TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Medicine and Health

Einfluss der Covid-19-Pandemie auf die Vorhersagegüte eines maschinellen Lernalgorithmus zur Prädiktion perioperativer Mortalität

Dimislav Ivanov Andonov

Vollständiger Abdruck der von der TUM School of Medicine and Health der Technischen Universität München zur Erlangung eines Doktors der Medizin genehmigten Dissertation.

Vorsitz: Prof. Dr. Gabriele Multhoff

Prüfer*innen der Dissertation:

1. Priv.-Doz. Dr. Simone Kagerbauer

2. apl. Prof. Dr. Gerhard Rammes

Die Dissertation wurde am 05.06.2023 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 30.09.2023 angenommen.

Inhaltsverzeichnis

Inhaltsverzeichnis 3
Abkürzungsverzeichnis 5
Abbildungsverzeichnis 6
Tabellenverzeichnis 7
1 Einleitung 8
1.1 Künstliche Intelligenz (KI) und maschinelles Lernen (ML) 8
1.2 Modellentwicklung und Anpassung 14
1.3 Ethische und gesetzliche Aspekte von ML in der Medizin 14
1.4 ML in der perioperativen Medizin 16
1.5 eXtreme Gradient Boosting (XGBoost) 17
2 Fragestellung und Zielsetzung 19
3 Patienten und Methoden 20
3.1 Studiendesign, Patienten und Ethik 20
3.2 Datenquellen 22
3.3 Endpunkte 22
3.4 Merkmale, Datenvorverarbeitung und Umgang mit fehlenden Daten 22
3.5 Stichprobenumfang 25
3.6 Entwicklung 25
3.7 Statistische Auswertung 27
4 Ergebnisse 28
4.1 Patientencharakteristika 28
4.2 Vergleich des Anteils fehlender Werte zu den Untersuchungszeitpunkten. 32
4.3 Vorhersagemodelle im Rahmen der Covid-19-Pandemie 32
4.4 Information Gain 38
5 Diskussion 43
5.1 Zusammenfassung und Interpretation der Ergebnisse 43

5.2 Beurteilung der Modelle und Einordnung in die Literatur 45
5.3 Einfluss der Covid-19 Pandemie auf die Variablen 48
5.4 Limitationen 51
6 Zusammenfassung 53
7 Literaturverzeichnis 54
8 Danksagung 59
9 Veröffentlichungen 60

Abkürzungsverzeichnis

ASA	American Society of Anesthesiologists Physical Score
AUROC	Area under the Receiver Operating Characteristic
ATC	Anatomical Therapeutic Chemical Classification System
CARES	Combined Assessment of Risk Encountered in Surgery
CART	Classification and Regression Tree
Covid-19	Coronavirus Disease 2019
GBT	Gradient-Boosted Trees
FPR	Falsch –Positive-Rate/Falsch-Positive-Rate
FN	Falsch –Negative/ False negative
KI	Künstliche Intelligenz
MEDS	Mortality in Emergency Department Sepsis
MEWS	Modified Early Warning Score
ML	Machine Learning
NPV	Negativer Vorhersagewert/ Negative predictive value
OPS	Operationen- und Prozedurenschlüssels
POSSUM	Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity
PRC	Precision-recall curve
PPV	Positiver Vorhersagewert/ Positive predictive value
ROC	Receiver Operating Characteristic
SIRS	Systemic Inflammatory Response Syndrome
SOFA	Sepsis-Related Organ Failure Assessment
TNR	Korrekt -Negative-Rate/True negative rate
TPR	Korrekt -Positive-Rate/True positive rate
XGBoost	Extreme Gradient Boosting

Abbildungsverzeichnis

Abbildung 1: Publikationen zur Anwendung von KI aus dem Jahr 2020 nach medizinischer Disziplin. 8
Abbildung 2: Das Kontinuum der Komplexität und Interpretierbarkeit von ML-Algorithmen. 10
Abbildung 3: Kategorisierung von Maschinellem Lernen. 11
Abbildung 4: Receiver-Operating-Characteristic-(ROC-) Kurve. 12
Abbildung 5: Konfusionsmatrix. 13
Abbildung 6: Perioperative Anwendungsmöglichkeiten von ML. 17
Abbildung 7: Vereinfachter Algorithmus der XGBoost-Technik. 18
Abbildung 8: Pandemieverlauf. 20
Abbildung 9: STROBE-Diagramm der Datenselektion. 21
Abbildung 10: Screenshot der Prämedikationsmaske. 23
Abbildung 11: AUROC-Kurven für das erste Modell. 34
Abbildung 12: PR-Kurven für das erste Modell. 34
Abbildung 13: AUROC-Kurven für das zweite Modell. 35
Abbildung 14: PR-Kurven für das zweite Modell. 35
Abbildung 15: AUROC-Kurve für das dritte Modell. 36
Abbildung 16: PR-Kurve für das dritte Modell. 36
Abbildung 17: Information Gain vor der Pandemie. 40
Abbildung 18: Information Gain einschließlich der ersten Covid-19-Welle. 41
Abbildung 19: Information Gain gesamter Zeitraum. 42

Tabellenverzeichnis

Tabelle 1: Codierung der Einweisungsarten als Beispiel für Bewegungsdaten aus dem Klinik-Informationssystem, die ebenfalls als Variablen in das Modell aufgenommen wurden. 25
Tabelle 2: Patientencharakteristika entsprechend der Pandemiewellen. 30
Tabelle 3: Validierungsdaten (AUROC, AUPR) der drei Modelle [52]. 37
Tabelle 4: Statistische Benchmarks zu den jeweiligen Cut-off-Werten [52]. 38

1 Einleitung

1.1 Künstliche Intelligenz (KI) und maschinelles Lernen (ML)

Künstliche Intelligenz (KI) beschreibt einen Wissenschaftsbereich, der sich mit der Entwicklung von Systemen oder Maschinen zur Reproduktion der menschlichen Intelligenz befasst [1, 2]. Für den medizinischen Bereich werden hierbei Technologien entwickelt, die der Risikoprädiktion und Entscheidungsfindung dienen. Mathur et al. evaluierten die Anzahl der Publikationen im Jahr 2020 zu KI in der Medizin, aufgeteilt nach medizinischer Disziplin [3] (Abbildung 1). Die Kategorie „Allgemein“ umfasst Übersichten über ethische und gesellschaftliche Herausforderungen der KI-Forschung im Gesundheitswesen, die Entwicklung spezifischer maschineller Lerntechniken und Studien zur Arzneimittelentwicklung. Insbesondere in der Radiologie und der Versorgung von Covid -19-Patienten, aber auch in der Anästhesiologie, der Chirurgie und vielen weiteren Disziplinen finden KI-basierte Technologien Anwendung.

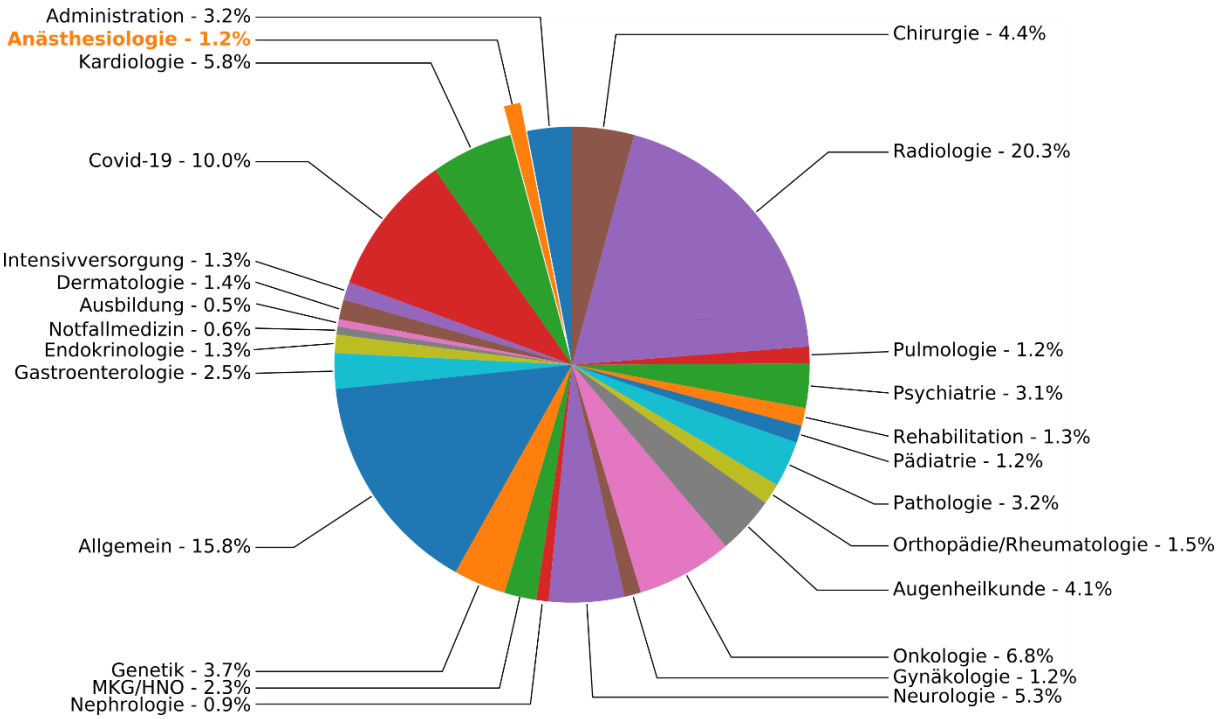

Abbildung 1: Publikationen zur Anwendung von KI aus dem Jahr 2020 nach medizinischer Disziplin. Eigene Darstellung, angelehnt an [3].

Unter Machine Learning (ML) versteht man Techniken, die es einem Computer ermöglichen, komplexe Aufgaben zu bewältigen und Big Data, also große Datenvolumina, schnell und variabel zu integrieren [4]. Im medizinischen Sinne setzt sich Big Data unter anderem aus den Daten der elektronischen Patientenakten, Daten der physiologischen Überwachung und Register- oder Studiendaten aus nationalen und internationalen Datenrepositorien zusammen. ML stellt einen Teilbereich der künstlichen Intelligenz dar, der Elemente der Informatik, Statistik und mathematische Algorithmen umfasst. ML-Methoden haben somit das Potenzial, sowohl in der medizinischen Forschung als auch der klinischen Versorgung komplexe Datenverarbeitungsprozesse durchzuführen und die Therapieentscheidung und die Vorhersage des Outcomes basierend auf großen Datenmengen zu erleichtern [5-7]. Im Gegensatz zur konventionellen Programmierung „lernt" die Maschine anhand der analysierten Daten selbständig und integriert diese neuen Erkenntnisse in ein Modell, das sich somit kontinuierlich verbessert. ML-Algorithmen müssen trainiert werden. Dies erfolgt durch Aufteilung eines Datensatzes in ein Trainings- und ein Testsegment. Anhand des Trainingsdatensatzes „lernt" der Algorithmus die Muster und Zusammenhänge in den Daten. Die Testdaten dienen der Kontrolle des Lernprozesses. Validierungsdaten sind von den Test- und Trainingsdaten unabhängig, sie dienen dazu, die Vorhersagequalität des erstellten Modells zu überprüfen. In manchen Arbeiten werden die Bezeichnungen „Testdatensatz" und „Validierungsdatensatz" auch vertauscht verwendet.

Die Komplexität von ML-Algorithmen hängt von der Fragestellung und der Art der zu bewertenden Beziehung zwischen Variablen und Ergebnis ab. Ein ML-Algorithmus kann interpretierbar sein oder eine sogenannte „black box" bilden (Abbildung 2). Leicht interpretierbare Algorithmen sind beispielsweise statistische Regressionsmodelle, die lineare Beziehungen zwischen dem Vorhersagefaktor und dem Ergebnis nutzen[8]. Aus Informatik-Sicht kann so ein simples Modell auf einem herkömmlichen Computer erstellt werden. Nicht interpretierbare ML-Algorithmen erstellen basierend auf Prädiktoren mit einer nichtlinearen Beziehung zum Ergebnis ein komplexes Modell, dessen Berechnung je nach Menge der Trainingsdaten mehrere Tage in Anspruch nehmen kann und somit Computersysteme mit hoher Rechenleistung erfordert. Derartige komplexe Algorithmen sind zum Beispiel Gradient Boosting, Support Vector Machines oder Neuronale Netze.

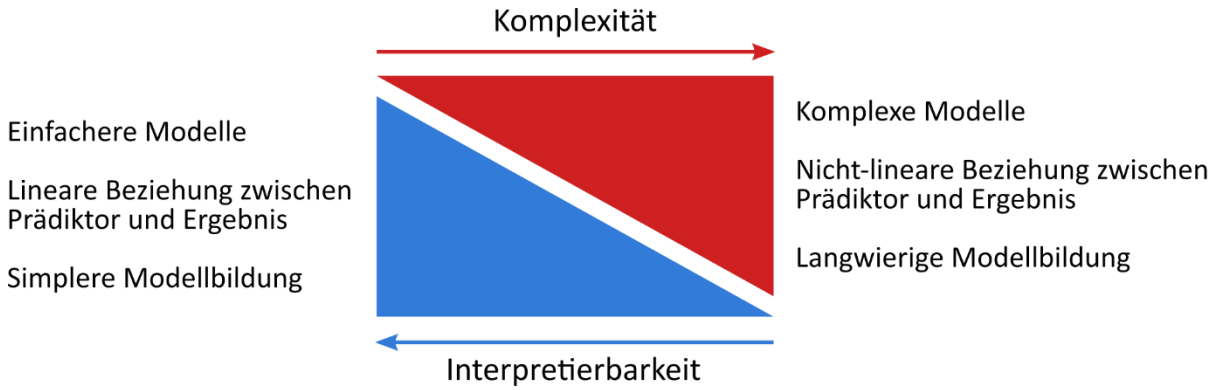

Abbildung 2: Das Kontinuum der Komplexität und Interpretierbarkeit von ML-Algorithmen. Eigene Darstellung, angelehnt an [6].

Eine weitere Kategorisierung von ML-Methoden sieht eine Unterteilung in supervised (überwachtes) ML und unsupervised (unüberwachtes) ML vor [9, 10]. Überwachtes ML beinhaltet das Training eines Modells basierend auf eingegebenen Daten und einem bekannten Ergebnis. In der Klinik kann ein solches Modell beispielsweise in der Prädiktion von Adipositas oder Diabetes mellitus basierend auf den Ernährungsgewohnheiten, dem Gewicht und der Körpergröße Anwendung finden [11, 12]. Sobald ein Algorithmus erfolgreich trainiert wurde, ist er in der Lage, Vorhersagen über das Ergebnis zu treffen, wenn er auf neue Daten angewendet wird. ML-Algorithmen können zur Vorhersage von Ergebnissen in diskreten Kategorien benutzt werden, beispielweise Tumorgrade oder Schmerzkategorien. Auch kontinuierliche Daten, wie beispielsweise die Lebenserwartung oder die optimale Dosis eines Medikamentes, können anhand eines überwachten Regressionsalgorithmus vorhergesagt werden[13, 14]. Die Vorhersage eines kontinuierlichen Wertes wird basierend auf einem Daten- bzw. Variablensatz ähnlich einer statistischen Regression ermöglicht.

Beim unüberwachten ML gibt es kein vordefiniertes Ergebnis, sondern der Algorithmus sucht im Datensatz explorativ nach Mustern oder Ähnlichkeiten zwischen Gruppen (Clustern) ohne Eingaben des Nutzers. Sie werden beispielweise eingesetzt, um Datensätze zu reduzieren und Einzeldaten zusammenzufassen, um sie zu gruppieren und so die Informationen zu komprimieren[15].

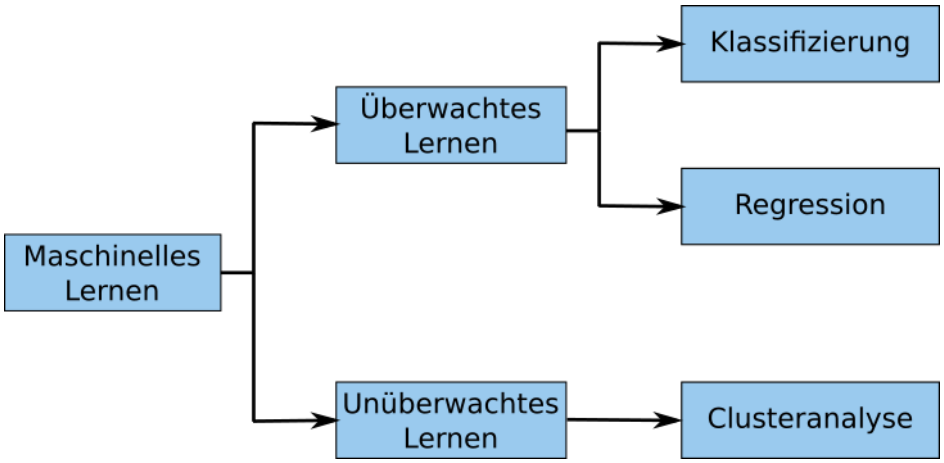

Abbildung 3: Kategorisierung von Maschinellem Lernen. Eigene Darstellung, angelehnt an [10].

Das eine Extrem eines ML-Modells wäre somit die Integration weniger Variablen für ein bedingt aussagekräftiges und wahrscheinlich ungenaues, aber leicht generierbares und auf andere Situationen übertragbares Modell. Das andere Extrem ist ein flexibles und hochkomplexes Modell, das auf belastbaren Testdaten mit hoher Genauigkeit beruht, jedoch große Datenmengen erfordert und eine langwierige Auswertung beinhaltet. Ein ML-Modell ist also im Idealfall ein Kompromiss zwischen der Komplexität des Modells und der Verallgemeinerbarkeit auf neue Datensätze [9].

Klassifizierungsmodelle sagen nicht das Ergebnis, sondern nur die Wahrscheinlichkeit des Auftretens voraus. Im Falle eines binären Klassifizierungsmodells, wie es in dieser Arbeit verwendet wird, sagt das Modell eine Wahrscheinlichkeit für das Eintreten oder Nicht-Eintreten des Ereignisses voraus. In diesem Fall kann die Wahrscheinlichkeit von 0 bis 1 variieren, so dass die Summe der positiven und negativen Vorhersagen gleich 1 ist. Wenn man die vorhergesagten Wahrscheinlichkeiten analysiert, kann man entscheiden, welches Ergebnis wahrscheinlicher ist[16].

Allerdings sollte man sich nicht völlig auf die vorhergesagten Wahrscheinlichkeiten verlassen, denn auch wenn das Modell eine hohe Wahrscheinlichkeit für ein bestimmtes Ergebnis angibt, kann das Ergebnis, das in der Realität eintritt, dennoch das Gegenteil sein. Um die Effizienz der Modelle abzuschätzen und sie miteinander zu vergleichen, müssen neu trainierte Modelle mit dem Testdatensatz überprüft werden.

Zur Charakterisierung der Effizienz von binären Klassifizierungsmodellen können die Receiver-Operating-Characteristic-(ROC-) Kurve und die Area under the ROC curve (AUROC) verwendet werden (Abbildung 4).

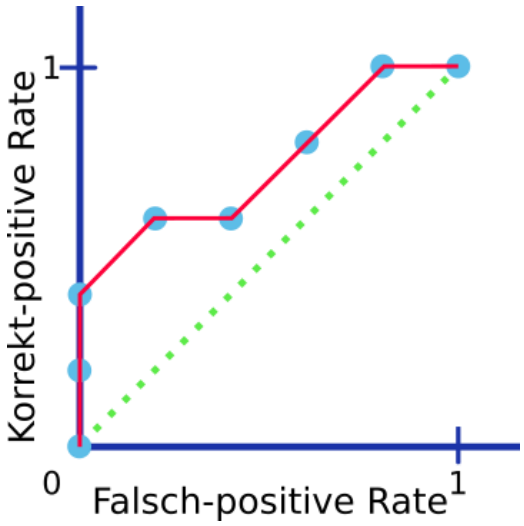

Abbildung 4: Receiver-Operating-Characteristic-(ROC-) Kurve. Eigene Darstellung, angelehnt an [17].

Bei der ROC-Kurve handelt es sich um ein Diagramm, bei dem die korrekt-positive Rate auf der y-Achse und die falsch-positive Rate auf der x-Achse liegt. Die Korrekt-Positiv-Rate (true positive rate, TPR) kann als „Entdeckungswahrscheinlichkeit“ (oder als Sensitivität) betrachtet werden und wird wie folgt berechnet:

$$TPR = \frac{TP}{TP + FN}$$

wobei *TP* der Anzahl der korrekten positiven Vorhersagen und *FN* der Anzahl der falsch negativen Vorhersagen entspricht.

