



AUS DEM NETZWERK EVIDENZBASIERTE MEDIZIN

Voraussetzung für die Nutzung von (KI-basierten) Prognosemodellen in der Praxis

VON MAXI SCHULZ, MSC, UND PROF. DR. RER. MEDIC. TIM MATHES IM AUFTRAG DES NETZWERKS
EVIDENZBASIERTE MEDIZIN E.V. (WWW.EBM-NETZWERK.DE)

P

Prädiktionsmodelle umfassen ein großes Forschungsgebiet im Gesundheitswesen. Diese Modelle analysieren Daten, um die Wahrscheinlichkeit bestimmter Gesundheitsausgänge oder Ereignisse für Patientinnen und Patienten abzuschätzen. Sie können somit ein wertvolles Instrument in der personalisierten medizinischen Entscheidungsfindung sein, häufig werden sie auch als klinische (elektronische) Entscheidungsunterstützungssysteme bezeichnet. Man unterscheidet zwischen zwei Arten von Prädiktionsmodellen: Diagnosemodelle, die dabei helfen, das Vorliegen einer Krankheit zu bestimmen, und Prognosemodelle, die zukünftige Gesundheitsentwicklungen vorhersagen. In diesem Artikel konzentrieren wir uns auf Prognosemodelle, die Ärztinnen und Ärzten bei der Vorhersage von Krankheitsverläufen und -ausgängen unterstützen können.

Für viele medizinische Anwendungsbereiche gibt es bereits eine Vielzahl an verschiedenen Prognosemodellen. Beispiele für erfolgreiche Prognosemodelle sind der Framingham Risk Score, der das Risiko einer Herz-Kreislauf-Erkrankung vorhersagt,[1] und der EuroSCORE, der das Sterblichkeitsrisiko nach einer Herzoperation quantifiziert.[2] Diese Modelle sind etablierte Instrumente in der Praxis und teilweise auch als elektronische Version verfügbar, wie zum Beispiel der webbasierte Rechner auf der EuroSCORE-Website.[3]

Trotz der Fülle an existierenden Modellen konzentrieren sich Forschungsgruppen oft auf die Entwicklung neuer Modelle. So gibt es allein für den Bereich der Herz-Kreislauf-Erkrankungen über 360 Modelle (Stand 2022).[4] Was häufig fehlt, ist die Überlegung, wie diese Modelle in die klinische Praxis überführt werden können. Dies legen auch die Ergebnisse systematischer Reviews nahe: Sie zeigen, dass eine Vielzahl der entwickelten Modelle erhebliche Mängel hinsichtlich Transparenz, Reproduzierbarkeit, Übertragbarkeit und ethischen Fragen aufweisen, und kommen zu dem Schluss, dass erhebliche Bedenken für ihre Anwendung in der klinischen Praxis bestehen.[4,5,6] Um erfolgreich in die Praxis überführt zu werden, müssen Prognosemodelle daher bestimmte Gütekriterien erfüllen.

Nach einer kurzen Einleitung zu Prognosemodellen gibt dieser Artikel einen Überblick über die Kriterien, die ein Prognosemodell erfüllen muss, um als relevant und vertrauenswürdig angesehen zu werden und die

somit vor der Anwendung im klinischen Alltag betrachtet werden sollten. Im Anschluss werden Aspekte für die erfolgreiche Implementierung eines Prognosemodells betrachtet. Abschließend wird kurz beleuchtet, was insbesondere bei KI-basierten Prognosemodellen zu beachten ist.

