baum SA, Zhang C, *et al.* A Method to Estimate Brain Volume from Head CT Images and Application to Detect Brain Atrophy in Alzheimer Disease. *Am J Neuroradiol* 2020; **41**: 224–30.

21 Brudfors M, Balbastre Y, Flandin G, Nachev P, Ashburner J. Flexible Bayesian Modelling for Nonlinear Image Registration. *Arxiv* 2020. DOI:10.48550/arxiv.2006.02338.

22 Jacobs I, Nadkarni V, Bahr J, *et al.* Cardiac Arrest and Cardiopulmonary Resuscitation Outcome Reports. *Circulation* 2004; **110**: 3385–97.

23 Langhelle A, Nolan J, Herlitz J, *et al.* Recommended guidelines for reviewing, reporting, and conducting research on post-resuscitation care: The Utstein style. *Resuscitation* 2005; **66**: 271–83.

24 Rittenberger JC, Raina K, Holm MB, Kim YJ, Callaway CW. Association between Cerebral Performance Category, Modified Rankin Scale, and discharge disposition after cardiac arrest. *Resuscitation* 2011; **82**: 1036–40.

25 Bernard SA, Gray TW, Buist MD, *et al.* Treatment of Comatose Survivors of Out-of-Hospital Cardiac Arrest with Induced Hypothermia. *New Engl J Medicine* 2002; **346**: 557–63.

26 Grossestreuer AV, Abella BS, Sheak KR, *et al.* Inter-rater reliability of post-arrest cerebral performance category (CPC) scores. *Resuscitation* 2016; **109**: 21–4.

27 Nakamoto T, Nawa K, Nishiyama K, *et al.* Neurological prognosis prediction for cardiac arrest patients using quantitative imaging biomarkers from brain computed tomography. *Phys Med* 2024; **125**: 103425.

28 Hoiland RL, Rikhraj KJK, Thiara S, *et al.* Neurologic Prognostication After Cardiac Arrest Using Brain Biomarkers. *JAMA Neurol* 2022; **79**: 390–8.

29 Sohn G, Kim SE. Measurement of thalamus and cortical damages in hypoxic ischemic encephalopathy. *IBRO Neurosci Rep* 2023; **15**: 179–85.

30 Fujioka M, Okuchi K, Sakaki T, Hiramatsu K, Miyamoto S, Iwasaki S. Specific changes in human brain following reperfusion after cardiac arrest. *Stroke* 2018; **25**: 2091–5.

31 Horn M, Schlote W. Delayed neuronal death and delayed neuronal recovery in the human brain following global ischemia. *Acta Neuropathol* 1992; **85**: 79–87.

32 Ng T, Graham DI, Adams JH, Ford I. Changes in the hippocampus and the cerebellum resulting from hypoxic insults: frequency and distribution. *Acta Neuropathol* 1989; **78**: 438–43.

33 Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. *arXiv* 2017.

6. Anhang

6.1. Tabellenverzeichnis

Tabelle 1: Kreuztabelle (Confusion Matrix)