Die Falsch-Positive-Rate kann als die Wahrscheinlichkeit eines falschen Alarms betrachtet werden und ergibt sich wie folgt:

$$FPR = \frac{FP}{FP + TN}$$

wobei *FP* der Anzahl der falsch-positiven Vorhersagen entspricht und *TN* der Anzahl der korrekten negativen Vorhersagen.

Die Korrekt-Negative-Rate (true negative rate, TNR) kann als die Spezifität betrachtet werden und ergibt sich wie folgt:

$$TNR = \frac{TN}{FP + TN}$$

wobei *FP* der Anzahl der falsch-positiven Vorhersagen entspricht und *TN* der Anzahl der korrekten negativen Vorhersagen.

Die ROC-Kurve wird durch Variation des Schwellenwerts für die Wahrscheinlichkeit erreicht, ab der das Ergebnis des Ereignisses als positiv zu betrachten ist. Bei jedem festgelegten Schwellenwert wird die Anzahl der richtigen und falschen Vorhersagen gezählt, in einer Konfusionsmatrix zusammengefasst und zur Berechnung eines Punktes auf der ROC-Kurve verwendet (Abbildung 5). So erhält man für jedes Modell eine ROC-Kurve. Anschließend kann die Effizienz der verschiedenen Modelle anhand der Fläche unter der entsprechenden ROC-Kurve (AUROC) verglichen werden. Im Idealfall muss der AUROC-Wert gleich 1 sein, was in den meisten Fällen unmöglich zu erreichen ist, da die Modelle mit einer begrenzten Datenmenge trainiert werden und nicht alle möglichen Fälle und Faktoren berücksichtigen können. Daher wählt man beim Vergleich der Modelle das Modell mit dem höchsten AUROC-Wert für die weitere Anwendung aus[18].

In der Precision-Recall-Kurve werden Präzisionswerte (y-Achse) und Recall-Werte (x-Achse) aufgetragen [19]. Die Präzision wird in diesem Zusammenhang auch als positiver Vorhersagewert (positive predictive value, PPV) bezeichnet und der Recall entspricht der True-Positive-Rate (TPR).

$$PPV = \frac{TP}{TP + FP}$$

Im Idealfall weisen sowohl Präzision als auch Recall hohe Werte auf. Oft geht man im Zusammenhang mit ML-Algorithmen einen Kompromiss zwischen beiden Faktoren ein [20].

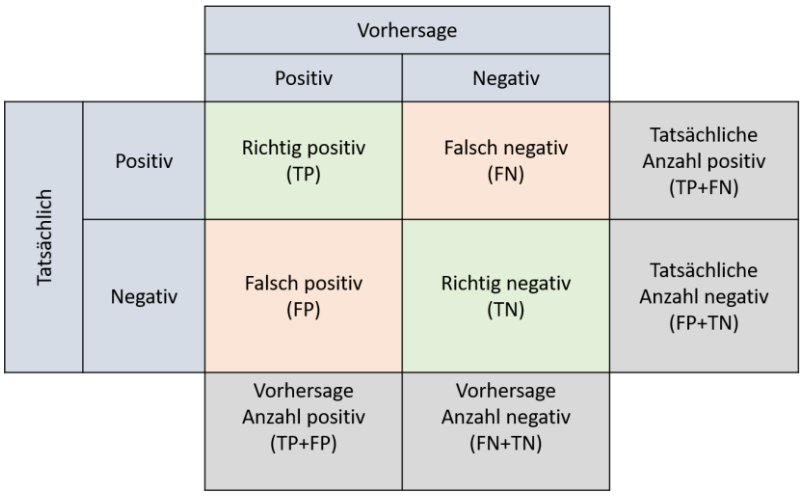

Abbildung 5: Konfusionsmatrix. Eigene Darstellung, angelehnt an [18].

1.2 Modellentwicklung und Anpassung

Wichtig ist vor allem, dass ein Modell nicht nur die für dessen Entwicklung verwendete Stichprobe beschreibt, sondern verallgemeinerbar ist. Das aus dem Trainingsdatensatz entwickelte Modell wird daher in einem Testdatensatz erprobt und weiter angepasst und optimiert, bis keine Verbesserung mehr zu sehen ist, weil z.B. die maximale AUROC erreicht ist. Damit soll zum einen „underfitting“ verhindert werden, das heißt, ein zu ungenaues „unterangepasstes“ Modell, das die Zielvariablen nicht ausreichend genau vorhersagt. Das Gegenteil davon ist „overfitting“: ein zu gut an die Trainingsdaten angepasstes Modell liefert im Testdatensatz eine wesentlich ungenauere Prädiktion. Das Modell wird so lange trainiert, wie die Genauigkeit der Prädiktion im Testdatensatz noch zu steigern ist. Fällt die Vorhersagegenauigkeit im Testdatensatz wieder, so ist das ein Zeichen für overfitting, und jede weitere Anpassung sollte unterbleiben [21].

Wurde nun ein optimales Modell entwickelt, so wird es in der Praxis auf bisher unbekannte Daten angewendet. Nun geht man bei der Vorhersage davon aus, dass die Art und Struktur der Daten sowie beispielsweise deren Wahrscheinlichkeitsverteilung gleichbleibt. Dies entspricht natürlich nicht der Wahrheit, denn Trends oder Begleitumstände ändern sich mit der Zeit, bestimmte Einflussfaktoren werden wichtiger, während andere an Bedeutung verlieren. Daher stellt sich in der Praxis stets die Frage, ob und wenn ja, wie oft, ein Modell neu angepasst werden muss [21]. Viele in der Wirtschaft angewendete Modelle erfahren automatische Updates in bestimmten Zeitabständen [22]. Gerade im Bereich der Medizin sind es demographische Veränderungen, neue Untersuchungs- und Behandlungsmethoden oder die Zu- und Abnahme von Inzidenzen, die die Vorhersagegüte eines Modells beeinflussen können. In jüngster Zeit war es die Covid-Pandemie, die zu einem Engpass in der Intensivversorgung und zu einer Verschiebung elektiver Operationen führte. Ziel der vorliegenden Arbeit war es daher, den Einfluss der durch die Covid-Pandemie bedingten Änderungen auf ein mit Daten vor der Pandemie erstellten Prädiktionsmodells der perioperativen Mortalität zu eruieren.

1.3 Ethische und gesetzliche Aspekte von ML in der Medizin

Die Anwendung von ML in der Medizin bedarf einer Abwägung der ethischen Verwendung persönlicher Daten und des klinischen Mehrwertes [23, 24]. Es stellen sich also sowohl ethische als auch gesetzliche Herausforderungen für die medizinische Anwendung von ML-Algorithmen, die Aspekte wie den Datenschutz, die Transparenz, die Haftung und die

Rechenschaftspflicht betreffen [23]. Hierbei lassen sich zwei Bereiche unterscheiden, zum einen die Quelle der zu verwendenden Daten, zum anderen die Entwicklung der Algorithmen und ihre Anwendung in der klinischen Praxis [25]. Medizinische Daten unterliegen grundsätzlich dem Datenschutz und dem Schutz der Privatsphäre, so dass der Verwendung der Daten für KI- Zwecke zugestimmt werden muss und ihr Schutz durch die geltenden Verwendungsbedingungen und Rechtsvorschriften (beispielsweise die Datenschutzgrundverordnung) gewährleistet sein muss [26].

Bei der Verwendung von ML-Algorithmen zur klinischen Entscheidungsunterstützung wird Wert auf die Interpretierbarkeit der Modelle gelegt. Komplexe, intransparente „black box"-Modelle sind dabei zu vermeiden[27].

Daher sollten für die medizinische Entscheidungsunterstützung Algorithmen verwendet werden, die für den Mediziner interpretierbar sind. Damit soll das Vertrauen des Patienten sowohl in die ärztliche Kompetenz als auch in die Zuverlässigkeit der ML-Technologien gewährleistet werden. Um ML-Algorithmen interpretierbar zu machen gibt es verschiedene mathematische und statistische Verfahren. In dieser Arbeit beispielsweise wird der Algorithmus durch die Auflistung der wichtigsten Faktoren und die Visualisierung ihres Anteils an der Prädiktion des Modells für den Arzt nachvollziehbar.

Die Verantwortlichkeit der KI setzt sich demzufolge aus der Rechenschaftspflicht (Erklärung der Handlungen), der Haftung (Maßnahmen des Rechtssystems), und der Schuldfrage (Bestrafung) zusammen [28]. Die Rechenschaftspflicht kann in interpretierbaren Systemen durch die Nachvollziehbarkeit der Ergebnisse erfüllt werden, aber die Fragen der Haftung und der Schuld erfordern einen bisher noch nicht klar definierten Konsens.

Bei Anwendung von ML-Algorithmen in der Klinik muss klar sein, wer im Schadensfalle haftet, wenn es zu einer Fehlentscheidung oder einem unerwarteten Ergebnis kommt. Dies kann durch verschiedene Fehlerquellen, wie beispielsweise fehlerhafte Algorithmen, mangelhafte Optimierung oder fehlerhaftem Erlernen neuer Ergebnisse, Viren oder Systemfehler eintreffen. Eine äquivalente rechtliche Würdigung der Konsequenzen von ML steht bisher noch aus [29]. Aus diesem Grund kommen allgemeine Haftungsregeln zum Einsatz, die sich in Abhängigkeit von der Situation aus einer Haftung des Herstellers nach dem Produkthaftungsgesetz und einer Haftung des Anwenders bzw. des behandelnden Arztes ergeben können [30].

1.4 ML in der perioperativen Medizin

Das Scheitern von lebenserhaltenden Maßnahmen nach chirurgischen Komplikationen („failure to rescue“) ist eine der Haupttodesursachen nach einem operativen Eingriff [31, 32]. Mehr als zwei Drittel der postoperativen Todesfälle geschehen auf einer normalen Station und nicht auf einer Intensivstation [33, 34]. Es ist somit anzunehmen, dass eine gezielte Identifizierung von Risikopatienten und eine engmaschige perioperative Überwachung der Patienten das Ergebnis verbessern kann. Eine solche Intensivüberwachung ist jedoch mit einem hohen Kostenaufwand verbunden, und die Zahl verfügbarer Intensivbetten ist begrenzt, sodass Methoden zur Selektion von Patienten, für die eine solche Überwachung sinnvoll und nötig ist, unabdinglich sind [35]. Einer effektive Triage von Patienten war insbesondere in der Covid-19-Pandemie, in der die Ressourcen aufgrund der hohen Patientenzahlen knapp und die Risikofaktoren noch nicht genau bekannt waren, von großer Relevanz [36]. ML-Systeme können sich hierbei gegenüber herkömmlichen Risikoscores wie ASA, POSSUM oder surgical APGAR überlegen zeigen, da sie genaue Vorhersagen auf individueller Ebene leisten können. [37, 38].

Der am häufigsten verwendete klassische Risikoscore ist der ASA-Score der American Society of Anesthesiologists, anhand dessen das perioperative Risiko in eine von sechs Risikokategorien eingeordnet wird [39]. Als problematisch wird beim ASA die subjektive Einschätzung des Untersuchenden diskutiert, die mit einer moderaten Interrater-Reliabilität einhergeht [40].

Die Mehrheit der ML-basierten Methoden in der Medizin konzentriert sich daher auf die Identifizierung von Risikopatienten. Die Anwendung eines solchen ML-Systems kann prä-, intra- oder postoperativ erfolgen (Abbildung 6) [41]. Die Einbeziehung von ML-Algorithmen in die präoperative Beurteilung umfasst die Auswertung der Ergebnisse bildgebender Verfahren, die Durchführung und Beurteilung von Endoskopien, sowie die Diagnosestellung basierend auf histologischen und pathologischen Daten und die daraus resultierende chirurgische Entscheidung. Das ML-gestützte intraoperative Management umfasst die Überwachung der Narkosetiefe, Infusionsgeräte, Intubationen, Operationsroboter, und die Früherkennung von narkose- und operationsbedingten Komplikationen.

Postoperativ dienen ML-Systeme unter anderem der Risikoprädiktion, der Identifikation von notwendigen Maßnahmen zur Minimierung von Komplikationen, der frühen Erkennung von Komplikationen und der Einschätzung und Behandlung von postoperativen Schmerzen [41].

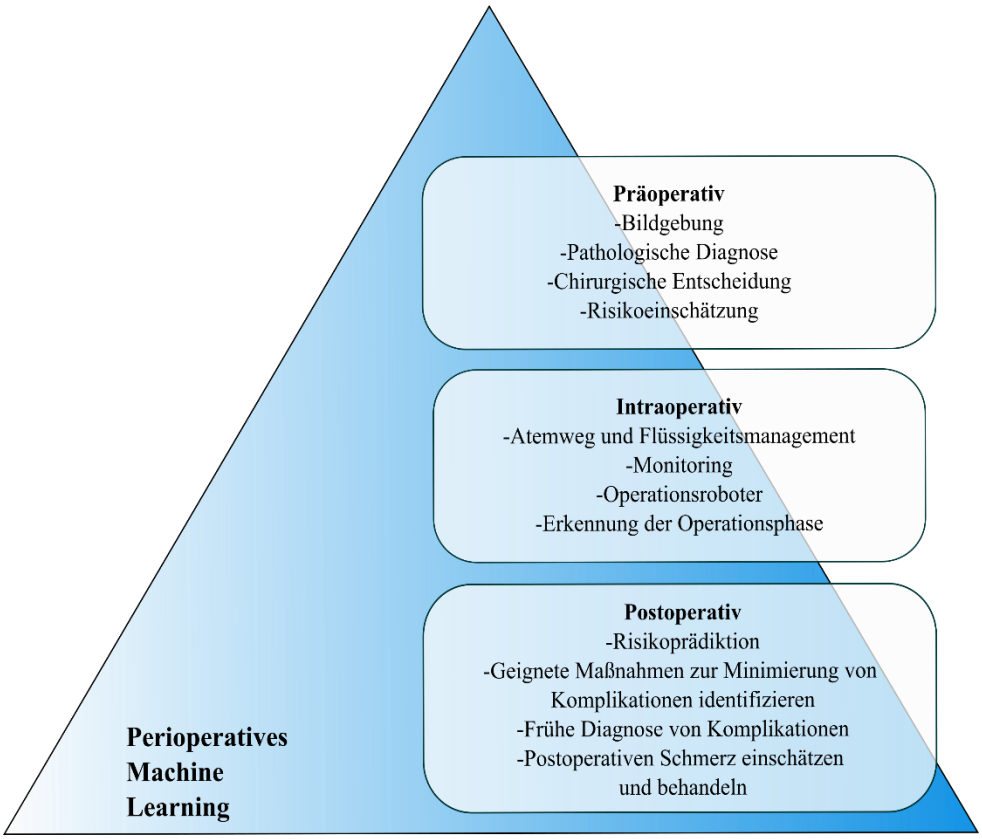

Abbildung 6: Perioperative Anwendungsmöglichkeiten von ML. Eigene Darstellung, angelehnt an [41].

1.5 eXtreme Gradient Boosting (XGBoost)

Das eXtreme Gradient Boosting (XGBoost) gehört zu den Techniken des maschinellen Lernens wie die Regressionsmodelle [42], Random Forests [43, 44] und Gradient Tree Boosting [45]. Bei XGBoost handelt es sich um eine neue Gradient Boosting Ensemble-Lernmethode mit C++-Implementierung des Gradient-Boosting-Baum-Algorithmus von Tianqi [46]. Dank seiner hohen Effizienz und Geschwindigkeit wird XGBoost häufig von Datenwissenschaftlern eingesetzt, um technische und wissenschaftliche Aufgaben zu lösen [47] sowie Wettbewerbe für maschinelles Lernen zu gewinnen [48]. Der Ansatz von XGBoost wurde bereits in mehreren Veröffentlichungen ausführlich beschrieben [46, 49]. XGBoost ist eine sogenannte Ensemble-Methode, das heißt, es werden mehrere Modelle miteinander kombiniert, um ein endgültiges Modell zu erhalten. Ein solches Einzelmodell in XGBoost ist in der Regel ein Entscheidungsbaum (decision tree), der für sich allein keine große Vorhersagekraft besitzt, weshalb er auch als „weak learner“ bezeichnet wird. Diese Entscheidungsbäume werden nun sequenziell miteinander kombiniert (Abbildung 7). Das Ergebnis wird jeweils mit einem Faktor gewichtet, so dass die Vorhersage sich mit jedem Durchlauf verbessert. Der Unterschied

zwischen Vorhersage und tatsächlichem Wert wird benutzt, um den sogenannten Gradienten zu berechnen, der hilft, den Fehler in der nächsten „Runde“ zu verkleinern. Sogenannte Hyperparameter, die vor Beginn des Trainings festgelegt werden und die Güte des Algorithmus bestimmen, sind bei XGBoost, ähnlich wie bei neuronalen Netzen die sogenannte Lernrate (die Schrittgröße, mit der der Gradient vermindert wird) und die Verlustfunktion, mit der der Unterschied zwischen den tatsächlichen Beobachtungen und der Vorhersage charakterisiert wird. Außerdem spielen weitere Hyperparameter eine Rolle, wie zum Beispiel die insgesamt verwendete Anzahl der Bäume, die Anzahl der Beobachtungen in jedem „Blatt“, also Endpunkt und die Komplexität, also die „Tiefe“ der Bäume.

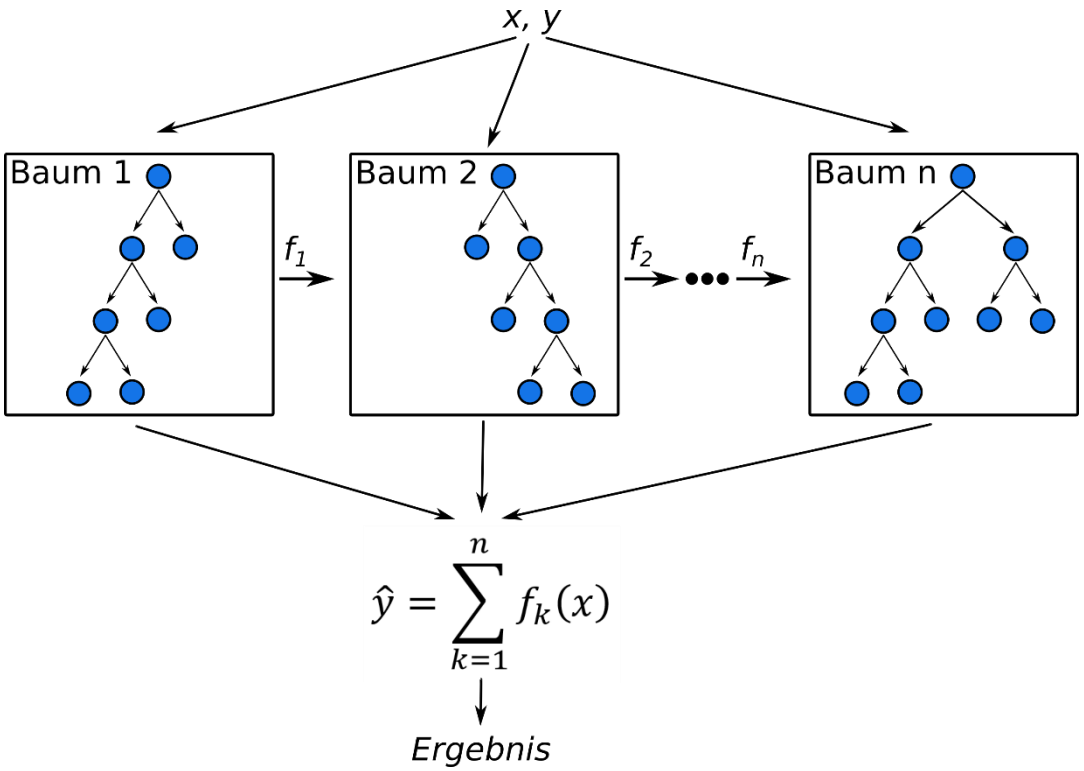

Abbildung 7: Vereinfachter Algorithmus der XGBoost-Technik. Eigene Darstellung, angelehnt an [50].