EINFÜHRUNG PROGNOSEMODELLE

Prognosemodelle zielen darauf ab, zukünftige Gesundheitsentwicklungen von Patientinnen und Patienten vorherzusagen, zum Beispiel den Ausgang einer Krankheit oder die Wahrscheinlichkeit, dass eine Krankheit auftritt. Das Ergebnis, das mittels eines Prognosemodells prognostiziert werden soll, kann binär (z.B. Tod, Genesung), multinomial (z.B. Art der Nebenwirkung, Lebensqualität) oder kontinuierlich (Blutdruck, Anzahl von schmerzempfindlichen Gelenken) sein. Bei binären Ergebnissen können die Vorhersagen als Wahrscheinlichkeiten (auch Risiken) ausgedrückt werden. Kontinuierliche Ergebnisse werden als Erwartungswerte dargestellt. Prognosemodelle verwenden Algorithmen, um diese Ergebnisse vorherzusagen. Zur Vorhersage ziehen sie prognostische Faktoren heran, also Variablen, deren Werte mit einer Veränderung des Risikos oder des Erwartungswertes assoziiert sind, wie zum Beispiel die Blutdruckwerte oder das Alter. Zur Entwicklung solcher Prognosemodelle können sowohl traditionelle statistische Methoden, wie zum Beispiel Regressionsanalysen, als auch moderne Verfahren der künstlichen Intelligenz

und des maschinellen Lernens zur Anwendung kommen.

KRITERIEN DER VERTRAUENSWÜRDIGKEIT VON PROGNOSEMODELLEN

RELEVANZ DES MODELLS: Ein Prognosemodell sollte eine für den klinischen Alltag relevante Fragestellung beantworten und sich nahtlos in den klinischen Entscheidungsprozess einfügen. Daher ist es von zentraler Bedeutung, bei der Entwicklung neuer Prognosemodelle von Anfang an die Relevanz des Modells und den Zeitpunkt, wann es in der Versorgung angewendet werden soll, zu definieren. Es sollte hinterfragt werden, warum ein neues Vorhersagemodell benötigt wird und welchen Mehrwert es liefert. Oftmals existieren in der Praxis bereits gut funktionierende Mechanismen, die zuverlässige Vorhersagen liefern oder es existieren bereits ähnliche Modelle für denselben Anwendungsfall, die eine vergleichbare Vorhersagegenauigkeit aufweisen, und somit kein neues Modell erforderlich machen. Ein Prognosemodell kann zudem nur dann relevant sein, wenn für die meisten Variablen ein sinnvoller kausaler Zusammenhang zwischen den prognostischen Faktoren und dem vorherzusagenden Ergebnis besteht. Zudem müssen die prognostischen Faktoren zum Zeitpunkt der Vorhersage verfügbar sein. Häufig werden bei der Modellentwicklung Faktoren berücksichtigt, die erst nach dem Auftreten des interessierenden Ergebnisses erhoben werden, was die Anwendbarkeit des Modells im Versorgungsalltag stark reduziert oder sogar unmög-



lich macht, da fehlende Variablen die Vorhersagegenauigkeit erheblich beeinträchtigen.

Ein weiterer wichtiger Aspekt ist die Verfügbarkeit des Modell-Ergebnisses für den Entscheidungsprozess. Es muss geklärt werden, wann das Ergebnis des Modells benötigt wird und welche Anforderungen die Anwenderinnen und Anwender an ein solches Unterstützungstool haben. Idealerweise sollte gemeinsam mit den Anwenderinnen und Anwendern sowie den Personen, die die Vorhersage betrifft (in der Regel Patientinnen und Patienten), gemeinsam erörtert werden, welche Anforderungen ein Prognosemodell erfüllen muss.

REPRÄSENTATIVITÄT UND QUALITÄT DER TRAININGSDATEN: Die Auswahl der Daten, auf deren Basis ein Prognosemodell entwickelt wird (auch Trainingsdaten), ist entscheidend für die Güte des Modells. Ein Modell kann nur so gut sein wie die Daten, auf denen das Modell entwickelt wurde. Diese Trainingsdaten müssen für das Anwendungsszenario, für das das Modell zur Anwendung kommen soll, repräsentativ sein. Wenn dies nicht der Fall ist, kann u. U. die systematische Benachteiligung von Personen bestimmter Gruppen drohen. Um dies zu vermeiden, muss der Trainingsdatensatz die Vielfalt des realen Anwendungsszenarios widerspiegeln. Insbesondere sind Charakteristika wie Geschlecht, Alter, Ethnizität usw. zu berücksichtigen. Ist die Vielfalt der Population in den Trainingsdaten nicht ausreichend widerspiegelt, kann die Vorhersage des Modells für bestimmte Gruppen unzuverlässig sein. Darüber hinaus können Vorurteile, die in der realen Gesundheitsversorgung zu Ungleichheiten führen, in den Daten reflektiert sein. Ein Beispiel: Algorithmen, die