2 Fragestellung und Zielsetzung

Das Ziel der vorliegenden Studie war es, XGBoost-Modelle zur Vorhersage der perioperativen Mortalität zu erstellen und den Einfluss von Veränderungen im Rahmen der Covid-Pandemie zu analysieren. Verwendet wurden Daten von Patienten, die am Klinikum rechts der Isar im Zeitraum Juni 2014 bis Oktober 2021 operativ versorgt wurden. Hierbei wurden die Daten so aufgeteilt, dass sie Patienten vor der Covid -19-Pandemie, während der ersten Welle der Covid-19-Pandemie, zwischen der ersten und zweiten Welle der Covid -19-Pandemie, während der zweiten Welle der Covid -19-Pandemie, und nach der zweiten Covid-19-Pandemiewelle umfassten. Die Modellerstellung erfolgte für ein erstes Modell mit den präpandemischen Daten, für ein zweites Modell mit Daten vor der Pandemie und der ersten Welle zusammengenommen, und in einem dritten Modell für den gesamten Zeitraum von Juni 2014 bis Oktober 2021. Die ersten beiden Modelle wurden anschließend auf die Datensätze der unterschiedlichen Pandemiephasen angewendet, so dass der Einfluss der Covid-19-Pandemie auf die Vorhersagekraft des jeweiligen Modells analysiert werden konnte. Das dritte Modell diente als Referenz.

Ziel unserer Untersuchung war, die folgenden Forschungsfragen zu beantworten:

- Gelingt es grundsätzlich, auf Basis präoperativ erhobener Routinedaten ein Modell zu entwickeln, das die perioperative Mortalität mit akzeptabler Genauigkeit vorhersagt?
- Wie gut sind die Vorhersagen des Modells, wenn sich die Begleitumstände ändern? In unserem Fall beinhaltet das zum Beispiel eine Verschiebung des OP-Spektrums hin zu Notfalleingriffen im Rahmen der Einschränkungen der Covid-Pandemie.
- Muss ein solches Vorhersagemodell auf jeden Fall an sich ändernde Begleitumstände angepasst werden? Wann ist der ideale Zeitpunkt für die Anpassung?

3 Patienten und Methoden

3.1 Studiendesign, Patienten und Ethik

Die Studie zur Erstellung des Prädiktionsmodells wurde durch die Ethikkommission der Medizinischen Fakultät der Technischen Universität München (TUM) genehmigt (253/19 S-SR vom 11. Juni 2019) sowie im Studienregister Clinical Trials registriert (NCT04092933) und am Universitätsklinikum rechts der Isar der TUM durchgeführt. Die definitive Analyse umfasste Daten aller Patienten, die sich zwischen Juni 2014 und Oktober 2021 einer nicht-kardiochirurgischen Operation unterzogen. Eingeschlossen wurde jeweils die erste OP eines Patienten innerhalb eines Klinikaufenthaltes, Folgeeingriffe wurden nicht weiter berücksichtigt. Patienten, die bereits vor der ersten Operation auf der Intensivstation aufgenommen wurden, wurden ausgeschlossen, ebenso Patienten mit einem nicht-chirurgischen Eingriff (z. B. Diagnostik) oder einer ambulanten Behandlung. Neben elektiven waren auch dringliche, Notfalleingriffe sowie Kindereingriffe in dem Datensatz enthalten.

Der Datensatz wurde in verschiedene Zeiträume unterteilt (Abbildung 8), und zwar nach Patienten, die vor der Covid -19-Pandemie behandelt wurden (06.2014 – 03.2020), solche, die während der ersten Pandemiewelle behandelt wurden (04.2020 – 05.2020), solche, die zwischen der ersten und der zweiten Pandemiewelle behandelt wurden (06.2020 – 09.2020), solche, die während der zweiten Pandemiewelle behandelt wurden (10.2020 – 05.2021) und solche, die nach der zweiten Pandemiewelle behandelt wurden (06.2021 – 10.2021).

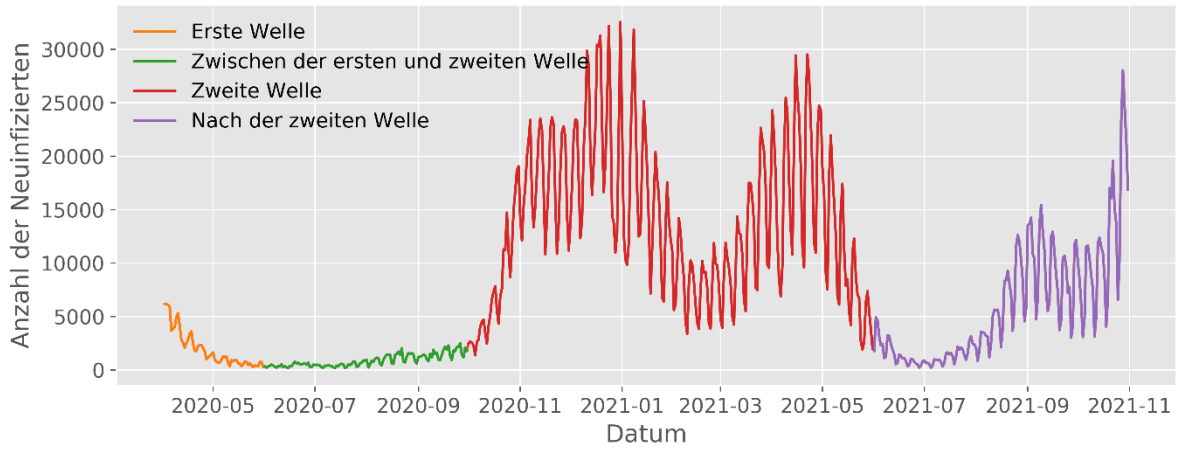

Abbildung 8: Pandemieverlauf. Eigene Darstellung, angelehnt an [51].

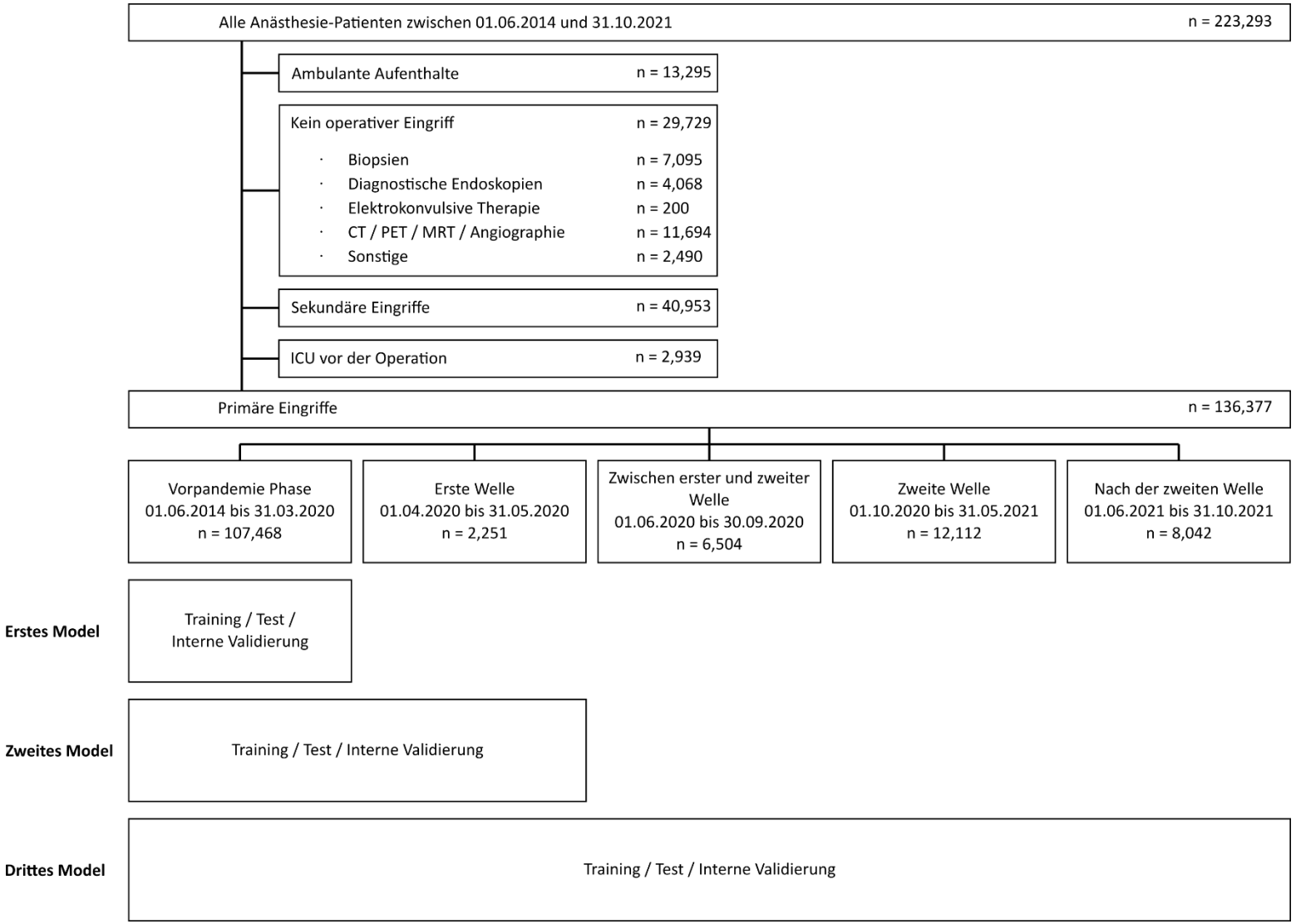

Abbildung 9: STROBE-Diagramm der Datenselektion. CT = Computer-Tomographie, ICU = Intensive Care Unit, MRT = Magnet-Resonanz-Tomographie, PET = Positronen-Emissions-Tomographie [52].

3.2 Datenquellen

Die für die vorliegende Arbeit verwendeten Daten stammten aus drei verschiedenen Quellen: dem Klinikinformationssystem (SAP i.s.h.med), dem Laborinformationssystem (Swisslab Lauris) und dem Patientendatenmanagementsystem der Anästhesie (QCare, HIM-Health Information Management GmbH, Bad Homburg, Deutschland)) des Universitätsklinikums rechts der Isar der TUM. Während sich die Datenpunkte des Klinikinformationssystems und des Patientendatenmanagementsystems kaum überschneiden, da beide Systeme weitestgehend unabhängig voneinander sind, werden die Datenpunkte des Laborinformationssystems durch eine technische Schnittstelle vollständig in das Klinikinformationssystem integriert. Die Daten wurden aus den Datenbanken der jeweiligen Informationssysteme im comma-separated values (csv)-Format extrahiert. Patientenidentifizierende Daten wie Name und Geburtsdatum waren zuvor entfernt worden. Die Verarbeitung erfolgte auf klinikinternen Rechnern ohne Zugriffsmöglichkeit für unbefugte Dritte.

3.3 Endpunkte

Der primäre Endpunkt des jeweiligen Modells war die Mortalität im Rahmen des Krankenhausaufenthaltes („in-hospital-mortality“) nach dem jeweils ersten stationär durchgeführten Eingriff eines Patienten. Auf Basis der aus den klinischen Informationssystemen extrahierten Daten wurde ein Algorithmus zur Vorhersage derselben entwickelt und validiert. Zielsetzung der Studie ist ein Vergleich der Performance der Modelle zur Mortalitätsprädiktion vor Beginn der Covid-19-Pandemie und zu verschiedenen Zeitpunkten während der Pandemie.

3.4 Merkmale, Datenvorverarbeitung und Umgang mit fehlenden Daten

Die für das Training des Modells verwendeten Merkmale („features“) lassen sich in sechs Kategorien einteilen: Patientenanamnese, chirurgisches Verfahren gemäß der Kodierung nach der jeweils gültigen Form des Operationen- und Prozedurenschlüssels (OPS-Codes), Bewegungsdaten (Details zur Aufnahme, Verlegung und Entlassung), Laborwerte, aktuelle Medikation und Blutanforderungen. Einige Daten wie Laborwerte, Anforderungen aus dem Blutdepot, OPS-Codes und Bewegungen innerhalb des Krankenhauses lagen bereits in tabellarischer und strukturierter Form vor. Sie mussten nicht oder nur geringfügig angepasst werden, um für die Erstellung des Modells verwendet werden zu können. Andere Daten wie

der präoperative Fragebogen und die aktuelle Medikation waren meist nur als Freitext verfügbar. Freitextdaten mussten weitgehend manuell vorverarbeitet werden.

Bei der Prämedikationsvisite erhobene Befunde und relevante Diagnosen werden im Prämedikations-PDMS (QCare, HIM-Health Information Management GmbH) dokumentiert. Hier stehen Drop-Down-Listen mit vorformulierten Antworten zur Verfügung, die um Freitexteingaben ergänzt werden können (Abbildung 10).

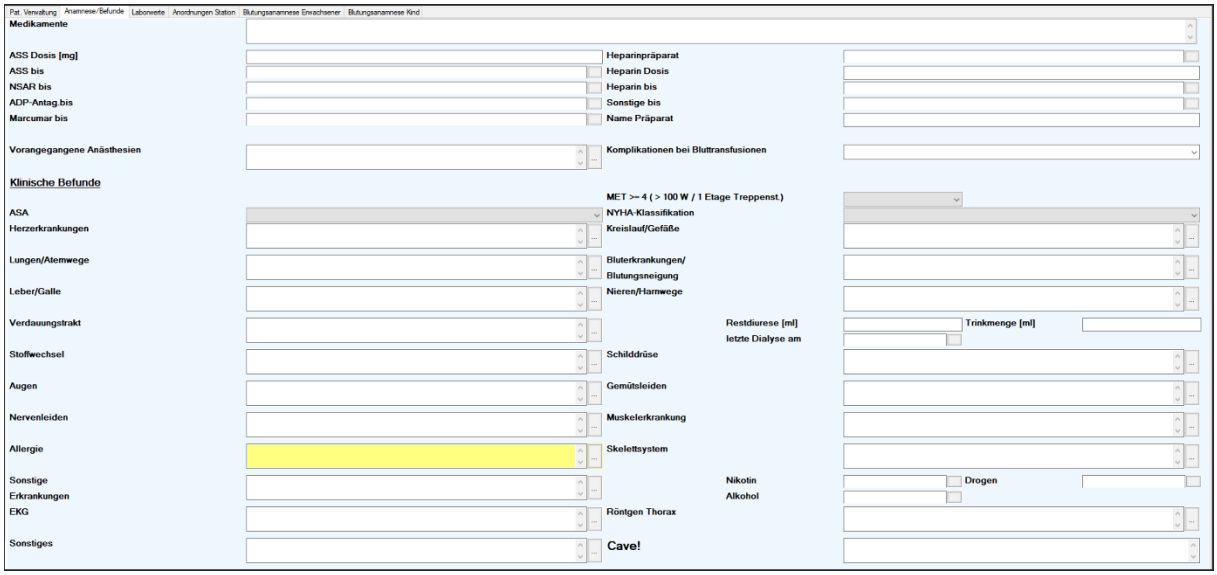

Abbildung 10: Screenshot der Prämedikationsmaske.

Zur Strukturierung der bei der Prämedikationsvisite eingegebenen Diagnosen und Befunde wurde zunächst eine Liste aus den in den Drop-Down-Listen der Prämedikationssoftware des PDMS verwendeten Begriffe erstellt. Die in den Freitexten vorkommenden Begriffe wurden außerdem nach ihrer Häufigkeit sortiert und Stoppwörter und Füllwörter entfernt. Die Liste wurde anschließend um diese Begriffe ergänzt. Mit dieser Liste wurde dann in den Freitexten nach diesen Begriffen und eventuellen Verneinungen in ihrem Zusammenhang gesucht. Für jeden gesuchten medizinischen Begriff wurde ein Merkmal „ja“, „nein“ und „nicht verfügbar“ erstellt.

Die präoperative Dauermedikation war als Freitextangabe in der Prämedikationsmaske verfügbar. Zur Strukturierung wurden zunächst alle Medikamentennamen aus dem Freitextfeld extrahiert und auf Rechtschreib- und Tippfehler analysiert. Es erfolgte eine Rechtschreibkorrektur, und die Medikamente wurden ihrem jeweiligen ATC-Code zugeordnet. Abschließend wurden Gruppen gebildet, die die Medikamente mit den jeweils gleichen ersten vier Stellen des ATC-Codes zusammenfassten [53].

Bei Patienten, die nicht in der Ambulanz, sondern auf Station prämediziert wurden, erfolgte die Dokumentation auf Papier. Die Eingabe der zur Qualitätssicherung relevanten Daten des Kerndatensatzes Anästhesie der Deutschen Gesellschaft für Anästhesie und Intensivmedizin (DGAI) und des Berufsverbandes Deutscher Anästhesisten (BDA) in das klinische Informationssystem erfolgt routinemäßig durch das Assistenzpersonal in der Prämedikationsambulanz. Der Kerndatensatz 3.0/2010 umfasst 66 Felder, die in 13 Gruppen aufgeteilt sind. Dieser standardisierte Datensatz liefert ein Dokumentationssystem zur einheitlichen Datenerfassung und Qualitätssicherung, das neben inhaltlichen auch rechtliche Aspekte des präoperativen Zustandes und des intraoperativen Anästhesieprotokolls beachtet. Die 13 großen Gruppen sind wie folgt: Header Informationen (Informationen zur Einrichtung), Allgemeine Daten, Risikobewertung, Zeiterfassung, Anästhesieverfahren, Luftweg, Atmung/Beatmung, Erweitertes Monitoring, Operationsart, Anästhesie-Verlaufs-Beobachtung (AVB), Entlassung, besondere Qualitätsmerkmale, Datensatztechnische Felder [54]. Somit war bei den auf Station prämedizierten Patienten eine wesentlich geringere Datenmenge vorhanden. Auch von diesen Patienten flossen aber alle verfügbaren Informationen in den Algorithmus ein.

Die Daten über die Bewegungen innerhalb des Krankenhauses und zur Beschreibung des Behandlungsprozesses (Art der Aufnahme, Verlegung, Entlassung, Intensivaufenthalt, Notwendigkeit einer postoperativen Nachbeatmung) lagen bereits in codierter Form vor (siehe Tabelle 1).

Zusätzlich wurden abgeleitete Variablen verwendet, z.B. konnte aus Datum und Uhrzeit ermittelt werden, ob der chirurgische Eingriff an einem Wochentag, am Wochenende oder an einem Feiertag, während der regulären Arbeitszeiten oder in der Nachtschicht durchgeführt wurde.

Jeder Patient und viele Variablen hatten fehlende Werte. Diese wurden nicht imputiert, sondern für jede Variable ein neues dichotomes Merkmal mit der Information über die Verfügbarkeit für jeden einzelnen Patienten (ja/nein/nicht bekannt) erstellt.

Zusammengenommen wurden über 12000 Parameter für die Modellerstellung verwendet, davon 9300 OPS-Codes und 780 Laborwerte. In den drei endgültigen Modellen wurden zusätzlich Parameter aus der Anamnese (241), Bewegungen innerhalb des Krankenhauses (24), Medikamente (199) und präoperative Bestellungen aus dem Blutdepot (13) verwendet. Die wichtigsten Parameter waren (in abnehmender Reihenfolge): Alter, bestellte Erythrozytenkonzentrate, Anzahl präoperativer Konsile, C-reaktives Protein (CRP),

Serumalbumin, Hämoglobinwerte, Aufnahmegrund, ASA-Kategorie, QUICK-Wert, Thrombozytenzahl, Fachabteilung, Body-Mass-Index (BMI), Hämatokrit, Erythrozytenzahl, präoperativ angefordertes gefrorenes Frischplasma (FFP) und Leukozytenzahl.

Code der Einweisung	Bezeichnung der Einweisung
A	Einweisung
AE	Aufnahme aus Ext.KH
AK	A. ext. KH < 24h
AR	Aufnahme aus Reha
B	Begleitperson
BM	Begleitperson med
E	Entbindung
EB	Einbestellt
EO	Organentnahme
HS	Hubschrauber
KT	Katastrophen-Fall
N	Neugeborenes
NA	Notarzt
NO	Notfall
NP	nicht verwenden
PO	Poliklinik
RW	Rettungswagen
S	Selbsteinweisung
ST	Studie
WD	Wiederkehrer

Tabelle 1: Codierung der Einweisungsarten als Beispiel für Bewegungsdaten aus dem Klinik-Informationssystem, die ebenfalls als Variablen in das Modell aufgenommen wurden.

3.5 Stichprobenumfang

Die Modelle wurden basierend auf Datensätzen von den ausgewählten Zeiträumen trainiert und getestet. Die jeweils dafür verwendeten Datensätze wurden als Trainings- und Test- und Validierungskohorte stratifiziert 3:1:1 aufgeteilt (60% der Daten wurde für das Training, 20% zum Testen und 20% für die Validierung verwendet).

3.6 Entwicklung

Die Vorhersagemodelle wurden mithilfe von Extreme Gradient Boosting (XGBoost) mit den Hyperparametern „Lernrate“, „minimale Verlustreduktion“, „maximale Tiefe jedes Baums“, „Anteil der Merkmale“, „Anteil der Trainingsstichproben“, „Skala der positiven Gewichte“ und „Minimum des Instanzgewichts“[46] entwickelt.