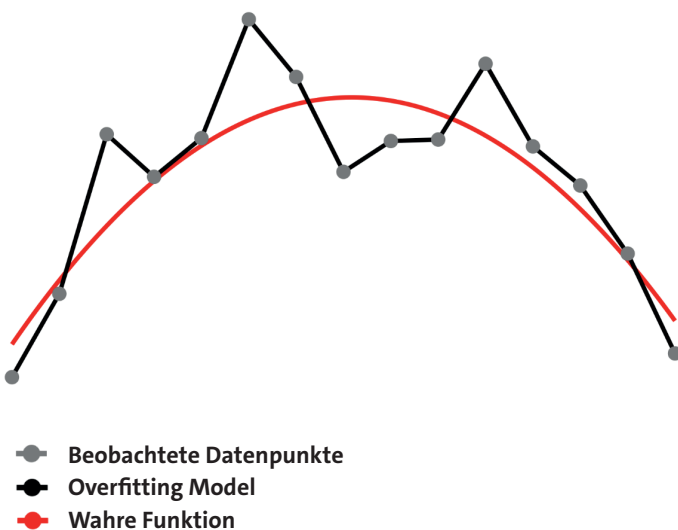
vorhersagen, ob eine Person operiert werden sollte, können auf diejenigen ausgerichtet sein, die sich überhaupt medizinisch beraten lassen. Werden solche Verzerrungen in den Daten bei der Modellerstellung nicht berücksichtigt, können Ungleichheiten durch die Vorhersage verstärkt werden.

Zeitgleich sollten Trainingsdaten ausreichend auf Qualität geprüft werden. Das Auftreten von fehlenden Werten (Missing Data) und warum diese fehlen, kann auf eine systematische Verzerrung in den Daten hinweisen, wenn z.B. Messungen nur für eine bestimmte Subgruppe vorliegen. Auch Fehler in den Messungen können erhebliche Konsequenzen für die Vorhersagegüte eines Modells sein. Wenn z.B. für das interessierende Ergebnis kein einheitlicher Standard existiert (z.B. standardisierte Eingabe), ist das Modell nicht ausreichend in der Lage, eine zuverlässige Prognose zu liefern. Zudem kann sich die Erfassung von Variablen über die Zeit verändern, und muss in der Datenaufbereitung berücksichtigt werden. Weiterhin muss der Datensatz eine ausreichende Fallzahl aufweisen, um robuste Ergebnisse zu liefern.

OVERFITTING: Overfitting, auf Deutsch ‚Überanpassung‘, bezeichnet ein Phänomen, bei dem ein Modell zu stark auf die Trainingsdaten passt. Es passt sich so sehr an die Trainingsdaten an, dass es nicht mehr auf eine externe Kohorte generalisierbar ist. Das bedeutet, dass das Modell zwar während der Trainingsphase eine gute (oder sogar hervorragende) Performance aufweisen kann, aber in der Validierung nicht mehr gut abschneidet. Somit ist es für Anwendungen in verschiedenen klinischen Situationen unbrauchbar. Overfitting kann

verschiedene Gründe haben, zum Beispiel die Verwendung vieler Faktoren bei kleinen Stichproben. Es kann jedoch durch geeignete Maßnahmen erkannt und während der Modellentwicklung verhindert werden. Vorsicht ist geboten, wenn nur bestimmte Ergebnisse berichtet werden, da diese ein Overfitting verschleiern können.

OVERFITTING



Darstellung Overfitting an einem fiktiven Beispiel. Die beobachteten Daten (graue Punkte) folgen einer Funktion (rote Linie). Das Modell (schwarze Linie) errechnet eine Funktion, die durch alle Datenpunkte geht; die wahre zugrundeliegende Funktion kann das Modell jedoch nicht beschreiben. Das Modell wird auf neue Datenpunkte nicht generalisieren. Es kommt zu Overfitting.

(EXTERNE) VALIDIERUNG: Nach der Entwicklung eines Prognosemodells muss dieses validiert werden. Validierung bedeutet, die Vorhersagegüte eines Modells zu beurteilen und zu überprüfen, wie gut es auf verschiedene Settings generalisierbar ist. Wenn ein Modell zu stark auf die Population passt, auf dem es entwickelt wurde (Overfitting), so kann dies in einer Validierung entdeckt werden.

Es gibt zwei Arten der Validierung: interne Validierung und externe Validierung. Bei der internen Validierung wird ein Datensatz verwendet, der aus der gleichen Population und dem gleichen Setting stammt wie der Trainingsdatensatz. Dies kann zum Beispiel ein Datensatz sein, der an Patientinnen und Patienten desselben Klinikums erhoben wurde. Zumeist wird zur internen Validierung bei der Modellentwicklung der gesamte vorliegende Datensatz in zwei Teile geteilt, einen Trainings- und einen Validierungsdatsatz. Die Daten, auf denen das Modell entwickelt wird und auf dem es intern validiert wird, stammen dann aus derselben Population. Im Gegensatz dazu bezieht sich die externe Validierung auf einen Datensatz, der völlig neu ist. Dieser Datensatz kann zeitlich (temporale Validierung) und räumlich (geografische Validierung) von der ursprünglichen Population verschieden sein, beispielsweise Daten aus einem anderen Klinikum. Die externe Validierung kann anhand retrospektiv oder prospektiv erhobener Daten durchgeführt werden. Eine prospektive Validierung hat den Vorteil, dass die Anwendbarkeit des Modells in der jeweiligen Anwendungspraxis besser repräsentiert und zudem potenzielle Anwendungsprobleme aufgedeckt werden können.

Eine Validierung sollte den Vergleich der Vorhersa-



gen des Prognosemodells mit dem aktuellen Standard der Versorgung, also ohne Prognosemodell, umfassen, um den tatsächlichen Nutzen des Vorhersagemodells zu quantifizieren. Dies könnte beispielsweise den Vergleich von Sensitivität und Spezifität der Entscheidungen von Ärztinnen und Ärzten und ohne Entscheidungstool umfassen.

MASSE FÜR DIE VORHERSAGEGENAUIGKEIT: Die beiden wichtigsten Metriken zur Einschätzung der Güte eines Prognosemodells sind Diskriminierung und Kalibrierung.

Diskriminierung bezeichnet die Fähigkeit eines Prognosemodells, zwischen Individuen mit und ohne das vorhergesagte Ergebnis zu unterscheiden. Dies ist relevant bei binären und multinomialen Klassifikationen. Zum Beispiel bei der Bestimmung des Risikos für das Wiederauftreten einer Krankheit zeigt die Diskriminierung, wie gut das Modell zwischen denjenigen unterscheiden kann, bei denen die Krankheit wieder aufgetreten ist, und denen, bei denen sie nicht wieder aufgetreten ist.

Kalibrierung beurteilt, inwiefern die vom Modell berechneten Werte mit den tatsächlichen beobachteten Ergebnissen übereinstimmen. Für das Beispiel der Wiederauftretenden Krankheit bedeutet dies, dass die beobachteten Wahrscheinlichkeiten für das Wiederauftreten der Krankheit mit den vom Modell prognostizierten Wahrscheinlichkeiten verglichen werden. Je besser die beobachteten und vom Modell prognostizierten Wahrscheinlichkeiten übereinstimmen, desto besser ist das Modell kalibriert.