Lernrate: Die Lernrate bestimmt die Schrittgröße bei jeder Iteration, während das Modell auf sein Ziel hin optimiert wird. Eine niedrige Lernrate macht die Berechnungen langsamer und erfordert mehr Runden, um die gleiche Verringerung des Restfehlers zu erreichen wie ein

Modell mit einer hohen Lernrate, aber sie maximiert die Chancen, das Optimum zu erreichen. Eine zu kleine Lernrate kann allerdings dazu führen, das globale Maximum nie zu erreichen.

Minimale Verlustreduktion: Minimale Verlustreduzierung, die erforderlich ist, um eine weitere Partition an einem Blattknoten des Baums vorzunehmen. Je größer sie ist, desto konservativer wird der Algorithmus sein.

Maximale Tiefe jedes Baums: hierbei wird die maximale Anzahl an möglichen Knotenpunkten vorgegeben. Dadurch wird versucht, eine Überanpassung des Modells zu verhindern.

Anteil der Merkmale: Er stellt den Anteil der Spalten dar, die für jeden Baum nach dem Zufallsprinzip abgetastet werden, um eine Überanpassung zu vermeiden.

Anteil der Trainingsstichproben. Das Teilprobenverhältnis der Trainingsinstanzen ist für die Vermeidung einer Überanpassung relevant. Das Subsampling wird einmal in jeder Boosting-Iteration durchgeführt. Es stellt den Anteil der Beobachtungen dar, die für jeden Baum in die Stichprobe aufgenommen werden. Ein niedriger Wert verhindert eine Überanpassung, kann aber zu einer Unteranpassung führen.

Skala der positiven Gewichte: Sie gewährleistet das Gleichgewicht der positiven und negativen Gewichte und ist besonders nützlich für unausgewogene Werteklassen.

Minimum des Instanzgewichts: Wenn die Baumpartitionierung zu einem Blattknoten führt, dessen Summe der Instanzgewichte kleiner ist als das Minimum des Instanzgewichts, dann wird der Aufbauprozess die weitere Partitionierung abbrechen.

Die Ergebnisse der XGBoost-Modelle sind stark von diesen Parametern abhängig. Die Optimierung der Parameter ist sowohl zeit- als auch rechenintensiv. Die Bayes‘sche Hyperparametersuche erfolgte mit dem R-Package ‚mlrMBO‘ und wurde zur Abstimmung der Parameter für die Area-Under-the-Receiver-Operating-Characteristic-(AUROC-)Kurve verwendet. In unserem Fall bestimmen die Hyperparameter die AUROC und sollen so optimiert werden, dass diese maximal wird. Dies geschieht mit Hilfe eines probabilistischen Modells, das durch die Hinzunahme weiterer Datenpunkte iterativ verbessert wird [55].

Insgesamt wurden drei Modelle entwickelt:

1. Ein Modell unter Verwendung des vorpandemischen Datensatzes (06/2014 – 03/2020)
2. Ein Modell aus dem vorpandemischen Datensatz und dem der ersten Welle (06/2014 – 05/2020)
3. Ein Modell mit Daten aus dem gesamten Zeitraum (06/2014-10/2021)

Die Hyperparameteroptimierung erfolgte für jedes Modell separat.

Die Grenzen der Hyperparameter wurden für alle Modelle einheitlich wie folgt festgelegt: Lernrate (0,01 – 0,2), minimale Verlustreduzierung (0 – 6), maximale Tiefe jedes Baums (3 – 30 Ebenen), Anteil der Merkmale (0,5 – 1), Anteil der Trainingsstichproben (0,5 – 1), Skala der positiven Gewichte (0,01 – 10) und minimale Summe der Instanzgewichte (0 – 20). Die Konfidenzintervalle für die einzelnen Vorhersagen wurden anhand von 100 Bootstrap-Stichproben berechnet.

3.7 Statistische Auswertung

Die drei Modelle wurden auf der Grundlage ihrer AUROC [95-%-Konfidenzintervall] und Precision-Recall-Kurve verglichen. Die Erstellung der Modelle erfolgte mit Hilfe des folgenden R-Packages:

- dplyr
- xgboost
- tidyverse
- rBayesianOptimization
- mlrMBO
- skimr
- purrr
- DiceKriging
- Rgenoud

Für jedes Modell wurde mit Hilfe des maximalen Youden-Index (Sensitivität + Spezifität -1) ein Cut-off-Wahrscheinlichkeitswert bestimmt. Der Cut-off-Wert bezeichnet das optimale Verhältnis von richtig negativen zu richtig positiven Ergebnissen für eine spezielle Frage. Bei einer neuen Fragestellung muss der Cut-off-Wert gegebenenfalls anhand der ROC-Kurve angepasst werden[56]. Auf der Grundlage dieses Cut-off-Werts wurden die Sensitivität, Spezifität, sowie ein positiver und negativer prädiktiver Wert in jedem Zeitraum für jedes der drei Modelle bestimmt, um die Performance der Modelle zu vergleichen. Der Cut-off Wert wurde jeweils mit seinem Konfidenzintervall angegeben. Alle weiteren Analysen wurden ebenfalls mit der statistikorientierten Programmiersprache R (Version 4.2.1, R Foundation for Statistical Computing; Wien) durchgeführt. Kontinuierliche Variablen werden mit Median und Interquartilsbereich dargestellt, diskrete Variablen mit Absolut- und Relativwerten.

4 Ergebnisse

4.1 Patientencharakteristika

Die Charakteristika der Patienten, deren Daten in die Modellentwicklung inkludiert wurden, sind in Tabelle 2 aufgeführt. Hierbei wurden die Patienten entsprechend den untersuchten Zeitpunkten vor und nach der Covid-19-Pandemie kategorisiert.

Die meisten Patienten fielen in die Gruppe der vor der Covid-19-Pandemie behandelten Patienten (n = 107468), gefolgt von Patienten, die in der zweiten Covid-19-Welle behandelt wurden (n = 12112). Die wenigsten Patienten wurden während der ersten Covid-19-Welle operiert (n = 2251).

Der prozentuale Anteil der verstorbenen Patienten reichte von 0,8 % vor und nach der Covid -19-Pandemie bis 1,0 % während der ersten und zweiten Covid-19-Pandemiewelle.

Es zeigten sich hinsichtlich der Patientencharakteristika einige Auffälligkeiten im Vergleich der Zeitspannen vor, während und nach den Covid -19-Pandemiewellen. Zum einen waren die Patienten während und nach der ersten Covid -19-Welle und während und nach der zweiten Covid -19-Welle im Median älter als die Patienten vor der Pandemie (57 oder 58 Jahre im Vergleich zu 56 Jahren).

Hinsichtlich der ASA-Klassifikation zeigten sich Unterschiede in den Häufigkeiten der ASA-Kategorien zwischen den Untersuchungszeitpunkten. Vor der Covid -19-Pandemie wurden mehr Patienten einer niedrigen ASA-Kategorie von 1 oder 2 zugeordnet als während der Pandemiewellen oder nach der zweiten Pandemiewelle. Umgekehrt fielen während beiden Pandemiewellen und nach der Pandemie mehr Patienten in die ASA-Kategorien 3 oder 4, wurden also als schwerer erkrankt eingestuft.

Aus der Tabelle ist ersichtlich, dass elektive Eingriffe während der Pandemiewellen im Vergleich zu vor der Pandemie deutlich zurückgingen, während gleichzeitig eine Zunahme von Notfällen zu beobachten war. Zudem wurde eine veränderte Häufigkeit von Patienten in bestimmten Fachabteilungen während der Covid -19-Pandemie beobachtet. So nahm während der ersten Covid-19-Pandemiewelle die relative Häufigkeit der Patienten in der Orthopädie und Unfallchirurgie ab (von 17,8 % vor der Pandemie auf 14,8 % während der ersten Pandemiewelle). Nach der ersten Covid-19-Pandemiewelle nahm die Anzahl dieser Patienten jedoch wieder zu, so dass sie etwa der Anzahl vor der ersten Welle entsprach (17,6 % zwischen

beiden Pandemiewellen, 16,8 % während der zweiten Pandemiewelle, und 17,8 % nach der zweiten Pandemiewelle). Eine ebensolche Abnahme der Patienten während der ersten Covid -19-Pandemiewelle wurde in den Abteilungen HNO, MKG und Augenheilkunde (von 23,4 % auf 17,9 %) beobachtet. Hier nahm die Patientenzahl nach der ersten Covid -19 Pandemiewelle wieder auf 21,2% zu, fiel jedoch in der zweiten Covid -19-Pandemiewelle wieder auf 19,9 % ab und blieb auch nach der zweiten Covid -19-Pandemiewelle im Vergleich zu der Situation vor der Covid -19-Pandemie niedriger (20,5 %).

Auffällig war eine Zunahme des Anteils der Patienten in der Gynäkologie und Geburtshilfe während der ersten Covid -19-Pandemiewelle im Vergleich zu der Situation vor der Covid-19-Pandemie (von 10,0 % auf 13,0 %). Auch dieser Anteil relativierte sich nach der ersten Covid-19-Pandemiewelle wieder auf das Niveau vor der Covid-19-Pandemie (10,0 % zwischen beiden Pandemiewellen, 11,2 % während der zweiten Pandemiewelle, und 10,0 % nach der zweiten Pandemiewelle).

Eine ebensolche Zunahme war für die Anzahl der Patienten in der Neurochirurgie zu beobachten. Hier wurden vor der Covid -19-Pandemie 9,6 % der Patienten operiert, während dies während der ersten Covid -19-Pandemiewelle 14,0 % der Patienten waren. Nach der ersten Covid -19-Pandemiewelle nahm dieser Anteil wieder etwas ab, blieb aber auch nach der zweiten Covid -19-Pandemiewelle höher als vor der Pandemie (11,1 %).

Die Anzahl der ambulant behandelten Patienten nahm ab der ersten Covid -19-Pandemiewelle kontinuierlich zu (erste Welle: 8,9 %, zwischen den Pandemiewellen: 10,9 %, zweite Welle: 11,1 %, nach der zweiten Pandemiewelle: 10,9 %).

Während der ersten Pandemiewelle war eine Zunahme von Operationen außerhalb der normalen Betriebszeiten (von 7,3 % auf 9,1 %) und an den Wochenenden (von 4,3 % auf 5,1 %) zu beobachten. Die Anzahl der Operationen außerhalb der normalen Betriebszeiten und am Wochenende ähnelte während der zweiten Covid -19-Pandemiewelle hingegen der Anzahl vor der Covid -19-Pandemie.

Die Bestellung von Erythrozytenkonzentraten nahm prozentual während der ersten Covid -19-Pandemiewelle zu (von 34,3 % auf 41,7 %). Zwischen der ersten und zweiten Welle nahm die Bestellung von Erythrozytenkonzentraten wieder ab (zwischen den Wellen: 36,8 %, zweite Welle: 36,5 %) und erreichte nach der zweiten Welle das niedrigste Niveau (31,0 %). Gleichzeitig nahmen während der ersten Covid -19-Pandemiewelle auch die Bestellungen von gefrorenem Frischplasma von 26,7 % auf 30,5 % zu. Nach der ersten Covid -19-Pandemiewelle

nahm dieser Anteil wieder ab und lag nach der zweiten Pandemiewelle unter dem relativen Anteil vor der Pandemie (17,6 %).

	Vor der Pandemie	**Erste Covid - 19-Welle**	**Zwischen der ersten und zweiten Covid - 19-Welle**	**Zweite Covid - 19-Welle**	**Nach der zweiten Covid-19-Welle**
Zeitraum	06.2014 – 03.2020	04.2020 – 05.2020	06.2020 – 09.2020	10.2020 – 05.2021	06.2021 – 10.2021
Anzahl der Patienten	107468	2251	6504	12112	8042
Anzahl der Verstorbenen (%)	883 (0.8)	22 (1.0)	61 (0.9)	125 (1.0)	65 (0.8)
Alter (Median [IQR])	56.00 [38.00, 70.00]	58.00 [40.00, 72.00]	57.00 [39.00, 71.00]	57.00 [39.00, 72.00]	57.00 [38.00, 72.00]
Geschlecht männlich (%)	49282 (45.9)	1065 (47.3)	2896 (44.5)	5815 (48.0)	3789 (47.1)
BMI vorhanden	34276 (31.9)	484 (21.5)	1114 (17.1)	2061 (17.0)	1329 (16.5)
BMI (median [IQR])	25.24 [22.50, 28.58]	24.86 [22.21, 28.40]	25.37 [22.64, 28.72]	25.26 [22.60, 28.72]	25.26 [22.51, 28.69]
Anzahl der Patienten ohne ASA	31407 (29.2)	440 (19.5)	1058 (16.3)	1917 (15.8)	1212 (15.1)
ASA (%)					
1	21204 (27.9)	415 (22.9)	1392 (25.6)	2498 (24.5)	1637 (24.0)
2	38999 (51.3)	889 (49.1)	2667 (49.0)	5021 (49.2)	3417 (50.0)
3	15113 (19.9)	474 (26.2)	1317 (24.2)	2561 (25.1)	1700 (24.9)
4	695 (0.9)	33 (1.8)	70 (1.3)	109 (1.1)	76 (1.1)
5	50 (0.1)	0 (0.0)	0 (0.0)	6 (0.1)	0 (0.0)
Anzahl der Patienten mit erfasstem Mallampati-Score (%)	42492 (39.5)	747 (33.2)	1634 (25.1)	2804 (23.2)	1624 (20.2)
Mallampati (%)					
I	27210 (41.9)	677 (45.0)	2358 (48.4)	4243 (45.6)	2906 (45.3)
II	28633 (44.1)	603 (40.1)	1893 (38.9)	3824 (41.1)	2666 (41.5)
III	7242 (11.1)	177 (11.8)	495 (10.2)	1043 (11.2)	720 (11.2)
IV	1891 (2.9)	47 (3.1)	124 (2.5)	198 (2.1)	126 (2.0)
Aufnahme (%)					
Aufnahme aus externem Krankenhaus	2269 (2.1)	54 (2.4)	119 (1.8)	233 (1.9)	142 (1.8)
Geburt	4236 (4.0)	103 (4.6)	258 (4.0)	496 (4.1)	327 (4.1)
Elektive Fälle	77843 (72.6)	1491 (66.8)	4704 (72.5)	8680 (71.9)	5689 (70.9)
Notfälle	19342 (18.0)	509 (22.8)	1190 (18.3)	2306 (19.1)	1543 (19.2)
Neugeborene	5 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Andere	661 (0.6)	21 (0.9)	61 (0.9)	151 (1.3)	166 (2.1)
Poliklinik	2853 (2.7)	54 (2.4)	154 (2.4)	213 (1.8)	159 (2.0)
Anzahl der Konsile vor OP (median [IQR])	2.00 [1.00, 3.00]	3.00 [2.00, 4.25]	3.00 [2.00, 4.00]	3.00 [2.00, 5.00]	3.00 [2.00, 5.00]

Tabelle 2: Patientencharakteristika entsprechend der Pandemiewellen.

	Vor der Pandemie	Erste Covid -19-Welle	Zwischen der ersten und zweiten Covid -19-Welle	Zweite Covid -19-Welle	Nach der zweiten Covid -19-Welle
Zeitraum	06.2014 – 03.2020	04.2020 – 05.2020	06.2020 – 09.2020	10.2020 – 05.2021	06.2021 – 10.2021
Abteilung (%)					
Orthopädie/ Unfallchirurgie	19153 (17.8)	334 (14.8)	1145 (17.6)	2037 (16.8)	1433 (17.8)
Gynäkologie/ Geburtshilfe	10692 (10.0)	293 (13.0)	647 (10.0)	1351 (11.2)	803 (10.0)
HNO/ MKG / Augenheilkunde	25171 (23.4)	402 (17.9)	1378 (21.2)	2406 (19.9)	1647 (20.5)
Neurochirurgie	10259 (9.6)	315 (14.0)	720 (11.1)	1353 (11.2)	895 (11.1)
Ambulant	9582 (8.9)	201 (8.9)	707 (10.9)	1347 (11.1)	880 (10.9)
Allgemeine Chirurgie	20397 (19.0)	421 (18.7)	1104 (17.0)	2087 (17.2)	1382 (17.2)
Urologie	12160 (11.3)	285 (12.7)	801 (12.3)	1530 (12.6)	1002 (12.5)
Anteil OPs außerhalb der Normal-betriebszeit (%)	7810 (7.3)	204 (9.1)	480 (7.4)	887 (7.3)	577 (7.2)
OPs an Wochenenden (%)	4658 (4.3)	115 (5.1)	242 (3.7)	516 (4.3)	323 (4.0)
Bestellung von Erythrozyten-konzentraten (%)	29212 (34.3)	849 (41.7)	2163 (36.8)	4023 (36.5)	2292 (31.0)
Wenn ja: Median [Q1, Q3] (Median [IQR])	4.00 [2.00, 4.00]	4.00 [2.00, 4.00]	4.00 [2.00, 4.00]	4.00 [2.00, 4.00]	2.00 [2.00, 4.00]
Bestellung von gefrorenem Frischplasma (%)	22724 (26.7)	621 (30.5)	1590 (27.0)	2676 (24.2)	1301 (17.6)
Wenn ja: Median [Q1, Q3] (Median [IQR])	4.00 [2.00, 4.00]	4.00 [2.00, 4.00]	4.00 [2.00, 4.00]	4.00 [2.00, 4.00]	4.00 [2.00, 4.00]
Bestellung von Prothrombin Komplex (%)	441 (0.5)	12 (0.6)	40 (0.7)	91 (0.8)	47 (0.6)
Wenn ja: Median [Q1, Q3] (Median [IQR])	4.00 [3.00, 4.00]	4.00 [4.00, 6.00]	4.00 [4.00, 6.00]	4.00 [3.00, 6.00]	4.00 [2.00, 4.00]
CRP nicht vorhanden (C react. Protein)	49466 (46.0)	879 (39.0)	2419 (37.2)	4579 (37.8)	2937 (36.5)
CRP mg/dL (C reakt. Protein) (Median [IQR])	0.30 [0.10, 1.00]	0.30 [0.10, 1.60]	0.30 [0.10, 1.00]	0.30 [0.10, 1.00]	0.30 [0.10, 1.00]
Leukozyten nicht vorhanden	25031 (23.3)	267 (11.9)	734 (11.3)	1312 (10.8)	758 (9.4)
Leukozyten 10^6/µl (Median [IQR])	7.26 [5.90, 9.12]	7.24 [5.76, 9.24]	7.16 [5.75, 9.02]	7.19 [5.79, 9.13]	7.17 [5.77, 9.08]
Albumin nicht vorhanden	98097 (91.3)	1976 (87.8)	5804 (89.2)	10726 (88.6)	7168 (89.1)
Albumin g/dL (Median [IQR])	4.40 [3.90, 4.60]	4.30 [3.75, 4.70]	4.40 [3.90, 4.70]	4.30 [3.70, 4.60]	4.30 [3.80, 4.60]
QUICK Wert nicht vorhanden	25443 (23.7)	283 (12.6)	772 (11.9)	1372 (11.3)	794 (9.9)
QUICK % (median [IQR])	103.00 [95.00, 112.00]	116.00 [107.00, 120.00]	108.00 [99.00, 115.00]	107.00 [98.00, 114.00]	106.00 [98.00, 113.00]

Tabelle 2: Patientencharakteristika entsprechend der Pandemiewellen (Fortsetzung).

4.2 Vergleich des Anteils fehlender Werte zu den Untersuchungszeitpunkten.

Zusätzlich zu den absoluten Werten wurde in der Datenerhebung auch erfasst, für wie viele Patienten Daten zu den CRP-Werten, den Leukozyten, den Albuminwerten, und dem QUICK-Wert fehlten.

Aus Tabelle 2 ist bereits ersichtlich, dass CRP-Wert, Albumin, Leukozytenzahl und QUICK-Werte während der ersten Pandemiewelle deutlich öfter angefordert wurden als vor der Covid-19-Pandemie. Lagen vor der Pandemie noch für 46,0 % der Patienten keine Daten zu den CRP-Werten vor, so waren es während der ersten Welle nur noch 39,0 % der Patienten. Die Anforderung der Leukozytenzahl aus dem Labor nahm von 23,3 % der Patienten vor der Pandemie auf 11,9 % während der ersten Pandemiewelle zu. Wurde vor der Pandemie für 23,7% der Patienten kein QUICK-Wert angefordert, so waren es während der ersten Pandemiewelle nur noch 12,6 %.

Im Gegensatz dazu nahm die Bedeutung der ASA-Klassifikation seit Beginn der Covid -19-Pandemie zu, da im Vergleich zu 29,2 % der Patienten vor der Pandemie nur noch für 19,5 % der Patienten während der ersten Covid -19-Pandemiewelle keine ASA-Kategorie vorhanden war. Der prozentuale Anteil dieser Patienten ohne eine dokumentierte ASA-Kategorie nahm nach der ersten Covid-19-Pandemiewelle kontinuierlich weiter ab (zwischen der ersten und zweiten Pandemiewelle: 16,3 %, während der zweiten Pandemiewelle 15,8 %, nach der zweiten Pandemiewelle: 15,1 %). Der Mallampati-Score dagegen wurde im Verlauf der Pandemie zunehmend seltener dokumentiert. Dies liegt wahrscheinlich auch daran, dass der Mallampati-Score nicht zu den Pflichtfeldern des Kerndatensatzes Anästhesie gehört und deswegen wahrscheinlich nicht immer aus der Papierdokumentation der auf Station prämedizierten Patienten übertragen wurde.

Die hinsichtlich des BMI der Patienten verfügbaren Daten zeigten, dass der BMI im Median zwischen 24,8 und 25,3 kg/m^2 lag. Auch die Dokumentation des BMI wurde im Verlauf schlechter.

4.3 Vorhersagemodelle im Rahmen der Covid-19-Pandemie

Insgesamt wurden drei Modelle mit den Daten von über 136000 Patienten erstellt. Die Modelle wurden entsprechend der Covid-19-Wellen wie folgt eingeteilt:

Das erste Modell beinhaltete die Daten von vor der Covid-19-Pandemie behandelten Patienten. Die Daten wurden in drei Datensätze unterteilt, wobei 60 % der Daten für Training verwendet wurden, 20 % der Daten für den Test, und 20 % der Daten für die interne Validierung. Das Modell wurde dann auf die Daten aus den vier Untersuchungszeiträumen angewendet und die jeweilige Vorhersagegüte analysiert.

Im zweiten Modell wurden die Daten von vor der Covid-19-Pandemie behandelten Patienten mit denen während der ersten Pandemiewelle behandelten Patienten einbezogen, und das Modell auf die Daten der restlichen drei Untersuchungszeiträume angewendet (externe Validierung).

In das dritte Modell wurden schließlich alle Daten des gesamten Untersuchungszeitraums inkorporiert, also vor der Pandemie, während der ersten Pandemiewelle, nach der ersten Pandemiewelle, während der zweiten Pandemiewelle und nach der zweiten Pandemiewelle. Die Ergebnisse der Modelle (AUROC und PR) sind in den Abbildungen 11 bis 16 und in Tabelle 3 dargestellt.

Die in der PR Grafiken eingezeichnete gestrichelte Linie (Baseline) entspricht der Vorhersagegüte eines Zufallsschätzers und liegt auf Grund der sehr niedrigen Mortalitätsrate deutlich unter 1 %.

Die Modelle wiesen AUROC-Werte zwischen 0,904 und 0,951 auf. Bei der Betrachtung der Kurve während der ersten Covid-19-Pandemiewelle ist ersichtlich, dass die pandemiebedingten Veränderungen die Vorhersagekraft der Modelle beeinflusste, da sowohl die AUROC als auch die AUPR in diesem Untersuchungszeitraum schlechter waren als vor der Covid -19-Pandemie (Tabelle 3).

Anhand des zweiten Modells (Abbildung 13) zeigt sich, dass die Vorhersagekraft während der zweiten Pandemiewelle der vor der Pandemie und während der ersten Welle überlegen ist. Modell 1 war hinsichtlich der AUROC und somit der Vorhersagekraft beiden anderen Modellen überlegen. Das zweite Modell wird in der zweiten Welle besser als im vorpandemischen Zeitraum und in der ersten Welle.

Die AUROC bleibt für jedes Modell zu jedem Zeitpunkt bei einem Wert über 0,9. Auffällig ist jedoch die Entwicklung der AUPR: Das präpandemische Modell verliert hier während der ersten Pandemiewelle etwa 50% seiner Vorhersagegüte und normalisiert sich dann im Verlauf der Pandemie wieder. Das mit Daten vor der Pandemie und aus der ersten Welle erstellte Modell reagiert mit der AUPR ebenso nach der ersten Welle. Das auf dem gesamten Datensatz

trainierte Modell weist in der externen Validierung Werte auf, die mit denen des präpandemischen Modells in der präpandemischen Phase vergleichbar sind.

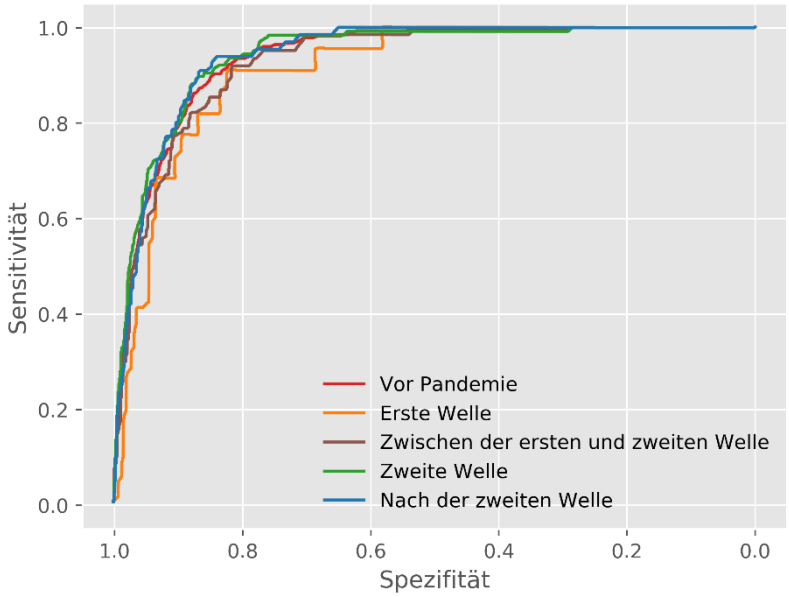

Abbildung 11: AUROC-Kurven für das erste Modell.

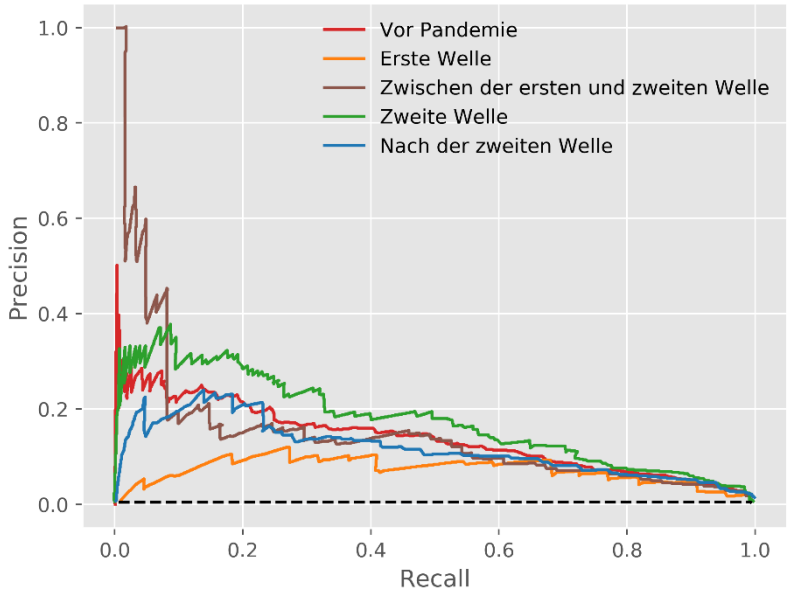

Abbildung 12: PR-Kurven für das erste Modell.

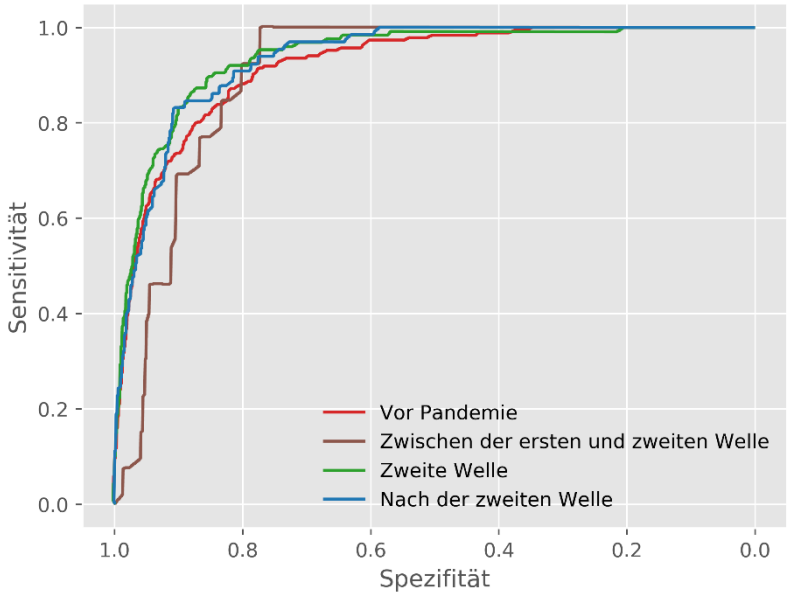

Abbildung 13: AUROC-Kurven für das zweite Modell.

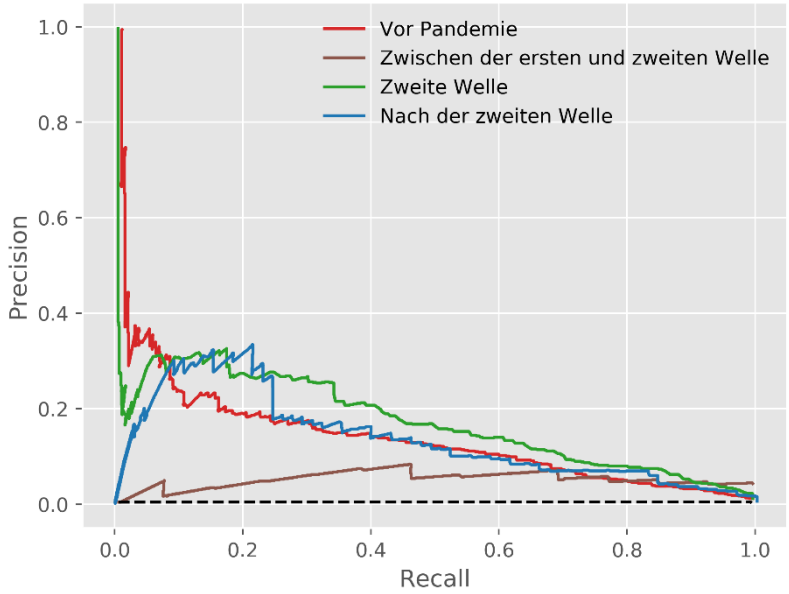

Abbildung 14: PR-Kurven für das zweite Modell.

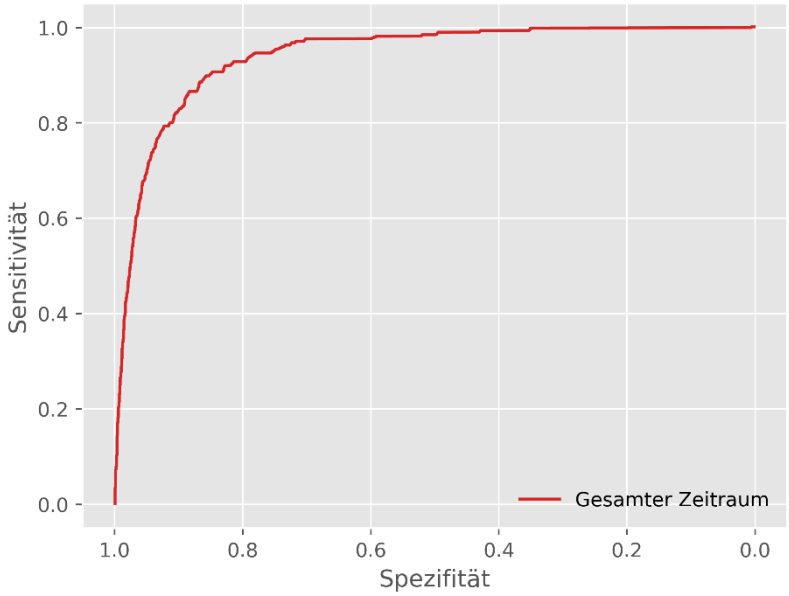

Abbildung 15: AUROC-Kurve für das dritte Modell.

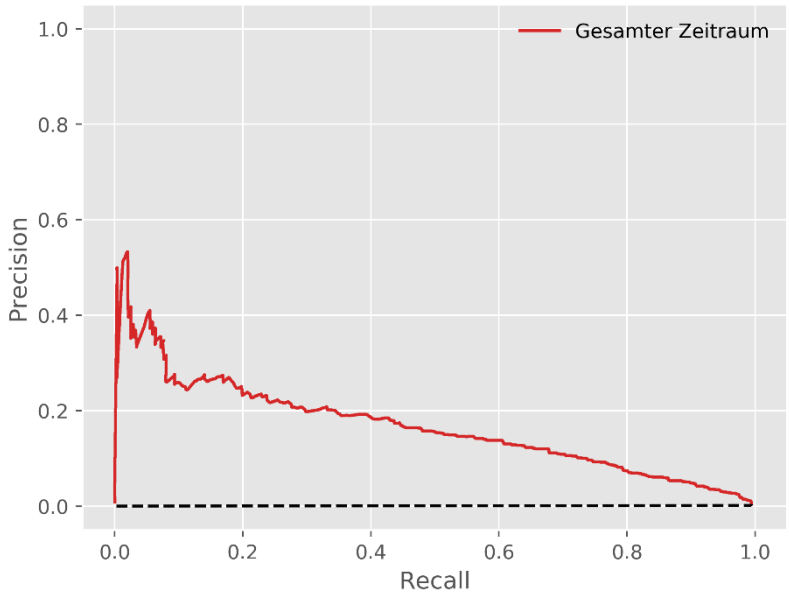

Abbildung 16: PR-Kurve für das dritte Modell.

Trainings-datensatz	Validierungs-Parameter	Validierungsdatensatz				
		Vor Pandemie 06.2014-03.2020	**Erste Welle 04.2020-05.2020**	**Zwischen der ersten und zweiten Welle 06.2020-09.2020**	**Zweite Welle 10.2020-05.2021**	**Nach der zweiten Welle 06.2021-10.2021**
Vor Pandemie 06.2014-03.2020	AUROC	0.951	0.914	0.931	0.944	0.944
	[95% CI]	[0.941 - 0.962]	[0.871 - 0.957]	[0.909 - 0.953]	[0.929 - 0.959]	[0.927 - 0.961]
	AUPR	0.144	0.074	0.150	0.177	0.118
	[95% CI]	[0.140 - 0.149]	[0.064 - 0.086]	[0.141 - 0.159]	[0.171 - 0.184]	[0.111 - 0.125]
Vor Pandemie & Erste Welle 06.2014-05.2020	AUROC	0.923		0.907	0.942	0.937
	[95% CI]	[0.907 - 0.940]		[0.870 - 0.943]	[0.924 - 0.959]	[0.917 - 0.958]
	AUPR	0.142		0.052	0.174	0.136
	[95% CI]	[0.138 - 0.147]		[0.041 - 0.066]	[0.169 - 0.179]	[0.129 - 0.144]
Gesamter Datensatz 06.2014-10.2021	AUROC	0.941				
	[95% CI]	[0.927 - 0.954]				
	AUPR	0.168				
	[95% CI]	[0.164 - 0.173]				

Tabelle 3: Validierungsdaten (AUROC, AUPR) der drei Modelle [52].

Die Cut-off-Werte der drei Modelle (Sensitivität, Spezifität, positiver prädiktiver Wert, negativer prädiktiver Wert für jeden Validierungszeitraum), die auf der Grundlage der Youden-Indizes ermittelt wurden, sind in Tabelle 4 aufgeführt. Liegt die vom Algorithmus ermittelte Wahrscheinlichkeit unter dem Cut-off-Wert, so wird der Patient als überlebend eingestuft, liegt er darüber, so wird er als verstorben eingestuft. Am positiven prädiktiven Wert erkennt man, dass der Algorithmus das Risiko bei vielen Patienten überschätzt, woraus sich viele falsch Positive ergeben. Am negativen prädiktiven Wert erkennt man, dass der Algorithmus die Patienten mit niedrigem Risiko sehr gut erkennt. Zu beachten ist hierbei, dass beim Risiko eine linksschiefe Verteilung vorlag, d.h. es gab sehr viele Patienten mit einem Risiko zwischen 0 und 1%. Bei guter Sensitivität und guter Spezifität war der positive prädiktive Wert durchgehend schwach. Die Spezifität des ersten Modells nahm bei der Anwendung auf die Daten der ersten Welle erheblich ab. Die restlichen zu beobachtenden Veränderungen waren aufgrund der breiten Konfidenzintervalle nicht signifikant, aber es ist anzunehmen, dass das zweite Modell an Sensitivität verliert, wenn es auf Daten aus der Zeit nach der zweiten Welle angewendet wird.

Trainings-datensatz	Validierungs-Parameter	Validierungsdatensatz					Cut-off
		Vor Pandemie 06.2014-03.2020	**Erste Welle 04.2020-05.2020**	**Zwischen der ersten und zweiten Welle 06.2020-09.2020**	**Zweite Welle 10.2020-05.2021**	**Nach der zweiten Welle 06.2021-10.2021**	
Vor Pandemie 06.2014-03.2020	Sensitivität [95% CI]	0.722 [0.664-0.774]	0.773 [0.546-0.922]	0.672 [0.540-0.787]	0.760 [0.675-0.832]	0.677 [0.549-0.788]	0.1611
	Spezifität [95% CI]	0.927 [0.924-0.930]	0.894 [0.880-0.906]	0.932 [0.925-0.938]	0.921 [0.916-0.926]	0.940 [0.935-0.946]	
	PPV [95% CI]	0.086 [0.075-0.098]	0.067 [0.039-0.105]	0.085 [0.062-0.114]	0.091 [0.074-0.110]	0.085 [0.062-0.112]	
	NPV [95% CI]	0.997 [0.996-0.998]	0.997 [0.994-0.999]	0.997 [0.995-0.998]	0.997 [0.996-0.998]	0.997 [0.996-0.998]	
Vor Pandemie & Erste Welle 06.2014-05.2020	Sensitivität [95% CI]	0.730 [0.660-0.792]		0.462 [0.192-0.749]	0.753 [0.685-0.812]	0.662 [0.534-0.774]	0.1805
	Spezifität [95% CI]	0.915 [0.912-0.919]		0.927 [0.911-0.941]	0.920 [0.916-0.923]	0.933 [0.927-0.938]	
	PPV [95% CI]	0.065 [0.055-0.076]		0.061 [0.023-0.129]	0.082 [0.069-0.096]	0.074 [0.054-0.099]	
	NPV [95% CI]	0.998 [0.997-0.998]		0.994 [0.988-0.998]	0.997 [0.997-0.998]	0.997 [0.996-0.998]	
Gesamter Datensatz 06.2014-10.2021	Sensitivität [95% CI]	0.258 [0.204-0.319]					0.1296
	Spezifität [95% CI]	0.992 [0.991-0.993]					
	PPV [95% CI]	0.221 [0.173-0.275]					
	NPV [95% CI]	0.994 [0.992-0.994]					

Tabelle 4: Statistische Benchmarks zu den jeweiligen Cut-off-Werten [52].

4.4 Information Gain

„Information Gain“ oder „importance“ bezeichnet den prozentualen Anteil der jeweiligen Variable an der Vorhersage [57], also dessen Wichtigkeit und wird in Abbildung 17, Abbildung 18 und Abbildung 19 für die untersuchten Zeiträume als sogenannter „Importance Plot“ dargestellt. Hier sind in abnehmender Reihenfolge der Wichtigkeit folgende Parameter aufgelistet: Alter, bestellte Erythrozytenkonzentrate, Anzahl präoperativer Konsile, CRP-Werte, Albuminwerte, Hämoglobinwerte, Aufnahmegrund, ASA-Kategorie, QUICK-Wert, Thrombozytenzahl, Fachabteilung, BMI, Hämatokritwert, Erythrozytenzahl, Anzahl bestellter

FFPs und Leukozytenzahl. Die Darstellung erfolgt für die Situation vor der Covid -19-Pandemie (Abbildung 17), für den Zeitraum vor der Pandemie und während der ersten Pandemiewelle (Abbildung 18) und für den gesamten Untersuchungszeitraum (Abbildung 19).