Bei der Betrachtung dieser Metriken sollte nicht nur

auf statistische Signifikanz geachtet werden, sondern vielmehr auf die Schätzung von Konfidenzintervallen. In bestimmten Anwendungsfällen können auch weitere Metriken von Relevanz sein, wie z.B. die Anzahl der falsch klassifizierten Fälle, wenn Über- bzw. Unterdiagnose von besonderer Bedeutung sind.

Ein trainiertes Modell wird immer besser auf den Daten funktionieren, auf denen es trainiert wurde (Trainingsdaten) als auf neuen Daten (siehe Overfitting). Wichtig ist daher, dass die Performance-Metriken für die Test- und/oder Validierungsdaten berichtet werden und keinesfalls ausschließlich auf den Trainingsdaten. Die auf den Test- und/oder Validierungsdaten basierenden Metriken sollten daher ausschlaggebend für die Beurteilung sein, ob ein Modell auch in der Praxis gut funktioniert. Im Idealfall sollten die Metriken auf einem Datensatz einer externen Kohorte überprüft werden (siehe externe Validierung).

REPRODUZIERBARKEIT: Um die Ergebnisse eines Prognosemodells in Validierungsstudien zu untersuchen und das Modell erfolgreich zu implementieren, ist eine transparente und nachvollziehbare Dokumentation der Modellentwicklung und -ergebnisse unerlässlich. Eine sorgfältige Dokumentation gewährleistet die Reproduzierbarkeit der Modellergebnisse und ermöglicht es Forschenden und Nutzen, die Ergebnisse zu überprüfen und das Modell weiter zu verbessern.

IMPLEMENTIERUNG UND INSTANDHALTUNG

INTEGRATION IN BESTEHENDE STRUKTUREN: Für eine erfolgreiche Implementierung muss das Modell zunächst in die bestehenden Systeme und Workflows integriert werden.

GLOSSAR

Ergebnis (Outcome): Das interessierende Ergebnis, für das eine Vorhersage mittels eines Modells getroffen werden soll. Kann binär, multinomial oder kontinuierlich sein.

Prognostischer Faktor: Faktor, der in sinnvoller Assoziation mit dem interessierenden Ergebnis steht.

Prädiktoren / Features: Prognostische Faktoren, die in das Prognosemodell einfließen und die Vorhersage des interessierenden Ergebnisses beeinflussen. Häufige Beispiele sind Alter, Geschlecht und Vorerkrankungen.

Trainingsdatensatz: Daten, auf deren Grundlage das Modell entwickelt und trainiert wird.

Validierungsdatensatz: Daten, auf deren Grundlage die Güte des Modells geprüft wird. Unterscheidung zwischen interner und externer Validierung.

Künstliche Intelligenz (KI): Fähigkeit von Computern, Aufgaben automatisiert auszuführen, die normalerweise von Menschen ausgeführt werden.

Maschinelles Lernen (ML): Methode der Künstlichen Intelligenz, bei der Computer aus Daten lernen, um Aufgaben auszuführen.

Deep Learning: Methode des Maschinellen Lernens, bei der Computer über sukzessive Schichten aus Daten lernen, um Aufgaben auszuführen.

Die Mensch-Modell-Interaktion sollte im Voraus genau definiert werden: Welche Informationen werden den Nutzenden zur Verfügung gestellt? Welche Informationen benötigt das Modell von Nutzerinnen und Nutzern? Es ist wichtig, Sicherheitsrisiken zu prüfen und Strategien zu entwickeln, um diese zu minimieren. Dazu gehören Risiken wie Manipulation und die Re-Identifizierung von Behandlungsinformationen. Darüber hinaus ist es wichtig zu klären, wie das Modell regelmäßig angepasst wird. Wie wird sichergestellt,

dass Modelle aktualisiert werden? Werden neue Daten verwendet, um das Modell regelmäßig zu aktualisieren? Was passiert, wenn sich Standards ändern oder bestimmte Variablen nicht mehr erfasst werden? Auch sollte regelmäßig überprüft werden, ob das Modell zu unerwünschten Vorhersagen führt, und ständige Qualitätskontrollen sollten durchgeführt werden. Dies ermöglicht es, das Modell kontinuierlich zu verbessern und sicherzustellen, dass es den Anforderungen der Nutzenden entspricht.