Hier zeigte sich, dass das Alter der Patienten in allen Untersuchungszeiträumen an oberster Stelle der Wichtigkeit stand, mit einem Anteil von 9,77% an der Vorhersage. Auch die Erythrozytenzahl, der CRP-Wert und die Anzahl der präoperativen Konsile waren in allen drei Untersuchungszeiträumen wichtig. Seit dem Beginn der Covid -19-Pandemie war eine Zunahme der Bedeutung der ASA-Kategorie zu verzeichnen, und auch die Wichtigkeit der Leukozytenzahl nahm zu. An den Abbildungen wird ferner deutlich, dass die jeweils 16 wichtigsten Faktoren nicht einmal die Hälfte der Prädiktion erklären. Es sind eine Vielzahl weiterer Einflussfaktoren (Im Modell 1: 588, im Modell 2: 276 und im Modell 3: 924 Faktoren) mit kleinen prozentualen Anteilen beteiligt.

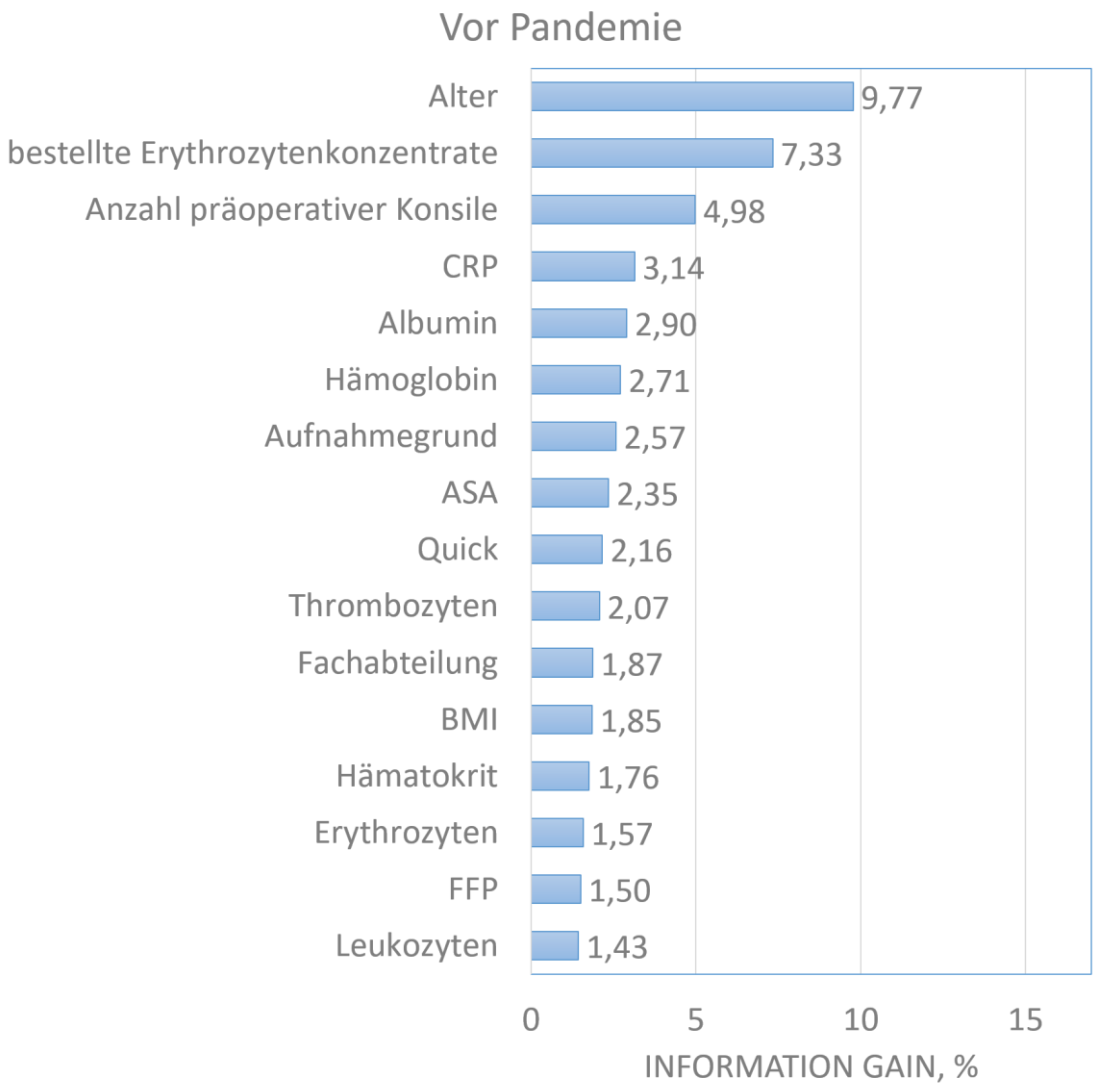

Abbildung 17: Information Gain vor der Pandemie. Der Anteil des jeweiligen Parameters an der Modellvorhersage, die sogenannte „Importance“ oder „Wichtigkeit“, wird hier für die 16 wichtigsten Faktoren in absteigender Reihenfolge dargestellt. ASA = American Society of Anaesthesiologists Physical Score, BMI = Body Mass Index, CRP = C-reaktives Protein, FFP = gefrorenes Frischplasma.

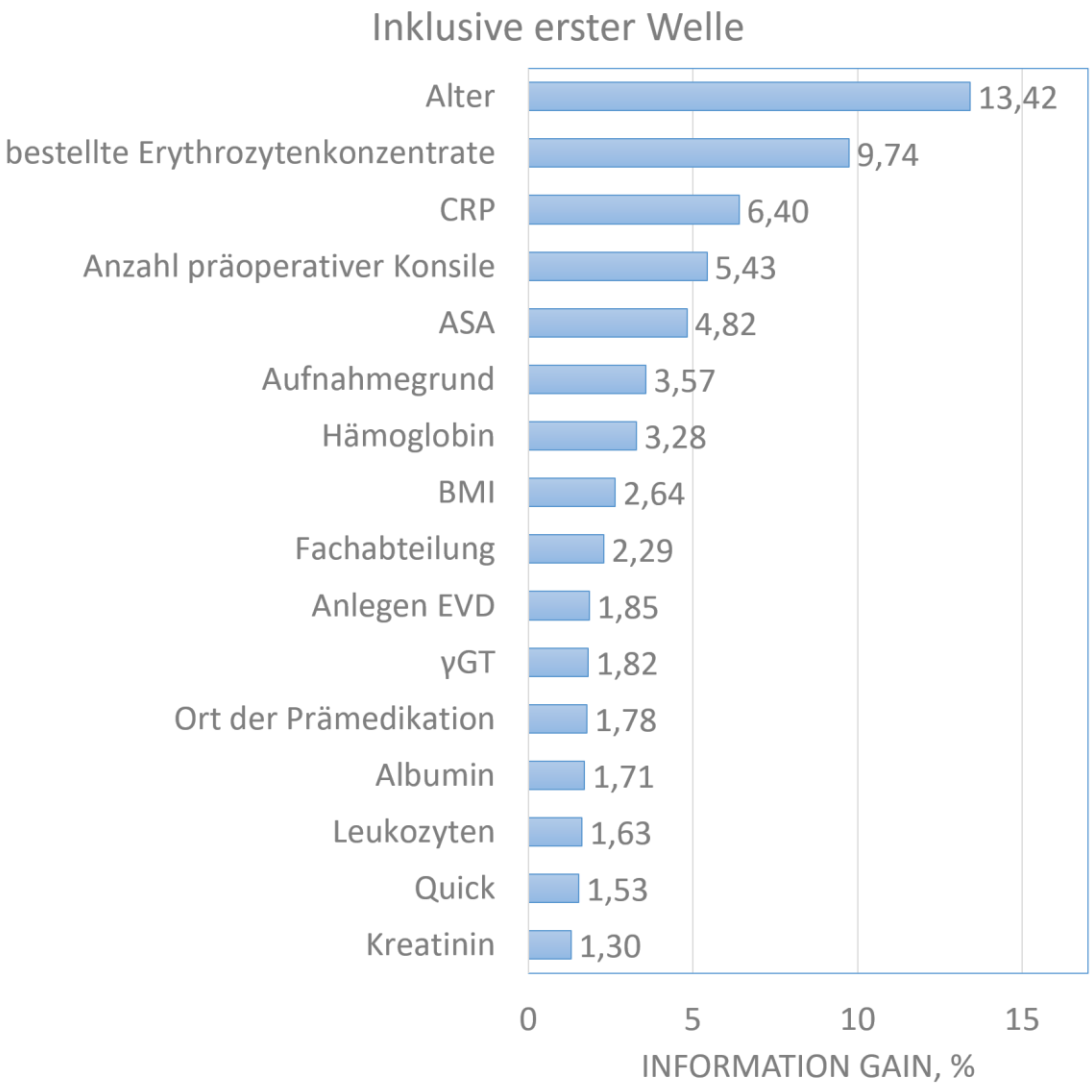

Abbildung 18: Information Gain einschließlich der ersten Covid-19-Welle. ASA = American Society of Anaesthesiologists Physical Score, BMI = Body Mass Index, CRP = C-reaktives Protein, EVD = Externe Ventrikeldrainage.

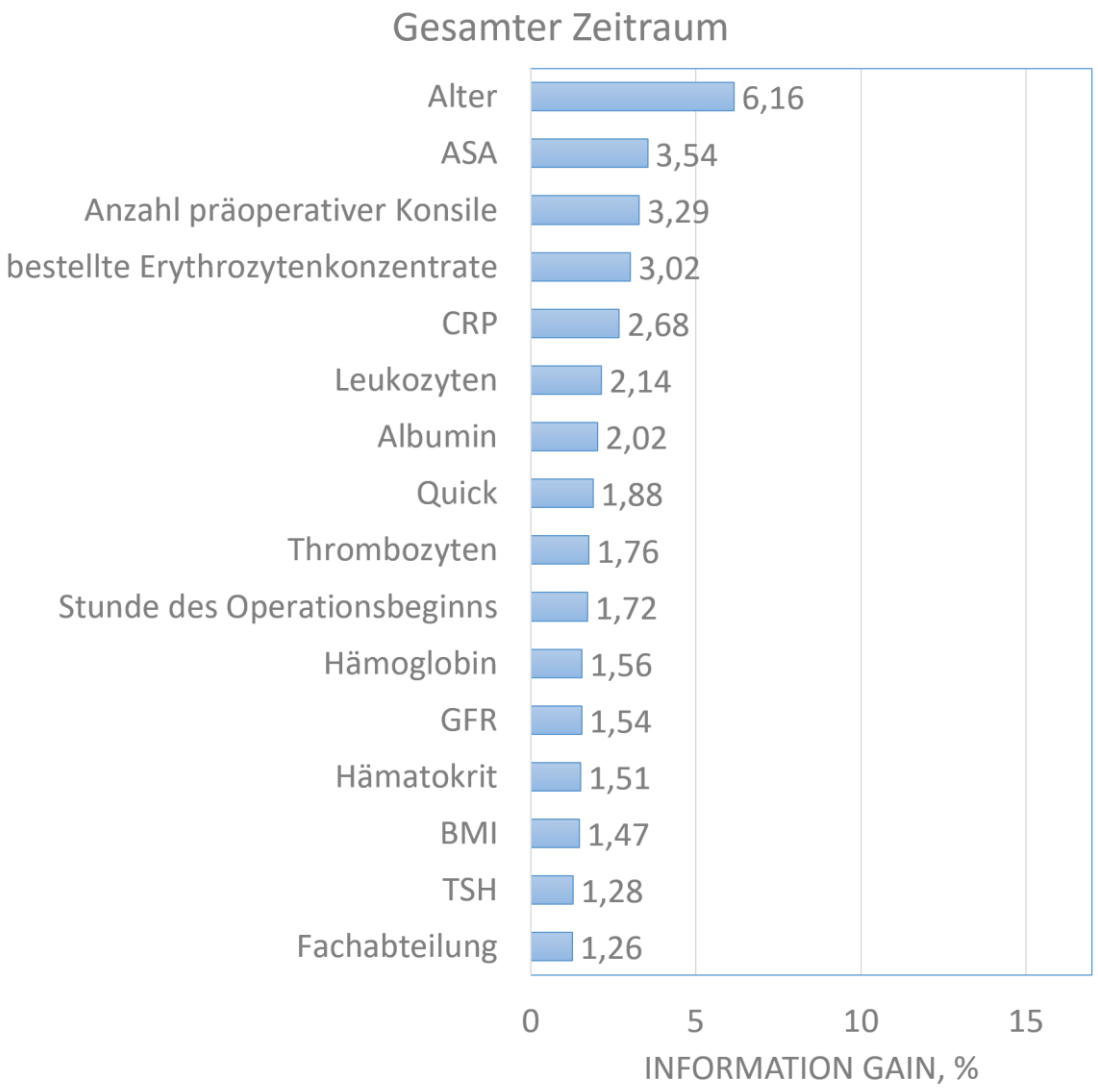

Abbildung 19: Information Gain gesamter Zeitraum. ASA = American Society of Anaesthesiologists Physical Score, BMI = Body Mass Index, CRP = C-reaktives Protein, GFR = Glomeruläre Filtrationsrate, TSH = Thyreoidea stimulierendes Hormon.

5 Diskussion

5.1 Zusammenfassung und Interpretation der Ergebnisse

Das Ziel der vorliegenden Studie war es, unter Verwendung eines XGBoost-Algorithmus zu untersuchen, inwiefern maschinelle Lernmodelle in der Lage sind, anhand präoperativer Daten eine zuverlässige Prädiktion zur perioperativen Mortalität zu treffen. Anschließend sollte die Auswirkung äußerer Einflüsse auf die Vorhersagekraft derartiger Modelle am Beispiel der Covid-Pandemie untersucht werden. Hierbei zeigte sich zwar eine durchwegs akzeptable Vorhersagegenauigkeit sowohl der unangepassten, als auch der mit neuen Daten trainierten Modelle für die Mortalität, allerdings ist der Einfluss der Covid-Pandemie deutlich zu erkennen.

Die AUROC-Werte aller drei in der vorliegenden Studie entwickelten Modelle war sehr hoch und lagen für alle Kurven über 0,900. Dies traf somit sowohl auf die Patienten vor der Covid -19-Pandemie zu als auch auf solche, die während der Pandemie behandelt wurden. Diese Ergebnisse deuten auf eine hohe Vorhersagekraft XGBoost-basierter Modelle hinsichtlich des Patientenergebnisses hin. Die AUROC-Werte wurden durch die Covid-19-Pandemie nur geringfügig beeinflusst und näherten sich nach der zweiten Welle bereits den AUROC-Werten aus der vorpandemischen Zeit an.

Das vor der Pandemie trainierte Modell zeigte hinsichtlich der AUROC keinen signifikanten Unterschied bei Anwendung während oder zwischen den Pandemiewellen. Das Modell war in der Lage, die überlebenden Patienten in den meisten Fällen richtig zu erkennen, was sich in den hohen AUROC-Werten zeigte. Allerdings wurde dies aufgrund des zugrundeliegenden unbalancierten Datensatzes bei insgesamt niedriger Mortalität erleichtert. Der Precision-Recall-Tradeoff, ersichtlich aus der AUPR, war allerdings während der ersten Welle deutlich schlechter. Das bedeutet, dass viele Patienten, für die das Modell ein hohes Mortalitätsrisiko errechnet hatte, trotzdem überlebten. Dies stellt sich auch in den schlechten positiven prädiktiven Werten am mittels des Youden-Index berechneten cut-off dar. Zusammengefasst erkennen die Modelle sehr gut die „Negativen“, also die überlebenden Patienten, ermitteln aber sehr viele „falsch Positive“, das heißt, Patienten, die überleben, obwohl der Algorithmus ein Risiko ermittelt hat, das am cut-off-Punkt zu einer Einteilung in die Gruppe der verstorbenen Patienten geführt hätte.

Interessanterweise war das auf den Gesamtdaten trainierte Modell nicht besser als das aus Daten vor der Pandemie erstellte Modell. Die AUPR-Werte der ersten beiden Modelle unterschieden sich nur minimal (Modell 1: AUPR = 0,144;

Modell 2: AUPR = 0,142), während im dritten Modell ein höherer AUPR-Wert von 0,168 zu verzeichnen war. Insgesamt sind die AUPR-Werte aufgrund der niedrigen Baseline bei niedriger Mortalität als akzeptabel einzuordnen, da sie immer noch um ein Vielfaches besser sind als zufälliges Raten. Bei dem AUROC der XGBoost-Modelle war ein deutlicher Abfall zwischen den ersten beiden Modellen (Modell 1: AUROC = 0,951; Modell 2: AUROC = 0,923) zu verzeichnen, während der AUROC von Modell 3: 0,941 wieder dem von Modell 1 ähnelte.

Die Corona-Pandemie mit den damit verbundenen Einschränkungen und einer Verschiebung des OP-Spektrums hin zu dringlichen und Notfalleingriffen brachte Änderungen des Patientenkollektivs. Es zeigten sich deutliche Unterschiede in den Patientencharakteristika vor Beginn der Covid-19-Pandemie im Vergleich zu der ersten Pandemiewelle und zwischen beiden Covid-19 Pandemiewellen. So nahm die Anzahl von Operationen in bestimmten Fachabteilungen ab, da zum Beispiel elektive Eingriffe in den Abteilungen HNO, Augenheilkunde und Orthopädie nur noch sehr bedingt stattfanden. Dies ist unter anderem darauf zurückzuführen, dass aufgrund der Covid-19-bedingten Gesetzesvorgaben elektive Eingriffe während der ersten Welle deutlich zurückgingen und sich auf Notfall- und dringliche Eingriffe konzentriert wurde. Hierfür spricht auch, dass die Anzahl der Behandlungen der operativen Eingriffe an den Wochenenden und außerhalb der normalen Dienstzeiten während der Pandemiewellen zunahm. Zudem ist es denkbar, dass Patienten ihre bereits terminierten Operationen aus Angst vor Covid-19 absagten.

Dies wirkte sich auf die Wichtigkeit der in den Modellen verwendeten Faktoren aus: Das Alter der Patienten war in allen Untersuchungszeiträumen an erster Stelle. Die Erythrozytenzahl, der CRP-Wert und die Anzahl der präoperativen Konsile waren stets unter den 5 wichtigsten Faktoren zu finden. Die Bedeutung der ASA-Kategorie nahm im Verlauf der Pandemie an Wichtigkeit zu ebenso wie die Leukozytenzahl.

Die Tatsache, dass während der ersten Pandemiewelle beobachtete Änderungen in der Anzahl der operativen Eingriffe in bestimmten Abteilungen im Rahmen der zweiten Pandemiewelle entweder wieder auf dem Niveau vor der Pandemie waren, oder sich aber zumindest im Vergleich zu der ersten Welle der Situation vor der Pandemie annäherten, deutet auf Unterschiede zwischen beiden Pandemiewellen hin. Dies könnte darauf zurückzuführen sein, dass die Ärzte während der ersten Pandemiewelle Erfahrungen mit der Situation sammeln und diese in der zweiten Welle entsprechend anwenden konnten. Es ist in diesem Zusammenhang auffällig, dass die Patientenparameter während der zweiten Pandemiewelle insgesamt eher der Situation vor der Pandemie ähneln als der Situation während der ersten Pandemiewelle.

Während der zweiten Welle unterschieden sich zudem aufgrund der Wiederaufnahme elektiver Eingriffe das Patientengut und das Operationsspektrum nicht mehr so stark von dem vor der Pandemie. Zudem mussten mit dem Fortschreiten der Pandemie aufgeschobene elektive Eingriffe, der so genannte „Surgical Backlog“, dennoch durchgeführt werden, was dazu geführt hat, dass sich die Zahl der Operationen wieder normalisierte [58].

Es wurde bei der Beurteilung der Wichtigkeit der erhobenen Parameter beobachtet, dass die Importance der ASA-Klassifikation während der Pandemie anstieg. Dies kann einerseits bedeuten, dass die Klassifikation tatsächlich als wichtiger eingeschätzt wird, seit die Covid-19-Pandemie begann, zugleich wurden jedoch die ASA-Daten häufiger erhoben, es gab also weniger Patienten mit fehlender Dokumentation des ASA in den präoperativen Unterlagen. Es könnte also sein, dass die Anästhesisten seit Beginn der Pandemie genauer darauf achteten, dass einem Patienten eine ASA-Kategorie zugeordnet wurde, da sie diesem Parameter mehr Bedeutung beimaßen. Das Patientenalter war für den gesamten Untersuchungszeitraum durchweg der wichtigste Parameter zur Vorhersage der Mortalität. Dieser Fakt spiegelt sich in vielen anderen Modellen und Scores wieder, z.B. im POSPOM, wo das Alter des Patienten auch einen wichtigen Faktor darstellt. [59, 60]

5.2 Beurteilung der Modelle und Einordnung in die Literatur

Maschinelle Lernmodelle erweisen sich häufig in Vergleichsstudien den konventionellen Screening-Tools überlegen. So konnten Kijoaisalratana et al. beispielsweise zeigen, dass ML-Algorithmen bei der Vorhersage einer Sepsis die konventionellen Scores qSOFA (quick sequential organ failure assessment) und MEWS (modified early warning score) übertreffen [61]. Auch Desautels et al. attestierten einem ML-Triage-System Überlegenheit gegenüber qSOFA, MEWS und SIRS (systemic inflammatory response syndrome) [62]. Taylor et al. fanden heraus, dass die Vorhersage der Krankenhaussterblichkeit anhand eines ML-Algorithmus herkömmlichen Entscheidungsfindungssystemen wie dem CART-Algorithmus (classification and regression tree) und MEDS (Mortality in Emergency Department Sepsis score) überlegen war [63].