SCHULUNG DER ANWENDERINNEN UND ANWENDER:

Nicht zuletzt hängt die erfolgreiche Implementierung eines Prognosemodells davon ab, wie Anwenderinnen und Anwender das System nutzen. Dies setzt voraus, dass sie in der Anwendung geschult werden. Ihnen sollte klar sein, wofür das System nutzbar ist, welche möglichen Verzerrungen dem System inhärent sind und wie sie die Entscheidungen nachvollziehbar machen können.

Die Schulung der Anwenderinnen und Anwender sollte folgende Ziele verfolgen:

- Verständnis der Funktionsweise des Prognosemodells
- Kenntnis der Stärken und Schwächen des Modells
- Fähigkeit, die Ergebnisse des Modells zu interpretieren und zu nutzen
- Verständnis der möglichen Verzerrungen und Limitationen des Modells

Spezifika und Herausforderungen für KI-basierte Modelle

In der Landschaft der klinischen Forschung hat die Integration von Methoden des maschinellen Lernens



(ML) und künstlicher Intelligenz (KI) zahlreiche neue Möglichkeiten eröffnet. Während der Begriff KI umgangssprachlich häufig mit generativen Sprachmodellen wie ChatGPT gleichgesetzt wird, bieten KI-basierte Methoden darüber hinaus auch Einsatzmöglichkeiten in der Entwicklung von Prognosemodellen. So erweisen sich KI-basierte Anwendungen in der Vorhersage und Risikobewertung für verschiedene Krebsarten, darunter Brust-, Hirn-, Lungen-, Leber- und Prostatakrebs, als äußerst effektiv und können zum Teil sogar eine größere Genauigkeit bei der Vorhersage erzielen als die Einschätzung von, z.B. Berufsanfängerinnen und Berufsanfängern.[7]

Die oben beschriebenen generellen methodischen Fallstricke treffen, wie bei klassischen Auswertungsmethoden, auch auf KI-basierte Prognosemodelle zu.[8,9] Es gibt jedoch spezifische Herausforderungen, die bei KI-basierten Prognosemodellen zusätzlich von Relevanz sind.[8,9,10,11]

Methoden der Künstlichen Intelligenz und des Maschinellen Lernens stellen im Gegensatz zu klassischen Verfahren wenige Annahmen an die Daten, auf denen sie lernen bzw. trainieren. Dies ermöglicht es, auch nicht nur strukturierte quantitative Daten zu berücksichtigen, sondern auch andere Datentypen, wie Textdaten oder Bilddaten auszuwerten. Dies ist vor allem bei großen Datenmengen von Vorteil.

Diese technischen Vorteile stehen jedoch Einschränkungen in Bezug auf die Interpretierbarkeit und Nachvollziehbarkeit von Ergebnissen gegenüber. Die Transparenz der Entscheidungsfindung bei Prognosemodellen ist entscheidend für das Vertrauen der Anwenderinnen und Anwender, unabhängig von der ver-

wendeten Methode. Bei komplexeren Methoden wird die Vermittlung, wie das Modell zu einer Entscheidung kommt, zunehmend schwieriger. Insbesondere bei Methoden der Künstlichen Intelligenz ist häufig nicht klar, wie diese Modelle zu einem Ergebnis gelangen. Der Algorithmus ist oft so komplex, dass selbst für Entwicklerinnen und Entwickler die Entscheidungsfindung häufig eine „Black Box“ darstellt. Bei KI-basierten Methoden können sehr komplexe Zusammenhänge zwischen verschiedenen Faktoren berechnet werden, die ebenfalls häufig schwer zu vermitteln sind. Für Nutzende kann dies eine immense Hürde in der Anwendung darstellen, da sie dem Ergebnis nicht vertrauen und auch von haftungsrechtlicher Relevanz sein können. Daher sollte in der Entwicklung von KI-basierten Prognosemodellen stets die Bemühung bestehen, die Ergebnisse nachvollziehbar zu machen (Explainable AI).