Der in der präoperativen Evaluation am häufigsten verwendete Score ist der ASA. Für ihn wird in der Literatur allerdings lediglich eine AUROC von 0,66 angegeben [64]. Der POSPOM als modernerer Score erreicht Werte über 0,9, während andere gängige Scores deutlich schlechter bei der Prädiktion der perioperativen Mortalität abschneiden [65]. Unser rein auf der Basis

präoperativ verfügbarer Daten erstelltes Modell ist somit gemäß der aktuellen Literatur den gängigen Scores überlegen [59]. Die AUROC-Werte unseres Modells waren mit > 0,95 sehr hoch und der Algorithmus scheint daher geeigneter zu sein, das Risiko der Patienten vorherzusagen, als ein etablierter Risikoscore wie der ASA. Außerdem können damit Parameter identifiziert werden, die einen besonders großen Anteil zur Prädiktion beitragen, während herkömmliche Scores lediglich eine Aussage über das Risiko, nicht aber über Einflussfaktoren zulassen.

Diese Identifikation von Risikofaktoren wird auch in der Intensivmedizin genutzt: Clark und Lettieri beispielsweise entwickelten einen ML-Algorithmus an 130 Intensivpatienten zur Vorhersage der postoperativen Beatmungspflichtigkeit [66] und identifizierten Parameter, die als Prädiktoren für eine derartige postoperative Komplikation geeignet sein könnten (Herzfrequenz über 110 bpm, Harnstoffwert über 25 mg/dL, Serum-pH unter 7,25, Serum-Kreatinin-Wert über 2,0 mg/dL, HCO_3-Wert unter 20 mEq/L).

Eine höhere Patientenzahl wurde von Chiew et al. (90.785 Patienten) verwendet, die anhand von ML-Modellen die postoperative 30-Tage-Mortalität und die Wahrscheinlichkeit einer Intensivversorgung untersuchten [67]. Mit den verschiedenen Modellen (Random Forest-, Adaptive Boosting-, Gradient Boosting- und Support-Vector-Machine-Algorithmen) wurden AUROC-Werte für die Vorhersage der Mortalität von mindestens 0,89 erzielt, wobei sich Gradient Boosting den anderen Modellen überlegen zeigte. Für die Vorhersage der Wahrscheinlichkeit eines Aufenthaltes auf der Intensivstation lagen die AUROC-Werte zwischen 0,80 und 0,95, je nachdem, welches Modell verwendet wurde. Diese Werte waren niedriger als die für die Mortalität, und auch die Sensitivität der Modelle für diesen Vorhersagefaktor war sehr niedrig. Basierend auf diesen Ergebnissen ist es denkbar, dass sich nicht alle Endpunkte gleichermaßen für die Beurteilung mittels ML-basierter Modelle eignen.

In der retrospektiven Studie von Mohamadlu et al., in der ML-Boosted Trees-Algorithmen für Mortalitätsvorhersagen zum Einsatz kamen, konnten ebenfalls sehr hohe AUROC-Werte von 0,94 und höher erzielt werden [68]. Da es sich um eine multizentrische Studie handelte, wurden die Daten von mehreren Kliniken gepoolt. Bei einer separaten Analyse der einzelnen Kliniken zeigten sich jedoch auch vereinzelt niedrigere AUROC-Werte unter 0,90. Diese Ergebnisse zeigen, dass die Vorhersage möglicherweise durch klinikinterne Faktoren beeinflusst wird und daher die Modelle nicht ohne weiteres auf verschiedene Kliniken übertragbar sind. Es ist daher auch davon auszugehen, dass der von uns entwickelte Algorithmus an Vorhersagekraft einbüßte, wenn er in einer anderen Klinik angewendet werden würde.

In der medizinischen Risikoprädiktion erfreuen sich XGBoost-Modelle zunehmender Beliebtheit, vor allem zur Ermittlung der Wahrscheinlichkeit des Auftretens von Komplikationen:

Mori et al. verwendeten ein XGBoost-Modell zur Vorhersage postoperativer Ereignisse nach koronaren Bypass-Operationen [69]. Die Daten von 378.572 Koronararterien-Bypass-Operationen wurden hinsichtlich der Prädiktion der operativen Mortalität und der postoperativen Komplikationen evaluiert. XGBoost wurde mit Regressionsmodellen verglichen. Anhand der Kalibrierungsdiagramme zeigte sich, dass das logistische Regressionsmodell das Risiko bei 9,8 % der Patienten unterschätzte und bei 10,6 % überschätzte, während das XGBoost-Modell das Risiko bei 6,4 % der Patienten unterschätzte und bei keinem Patienten überschätzte.

Chen et al. wendeten verschiedene ML-Modelle, unter anderem XGBoost an, um Lungenentzündungen als Indikator für Morbidität und Mortalität nach Lebertransplantation vorherzusagen [70]. Von den sechs ML-Modellen schnitt das XGBoost-Modell am besten ab, mit einer AUROC von 0,734. Das XGBoost-Modell war in der Lage, die postoperative Pneumonie anhand von 14 Variablen vorherzusagen.

Einige Studien untersuchten auch die Eignung XGBoost-basierter Modelle im Zusammenhang mit der Covid -19-Pandemie.

Bolourani et al. analysierten ein frühzeitiges Atemversagen aufgrund von Covid -19 als Indikator der Morbidität und Mortalität anhand eines ML-Modells [71]. Drei prädiktive Modelle wurden trainiert und evaluiert. Zwei der Modelle basierten auf Extreme Gradient Boosting (XGBoost), eines auf logistischer Regression. Die Autoren verglichen die Modellleistung zwischen den beiden XGBoost-Modellen und des Regressionsmodells mit einem etablierten Frühwarnscore (Modified Early Warning Score) anhand von AUROC-Kurven, AUPRC-Kurven und anderen Metriken. Das XGBoost-Modell hatte die höchste mittlere Genauigkeit (0,919; AUROC = 0,77) und übertraf die beiden anderen Modelle sowie den Modified Early Warning Score. Das XGBoost-Modell hatte somit eine hohe Vorhersagegenauigkeit und war anderen Frühwarnscores zur Ermittlung des postoperativen Risikos überlegen.

Vaid et al. wandten XGBoost im aktuellen Kontext der Covid -19-Pandemie an, um Hochrisikopatienten zu identifizieren [72]. XGBoost wurde mit anderen Risikoprädiktionsmodellen zur Vorhersage der Sterblichkeit im Krankenhaus und von

kritischen Ereignissen nach der Aufnahme verglichen. Anhand des Modells konnten Risikofaktoren für die Vorhersage kritischer Ereignisse identifiziert werden.

Hong et al. verglichen XG-Boost- und Random Forest-Modelle hinsichtlich ihrer Eignung für die Unterscheidung zwischen dem Zustand von Patienten (kritisch vs. unkritisch) mit Covid -19-assoziierter Pneumonie [73]. Die Ergebnisse zeigten die höchste Vorhersagekraft für das XGBoost-Modell bei der Differenzierung zwischen kritisch-kranken und nicht kritisch kranken Covid-19-Patienten im Vergleich zu den anderen beiden ML-Modellen.

Li et al. verwendeten XG-Boost-Modelle für die Diagnose einer Covid -19-Erkrankung, speziell als Abgrenzung dieser Patienten von an Influenza erkrankten Patienten. Anhand dieses Modells konnten einige diagnostische Variablen identifiziert und die Patienten eindeutig als Covid -19-Patienten klassifiziert werden [74].

Diese Ergebnisse deuten darauf hin, dass Risikomodelle auf der Grundlage von XGBoost eine gute Vorhersage von unerwünschten Ereignissen ermöglichen und das Potenzial besitzen, die klinische Versorgung zu verbessern. Dies wird durch unsere Ergebnisse mit durchwegs hohen AUROC-Werten bestätigt. Leider wird in vielen Studien keine Aussage zum Precision-Recall-Tradeoff gemacht, obwohl dieser, bei derart unbalansierten Datensätzen wie in unserer Studie mit einer Ereignisrate von ca. 1%, noch relevante Zusatzinformationen zur Vorhersagekraft geben würde [75].

5.3 Einfluss der Covid-19 Pandemie auf die Variablen

In unserer Untersuchung zeigten sich die Auswirkungen der Covid-19-Pandemie auf die klinischen Rahmenbedingungen deutlich, in erster Linie in Form einer Abnahme elektiver Eingriffe und einer deutlichen Zunahme von Notfällen. Dies ist im Einklang mit in der Literatur beschriebenen Einflüssen der Covid -19-Pandemie auf die klinische Versorgung, die von einer Abnahme elektiver Eingriffe berichten [76-78]. Bezüglich der Notfallpatienten zeigen sich widersprüchliche Ergebnisse in der Literatur, wobei manche Studien von einer Zunahme der Notfalleingriffe berichten, andere von einer Abnahme [79-82]. Balvardi et al. beobachteten einen Rückgang der Patienten in der Notaufnahme und einen kürzeren Verbleib in der Notaufnahme zu Beginn der Pandemie im März 2020 im Vergleich zum Vorjahr, jedoch blieb die Anzahl der operativen Eingriffe gleich [79]. Im Gegensatz dazu fanden Sá et al. in einer retrospektiven Untersuchung von Notoperationsdaten aus den Jahren 2019 und 2020 im Vergleich, dass die Anzahl aller Notfalloperationen zu Beginn der Covid -19-Pandemie im Jahr

2020 gegenüber dem Vorjahr um 30 % abnahm [80]. Tebala et al. fanden ebenfalls einen signifikanten Rückgang der Einweisungen in die Notaufnahme für Notoperationen zu Beginn der Covid -19-Pandemie im Jahr 2020 im Vergleich zu denselben Monaten im Jahr zuvor [81]. Karlafti et al. schlussfolgerten anhand ihrer Literaturübersicht, dass die Anzahl der Patienten in der allgemeinen Notfallchirurgie im Frühjahr 2020 im Vergleich zum Frühjahr 2019 signifikant abnahm [82]. Im Vergleich zu den Untersuchungen der vorliegenden Studie muss angemerkt werden, dass diese Studien immer nur die erste, frühe Phase der Covid -19-Pandemie (März bis April oder Mai 2020) mit denselben Monaten des Vorjahres verglichen, so dass sie nur die initialen Monate nach Beginn der Pandemie abbilden (Welle 1 in dieser Studie), während in der vorliegenden Untersuchung ein längerer Zeitraum vor Beginn der Pandemie, der Zeitraum zwischen der ersten und der zweiten Welle, die zweite Welle, und die Zeit danach abgedeckt werden.

Die beobachteten Veränderungen in der Patientenkohorte bezüglich elektiver Operationen und Notfalleingriffen während der Covid -19-Pandemie sind im Einklang mit den Ergebnissen ähnlicher Studien. So analysierten Lane et al. beispielsweise die Anzahl von Notfallpatienten in Kanada während der Pandemie im direkten Vergleich von Patienten, die einen Notarzt riefen und solchen, die eine Notaufnahme aufsuchten [83]. Sie fanden eine Zunahme von Patienten, die einen Notarzt riefen um 61 % und eine Abnahme um 35 % von Patienten, die eigenständig in die Notaufnahme kamen. Dies verdeutlicht, dass die Patientencharakteristika sich auch möglicherweise deshalb ändern, weil Patienten zögerlicher mit dem Aufsuchen eines Krankenhauses sind, vermutlich aus Angst vor einer Ansteckung.

Die Änderungen während der Pandemie im Vergleich zu der Situation vor der Pandemie in der vorliegenden Studie waren sowohl an Unterschieden im Patientenkollektiv und an der Veränderung der Wichtigkeit einzelner Faktoren - wie in den Importance Plots dargestellt - erkennbar.

Duckworth und Kollegen haben ein XGBoost-Modell für die Vorhersage von Krankenhauseinweisungen aus der Notaufnahme entwickelt und den durch die Covid-19-Pandemie verursachten Datendrift untersucht. Sie fanden einen Abfall der AUROC nach Ausbruch der Pandemie, während die AUPR anstieg. [84] Diese Veränderungen waren als Folge einer Verschiebung ihrer Zielvariablen, der Einweisungsrate, verursacht, die während der Pandemie deutlich anstieg. Im Vergleich dazu zeigte die Zielvariable in unserer Studie, die Mortalität, nur Schwankungen zwischen 0,8 und 1,0 %.

In der oben genannten Studie an Notaufnahmedaten wurden auch Veränderungen der Importance der einzelnen Variablen festgestellt: so nahm die Wichtigkeit der Atemfrequenz zu Beginn des Lockdowns zu, im weiteren Verlauf wieder ab.

Was die Bedeutung der Merkmale in unseren XGBoost-Modellen aus drei verschiedenen Zeiträumen betrifft, so finden sich dort die gleichen Variablen unter den wichtigsten Einflussfaktoren, allerdings in einer anderen Reihenfolge. Diese wichtigen Einflussfaktoren sind auch klinisch plausibel: Die Anzahl der präoperativen Konsile spiegelt die Komorbiditäten eines Patienten wider, die ebenso wie der ASA-Score mit der Sterblichkeit korrelieren [85], während die Anzahl der transfundierten Erythrozytenkonzentrate mit der Schwere der Operation in Zusammenhang steht.

In der vorliegenden Studie untersuchten wir auch die Fragestellung, ob ein ML Modell aktualisiert werden muss, wenn sich die Begleitumstände ändern, um eine Verschlechterung der Vorhersagegüte zu verhindern. Die Covid -19-Pandemie wurde zur Beantwortung dieser Forschungsfrage beispielhaft herangezogen, da die damit verbundenen Änderungen des Patientenkollektivs mit einer Verschiebung hin zu älteren und vorerkrankten Patienten einen sogenannten Covariate-Shift verursachte, also eine Änderung der Wahrscheinlichkeitsverteilung der Variablen zwischen Trainings- und Testdatensatz auf der einen und Validierungsdatensatz auf der anderen Seite[86].

In einigen Studien wurde beobachtet, dass die Zuverlässigkeit von ML-Modellen über die Zeit abnehmen kann, so dass eine Anpassung an veränderliche Patientenkohorten erforderlich sein könnte [87]. Dieses Problem ist nicht neu und bereits auch in der Medizin bekannt. Es handelt sich häufig um schleichende Veränderungen, die sich langsam entwickeln. Aus den Wirtschaftswissenschaften sind gängige Verfahren bekannt, zum Beispiel das Entfernen oder eine schwächere Gewichtung alter Daten [88]. Der Umgang mit abrupten Veränderungen, wie sie beispielsweise durch die Covid-Pandemie verursacht wurde, ist vor allem im medizinischen Kontext kaum untersucht.

Unsere Ergebnisse deuten darauf hin, dass sich ein Modell nach der Anfangsphase einer akuten Veränderung verschlechtern kann. Die Ursache dafür kann in einer veränderten Häufigkeit des Endpunktes, einem Shift der Covariaten oder einer Änderung der Beziehung zwischen Endpunkt und Covariaten liegen [86]. In unserem Fall ist die Verschlechterung am ehesten durch den Covariate-Shift verursacht. Aktualisiert man die Datenbasis und normalisieren sich dann die Begleitumstände wieder, so kann eine zu frühe Aktualisierung eines Modells zu einer

spürbaren Verschlechterung der Vorhersage führen. In unserer Studie änderte sich der Endpunkt, die Mortalität, nicht wesentlich (Häufigkeit 0.8-1.0%), aber die Variablen. Diese Änderung fand nicht allmählich statt, sondern plötzlich, und wirkte sich trotz des relativ geringen Anteils an Patienten signifikant auf das Modell aus. Ein zu frühes Re-training mit Erweiterung der präpandemischen Daten durch Daten aus der ersten Pandemiewelle führte dann prompt zu einer Verschlechterung der Vorhersage nach der ersten Welle.

Die Ergebnisse der vorliegenden Studie weisen darauf hin, dass ML-Modelle im medizinischen Kontext, speziell auch XG-Boost, zwar eine gute Vorhersagekraft ausweisen, jedoch anfällig sind gegenüber Änderungen der klinischen Rahmenbedingungen, die einen Covariate-Shift verursachen. Dies muss insbesondere im Rahmen von Ausnahmesituationen wie Covid -19 berücksichtigt werden, weil gerade da wichtig ist, dass diese Vorhersagen funktionieren, um eine optimale Intensivbetten- und OP-Planung bei knappen Ressourcen zu gewährleisten

Grundsätzlich ist der Vorteil eines ML-Algorithmus gegenüber herkömmlichen Vorhersagescores die Möglichkeit, eine Vielzahl von klinisch relevanten Variablen gleichzeitig zu identifizieren und zu bewerten. Die Covid-19 Pandemie war allerdings eine Belastungssituation für solche klinischen Vorhersagemodelle. Aufgrund fehlender Studien gibt es kein Patentrezept für den Umgang mit derartigen Umständen. Es erscheint aber praktikabel, in Phasen rascher Veränderungen Modelle zu re-trainieren und dann sowohl das alte als auch das aktualisierte Modell nebeneinander anzuwenden und zu vergleichen. Die Vorhersagen beider Modelle müssen engmaschig überwacht und kritisch bewertet werden.

5.4 Limitationen

Die vorliegende Studie weist einige Limitationen auf, die im Folgenden adressiert werden sollen. Eine erste Limitation ist die Tatsache, dass für die Zeit vor der Pandemie deutlich mehr Patientendaten zur Verfügung standen als während der ersten Pandemiewelle, nach der ersten Pandemiewelle, während der zweiten Pandemiewelle und nach der zweiten Pandemiewelle. Dies könnte die Aussagekraft der Daten verzerren, da > 100000 Patienten für die Situation vor der Pandemie verfügbar waren, während es maximal 12000 für die Situation nach Beginn der Pandemie waren. Für eine Validierung sollten diese Zahlen jedoch ausreichend sein.

Die Zeiträume, für die die Modelle entwickelt wurden, waren unterschiedlich lang. Die Patientenzahlen während der Pandemie beziehungsweise ab der ersten Pandemiewelle waren zwar insgesamt geringer als vor der Pandemie, jedoch wurden jeweils immer noch mehrere tausend Patienten in jedem Zeitraum analysiert. Die Daten während der ersten Welle reichten

immerhin aus, um die AUROC von 0,951 (nur vor der Pandemie) auf 0,923 (vor der Pandemie und während der ersten Welle) zu reduzieren. Es lässt sich daraus schlussfolgern, dass insbesondere die Daten aus der ersten Welle die Vorhersagekraft des Modells beeinträchtigt haben.

Zweitens muss bei der Auswertung von AUROCs berücksichtigt werden, dass sie zwar eingesetzt werden können, um die Leistung von binären Klassifikatoren zu beurteilen, aber auch irreführend sein können, wenn mit unausgewogenen Datensätzen gearbeitet wird, da sie die Prävalenz nicht berücksichtigen. So ist bei unseren Modellen die AUROC meist hervorragend mit Werten >0,9. Die Anwendung der Precision-Recall-Plots offenbart dann die Schwäche der Modelle: Die hohe Rate an Falsch Positiven. Das bedeutet, dass Patienten, für die ein niedriges Mortalitätsrisiko vorhergesagt wird, nahezu immer überleben, viele Patienten mit einem vorhergesagten hohen Mortalitätsrisiko überleben aber auch.

Drittens basieren die Ergebnisse der vorliegenden Studie auf den Daten einer einzelnen Klinik. Sie umfassen zwar einen Zeitraum von mehreren Jahren, jedoch ist durch derartige monozentrische Studien die Übertragbarkeit auf die Situation in anderen Zentren nicht gewährleistet, so dass im Idealfall die Modelle mit den Daten zusätzlicher Kliniken oder mit anderen veränderlichen Begleitumständen optimiert werden sollten. Hierbei stellt sich das Problem einer sehr aufwändigen Datenvorverarbeitung, der unterschiedlichen Datenerfassungs- und Datenverarbeitungssystemen in den Kliniken, sowie eine immer noch umfassende papiergebundene Dokumentation von Patientendaten, so dass eine parallele Auswertung der Daten unterschiedlicher Einrichtungen in einem großen Datenumfang kaum machbar ist.