Diese Nachteile sollten insbesondere auch vor dem Hintergrund betrachtet werden, dass KI-basierte Prädiktionsmodelle oftmals nicht besser sind als traditionelle statistische Modelle. So zeigt ein systematischer Review von Prädiktionsmodellen aus verschiedenen Disziplinen des Gesundheitswesens, dass ML-basierte Prädiktionsmodelle nicht besser funktionieren als regressionsbasierte Modelle.[12] Während dies sicherlich nicht auf alle ML-/KI-basierten Modelle zutreffen wird, bedeutet es jedoch, dass ein Zugewinn von ML/KI-Methoden nicht immer zwingend gegeben ist. Ein einfaches Modell, das nachvollziehbar und leichter zu vermitteln ist, sollte aus genannten Gründen bei gleichwertiger Güte, jedoch immer bevorzugt werden gegenüber einem komplexen Algorithmus.

FAZIT

Systematische Reviews zeigen, dass viele Prognosemodelle methodisch mangelhaft sind und somit ein Risiko für systematische Verzerrungen bergen.[4,5,6] Dabei werden Probleme wie Overfitting, unrepräsentative Daten und unzureichende Validierung häufig bemängelt. Dies gilt für klassische wie für KI-basierte Modelle. Diese mangelhaften Modelle sind i. d. R. für die klini-

sche Praxis nicht nutzbar. Daher sollten Prognosemodelle kritisch beurteilt werden, bevor sie für die Anwendung in der klinischen Praxis in Frage kommen. Zudem gibt es eine Reihe an Dingen zu beachten für die tatsächliche Implementierung, wie die Integration in den bestehenden Workflow, die Qualitätssicherung und Wartung, sowie die Schulung von Anwenderinnen und Anwendern. ■

**MAXI SCHULZ**

MSc, Wissenschaftliche
Mitarbeiterin, Institut für
Medizinische Statistik,
Universitätsmedizin Göttingen,
Humboldtallee 32
37073 Göttingen
maxi.schulz@med.uni-goettingen.de

**PROF. DR. RER. MEDIC.****TIM MATHES**

Gruppenleiter AG
„Klinische Epidemiologie und Ge-
sundheitsökonomie“,
Institut für Medizinische Statistik,
Universitätsmedizin Göttingen,
Humboldtallee 32, 37073 Göttingen

Literatur:

- [1] Framingham Risk Score: <https://doi.org/10.1161/CIRCULATIONAHA.107.699579>
 [2] EuroSCORE: [https://doi.org/10.1016/S0195-668X\(02\)00799-6](https://doi.org/10.1016/S0195-668X(02)00799-6)
 [3] EuroSCORE digital: <https://www.euroscore.org/index.php?id=17>, zuletzt abgerufen am 01.07.2024
 [4] <https://doi.org/10.1093/eurheartj/ehac238>
 [5] <https://doi.org/10.1371/journal.pmed.1001221>
 [6] <https://doi.org/10.1016/j.jclinepi.2012.06.020>
 [7] <https://doi.org/10.1136/bmj.m1328>
 [8] <https://doi.org/10.2147/JMDH.S410301>
 [9] <https://doi.org/10.1136/bmj.l6927>
 [10] <https://doi.org/10.1093/ehjdh/ztob054>
 [11] <https://doi.org/10.1136/bmjhci-2023-100784>
 [12] <https://doi.org/10.1038/s41746-021-00549-7>
 [13] <https://doi.org/10.1016/j.jclinepi.2019.02.004>
 Weitere Literatur: <https://doi.org/10.7326/M19-0872>