Eine weitere Limitation dieser Studie ist die Beschränkung auf präoperative Daten. Als mögliche Verbesserung der Vorhersagegenauigkeit der Modelle für zukünftige Studien könnte der Einbezug intraoperativer Daten hilfreich sein. Der Einbezug solcher Daten könnte dazu beitragen, die Vorhersagegenauigkeit vor allem bezüglich des Precision-Recall-Tradeoffs zu erhöhen, indem beispielsweise der intraoperative Blutdruck, die Dauer des Eingriffs und der Blutverlust mitberücksichtigt werden.

6 Zusammenfassung

Machine-Learning-Methoden erhalten zunehmend Einzug in die medizinische Versorgung. Für die Diagnostik und Klinik ist es gleichermaßen wichtig, zuverlässige Vorhersagen zu treffen um die Patienten kategorisieren und angemessen behandeln zu können. XGBoost-Modelle werden im medizinischen Kontext häufig verwendet. Ihre Zuverlässigkeit im Rahmen der Covid -19-Pandemie wurde jedoch noch nicht ausreichend untersucht.

In der vorliegenden Studie kamen daher drei verschiedene XG-Boost-Modelle, die auf Datensätzen aus verschiedenen Zeiträumen trainiert wurden zum Einsatz, um den Einfluss der Covid -19-Pandemie auf die Vorhersagekraft derartiger Modelle zu evaluieren.

Diese erstellten Modelle können die perioperative Mortalität besser vorhersagen als herkömmliche Scores wie der ASA-Score.

Seit Beginn der Covid-19-Pandemie kam es zu veränderten gesetzlichen Richtlinien hinsichtlich der Patientenversorgung, so dass sich nicht nur das Patientengut, sondern auch die Art der Eingriffe verschob hin zu Patienten mit schwereren Vorerkrankungen und einem höheren Anteil an Notfalleingriffen. Entsprechend änderte sich die Wichtigkeit der im Modell verwendeten Parameter: Beispielsweise nahm die Bedeutung des ASA-Scores zu.

Dieser durch die Covid-Pandemie verursachte Covariate-Shift führt zu einer Beeinträchtigung der Vorhersagegenauigkeit Dies wurde deutlich in einem Abfall der AUROC und, noch ausgeprägter, in einem Abfall der AUPRC. Ein mit präpandemischen Daten trainiertes Modell verschlechterte sich deutlich in der ersten Welle. Ein Re-Training des präpandemischen Modells mit zusätzlichen Daten aus der ersten Welle führte zu einer Verschlechterung zwischen der ersten und zweiten Welle. Im Verlauf der Pandemie mit zunehmender Normalisierung der Variablen verbesserte sich die Vorhersagegenauigkeit dann wieder.

Die vorliegende Studie verdeutlicht die Eignung XGBoost-basierter ML-Modelle für die Kategorisierung von Patienten, beispielsweise im Rahmen einer Pandemie. Ein Einsatz derartiger Modelle in der Routineversorgung ist derzeit jedoch noch nicht absehbar.

Bei der Anwendung von ML-Modellen sollte man sich der Tatsache bewusst sein, dass ein einmal erstelltes Modell an Vorhersagegüte verlieren kann, wenn sich die äußeren Umstände ändern. Daher ist eine engmaschige Überwachung und Überprüfung derartiger Modelle unabdingbar.

7 Literaturverzeichnis

1. Amisha, et al., *Overview of artificial intelligence in medicine.* Journal of family medicine and primary care, 2019. **8**(7): p. 2328-2331.
2. Mintz, Y. and R. Brodie, *Introduction to artificial intelligence in medicine.* Minim Invasive Ther Allied Technol, 2019. **28**(2): p. 73-81.
3. Mathur, P., et al., *Artificial Intelligence in Healthcare: 2020 Year in Review*. 2021.
4. Beam, A.L. and I.S. Kohane, *Big Data and Machine Learning in Health Care.* JAMA, 2018. **319**(13): p. 1317-1318.
5. Clifton, D.A., et al., *Health Informatics via Machine Learning for the Clinical Management of Patients.* Yearb Med Inform, 2015. **10**(1): p. 38-43.
6. Sidey-Gibbons, J.A.M. and C.J. Sidey-Gibbons, *Machine learning in medicine: a practical introduction.* BMC Medical Research Methodology, 2019. **19**(1): p. 64.
7. Rajkomar, A., J. Dean, and I. Kohane, *Machine Learning in Medicine.* New England Journal of Medicine, 2019. **380**(14): p. 1347-1358.
8. Maulud, D. and A.M. Abdulazeez, *A Review on Linear Regression Comprehensive in Machine Learning.* Journal of Applied Science and Technology Trends, 2020. **1**(4): p. 140-147.
9. Deo, R.C., *Machine Learning in Medicine.* Circulation, 2015. **132**(20): p. 1920-1930.
10. Kagerbauer, S., et al., *Tomorrow is already here.* Anaesthesia Intensivmed, 2020. **61**: p. 85-94.
11. Ferdowsy, F., et al., *A machine learning approach for obesity risk prediction.* Current Research in Behavioral Sciences, 2021. **2**: p. 100053.
12. Khanam, J.J. and S.Y. Foo, *A comparison of machine learning algorithms for diabetes prediction.* ICT Express, 2021. **7**(4): p. 432-439.
13. Liu, C.Y., et al., *Mortality Evaluation and Life Expectancy Prediction of Patients with Hepatocellular Carcinoma with Data Mining.* Healthcare, 2023. **11**(6).
14. Roche-Lima, A., et al., *Machine Learning Algorithm for Predicting Warfarin Dose in Caribbean Hispanics Using Pharmacogenetic Data.* Frontiers in Pharmacology, 2019. **10**: p. 1550.
15. Ezugwu, A.E., et al., *A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects.* Engineering Applications of Artificial Intelligence, 2022. **110**(104743).
16. Kotsiantis, S.B., *Supervised machine learning: A review of classification techniques.* Informatica, 2007. **31**: p. 249-268.
17. Tape, T., *Using the Receiver Operating Characteristic (ROC) curve to analyze a classification model.* Univ. Nebraska Med. Cent, 2000: p. 1-3.
18. Fawcett, T., *An introduction to ROC analysis.* Pattern Recognition Letters, 2006. **27**(8): p. 861-874.
19. Keilwagen, J., I. Grosse, and J. Grau, *Area under precision-recall curves for weighted and unweighted data.* PloS one, 2014. **9**(3): p. e92209-e92209.
20. Cook, J. and V. Ramadas, *When to consult precision-recall curves.* The Stata Journal, 2020. **20**(1): p. 131-148.
21. Agarwal, N. *The Ultimate Guide To Different Word Embedding Techniques In NLP*. 06.11.2022]; Available from: https://www.kdnuggets.com/2019/12/ultimate-guide-model-retraining.html.
22. Dral, E. and E. Samuylova. *To retrain, or not to retrain? Let's get analytical about ML model updates*. 18.10.2022; Available from: https://www.quora.com/Should-a-machine-learning-model-be-retrained-each-time-new-observations-are-available.

23. Vayena, E., A. Blasimme, and I.G. Cohen, *Machine learning in medicine: Addressing ethical challenges.* PLOS Medicine, 2018. **15**(11): p. e1002689.
24. Pinto Dos Santos, D., et al., *Medical students' attitude towards artificial intelligence: a multicentre survey.* Eur Radiol, 2019. **29**(4): p. 1640-1646.
25. Sonar, A. and K. Weber, *KI gestern und heute: Einsichten aus der Frühgeschichte der KI für aktuelle ethische Überlegungen zum Einsatz von KI in der Medizin.* Arbeit, 2020. **29**(2): p. 105-122.
26. Murdoch, B., *Privacy and artificial intelligence: challenges for protecting health information in a new era.* BMC Med Ethics, 2021. **22**(1): p. 122.
27. *Künstliche Intelligenz: Ethikrat empfiehlt strenge Vorgaben in der Medizin*. 2023 07.05.2023]; Available from: https://www.aerzteblatt.de/nachrichten/141824/Kuenstliche-Intelligenz-Ethikrat-empfiehlt-strenge-Vorgaben-in-der-Medizin.
28. O'Sullivan, S., et al., *Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery.* Int J Med Robot, 2019. **15**(1): p. e1968.
29. Vokinger, K., et al., *Artificial Intelligence und Machine Learning in der Medizin: eine medizinische und rechtliche Würdigung am Beispiel der Radiologie.* Jusletter, 2017.
30. Helle, K., *Intelligente Medizinprodukte: Ist der geltende Rechtsrahmen noch aktuell?* Medizinrecht, 2020. **38**(12): p. 993-1000.
31. Nepogodiev, D., et al., *Global burden of postoperative death.* Lancet, 2019. **393**(10170): p. 401.
32. Ahmad, T., et al., *Use of failure-to-rescue to identify international variation in postoperative care in low-, middle- and high-income countries: a 7-day cohort study of elective surgery.* Br J Anaesth, 2017. **119**(2): p. 258-266.
33. Pearse, R.M., et al., *Mortality after surgery in Europe: a 7 day cohort study.* Lancet, 2012. **380**(9847): p. 1059-65.
34. ISOS, *Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries.* Br J Anaesth, 2016. **117**(5): p. 601-609.
35. Thevathasan, T., et al., *The Impact of Postoperative Intensive Care Unit Admission on Postoperative Hospital Length of Stay and Costs: A Prespecified Propensity-Matched Cohort Study.* Anesth Analg, 2019. **129**(3): p. 753-761.
36. Liang, W., et al., *Early triage of critically ill COVID-19 patients using deep learning.* Nature Communications, 2020. **11**(1): p. 3543.
37. Scott, S., et al., *An evaluation of POSSUM and P-POSSUM scoring in predicting post-operative mortality in a level 1 critical care setting.* BMC Anesthesiology, 2014. **14**(1): p. 104.
38. Merad, F., et al., *Prospective evaluation of in-hospital mortality with the P-POSSUM scoring system in patients undergoing major digestive surgery.* World J Surg, 2012. **36**(10): p. 2320-7.
39. Mayhew, D., V. Mendonca, and B.V.S. Murthy, *A review of ASA physical status – historical perspectives and modern developments.* Anaesthesia, 2019. **74**(3): p. 373-379.
40. Sankar, A., et al., *Reliability of the American Society of Anesthesiologists physical status scale in clinical practice.* BJA: British Journal of Anaesthesia, 2014. **113**(3): p. 424-432.
41. Solanki, S.L., et al., *Artificial intelligence in perioperative management of major gastrointestinal surgeries.* World J Gastroenterol, 2021. **27**(21): p. 2758-2770.
42. Goodfellow, I., B. Yoshua, and A. Courville, *Machine learning basics.* Deep learning 1.7, 2016: p. 98-164.

43. Liu, Y., Y. Wang, and J. Zhang. *New machine learning algorithm: Random forest*. in *International Conference on Information Computing and Applications*. 2012. Springer.
44. Zimmerman, N., et al., *A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring.* Atmospheric Measurement Techniques, 2018. **11**(1): p. 291-313.
45. Natekin, A. and A. Knoll, *Gradient boosting machines, a tutorial.* Front Neurorobot, 2013. **7**: p. 21.
46. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 785–794.
47. Guolin Ke, et al., *LightGBM: A Highly Efficient Gradient Boosting Decision Tree.* Advances in neural information processing systems 2017. **30**: p. 3146-3154.
48. Bennett, J. and S. Lanning, *The Netflix Prize.* Proceedings of KDD cup and workshop, 2007. **2007**: p. 35.
49. Chen, T., et al., *"Xgboost: extreme gradient boosting." R package version 0.4-2 1.4.* 2015: p. 1-4.
50. Wang, Y., et al., *A hybrid ensemble method for pulsar candidate classification.* Astrophysics and Space Science, 2019. **364**(8).
51. COVID-19-Dashboard, R.K.-I. 2020; Available from: https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4/page/page_1/.
52. Andonov, D.I., et al., *Impact of the Covid-19 pandemic on the performance of machine learning algorithms for predicting perioperative mortality.* BMC Med Inform Decis Mak, 2023. **23**(1): p. 67.
53. Zawinell, A. and K. Niepraschk-von Dollen. *ATC-Klassifikation für den deutschen Arzneimittelmarkt*. 05.2022; Available from: https://www.wido.de/publikationen-produkte/arzneimittel-klassifikation/.
54. Blumrich, W., et al., *Kerndatensatz Anästhesie Version 3.0 / 2010.* Anästh. Intensivmed., 2010. **51**: p. Anästh. Intensivmed. 51 (2010) S33 - S55.
55. Witte, R. *Bayessche Optimierung: Theoretische Grundlagen der Hyperparameteroptimierung*. Available from: https://robinwitte.com/wp-content/uploads/2019/11/BayesscheOptimierung.pdf.
56. Hoyer, A. and A. Zapf, *Studien zur Evaluation diagnostischer Verfahren.* Deutsches Ärzteblatt, 2021. **118**(33-34): p. 555-560.
57. Mitchell, T.M., *Machine learning*. Vol. 1. 2007: McGraw-Hill New York.
58. Mehta, A., et al., *Elective surgeries during and after the COVID-19 pandemic: Case burden and physician shortage concerns.* Ann Med Surg (Lond), 2022. **81**: p. 104395.
59. Le Manach, Y., et al., *Preoperative Score to Predict Postoperative Mortality (POSPOM): Derivation and Validation.* Anesthesiology, 2016. **124**(3): p. 570-9.
60. Moll, M., et al., *Machine Learning and Prediction of All-Cause Mortality in COPD.* Chest, 2020. **158**(3): p. 952-964.
61. Kijpaisalratana, N., et al., *Machine learning algorithms for early sepsis detection in the emergency department: A retrospective study.* Int J Med Inform, 2022. **160**: p. 104689.
62. Desautels, T., et al., *Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach.* JMIR Med Inform, 2016. **4**(3): p. e28.
63. Taylor, R.A., et al., *Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach.* Acad Emerg Med, 2016. **23**(3): p. 269-78.

64. Moreno, R.P., R. Pearse, and A. Rhodes, *American Society of Anesthesiologists Score: still useful after 60 years? Results of the EuSOS Study.* Rev Bras Ter Intensiva, 2015. **27**(2): p. 105-12.
65. Yurtlu, D.A., et al., *Comparison of Risk Scoring Systems to Predict the Outcome in ASA-PS V Patients Undergoing Surgery: A Retrospective Cohort Study.* Medicine (Baltimore), 2016. **95**(13): p. e3238.
66. Clark, P.A. and C.J. Lettieri, *Clinical model for predicting prolonged mechanical ventilation.* J Crit Care, 2013. **28**(5): p. 880.e1-7.
67. Chiew, C.J., et al., *Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission.* Ann Surg, 2020. **272**(6): p. 1133-1139.
68. Mohamadlou, H., et al., *Multicenter validation of a machine-learning algorithm for 48-h all-cause mortality prediction.* Health Informatics J, 2020. **26**(3): p. 1912-1925.
69. Mori, M., et al., *Toward Dynamic Risk Prediction of Outcomes After Coronary Artery Bypass Graft: Improving Risk Prediction With Intraoperative Events Using Gradient Boosting.* Circ Cardiovasc Qual Outcomes, 2021. **14**(6): p. e007363.
70. Chen, C., et al., *Development and performance assessment of novel machine learning models to predict pneumonia after liver transplantation.* Respir Res, 2021. **22**(1): p. 94.
71. Bolourani, S., et al., *A Machine Learning Prediction Model of Respiratory Failure Within 48 Hours of Patient Admission for COVID-19: Model Development and Validation.* J Med Internet Res, 2021. **23**(2): p. e24246.
72. Vaid, A., et al., *Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients With COVID-19 in New York City: Model Development and Validation.* J Med Internet Res, 2020. **22**(11): p. e24018.
73. Hong, W., et al., *A Comparison of XGBoost, Random Forest, and Nomograph for the Prediction of Disease Severity in Patients With COVID-19 Pneumonia: Implications of Cytokine and Immune Cell Profile.* Frontiers in Cellular and Infection Microbiology, 2022. **12**.
74. Li, W.T., et al., *Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis.* BMC Medical Informatics and Decision Making, 2020. **20**: p. 247.
75. Saito, T. and M. Rehmsmeier, *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.* PLOS ONE, 2015. **10**(3): p. e0118432.
76. Martin, B.I., et al., *The Impact of Halting Elective Admissions in Anticipation of a Demand Surge Due to the Coronavirus Pandemic (COVID-19).* Med Care, 2021. **59**(3): p. 213-219.
77. Frio, G.S., et al., *The disruption of elective procedures due to COVID-19 in Brazil in 2020.* Scientific Reports, 2022. **12**(1): p. 10942.
78. Best, M.J., et al., *The likely economic impact of fewer elective surgical procedures on US hospitals during the COVID-19 pandemic.* Surgery, 2020. **168**(5): p. 962-967.
79. Balvardi, S., et al., *Impact of the Covid-19 pandemic on rates of emergency department utilization and hospital admission due to general surgery conditions.* Surgical Endoscopy, 2022. **36**(9): p. 6751-6759.
80. Sá, A.F., et al., *Urgent/emergency surgery during COVID-19 state of emergency in Portugal: a retrospective and observational study.* Brazilian Journal of Anesthesiology (English Edition), 2021. **71**(2): p. 123-128.
81. Tebala, G.D., et al., *Emergency surgery admissions and the COVID-19 pandemic: did the first wave really change our practice? Results of an ACOI/WSES international*

retrospective cohort audit on 6263 patients. World Journal of Emergency Surgery, 2022. **17**(1): p. 8.
82. Karlafti, E., et al., *Emergency General Surgery and COVID-19 Pandemic: Are There Any Changes? A Scoping Review.* Medicina, 2022. **58**(9): p. 1197.
83. Lane, D.J., et al., *Changes in presentation, presenting severity and disposition among patients accessing emergency services during the first months of the COVID-19 pandemic in Calgary, Alberta: a descriptive study.* CMAJ Open, 2021. **9**(2): p. E592-e601.
84. Duckworth, C., et al., *Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19.* Sci Rep, 2021. **11**(1): p. 23017.
85. Hackett, N.J., et al., *ASA class is a reliable independent predictor of medical complications and mortality following surgery.* Int J Surg, 2015. **18**: p. 184-90.
86. Jain, S. *Covariate Shift – Unearthing hidden problems in Real World Data Science*. 2017 25.07.2022 26.02.2023]; Available from: https://www.analyticsvidhya.com/blog/2017/07/covariate-shift-the-hidden-problem-of-real-world-data-science/.
87. Vela, D., et al., *Temporal quality degradation in AI models.* Scientific Reports, 2022. **12**(1): p. 11654.
88. Tsymbal, A., *The problem of concept drift: definitions and related work.* Computer Science Department, Trinity College Dublin, 2004. **106**(2): p. 58.

8 Danksagung

Ich möchte mich an dieser Stelle bei allen bedanken, die mich bei der Anfertigung meiner Dissertation unterstützt und begleitet haben.

Mein besonderer Dank gilt Herrn Professor Dr. Blobner, der mir die Möglichkeit gegeben hat, die Arbeit in seiner renommierten Forschungsgruppe durchzuführen.

Ebenso gilt mein herzlicher Dank meiner Betreuerin PD Dr. med. Kagerbauer, für ihre Geduld und Zeit, für die zahlreichen konstruktiven Kommentare und hilfreichen Ideen, sowie für ihre ständige Hilfsbereitschaft während der Fertigstellung dieser Arbeit. Für die fachkundige Anleitung, die endlosen Diskussionen und die wertvollen Anregungen ein herzliches Dankeschön.

Weiterhin möchte ich mich speziell bei Bernhard Ulm bedanken, der eine großartige Vorarbeit geleistet hat und eine große Unterstützung bei der Auswertung und statistischen Analyse der Anästhesiedaten war.

Last but not least möchte ich mich bei meiner Familie und meinen Freunden bedanken, die mich immer in einzigartiger Weise und liebevoll unterstützt haben.

9 Veröffentlichungen

Teile dieser Dissertation wurden bereits in folgendem Artikel und Kongressbeitrag veröffentlicht:

- Andonov, D.I., et al., *Impact of the Covid-19 pandemic on the performance of machine learning algorithms for predicting perioperative mortality.* BMC Medical Informatics and Decision Making, 2023. **23(1)**: p. 1-12.
- Kagerbauer S. M., et al., *Machine-learning-based algorithm for prediction of postoperative death of non-ICU patients.* Anesthesia and Analgesia, 2020. **130**: p. 821-822.